Algorithms for Learning Latent Models : Establishing Tractability to Approaching Optimality

Ainesh Bakshi

CMU-CS-22-146

August 2022

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Pravesh K. Kothari, Co-Chair
David P. Woodruff, Co-Chair
Ryan O'Donnell
Boaz Barak (Harvard)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was sponsored by the Office of Naval Research under award number N000141812562, the Air Force Office of Scientific Research under award number FA870215D0002, and the National Science Foundation under award numbers CCF-1815840 and CCF-2047933. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Latent Models, Robust Statistics, Sum-of-Squares, Numerical Linear Algebra

To my Grandparents,
who woke up as refugees in an independent India
and toiled to provide a better life for their kids.

Abstract

Modern machine learning relies on algorithms that fit expressive latent models to large datasets. While such tasks are easy in low dimensions, real-world datasets are truly high-dimensional, often leading to computational intractability. Additionally, a prerequisite to deploying models in real-world systems is to ensure that their behavior degrades gracefully when the modeling assumptions no longer hold. Therefore, there is a growing need for *efficient algorithms* that fit reliable and robust models to data and are accompanied with provable guarantees.

In this thesis, we focus on designing such efficient and robust algorithms for learning latent variable models. In particular, we investigate two complementary regimes arising in learning latent models: *establishing computational tractability* and *approaching computational optimality*. The first regime considers learning high-dimensional latent models where no efficient algorithms were known. We resolve several central open questions in this regime, by providing the first polynomial time algorithms for robustly learning a mixture of Gaussians, robust linear regression and learning two-layer neural networks. The second regime considers models where polynomial time algorithms were already well-established. Here, we show that we can obtain algorithms with information-theoretically minimal running time and sample complexity. In particular, we show that for several low-rank models there is no *statistical vs. computational* trade-off.

Contents

Acknowledgments

First and foremost, I must acknowledge how grateful I am to my advisors, Pravesh Kothari and David Woodruff, for providing an incredibly enriching academic experience. I could not have asked for a better co-advising duo. David was extremely influential in shaping me as a researcher early on, and was incredibly patient, constantly involved, and always ready to jump into the technical weeds for hours on end. In addition to providing exciting research directions, David was receptive to me coming up with my own questions and research directions. He approached every project we worked on with his characteristic vigor and unwavering support that has made each collaboration a cherished one. Pravesh took me on as a student at a crucial juncture during my PhD, and introduced me to the fascinating world of convex hierarchies. Pravesh's curiosity and breath of knowledge across theoretical computer science sets the bar for young researchers. His infectious enthusiasm and relentless optimism are traits I hope to carry with me for the rest of my career. Finally, I am really grateful to my advisors for constantly believing in me, even when I did not, and giving me the time and space to pursue directions I found interesting.

Next, I want to thank Yury Makarychev and Madhur Tulsiani for an exhilarating summer internship at TTI-Chicago in the summer of 2021. Despite the pandemic, this was an incredible internship, and one that broke the monotony of working from home. Over the course of the summer, I also started working closely with Aravindan Vijayaraghavan and Goutham Rajendran. The weekly Friday night meetings with Madhur, Aravindan, Goutham and Xue became a beacon of positivity during an otherwise tumultuous period. The meetings were filled with unrelenting optimism and camaraderie. Over the course of this internship and the collaborations that came out of it, I got a glimpse of what academic life would actually be like, and I am very grateful for it. I would also like to extend special gratitude towards Madhur and Aravindan, who were always available to answer my questions, to engage with my arcane suggestions, and to host me in Chicago multiple times. I would also like to thank Ken Clarkson for an excellent summer internship at IBM Research in the summer of 2020. Ken provided stimulating research questions that have kept me busy ever since. I would also like to thank Ewin Tang for closely collaborating with me on questions that came out of this internship, for her exuberance on dull pandemic days, and for her knack for formulating beautiful conjectures along the way.

Additionally, I'd like to thank Santosh Vempala for constant encouragement over

the course of my PhD and the numerous meetings wherein we would painstakingly go over an algorithm and analysis that required a flow-chart to keep track of. I'd also like to thank Ryan O'Donnell for always hearing out my most obscure math questions and pointing me to all the right places, and Daniel Kane for being the closest approximation to an oracle.

I am grateful to Ryan O'Donnell and Boaz Barak for agreeing to serve on my thesis committee, and providing insightful feedback and a fresh perspective on the results that appear in this thesis. I also owe a great deal to the faculty in the Computer Science Department at CMU for their support and invaluable advice, especially Anupam Guptam, Gary Miller, Ryan O'Donnell and Andrej Risteski.

I am also immensely grateful for having had incredible collaborators over the past years: Pranjal Awasthi, Nina Balcan, Michael Bender, Chiranjeeb Bhattacharya, Xue Chen, Nadiia Chepurko, Ken Clarkson, Alex Conway, Ilias Diakonikolas, Martin Farach-Colton, Piotr Indyk, Rajesh Jayaram, He Jia, Praneeth Kaccham, Daniel Kane, Ravi Kannan, Pravesh Kothari, Jerry Li, Sidhanth Mohanty Adarsh Prasad, Goutham Rajendran, Sandeep Silwal, Ewin Tang, Madhur Tulsiani, Santosh Vempala, Aravindan Vijayaraghavan, Colin White, David Woodruff and Samson Zhou.

Getting to graduate school was a long and winding road for me. I started out doing chemical engineering in India, and two years in, I realized this was not the path for me. I transfered to Rutgers on a whim, and started doing Computer Science. I never considered doing a PhD until I took Martin Farach-Colton's graduate algorithms course. Apart from sparking my love for algorithm design, Martin provided the blueprint of how awesome academic life can be. Taking Eric Allender's Graduate Complexity class convinced me that I wanted to do theoretical computer science, and I applied to graduate school two years into my CS degree. I got rejected from every school, and if it wasn't for a singular conversation with Michael Bender, I'd quit research. I owe a great deal to Martin, Eric and Pranjal Awasthi for encouraging me to try again, and taking me on as a research assistant straight out of undergrad. I am specially grateful to Martin for being a sounding board for every step of my academic journey, and perhaps the most influential person after my advisors.

It takes a village to raise a PhD student, and in addition to senior academics mentioned above, I am deeply grateful to my 'academic' friends, who filled my time as a PhD student with much joy. In particular, I'd like to thank Vijay Bhattiprolu, Gautam Kamath, Sahil Singla, David Wajc and Erik Waingarten for always listening to my complaints and questions, and calming my academic anxieties. In addition,

I would like to thank the vibrant Theory Group at CMU, and in particular, Costin Badescu, Naama Ben-David, Vijay Bhattiprolu, Tim Chu, Laxman Dhulipala, Guru Guruganesh, Paul Goelz, Isaac Grossof, Nika Haghtalab, Ellis Hershkowitz, Rajesh Jayaram, Praneeth Kachham, Greg Kehne, Misha Khodak, Roie Levin, Jason Li, Peter Manohar, Pedro Paredes, Siddharth Prasad, Nic Resch, Andrii Riazanov, Michael Rudow, Sai Sandeep, Anish Sevakari, Sahil Singla, Ellen Vitercik, Alex Wang, Ruosong Wang, David Wajc, Jalani Williams, Colin White, Justin Whitehouse, Xinyu Wu, Jeff Xu, Samson Zhou and Goran Zuzic.

In addition to having fantastic collaborators and mentors over the years, I've been tremendously lucky to meet some of the most incredible people during my time in Pittsburgh. I'd like to thank Vijay Bhattiprolu, Emily Black, Shreya Bhatia, Sofia Bosch, Liting Chen, Nadiia Chepurko, Tim Chu, Marina DiMarco, Rajesh Jayaram, Greg Kehne, Klas Leino, Roie Levin, Pedro Paredes, Filipe Perez, Kevin Pratt, Aria Wang, Zoe Wellner and Goran Zuzic for keeping me sane outside of work, and organizing numerous social events over the years. Some of the best memories of my life have been with you folks and I will cherish them forever. If any of you are reading this, I am truly sorry for the terrible jokes I made. A special shout out to Pedro and Zoe for treating me like family and always having me over for dinner. To Tim for the extra-ordinary stories, full of life lessons, and Goran for unfiltered expositions about every-day life. To Pedro for *always* making time and competing in obscure sports. And finally to Raj for being a collaborator, travel companion, gambler, and the sibling I never had.

Next, I'd like to thank my lifelong friends Akshat Agarwal, Anant Agarwal, Spardha Angra, Khushi Mehra and Karan Saharya for always keeping me grounded, and in general, sticking around for over a decade and a half.

I would also like to thank my family, without whom I would not be here today. Geeta Massi, Lewis Uncle and Nina provided a home away from home, here in the US. My grandparents, to whom this thesis is dedicated, Swaran and Arbindo Kwatra, and Shakuntala and Manmohan Bakshi, came to modern day India as refugees in 1947. They started from scratch in a new land, but instilled the importance of education and hard work in their kids. I've been fortunate to spend a lot of my childhood in their presence, and I am extremely grateful for all the stories, from the world war to Indian mythology. Finally, I'd like to thank my parents, Jappy and Rakesh Bakshi, without whom none of this would be possible. I am the person I am today due to them, and I am so grateful to them for being my cornerstone. I am also grateful to

them for spending countless hours explaining how and why *everything* around me works. In particular, I'd like to thank my mom instilling a passion for learning from a young age and always indulging my curiosity for the unknown. I'd like to thank my dad for instilling rationality as virtue, and setting the example for what it means to be humble, honest, hardworking, and for the countless hours we sunk into playing cricket and table tennis over the years. Mom and Dad, if you are reading this, thank you for all your sacrifice, you are the best parents anyone could ask for.

Chapter 1

Introduction

The unreasonable success of modern machine learning relies on algorithms that fit expressive latent models to large datasets. While such tasks are easy in low dimensions, real-world datasets are truly high-dimensional, often leading to computational intractability. Additionally, a prerequisite to deploying such models in real-world systems is to ensure that their behavior degrades gracefully when the modeling assumptions no longer hold. Therefore, there is a growing need for *efficient algorithms* that fit reliable and robust latent models to data and are accompanied with provable guarantees on their performance.

This thesis focuses on the burgeoning area of designing efficient, robust and provable algorithms for fundamental tasks arising in machine learning. In particular, we focus on two complementary regimes for algorithm design: *establishing tractability* and *approaching optimality*. The first regime tackles learning latent models where no efficient algorithms were known when the dimension is large. We begin by considering the most well-known and widely studied statistical model: the Gaussian Mixture Model (GMM). A long-standing open question in algorithmic statistics asks whether there exists any efficient algorithm to provably learn the parameters of a GMM in the presence of a small fraction of outliers. We completely resolve this problem and dedicate the first two chapters of the thesis to describing our result. Next, we consider learning a hyperplane and a subspace in the presence of outliers, and characterize the family of distributions that admit efficient algorithms for these problems. We show that for fitting such simple models (linear or low-rank), we can handle a much larger family of distributions, often including heavy tailed and log-concave distributions. Finally, we consider learning the parameters of a two-layer neural network with non-linear activations. Here, the input is drawn from a sub-Gaussian distribution and the network may be under-parameterized. In this setting, we obtain

the first polynomial time algorithms to recover the weight parameters of the network. Therefore, in the first part of this thesis, we establish the computational tractability for (a) robustly learning any Gaussian Mixture Models, (b) robust linear regression and subspace recovery for a broad family of distributions and (c) learning the parameters of a two-layer neural network.

The second regime tackles latent models which already admit efficient (polynomial time) algorithms, and our goal is to obtain nearly optimal (information-theoretically) algorithms. We begin by considering the low-rank approximation problem (also known as PCA), where the objective function is any Schatten norm, including well-studied objectives such as Frobenius norm, Operator norm and Nuclear norm. We resolve the matrix-vector product complexity of low-rank approximation for a large class of Schatten norms, obtaining information-theoretically optimal bounds, and in turn the fastest iterative algorithms for this class of latent models. Next, we consider the setting where we fit low-rank models with additional structure, in particular, positive semi-definiteness (PSD). We show that if the input is promised to be PSD, then we can obtain a low-rank approximation without reading most of the input. Our algorithm runs in *sub-linear* time and reads the information-theoretically minimal number of entries required. Finally, we consider the problem of learning a latent simplex, a formulation that captures several latent models such as the stochastic block model, clustering, latent Dirichlet allocation (topic modeling). We obtain truly input-sparsity (nearly linear time) algorithms for learning a latent simplex. Therefore, in the second part of this thesis we obtain nearly optimal algorithms for (a) low-rank approximation under Schatten norms, (b) low-rank approximation of PSD matrices and (c) learning a latent simplex.

The algorithms we develop draw upon tools from convex and polynomial optimization, high-dimensional probability, random matrix theory, functional analysis and convex geometry. In each setting, the algorithms we obtain are accompanied with provable guarantees on their correctness and performance. We focus on obtaining the most general theorems possible and identify techniques that may be of interest beyond the specific problems we consider. Next, we describe our results at a technical level, and explain the new ideas we introduce in each corresponding paper.

1.1 Establishing Tractability of Learning Latent Models

Given a collection of observations and a class of latent models, the objective of a typical learning algorithm is to find the model in the class that best fits the data. The classes of latent models we consider in this section are (a) Gaussian mixture models, (b) linear models, (c) low-rank models

and (d) two-layer neural networks. For linear and low-rank models, folklore algorithms such as least-squares regression on the empirical samples suffices to learn the optimal hyperplane or subspace efficiently. For GMMs, the first efficient algorithms were obtained in breakthrough works more than a decade ago [MV10, BS15]. For learning two-layer neural networks, even under Gaussian input, there were no provably efficient algorithms to find the model that best fits the data.

The aforementioned algorithms all assume that the input data are i.i.d. samples generated by a statistical model in the given class. However, as early as the 60's, statisticians already realized that real-world datasets are noisy and are unlikely to fit idealized statistics models [Hub64]. The sources of such noise can range from systematic bias and error in data collection to malicious tampering. Robust statistics [Hub04, HRRS11] challenges this assumption by focusing on the design of *outlier-robust* estimators – algorithms that can tolerate a *constant fraction* of corrupted datapoints, and achieve error that is independent of the dimension. Despite significant effort over several decades starting with important early works of Tukey and Huber in the 60s, until fairly recently, even for the most basic high-dimensional estimation tasks, all known computationally efficient estimators were highly sensitive to outliers.

In the first part of this thesis, we establish the computational tractability of learning GMMs, linear models and low-rank models under adversarial outliers. Subsequently, we provide the first polynomial time algorithm for fitting a two-layer neural network in the non-robust setting. A robust variant of this algorithm remains an outstanding open question. We discuss the historical context, related work and technical details of each of these results below.

1.1.1 Gaussian Mixture Models

The Gaussian Mixture Model (GMM) has been the subject of a century-old line of research beginning with Pearson [Pea94]. Progress on provable algorithms for learning GMMs began with the influential work of Dasgupta [Das99], yielding clustering algorithms that succeed under various separation assumptions [AK05, VW04, AM05, BV08]. These assumptions, however, do not capture natural separated instances of Gaussians, such as separation in distribution (total variation) distance. A more general approach [MV10, BS15] circumvents clustering altogether by giving an efficient algorithm for parameter estimation without any separation assumptions. However, this approach is brittle to even adversarially corrupting a single input point and crucially relies on the algebraic structure of Gaussians. A natural question to ask is then as follows:

Question 1. *Is there an efficient and robust algorithm to learn the parameters of arbitrary mixtures of k Gaussians?*

This question, and several special cases has received a lot of attention over the years. Finding an efficient algorithm for this task was also highlighted as a central open problem at the Foundations of Big Data workshop at the Simons Institute [DVW18]. Clustering a mixture of k Gaussians is an important special case of this problem, where each pair of components of the mixture is nearly completely separated in total variation distance. Until recently, no efficient robust algorithm was known even for clustering a mixture of *two* well-separated Gaussians.

We begin by formally defining a Gaussian Mixture model:

Definition 1.1.1 (Gaussian Mixture Model)**.** *A mixture of k Gaussians is a probability distributon, denoted by $\mathcal{D} = \sum_{i\in[k]} p_i \cdot \mathcal{N}(\mu_i, \Sigma_i)$, where for all $i \in [k]$, $\mu_i \in \mathbb{R}^d$ and $\Sigma_i \in \mathbb{R}^{d\times d}$ is a set of k means and covariances respectively, $p_i \geqslant 0$, and $\sum_{i\in[k]} p_i = 1$. A sample from $\mathcal{D}$ is generated by picking component i with probability p_i and then outputting an i.i.d. sample from $\mathcal{N}(\mu_i, \Sigma_i)$.*

Additionally, to measure closeness between two distributions, we use total variation (TV) distance.

Definition 1.1.2 (Total Variation Distance)**.** *Given two distributions p and q, we define the total variation distance between them as follows:*

$$d_{TV}(p, q) = \frac{1}{2}\int_{-\infty}^{\infty} |p(x) - q(x)|dx.$$

We first consider the special case where the input mixture is clusterable, i.e. all components of the mixture are pairwise separated in TV distance.

Robustly Clustering a Mixture of Gaussians

In recent work with Pravesh Kothari [BK20b], we obtained the first polynomial-time algorithm based on the sum-of-squares (SoS) method for clustering TV-separated k-GMMs in the presence of a small fraction of fully adversarial outliers. We begin by precisely defining the corruption model we consider. We work in the strong contamination model, which generalized several well-studied noise models, including the Huber contamination model [Hub64].

Definition 1.1.3 (Strong Contamination Model)**.** *Given a parameter $\epsilon \in (0, 1/2)$ and a class of distributions $\mathcal{D}$ over $\mathbb{R}^d$, the adversary is computationally unbounded and operates as follows: the algorithm specifies a number of samples, n, and n i.i.d. samples are drawn from some unknown $D \in \mathcal{D}$. The adversary is allowed to inspect the samples, remove up to ϵn samples and replace them with arbitrary points in $\mathbb{R}^d$. The modified set is given as input to the algorithm. We call such a set an ϵ-corrupted sample.*

Various communities have also considered less powerful adversaries, giving rise to weaker contamination models. For instance, an adversary may be adaptive or oblivious to the inliers, only allowed to add outliers, or only allowed to remove inliers.

Formally, our main result is as follows:

Theorem 1 (Outlier-Robust Clustering of k-GMMs, [BK20b])**.** *Fix $\eta, \epsilon > 0$. Let $\mathcal{D}$ be an equi-weighted k-GMM such that for all $i \neq i'$, $d_{TV}\left(\mathcal{N}\left(\mu_i, \Sigma_i\right), \mathcal{N}\left(\mu_i, \Sigma_i\right)\right) \geqslant 1 - \exp\left(-(k/\eta)^c\right)$, for a fixed constant c. Then, there exists an algorithm that takes input an ϵ-corruption Y of a sample $X \sim \mathcal{D}$ such that $X = C_1 \cup C_2 \cup \ldots \cup C_k$, with equal sized clusters C_i corresponding to points drawn from $\mathcal{N}\left(\mu_i, \Sigma_i\right)$, and with probability at least 0.99, outputs an approximate clustering $Y = \hat{C}_1 \cup \hat{C}_2 \cup \ldots \cup \hat{C}_k$ satisfying $\min_{i \leqslant k} \frac{|\hat{C}_i \cap C_i|}{|C_i|} \geqslant 1 - O(k^{2k})(\epsilon + \eta)$. The algorithm succeeds whenever $n = |X| \geqslant d^{poly(k/\eta)}$ and runs in time $n^{poly(k/\eta)}$.*

We can use off-the-shelf robust estimators for mean and covariance of Gaussians([DKK+19]) in order to get statistically optimal estimates of the mean and covariances of the target k-GMM.

Corollary 1.1.4 (Parameter Recovery from Clustering, [BK20b])**.** *In the setting of Theorem 1, with the same running time, sample complexity and success probability, our algorithm can output $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i \in [k]}$ such that for some permutation $\pi : [k] \to [k]$,*

$$d_{TV}\left(\mathcal{N}\left(\mu_i, \Sigma_i\right), \mathcal{N}\left(\hat{\mu}_{\pi(i)}, \hat{\Sigma}_{\pi(i)}\right)\right) \leqslant \tilde{O}\left(k^{2k}(\epsilon + \eta)\right),$$

where $\tilde{O}$ surpresses polylogarithmic factors in k, η and ϵ.

We note that a similar result was independently and concurrently obtained by [DHKK20] resulting in a merge [BDH+20].

Discussion. We obtain the first outlier-robust algorithm that works for clustering k-GMMs under information-theoretically minimal separation assumptions. Such results were not known even for $k = 2$. To discuss the bottlenecks in prior works, it is helpful to use following con-

sequence of two Gaussians with means μ_1, μ_2 and covariances Σ_1, Σ_2 being at a TV distance $\geqslant 1 - \exp(-O(\Delta^2))$ in terms of the distance between their parameters.

Definition 1.1.5 (Δ-Separated Mixture Model)**.** *An equi-weighted mixture $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_k$ with parameters $\{\mu_i, \Sigma_i\}_{i\in[k]}$ is Δ-separated if for every pair of distinct components i, j, one of the following three conditions hold ($\Sigma^{\dagger/2}$ is the square root of pseudo-inverse of Σ):*

1. ***Mean-Separation:*** *$\exists v \in \mathcal{R}^d$ such that*

$$\langle \mu_i - \mu_j, v\rangle^2 > \Delta^2 \cdot v^\top(\Sigma_i + \Sigma_j)v,$$

2. ***Spectral-Separation:*** *$\exists v \in \mathcal{R}^d$ such that*

$$v^\top \Sigma_i v > \Delta \cdot v^\top \Sigma_j v,$$

3. ***Relative-Frobenius Separation:***[1] *Σ_i and Σ_j have the same range space and*

$$\left\|\Sigma_i^{\dagger/2}\Sigma_j\Sigma_i^{\dagger/2} - I\right\|_F^2 > \Delta^2 \cdot \left\|\Sigma_i^{\dagger/2}\Sigma_j^{1/2}\right\|_{op}^4.$$

We show that two Gaussians separated in TV distance can be separated in any of the aforementioned notions of parameter distance. The key bottleneck for known algorithms prior to our work was handling separation in Spectral and Relative Frobenius distance (cases 2 and 3 above).

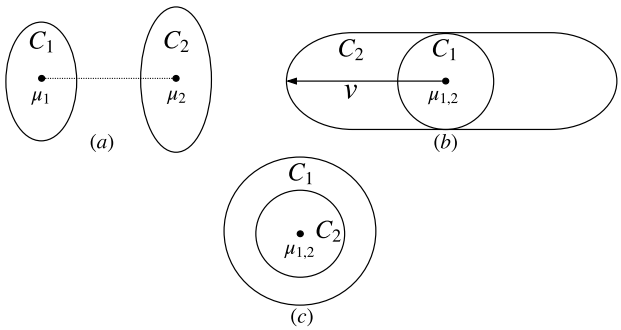

Figure 1.1: (a) Mean Separation (b) Spectral Separation (c) Relative Frobenius Separation

Often real-world data need not be Gaussian, and our algorithm does not overfit to this as-

[1]Unlike the other two distances, relative Frobenius distance is meaningful only for high-dimensional Gaussians. As an illustrative example, consider two 0 mean Gaussians with covariances $\Sigma_1 = I$ and $\Sigma_2 = (1 + \Theta(1/\sqrt{d}))I$. Then, for large enough d, the parameters are separated in relative Frobenius distance but not spectral or mean distance.

sumption. It succeeds for mixtures of all distributions that satisfy two well-studied analytic conditions: *anti-concentration* and *hypercontractivity*. In particular, we formulate these conditions as polynomial inequalities and obtain algorithms that can efficiently verify them. We thus move beyond Pearson's method of moments and consider identifying clean analytic conditions that enable the existence of efficient and robust clustering algorithms an important contribution of our work. We note that such a result for non-Gaussian distributions was not known, even with access to unbounded computation.

Next, we define the precise analytic conditions we require.

Definition 1.1.6 (Certifiable Hypercontractivity of Degree-2 polynomials)**.** *An isotropic distribution $\mathcal{D}$ on $\mathcal{R}^d$ is said to be h-certifiably C-hypercontractive if there's a degree h sum-of-squares proof of the following unconstrained polynomial inequality in $d \times d$ matrix-valued indeterminate Q:*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx\right]\right)^h\right] \leqslant (Ch)^h \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx\right]\right)^2\right]\right)^{h/2},$$

A set of points $X \subseteq \mathcal{R}^d$ is said to be C-certifiably hypercontractive if the uniform distribution on X is h-certifiably C-hypercontractive.

Hypercontractivity is an important notion in high-dimensional probability and analysis on product spaces [O'D14]. Kauers, O'Donnell, Tan and Zhou [KOTZ14] showed certifiable hypercontractivity of Gaussians and more generally product distributions with subgaussian marginals. Certifiable hypercontractivity strictly generalizes the better known *certifiable subgaussianity* property (studied first in [KSS18]) that controls higher moments of linear polynomials.

In contrast to hypercontractivity, anti-concentration forces *lower-bounds* of the form $\mathbf{Pr}[\langle x, v\rangle^2 \geqslant \delta \|v\|_2^2] \geqslant \delta'$, for all directions v. Certifiable anti-concentration was recently introduced in independent works of Karmalkar, Klivans and Kothari [KKK19] and Raghavendra and Yau [RY20a] and later used [BK21],[RY20b] for the related problems of list-decodable linear regression and subspace recovery[2].

Following [KKK19], we formulate certifiable anti-concentration via a univariate, even polynomial $p_{\delta,\Sigma}$ that uniformly approximates the 0-1 core-indicator $1(\langle x, v\rangle^2 \geqslant \delta v^\top \Sigma v)$ over a large enough interval around 0. Let $q_{\delta,\Sigma}(x, v)$ be a multivariate (in v) polynomial defined by

[2]List-decodable versions of these problems generalize the "mixture" variants - mixed linear regression and subspace clustering - that are easily seen to be special cases of mixtures of k-Gaussians with TV separation 1.

$q_{\delta,\Sigma}(x,v) = \left(v^\top \Sigma v\right)^{2s} p_{\delta,\Sigma}\left(\frac{\langle x,v\rangle}{\sqrt{v^\top \Sigma v}}\right)$.Since $p_{\delta,\Sigma}$ is an even polynomial, $q_{\delta,\Sigma}$ is a polynomial in v.

Definition 1.1.7 (Certifiable Anti-Concentration)**.** *A mean* 0 *distribution* D *with covariance* Σ *is* $2s$*-certifiably* $(\delta, C\delta)$*-anti-concentrated if for* $q_{\delta,\Sigma}(x,v)$ *defined above, there exists a degree-*$2s$ *sum-of-squares proof of the following two unconstrained polynomial inequalities in indeterminate* v*:*

$$\left\{\langle x,v\rangle^{2s} + \delta^{2s} q_{\delta,\Sigma}(x,v)^2 \geqslant \delta^{2s}\left(v^\top \Sigma v\right)^{2s}\right\}, \left\{\mathop{\mathbb{E}}_{x\sim D}\left[q_{\delta,\Sigma}(x,v)^2\right] \leqslant C\delta\left(v^\top \Sigma v\right)^{2s}\right\},$$

An isotropic subset $X \subseteq \mathcal{R}^d$ *is* $2s$*-certifiably* $(\delta, C\delta)$*-anti-concentrated if the uniform distribution on* X *is* $2s$*-certifiably* $(\delta, C\delta)$*-anti-concentrated.*

Remark 2. For natural examples, $s(\delta) \leqslant 1/\delta^c$ for some fixed constant c. For e.g., $s(\delta) = O(\frac{1}{\delta^2})$ for standard Gaussian distribution and the uniform distribution on the unit sphere (see [KKK19] and [BK21]). To simplify notation, we will assume $s(\delta) \leqslant \text{poly}(1/\delta)$ in the statement of our results.

Additionally, we need that the variance of degree-2 polynomials is bounded in terms of the Frobenius norm of the coefficients of the polynomial. Formally,

Definition 1.1.8 (Degree-2 Polynomials with Certifiably Bounded Variance)**.** *A mean* 0 *distribution* $\mathcal{D}$ *with covariance* Σ *certifiably bounded variance degree* 2 *polynomials if there is a degree* 2 *sum-of-squares proof of the following inequality in the indeterminate* $Q \in \mathbb{R}^{d\times d}$

$$\left\{\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Q x - \mathbf{E}_{x\sim\mathcal{D}} x^\top Q x\right)^2\right] \leqslant C\left\|\Sigma^{1/2} Q \Sigma^{1/2}\right\|_F^2\right\},$$

Our general result gives an outlier-robust clustering algorithm for separated mixtures of *reasonable* distributions, i.e., distributions that satisfies both certifiable hypercontractivity, anti-concentration and have bounded variance of degree-2 polynomials. Even the information-theoretic (and without outliers, i.e., $\epsilon = 0$) clusterability of such distributions was not known prior to our work.

Theorem 3 (Outlier-Robust Clustering of Reasonable Mixtures, [BK20b])**.** *Fix* $\eta > 0, \epsilon > 0$. *Let* $\mathcal{D}$ *be a* Δ*-separated mixture of reasonable distributions. Then, there exists an algorithm that takes input an* ϵ*-corruption* Y *of a sample* $X = C_1 \cup C_2 \cup \ldots C_k$*, with true clusters* C_i *of size* n/k *drawn i.i.d. from* $\mathcal{D}$ *and outputs an approximate clustering* $Y = \hat{C}_1 \cup \hat{C}_2 \cup \ldots \cup \hat{C}_k$ *satisfying* $\min_{i\leqslant k} \frac{|\hat{C}_i \cap C_i|}{|C_i|} \geqslant 1 - O(k^{2k})(\epsilon + \eta)$. *The algorithm succeeds with probability at least* 0.99 *over*

the draw of the original sample X whenever $n \geqslant d^{poly(k/\eta)}$ and runs in time $n^{poly(k/\eta))}$ whenever $\Delta \geqslant poly(k/\eta)^k$.

Overview. Our work is naturally related to the recent progress (see Chapter 4 [FKP+19] for an exposition) on learning spherical mixtures[3] of Gaussians [DKS18, KSS18, HL18] and more generally, all Poincaré distributions [KSS18]. These results rely on subgaussian moment *upper bounds* and extend to the outlier-robust setting. However, moment upper bounds are inherently insufficient to cluster non-spherical mixtures. Informally, this is because the property of having subgaussian moment upper bounds is closed under taking mixtures and thus cannot distinguish between a single Gaussian and mixture of a few.

Indeed, it was "folklore" that obtaining generalization of the results above to non-spherical mixtures will likely require algorithmic use of *moment lower bounds*. A recent line of work begun by [KKK19, RY20a] and further built on in [BK20a, RY20b] introduced *certifiable anti-concentration* that allows algorithmically accessing moment lower-bounds to solve list-decodable variants (harsher outlier model than ours) of regression and subspace recovery. An important technical contribution of our work is to show that moment lower-bounds, inferred from anti-concentration inequalities along with certifiable hypercontractivity and bounded variance of degree-2 polynomials are enough to obtain the desired generalization for clustering of all TV-separated mixtures.

The key technical contribution of our work is a low-degree sum-of-squares proof of a basic statistical statement that gives a strong, dimension-independent bound relating closeness of distribution in *total variation distance* (TV) to an appropriate *parameter distance* between their means and covariances. Our proof of this basic result works for all distributions that satisfy (certifiable) anti-concentration and hypercontractivity of degree-2 polynomials. To the best of our knowledge, even the information-theoretic relationship between total variation and parameter distances of such distributions was not known prior to our work. Further, in Chapter 2.12, we give a simple proof by exhibiting two (certifiably) hypercontractive (and, thus, also subgaussian) distributions that are $(1-\eta)$-close in TV distance but arbitrarily far in parameter distance showing that moment upper bounds are provably not enough for the TV vs parameter distance relationships to hold.

Along the way, we grow the general purpose SoS toolkit for algorithm design. For instance, we give low-degree sum-of-squares formulations of *conditional* arguments using uni-

[3]More generally, the SoS-based algorithms succeed when the means of the components are separated when compared to the maximum variance of the components in any direction.

form polynomial approximators and basic matrix analytic facts. As another application of our techniques, we give an outlier-robust algorithm for covariance estimation of all certifiable hypercontractive distributions with $\tilde{O}(\epsilon)$ relative Frobenius error guarantee. All prior works [DKK+19, LRV16] either gave error guarantees in spectral norm, which only translate into dimension dependent guarantees for relative Frobenius distance, or worked only for the Gaussian distribution [DKK+19]). Combined with our outlier-robust clustering algorithm, we obtain a statistically optimal outlier-robust parameter estimation algorithms for mixtures of Gaussians.

Future Directions. The class of distributions that satisfy certifiable hypercontractivity of degree-2 polynomials is quite broad, and includes all strongly log-concave distributions. However, all existing approaches can establish certifiable anti-concentration only for rotationally invariant distributions and affine transformations thereof. Therefore, a natural open question is as follows:

Open Question 4 (Charecterizing Certifiable Anti-Concentration)**.** What class of distributions (beyond rotationally invariant distributions) admit low-degree sum-of-squares certificates ?

Further, the certificates we establish, even for Gaussian distributions, require a degree that grows polynomialy with δ, the bound on the expectation. The running time of our algorithm scales exponentially in the degree required above and thus improved bounds lead to significantly faster algorithms. Moreover, such an improvement would lead to milder assumptions on the TV separation between the components.

Open Question 5 (Degree of Certifiable Anti-Concentration)**.** What is the minimum degree required to establish $(\delta, C\delta)$-certifiable anti-concentration for Gaussian distributions? Is a polynomial dependence on δ neccessary?

Robustly Learning a Mixture of Arbitrary Gaussians

Building on [BK20b], in joint work with Diakonikolas, Jia, Kane, Kothari and Vempala, [BDJ+22] we were able to completely answer the aforementioned central question (Question 1) in the affirmative, by providing an efficient and robust algorithm that learns the parameters of all mixtures of k Gaussians, thereby resolving this central question in high-dimensional statistics. Our result requires the information-theoretically minimum assumptions on the input mixture, is robust a small fraction of adversarial corruptions and is provably faster than the existing non-robust algorithm of Moitra-Valiant [MV10]. Formally,

Theorem 6 (Robustly Learning k Arbitrary Gaussians, [BDJ+22])**.** *There is an algorithm with*

the following behavior: Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)}\,(1/\epsilon)^{c_k}$ samples from a distribution F on $\mathcal{R}^d$ such that $d_{TV}(F, \mathcal{M}) \leqslant \epsilon$, for an unknown target k-GMM $\mathcal{D} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time poly$(n)\,(1/\epsilon)^{c'_k}$ and outputs a k-GMM hypothesis $\widehat{\mathcal{D}} = \sum_{i=1}^{k} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ such that with high probability we have that $d_{TV}\left(\widehat{\mathcal{D}}, \mathcal{D}\right) \leqslant \mathcal{O}\left(\varepsilon^{1/c''_k}\right)$, where c_k, c'_k, c''_k depends only on k.

A number of works have made algorithmic progress on important special cases of the above problem, including faster robust clustering for the spherical case under minimal separation conditions [HL18, KSS18, DKS18], robust clustering for separated (and potentially non-spherical) Gaussian mixtures [BDH+20], and robustly learning *uniform* mixtures of two arbitrary Gaussian components [Kan20]. A similar result was independently and concurrently obtained by [LM21], under slightly stronger assumptions, and using completely different techniques.

Theorem 6 gives the first polynomial-time *robust proper learning algorithm*, with dimension-independent error guarantee, for *arbitrary* k-GMMs, for any fixed k. Known Statistical Query lower bounds [DKS17] suggest that $d^{\Omega(k)}$ samples are necessary for efficiently learning GMMs, for approximation to constant accuracy, even in the (much simpler) noiseless setting and when the components are pairwise well-separated in total variation distance. This provides evidence that the sample-time tradeoff achieved by our result is qualitatively optimal.

Further, we show that *the same algorithm* also achieves the stronger parameter estimation guarantee. We note that parameter estimation requires some assumptions on the underlying mixture. The following corollary applies under the standard assumption that any pair of components in the unknown mixture has total variation distance at least ϵ^{c_k}, where c_k only depends on k.

Corollary 1.1.9 (Robust Parameter Estimation, [BDJ+22])**.** *Let $\mathcal{D} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$ be an unknown target k-GMM satisfying the following conditions: (i) $d_{TV}(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j)) \geqslant \varepsilon^{f_1(k)}$ for all $i \neq j$, and (ii) $S = \{i \in [k] : w_i \geqslant \varepsilon^{f_2(k)}\}$ is a subset of $[k]$, where $f_1(k), f_2(k)$ are sufficiently small functions of k. Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)}\,(1/\varepsilon)^{c_k}$ samples from a distribution F on $\mathcal{R}^d$ such that $d_{TV}(F, \mathcal{D}) \leqslant \epsilon$, there exists an algorithm that runs in time poly$(n)\,(1/\varepsilon)^{c'_k}$ and outputs a k'-GMM hypothesis $\widehat{\mathcal{D}} = \sum_{i=1}^{k'} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ with $k' \leqslant k$ such that with high probability there exists a bijection $\pi : S \to [k']$ satisfying the following: For all $i \in S$, it holds that $|w_i - \widehat{w}_{\pi(i)}| \leqslant$ poly$_k(\epsilon)$ and $d_{TV}\left(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\widehat{\mu}_{\pi(i)}, \widehat{\Sigma}_{\pi(i)})\right) \leqslant \epsilon^{1/c''_k}$.*

Discussion. *Handling Arbitrary Weights*: Our algorithm succeeds without any assumptions on the weights of the mixture components. We emphasize that this is an important feature and not a technicality. Prior and concurrent work cannot handle the case of general weights – even for

the case of $k = 2$ components! Obtaining a fully polynomial-time algorithm for the general case (i.e., one not incurring an exponential cost in $1/w_{\min}$) requires genuinely new algorithmic ideas and is one of the key technical innovations of the aforementioned result.

Handling Arbitrary Covariances: Our algorithm does not require assumptions on the eigenvalues of the component covariances, modulo basic limitations posed by numerical computation issues. Specifically, our algorithm works even if some of the component covariances are rank-deficient (i.e., have directions of 0 variance) with running time scaling polynomially in the bit-complexity of the unknown component means and covariances. Such a dependence on the bit complexity of the input parameters is unavoidable – there exist[4] examples of rank-deficient covariances with irrational entries such that the total variation distance between the corresponding Gaussian and every Gaussian with covariance matrix of rational entries is the maximum possible value of one.

Overview. In the non-robust setting (i.e., for $\epsilon = 0$), the algorithm of [MV10] solves this learning problem. The key idea of [MV10] is to observe that if a mixture of k Gaussians has every pair of components separated in total variation distance by at least δ, then a random univariate projection of the mixture has a pair of components that are $\delta/\sqrt{d}$-separated in total variation distance. Their algorithm uses this observation to piece together estimates of the mixture when projected to several carefully chosen directions to get an estimate of the high-dimensional mixture. Notice, however, that such a strategy meets with instant roadblock in the presence of outliers: the fraction of outliers, being a dimension-independent constant, completely overwhelms the total variation distance between components in any one direction making them indistinguishable.

Our robust algorithm is based on three new ingredients:

1. a new and efficient *partial clustering algorithm* based on the sum-of-squares (SoS) method,
2. a novel *list-decodable tensor decomposition* method, and
3. a recursive *spectral separation* method.

We briefly describe these ideas below and how they can be interleaved to obtain our algorithm.

Efficient Partial Clustering. We call a mixture partially clusterable if it contains a pair of components at total variation distance larger than $1 - \Omega_k(1)$. Interestingly, it turns out that the clustering algorithm of [BK20b] (Theorem 3) can be generalized to the partial clustering setting, i.e., the setting where we are guaranteed to have a pair of components that are well-separated (with no guarantees on the remaining components). For a mixture with minimum

[4]For example, for the unit vector $v = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, 0, \ldots, 0)$, for every choice of rational covariance Σ, the total variation distance between $\mathcal{N}(0, I - vv^\top)$ and $\mathcal{N}(0, \Sigma)$ is one.

mixing weight α, this gives an algorithm with running time of $d^{(k/\alpha)^{O(k)}}$ to partition the input sample into components so that each piece of the partition is (effectively) a $(\text{poly}(\alpha/k)+\epsilon)$-corrupted sample from disjoint sub-mixtures.

By applying the above partial-clustering algorithm, we can effectively assume that the input is an ϵ-corrupted sample from a mixture with every pair of components *at most* $(1-\Omega_k(1))$-far in total variation distance. Then, we use robust covariance estimation (see Theorem 7.1 in [BK20b]) to make the mixture approximately isotropic, i.e. the mean of the mixture is ≈ 0 and the covariance of the mixture is $\approx I$ (in Frobenius norm).

After partial clustering and an approximate isotropic transformation, every pair of components are close in TV distance. Under this condition, in order to learn the unknown mixture with error guarantees in total variation distance, it suffices to obtain $\text{poly}_k(\epsilon)$-error estimates of the μ_i, Σ_i's in Frobenius norm. As we will see soon, this will suffice for our weaker result that has an exponential dependence on the minimum mixing weight.

To get a fully polynomial algorithm, we delve a bit deeper: the exponential dependence on the minimum mixing weight is incurred only when two components are spectrally separated (see Definition 1.1.5, which in turn relies on the degree required for certifiable anti-concentration. Instead, we give a new partial clustering algorithm that works in fixed polynomial time, whenever there is a pair of Gaussian components separated either via their means or the relative Frobenius distance. The resulting clusters might now have components that are spectrally separated, a difficulty that we address later.

List-decodable Tensor Decomposition. Kane [Kan20] gave a polynomial-time algorithm to robustly learn an *equiweighted* mixture of two Gaussians. For this special case, after isotropic transformation, one can effectively assume that the two means are $\pm\mu$ and the two covariances are $I \pm \Sigma$. Kane's idea was to use the Hermite tensor (which can be built using the 4-th and 6-th raw moments of the mixture). Since we must use outlier-robust estimates of these tensors, we can only obtain estimates that are accurate up to constant error in Frobenius norm of the tensor. Kane's key observation is that for the special case of $k=2$ components, one can build two different Hermite tensors, one of which is rank-one with component $\approx \mu$ (and thus one can immediately "read off" μ); the other only has a tensor power of Σ. This second tensor is of the form $\hat{T}_4 = \text{Sym}((\Sigma - I)\otimes(\Sigma - I)) + E$, where $\|E\|_F = O_k(\sqrt{\epsilon})$ and Sym refers to symmetrizing over all possible permutations of the "4 modes of the tensor". Unlike the case of the mean, one cannot simply "read-off"[5] Σ from T_4, but Kane gives a simple method to accomplish this. As

[5]It is helpful to visualize a single entry of this tensor for, say, the case when i,j,k,ℓ are all distinct: $\hat{T}_4(i,j,k,\ell) = \frac{1}{3}(\Sigma(i,j)\Sigma(k,\ell) + \Sigma(i,k)\Sigma(j,\ell) + \Sigma(i,\ell)\Sigma(j,k)) + error$. Notice that obtaining entries of Σ

noted in [Kan20], it is not clear how to extend this to non-equiweighted mixtures of $k = 2$ Gaussians, and going to even $k = 3$ components requires substantially new ideas.

The surprising fact that we establish is that by looking at only the first four moments of our mixture, we can learn all of the components up to low-rank error, i.e., up to errors along a bounded number of hidden directions. Thus, the new tensor decomposition has both Frobenius norm error and low-rank error. To see the idea, it is helpful to focus on the simpler case where all the means are zero. In this case, the estimated 4th Hermite tensor (built from estimated raw moments of degree at most 4 of the mixture) has the following :

$$\hat{T}_4 = \sum_{i=1}^{k} w_i \mathrm{Sym}((\Sigma_i - I) \otimes (\Sigma_i - I) + E) \ ,$$

where E is a 4-tensor with $\|E\|_F = O_k(\sqrt{\epsilon})$.

Given the form of this tensor, it is natural to consider tensor decomposition algorithms, by thinking of $\Sigma_i - I$ as a d^2-dimensional vector. However, we run into the issue of uniqueness of tensor decomposition, since we are dealing with 2nd order tensors (once we view $\Sigma_i - I$ as a d^2-dimensional vector). One might imagine computing higher-order tensors of similar forms to overcome the uniqueness issues, but this runs into two major complications: first, the symmetrization operation introduces spurious terms that do not have the sum of tensor-power structure required for such an algorithm to succeed.

Second, even if one were to get hold of the tensor without the symmetrization operation, the only applicable tensor decomposition algorithm (recall that we do not make *any* genericity assumptions on the components that are typically required by tensor decomposition algorithms) is the result of Barak, Kelner, and Steurer [BKS15]. However, the [BKS15] result, while being efficient in its dependence on the number of components, has exponential dependence on the target error, which is prohibitively expensive for our application.

Rather than recovering the unique decomposition of the tensor $\hat{T}_4$ above, we instead produce a list of candidate decompositions. To do this, we start by applying an operation that is a common trick in most tensor decomposition algorithms. In our context, this trick amounts to taking a random matrix (with independent standard Gaussian entries) P and "collapsing" the last two modes of $\hat{T}_4$ with P (i.e., computing $\hat{S}(i,j) = \sum_{k,\ell} \hat{T}_4(i,j,k,\ell)P(k,\ell)$) to obtain a matrix Q. In the usual tensor decomposition procedures, we are interested in proving that one can recover all the information about the components of the tensor from Q.

from T_4 is formally a task of solving noisy quadratic equations.

Spectral Separation of Thin Components. While the running time of our partial clustering and tensor decomposition algorithms are now polynomial, the guarantees of the tensor decomposition subroutine we discussed above are no longer enough to guarantee a recovery of parameters that result in a mixture close in total variation distance. Because of the three conditions that we assumed in the working of the tensor decomposition algorithm, we can no longer guarantee the third one that gives a lower bound on the smallest eigenvalue of every covariance (relative to the covariance of the mixture). In particular, we can end up in a situation where, even though we have a list of parameters that contain Frobenius-norm-close estimates of the covariances, the estimates do not provide a total variation distance guarantee. (Consider a "skinny" direction where the variance of some component is very small, or even 0, forcing us to learn the parameters more precisely!)

It turns out that the above is the only way the algorithm can fail at this point — one or more covariance matrices have a very small eigenvalue (if not, the Frobenius norm error would imply TV-distance error). But since we have estimates of the covariances, we can find such a small eigenvector. Now we observe that since the mixture is nearly isotropic (i.e., the overall variance in each direction is ~ 1), if some component has very small variance along a direction, then the components must be separable along this direction. We show that it is possible to efficiently cluster the mixture after projecting it to this direction, so that each cluster has strictly fewer components. We then recursively apply the entire algorithm on the clusters obtained, which will each have strictly fewer components.

Future Directions. A natural question arising from our work is to characterize the class of distributions such that their mixtures can be learned, even information-theoretically.

Open Question 7. Are there mixtures of non-Gaussian distributions that can be learned robustly/non-robustly and information-theoretically/efficiently? Is there a statistical-computational gap between any of these settings?

We note that the aforementioned algorithm is not entirely captured by the sum-of-squares proof system. This leads to the following question:

Open Question 8. Can the sum-of-squares proof system efficiently learn a mixutre of k arbitrary Gaussians?

We hope that answering some of these question, along with the techniques we have developed can pave the way for robustly learning various popular latent variable models.

1.1.2 Robust Linear Regression

Regression continues to be extensively studied under various models, including realizable regression (no noise), true linear models (independent noise), asymmetric noise, agnostic regression and generalized linear models (see [Wei05] and references therein). In each model, a variety of distributional assumptions are considered over the covariates and the noise. As a consequence, there exist innumerable estimators for regression achieving various trade-offs between sample complexity, running time and rate of convergence. The presence of adversarial outliers adds yet another dimension to design and compare estimators.

Seminal works on robust regression focused on designing non-convex loss functions, including M-estimators [Hub04], Theil-Sen estimators [The92, Sen68], R-estimators [Jae72], Least-Median-Squares [Rou84] and S-estimators [RY84]. These estimators have desirable statistical properties under disparate assumptions, yet remain computationally intractable in high dimensions. Further, recent works show that it is information-theoretically impossible to design robust estimators for linear regression without distributional assumptions [KKM18].

An influential recent line of work showed that when the data is drawn from the well studied and highly general class of *hypercontractive* distributions (see Definition 1.1.6), there exist robust and computationally efficient estimators for regression [KKM18, PSBR20, DKS19]. Several families of natural distributions fall into this category, including Gaussians, strongly log-concave distributions and product distributions on the hypercube. However, both estimators converge to the the true hyperplane (in ℓ_2-norm) at a sub-optimal rate, as a function of the fraction of corrupted points.

Given the vast literature on ad-hoc and often incomparable estimators for high-dimensional robust regression, the central question we address in this work is as follows:

Does there exist a unified approach to design robust and computationally efficient estimators achieving optimal rates for all linear regression models under mild distributional assumptions?

We address the aforementioned question by introducing a framework to design robust estimators for linear regression when the input is drawn from a *hypercontractive* distribution. Our estimators converge to the true hyperplanes at the information-theoretically optimal rate (as a function of the fraction of corrupted data) under various well-studied noise models, including independent and agnostic noise. Further, we show that our estimators can be computed in polynomial time using the *sum-of-squares* convex hierarchy.

In classical regression, we assume $\mathcal{D}$ is a distribution over $\mathcal{R}^d \times \mathcal{R}$ and for a vector $\Theta \in \mathcal{R}^d$, the least-squares loss is given by $\mathsf{err}_{\mathcal{D}}(\Theta) = \mathbb{E}_{x,y\sim\mathcal{D}}\left[\left(y - x^\top\Theta\right)^2\right]$. The goal is to learn $\Theta^* = \arg\min_\Theta \mathsf{err}_{\mathcal{D}}(\Theta)$. We assume sample access to $\mathcal{D}$, and given n i.i.d. samples, we want to obtain a vector Θ that approximately achieves optimal error, $\mathsf{err}_{\mathcal{D}}(\Theta^*)$. In contrast to the classical setting, we work in the *strong contamination model*, defined above.

Model 9 (Robust Regression Model)**.** Let $\mathcal{D}$ be a distribution over $\mathcal{R}^d \times R$ such that the marginal distribution over $\mathcal{R}^d$ is centered and has covariance Σ^* and let $\Theta^* = \arg\min_\Theta \mathbb{E}_{x,y\sim\mathcal{D}}\left[(y - \langle\Theta, x\rangle)^2\right]$ be the optimal hyperplane for $\mathcal{D}$. Let $\{(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots (x_n^*, y_n^*)\}$ be n i.i.d. random variables drawn from $\mathcal{D}$. Given $\epsilon > 0$, the robust regression model $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$ outputs a set of n samples $\{(x_1, y_1), \ldots (x_n, y_n)\}$ such that for at least $(1-\epsilon)n$ points $x_i = x_i^*$ and $y_i = y_i^*$. The remaining ϵn points are arbitrary, and potentially adversarial w.r.t. the input and estimator.

Our Results. A natural starting point is to assume that the marginal distribution over the covariates (the x's above) is heavy-tailed and has bounded, finite covariance. However, we show that there is no robust estimator in this setting, even when the linear model has no noise and the uncorrupted points lie on a line.

Theorem 10 (Bounded Covariance does not suffice [BP21])**.** *For all $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^d \times \mathcal{R}$ such that $d_{\mathsf{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leqslant \epsilon$ and the marginal distribution over the covariates has bounded covariance, denoted by $I \preceq \Sigma \preceq O(1)I$, yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega(1)$, where Θ_1 and Θ_2 are the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$.*

The aforementioned result precludes any statistical estimator that converges to the true hyperplane as the fraction of corrupted points tends to 0. Therefore, we strengthen the distributional assumption consider hypercontractive distributions instead.

Definition 1.1.10 (Certifiable Hypercontractivity)**.** *A distribution $\mathcal{D}$ on $\mathcal{R}^d$ is (c_k, k)-certifiably hypercontractive if for all $r \leqslant k/2$, there exists a degree $\mathcal{O}(k)$ sum-of-squares proof (defined below) of the following inequality in the variable v*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\langle x, v\rangle^{2r}\right] \leqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[c_r \langle x, v\rangle^2\right]^r$$

such that $c_r \leqslant c_k$.

Remark 11. Hypercontractivity captures a broad class of distributions, including Gaussian distributions, uniform distributions over the hypercube and sphere, affine transformations of isotropic

distributions satisfying Poincare inequalities [KSS18] and strongly log-concave distributions. Further, hypercontractivity is preserved under natural closure properties like affine transformations, products and weighted mixtures [KSS18].

In this work we focus on the *rate of convergence* of our estimators to the true hyperplane, Θ^*, as a function of the fraction of corrupted points, denoted by ϵ. We measure convergence in both parameter distance (ℓ_2-distance between the hyperplanes) and least-squares error on the true distribution ($\mathsf{err}_{\mathcal{D}}$).

We introduce a simple analytic condition on the relationship between the noise (marginal distribution over $y - x^\top\Theta^*$) and covariates (marginal distribution over x) that can be considered as a proxy for independence of $y - x^\top\Theta^*$ and x :

Definition 1.1.11 (Negatively Correlated Moments)**.** *Given a distribution $\mathcal{D}$ over $\mathcal{R}^d \times \mathcal{R}$, such that the marginal distribution on $\mathcal{R}^d$ is (c_k, k)-hypercontractive, the corresponding regression instance has negatively correlated moments if for all $r \leqslant k$, and for all v,*

$$\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\langle v, x\rangle^r \left(y - x^\top\Theta^*\right)^r\right] \leqslant \mathcal{O}(1) \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\langle v, x\rangle^r\right] \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y - x^\top\Theta^*\right)^r\right]$$

Informally, the *negatively correlated moments* condition can be viewed as a polynomial relaxation of independence of random variables. Note, it is easy to see that when the noise is independent of the covariates, the above definition is satisfied.

Remark 12. We show that when this condition is satisfied by the true distribution, $\mathcal{D}$, we obtain rates that match the information theoretically optimal rate in a *true linear model*, where the noise (marginal distribution over $y - x^\top\Theta^*$) is independent of the covariates (marginal distribution over x). Further, when this condition is not satisfied, we show that there exist distributions for which obtaining rates matching the *true linear model* is impossible.

When the distribution over the input is hypercontractive and has negatively correlated moments, we obtain an estimator achieving *rate* proportional to $\epsilon^{1-1/k}$ for parameter recovery. Further, our estimator can be computed efficiently. Thus, our main algorithmic result is as follows:

Theorem 13 (Robust Regresssion with Negatively Correlated Noise, [BP21])**.** *Given $\epsilon > 0, k \geqslant 4$, and $n \geqslant (d\log(d))^{\mathcal{O}(k)}$ samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, such that $\mathcal{D}$ is (c, k)-certifiably hypercontractive and has negatively correlated moments, there exists an algorithm that runs in $n^{\mathcal{O}(k)}$*

time and outputs an estimator $\tilde{\Theta}$ such that with high probability,

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \tilde{\Theta}\right)\right\|_2 \leqslant \mathcal{O}\left(\epsilon^{1-1/k}\right)\left(\mathit{err}_{\mathcal{D}}(\Theta^*)^{1/2}\right)$$

and,

$$\mathit{err}_{\mathcal{D}}(\tilde{\Theta}) \leqslant \left(1 + \mathcal{O}\left(\epsilon^{2-2/k}\right)\right)\mathit{err}_{\mathcal{D}}(\Theta^*)$$

Remark 14. We note that prior work does not draw a distinction between the independent and dependent noise models. In comparison (see Table 4.1), Klivans, Kothari and Meka [KKM18] obtained a sub-optimal least-squares error scales proportional to $\epsilon^{1-2/k}$. For the special case of $k = 4$, Prasad et. al. [PSBR20] obtain least squares error proportional to $O(\epsilon\kappa^2(\Sigma))$, where κ is the condition number. In very recent independent work Zhu, Jiao and Steinhardt [ZJS20] obtained a sub-optimal least-squares error scales proportional to $\epsilon^{2-4/k}$.

Further, we show that the rate we obtained in Theorem 13 is information-theoretically optimal, even when the noise and covariates are independent:

Theorem 15 (Lower Bound for Independent Noise, [BP21])**.** *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^2 \times \mathcal{R}$ such that the marginal distribution over $\mathcal{R}^2$ has covariance Σ and is (c, k)-hypercontractive for both distributions, and yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\epsilon^{1-1/k}\sigma\right)$, where Θ_1, Θ_2 are the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, $\sigma = \max(\mathit{err}_{\mathcal{D}_1}(\Theta_1), \mathit{err}_{\mathcal{D}_2}(\Theta_2))$ and the noise is uniform over $[-\sigma, \sigma]$. Further, $|\mathit{err}_{\mathcal{D}_1}(\Theta_2) - \mathit{err}_{\mathcal{D}_1}(\Theta_1)| = \Omega\left(\epsilon^{2-2/k}\sigma^2\right)$.*

Next, we consider the setting where the noise is allowed to arbitrary, and need not have negatively correlated moments with the covariates. A simple modification to our algorithm and analysis yields an efficient estimator that obtains rate proportional to $\epsilon^{1-2/k}$ for parameter recovery.

Corollary 1.1.12 (Robust Regresssion with Dependent Noise, [BP21])**.** *Given $\epsilon > 0, k \geqslant 4$ and $n \geqslant (d\log(d))^{\mathcal{O}(k)}$ samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, such that $\mathcal{D}$ is (c, k)-certifiably hypercontractive, there exists an algorithm that runs in $n^{\mathcal{O}(k)}$ time and outputs an estimator $\tilde{\Theta}$ such that with probability 9/10,*

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \tilde{\Theta}\right)\right\|_2 \leqslant \mathcal{O}\left(\epsilon^{1-2/k}\right)\left(\mathit{err}_{\mathcal{D}}(\Theta^*)^{1/2}\right),$$

and,

$$\mathit{err}_{\mathcal{D}}(\tilde{\Theta}) \leqslant \left(1 + \mathcal{O}\left(\epsilon^{2-4/k}\right)\right)\mathit{err}_{\mathcal{D}}(\Theta^*).$$

Further, we show that the dependence on ϵ is again information-theoretically optimal:

Estimator	Independent Noise	Arbitrary Noise
Prasad et. al. [PSBR20], Diakonikolas et. al. [DKK+18]	$\epsilon\,\kappa^2$ (only $k=4$)	$\epsilon\,\kappa^2$ (only $k=4$)
Klivans, Kothari and Meka [KKM18]	$\epsilon^{1-2/k}$	$\epsilon^{1-2/k}$
Zhu, Jiao and Steinhardt [ZJS20]	$\epsilon^{2-4/k}$	$\epsilon^{2-4/k}$
Our Work Thm 13, Cor 4.1.3	$\epsilon^{2-2/k}$	$\epsilon^{2-4/k}$
Lower Bounds Thm 15, Thm 16	$\epsilon^{2-2/k}$	$\epsilon^{2-4/k}$

Table 1.1: Comparison of convergence rate (for least-squares error) achieved by various computationally efficient estimators for Robust Regression, when the underlying distribution is (c_k, k)-hypercontractive.

Theorem 16 (Lower Bound for Dependent Noise, [BP21])**.** *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^2 \times \mathcal{R}$ such that the marginal distribution over $\mathcal{R}^2$ has covariance Σ and is (c,k)-hypercontractive for both distributions, and yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\epsilon^{1-2/k}\sigma\right)$, where Θ_1, Θ_2 be the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively and $\sigma = \max(\mathsf{err}_{\mathcal{D}_1}(\Theta_1), \mathsf{err}_{\mathcal{D}_2}(\Theta_2))$. Further, $|\mathsf{err}_{\mathcal{D}_1}(\Theta_2) - \mathsf{err}_{\mathcal{D}_1}(\Theta_1)| = \Omega\left(\epsilon^{2-4/k}\sigma^2\right)$.*

Overview. Consider two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over $\mathcal{R}^d \times \mathcal{R}$ such that the total variation distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ is ϵ and the marginals for both distributions over $\mathcal{R}^d$ are (c_k, k)-hypercontractive and have covariance Σ. Ignoring computational and sample complexity concerns, we can obtain the optimal hyperplanes corresponding to each distribution. Note, these hyperplanes need not be unique and are simply charecterized as minimzers of the least-squares loss : for $i \in \{1,2\}$,

$$\Theta_i = \arg\min_{\Theta} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_i}\left[\left(y - x^\top\Theta\right)^2\right]$$

Our central contribution is to obtain an information theoretic proof that the optimal hyperplanes are indeed close in scaled ℓ_2 norm, i.e.

$$\left\|\Sigma^{1/2}\left(\Theta_1 - \Theta_2\right)\right\|_2 \leqslant \mathcal{O}\left(\epsilon^{1-1/k}\right)\left(\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_1}\left[\left(y - x^\top\Theta_1\right)^2\right]^{1/2} + \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_2}\left[\left(y - x^\top\Theta_2\right)^2\right]^{1/2}\right)$$

Further, we show that the $\epsilon^{1-1/k}$ dependence can be achieved even when the noise is not completely independent of the covariates but satisfies an analytic condition which we refer to as *negatively correlated moments* (see Definition 1.1.11). We provide an outline of the proof as it

illustrates the techniques we introduced in this work.

Coupling and Decoupling. We begin by considering a maximal coupling, $\mathcal{G}$, between distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ such that they disagree on at most an ϵ-measure support (ϵ-fraction of the points for a discrete distribution). Let $(x, y) \sim \mathcal{D}_1$ and $(x', y') \sim \mathcal{D}_2$. Then, observe for any vector v,

$$\begin{aligned}\langle v, \Sigma(\Theta_1 - \Theta_2)\rangle &= \left\langle v, \mathop{\mathbb{E}}_{\mathcal{G}}\left[xx^\top\right](\Theta_1 - \Theta_2)\right\rangle \\ &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(x^\top\Theta_1 - y\right)\right\rangle\right] + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(y - x^\top\Theta_2\right)\right\rangle\right]\end{aligned} \tag{1.1}$$

While the first term in Equation (1.1) depends completely on $\mathcal{D}_1$, the second term requires using the properties of the maximal coupling. Since $1 = 1_{(x,y)=(x',y')} + 1_{(x,y)\neq(x',y')}$, we can rewrite the second term in Equation (1.1) as follows:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(y - x^\top\Theta_2\right)\right\rangle\right] &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - (x')^\top\Theta_2\right)\right\rangle 1_{(x,y)=(x',y')}\right] \\ &\quad + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(y - x^\top\Theta_2\right)\right\rangle 1_{(x,y)\neq(x',y')}\right]\end{aligned} \tag{1.2}$$

With a bit of effort, we can combine Equations (1.1) and (1.2), and upper bound them as follows:

$$\begin{aligned}\langle v, \Sigma(\Theta_1 - \Theta_2)\rangle \leqslant \mathcal{O}(1)\Bigg(&\underbrace{\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(x^\top\Theta_1 - y\right)\right\rangle\right]}_{\text{(i)}} + \underbrace{\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left((x')^\top\Theta_2 - y'\right)\right\rangle\right]}_{\text{(ii)}} \\ &+ \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x\left(y - x^\top\Theta_1\right)\right\rangle 1_{(x,y)\neq(x',y')}\right] \\ &+ \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x'\left(y' - (x')^\top\Theta_2\right)\right\rangle 1_{(x,y)\neq(x',y')}\right]\Bigg)\end{aligned} \tag{1.3}$$

Observe, since we have a maximal coupling, the last two terms appearing in Equation (1.3) are non-zero only on an ϵ-measure support. To bound them, we decouple the indicator using Hölder's inequality,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, x(y - x^\top\Theta_1)\right\rangle 1_{(x,y)\neq(x',y')}\right] &\leqslant \mathbb{E}\left[1_{(x,y)\neq(x',y')}\right]^{\frac{k-1}{k}} \mathbb{E}\left[\langle v, x\rangle^k \left(y - x^\top\Theta_1\right)^k\right]^{\frac{1}{k}} \\ &\leqslant \epsilon^{1-1/k} \cdot \underbrace{\mathbb{E}\left[\langle v, x\rangle^k \left(y - x^\top\Theta_1\right)^k\right]^{\frac{1}{k}}}_{\text{(iii)}} \end{aligned} \tag{1.4}$$

where we used the maximality of the coupling $\mathcal{G}$ to bound $\mathbb{E}\left[1_{(x,y)\neq(x',y')}\right] \leqslant \epsilon$. The above analysis can be repeated verbatim for the second term in (1.3) as well. Going forward, we focus on bounding terms (i), (ii) and (iii).

Gradient Conditions. To bound terms (i) and (ii) in Equation (1.3), we crucially rely on *gradient information* provided by the least-squares objective. Concretely, a key observation in our information-theoretic proof is that the candidate hyperplanes are locally optimal: given least-squares loss, for $i \in \{1, 2\}$ for all vectors v,

$$\left\langle \nabla \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_i}\left[\left(y - x^\top\Theta_i\right)^2\right], v\right\rangle = \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_i}\left[\left\langle v, xx^\top\Theta_i - xy\right\rangle\right] = 0$$

where Θ_1 and Θ_2 are the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. Therefore, both (i) and (ii) are identically 0. It remains to bound (iii).

Independence and Negatively Correlated Moments. We observe that term (iii) can be interpreted as the k-th order correlation between the distribution of the covariates projected along v and the distribution of the noise in the linear model. Here, we observe that if the linear model satisfies the *negatively correlated moments* condition (Definition 1.1.11), we can decouple the expectation and bound each term independently:

$$\mathbb{E}\left[\langle v, x\rangle^k \left(y - x^\top\Theta_1\right)^k\right]^{1/k} \leqslant \mathbb{E}\left[\langle v, x\rangle^k\right]^{1/k} \mathbb{E}\left[\left(y - x^\top\Theta_1\right)^k\right]^{1/k} \tag{1.5}$$

Observe, when the underlying linear model has independent noise, Equation (1.5) follows for any k. We thus crucially exploit the structure of the noise and require a considerably weaker notion than independence. Further, if the *negatively correlated moments* property is not satisfied, we can use Cauchy-Schwarz to decouple the expectation in Equation (1.5) and incur a $\epsilon^{1-2/k}$ dependence. Conceptually, we emphasize that the *negatively correlated moments* condition may be of independent interest to design estimators that exploit independence in various statistics problems.

Hypercontractivity. To bound the RHS in Equation (1.5), we use our central distributional assumption of hypercontractive k-th moments (Definition 1.1.6) of the covariates :

$$\mathbb{E}\left[\langle v, x\rangle^k\right]^{1/k} \leqslant \sqrt{c_k}\,\mathbb{E}\left[\langle v, x\rangle^2\right]^{1/2} = \sqrt{c_k}\,\langle v, \Sigma v\rangle^{1/2}$$

We can bound the noise similarly, by assuming that the noise is hypercontractive and this considerably simplifies our statements. However, hypercontractivity of the noise is not a necessary assumption and prior work indeed incurs a term proportional to the k-th moment of the noise. Assuming boundedness of the regression vectors, Klivans, Kothari and Meka [KKM18] obtained a uniform upper bound on k-th moment of the noise by truncating large samples. We note that the same holds for our estimators and we refer the reader to Section 5.2.3 in their paper. Finally, substituting $v = \Theta_1 - \Theta_2$ and rearranging, completes the information-theoretic proof.

We note that our approach already differs from prior work [KKM18, PSBR20, ZJS19] and to our knowledge, we obtain the first information theoretic proof that being ϵ-close in TV distance implies that the optimal hyperplanes are $\mathcal{O}\left(\epsilon^{1-1/k}\right)$ close in ℓ_2 distance.

Future Directions. We note that our estimators obtain the rate matching recent work for Gaussians, albeit in quasi-polynomial time. In comparison, Diakonikolas, Kong and Stewart [DKS18] obtain the same rate in polynomial time, when the noise is independent of the covariates. This leads to the following question

Open Question 17 (Sub-Gaussian Rates in Polynomial Time)**.** Is there a polynomial time algorithm that achieves $O\left(\epsilon \cdot \text{poly}\left(\log(1/\epsilon)\right)\right)$ rates for all sub-Gaussian distributions? Is any extra $\log(1/\epsilon)$ factor neccesary?

Further, the sample complexity of our estimators scales proportional to $d^{\Omega(k)}$. Such large sample complexity may not be neccesary.

Open Question 18 (Sub-Gaussian Rates in Polynomial Time)**.** Can we achieve the optimal tradeoff between sample complexity, running time and rate for all hypercontractive distributions?

A natural generalization of our work is to consider robust algorithms for Generalized Linear Models, which capture linear, logistic and multi-response regression. Further, such algorithms would pave the way for robust estimators for learning Graphical Models that have received significant attention in various machine learning and computational biology domains. Thus far, obtaining the statistically optimal rate for learning simple Graphical Models remains open, even with unbounded computation [LSS$^+$]. A closely related problem is that of list-decodable re-

gression and subspace recovery, where an overwhelming fraction of data is corrupted (see for example [KKK19, RY20a, RY20b][BK21]). Studying variants of regression and latent variable models in the list-decodable setting is ripe for future work.

1.1.3 List-Decodable Subspace Recovery

List-decodable learning is a strict generalization of related and well-studied *clustering* problems (for e.g., list-decodable mean estimation generalizes clustering spherical mixture models, list-decodable regression generalizes mixed linear regression). In our case, list-decodable subspace recovery generalizes the well-studied problem of subspace clustering where given a mixture of k distributions with covariances non-zero in different subspaces, the goal is to recover the underlying k subspaces [AGGR05, CFZ99, PJAM02]. Algorithms in this model thus naturally yield robust algorithms for the related clustering formulations. In contrast to known results, such algorithms allow "partial recovery" (e.g. for example recovering $k-1$ or fewer clusters) even in the presence of outliers that garble up one or more clusters completely.

Another important implication of list-decodable estimation is algorithms for the *small outlier* model that work whenever the fraction of inliers $\alpha > 1/2$ - the information-theoretic minimum for unique recovery. As a specific corollary, we obtain an algorithm for (uniquely) estimating the subspace spanned by the inlier distribution D whenever $\alpha > 1/2$. We note that if α is sufficiently close to 1, such a result follows from outlier-robust covariance estimation algorithms [DKK+19, LRV16]. While prior works do not specify precise constants, all known works appear to require α at least > 0.75.

List-decodable learning was first proposed in the context of clustering by Balcan, Blum and Vempala [BBV08]. In a recent work, Charikar, Steinhardt and Valiant [CSV17] rejuvenated it as a natural model for algorithmic robust statistics. Most recent works in algorithmic robust statistics have focused on the related but less harsh model of where input data is corrupted by an $\epsilon < 1/2$ fraction outliers. This line of work boasts of some remarkable successes including robust algorithms for computing mean, covariance and higher moments of distributions, clustering mixture models, and performing linear regression in the presence of a small ϵ fraction of adversarial outliers.

While the success hasn't been of the same scale, there has been quite a bit of progress on list-decodable learning that surmount the challenges that arise in dealing with overwhelmingly corrupted data. Recent sequence of works have arrived at a blueprint using the *sum-of-squares method* for list-decodable estimation with applications to list-decodable mean estima-

tion [DKS18, KSS18] and linear regression [KKK19, RY20a].

In the list-decodable subspace recovery problem, our input is a collection of samples $\{x_i\}_{i\in[n]} \in \mathcal{R}^d$, an αn of which are drawn i.i.d. from a distribution $\mathcal{D}$ with mean 0 and unknown projective covariance Π_* of rank k. The main idea of the algorithm is to encode finding the "inliers" in the input sample via a polynomial program. To do this, we introduce variables $w_1, w_2, \ldots, w_n$ that are supposed to indicate the samples that correspond to the inliers. Thus, we force $w_i^2 = w_i$ (i.e. $w_i \in \{0, 1\}$) and $\sum_{i\leqslant n} w_i = \alpha n$ as constraints on w. We also introduce a variable Π that stands for the covariance of the inliers and add constraints that force it to be a projection matrix. To this end, it suffices to constraint $\Pi^2 = \Pi$ and $\Pi \succeq 0$. Further, we require that each of the samples indicated by w are in the subspace described by Π: $w_i(\Pi x_i - x_i) = 0$ for every i.

Recall, an adversary can create multiple rank-k subspaces that satisfy all the aforementioned constraints, and a priori, a solution to the above polynomial program need not tell us anything about the *true* inliers. Therefore, we must force w to share some property that $\mathcal{D}$ satisfies so that we can guarantee a solution to the program contains some information about the inliers. What property should this be? In the context of list-decodable regression [KKK19, RY20a], it turns out that it was both necessary and sufficient (up to the additional qualifier of "certifiability") for w (and $\mathcal{D}$) to be anti-concentrated. Anti-concentration is also *sufficient* to get some guarantees for subspace recovery as shown in [RY20b, BK21]. Is it necessary? And if not, is there a property satisfied by a larger class of distributions that might be sufficient?

Subspace Clustering. A closely related (and formally easier[6]) problem to list-decodable subspace recovery is subspace clustering [EV13, PHL04, SEC14]. Known algorithms with provable guarantees for this problem either require runing time exponential in the ambient dimension, such as RANSAC [FB81], algebraic subspace clustering [VMS05] and spectral curvature clustering [LLY+12], or require the co-dimension to be a constant fraction of the ambient dimension [CSV13, LMZ+12, TV17, ZWR+18].

Robust Subspace Recovery. Our setting also superficially resembles *robust subspace recovery* (see [LM18a] for a survey), where the goal is to recover a set of inliers that span a single low-dimensional space. In this setting, α is assumed to be close to 1. Prior works on this problem identify some tractable special cases (see [VN18]) while no provable guarantees are known for the general setting. Further, Hardt and Moitra [HM13] (see also the recent work of Bhaskara, Chen, Perreault and Vijayraghavan [BCPV19]) provide a polynomial time random-

[6]One can think of input to subspace clustering as the special case in list-decodable subspace recovery where the input sample is a mixture of $k = 1/\alpha$ distributions each with a covariance restricted to some subspace.

ized algorithm, where both the inliers and outliers are required to be in general position and their algorithm works as long as the inliers constitute an $\alpha = r/d$ fraction, where r is the rank of the subspace and d is the ambient dimension. This is contrast to our work where the outliers are completely arbitrary and potentially adversarial with respect to the inliers.

Previous Versions and Concurrent Work. A previous version of this work [BK21] appeared concurrently with [RY20b] and gave a $d^{O(\log r)/\alpha^4}$ time algorithm to output a $O(1/\alpha)$ size list that contains a candidate projection matrix that is $O(\frac{\log r}{\alpha})$-Frobenius close to the rank r projection matrix of the true subspace. The algorithm worked whenever the inlier distribution $\mathcal{D}$ satisfies *certifiable anti-concentration*. This version of the work combines the ideas in [BK21] with multiple new insights to obtain the improved results.

Our Results. Our results apply to input samples generated according to the following model:

Model 19 (Robust Subspace Recovery with Large Outliers)**.** For $0 < \alpha < 1$ and $r < d$, let $\Pi_* \in \mathbb{R}^{d\times d}$ be a projector to a subspace of dimension $r \leqslant d$ and let $\mathcal{D}$ be a distribution on $\mathcal{R}^d$ with mean μ_* and covariance Π_*. Let $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Pi_*)$ denote the following probabilistic process to generate n samples, $x_1, x_2 \dots x_n$ with αn inliers $\mathcal{I}$ and $(1-\alpha)n$ outliers $\mathcal{O}$:

1. Construct $\mathcal{I}$ by choosing αn i.i.d. samples from $\mathcal{D}$.
2. Construct $\mathcal{O}$ by choosing the remaining $(1-\alpha)n$ points arbitrarily and potentially adversarially w.r.t. the inliers.

Remark 20. We will mainly focus on the case when $\mu_* = 0$. The case of non-zero μ_* can be easily reduced to the case of $\mu_* = 0$ by modifying samples by randomly pairing them up and subtracting off samples in each pair (this changes the fraction of inliers from α to α^2).

Our main result is a *fixed* (i.e. exponent of the polynomial does not depend on α) polynomial time algorithm with *dimension-independent* error in *Frobenius norm* - the strongest notion of closeness that implies other guarantees such as the principal angle and spectral distance between subspaces - for list-decodable subspace recovery that succeeds whenever $\mathcal{D}$ has certifiably hypercontractive degree-2 polynomials:

Definition 1.1.13 (Certifiably Hypercontractivity)**.** *A distribution $\mathcal{D}$ on $\mathcal{R}^d$ is said to have $(C, 2h)$-certifiably hypercontractive polynomials if there is a degree-$2h$ sum-of-squares proof in the $d \times d$*

matrix-valued indeterminate Q of the following inequality:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx - \mathbf{E}_{x\sim\mathcal{D}}x^\top Qx\right)^{2h}\right] \leqslant (Ch)^{2h}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx - \mathbf{E}_{x\sim\mathcal{D}}x^\top Qx\right)^{2}\right]^{h}\right),$$

Many natural distributions are certifiably hypercontractive including linear transforms of uniform distribution on the Boolean hypercube and unit sphere, Gaussian distributions, and product distributions with subgaussian marginals. In particular, the set of certifiably hypercontractive distributions is strictly larger than the currently known list of certifiably anti-concentrated distributions (that essentially only holds for rotationally symmetric distributions with sufficiently light tails).

We are now ready to state our main result.

Theorem 21 (Dimension-Independent List-Decodable Subspace Recovery, [BK21])**.** *Let Π_* be a projection matrix for a subspace of dimension r. Let $\mathcal{D}$ be a distribution with mean 0, covariance Π_*, and certifiably $(C, 8)$-hypercontractive polynomials.*

Then, there exists an algorithm takes as input $n = n_0 \geqslant \Omega\left((d\log(d)/\alpha)^{16}\right)$ samples from $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Pi_*)$ and in $O(n^{18})$ time, outputs a list $\mathcal{L}$ of $O(1/\alpha)$ projection matrices such that with probability at least 0.99 over the draw of the sample and the randomness of the algorithm, there is a $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F \leqslant O(1/\alpha)$.

As an immediate corollary, this gives an algorithm for list-decodable subspace recovery when $\mathcal{D}$ is Gaussian, uniform on the unit sphere, uniform on the discrete hypercube/q-ary cube, product distribution with subgaussian marginals and their affine transforms.

Discussion and Comparison with Prior Works Theorem 21 improves on a previous version of [BK21] and on [RY20b] in running time, error guarantees and the generality of the distribution $\mathcal{D}$. In particular, it strictly improves on the work of Raghavendra and Yau who gave an error guarantee of $O(r/\alpha^5)$ in polynomial time by relying on certifiable anti-concentration.[7] It also improves on the guarantee in a previous version of this work for Gaussians that relied on certifiable anti-concentration and an exponential error reduction method to give an error of $O(\log(r)/\alpha)$ in $d^{O(\log r/\alpha^4)}$ time. Unlike Theorem 21, both these algorithms provably cannot extend to the uniform distribution on the hypercube.

A priori, our result might appear surprising and almost too-good-to-be-true. Indeed, prior

[7]The results in [RY20a] handle a small amount of additive noise. The algorithm in this paper can be extended to handle a similar amount of noise but we do not focus on that aspect in this paper.

works identified anti-concentration as a information-theoretic necessary condition on $\mathcal{D}$ for list-decodable regression (a special case of list-decodable subspace recovery) to be feasible. Specifically, Karmalkar, Klivans and Kothari [KKK19] show:

Fact 1.1.14 (Theorem 6.1, Page 19 in [KKK19])**.** *For any constant $\alpha > 0$, there exists a distribution $\mathcal{D}$ (uniform distribution on $\{0,1\}^n$) that is $(\alpha+\epsilon)$-anti-concentrated for every $\epsilon > 0$ but there is no algorithm for $\alpha/2$-approximate list-decodable subspace recovery with rank $r = d-1$ that outputs a list of size $< d$.*

On the other hand, note that discrete product distributions such as uniform distribution on the hypercube/q-ary cube satisfy certifiable hypercontractivity (see [KOTZ14]) so our Theorem 21 applies. This is not a contradiction because of the error guarantees - observe that the Frobenius error bound of $O(1/\alpha)$ provided by Theorem 21 translates to a ℓ_2-norm bound of $O(1/\alpha)$ for linear regression. This is not meaningful for unit vectors, whenever $\alpha \leqslant 1/2$, since even a random unit vector achieves an error of at most $\sqrt{2}$ in this setting. On the other hand, for subspace recovery, this is a non-trivial guarantee whenever the dimension and the co-dimension of the unknown subspace are $\gg 1/\alpha$.

High-Accuracy Subspace Recovery. Our first result naturally raises the question of algorithms obtaining arbitrarily tiny error (instead of $O(1/\alpha)$). For sufficiently small errors ($\ll 1$), $\mathcal{D}$ must necessarily be anti-concentrated, given the lower-bound from Fact 1.1.14 above. Our next result confirms that *certifiable anti-concentration* of $\mathcal{D}$ is sufficient to obtain an arbitrarily small error while still maintaining a list-size of an absolute constant (but of size $1/\alpha^{O(\log(1/\alpha))}$) independent of the dimension.

To state our result, we first recall certifiable anti-concentration from the previous subsection.

Definition 1.1.15 (Certifiable Anti-Concentration)**.** *A zero-mean distribution D with covariance Σ is $2t$-certifiably $(\delta, C\delta)$-anti-concentrated if there exists a univariate polynomial p of degree t such that there is a degree $2t$ sum-of-squares proof in variable v of the following inequalities:*

1. $\|v\|_2^{2t-2} \langle x, v\rangle^2 + \delta^2 p^2(\langle x, v\rangle) \geqslant \frac{\delta^2 \|\Sigma^{1/2} v\|_2^{2t}}{2}$.
2. $\mathbb{E}_{x\sim D}[p^2(\langle x, v\rangle)] \leqslant C\delta \|\Sigma^{1/2} v\|_2^{2t}$.

A subset $\mathcal{S} \subseteq \mathcal{R}^d$ is $2t$-certifiably $(\delta, C\delta)$-anti-concentrated if the uniform distribution on $\mathcal{S}$ is $2t$-certifiably $(\delta, C\delta)$-anti-concentrated.

Gaussian distributions and spherically symmetric distributions with subgaussian tails are

$O(1/\delta^2)$-certifiably $(2,\delta)$-anti-concentrated for every $\delta > 0$ (see Section 5.5).

Theorem 22 (High-Accuracy Subspace Recovery, [BK21])**.** *Let Π_* be a projector to a subspace of dimension r. Let $\mathcal{D}$ be a k-certifiably $(C, \alpha/2C)$-anti-concentrated distribution with certifiably C-hypercontractive degree 2 polynomials.*

Then, there exists an algorithm that takes as input $n = n_0 \geqslant (d\log(d)/\alpha)^{O(k)}$ samples from $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Pi_*)$ and in $n^{O(k+\log(1/\eta))}$ time, outputs a list $\mathcal{L}$ of $O(1/\alpha^{\log k+\log(1/\eta)})$ projection matrices such that with probability at least 0.99 over the draw of the sample and the randomness of the algorithm, there is a $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F \leqslant \eta$.

The proof of Theorem 22 is based on new argument using certifiable anti-concentration that bootstraps our first result with an exponential error reduction mechanism within the sum-of-squares proof system. This improves on the result in a previous version of this work that gave a $d^{O(\log d/\alpha^4)}$ algorithm with a dimension-dependent list size of $d^{O(\log(1/\alpha)}$ based on a somewhat complicated pruning procedure.

Using $O(1/\delta^2)$-certifiable $(\delta, C\delta)$-anti-concentration of Gaussians and spherically symmetric distribution with subgaussian tails, we obtain:

Corollary 1.1.16 (Subspace Recover for Gaussian Inliers, [BK21])**.** *Let Π_* be a projector a subspace of dimension r. Let $\mathcal{D}$ be a mean 0 Gaussian or a spherically symmetric distribution with subgaussian tails with covariance Π_*.*

Then, there exists an algorithm that takes as input $n = n_0 \geqslant (d\log(d)/\alpha^2)^{O(1/\alpha^2)}$ samples from $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Pi_*)$ and in $n^{\log(1/\alpha\eta)/\alpha^4}$ time, outputs a list $\mathcal{L}$ of $O(1/\alpha^{\log 1/\alpha+\log(1/\eta)})$ projection matrices such that with probability at least 0.99 over the draw of the samples and the randomness of the algorithm, there is a $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F \leqslant \eta$.

Further uses of exponential error reduction. Our exponential error reduction method is likely to be of wider use. As an example, we observe the following immediate consequence to list-decodable linear regression by obtaining an improved running time (with a large constant list-size) as a function of the target accuracy.

Corollary 1.1.17 (List-Decodable Regression, [BK21])**.** *Let $\mathcal{D}$ be k-certifiably $(\alpha/2C)$-anti concentrated distribution with mean 0 and covariance I. Then, there exists an algorithm that takes as input $n = n_0 \geqslant (d\log(d)/\alpha)^{\widetilde{O}(k)}$ labeled samples where an αn samples (x, y) are i.i.d. with $x \sim \mathcal{D}$ and $y = \langle \ell_*, x\rangle$ for some unknown, unit vector ℓ_* and outputs a list $\mathcal{L}$ of $O(1/\alpha^{O(\log(k)+\log(1/\eta))})$ regressors such that with probability at least 0.99 over the draw of the*

samples and the randomness of the algorithm, there is a regressor $\hat{\ell} \in \mathcal{L}$ satisfying $\|\hat{\ell} - \ell_*\|_2^2 \leqslant \eta$. The algorithm has time complexity at most $n^{O(k+\log(1/\eta))}$.

Prior works [KKK19, RY20a] needed $n^{O(k^2/\eta^2)}$ time but computed a smaller list of size $O(1/\alpha)$).

Future Directions. Given the resurgence of interest in list-decodable learning, and the limited algorithmic results mentioned above, this area is ripe for future work. In addition to obtaining statistically optimal rates and parameter dependence, it would be interesting to develop a general theory for when list-decodable learning can be performed efficiently. For high-accuracy learning, anti-concentration of the underlying distribution appears to be the key ingredient underlying all efficient estimators.

Open Question 23. Does anti-concentration characterize high-accuracy list-decodable learning?

1.1.4 Learning a Two-Layer Neural Network

Neural networks have achieved remarkable success in solving many modern machine learning problems which were previously considered to be intractable. With the use of neural networks now being wide-spread in numerous communities, the optimization of neural networks is an object of intensive study.

Common usage of neural networks involves running stochastic gradient descent (SGD) with simple non-linear activation functions, such as the extremely popular ReLU function, to learn an incredibly large set of weights. This technique has enjoyed immense success in solving complicated classification tasks with record-breaking accuracy. However, theoretically the behavior and convergence properties of SGD are very poorly understood, and few techniques are known which achieve provable bounds for the training of large neural networks. This is partially due to the hardness of the problem – there are numerous formulations where the problem is known to be NP-hard [BR92, Jud88, BDL18, MR18]. Nevertheless, given the importance and success in solving this problem in practice, it is important to understand the source of this hardness.

Typically a neural network can be written in the following form: $\mathbf{A} = \boldsymbol{U}^i(\cdots \boldsymbol{U}^3 f(\boldsymbol{U}^2 f(\boldsymbol{U}^1 \mathcal{X}))$, where i is the depth of the network, $\mathcal{X} \in \mathbb{R}^{d \times n}$ is a matrix with columns corresponding to individual d-dimensional input samples, and $\mathbf{A}$ is the output labeling of $\mathcal{X}$. The functions f are applied entry-wise to a matrix, and are typically non-linear. Perhaps the most popular activation used in practice is the ReLU, given by $f(x) = \max\{0, x\}$. Here each $\boldsymbol{U}^i$ is an unknown linear

map, representing the "weights", which maps inputs from one layer to the next layer. In the reconstruction problem, when it is known that $\mathbf{A}$ and $\mathcal{X}$ are generated via the above model, the goal is to recover the matrices $\boldsymbol{U}^1, \ldots, \boldsymbol{U}^i$.

In this work, we consider the problem of learning the weights of two layer networks with a single non-linear layer. Such a network can be specified by two weight matrices $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$ and $\boldsymbol{V}^* \in \mathcal{R}^{k \times d}$, such that, on a d-dimensional input vector $x \in \mathcal{R}^d$, the classification of the network is given by $\boldsymbol{U}^* f(\boldsymbol{V}^* x) \in \mathcal{R}^m$. Given a training set $\mathcal{X} \in \mathcal{R}^{d \times n}$ of n examples, along with their labeling $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$, where $\mathbf{E}$ is a (possibly zero) noise matrix, the learning problem is to find $\boldsymbol{U}$ and $\boldsymbol{V}$ for which

$$\|\boldsymbol{U} - \boldsymbol{U}^*\|_F + \|\boldsymbol{V} - \boldsymbol{V}^*\|_F \leqslant \varepsilon$$

We consider two versions of this problem. First, in the noiseless (or realizable) case, we observe $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ precisely. In this setting, we demonstrate that exact recovery of the matrices $\boldsymbol{U}^*, \boldsymbol{V}^*$ is possible in polynomial time. Our algorithms, rather than exploiting smoothness of activation functions, exploit combinatorial properties of rectified activation functions. Additionally, we consider the more general noisy case, where we instead observe $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$, where $\mathbf{E}$ is a noise matrix which can satisfy various conditions. Perhaps the most common assumption in the literature [GKLW18, GLM17, JSA15] is that $\mathbf{E}$ has mean 0 and is sub-Gaussian. Observe that the first condition is equivalent to the statement that $\mathbb{E}[\mathbf{A} \mid \mathcal{X}] = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$. While we primarily focus on designing polynomial time algorithms for this model of noise, in Section 6.7 we demonstrate fixed-parameter tractable (in the number k of ReLUs) algorithms to learn the underlying neural network for a much wider class of noise matrices $\mathbf{E}$. We predominantly consider the *identifiable* case where $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$ has full column rank, however we also provide supplementary algorithms for the exact case when $m < k$. Our algorithms are robust to the behavior of $f(x)$ for positive x, and therefore generalize beyond the ReLU to a wider class of rectified functions f such that $f(x) = 0$ for $x \leqslant 0$ and $f(x) > 0$ otherwise.

It is known that stochastic gradient descent cannot converge to the ground truth parameters when f is ReLU and $\boldsymbol{V}^*$ is orthonormal, even if we have access to an infinite number of samples [LSSS14]. This is consistent with empirical observations and theory, which states that overparameterization is crucial to train neural networks successfully [Har14, SC16]. In contrast, in this work we demonstrate that we can approximate the optimal parameters in the noisy case, and obtain the optimal parameters exactly in the realizable case, in polynomial time, without overparameterization. In other words, we provide algorithms that do not succumb to spurious local

minima, and can converge to the global optimum efficiently, without over-parametrization.

Our Results. We now state our results more formally. We consider 2-layer neural networks with ReLU-activation functions f. Such a neural network is specified by matrices $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$ and $\boldsymbol{V}^* \in \mathcal{R}^{k \times d}$. We are given d-dimensional input examples $x^i \in \mathcal{R}^d$, which form the columns of our input matrix $\mathcal{X}$, and also give the network's m-dimensional classification of $\mathcal{X}$, which is $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$, where f is applied entry-wise. We note that our formulation corresponds to having one non-linear layer.

In the worst case setting, no properties are assumed on the inputs $\mathcal{X}, \mathbf{A}$. While this problem is generally assumed to be intractable, we show, perhaps surprisingly, that when $\text{rank}(\mathbf{A}) = k$ and $k = O(1)$, polynomial time exact algorithms do exist. One of our primary techniques throughout this work is the leveraging of combinatorial aspects of the ReLU function. For a row $f(\boldsymbol{V}^* \mathcal{X})_{i,*}$, we define a *sign pattern* of this row to simply be the subset of positive entries of the row. Thus, a sign pattern of a vector in $\mathcal{R}^n$ is simply given by the orthant of $\mathcal{R}^n$ in which it lies. We first prove an upper bound of $O(n^k)$ on the number of orthants which intersect with an arbitrary k-dimensional subspace of $\mathcal{R}^n$. Next, we show how to enumerate these sign patterns in time $n^{k+O(1)}$.

We use this result to give an $n^{O(k)}$ time algorithm for the neural network learning problem in the *realizable case*, where $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ for some fixed *rank-k* matrices $\boldsymbol{U}^*, \boldsymbol{V}^*$. After fixing a sign pattern of $f(\boldsymbol{V}^* \mathcal{X})$, we can effectively "remove" the non-linearity of f. Even so, the learning problem is still non-convex, and cannot be solved in polynomial time in the general case (even for fixed k). We show, however, that if the *rank* of $\mathbf{A}$ is k, then it is possible to use a sequence of linear programs to recover $\boldsymbol{U}^*, \boldsymbol{V}^*$ in polynomial time given the sign pattern, which allows for an $n^{O(k)}$ overall running time. Our theorem is stated below.

Since non-convex optimization problems are known to be NP-hard in general, it is, perhaps, unsatisfying to settle for worst-case results. Typically, in the learning community, to make problems tractable it is assumed that the input data is drawn from some underlying distribution that may be unknown to the algorithm. So, in the spirit of learning problems, we make the common step of assuming that the samples in $\mathcal{X}$ have a standard Gaussian distribution. More generally, our algorithms work for arbitrary multi-variate Gaussian distributions over the columns of $\mathcal{X}$, as long as the covariance matrix is non-degenerate, i.e., full rank. In this case, our running time and sample complexity will blow up by the condition number of the covariance matrix, which we can estimate first using standard techniques. For simplicity, we state our results here for $\boldsymbol{\Sigma} = \mathbb{I}$, though, for the above reasons, all of our results for Gaussian inputs $\mathcal{X}$ extend to all full rank $\boldsymbol{\Sigma}$

Furthermore, because many of our primary results utilize the combinatorial sparsity patterns of $f(\boldsymbol{V}\mathcal{X})$, where $\mathcal{X}$ is a Gaussian matrix, we do not rely on the fact that $f(x)$ is linear for $x > 0$. For this reason, our results generalize easily to other *non-linear* rectified functions f. In other words, any function f given by

$$f(x) = \begin{cases} 0 & \text{if } x \leqslant 0 \\ \phi(x) & \text{otherwise} \end{cases}$$

where $\phi(x) : [0, \infty] \to [0, \infty]$ is a continuous, injective function. In particular, our bounds do not change for polynomial valued $\phi(x) = x^c$ for $c \in \mathbb{N}$. Note, however, that our worst-case, non-distributional algorithms (stated earlier), where $\mathcal{X}$ is a fixed matrix, do not generalize to non-linear $\phi(x)$.

We first consider the noiseless setting, also referred to as the exact or realizable setting. Here $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ is given for rank k matrices $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$, where $\mathcal{X}$ has non-degenerate Gaussian marginals. The goal is then to recover the weights $(\boldsymbol{U}^*)^T, \boldsymbol{V}^*$ exactly up to a permutation of their rows (since one can always permute both sets of rows without effecting the output of the network). Note that for any positive diagonal matrix $\boldsymbol{D}$, $\boldsymbol{U}^* f(\boldsymbol{D}\boldsymbol{V}^*\mathcal{X}) = \boldsymbol{U}^*\boldsymbol{D} f(\boldsymbol{V}^*\mathcal{X})$ when f is the ReLU. Thus recovery of $(\boldsymbol{U}^*)^T, \boldsymbol{V}^*$ is always only possible up to a permutation and positive scaling. We now state our main theorem for the exact recovery of the weights in the realizable (noiseless) setting.

Theorem 24 (Exact Parameter Recovery, [BJW19])**.** *Suppose* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ *where* $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ *are both rank-*k*, and such that the columns of* $\mathcal{X} \in \mathcal{R}^{d\times n}$ *are mean* 0 *i.i.d. Gaussian. Then if* $n = \Omega(\text{poly}(d, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*)))$*, then there is a* $\text{poly}(n)$*-time algorithm which recovers* $(\boldsymbol{U}^*)^T, \boldsymbol{V}^*$ *exactly up to a permutation of the rows with high probability.*

To the best of our knowledge, this is the first algorithm which learns the weights matrices of a two-layer neural network with ReLU activation *exactly* in the noiseless case and with Gaussian inputs $\mathcal{X}$. Our algorithm first obtains good approximations to the weights $\boldsymbol{U}^*, \boldsymbol{V}^*$, and concludes by solving a system of judiciously chosen linear equations, which we solve using Gaussian elimination. Therefore, we obtain exact solutions in polynomial time, without needing to deal with convergence guarantees of continuous optimization primitives. Furthermore, to demonstrate the robustness of our techniques, we show that using results introduced in the concurrent and independent work of Ge et. al. [GKLW18], we can extend Theorem 24 to hold for inputs sampled from symmetric distributions. We note that [GKLW18] recovers the weight matrices up to additive error ε and runs in poly $\left(\frac{1}{\varepsilon}\right)$-time, whereas our algorithm has no ε dependency.

The runtime of our algorithm depends on the condition number $\kappa(\boldsymbol{V}^*)$ of $\boldsymbol{V}^*$, which is a fairly ubiquitous requirement in the literature for learning neural networks, and optimization in general [GKLW18, JSA15, CMTV17, AGMR17, ZSJ+17, SJA16]. To address this dependency, we give a lower bound which shows at least a linear dependence on $\kappa(\boldsymbol{V}^*)$ is necessary in the sample and time complexity.

Next, we introduce an algorithm for approximate recovery of the weight matrices $\boldsymbol{U}^*, \boldsymbol{V}^*$ when $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$ for Gaussian marginals $\mathcal{X}$ and an i.i.d. sub-Gaussian mean-zero noise matrix $\mathbf{E}$ with variance σ^2.

Theorem 25 (Noisy Parameter Recovery, [BJW19])**.** *Let $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$ be given, where $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}, \boldsymbol{V}^* \in \mathcal{R}^{k \times d}$ are rank-k, $\mathbf{E}$ is a matrix of i.i.d. mean-zero sub-Gaussian random variables with variance σ^2, and such that the columns of $\mathcal{X} \in \mathcal{R}^{d \times n}$ are i.i.d. Gaussian. Then given $n = \Omega\Big(poly\Big(d, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*), \sigma, \frac{1}{\varepsilon}\Big)\Big)$, there is an algorithm that runs in poly(n) time and w.h.p. outputs $\boldsymbol{V}, \boldsymbol{U}$ such that*

$$\|\boldsymbol{U} - \boldsymbol{U}^*\|_F \leqslant \varepsilon, \qquad \|\boldsymbol{V} - \boldsymbol{V}^*\|_F \leqslant \varepsilon$$

Again, to the best of our knowledge, this work is the first which learns the weights of a 2-layer network in this noisy setting without additional constraints, such as the restriction that $\boldsymbol{U}$ be positive. Recent independent and concurrent work, using different techniques, achieves similar approximate recovery results in the noisy setting [GKLW18]. We note that the algorithm of Goel et. al. [GK17] that [GKLW18] uses, crucially requires the linearity of the ReLU for $x > 0$, and thus the work of [GKLW18] does not generalize to the larger class of rectified functions which we handle. We also note that the algorithm of [GLM17] requires $\boldsymbol{U}^*$ to be non-negative. Finally, the algorithms presented in [JSA15] work for activation functions that are thrice differentiable and can only recover rows of $\boldsymbol{V}^*$ up to ± 1 scaling. Note, for the ReLU activation function, we need to resolve the signs of each row.

One of the primary technical contributions of this work is the utilization of the combinatorial structure of sparsity patterns of the rows of $f(\boldsymbol{V}\mathcal{X})$, where f is a rectified function, to solve learning problems. Here, a sparsity pattern refers to the subset of coordinates of $f(\boldsymbol{V}\mathcal{X})$ which are non-zero, and a rectified function f is one which satisfies $f(x) = 0$ for $x \leqslant 0$, and $f(x) > 0$ otherwise.

Overview. In response to the aforementioned hardness results, we relax to the case where the input $\mathcal{X}$ has Gaussian marginals. In the noiseless case, we *exactly* learn the weights $\boldsymbol{U}^*, \boldsymbol{V}^*$

given $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ (up to a positive scaling and permutation). As mentioned, our results utilize analysis of the sparsity patterns in the row-span of $\mathbf{A}$. One benefit of these techniques is that they are largely insensitive to the behavior of $f(x)$ for positive x, and instead rely on the rectified property $f(\cdot)$. Hence, this can include even exponential functions, and not solely the ReLU.

Our exact recovery algorithms proceed in two steps. First, we obtain an approximate version of the matrix $f(\boldsymbol{V}^* \mathcal{X})$. For a good enough approximation, we can exactly recover the sparsity pattern of $f(\boldsymbol{V}^* \mathcal{X})$. Our main insight is, roughly, that the only sparse vectors in the row span of $\mathbf{A}$ are precisely the rows of $f(\boldsymbol{V}^* \mathcal{X})$. Specifically, we show that the only vectors in the row span which have the same sparsity pattern as a row of $f(\boldsymbol{V}^* \mathcal{X})$ are scalar multiples of that row. Moreover, we show that no vector in the row span of $\mathbf{A}$ is supported on a strict subset of the support of a given row of $f(\boldsymbol{V}^* \mathcal{X})$. Using these facts, we can then set up a judiciously designed linear system to find these vectors, which allows us to recover $f(\boldsymbol{V}^* \mathcal{X})$ and then $\boldsymbol{V}^*$ exactly. By solving linear systems, we avoid using iterative continuous optimization methods, which recover a solution up to additive error ε and would only provide rates of convergence in terms of ε. In contrast, Gaussian elimination yields exact solutions in a polynomial number of arithmetic operations.

The first step, finding a good approximation of $f(\boldsymbol{V}^* \mathcal{X})$, can be approached from multiple angles. In this work, we demonstrate two different techniques to obtain these approximations, the first being Independent Component Analysis (ICA), and the second being tensor decomposition. To illustrate the robustness of our exact recovery procedure once a good estimate of $f(\boldsymbol{V}^* \mathcal{X})$ is known, we show in Section 6.4.3 how we can bootstrap the estimators of recent, concurrent and independent work [GKLW18], to improve them from approximate recovery to exact recovery.

In the restricted case when $\boldsymbol{V}^*$ is orthonormal, we show that our problem can be modeled as a special case of *Independent Component Analysis* (ICA). The ICA problem approximately recovers a subspace $\mathbf{B}$, given that the algorithm observes samples of the form $y = \mathbf{B}x + \zeta$, where x is i.i.d. and drawn from a distribution that has moments bounded away from Gaussians, and ζ is a Gaussian noise vector. Intuitively, the goal of ICA is to find a linear transformation of the data such that each of the coordinates or features are as independent as possible. By rotational invariance of Gaussians, in this case $\boldsymbol{V}^* \mathcal{X}$ is also i.i.d. Gaussian, and we know that the columns of $f(\boldsymbol{V}^* \mathcal{X})$ have independent components and moments bounded away from a Gaussian. Thus, in the orthonormal case, our problem is well suited for the ICA framework.

A second, more general approach to approximating $f(\boldsymbol{V}^* \mathcal{X})$ is to utilize techniques from *tensor decomposition*. Our starting point is the generative model considered by Janzamin et.

al. [JSA15], which matches our setting, i.e., $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$. The main idea behind this algorithm is to construct a tensor that is a function of both $\mathbf{A}, \mathcal{X}$ and captures non-linear correlations between them. A key step is to show that the resulting tensor has low CP-rank and the low-rank components actually capture the rows of the weight matrix $\boldsymbol{V}^*$. Intuitively, working with higher order tensors is necessary since matrix decompositions are only identifiable up to orthogonal components, whereas tensors have identifiable non-orthogonal components, and we are specifically interested in recovering approximations for non-orthonormal $\boldsymbol{V}^*$.

Next, we run a tensor decomposition algorithm to recover the low-rank components of the resulting tensor. While computing a tensor decomposition is NP-hard in general [HL13], there is a plethora of work on special cases, where computing such decompositions is tractable [BCMV14, SWZ16, WA16, GVX14, GM15, BM16]. Tensor decomposition algorithms have recently become an invaluable algorithmic primitive and with applications in statistical and machine learning [JSA15, JSA14, GLM17, AGHK14a, BKS15].

However, there are several technical hurdles involved in utilizing tensor decompositions to obtain estimates of $\boldsymbol{V}^*$. The first is that standard analysis of these methods utilizes a generalized version of *Stein's Lemma* to compute the expected value of the tensor, which relies on the smoothness of the activation function. Thus, we first approximate $f(\cdot)$ closely using a Chebyshev polynomial $p(\cdot)$ on a sufficiently large domain. However, we cannot algorithmically manipulate the input to demand that $\mathbf{A}$ instead be generated as $\boldsymbol{U}^* p(\boldsymbol{V}^* \mathcal{X})$. Instead, we add a small mean-zero Gaussian perturbation to our samples and analyze the variation distance between $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \boldsymbol{G}$ and $\boldsymbol{U}^* p(\boldsymbol{V}^* \mathcal{X}) + \boldsymbol{G}$. For a good enough approximation p, this variation distance will be too small for any algorithm to distinguish between them, thus standard arguments imply the success of tensor decomposition algorithms when given the inputs $\mathbf{A} + \boldsymbol{G}$ and $\mathcal{X}$.

Next, a key step is to construct a non-linear transformation of the input by utilizing knowledge about the underlying density function for the distribution of $\mathcal{X}$, which we denote by $p(x)$. The non-linear function considered is the so-called Score Function, defined in [JSA14], which is the normalized m-th order derivative of the input probability distribution function $p(x)$. Computing the score function for an arbitrary distribution can be computationally challenging. However, as mentioned in [JSA14], we can use Hermite polynomials that help us compute a closed form for the score function, in the special case when $x \sim \mathcal{N}(0, \mathbf{I})$.

A further complication arises due to the fact that this form of tensor decomposition is agnostic to the signs of $\boldsymbol{V}$. Namely, we are guaranteed vectors v_i from tensor decomposition such that $\|v_i - \xi_i \boldsymbol{V}^*_{i,*}\|_F < \varepsilon$, where $\xi_i \in \{1, -1\}$ is some unknown sign. Prior works have dealt with

this issue by considering restricted classes of smooth activation functions which satisfy $f(x) = 1 - f(-x)$ [JSA15]. For such functions, one can compensate for not knowing the signs by allowing for an additional affine transformation in the neural network. Since we consider non-affine networks and rectified functions $f(\cdot)$ which do not satisfy this restriction, we must develop new methods to recover the signs ξ_i to avoid the exponential blow-up needed to simply guess them.

For the noiseless case, if v_i is close enough to $\xi_i \boldsymbol{V}^*_{i,*}$, we can employ our previous results on the uniqueness of sparsity patterns in the row-span of $\mathbf{A}$. Namely, we can show that the sparsity pattern of $f(\xi v_i)$ will in fact be feasible in the row-span of $\mathbf{A}$, whereas the sparsity pattern of $f(-\xi v_i)$ will not, from which we recover the signs ξ_i via a linear system.

In the presence of noise, however, the problem becomes substantially more complicated. Because we do not have the true row-span of $f(\boldsymbol{V}^*\mathcal{X})$, but instead a noisy row-span given by $\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X}) + \mathbf{E}$, we cannot recover the ξ_i's by feasibility arguments involving sparsity patterns. Our solution to the sign ambiguity in the noisy case is a projection-based scheme. Our scheme for determining ξ_i involves constructing a $2k-2$ dimensional subspace S, spanned by vectors of the form $f(\pm v_j \mathcal{X})$ for all $j \neq i$. We augment this subspace as $S^1 = S \cup \{f(v_i\mathcal{X})\}$ and $S^{-1} = S \cup \{f(-v_i\mathcal{X})\}$. We then claim that the length of the projections of the rows of $\mathbf{A}$ onto the S^ξ will be *smaller* for $\xi = \xi_i$ than for $\xi = -\xi_i$. Thus by averaging the projections of the rows of $\mathbf{A}$ onto these subspaces and finding the subspace which has the smaller projection length on average, we can recover the ξ_i's with high probability. Our analysis involves bounds on projections onto perturbed subspaces, and a spectral analysis of the matrices $f(\boldsymbol{W}\mathcal{X})$, where $\boldsymbol{W}$ is composed of up to $2k$ rows of the form $\boldsymbol{V}^*_{i,*}$ and $-\boldsymbol{V}^*_{i,*}$.

Future Directions. There has been a significant amount of progress on designing algorithms for provably learning the parameters of two layer (and deeper) neural networks in the years since our work was published [AZL19, JMM20, DK20, DKKZ20, ATV21, AAK21, CGKM22, CKM22]. However, some basic algorithmic questions in the simplest possible setting remain open:

Open Question 26. Given a two layer neural network $y = U^* f(V^* X) + \zeta$, where X is drawn from a sub-Gaussian distribution, ζ is mean zero independent noise, and U^* is a $1 \times k$ matrix, can we learn some neural network that has small labeling error in time that is polynomial in all input parameters, and independent of the condition number?

We note that our results require the output dimension to be larger than the number of neurons

in the hidden layer, and the running time scales proportional to the condition number of U^* and V^*. In the easier PAC learning setting, we do not require recovering the parameters up to small error, and thus could get away without incurring any condition number dependence. For deeper layers, the best known algorithm [] requires an exponential dependence on the lipschitz constant of the network, and obtaining a polynomial dependence remains open. We conclude with another open question on robustly learning two-layer neural networks, which may be of significant practical interest as well.

Open Question 27. Is there a polynomial time algorithm to learn a two-layer neural network under the strong contamination model, i.e. $(1-\epsilon)$-fraction of the samples are drawn i.i.d. from a Gaussian (or any other known) distribution and the remaining ϵ-fraction are arbitrarily chosen by an adversary?

Natural variants of the above formulation are also open.

1.2 Nearly Optimal Algorithms for Learning Latent Models

In the second half of this thesis we consider learning simple latent models that already admit polynomial time algorithms. We show that we can obtain nearly optimal algorithms for (a) low-rank approximation under any Schatten-p norm , (b) low-rank approximation of positive semi-definite and Euclidean distance matrices and (c) learning a latent simplex. Low-rank approximation under various Schatten-p norms has been vastly studied over the last two decades and the numerous algorithms have been obtained based on *sketching methods* and *iterative methods*, however obtaining optimal algorithms for this family of optimization problems has remained a central open question [Woo14a]. Studying questions in numerical linear algebra where the input matrix is drawn from a structured family has also received a lot of attention in recent years. In particular, several works have considered solving linear systems for Laplacian/Diagonally Dominant matrices [ST14, KOSZ13, KMP14] and Block Henkel matrices [PV21], covariance estimation of Toeplitz matrices [ELMM20], and approximation the permanent of boolean [JS89], non-negative matrices [JSV04] and PSD [AGGS17, YP21] matrices, and low-rank approximation for PSD [MW17b] and distance matrices [BW18]. However, obtaining optimal algorithms for these questions remain open. Finally, the latent simplex framework was recently formalized as a way to capture several well-studied latent models, such as the stochastic block model and latent dirichelet allocation [BK20d]. However, obtaining a truly input-sparsity time algorithm for this problem remained open.

We begin by providing an overview of an optimal matrix-vector product algorithm for low-rank approximation under Schatten-p norms, for all constant p. Next, we describe an optimal *sub-linear time* algorithm for computing a low-rank approximation when the input matrix is promised to be PSD or a Euclidean distance matrix. We then conclude by describing a truly *input-sparsity* time algorithm for learning a latent simplex.

1.2.1 Low-Rank Approximation for Schatten Norms

Iterative methods, and in particular Krylov subspace methods, are ubiquitous in scientific computing. Algorithms such as power iteration, Golub-Kahan Bidiagonalization, Arnoldi iteration, and the Lanczos iteration, are used in basic subroutines for matrix inversion, solving linear systems, linear programming, low-rank approximation, and numerous other fundamental linear algebra primitives [Saa81, LS13]. A common technique in the analysis of Krylov methods is the use of Chebyshev polynomials, which can be applied to the singular values of a matrix to implement an approximate interval or step function [MH02, Riv20]. Further, Chebyshev polynomials reduce the degree required to accurately approximate such functions, leading to significantly fewer iterations and faster running time. We investigate the power of Krylov methods for low-rank approximation in the matrix-vector product model.

The Matrix-Vector Product Model. In this model, there is an underlying matrix $\mathbf{A}$, which is often implicit, and for which the only access to $\mathbf{A}$ is via matrix-vector products. Namely, the algorithm chooses a query vector v^1, obtains the product $\mathbf{A} \cdot v^1$, chooses the next query vector v^2, which is any randomized function of v^1 and $\mathbf{A} \cdot v^1$, then receives $\mathbf{A} \cdot v^2$, and so on. If $\mathbf{A}$ is a non-symmetric matrix, we assume access to products of the form $\mathbf{A}^\top v$ as well. We refer to the minimal number q of queries needed by the algorithm to solve a problem with constant probability as the *query complexity*. We note that upper bounds on the query complexity immediately translate to running time bounds for the RAM model, when $\mathbf{A}$ is explicit, since a matrix-vector product can be implemented in $\text{nnz}(\mathbf{A})$ time, i.e., the number of non-zero entries in the matrix. Since this model captures a large family of iterative methods, it is natural to ask whether Krylov subspace based methods yield optimal algorithms, where the complexity measure of interest is the number of matrix-vector products.

This model and related vector-matrix-vector query models were formalized for a number of problems in [SWYZ19, RWZ20], though the model is standard for measuring efficiency in scientific computing and numerical linear algebra, see, e.g., [BFG96]; in that literature, methods that use only matrix-vector products are called *matrix-free*. Subsequently, for the problem

of estimating the top eigenvector, nearly tight bounds were obtained in [SAR18, BHSW20]. Also, for the problem of estimating the trace of a positive semidefinite matrix, tight bounds were obtained in [MMMW21] (see, also [WWZ14], where tight bounds were shown in the related vector-matrix-vector query model). For recovering a planted clique from a random graph, upper and lower bounds were obtained in [RWYZ21]. In the non-adaptive setting, where $v^1, \ldots, v^q$, are chosen before making any queries to $\mathbf{A}$, this is equivalent to the *sketching model*, which is thoroughly studied on its own (see, e.g., [Nel11, Woo14b]), and in the context of data streams [Mut05, LNW14b].

Why is the matrix $\mathbf{A}$ implicit? A small query complexity q leads to an algorithm running in time $\mathcal{O}(T(\mathbf{A}) \cdot q + P(n, d, q))$, where $T(\mathbf{A})$ is the time to multiply the $n \times d$ matrix $\mathbf{A}$ by an arbitrary vector, and $P(n, d, q)$ is the time needed to form the queries and process the query responses, which is typically small. When the matrix $\mathbf{A}$ is given as a list of $\text{nnz}(\mathbf{A})$ non-zero entries, then $T(\mathbf{A}) \leqslant \text{nnz}(\mathbf{A})$. However, in many problems $\mathbf{A}$ is not given explicitly, and it is too expensive to write $\mathbf{A}$ down. Indeed, one may be given $\mathbf{A}$ but want to compute a low-rank approximation to the "covariance" (Gram) matrix $\mathbf{A}^\top\mathbf{A}$, and computing $\mathbf{A}^\top\mathbf{A}$ is too slow [MW17a]. More generally, one may be given $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and a function $f : \mathbb{R} \to \mathbb{R}$, and want to compute matrix-vector products with the generalized matrix function $f(\mathbf{A}) = \mathbf{U}f(\mathbf{\Sigma})\mathbf{V}^\top$, where $\mathbf{U}$ has orthonormal columns, $\mathbf{V}^\top$ has orthonormal rows, $\mathbf{\Sigma}$ is a diagonal matrix, and f is applied entry-wise to each entry on the diagonal.

The covariance matrix corresponds to $f(x) = x^2$, and other common functions f include the matrix exponential $f(x) = e^x$ and low-degree polynomials. For instance, when $\mathbf{A}$ is the adjacency matrix of an undirected graph, $f(x) = x^3/6$ is used to count the number of triangles [Tso08, Avr10]. Yet another example is when $\mathbf{A}$ is the Hessian $\mathbf{H}$ of a neural network with a huge number of parameters, for which it is often impossible to compute or store the entire Hessian [GKX19]. Typically $\mathbf{H} \cdot v$, for any chosen vector v, is computed using Pearlmutter's trick [Pea94]. However, even with Pearlmutter's trick and distributed computation on modern GPUs, it takes 20 hours to compute the eigendensity of a single Hessian $\mathbf{H}$ with respect to the cross-entropy loss on the CIFAR-10 dataset from a set of fixed weights for ResNet-18 [KH+09], which has approximately 11 million parameters [HZRS16, GKX19]. This time is directly proportional to the number of matrix-vector products, and therefore minimizing this quantity is crucial.

Algorithms and Lower Bounds for Low-Rank Approximation. The low-rank approximation problem is well studied in numerical linear algebra, with countless applications to clustering,

data mining, principal component analysis, recommendation systems, and many more. (For surveys on low-rank approximation, see the monographs [KV09, Mah11, Woo14b] and references therein.) In this problem, given an implicit $n \times d$ matrix $\mathbf{A}$, the goal is to output a matrix $\mathbf{Z} \in \mathbb{R}^{d \times k}$ with orthonormal columns such that

$$\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_X \leqslant (1+\epsilon) \min_{\mathbf{U}:\mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \|\mathbf{A}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\|_X, \tag{1.6}$$

where $\|\cdot\|_X$ denotes some norm. Note that given $\mathbf{Z}$, one can compute $\mathbf{AZ}$ with an additional k queries, which will be negligible, and then $(\mathbf{AZ}) \cdot \mathbf{Z}^\top$ is a rank-k matrix written in factored form, i.e., as the product of an $n \times k$ matrix and a $k \times d$ matrix. Among other things, low-rank approximation provides (1) a compression of $\mathbf{A}$ from nd parameters to $(n+d)k$ parameters, (2) faster matrix-vector products, since $\mathbf{AZ} \cdot \mathbf{Z}^\top \cdot y$ can be computed in $O((n+d)k)$ time for an arbitrary vector y, as opposed to the $O(nd)$ time needed to compute $\mathbf{A} \cdot y$, and (3) de-noising, as often matrices $\mathbf{A}$ are close to low-rank (e.g., they are the product of latent factors) but only high rank due to noise.

Despite its tremendous importance, the optimal matrix-vector product complexity of low-rank approximation is unknown for any commonly used norm. The best known upper bound is due to Musco and Musco [MM15], who achieve $\tilde{\mathcal{O}}(k/\epsilon^{1/2})$ queries[8] for both the case when $\|\cdot\|_X$ is the commonly studied Frobenius norm $\|\mathbf{B}\|_F = \left(\sum_{i,j} \mathbf{B}_{i,j}^2\right)^{1/2}$ as well as when $\|\cdot\|_X$ is the Spectral (operator) norm $\|\mathbf{B}\|_2 = \sup_{\|y\|_2=1} \|\mathbf{B}y\|_2$.

On the lower bound front, there is a trivial lower bound of k, since $\mathbf{A}$ may be full rank and achieving (7.1) requires k matrix-vector products since one must reconstruct the column span of $\mathbf{A}$ exactly. However, *no lower bounds in terms of the approximation factor ϵ were known*. We note that Simchowitz, Alaoui and Recht [SAR18] prove lower bounds for approximating the top r eigenvalues of a symmetric matrix; however these guarantees are incomparable to those that follow from a low-rank approximation, even when the norm $\|\cdot\|_X$ is the operator norm.

Relationship to the Sketching Literature. Low-rank approximation has been extensively studied in the sketching literature which, when $\mathbf{A}$ is given explicitly, can achieve $\mathcal{O}(\text{nnz}(\mathbf{A}))$ time both for the Frobenius norm [CW13, MM13a, NN13a], as well as for Schatten-p norms [LW20]. However, these works require reading all of the entries in $\mathbf{A}$, and thus do not apply to any of the settings mentioned above. Further, the matrix-vector query model is especially important for problems such as trace estimation, where a low-rank approximation is used to first reduce the variance [MMMW21]. As trace estimation is often applied to implicit matri-

[8] We let $\tilde{\mathcal{O}}(f) = f \cdot \text{poly}(\log(dk/\epsilon))$.

ces, e.g., in computing Stochastic Lanczos Quadrature (SLQ) for Hessian eigendensity estimation [GKX19], in studying the effects of batch normalization and residual connections in neural networks [YGKM20], and in computing a disentanglement regularizer for deep generative models [PPZ+20], sketching algorithms for low-rank approximation often do not apply.

Another important application is low-rank approximation of covariance matrices [MW17a], for which the covariance matrix is not given explicitly. Here, we have a data matrix $\mathbf{A}$ and we want a low-rank approximation for $\mathbf{A}\mathbf{A}^\top$. Even when $\mathbf{S}$ is a sparse sketching matrix, the matrix $\mathbf{S}\mathbf{A}$ is no longer sparse, and one needs to multiply $\mathbf{S}\mathbf{A}$ by $\mathbf{A}^\top$ to obtain a sketch of $\mathbf{S}\mathbf{A}\mathbf{A}^\top$, which is a dense matrix-matrix multiplication. Moreover, when viewed in the matrix-vector product model, sketching algorithms obtain provably worse query complexity than existing iterative algorithms (see Table 1.2 for a comparison). Further, as modern GPUs often do not exploit sparsity, *even when the matrix* $\mathbf{A}$ *is given, a GPU may not be able to take advantage of sparse queries*, which means the total time taken is proportional to the number of matrix-vector products.

Motivating Schatten-p Norms. The Schatten norms for $1 \leqslant p < 2$ are more robust than the Frobenius norm, as they dampen the effect of large singular values. In particular, the Schatten-1 norm, also known as the nuclear norm, has been widely used for robust PCA [XCS10, CLMW11, YPCC16] as well as a convex relaxation of matrix rank in matrix completion [CR09, CP10], low-dimensional Euclidean embeddings [RFP10, TDSL00, RS00], image denoising [GZZF14, GXM+17] and tensor completion [YZ16]. In contrast, for $p > 2$, Schatten norms are more sensitive to large singular values and provide an approximation to the operator norm. In particular, for a rank r matrix, it is easy to see that setting $p = \log(r)/\eta$ yields a $(1+\eta)$-approximation to the operator norm (i.e., $p = \infty$). While the Block Krylov algorithm of Musco and Musco [MM15] implies a matrix-vector query upper bound of $\tilde{\mathcal{O}}\left(k/\epsilon^{1/2}\right)$ for Schatten-∞ low-rank approximation, the exact complexity of this problem remains an outstanding open problem. When $p > 2$, we can interpolate between Frobenius and operator norm, and setting p to be a large fixed constant can be a proxy for Schatten-∞ low-rank approximation, with significantly fewer matrix-vector products (see Theorem 28).

Our Central Question. The main question of our work is:

What is the matrix-vector product complexity of low-rank approximation for the Frobenius norm, and more generally, for other matrix norms?

Problem	Frobenius	Schatten-p, $p \in [1,2)$	Schatten-p, $p > 2$
Sketching [CW09, LW20]	$\Theta(k/\epsilon)$	$\Omega(k^{2/p}/\epsilon^{4/p+1})$	$\Omega(\min(n,d)^{1-2/p})$
Block Krylov [MM15]	$\tilde{\mathcal{O}}(k/\epsilon^{1/2})$	N/A	N/A
Our Upper Bound	$\tilde{\mathcal{O}}(k/\epsilon^{1/3})$	$\tilde{\mathcal{O}}(k/\epsilon^{1/3})$	$\tilde{\mathcal{O}}(kp^{1/6}/\epsilon^{1/3})$
Our Lower Bound	$\Omega(1/\epsilon^{1/3})$	$\Omega(1/\epsilon^{1/3})$	$\Omega(1/\epsilon^{1/3})$

Figure 1.2: Prior Upper and Lower Bounds on the Matrix Vector Product Complexity for Frobenius and Schatten-p low-rank Approximation. The poly(k/ϵ) factors in prior sketching work for Schatten-p are not explicit, but we have computed lower bounds on them to illustrate our improvements. Our bounds are optimal, up to logarithmic factors, for constant k. For $p > \log(d)/\epsilon$, spectral low-rank approximation [MM15] implies an $\tilde{\mathcal{O}}\left(k/\sqrt{\epsilon}\right)$ upper bound.

Our Results. We begin by stating our results for Frobenius and more generally, Schatten-p norm low-rank approximation for any $p \geqslant 1$; see Table 1.2 for a summary.

Theorem 28 (Query Upper Bound, [BCW22])**.** *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, a target rank $k \in [d]$, an accuracy parameter $\epsilon \in (0,1)$ and any (not necessarily constant) $p \in [1, \mathcal{O}(\log(d)/\epsilon)]$, there exists an algorithm that uses $\tilde{\mathcal{O}}\left(kp^{1/6}/\epsilon^{1/3}\right)$ matrix-vector products and outputs a $d \times k$ matrix $\mathbf{Z}$ with orthonormal columns such that with probability at least $99/100$,*

$$\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} \leqslant (1+\epsilon) \min_{\mathbf{U}:\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \|\mathbf{A}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\|_{\mathcal{S}_p}.$$

When $p \geqslant \log(d)/\epsilon$, we get $\tilde{\mathcal{O}}\left(k/\sqrt{\epsilon}\right)$ matrix-vector products.

We note that for Frobenius norm low-rank approximation (Schatten p for $p = 2$), we improve the prior matrix-vector product bound of $\tilde{\mathcal{O}}(k/\epsilon^{1/2})$ by Musco and Musco [MM15] to $\tilde{\mathcal{O}}(k/\epsilon^{1/3})$. For Schatten-$p$ low-rank approximation for $p \in [1,2)$, we improve work of Li and Woodruff [LW20] who require query complexity at least $\Omega(k^{2/p}/\epsilon^{4/p+1})$, which is a polynomial factor worse in both k and $1/\epsilon$ than our $\tilde{\mathcal{O}}(k/\epsilon^{1/3})$ bound.

For $p > 2$, [LW20] obtain a query complexity of $\Omega(\min(n,d)^{1-2/p})$. We drastically improve this to $\tilde{\mathcal{O}}(k/\epsilon^{1/3})$, which does not depend on d or n at all. Setting $p = \log(d)/\epsilon$ suffices to obtain a $(1+\epsilon)$-approximation to the spectral norm ($p = \infty$), and we obtain an $\tilde{\mathcal{O}}\left(k/\sqrt{\epsilon}\right)$ query algorithm, matching the best known bounds for spectral low-rank approximation [MM15]. When $p > \log(d)/\epsilon$, we can simply run Block Krylov for $p = \infty$.

Remark 29 (Comments on the RAM Model)**.** Although our focus is on minimizing the num-

ber of matrix-vector products, which is the key resource in the applications described above, our bounds also improve the running time of low-rank approximation algorithms when the matrix $\mathbf{A}$ has a small number of non-zero entries and is explicitly given. For simplicity, we state our bounds and those of previous work without using algorithms for fast matrix multiplication; similar improvements hold when using such algorithms. When $\text{nnz}(\mathbf{A}) = O(n)$, for Frobenius norm low-rank approximation, work in the sketching literature, and in particular [ACW17] (building off of [CW13, NN13a, Coh16]), achieves $O(nk^2/\epsilon)$ time. In contrast, in this setting our runtime is $\tilde{\mathcal{O}}(nk^2/\epsilon^{2/3})$. Similarly, for Schatten-p low-rank approximation for $p \in [1, 2)$, the previous best [LW20] requires $\tilde{\Omega}(nk^{4/p}/\epsilon^{(8/p-2)})$ time, while for $p > 2$ [LW20] requires $\tilde{\Omega}(nd^{2(1-2/p)}(k/\epsilon)^{4/p})$ time. In both cases our runtime is only $\tilde{\mathcal{O}}(nk^2p^{1/3}/\epsilon^{2/3})$. We obtain analogous improvements when the sparsity $\text{nnz}(\mathbf{A})$ is allowed to be $n(k/\epsilon)^C$ for a small constant $C > 0$.

Next, we state our lower bounds on the matrix-vector query complexity of Schatten-p low-rank approximation.

Theorem 30 (Query Lower Bound for constant p, [BCW22])**.** *Given $\varepsilon > 0$, and a fixed constant $p \geqslant 1$, there exists a distribution $\mathcal{D}$ over $n \times n$ matrices such that for $\mathbf{A} \sim \mathcal{D}$, any algorithm that with at least constant probability outputs a unit vector v such that $\|\mathbf{A}\left(\mathbf{I} - vv^\top\right)\|_{\mathcal{S}_p}^p \leqslant (1+\varepsilon)\min_{\|u\|_2=1}\|\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p}^p$ must perform $\Omega(1/\varepsilon^{1/3})$ matrix-vector queries to $\mathbf{A}$.*

Remark 31. We note that this is the first lower bound as a function of ϵ for this problem, even for the well-studied case of $p = 2$, achieving an $\Omega(1/\epsilon^{1/3})$ bound, which is tight for any constant k, simultaneously for all constant $p \geqslant 1$.

Remark 32. Braverman, Hazan, Simchowitz and Woodworth [BHSW20] and Simchowitz, Alaoui and Recht [SAR18] establish eigenvalue estimation lower bounds that we use in our arguments, but their results do not directly imply low-rank approximation lower bounds for any matrix norm that we are aware of, including spectral low-rank approximation, i.e., $p = \infty$.

Overview. We first describe our algorithm for the special case of rank-1 approximation in the Frobenius norm, i.e., $p = 2$. Our algorithm is inspired by the Block Krylov algorithm of Musco and Musco [MM15]. Briefly, their algorithm begins with a random starting vector g (block size is 1) and computes the Krylov subspace $\mathbb{K} = [\mathbf{A}g; \mathbf{A}^2g; \ldots; \mathbf{A}^qg]$, for $q = \mathcal{O}\left(1/\epsilon^{1/2}\right)$. Next, their algorithm computes an orthonormal basis for the column span of $\mathbb{K}$, denoted by a matrix $\mathbf{Q}$, and outputs the top singular vector of $\mathbf{Q}^\top\mathbf{A}^2\mathbf{Q}$, denoted by z (see Algorithm 152 for a formal

description). It follows from Theorem 1, guarantee (1) in [MM15] that

$$\|\mathbf{A}\left(\mathbf{I}-zz^{\top}\right)\|_F^2 \leqslant (1+\epsilon)\min_{\|u\|_2=1}\|\mathbf{A}\left(\mathbf{I}-uu^{\top}\right)\|_F^2, \tag{1.7}$$

and it is easy to see that this algorithm requires $\Theta\left(1/\epsilon^{1/2}\right)$ matrix-vector products. A naïve analysis requires an $\mathcal{O}(1/\epsilon)$-degree polynomial in the matrix $\mathbf{A}$ to obtain (1.7), while [MM15] use Chebyshev polynomials to approximate the threshold function between first and second singular value, and save a quadratic factor in the degree. The guarantee in (1.7) then follows from observing that the best vector in the Krylov subspace is at least as good as the one that exists using Chebyshev polynomial approximation.

Algorithm 33 (Algorithm Sketch for Frobenius rank-1 LRA)**.**

Input: *An $n \times n$ symmetric matrix $\mathbf{A}$, accuracy parameter $0 < \varepsilon < 1$.*

1. *Run Block Krylov for $\mathcal{O}\left(1/\epsilon^{1/3}\right)$ iterations with a random starting vector g. Let z_1 be the resulting output.*
2. *Run Block Krylov for $\mathcal{O}(\log(n/\epsilon))$ iterations, but initialize with an $n \times b$ random matrix $\mathbf{G}$, where $b = \mathcal{O}\left(1/\epsilon^{1/3}\right)$. Let z_2 be the resulting output.*

Output: $z = \arg\max_{z_1,z_2}\left(\|\mathbf{A}z_1\|_2^2, \|\mathbf{A}z_2\|_2^2\right)$.

Our starting point is the observation that while we require degree $\Theta\left(1/\epsilon^{1/2}\right)$ to separate the first and second singular values, if any subsequent singular value is sufficiently separated from σ_1, a significantly smaller degree polynomial suffices. In the context of Krylov methods, this translates to the intuition that starting with a matrix $\mathbf{G}$ with b columns (block size is b) should result in fewer iterations to find some vector in the top b subspace of $\mathbf{A}$. On the other hand, if no such singular value exists, the norm of the tail must be large and we can get away with a less accurate solution. We show that we can indeed exploit this trade-off by running Block Krylov on two different scales in parallel and then combine the solution. In particular, we use Algorithm 33.

Algorithm 33 captures the extreme points of the trade-off between the size of the starting matrix and the number of iterations, such that the total number of matrix-vector products is at most $\tilde{\mathcal{O}}(1/\epsilon^{1/3})$. Further, we can compute the squared Euclidean norms of $\mathbf{A}z_1$ and $\mathbf{A}z_2$ with an additional matrix-vector product, and it remains to analyze the Frobenius cost of projecting $\mathbf{A}$ on the subspace $\mathbf{I} - zz^{\top}$, where z is the unit vector output by Algorithm 33.

Using gap-independent guarantees for Block Krylov [MM15], it follows that with $\mathcal{O}\left(1/\epsilon^{1/3}\right)$ iterations, we have

$$\|\mathbf{A}z_1\|_2^2 \geqslant \sigma_1^2(\mathbf{A}) - \epsilon^{2/3}\sigma_2^2(\mathbf{A}). \tag{1.8}$$

In contrast, using gap-dependent guarantees for Block Krylov [MM15] initialized with block size b, it follows that for any $\gamma > 0$, running $q = \log(1/\gamma) \cdot \sqrt{\sigma_1(\mathbf{A})/\left(\sigma_1(\mathbf{A}) - \sigma_b(\mathbf{A})\right)}$ iterations results in z_2 such that

$$\|\mathbf{A}z_2\|_2^2 \geqslant \sigma_1^2(\mathbf{A}) - \gamma\sigma_2^2(\mathbf{A}). \tag{1.9}$$

If $\sigma_b(\mathbf{A}) \leqslant \sigma_1(\mathbf{A})/2$, we can set $\gamma = \epsilon/n$ in Equation (1.9) to obtain a highly accurate solution. Further, regardless of the input instance, Step 3 in Algorithm 33 ensures that we get the best of both guarantees, (1.8) and (1.9). Then, observing that $\mathbf{I} - zz^\top$ is an orthogonal projection matrix (see Definition 7.3.1) and using the Pythagorean Theorem for Euclidean space we have:

$$\|\mathbf{A}\left(\mathbf{I} - zz^\top\right)\|_F^2 = \|\mathbf{A}\|_F^2 - \|\mathbf{A}zz^\top\|_F^2 = \|\mathbf{A}\|_F^2 - \|\mathbf{A}z\|_2^2, \tag{1.10}$$

where the second inequality follows from unitary invariance of the Frobenius norm and that the squared Frobenius norm of a rank-1 matrix $\mathbf{A}z$ (vector) is equal to its squared Euclidean norm. If it happens that $\sigma_2(\mathbf{A}) \leqslant \sigma_1(\mathbf{A})/2$, i.e., a constant gap exists between the first two singular values, then since guarantee (1.9) implies that $\|\mathbf{A}z\|_2^2 \geqslant \sigma_1^2(\mathbf{A}) - (\epsilon/n)\sigma_2^2(\mathbf{A})$, we can plug this into (1.10) to yield a $(1+\epsilon/n)$-approximate solution. Hence, we focus on instances where $\sigma_2(\mathbf{A}) > \sigma_1(\mathbf{A})/2$.

Consider the case where the Frobenius norm of the tail is large, i.e., $\|\mathbf{A}-\mathbf{A}_1\|_F^2 \geqslant \sigma_2^2(\mathbf{A})/\epsilon^{1/3}$, where $\mathbf{A}_1$ is the best rank-1 approximation to $\mathbf{A}$. Then we only require an $\epsilon^{2/3}$-approximate solution (plugging guarantee (1.8) into (1.10)) since

$$\|\mathbf{A}\left(\mathbf{I} - z_1z_1^\top\right)\|_F^2 \leqslant \|\mathbf{A}\|_F^2 - \sigma_1^2(\mathbf{A}) + \epsilon^{2/3}\sigma_2^2(\mathbf{A}) \leqslant \|\mathbf{A} - \mathbf{A}_1\|_F^2 + \epsilon\|\mathbf{A} - \mathbf{A}_1\|_F^2. \tag{1.11}$$

Otherwise, $\sum_{i=2}^n \sigma_i^2(\mathbf{A}) < \sigma_2^2(\mathbf{A})/\epsilon^{1/3}$, which implies that there is a constant gap between the second and b-th singular values, where $b = \mathcal{O}\left(1/\epsilon^{1/3}\right)$. To see this, observe if $\sigma_b(\mathbf{A}) > \sigma_2(\mathbf{A})/4$, then $\sum_{i=2}^n \sigma_i^2(\mathbf{A}) \geqslant \sum_{i=2}^b \sigma_i^2(\mathbf{A}) \geqslant b\sigma_2^2(\mathbf{A})/4$, which is a contradiction when $b > 10/\epsilon^{1/3}$, and thus $\sigma_b(\mathbf{A}) \leqslant \sigma_2(\mathbf{A})/4 < \sigma_1/2$. Now we can apply guarantee (1.9) with $q = \mathcal{O}(\log(n/\epsilon))$ and conclude $\|\mathbf{A}z\|_2^2 \geqslant \sigma_1^2(\mathbf{A}) - (\epsilon/n)\sigma_2^2(\mathbf{A})$, yielding a highly accurate solution yet again. Overall, this suffices to obtain a $(1+\epsilon)$-approximate solution with $\tilde{\mathcal{O}}(1/\epsilon^{1/3})$ matrix-vector queries.

Challenges in generalizing to Schatten $p \neq 2$ and rank $k > 1$. The outline above crucially relies on the norm of interest being Frobenius. In particular, we use the Pythagorean Theorem to

analyze the cost of the candidate solution in Equation (1.10); however, the Pythagorean Theorem does not hold for non-Euclidean spaces. Therefore, a priori, it is unclear how to analyze the Schatten-p norm of a candidate rank-1 approximation. A proxy for the Pythagorean Theorem that holds for Schatten-p norms is Mahler's operator inequality (see Fact 7.3.11), which is in the right direction but holds only for $p \geqslant 2$, whereas we would like to handle all $p \geqslant 1$. Separately, for $p > 2$, the case where the tail is small corresponds to $\|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p \leqslant \sigma_2^p(\mathbf{A}) / \epsilon^{1/3}$. Therefore, naïvely extending the above argument requires picking a block size that scales proportional to $\mathcal{O}\left(2^p/\epsilon^{1/3}\right)$ to induce a constant gap between σ_1 and σ_b, and the number of matrix-vector products scales exponentially in p.

Finally, in the above outline, we also crucially use that $\|\mathbf{A}zz^\top\|_F^2 = \|\mathbf{A}z\|_2^2$. Observe that this no longer holds if we replace z with a matrix $\mathbf{Z}$ that has k orthonormal columns. Therefore, it remains unclear how to relate $\|\mathbf{AZ}\|_{\mathcal{S}_p}^p$ to $\|\mathbf{AZ}_{*,i}\|_2^2$, yet the vector-by-vector error guarantee obtained by Block Krylov only bounds the latter.

Handling all Schatten-p Norms and $k > 1$. We modify our algorithm to run Block Krylov on $\mathbf{A}^\top$ and obtain an orthonormal matrix $\mathbf{W}$ such that for all $i \in [k]$,

$$\|\mathbf{A}^\top \mathbf{W}_{*,i}\|^2 \geqslant \sigma_i^2(\mathbf{A}) - \gamma \sigma_{k+1}^2(\mathbf{A}), \tag{1.12}$$

for some $\gamma > 0$. We then analyze the cost $\|\mathbf{A}\left(\mathbf{I} - \mathbf{ZZ}^\top\right)\|_{\mathcal{S}_p}^p$, where $\mathbf{Z}$ is a basis for $\mathbf{A}^\top\mathbf{W}$. Our key insight is to interpret the input matrix $\mathbf{A}$ as a partitioned operator (block matrix) and invoke *pinching inequalities* for such operators. Pinching inequalities were originally introduced to understand unitarily invariant norms over direct sums of Hilbert spaces [VN37, Sch60]. In our setting, given a block matrix $\mathbf{M} = \begin{pmatrix} \mathbf{M}^{(1)} & \mathbf{M}^{(2)} \\ \mathbf{M}^{(3)} & \mathbf{M}^{(4)} \end{pmatrix}$, the *pinching inequality* (see Fact 7.3.13) implies that for all $p \geqslant 1$,

$$\|\mathbf{M}\|_{\mathcal{S}_p}^p \geqslant \|\mathbf{M}^{(1)}\|_{\mathcal{S}_p}^p + \|\mathbf{M}^{(4)}\|_{\mathcal{S}_p}^p. \tag{1.13}$$

A priori, it is unclear how to use Equation (1.13) to bound $\|\mathbf{A}\left(\mathbf{I} - \mathbf{ZZ}^\top\right)\|_{\mathcal{S}_p}^p$. First, we establish a general inequality for the Schatten norm of a matrix times an orthogonal projection. Let $\mathbf{P}$ and $\mathbf{Q}$ be any $n \times n$ orthogonal projection matrices with rank k (see Definition 7.3.1). Then, we prove that for any matrix $\mathbf{A}$,

$$\|\mathbf{A}\|_{\mathcal{S}_p}^p \geqslant \|\mathbf{PAQ}\|_{\mathcal{S}_p}^p + \|(\mathbf{I} - \mathbf{P})\,\mathbf{A}\,(\mathbf{I} - \mathbf{Q})\|_{\mathcal{S}_p}^p. \tag{1.14}$$

To obtain this inequality, we use a rotation argument along with the fact that the Schatten-p norms are unitarily invariant to show that $\|\mathbf{A}\|_{\mathcal{S}_p}^p = \|\begin{pmatrix} \mathbf{A}^{(1)} & \mathbf{A}^{(2)} \\ \mathbf{A}^{(3)} & \mathbf{A}^{(4)} \end{pmatrix}\|_{\mathcal{S}_p}^p$, where $\|\mathbf{A}^{(1)}\|_{\mathcal{S}_p} = \|\mathbf{PAQ}\|_{\mathcal{S}_p}$ and $\|\mathbf{A}^{(4)}\|_{\mathcal{S}_p} = \|(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q})\|_{\mathcal{S}_p}$, and then we can apply Equation (1.13) to the block matrix above.

Once we have established Equation (1.14), we can set $\mathbf{P} = \mathbf{W}\mathbf{W}^\top$ and set $\mathbf{Q} = \mathbf{Z}\mathbf{Z}^\top$ to be the projection matrix corresponding to the column span of $\mathbf{A}^\top\mathbf{W}\mathbf{W}^\top$. Then, we have that $\mathbf{PAQ} = \mathbf{W}\mathbf{W}^\top\mathbf{A}$ and $(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q}) = \mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)$, and combined with (1.14) this yields

$$\|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p \leqslant \|\mathbf{A}\|_{\mathcal{S}_p}^p - \|\mathbf{W}\mathbf{W}^\top\mathbf{A}\|_{\mathcal{S}_p}^p. \tag{1.15}$$

To obtain a bound on $\|\mathbf{W}\mathbf{W}^\top\mathbf{A}\|_{\mathcal{S}_p}^p$, we appeal to the per-vector guarantees in Equation (1.12). However, translating from ℓ_2^2 error to $\sigma_p^p\left(\mathbf{W}^\top\mathbf{A}\right)$ incurs a mixed guarantee:

$$\|\mathbf{W}\mathbf{W}^\top\mathbf{A}\|_{\mathcal{S}_p}^p \geqslant \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - \mathcal{O}(\gamma p)\sum_{i\in[k]}\sigma_{k+1}^2(\mathbf{A})\,\sigma_i^{p-2}(\mathbf{A}).$$

To use this bound, we require $\sigma_1(\mathbf{A})$ to be comparable to $\sigma_{k+1}(\mathbf{A})$ and thus we require an involved case analysis, which appears in the proof of Theorem 28.

Avoiding an exponential dependence on p. Our main insight here is that we do not require a block size that induces a constant gap between singular values. Instead, we first observe that if the block size b is large enough such that $\sigma_b \leqslant \sigma_2/(1+1/p)$, then $\mathcal{O}\Big(\log(n/\epsilon)\sqrt{p}\Big)$ iterations suffice to obtain a vector z such that $\|\mathbf{A}z\|_2^2 \geqslant \sigma_1^2(\mathbf{A}) - (\epsilon/n)\,\sigma_2^2(\mathbf{A})$. Therefore, we can trade-off the threshold for the Schatten norm of the tail with the number of iterations as follows: if $\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p \leqslant \frac{1}{p^{1/3}\epsilon^{1/3}}\sigma_2^p(\mathbf{A})$, then setting $b = (1+1/p)^p/(\epsilon p)^{1/3} = \Theta(1/(\epsilon p)^{1/3})$ suffices to induce a gap of $1+1/p$ with block size b. The total number of matrix-vector products is $\mathcal{O}\Big(b\cdot\log(n/\epsilon)\sqrt{p}\Big) = \tilde{\mathcal{O}}(p^{1/6}/\epsilon^{1/3})$, since p can be assumed to be at most $(\log n)/\epsilon$. Otherwise, $\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p > \frac{1}{p^{1/3}\epsilon^{1/3}}\sigma_2^p(\mathbf{A})$, and we only require a $(1+\epsilon^{2/3}/p^{1/3})$-approximate solution instead (compare with Equation (1.11)). Using gap-independent bounds (see Lemma 7.4.1), it suffices to start with block size 1 and run $\mathcal{O}\Big(\log(n/\epsilon)p^{1/6}/\epsilon^{1/3}\Big)$ iterations to obtain a $(1+\epsilon^{2/3}/p^{1/3})$-approximate solution.

Avoiding a Gap-Dependent Bound. We note that even when there is a constant gap between the first and second singular values, and the per vector guarantee is highly accurate, i.e., for all $i\in[k]$, $\|\mathbf{A}\mathbf{Z}_{*,i}\|^2 \geqslant \sigma_i^2(\mathbf{A}) - \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2(\mathbf{A})$, it is not clear how to lower bound $\|\mathbf{A}\mathbf{Z}\|_{\mathcal{S}_p}^p$ in

Equation 1.15. In general, the best bound we can obtain using the above equation is

$$\|\mathbf{AZ}\|_{\mathcal{S}_p}^p \geqslant \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - \mathcal{O}\left(\frac{\epsilon}{\text{poly}(d)}\right)\sigma_{k+1}^2 \cdot \sum_{i\in[k]} \sigma_i^{p-2}, \tag{1.16}$$

which may be vacuous when the top k singular values are significantly larger than σ_{k+1} and $p > 2$. One could revert to a gap-dependent bound, where the error is in terms of the gap between σ_1 and σ_{k+1}, which one could account for by running an extra factor of $\mathcal{O}(\log(\sigma_1/\sigma_{k+1}))$ iterations.

To avoid this gap-dependent bound, we split $\mathbf{A}$ into a head part $\mathbf{A}_H$ and a tail part $\mathbf{A}_T$, such that $\mathbf{A}_H$ has all singular values that are at least $(1+1/d)\,\sigma_{k+1}$ and $\mathbf{A}_T$ has the remaining singular values. We then bound $\|\mathbf{A}_H\left(\mathbf{I}-\mathbf{ZZ}^\top\right)\|_{\mathcal{S}_p}$ and $\|\mathbf{A}_T\left(\mathbf{I}-\mathbf{ZZ}^\top\right)\|_{\mathcal{S}_p}$ separately. Repeating the above analysis, we can obtain Equation (1.16) for $\mathbf{A}_T$ instead, and since all singular values larger than σ_{k+1} in $\mathbf{A}_T$ are bounded, we can obtain $\|\mathbf{A}_T\left(\mathbf{I}-\mathbf{ZZ}^\top\right)\|_{\mathcal{S}_p}^p \leqslant \mathcal{O}(\epsilon k/\text{poly}(d))\,\sigma_{k+1}^p$. To adapt the analysis for $\mathbf{A}_T$ and obtain this bound, we use Cauchy's interlacing theorem to relate the j-th singular value of $\mathbf{A}_T\left(\mathbf{I}-\mathbf{ZZ}^\top\right)$ to the (i^*+j)-th singular value of $\mathbf{A}\left(\mathbf{I}-\mathbf{ZZ}^\top\right)$, where i^* is the rank of $\mathbf{A}_H$. We lower bound the (i^*+j)-th singular value of $\mathbf{A}\left(\mathbf{I}-\mathbf{ZZ}^\top\right)$ using the per vector guarantee of [MM15].

To bound $\|\mathbf{A}_H\left(\mathbf{I}-\mathbf{ZZ}^\top\right)\|_{\mathcal{S}_p}$, we observe it has rank at most k and thus

$$\|\mathbf{A}_H\left(\mathbf{I}-\mathbf{ZZ}^\top\right)\|_{\mathcal{S}_p} \leqslant \sqrt{k}\cdot\|\mathbf{A}_H\left(\mathbf{I}-\mathbf{ZZ}^\top\right)\|_F = \sqrt{k}\cdot\sqrt{\|\mathbf{A}_H\|_F^2 - \|\mathbf{A}_H\mathbf{Z}\|_F^2},$$

Intuitively, while the k-dimensional subspace that we find can "swap out" singular vectors corresponding to singular values σ_i for which σ_i is very close to σ_{k+1}, since they serve equally well for a Schatten-p low-rank approximation, for singular values σ_i that are a bit larger than σ_{k+1}, the k-dimensional subspace we find cannot do this. More precisely, if y is a singular vector of $\mathbf{A}_H$ with singular value σ_i, then the projection of y onto the k-dimensional subspace that our algorithm finds (namely, $\mathbf{Z}$) must be at least $1-\sigma_{k+1}^2/((\sigma_i^2-\sigma_{k+1}^2)\text{poly}(d))$, which suffices to bound the above since the additive error is inversely proportional to σ_i^2 when $\sigma_i^2 \gg \sigma_{k+1}^2$, and so the very tiny additive error negates the effect of very large singular values.

Future Directions. In terms of concrete open questions, we note that our lower bounds are tight only when the target rank k and Schatten norm p are fixed constants. In particular, it is open to obtain matrix-vector lower bounds that grow as a function of k, p and $1/\epsilon$.

Open Question 34. What is the optimal matrix-vector complexity of low-rank approximation as

a function of k, p and ϵ?

For the important special case of Spectral low-rank approximation ($p = \infty$), it is open to obtain any lower bound that grows as a function of $1/\epsilon$, even when the target rank $k = 1$. We also note that improving our upper bound to even $p^{1/6-o(1)}$ would imply a faster algorithm for Spectral low-rank approximation, addressing the main open question in [Woo14b].

In addition, more open ended questions include determining the matrix-vector product complexity of several fundamental problems in numerical linear algebra such as regression, PSD testing and estimating schatten norms, and finding structured optimization problems where we can beat the square-root speedup obtained by Chebyshev polynomials.

1.2.2 Low-Rank Approximation for PSD Matrices

As mentioned above, a large body of work over the past two decades has studied *relative-error* low-rank approximation, whereby given an $n \times n$ matrix $\mathbf{A}$, an accuracy parameter $\epsilon > 0$, and a rank parameter k, one seeks to output a rank-k matrix $\mathbf{B}$ for which

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2, \tag{1.17}$$

where for a matrix $\mathbf{C}$, $\|\mathbf{C}\|_F^2 = \sum_{i,j} \mathbf{C}_{i,j}^2$, and $\mathbf{A}_k$ denotes the best rank-k approximation to $\mathbf{A}$ in Frobenius norm. $\mathbf{A}_k$ can be computed exactly using the singular value decomposition, but takes time $O(n^\omega)$, where ω is the matrix multiplication constant. We refer the reader to the survey [Woo14a] and references therein.

For worst-case matrices, it is not hard to see that any algorithm achieving (8.1) must spend at least $\Omega(\text{nnz}(\mathbf{A}))$ time, where $\text{nnz}(\mathbf{A})$ denotes the number of non-zero entries (sparsity) of $\mathbf{A}$. Indeed, without reading most of the non-zero entries of $\mathbf{A}$, one could fail to read a single large entry, thus making one's output matrix $\mathbf{B}$ an arbitrarily bad approximation.

A flurry of recent work [KP16, MW17c, CLW18, Tan19, RSML18, GLT18, IVWW19, SW19, GSLW19] has looked at the possibility of achieving *sublinear* time algorithms (classical and quantum) for low-rank approximation. In particular, Musco and Woodruff [MW17c] consider the important case of positive-semidefinite (PSD) matrices. PSD matrices include as special cases covariance matrices, correlation matrices, graph Laplacians, kernel matrices and random dot product models. Further, the special case where the input itself is low-rank (PSD Matrix Completion) has applications in quantum state tomography [GLF+10]. Subsequently, Bakshi and

Woodruff [BW18] considered low-rank approximation of the closely related family of Negative-type (Euclidean Squared) distance matrices. Negative-type metrics include as special cases ℓ_1 and ℓ_2 metrics, spherical metrics and hypermetrics, as well as effective resistances in graphs [DL09, TD87, CRR+96, CKM+11]. Negative-type metrics have found various applications in algorithm design and optimization [ALN08, SS11, KMP14].

Musco and Woodruff show that it is possible to output a low-rank matrix $\mathbf{B}$ in factored form achieving (8.1) in $\widetilde{O}(nk/\epsilon^{2.5} + nk^{\omega-1}/\epsilon^{2(\omega-1)})$ time, while reading only $\widetilde{O}(nk/\epsilon^{2.5})$ entries of $\mathbf{A}$. They also showed a lower bound that any algorithm achieving (8.1) must read $\Omega(nk/\epsilon)$ entries, and closing the gap between these bounds has remained an open question. Similarly, in joint work with David Woodruff, we exploit the structure of Negative-type metrics to reduce to the PSD case and obtain a bi-criteria algorithm that requires $\widetilde{O}(nk/\epsilon^{2.5})$ queries. The gap in the sample complexity and the requirement of a bi-criteria guarantee remained open. We resolve these both these questions here.

Next we consider PSD matrices that have been corrupted by a small amount of noise. A drawback of algorithms achieving (8.1) is that they cannot tolerate any amount of unstructured noise. For instance, if one slightly corrupts a few off-diagonal entries, making the input matrix $\mathbf{A}$ no longer PSD, then it is impossible to detect such corruptions in sublinear time, making the relative-error guarantee (8.1) information-theoretically impossible. Motivated by this, we also introduce a new framework where an adversary corrupts the input by adding a noise matrix $\mathbf{N}$ to a psd matrix $\mathbf{A}$. We assume that the Frobenius norm of the corruption is bounded relative to the Frobenius norm of $\mathbf{A}$, i.e., $\|\mathbf{N}\|_F^2 \leqslant \eta\|\mathbf{A}\|_F^2$. We also assume the corruption is well-spread, i.e., each row of $\mathbf{N}$ has ℓ_2^2-norm at most a fixed constant factor larger than ℓ_2^2-norm of the corresponding row of $\mathbf{A}$.

This model captures small perturbations to PSD matrices that we may observe in real-world datasets, as a consequence of round-off or numerical errors in tasks such as computing Laplacian pseudoinverses, and systematic measurement errors when computing a covariance matrix. One important application captured by our model is low-rank approximation of corrupted *correlation matrices*. Finding a low-rank approximation of such matrices occurs when measured correlations are asynchronous or incomplete, or when models are stress-tested by adjusting individual correlations. Low-rank approximation of correlation matrices also has many applications in finance [Hig02].

Given that it is information-theoretically impossible to obtain the relative-error guarantee (8.1) in the *robust model*, we relax our notion of approximation to the following well-studied

Problem	Prior Work		Our Results		Query Lower Bound
	Query	Run Time	Query	Run Time	
PSD LRA	$O\left(\frac{nk}{\epsilon^{2.5}}\right)$	$O\left(\frac{nk^{\omega-1}}{\epsilon^{2\omega-2}}+\frac{nk}{\epsilon^{2.5}}\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	[MW17c]		Thm. 35		[MW17c]
PSD LRA PSD Output	$O\left(\frac{nk^2}{\epsilon^2}\right)$	$O\left(nk^{\omega-1}\left(\frac{k}{\epsilon^\omega}+\frac{1}{\epsilon^{3\omega-3}}\right)\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	[MW17c]		Thm. 35		[MW17c]
Negative-Type LRA	$O\left(\frac{nk}{\epsilon^{2.5}}\right)$	$O\left(\frac{nk^{\omega-1}}{\epsilon^{2\omega-2}}+\frac{nk}{\epsilon^{2.5}}\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	Bi-criteria, [BW18]		No Bi-criteria, Thm. 38		[BW18]
Coreset Ridge Regression	$O\left(\frac{ns_\lambda^2}{\epsilon^4}\right)$	$O\left(\frac{ns_\lambda^\omega}{\epsilon^\omega}\right)$	$O^*\left(\frac{ns_\lambda}{\epsilon^2}\right)$	$O^\dagger\left(\frac{ns_\lambda^{\omega-1}}{\epsilon^{2\omega-2}}\right)$	$\Omega\left(\frac{ns_\lambda}{\epsilon^2}\right)$
	[MW17c]		Thm. 40		

Table 1.2: Comparison with prior work. The notation O^* and $O^\dagger$ represent existence of matching lower bounds for query complexity and running time (assuming the fast matrix multiplication exponent ω is 2) respectively. The notation s_λ is used to denote the statistical dimension of ridge regression. All bounds are stated ignoring polylogarithmic factors in n, k and ϵ.

additive-error guarantee:

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant \|\mathbf{A}-\mathbf{A}_k\|_F^2+(\epsilon+\eta)\|\mathbf{A}\|_F^2. \tag{1.18}$$

This additive-error guarantee was introduced by the seminal work of Frieze et. al. [FKV04b], and triggered a long line of work on low-rank approximation from a computational perspective. Frieze et al. showed that it is possible to achieve (8.2) in $O(\text{nnz}(\mathbf{A}))$ time. Further, given access to an oracle for computing row norms of $\mathbf{A}$, 8.2 is achievable in sublinear time. More recently, the same notion of approximation was used to obtain sublinear sample complexity and running time algorithms for *distance matrices* [BW18],[IVWW19], and a quantum algorithm for recommendation systems [KP16], which was subsequently dequantized [Tan19].

This raises the question of how robust are our sublinear low-rank approximation algorithms for structured matrices, if we relax to additive-error guarantees and allow for corruption. In particular, can we obtain additive-error low-rank approximation algorithms for PSD matrices that achieve sublinear time and sample complexity in the presence of noise? We characterize when such robust algorithms are achievable in sublinear time.

Our Results. We begin with stating our results for low-rank approximation for structured matrices. Our main result is an optimal algorithm for low-rank approximation of PSD matrices:

Theorem 35 (Sample-Optimal PSD LRA)**.** *Given a PSD matrix* $\mathbf{A}$*, there exists an algorithm that queries* $\widetilde{O}(nk/\epsilon)$ *entries in* $\mathbf{A}$ *and outputs a rank* k *matrix* $\mathbf{B}$ *such that with probability* $99/100$*,*

$\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$, *and the algorithm runs in time* $\widetilde{O}(n\cdot(k/\epsilon)^{\omega-1})$.

Remark 36. Our algorithm matches the sample complexity lower bound of Musco and Woodruff, up to logarithmic factors, which shows that any randomized algorithm that outputs a $(1+\epsilon)$-relative-error low-rank approximation for a PSD matrix $\mathbf{A}$ must read $\Omega(nk/\epsilon)$ entries. Our running time also improves that of Musco and Woodruff and is optimal if the matrix multiplication exponent ω is 2.

Remark 37. We can extend our algorithm such that the low-rank matrix $\mathbf{B}$ we output is also PSD with the same query complexity and running time. In comparison, the algorithm of Musco and Woodruff accesses $\widetilde{O}(nk/\epsilon^3+nk^2/\epsilon^2)$ entries in $\mathbf{A}$ and runs in time $\widetilde{O}(n(k/\epsilon)^{\omega}+nk^{\omega-1}/\epsilon^{3(\omega-1)})$.

At the core of our analysis is a sample optimal algorithm for Spectral Regression: $\min_{\mathbf{X}}\|\mathbf{DX}-\mathbf{E}\|_2^2$. We show that when $\mathbf{D}$ has orthonormal columns and $\mathbf{E}$ is arbitrary, we can sketch the problem by sampling rows proportional to the leverage scores of $\mathbf{D}$ and approximately preserve the minimum cost. This is particularly surprising since our sketch only computes sampling probabilities by reading entries in $\mathbf{D}$, while being completely agnostic to the entries in $\mathbf{E}$. Here, we also prove a spectral approximate matrix product guarantee for our one-sided leverage score sketch, which may be of independent interest. We note that such a guarantee for leverage score sampling does not appear in prior work, and we discuss the technical challenges we need to overcome in the subsequent section.

The techniques we develop for PSD low-rank approximation also extend to computing a low-rank approximation for distance matrices that arise from negative-type (Euclidean-squared) metrics. Here, our input is a pair-wise distance matrix $\mathbf{A}$ corresponding to a point set $\mathcal{P}=\{x_1,x_2,\ldots,x_n\}\in\mathbb{R}^d$ such that $\mathbf{A}_{i,j}=\|x_i-x_j\|_2^2$. We obtain an optimal algorithm for computing a low-rank approximation of such matrices:

Theorem 38 (Sample-Optimal LRA for Negative-Type Metrics)**.** *Given a negative-type distance matrix* $\mathbf{A}$*, there exists an algorithm that queries* $\widetilde{O}(nk/\epsilon)$ *entries in* $\mathbf{A}$ *and outputs a rank* k *matrix* $\mathbf{B}$ *such that with probability* $99/100$*,* $\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$*, and the algorithm runs in time* $\widetilde{O}(n\cdot(k/\epsilon)^{\omega-1})$.

Remark 39. In prior work with David Woodruff [BW18], we obtained a $\widetilde{O}(nk/\epsilon^{2.5})$ query algorithm that outputs a rank-$(k+4)$ matrix $\mathbf{B}$ such that $\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$. We show that the bi-criteria guarantee is not necessary, thereby resolving an open question in their paper.

Structured Regression. The sample-optimal algorithm for PSD Low-Rank Approximation

also leads to a faster algorithm for Ridge Regression, when the design matrix is PSD. Given a PSD matrix $\mathbf{A}$, a vector y and a regularization parameter λ, we consider the following optimization problem: $\min_{x\in\mathbb{R}^n} \|\mathbf{A}x - y\|_2^2 + \lambda\|x\|_2^2$. This problem is often referred to as Ridge Regression and has been the focus of numerous theoretical and practical works.

Theorem 40 (PSD Ridge Regression.)**.** *Given a PSD matrix $\mathbf{A}$, a regularization parameter λ and statistical dimension $s_\lambda = Tr(\mathbf{A}^2 + \lambda\mathbf{I})^{-1}\mathbf{A}^2$, there exists an algorithm that queries $\widetilde{O}(ns_\lambda/\epsilon^2)$ entries of $\mathbf{A}$ and with probability $99/100$ outputs a $(1+\epsilon)$ approximate solution to the Ridge Regression objective and runs in $\widetilde{O}(n(s_\lambda/\epsilon^2)^{\omega-1})$ time.*

Remark 41. Our result improves on prior work by Musco and Woodruff [MW17c], who obtain an algorithm that queries $\widetilde{O}(ns_\lambda^2/\epsilon^4)$ entries in $\mathbf{A}$ and runs in $\widetilde{O}(n(s_\lambda/\epsilon^2)^\omega)$ time.

Robust Low-Rank Approximation. Next, we consider a robust form of low-rank approximation problem, where the input is a PSD matrix corrupted by noise. In this setting, we have query access to the corrupted matrix $\mathbf{A} + \mathbf{N}$, where $\mathbf{A}$ is PSD and $\mathbf{N}$ is such that $\|\mathbf{N}\|_F^2 \leqslant \eta\|\mathbf{A}\|_F^2$. Further, for all $i \in [n]$ $\|\mathbf{N}_{i,*}\|_2^2 \leqslant c\|\mathbf{A}_{i,*}\|_2^2$, for a fixed constant c. The diagonal of a PSD matrix carries crucial information since the largest diagonal entry upper bounds all off-diagonal entries. Therefore, a reasonable adversarial strategy is to corrupt the largest diagonal entries and make them close to the small diagonal entries, which enables the resulting matrix to have large off-diagonal entries that are hard to find. Capturing this intuition we parameterize our algorithms and lower bounds by the largest ratio between a diagonal entry of $\mathbf{A}$ and $\mathbf{A} + \mathbf{N}$, denoted by $\phi_{\max} = \max_{j\in[n]} \mathbf{A}_{j,j}/|(\mathbf{A} + \mathbf{N})_{j,j}|$.

Theorem 42 (Robust LRA Lower Bound)**.** *Let $\epsilon > \eta > 0$. Given $\mathbf{A} + \mathbf{N}$ such that $\mathbf{A}$ is PSD and $\mathbf{N}$ is a corruption matrix as defined above, any randomized algorithm that with probability at least $2/3$ outputs a rank-k approximation up to additive error $(\epsilon + \eta)\|\mathbf{A}\|_F^2$ must read $\Omega\left(\phi_{\max}^2 nk/\epsilon\right)$ entries of $\mathbf{A} + \mathbf{N}$.*

Remark 43. Any algorithm must incur additive error $\eta\|\mathbf{A}\|_F^2$, since $\mathbf{A}$ is not even identifiable below additive-error $\eta\|\mathbf{A}\|_F^2$.

Remark 44. In our hard instance, $\phi_{\max}^2$ can be as large as $\epsilon n/k$, which implies a sample-complexity lower bound of $\Omega(n^2)$. While this lower bound precludes sublinear algorithms for arbitrary PSD matrices, we observe that in many applications $\phi_{\max}$ can be significantly smaller. For instance, if $\mathbf{A}$ is a correlation matrix, we know that the true diagonal entries of $\mathbf{A} + \mathbf{N}$ are 1 and can ignore any corruption on them to bound $\phi_{\max}$ by 1.

Motivated by the aforementioned observation, we introduce algorithms for robust low-rank approximation, parameterized by the corruption on the diagonal entries. We obtain the following theorem:

Theorem 45 (Robust Low-Rank Approximation)**.** *Given* $\mathbf{A}+\mathbf{N}$*, which satisfies our noise model, there exists an algorithm that queries* $\widetilde{O}(\phi_{\max}^2 nk/\epsilon)$ *entries in* $\mathbf{A}+\mathbf{N}$ *and computes a rank* k *matrix* $\mathbf{B}$ *such that with probability at least* $99/100$*,* $\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant \|\mathbf{A}-\mathbf{A}_k\|_F^2 + (\epsilon+\sqrt{\eta})\|\mathbf{A}\|_F^2$.

Remark 46. While the sample complexity of this algorithm matches the sample complexity in the lower bound, it incurs additive-error $\sqrt{\eta}\|\mathbf{A}\|_F^2$ as opposed to $\eta\|\mathbf{A}\|_F^2$. An interesting open question here is whether we can achieve additive-error $o(\sqrt{\eta}\|\mathbf{A}\|_F^2)$, though we note that when $\eta^2 \leqslant \epsilon$, this just changes the additive error guarantee of our low-rank approximation by a constant factor.

Remark 47. Our techniques extend to low-rank approximation of correlation matrices, and we obtain a sample complexity of $\widetilde{O}(nk/\epsilon)$, which is optimal. In fact, the hard instance in [MW17c] implies an $\Omega(nk/\epsilon)$ lower bound on the sample complexity, even in the presence of no noise. Surprisingly, corrupting a correlation matrix does not increase the sample complexity and only incurs an additive error of $\sqrt{\eta}\|\mathbf{A}\|_F^2$.

Future Directions. A nascent area in algorithm design is developing fast algorithms for structured linear algebra problems. This area has seen rapid progress for problems including low-rank approximation (see above), regression and covariance estimation. Considering structured matrices can also be an avenue for progress on major open problems like spectral low-rank approximation. An open ended research direction is as follows:

Open Question 48. When does structure in the input lead to faster algorithms for fundamental problems in numerical linear algebra? How robust are the corresponding algorithms to perturbations of the structure in the input?

As mentioned above, exploiting structure of the input matrices has lead to several algorithmic breakthroughts: solving linear systems for Laplacian/Diagonally Dominant matrices [ST14, KOSZ13, KMP14] and Block Henkel matrices [PV21], covariance estimation of Toeplitz matrices [ELMM20], and approximation the permanent of boolean [JS89], non-negative matrices [JSV04] and PSD [AGGS17, YP21] matrices. Obtaining provable guarantees for the aforementioned tasks, even when the input matrix is perturbed by noise, is an intruiging research direction.

More broadly, the tools we developed in these works have been useful for a myriad of machine learning applications, including provable guarantees for training two layer ReLU networks [BJW19], distriuted clustering [ABB+19] , learning a latent simplex in input sparsity time [BBK+21a], and quantum-inspired algorithms for machine learning [CCH+20]. Looking forward, we hope to understand the power and applicability of these tools to learning other latent models as well as quantum-inspired algorithms.

1.2.3 Learning a Latent Simplex

We also study the problem of learning k vertices $\mathbf{M}_{*,1}, \ldots, \mathbf{M}_{*,k}$ of a latent k-dimensional simplex $\mathcal{K}$ in $\mathbb{R}^d$ using n data points generated from $\mathcal{K}$ and then possibly perturbed by a stochastic, deterministic, or adversarial source before given to the algorithm. In particular, the resulting points observed as input data could be heavily perturbed so that the initial points may no longer be discernible or they could be outside the simplex $\mathcal{K}$. Recent work of Bhattacharyya and Kannan [BK20c] unifies several stochastic models for unsupervised learning problems, including k-means clustering [CG92, GH+96, Web03, WT10, Dua20], topic models [BJ03, SG07, BL06a, Ble12, AGH+13a], mixed membership stochastic block models [ABFX08, MJG09, XFS+10, FSX09, ABEF14, LAW16, FXC16] and Non-negative Matrix Factorization [AGH+13b, GV14, Gil20] under the problem of learning a latent simplex. In general, identifying the latent simplex can be computationally intractable. However many special applications do not require the full generality. For example, in a mixture model like Gaussian mixtures, the data is assumed to be generated from a convex combination of density functions. Thus, it may be possible to efficiently approximately learn the latent simplex given certain distributional properties in these models.

Indeed, Bhattacharyya and Kannan showed that given certain reasonable geometric assumptions that are typically satisfied for real-world instances of Latent Dirichlet Allocation, Stochastic Block Models and Clustering, there exists an $\widetilde{O}(k \cdot \mathrm{nnz}(\mathbf{A}))$ [9] time algorithm for recovering the vertices of the underlying simplex. We show that, given an additional natural assumption, we can remove the dependency on k and obtain a true input sparsity time algorithm. We begin by defining the model along with our new assumption:

Definition 1.2.1 (Latent Simplex Model)**.** *Let $\mathbf{M}$ be a $d \times k$ matrix such that $\mathbf{M}_{*,1}, \mathbf{M}_{*,2}, \ldots, \mathbf{M}_{*,k} \in \mathbb{R}^d$ denote the vertices of a k-simplex, $\mathcal{K}$. Let $\mathbf{P}$ be a $d \times n$ matrix such that $\mathbf{P}_{*,1}, \mathbf{P}_{*,2} \ldots \mathbf{P}_{*,n} \in \mathbb{R}^d$ are n points in the convex hull of $\mathcal{K}$. Given $\sigma > 0$, we observe a $d \times n$ matrix $\mathbf{A}$, such that*

[9] Throughout the paper we use the notation $\widetilde{O}$ to suppress poly-logarithmic factors.

$\|\mathbf{A}-\mathbf{P}\|_2 \leqslant \sigma\sqrt{n}$. *Further, we make the following assumptions on the data generation process:*

1. ***Well-Separateness.*** *For all $\ell \in [k]$, $\mathbf{M}_{*,\ell}$ has non-trivial mass in the orthogonal complement of the span of the remaining vectors, i.e., for all $\ell \in [k]$, $|Proj(\mathbf{M}_{*,\ell}, Null(\mathbf{M} \setminus \mathbf{M}_{*,\ell}))| \geqslant \alpha \max_\ell \|\mathbf{M}_{*,\ell}\|_2$ where $Proj(x, U)$ denotes the orthogonal projection of x to the subspace U and $\mathbf{M} \setminus \mathbf{M}_{*,\ell}$ is the matrix $\mathbf{M}$ with the ℓ-th column removed.*

2. ***Proximate Latent Points.*** *Given $\delta \in (0,1)$, for all $\ell \in [k]$, there exists a set $\mathcal{S}_\ell \subseteq [n]$ such that $|\mathcal{S}_\ell| \geqslant \delta n$ and for all $j \in \mathcal{S}_\ell$, $\|\mathbf{M}_{*,\ell} - \mathbf{P}_{*,j}\|_2 \leqslant 4\sigma/\delta$.*

3. ***Spectrally Bounded Perturbation.*** *The spectrum of $\mathbf{A}-\mathbf{P}$ is bounded, i.e., for a sufficiently large constant c, $\sigma/\sqrt{\delta} \leqslant \alpha^2 \min_\ell \|\mathbf{M}_{*,\ell}\|_2/ck^9$.*

4. ***Significant Singular Values.*** *Let $\mathbf{A} = \sum_{i\in[d]} \sigma_i u_i v_i^T$ be the singular value decomposition and let $0 < \phi \leqslant nnz(\mathbf{A})/(n \cdot poly(k))$. We assume that for all $i \in [k]$, $\sigma_i > \phi \cdot \sigma_{k+1}$ and $\|\mathbf{A} - \mathbf{A}_k\|_F^2 \leqslant \phi\|\mathbf{A} - \mathbf{A}_k\|_2^2$.*

These assumptions are natural across many interesting applications. [BK20c] introduced the Well-Separateness (1), Proximate Latent Points (2) and Spectrally Bounded Perturbation (3) assumptions. We include an additional Significant Singular Values assumption (4), which is crucial for obtaining a faster running time; we discuss this in more detail below. Our main algorithmic result can then be stated as follows:

Theorem 49 (Learning a Latent Simplex in Input-Sparsity Time)**.** *Given $k \geqslant 2$ and $\mathbf{A} \in \mathbb{R}^{d\times n}$ from the Latent Simplex Model (Definition 1.2.1), there exists an algorithm that runs in $\widetilde{O}\left(nnz(\mathbf{A}) + (n+d)poly(k/\phi)\right)$ time to output subsets $\mathbf{A}_{\mathcal{R}_1}, \ldots, \mathbf{A}_{\mathcal{R}_k}$ such that upon permuting the columns of $\mathbf{M}$, with probability at least $1 - 1/\Omega(\sqrt{k})$, for all $\ell \in [k]$, we have $\|\mathbf{A}_{\mathcal{R}_\ell} - \mathbf{M}_{*,\ell}\|_2 \leqslant 300k^4\sigma/(\alpha\sqrt{\delta})$.*

Our result implies faster algorithms for various stochastic models that can be formulated as special cases of the Latent Simplex Model, including Latent Dirichlet Allocation for Topic Modeling, Mixed Membership Stochastic Block Models and Adversarial Clustering. We summarize the connections to these applications below. We describe our algorithm and provide an outline to our analysis; we defer all formal proofs to the supplementary material.

We first formalize the connection between the Latent Simplex Model (Definition 1.2.1) and numerous stochastic models. In particular, we show that topic models like Latent Dirichlet Allocation (LDA), Stochastic Block Models and Adversarial Clustering can be viewed as special cases of the Latent Simplex Model. We also show how our assumptions are natural in each of

these applications.

Topic Models. Probabilistic Topic Models attempt to identify abstract topics in a collection of documents by discovering latent semantic structure [BJ03, BL06b, HBB10, ZAX12, Ble12]. Each document in the corpus is represented by a bag-of-words vectorization with the corresponding word frequencies. The standard statistical assumption is that the generative process for the corpus is a joint probability distribution over both the observed and hidden random variables. The hidden random variables can be interpreted as representative documents for each topic. The goal is to then design algorithms that can learn the underlying topics. The topics can be viewed geometrically as k latent vectors $\mathbf{M}_{*,1}, \mathbf{M}_{*,2}, \ldots, \mathbf{M}_{*,k} \in \mathbb{R}^d$, where d is the size of the dictionary and $\mathbf{M}_{i,\ell}$ is the expected frequency of word i in topic ℓ. Since each vector $\mathbf{M}_{*,\ell}$ represents a probability distribution, $\sum_i \mathbf{M}_{i,\ell} = 1$. Let $\mathbf{M}$ be the corresponding $d \times k$ matrix. One important stochastic model is Latent Dirichlet Allocation (LDA) [BNJ03], where each document consists of m words is generated as follows :

- For all $\ell \in [k]$, we pick topic weights $\mathbf{W}_{j,\ell} \sim \mathsf{Dir}(1/k)$, where $\mathsf{Dir}(1/k)$ is the Dirichlet distribution over the unit simplex. The topic distribution of document j is decided by the topic weights, $\mathbf{W}_{j,\ell}$, and given by $\mathbf{P}_{*,j} = \sum_{\ell \in [k]} \mathbf{W}_{j,\ell} \cdot \mathbf{M}_{*,\ell}$, where $\mathbf{P}_{*,j}$ are latent points.
- We then generate the j-th document with m words by taking i.i.d. samples from $\mathsf{Mult}(\mathbf{P}_{*,j})$, the multinomial distribution with $\mathbf{P}_{*,j}$ as the probability vector. The resulting document observed is denoted by the vector $\mathbf{A}_{*,j}$, where for all $i \in [d]$ $\mathbf{A}_{i,j} = \frac{1}{m} \sum_{t=1}^{m} \mathbf{X}_{ij}^{(t)}$,, such that $\mathbf{X}_{ij}^{(t)} \sim \mathsf{Bern}(\mathbf{P}_{ij})$, where $\mathbf{X}_{ij}^{(t)} = 1$ if the i-th word was chosen in the t-th draw while generating the j-th document, and 0 otherwise.

The data generation process of LDA can be viewed as a special case of the Latent Simplex Model, where the j-th document is the data point $\mathbf{A}_{*,j}$ generated from the stochastic vector $\mathbf{P}_{*,j}$, a point in the simplex $\mathcal{K}$. The vertices of the simplex are the k topic vectors $\mathbf{M}_{*,1}, \ldots, \mathbf{M}_{*,k}$; the goal is then to recover the vertices of $\mathcal{K}$. [BK20c] remark that the Well-Separateness condition holds for LDA if we assume a Dirichlet prior on $\mathbf{M}$. We note that while $\mathcal{K}$ is a k-dimensional simplex, $d \ll k$ and the observed points need not lie inside the simplex. On the contrary, [BK20c] show that the data often lies significantly outside of $\mathcal{K}$. However, they show that the smoothed simplex obtained by taking the averages of all δn sized subsets of observed points results in a polytope $K_{\mathcal{S}}$ that is close to $\mathcal{K}$.

We formally justify our assumptions below.

Lemma 1.2.2 (LDA as a Latent Simplex)**.** *Given* $\mathbf{A}, \mathbf{P}, \mathbf{M}$ *following the LDA model as described*

above, such that for all $\ell \in [k]$, $\|\mathbf{M}_{*,\ell}\|_2 = \Omega(1)$, $m, n = \Omega(poly(k/\alpha))$ *and* $\delta = c\sigma/\sqrt{k}$, *assumptions* (2),(3) *and* (4) *from Definition 1.2.1 are satisfied with high probability.*

Proof. Assumptions (2) and (3) follow from Lemma 7.1 in [BK20c]. By Claim 8.1 in [BK20c], $\sigma_k(\mathbf{A}) \geqslant c\alpha\sqrt{\delta/k}\min_\ell \mathbf{M}_{*,\ell}$. Each column of $\mathbf{A}$ sums to 1, so $\|\mathbf{A}\|_F^2 = O(n)$ and $\sigma_k(\mathbf{A}) \geqslant \alpha\sqrt{\delta/k}\|\mathbf{A}\|_F$. Since $\|\mathbf{A} - \mathbf{P}\|_2 \leqslant \sigma\sqrt{n}$ by definition of σ, and $\mathbf{P}$ consists of n point in the convex hull of k points and thus $\sigma_{k+1}(\mathbf{P}) = 0$, we have $\sigma_{k+1}(\mathbf{A}) \leqslant \sigma_{k+1}(\mathbf{P}) + \|\mathbf{A} - \mathbf{P}\|_2 \leqslant \sigma\sqrt{n} \leqslant \sigma\|\mathbf{A}\|_F$. Thus if $\sigma \leqslant \alpha\sqrt{\delta}/\text{poly}(k)$ for a large enough poly(k), our Significant Singular Values assumption holds. □

Mixed Membership Stochastic Block Models. The Stochastic Block Model is a well-studied stochastic model for generating random graphs, where the vertices are partitioned into k communities and edges within each community are more likely to occur than edges across communities. Given communities $C_1, C_2, \ldots C_k$, there exists a $k \times k$ symmetric latent matrix $\mathbf{B}$, where, $\mathbf{B}_{\ell_1,\ell_2}$ is the probability that there exists an edge between vertices in C_{ℓ_1} and C_{ℓ_2}. The MMBM can be formalized as the following stochastic process:

- For $j \in [n]$, vertex j picks a probability vector $\mathbf{W}_{*,j} \in \mathbb{R}^k$ representing community membership probabilities that sum to 1, i.e., $\mathbf{W}_{i,j} \sim \mathsf{Dir}(1/k)$ for all $i \in [k]$.
- For all pairs $(j_1, j_2) \in [n]$, vertex j_1 picks a community ℓ_1 proportional to $\mathsf{Mult}(\mathbf{W}_{*,j_1})$ and j_2 picks a community ℓ_2 proportional to $\mathsf{Mult}(\mathbf{W}_{*,j_2})$. The edge (j_1, j_2) is included in the graph with probability $\mathbf{B}_{\ell_1,\ell_2}$. Since $\sum_{\ell_1,\ell_2} \mathbf{W}_{\ell_1,j_1}\mathbf{B}_{\ell_1,\ell_2}\mathbf{W}_{\ell_2,j_2}$ represents the edge probability of the edge (j_1, j_2), the latent variable matrix $\mathbf{P}$ of edge probabilities can be represented as $\mathbf{P} = \mathbf{W}^T\mathbf{B}\mathbf{W}^T$.

However, our reduction is not straightforward since now $\boldsymbol{P}$ depends quadratically on $\mathbf{W}$ and the only polynomial time algorithms for $\mathbf{B}$ directly rely on semidefinite programming. Further, they require non-degeneracy assumptions in order to compute a tensor decomposition provably in polynomial time [AGHK14b, HS17]. However, we can pose the problem of recovery of the k underlying communities differently and first pick at random a subset $V_1 \subset [n]$ of d vertices and represent the ℓ-th community by a d-dimensional vector that represents the probabilities of vertices in $[n] \setminus V_1$ belonging to community ℓ and having an edge with each of the d vertices in V_1. We now define $\mathbf{W}_{(1)}$ to be a $k \times d$ matrix representing the fractional membership of weights of vertices in V_1 and $\mathbf{W}_{(2)}$ to be the analogous $k \times n$ matrix for vertices in $[n] \setminus V_1$. Observe that the probability matrix $\mathbf{P}$ can now be represented as $\mathbf{W}_{(1)}^T\mathbf{B}\mathbf{W}_{(2)}$.

The reduction to the Latent Simplex Model can now be stated as follows: given a data matrix

$\mathbf{A}$ which is the adjacency matrix of the community graph, and the latent variable matrix $\mathbf{P}$, recover the simplex $\mathbf{M} = \mathbf{W}_{(1)}^T\mathbf{B}$. Further, [ABFX08] assumes that each column of $\mathbf{W}_{(2)}$ is picked from the Dirichlet distribution with parameter $1/k$. Combined with tools from random matrix theory [Ver10a], [BK20c] (Lemma 7.2) shows that the Proximate Latent Points and Spectrally Bounded assumptions hold for Stochastic Block Models. As for the Significant Singular Values assumption, it is satisfied when σ is a small enough polynomial in k.

Justifying Significant Singular Values. We give the following further justification for assumption (4) in Section 9.5: a faster algorithm only using the assumptions appearing in [BK20c] would imply an algorithmic breakthrough for spectral low-rank approximation and partially resolve the first open question of [Woo14b].

Theorem 50 (Spectral LRA and Learning a Simplex (informal))**.** *There exists a distribution over instances such that learning a latent simplex in $o(nnz(\mathbf{A}) \cdot k)$ time with good probability implies a constant factor spectral low-rank approximation algorithm in the same running time.*

Adversarial Clustering. We consider clustering problems that arise naturally from stochastic mixture models such as Gaussian, Mallows, categorical and so on [SK01, VW04, LB11, CSV17, DKS18, LM18b]. We can then formulate such a clustering problem in the Latent Simplex Model as follows: Given n data points $\mathbf{A}_{*,1}, \mathbf{A}_{*,2}, \ldots, \mathbf{A}_{*,n} \in \mathbb{R}^d$, such that the data is a mixture of k distinct clusters, $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k$, with means $\mathbf{M}_{*,1}, \mathbf{M}_{*,2}, \ldots, \mathbf{M}_{*,k}$, the goal is to approximately learn the means. Further, we can set the n latent vectors $\mathbf{P}_{*,j}$ to denote the mean of the cluster point $\mathbf{A}_{*,j}$ belongs to, and thus $\mathbf{P}_{*,j} \in \{\mathbf{M}_{*,1}, \mathbf{M}_{*,2}, \ldots, \mathbf{M}_{*,k}\}$. Prior work of [KK10] and [AS12] shows that if the minimum cluster size if δn and for all $\ell \neq \ell'$, $\|\mathbf{M}_{*,\ell} - \mathbf{M}_{*,\ell'}\| \geqslant ck\frac{\sigma}{\sqrt{\delta}}$ the $\mathbf{M}_{*,\ell}$ can be found within error $O(\sqrt{k}\sigma/\sqrt{\delta})$.

However, the aforementioned algorithms are not robust to adversarial perturbations. Therefore, we describe the perturbations we can handle in the Latent Simplex Model. The adversarial model is the same as the one considered in [BK20c]. The adversary is allowed to selected a subset S_ℓ of each cluster $\mathbf{C}_\ell$ of cardinality at most δn and perturb each point $\mathbf{A}_{*,j}$ for $j \in S_\ell$ by Δ_j such that :

- $\mathbf{P}_{*,j} + \Delta_j$ is still in the Convex Hull of $\mathbf{M}_{*,1}, \mathbf{M}_{*,2}, \ldots, \mathbf{M}_{*,k}$
- The norm of the perturbation is bounded, i.e., $|\Delta_j|_2 \leqslant 4\sigma/\sqrt{\delta}$.

Intuitively, the adversary can move a $1 - \delta$ fraction of the data points in each cluster an arbitrary amount towards the convex hull of the means of the remaining clusters. For the remaining δn, the perturbation should have norm at most $O(\sigma/\sqrt{\delta})$. The goal is to still learn the means

$\mathrm{M}_{*,\ell}$ approximately. [BK20c] shows that the aforementioned model satisfies Well-Separateness, Proximate Latent Points and Spectrally Bounded Perturbations assumptions. The proof for the Significant Singular Values assumption follows from Lemma 1.2.2. We note that there has been a flurry of recent progress on adversarial clustering in the strong contamination model, where the input data points are sampled from a mixture of Gaussians distribution and the adversary can corrupt a small fraction of the samples arbitrarily [DKS18, HL18, KSS18, DHKK20, BK20b]. In our setting, there is no distribution assumption on the data points but the adversary is constrained as the norm of the perturbation is bounded.

1.3 Roadmap of the Thesis

This thesis is divided into two parts, each focusing on one of the two distinct regimes of learning latent models, as discussed in this section. We note that the goals, motivations and technical ideas we use in the separate parts varies considerably. However, each chapter in the two parts is designed to be self-contained and thus introduces the notation, background and preliminaries used in that chapter. While this leads to some redundancy of definitions across chapters, we believe it vastly improves the readability of each chapter. Next, we outline the chapters in each part and the paper corresponding to that chapter:

Part I : Establishing Tractability of Latent Models

1. Chapter 2: Outlier-Robust Clustering of Non-spherical Mixtures [BK20b], with Pravesh Kothari. FOCS '20.
2. Chapter 3: Robustly Learning a Mixture of k Arbitrary Gaussians [BDJ+22], with Ilias Diakonikolas, He Jia, Daniel Kane, Pravesh Kothari and Santosh Vempala. STOC '22.
3. Chapter 4: Robust Linear Regression: Optimal Rates in Polynomial Time [BP21], with Adarsh Prasad. STOC '21.
4. Chapter 5: List-Decodable Subspace Recovery [BK21], with Pravesh Kothari. SODA '21.
5. Chapter 6: Learning a Two-Layer Neural Network in Polynomial Time [BJW19], with Rajesh Jayaram and David Woodruff. COLT '18.

Part II : Nearly Optimal Algorithms for Learning Latent Models

1. Chapter 7: Low-Rank Approximation with $1/\epsilon^{1/3}$ Matrix-Vector Products [BCW22], with Ken Clarkson and David Woodruff. STOC '22.

2. Chapter 8: PSD Low-Rank Approximation [BCW20a], with Nadiia Chepurko and David Woodruff. FOCS '20.
3. Chapter 9: Learning a Latent Simplex in Truly Input-Sparsity Time [BBK+21a], with Chiranjeeb Bhattacharya, Ravi Kannan, David Woodruff and Samson Zhou. ICLR '21.

We note that [BW18, ABB+19, BCJ20] and [BCW19] do not appear in this thesis.

Part I

Establishing Tractability of Latent Models

Chapter 2

Outlier-Robust Clustering of Non-Spherical Mixtures

2.1 Introduction

In this chapter, we study outlier-robust *clustering* of mixtures of distributions that exhibit mean or covariance separation. As a corollary, we obtain a polynomial time outlier-robust algorithm for clustering mixtures of k-Gaussians (k-GMMs) when each pair of components is separated in total variation (TV)[1] distance. This is the information-theoretically weakest notion of separation, allows components of same mean but variances differing in an unknown direction[2] or covariances separated in *relative* Frobenius distance (see Fig 2.1) and includes well-studied problems such as *mixed linear regression* and *subspace clustering* as special cases.

Clustering all Hypercontractive and Anti-Concentrated Distributions. The Gaussian Mixture Model has been the subject of a century-old line of research beginning with Pearson [Pea94]. A k-GMM $\sum_{r\leqslant k} p_r \mathcal{N}(\mu(r), \Sigma(r))$ is a probability distribution sampled by choosing a component $r \sim [k]$ with probability p_r and outputting a sample from the Gaussian distribution with mean $\mu(r)$ and covariance $\Sigma(r)$. In the k-GMM learning problem, the goal is to output an approximate *clustering* of the input sample or estimate the parameters (the mean and covariances) of the components. Progress on provable algorithms for learning k-GMMs began with the influential

[1]The TV distance between distributions with PDFs p, q is defined as $\frac{1}{2}\int_{-\infty}^{\infty} |p(x) - q(x)|dx$.

[2]As an interesting example, consider the case of subspace clustering: mixture of standard Gaussians restricted to unknown distinct subspaces. The components have a TV distance of 1 regardless of how close the subspaces are and thus satisfy our assumptions.

work of Dasgupta [Das99] followed up by [SK01, VW04, BV08] yielding clustering algorithms that succeed under various separation assumptions. These assumptions, however, do not capture natural separated instances of Gaussians (e.g., see (b) or (c) in Fig 2.1). A more general approach [KMV10, MV10, BS15] circumvents clustering altogether by giving an efficient algorithm (time $\sim d^{\text{poly}(k)}$) for parameter estimation without any separation assumptions.

Our main result is a polynomial-time algorithm based on the sum-of-squares (SoS) method for clustering TV-separated k-GMMs in the presence of an ϵ-fraction of fully adversarial outliers. Such a result was not known prior to our work even for $k = 2$. Our algorithms actually succeed more generally for mixtures of all distributions that satisfy two well-studied analytic conditions: certifiable *anti-concentration* and certifiable *hypercontractivity* and thus apply, for e.g., to clustering mixtures of arbitrary affine transforms of uniform distribution on the unit sphere. We consider identifying clean analytic conditions that enable the existence of efficient clustering algorithms an important contribution of our work.

2.1.1 Our Results

Outlier-Robust Clustering of k-GMMs. Our main result is an efficient algorithm for outlier-robust clustering of k-GMMs whenever every pair of components of the mixture are separated in total variation distance. Formally, our algorithms work in the *strong contamination* model studied in the bulk of the prior works on robust estimation where an adversary changes an arbitrary, potentially adversarially chosen ϵ-fraction of the input sample before passing it on to the algorithm.

Theorem 51 (Main Result, Outlier-Robust Clustering of k-GMMs)**.** *Fix $\eta, \epsilon > 0$. Let $\mathcal{D}_r = \mathcal{N}(\mu(r), \Sigma(r))$ for $r \leqslant k$ be k-Gaussians such that $d_{TV}(\mathcal{D}_r, \mathcal{D}_{r'}) \geqslant 1 - \exp(-\text{poly}(k/\eta))$ whenever $r \neq r'$. Then, there exists an algorithm that takes input an ϵ-corruption Y of a sample $X = C_1 \cup C_2 \cup \ldots \cup C_k$ of size n, with equal sized clusters C_i drawn i.i.d. from $\mathcal{D}_i$ for each $r \leqslant k$, and with probability at least 0.99, outputs an approximate clustering $Y = \hat{C}_1 \cup \hat{C}_2 \cup \ldots \cup \hat{C}_k$ satisfying $\min_{i \leqslant k} \frac{|\hat{C}_i \cap C_i|}{|C_i|} \geqslant 1 - O(k^{2k})(\epsilon + \eta)$. The algorithm succeeds whenever $n \geqslant d^{O(\text{poly}(k/\eta))}$ and runs in time $n^{O(\text{poly}(k/\eta))}$.*

We can use off-the-shelf robust estimators for mean and covariance of Gaussians([DKK+19]) in order to get statistically optimal estimates of the mean and covariances of the target k-GMM.

Corollary 2.1.1 (Parameter Recovery from Clustering)**.** *In the setting of Theorem 51, with*

the same running time, sample complexity and success probability, our algorithm can output $\{\hat{\mu}(r), \hat{\Sigma}(r)\}_{r \leqslant k}$ *such that for some permutation* $\pi : [k] \to [k]$*,*

$$d_{TV}\left(\mathcal{N}\left(\mu(r), \Sigma(r)\right), \mathcal{N}\left(\hat{\mu}(\pi(r)), \hat{\Sigma}(\pi(r)\right)\right) \leqslant \tilde{O}(k^{2k}(\epsilon + \eta)).$$

Discussion. These are the first outlier-robust algorithms that work for clustering k-GMMs under information-theoretically optimal separation assumptions. Such results were not known even for $k = 2$. To discuss the bottlenecks in prior works, it is helpful to use (see Prop 2.9.1 in Section 2.9 for a proof) following consequence of two Gaussians with means $\mu(1), \mu(2)$ and covariances $\Sigma(1), \Sigma(2)$ being at a TV distance $\geqslant 1 - \exp(-O(\Delta^2))$ in terms of the distance between their parameters.

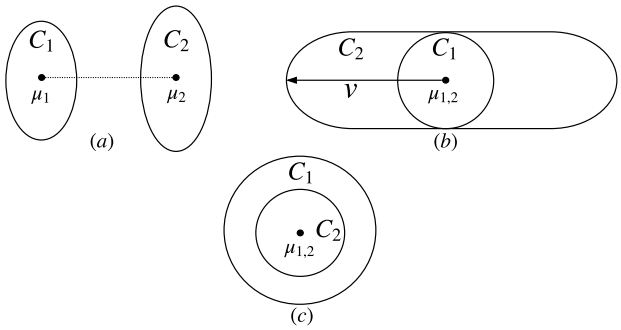

Figure 2.1: (a) Mean Separation (b) Spectral Separation (c) Relative Frobenius Separation

Definition 2.1.2 (Δ-Separated Mixture Model)**.** *An equi-weighted mixture* $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ *with parameters* $\{\mu(i), \Sigma(i)\}_{i \leqslant k}$ *is* Δ*-separated if for every pair of distinct components* i, j*, one of the following three conditions hold (*$\Sigma^{\dagger/2}$ *is the square root of pseudo-inverse of* Σ*):*

1. ***Mean-Separation:*** $\exists v \in \mathcal{R}^d$ *such that* $\langle \mu(i) - \mu(j), v \rangle^2 > \Delta^2 v^\top (\Sigma(i) + \Sigma(j))v$*,*
2. ***Spectral-Separation:*** $\exists v \in \mathcal{R}^d$ *such that* $v^\top \Sigma(i) v > \Delta v^\top \Sigma(j) v$*,*
3. ***Relative-Frobenius Separation:***[3] $\Sigma(i)$ *and* $\Sigma(j)$ *have the same range space and*

$$\left\|\Sigma(i)^{\dagger/2}\Sigma(j)\Sigma(i)^{\dagger/2} - I\right\|_F^2 > \Delta^2 \left\|\Sigma(i)^{\dagger/2}\Sigma(j)^{1/2}\right\|_{op}^4.$$

The key bottleneck for known algorithms was handling separation in cases 2 and 3 above.

[3]Unlike the other two distances, relative Frobenius distance is meaningful only for high-dimensional Gaussians. As an illustrative example, consider two 0 mean Gaussians with covariances $\Sigma_1 = I$ and $\Sigma_2 = (1 + \Theta(1/\sqrt{d}))I$. Then, for large enough d, the parameters are separated in relative Frobenius distance but not spectral or mean distance.

Dependence on k. The dependence on the number of components k in our result is doubly exponential. A singly exponential lower bound in the statistical query model (for even the non-robust variant) was shown by Diakonikolas, Kane and Stewart [DKS17].

Dependence on ϵ: While the information-theoretically optimal bound on fraction of misclassified samples is $O(k\epsilon)$, we only obtain the weaker bound of $k^{O(k)}\epsilon$. Our algorithms in Sections 2.3, 2.4 do obtain this the stronger $O(k\epsilon)$ guarantee at the cost of a larger running time. We believe it should be possible to match the optimal recovery guarantee without incurring this running time penalty.

Handling General Weights: While we have not attempted to do it in this work, it seems possible to generalize our techniques to handle arbitrary mixing weights albeit with an exponential dependence on the reciprocal of the smallest mixing weight in both the running time and sample complexity on the algorithm.

Clustering and Parameter Recovery for all Reasonable Distributions. Our results apply more generally to mixture models where each component distribution $\mathcal{D}$ satisfies two natural and well-studied analytic conditions: *hypercontractivity* and bounded variance of degree 2 polynomials and *anti-concentration* of all directional marginals. Our algorithmic results hold for distributions (e.g. Gaussians and affine transforms of uniform distribution on the unit sphere) that admit efficiently verifiable analogs (in the SoS proof system, see Sec 5.3) of these properties.

Definition 2.1.3 (Certifiable Hypercontractivity)**.** *An isotropic distribution $\mathcal{D}$ on $\mathcal{R}^d$ is said to be h-certifiably C-hypercontractive if there's a degree h sum-of-squares proof of the following unconstrained polynomial inequality in $d \times d$ matrix-valued indeterminate Q:*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx - \mathbf{E}_{x\sim\mathcal{D}}x^\top Qx\right]^h \leqslant (Ch)^h \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx - \mathbf{E}_{x\sim\mathcal{D}}x^\top Qx\right]^2\right)^{h/2}.$$

A set of points $X \subseteq \mathcal{R}^d$ is said to be C-certifiably hypercontractive if the uniform distribution on X is h-certifiably C-hypercontractive.

Hypercontractivity is an important notion in high-dimensional probability and analysis on product spaces [O'D14]. Kauers, O'Donnell, Tan and Zhou [KOTZ14] showed certifiable hypercontractivity of Gaussians and more generally product distributions with subgaussian marginals. Certifiable hypercontractivity strictly generalizes the better known *certifiable subgaussianity* property (studied first in [KSS18]) that controls higher moments of linear polynomials.

Certifiable anti-concentration. In contrast to subgaussianity, anti-concentration forces *lower-bounds* of the form $\mathbf{Pr}[\langle x, v\rangle^2 \geqslant \delta \|v\|_2^2] \geqslant \delta'$ for all directions v. Certifiable anti-concentration was recently introduced in independent works of Karmalkar, Klivans and Kothari [KKK19] and Raghavendra and Yau [RY20a] and later used [BK20a, RY20b] for the related problems of list-decodable linear regression and subspace recovery[4].

Following [KKK19], we formulate certifiable anti-concentration via a univariate, even polynomial $p_{\delta,\Sigma}$ that uniformly approximates the 0-1 core-indicator $\mathbf{1}(\langle x, v\rangle^2 \geqslant \delta v^\top \Sigma v)$ over a large enough interval around 0. Let $q_{\delta,\Sigma}(x, v)$ be a multivariate (in v) polynomial defined by $q_{\delta,\Sigma}(x, v) = \left(v^\top \Sigma v\right)^{2s} p_{\delta,\Sigma}\left(\frac{\langle x,v\rangle}{\sqrt{v^\top \Sigma v}}\right)$.Since $p_{\delta,\Sigma}$ is an even polynomial, $q_{\delta,\Sigma}$ is a polynomial in v.

Definition 2.1.4 (Certifiable Anti-Concentration)**.** *A mean 0 distribution D with covariance Σ is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated if for $q_{\delta,\Sigma}(x, v)$ defined above, there exists a degree $2s$ sum-of-squares proof of the following two unconstrained polynomial inequalities in indeterminate v:*

$$\left\{\langle x, v\rangle^{2s} + \delta^{2s} q_{\delta,\Sigma}(x, v)^2 \geqslant \delta^{2s}\left(v^\top \Sigma v\right)^{2s}\right\}, \left\{\mathbf{E}_{x\sim D} q_{\delta,\Sigma}(x, v)^2 \leqslant C\delta\left(v^\top \Sigma v\right)^{2s}\right\}.$$

An isotropic subset $X \subseteq \mathcal{R}^d$ is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated if the uniform distribution on X is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated.

Remark 52. For natural examples, $s(\delta) \leqslant 1/\delta^c$ for some fixed constant c. For e.g., $s(\delta) = O(\frac{1}{\delta^2})$ for standard Gaussian distribution and the uniform distribution on the unit sphere (see [KKK19] and [BK20a]). To simplify notation, we will assume $s(\delta) \leqslant \text{poly}(1/\delta)$ in the statement of our results.

Additionally, we need that the variance of degree-2 polynomials is bounded in terms of the Frobenius norm of the coefficients of the polynomial. Formally,

Definition 2.1.5 (Degree-2 Polynomials with Certifiably Bounded Variance)**.** *A mean 0 distribution $\mathcal{D}$ with covariance Σ certifiably bounded variance degree 2 polynomials if there is a degree 2 sum-of-squares proof of the following inequality in the indeterminate $Q \in \mathbb{R}^{d\times d}$*

$$\left\{\mathbf{E}_{x\sim\mathcal{D}}\left(x^\top Q x - \mathbf{E}_{x\sim\mathcal{D}} x^\top Q x\right)^2 \leqslant C\left\|\Sigma^{1/2} Q \Sigma^{1/2}\right\|_F^2\right\}.$$

[4]List-decodable versions of these problems generalize the "mixture" variants - mixed linear regression and subspace clustering - that are easily seen to be special cases of mixtures of k-Gaussians with TV separation 1.

Our general result gives an outlier-robust clustering algorithm for separated mixtures of *reasonable* distributions, i.e., one that satisfies both certifiable hypercontractivity, anti-concentration and has bounded variance of degree-2 polynomials. Even the information-theoretic (and without outliers, i.e., $\epsilon = 0$) clusterability of such distributions was not known prior to our work.

Theorem 53 (Outlier-Robust Clustering of Separated Mixtures, see Theorem 61 for precise bounds)**.** *Fix $\eta > 0, \epsilon > 0$. Let $\mathcal{D}_r$ be a Δ-separated mixture of reasonable distributions. Then, there exists an algorithm that takes input an ϵ-corruption Y of a sample $X = C_1 \cup C_2 \cup \ldots C_k$, with true clusters C_i of size n/k drawn i.i.d. from $\mathcal{D}_r$ for each $r \leqslant k$, and outputs an approximate clustering $Y = \hat{C}_1 \cup \hat{C}_2 \cup \ldots \cup \hat{C}_k$ satisfying $\min_{i \leqslant k} \frac{|\hat{C}_i \cap C_i|}{|C_i|} \geqslant 1 - O(k^{2k})(\epsilon + \eta)$. The algorithm succeeds with probability at least 0.99 over the draw of the original sample X whenever $n \geqslant d^{poly(k/\eta)}$ and runs in time $n^{poly(k/\eta))}$ whenever $\Delta \geqslant poly(k/\eta)^k$.*

Robust Covariance Estimation in Relative Frobenius Distance. In Section 2.6, we give an outlier-robust algorithm for covariance estimation for all certifiably hypercontractive distributions.

Theorem 54 (Robust Parameter Covariance Estimation for Certifiably Hypercontractive Distributions)**.** *Fix an $\epsilon > 0$ small enough fixed constant so that $Ct\epsilon^{1-1/t} \ll 1$[5]. For every even $t \in \mathbb{N}$, there's an algorithm that takes input Y be an ϵ-corruption of a sample X of size $n \geqslant n_0 = d^{O(t)}/\epsilon^2$ from a $2t$-certifiably C-hypercontractive and certifiably C-bounded variance with unknown mean μ_* and covariance Σ_* respectively and in time $n^{O(t)}$ outputs an estimate $\hat{\mu}$ and $\hat{\Sigma}$ satisfying:*

1. $\left\|\Sigma^{-1/2}(\mu_* - \hat{\mu})\right\|_2 \leqslant O(Ct)^{1/2}\epsilon^{1-1/t}$,
2. $(1-\eta)\Sigma_* \preceq \hat{\Sigma} \preceq (1+\eta)\Sigma_*$ *for* $\eta \leqslant O(Ck)\epsilon^{1-1/t}$, *and,*
3. $\left\|\Sigma_*^{-1/2}\hat{\Sigma}\Sigma_*^{-1/2} - I\right\|_F \leqslant (Ct)O(\epsilon^{1-1/t})$.

In particular, letting $t = O(\log(1/\epsilon))$ results in the error bounds of $\tilde{O}(\epsilon)$ in all the three inequalities above.

The first two guarantees above were shown in [KSS18] for all certifiably subgaussian distributions. [KSS18] also observed (see last paragraph of page 6 for a counter example) that it is provably impossible to obtain dimension-independent error bounds in relative Frobenius distance assuming only certifiable subgaussianity. We prove that under the stronger assumption of certifi-

[5]This notation means that we needed $Ct\epsilon^{1-1/t}$ to be at most c_0 for some absolute constant $c_0 > 0$

able *hypercontractivity* along with certifiably bounded variance of degree 2 polynomials, we can indeed obtain dimension-independent, information-theoretically optimal (for e.g. for Gaussians) error guarantees in relative Frobenius error. Prior works either obtained the weaker spectral error guarantee (that incurs a loss of $\sqrt{d}$ factor when translating into relative Frobenius distance) or worked only for Gaussians[6].

Combining this theorem with our clustering results above yields:

Corollary 2.1.6 (Parameter Recovery from Clustering, General Case)**.** *In the setting of either Theorem 53, there's an algorithm with same bounds on running time and sample complexity, that with probability at least* 0.99*, outputs* $\{\hat{\mu}(r), \hat{\Sigma}(r)\}_{r \leqslant k}$ *such that for some permutation* $\pi : [k] \to [k]$*, for every* i*,* $\hat{\mu}(\pi(i)), \hat{\Sigma}(\pi(i))$ *is* Δ*-close to* μ, Σ *in the three distances defined in Definition 3.4.1 for* $\Delta = \tilde{O}(k^{O(k)}(\epsilon + \eta))$*.*

2.2 Preliminaries

Throughout this paper, for a vector v, we use $\|v\|_2$ to denote the Euclidean norm of v. For a $n \times m$ matrix M, we use $\|M\|_2 = \max_{\|x\|_2=1} \|Mx\|_2$ to denote the spectral norm of M and $\|M\|_F = \sqrt{\sum_{i,j} M_{i,j}^2}$ to denote the Frobenius norm of M. For symmetric matrices we use $\succeq$ to denote the PSD/Löwner ordering over eigenvalues of M. For a $n \times n$, rank-r symmetric matrix M, we use $U\Lambda U^\top$ to denote the Eigenvalue Decomposition, where U is a $n \times r$ matrix with orthonormal columns and Λ is a $r \times r$ diagonal matrix denoting the eigenvalues. We use $M^\dagger = U\Lambda^\dagger U^\top$ to denote the Moore-Penrose pseudoinverse, where $\Lambda^\dagger$ inverts the non-zero eigenvalues of M. If $M \succeq 0$, we use $M^{\dagger/2} = U\Lambda^{\dagger/2}U^\top$ to denote taking the square-root of the non-zero eigenvalues. We use $\Pi = UU^\top$ to denote the Projection matrix corresponding to the column/row span of M. Since $\Pi = \Pi^2$, the pseudo-inverse of Π is itself, i.e. $\Pi^\dagger = \Pi$.

Definition 2.2.1 (σ-Sub-gaussian Distribution)**.** *A random variable* x *is drawn from a* σ*-Sub-gaussian distribution if for all* $t \geqslant 0$*,* $\mathbf{Pr}\left[|x| \geqslant t\right] \leqslant 2\exp(-t^2/\sigma^2)$*.*

We work with 1-Sub-gaussian distributions unless otherwise specified and drop the 1 when clear from context.

[6]We note that the algorithm of [DKK+19] for Gaussian distributions works in fixed polynomial time to obtain $\tilde{O}(\epsilon)$ error-estimate of the covariance in relative Frobenius distance whereas our algorithm works more generally for all certifiably hypercontractive distributions but runs in time $d^{O(log^2(1/\epsilon))}$.

Probability Preliminaries. We begin with standard convergence results for mean and covariance.

Fact 2.2.2 (Empirical Mean for Sub-gaussians)**.** *Let $\mathcal{D}$ be a Sub-gaussian distribution on $\mathcal{R}^d$ with mean μ and covariance Σ and let $x_1, x_2, \ldots x_n \sim \mathcal{D}$. Then, with probability $1-\delta$,*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} x_i - \mu \right\|_2 \leqslant \sqrt{\frac{\textit{Tr}\left[(\right]\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_2 \log(1/\delta)}{n}}$$

Fact 2.2.3 (Empirical Covariance for Sub-gaussians, Proposition 2.1 [Ver18])**.** *Let $\mathcal{D}$ be a Sub-gaussian distribution on $\mathbb{R}^d$ with mean μ and covariance Σ and let $x_1, x_2, \ldots x_n \sim \mathcal{D}$. Then, with probability $1-\delta$,*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top - \Sigma \right\|_2 \leqslant c\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Definition 2.2.4 (Hellinger Distance)**.** *For probability distribution p, q on $\mathcal{R}^d$, let*

$$h(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\int_{\mathcal{R}^d}\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx}$$

be the Hellinger distance between them.

Remark 55. Hellinger distance between p, q satisfies: $h(p,q)^2 \leqslant d_{\mathsf{TV}}(p,q) \leqslant h(p,q)\sqrt{2-h(p,q)^2}$.

Fact 2.2.5 (Hellinger Distance between Gaussians)**.**

$$h(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\mu',\Sigma'))^2 = 1 - \frac{\det(\Sigma)^{1/4}\det(\Sigma')^{1/4}}{\det\left(\frac{\Sigma+\Sigma'}{2}\right)^{\frac{1}{2}}}\exp\left(-\frac{1}{8}(\mu-\mu)^\top\left(\frac{\Sigma+\Sigma'}{2}\right)^{-1}(\mu-\mu')\right)$$

Basic Sum-of-Squares Proofs

Fact 2.2.6 (Operator norm Bound)**.** *Let A be a symmetric $d \times d$ matrix and v be a vector in $\mathbb{R}^d$. Then,*

$$\left\vert\frac{v}{2}\right\{ v^\top A v \leqslant \|A\|_2 \|v\|_2^2 \right\}$$

Fact 2.2.7 (SoS Hölder's Inequality)**.** *Let f_i, g_i for $1 \leqslant i \leqslant s$ be indeterminates. Let p be an*

even positive integer. Then,

$$\left|\frac{f,g}{p^2}\right. \left\{ \left(\frac{1}{s}\sum_{i=1}^{s} f_i g_i^{p-1}\right)^p \leqslant \left(\frac{1}{s}\sum_{i=1}^{s} f_i^p\right)^q \left(\frac{1}{s}\sum_{i=1}^{s} g_i^p\right)^{p-1} \right\}.$$

Observe that using $p = 2$ yields the SoS Cauchy-Schwarz inequality.

Fact 2.2.8 (SoS Almost Triangle Inequality)**.** *Let $f_1, f_2, \ldots, f_r$ be indeterminates. Then,*

$$\left|\frac{f_1,f_2,\ldots,f_r}{2t}\right. \left\{ \left(\sum_{i\leqslant r} f_i\right)^{2t} \leqslant r^{2t-1}\left(\sum_{i=1}^{r} f_i^{2t}\right) \right\}.$$

Fact 2.2.9 (SoS AM-GM Inequality, see Appendix A of [BKS15])**.** *Let $f_1, f_2, \ldots, f_m$ be indeterminates. Then,*

$$\left|\frac{f_1,f_2,\ldots,f_m}{m}\right. \left\{ \left(\frac{1}{m}\sum_{i=1}^{n} f_i\right)^m \geqslant \Pi_{i\leqslant m} f_i \right\}.$$

The following fact is a simple corollary of the fundamental theorem of algebra:

Fact 2.2.10. *For any univariate degree d polynomial $p(x) \geqslant 0$ for all $x \in \mathcal{R}$, $\left|\frac{x}{d}\right. \{p(x) \geqslant 0\}$.*

This can be extended to univariate polynomial inequalities over intervals of $\mathcal{R}$. 2

Fact 2.2.11 (Fekete and Markov-Lukacs, see [Lau09])**.** *For any univariate degree d polynomial $p(x) \geqslant 0$ for $x \in [a, b]$, $\{x \geqslant a, x \leqslant b\}$ $\left|\frac{x}{d}\right. \{p(x) \geqslant 0\}$.*

2.2.1 Certifiable Anti-Concentration

This definition is a homogenous variant of the one proposed in [KKK19].

Definition 2.2.12 (Certifiable Anti-Concentration)**.** *A zero-mean distribution D with covariance Σ is $2k$-certifiably $(\delta, C\delta)$-anti-concentrated if there exists a univariate polynomial p of degree k such that:*

1. $\left|\frac{v}{2k}\right. \left\{\|v\|_2^{2k-2}\left\langle \Sigma^{-1/2}x, v\right\rangle^2 + \delta^2 p^2\left(\left\langle \Sigma^{-1/2}x, v\right\rangle\right) \geqslant \delta^2\|v\|_2^{2k}\right\}$.
2. $\left|\frac{v}{2k}\right. \left\{\mathbf{E}_{\Sigma^{-1/2}x\sim D} p^2\left(\left\langle \Sigma^{-1/2}x, v\right\rangle\right) \leqslant C\delta\|v\|_2^{2k}\right\}$.

A subset $X \subseteq \mathcal{R}^d$ is $2k$-certifiably $(\delta, C\delta)$-anti-concentrated if the uniform distribution on X is

$2k$-cerfiably $(\delta, C\delta)$-anti-concentrated.

Definition 2.2.13 (Certifiable Anti-Concentration)**.** *A random variable (and its distribution) Y has a k-certifiably (C, δ)-anti-concentrated distribution if there is a univariate polynomial p satisfying $p(0) = 1$ such that there is a degree k sum-of-squares proof of the following two inequalities:*

1. $\langle Y, v\rangle^2 \leqslant \delta^2 \mathbf{E}\langle Y, v\rangle^2$ *implies* $(p(\langle Y, v\rangle) - 1)^2 \leqslant \delta^2$.
2. $\forall v$, $\mathbf{E}\langle Y, v\rangle^2 > 0$ *implies* $\mathbf{E}\left[\mathbf{E}\langle Y, v\rangle^2 p^2(\langle Y, v\rangle)\right] \leqslant C\delta \langle Y, v\rangle^2$.

A set of points $S \subseteq \mathcal{R}^d$ are said to be k-certifiably (C, δ)-anti-concentrated if uniform distribution on S is k-certifiably (C, δ)-anti-concentrated.

2.3 Clustering Mixtures of Reasonable Distributions

In this section, we provide algorithm for clustering mixtures of *reasonable* distributions. The main results of this section are *simultaneous intersection bounds* (Lemmas 2.3.5, 2.3.13, and 2.3.4) that we'll rely on in the next two sections. We then use these bounds to immediately derive an algorithm (via the rounding used in Chapter 4.3 of [FKP+19]) for clustering that runs in time $d^{\text{poly}(k)\log(\kappa)}$ where κ is the *spread* of the mixture defined as the maximum of $\frac{v^\top \Sigma(j) v}{v^\top \Sigma(i) v}$ over all $i, j \leqslant k$. In Section 2.5, we will show how to improve the running time of this algorithm to have no dependence on the spread and prove our main result (Theorem 53).

Theorem 56 (Clustering Mixtures of Separated Reasonable Distributions)**.** *For any $\eta > 0$, there exists an algorithm that takes input a sample of size n from Δ-separated equi-weighted mixture of reasonable distributions $\mathcal{D}(\mu(r), \Sigma(r))$ for $r \leqslant k$ with true clusters $C_1, C_2, \ldots, C_k$ and outputs $\hat{C}_1, \hat{C}_2, \ldots \hat{C}_k$ such that there exists a permutation $\pi : [k] \to [k]$ satisfying*

$$\min_{i \leqslant k} \frac{|C_i \cap \hat{C}_{\pi(i)}|}{|C_i|} \geqslant 1 - O(\eta)\,.$$

The algorithm succeeds with probability at least $1 - 1/k$ whenever $\Delta = \Omega\left((k/\eta)^c\right)$, for a large enough fixed universal constant c, needs $d^{poly(k/\eta)}$ samples and runs in time $n^{poly(k/\eta)\log(\kappa)}$ where $\kappa = \sup_{v \in \mathbb{R}^d} \max_{i,j \in [k]} \frac{v^\top \Sigma(j) v}{v^\top \Sigma(i) v}$ is spread of the mixture.

2.3.1 Algorithm

Our constraint system $\mathcal{A}$ uses polynomial inequalities to describe a subset $\hat{C}$ of size αn of the input sample X. We impose constraints on $\hat{C}$ so that the uniform distribution on $\hat{C}$ satisfies certifiable anti-concentration and hypercontractivity of degree-2 polynomials. We intend the true clusters $C_1, C_2, \ldots, C_r$ to be the only solutions for $\hat{C}$. Proving that this statement holds and that it has a low-degree SoS proof is the bulk of our technical work in this section.

We describe the specific formulation next. Throughout this section, we use the notation $Q(x)$ to denote $x^\top Q x$ for $d \times d$ matrix valued indeterminate Q. For ease of exposition, we break our constraint system $\mathcal{A}$ into natural categories $\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_5$. Our constraint system relies on parameter τ, δ that we will set in proof of Theorem 56 below.

For our argument, we will need access to the square root of the indeterminate Σ. So we introduce the constraint system $\mathcal{A}_1$ with an extra matrix valued indeterminate Π (with auxiliary matrix-valued indeterminate U) that satisfies the polynomial equality constraints corresponding to Π being the square root of Σ. Note that the first constraint is equivalent to $\Pi \succeq 0$ in "ordinary math".

$$\text{Square-Root Constraints: } \mathcal{A}_1 = \left\{ \begin{array}{r} \Pi = UU^\top \\ \Pi^2 = \Sigma\,. \end{array} \right\} \tag{2.1}$$

Next, we formulate intersection constraints that identify the subset $\hat{C}$ of size αn.

$$\text{Subset Constraints: } \mathcal{A}_2 = \left\{ \begin{array}{ll} \forall i \in [n] & w_i^2 = w_i \\ & \sum_{i \in [n]} w_i = \dfrac{n}{k}\,. \end{array} \right\} \tag{2.2}$$

Next, we enforce that $\hat{C}$ must have mean μ and covariance Σ, where both μ and Σ are indeterminates.

$$\text{Parameter Constraints: } \mathcal{A}_3 = \left\{ \begin{array}{r} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} w_i x_i = \mu \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} w_i \left(x_i - \mu\right)\left(x_i - \mu\right)^\top = \Sigma\,. \end{array} \right\} \tag{2.3}$$

Finally, we enforce certifiable anti-concentration at two slightly different parameter regimes

(characterized by $\tau \leqslant \delta$) along with the hypercontractivity of $\hat{C}$.

$$\text{Certifiable Anti-Concentration} : \mathcal{A}_4 = \left\{ \begin{array}{l} \frac{k^2}{n^2} \sum_{i,j=1}^{n} w_i w_j q_{\delta,2\Sigma}^2 \left((x_i - x_j), v \right) \leqslant 2^{s(\delta)} C\delta \left(v^\top \Sigma v \right)^{s(\delta)} \\ \frac{k^2}{n^2} \sum_{i,j=1}^{n} w_i w_j q_{\tau,2\Sigma}^2 \left((x_i - x_j), v \right) \leqslant 2^{s(\tau)} C\tau \left(v^\top \Sigma v \right)^{s(\tau)} \end{array} \right\}, \tag{2.4}$$

where $s(x) = \tilde{O}\left(1/x^2\right)$. Certifiable Hypercontractivity: $\mathcal{A}_5 =$

$$\left\{ \begin{array}{l} \forall h \leqslant 2s, \quad \frac{k^2}{n^2} \sum_{i,j \leqslant n} w_i w_j \left(Q(x_i - x_\ell) - \frac{k^2}{n^2} \sum_{i,\ell \leqslant n} w_i w_\ell Q(x_i - x_\ell) \right)^{2h} \\ \qquad\qquad \leqslant (Ch)^{2h} \left(\frac{k^2}{n^2} \sum_{i,\ell \leqslant n} w_i w_\ell \left(Q(x_i - x_\ell) - \frac{k^2}{n^2} \sum_{i,\ell \leqslant n} w_i w_\ell Q(x_i - x_\ell) \right)^2 \right)^h \end{array} \right\}. \tag{2.5}$$

Certifiable Bounded Variance: $\mathcal{A}_6 =$

$$\left\{ \frac{k^2}{n^2} \sum_{i,\ell \leqslant n} w_i w_\ell \left(Q(x_i - x_\ell) - \frac{k^2}{n^2} \sum_{i,\ell \leqslant n} w_i w_\ell Q(x_i - x_\ell) \right)^2 \leqslant C \left\| \Pi Q \Pi \right\|_F^2 . \right\} \tag{2.6}$$

Algorithm. We are now ready to describe our algorithm. Our algorithm follows the same outline as the simplified proof for clustering spherical mixtures presented in [FKP+19] (Chapter 4.3). The idea is to find a pseudo-distribution $\tilde{\zeta}$ that minimizes the objective $\left\| \tilde{\mathbb{E}}[w] \right\|_2$ and is consistent with the constraint system $\mathcal{A}$.

It is simple to round the resulting solution to true clusters: our analysis yields that the matrix $\tilde{\mathbb{E}}[ww^\top]$ is approximately block diagonal with the blocks approximately corresponding to the true clusters $C_1, C_2, \ldots, C_k$. We can then recover a cluster by a repeatedly greedily selecting n/k largest entries in a random row, removing those columns off and repeating. We describe this algorithm below.

Algorithm 57 (Clustering General Mixtures)**.**

Given: *A sample X of size n with true clusters $C_1, C_2, \ldots, C_k$ of size n/k each, accuracy parameter $\eta > 0$.*

Output: *A partition of X into an approximately correct clusters $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_k$.*

Operation:

1. *Find a pseudo-distribution $\tilde{\zeta}$ satisfying $\mathcal{A}$ with $s = \log(\kappa)$poly(k/η), $\delta = \eta^6/k^{12}$, and $\tau = 1/(C$poly$(k))$, and minimizing $\left\|\tilde{\mathbb{E}}[w]\right\|_2^2$.*
2. *For $M = \tilde{\mathbb{E}}_{w\sim\tilde{\zeta}}[ww^\top]$, repeat for $1 \leqslant \ell \leqslant k$:*
   (a) *Choose a uniformly random row i of M.*
   (b) *Let $\hat{C}_\ell$ be the set of points indexed by the largest $\frac{n}{k}$ entries in the ith row of M.*
   (c) *Remove the rows and columns with indices in $\hat{C}_\ell$.*

Analysis of the Algorithm. We first show that the sample X inherits the relevant properties of the distributions. Towards this, we make the following definition.

Definition 2.3.1 ("Good" Sample)**.** *A sample $X \subseteq \mathcal{R}^d$ of size n is said to be a good sample from a Δ-separated mixture of $\mathcal{D}(\mu(r), \Sigma(r))$ for $r \leqslant k$ if there exists a partition $X = C_1 \cup C_2 \cup \cdots C_k \subseteq \mathcal{R}^d$ with the corresponding empirical means and covariances $\hat{\mu}(1), \hat{\Sigma}(1), \ldots, \hat{\mu}(k), \hat{\Sigma}(k)$ such that for all $r \in [k]$ and $s = \log(\kappa)$poly(k/η) ,*

1. *Empirical mean: $\langle \hat{\mu}(r) - \mu(r), v\rangle^2 \leqslant 0.1 \cdot v^\top \Sigma(r) v$*
2. *Empirical covariance: $\left(1 - \frac{1}{2^{2s}}\right)\Sigma(r) \preceq \hat{\Sigma}(r) \preceq \left(1 + \frac{1}{2^{2s}}\right)\Sigma(r)$.*
3. *Certifiable Anti-concentration: For all $\tau \geqslant$ poly(η/Ck),*

$$\left|\frac{v}{2s}\right. \left\{ \frac{k^2}{n^2} \sum_{i\neq j \in C_r} q^2_{\tau,\hat{\Sigma}(r)} (x_i - x_j, v) \leqslant 10C\tau \left(v^\top \hat{\Sigma}(r) v\right)_2^{2s} \right\}.$$

$$\left|\frac{v}{2s}\right. \left\{ \frac{k}{n} \sum_{i_1,i_2\in C_r, j_1,j_2 \in C_{r'}} q^2_{\tau,\hat{\Sigma}(r)} (x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v) \leqslant 10C\tau \left(v^\top (\hat{\Sigma}(r) + \hat{\Sigma}(r') v\right)_2^{2s} \right\}.$$

4. *Certifiable Hypercontractivity: For every $j \leqslant s$,*

$$\left|\frac{Q}{2s}\right. \left\{ \frac{k^2}{n^2} \sum_{i\neq\ell\in C_r} \left(Q(x_i - x_\ell) - \frac{k^2}{n^2} \sum_{i\neq\ell\in C_r} Q(x_i - x_\ell) \right)^{2j} \right.$$
$$\left. \leqslant (2Cj)^{2j} \left(\frac{k^2}{n^2} \sum_{i\neq\ell\in C_r} \left(Q(x_i - x_\ell) - \frac{k^2}{n^2} \sum_{i\neq\ell\in C_r} Q(x_i - x_\ell) \right) \right)^j \right\}.$$

5. *Certifiable Bounded-Variance:*

$$\left|\frac{Q}{2}\right. \left\{ \frac{k^2}{n^2} \sum_{i \neq \ell \in C_r} \left(Q(x_i - x_\ell) - \frac{k^2}{n^2} \sum_{i \neq \ell \in C_r} Q(x_i - x_\ell) \right)^2 \leqslant C \left\| \Sigma(r)^{1/2} Q \Sigma(r)^{1/2} \right\|_F^2 \right\} .$$

Via standard concentration arguments, it is straightforward (See Section 2.10 of Appendix) to verify that a large enough sample X from a Δ-separated mixture of reasonable distributions is a good.

Lemma 2.3.2 (Typical samples are good)**.** *Let X be a sample of size n from a equi-weighted Δ-separated mixture $\mathcal{D}(\mu(r), \Sigma(r))$ for $r \leqslant k$. Then, for $n_0 = \Omega\left((poly(k/\eta)d)^{poly(k/\eta)}\, k \log k\right)$ and any $n \geqslant n_0$, X is good with probability at least $1 - 1/d$. Further, the the uniform distribution on $C_1, C_2, \ldots, C_k$ are pairwise $\Delta/2$-separated.*

As in the spherical case [FKP+19], the heart of the analysis involves showing that $\tilde{\mathbb{E}}_{\tilde{\zeta}}[ww^\top]$ is indeed approximately block diagonal whenever $\tilde{\zeta}$ satisfies $\mathcal{A}$. This follows immediately from the following lemma that shows that that there's a low-degree SoS proof that shows that the subset indicated by w cannot simultaneously have large intersections with two distinct clusters $C_r, C_{r'}$.

Lemma 2.3.3 (Simultaneous Intersection Bounds from Separation)**.** *Let X be a good sample of size n from a Δ-separated, equi-weighted mixture of affine transforms of a reasonable distribution $\mathcal{D}$ with true clusters $C_1, C_2, \ldots, C_k$. For all $r \in [k]$, let $w(C_r)$ denote the linear polynomial $\frac{k}{n} \sum_{i \in C_r} w_i$. Then, for every $r \neq r'$ and $\delta > 0$,*

$$\mathcal{A} \left|\frac{w}{O(\log \kappa / \delta^4)}\right. \left\{ w(C_r) w(C_{r'}) \leqslant O(\delta^{1/3}) \right\},$$

where $\kappa = \sup_{v \in \mathbb{R}^d} \max_{i,j} \frac{v^\top \Sigma(i) v}{v^\top \Sigma(j) v}$.

For the special case of $k = 2$, we obtain the following improved version with no dependence on κ in the degree.

Lemma 2.3.4 (Simultaneous Intersection Bounds from Separation, Two Components)**.** *Let $X = C_1 \cup C_2$ be a good sample with true clusters C_1, C_2 of size $n/2$ from a Δ-separated, equi-weighted mixture of affine transforms of a reasonable distribution $\mathcal{D}$. Let $w(C_r)$ denote the linear polynomial $\frac{k}{n} \sum_{i \in C_r} w_i$ for every $r \leqslant 2$. Then, for any $\delta > 0$,*

$$\mathcal{A} \left|\frac{w}{O(1/\delta^4)}\right. \left\{ w(C_1) w(C_2) \leqslant O(\delta^{1/3}) \right\} .$$

It is easy to finish the analysis of the algorithm given Lemma 2.3.3.

Proof of Theorem 56. **Enforcing Constraints.** First, we argue that the number of constraints in the SDP we need to solve to find $\tilde{\zeta}$ in Step 1 above is $d^{O(\log(\kappa)(1/\delta)^4)}$. For this, it is enough to show that the number of polynomial inequalities needed to enforce $\mathcal{A}$ is appropriately bounded. $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ encode $O(d^2)$ inequalities by direct inspection. $\mathcal{A}_4, \mathcal{A}_5$ superficially seem to encode infinitely many constraints. However, by applying the quantifier alternation technique that only requires SoS certifiability, (first used in [KSS18, HL18], see Page 131 of [FKP+19] for an exposition), to compress such constraints by leveraging low-degree SoS proofs allows us to encode them into $d^{O(1/\delta^4)}$ polynomial inequalities.

Minimizing Norm. Observe that $\left\|\tilde{\mathbb{E}}[w]\right\|_2$ is a convex function in $\tilde{\mathbb{E}}[w]$ and thus, a pseudo-distribution minimizing $\left\|\tilde{\mathbb{E}}[w]\right\|_2$ consistent with $\mathcal{A}$ can be found in time $n^{O(\log(\kappa)/\delta^4)}$ if it exists using the ellipsoid method. The rounding itself is easily seen to take at most $O(n^2)$ time. This completes the analysis of the running time.

Feasibility of the SDP. In the remaining part of the analysis, we condition on the event that the input X is a good sample. We show that the SDP for computing the pseudo-distribution in Step 1 of the algorithm is feasible. We exhibit a feasible solution by describing a natural setting of the indeterminates in our constraint program. Let ζ be the uniform distribution (thus, also a pseudo-distribution of degree ∞) on $\mathbf{1}(C_r)$, for all $r \in k$. That is, ζ is uniformly distributed on the true clusters. Lemma 2.3.2 implies that setting $w = \mathbf{1}(C_r)$ satisfies all the constraints in $\mathcal{A}$. Thus, $\tilde{\zeta}$ is indeed a feasible for the SDP. Observe further that for every i, $\tilde{\mathbb{E}}_{\zeta}[w_i] = 1/k$.

Analysis of the SDP Solution. Now, let $\tilde{\zeta}$ be the pseudo-distribution computed in Step 1 of the algorithm. First, observe that by Cauchy-Schwarz inequality, $\|\tilde{\mathbb{E}}_{\tilde{\zeta}}[w]\|_2^2 = \sum_{i \leqslant n} \tilde{\mathbb{E}}_{\tilde{\zeta}}[w_i]^2 \geqslant \frac{1}{n}\left(\sum_{i \leqslant n} \tilde{\mathbb{E}}_{\tilde{\zeta}}[w_i]\right)^2 = \frac{n}{k^2}$ where we used that $\mathcal{A} \vdash \left\{\frac{k}{n}\sum_{i=1}^n w_i = 1\right\}$. On the other hand, we exhibited a feasible pseudo-distribution ζ above with $\left\|\tilde{\mathbb{E}}_{\zeta}[w]\right\|_2^2 = \frac{n}{k^2}$. Together, we obtain that the output $\tilde{\zeta}$ obtained by solving the SDP relaxation must satisfy $\|\tilde{\mathbb{E}}_{\tilde{\zeta}}[w]\|_2^2 = \frac{n}{k^2}$. Observe that this is equivalent to $\tilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] = 1/k$ for every $i \leqslant n$. Thus, we can assume in the following that $\tilde{\mathbb{E}}_{\tilde{\zeta}}[w_i] = 1/k$ for all i. Our analysis is similar to the proofs of Lemmas 4.21 and Lemma 4.23 in [FKP+19].

Let $M = \tilde{\mathbb{E}}[ww^\top]$. Let's understand the entries of M more carefully. First, since $\tilde{\mathbb{E}}[w_i w_j] = \tilde{\mathbb{E}}[w_i^2 w_j^2] \geqslant 0$, $M(i,j)$ is non-negative. The diagonals of M are $\tilde{\mathbb{E}}[w_i^2] = \tilde{\mathbb{E}}[w_i] = 1/k$. By the Cauchy-Schwarz inequality for pseudo-distributions (Fact 3.2.14), $M(i,j) = \tilde{\mathbb{E}}[w_i w_j] \leqslant$

$\sqrt{\tilde{\mathbb{E}}[w_i^2]}\sqrt{\tilde{\mathbb{E}}[w_j^2]} \leqslant 1/k$. Thus, the entries of M are between 0 and $1/k$. Next, observe that since $\mathcal{A} \vdash \left\{w_i \frac{k}{n} \sum_{j \leqslant n} w_j = w_i\right\}$, taking pseudo-expectations and rearranging yields that for every i, $\mathbf{E}_{j \sim [n]} M(i,j) = \frac{1}{k^2}$.

For $\eta' = \eta^2/k^3$, choose $\delta = \eta'^3/k^3$. Then, applying Lemma 2.3.3 and using Fact 3.2.18, we have that for every r, $\mathbf{E}_{i \in C_r} \mathbf{E}_{j \notin C_{r'}} M(i,j) = \sum_{r' \neq r} \mathbf{E}_{i \in C_r} \mathbf{E}_{j \in C_{r'}} \tilde{\mathbb{E}}[w_i w_j] = \tilde{\mathbb{E}}[w(C_r) w(C_{r'})] \leqslant O(\eta')$.

Fix any cluster C_r. Call an entry of M large if it exceeds η/k^2. Using the above estimates, we obtain that, the fraction of entries in the ith row that exceed η/k^2 is at least $(1-\eta)/k$.

On the other hand, by Markov's inequality applied to the calculation above, we obtain that with probability $1 - 1/k^2$ over the uniformly random choice of $i \in C_r$, $\mathbf{E}_{j \notin C_r} M(i,j) \leqslant O(\eta') = O(\eta^2/k^3)$. Call an $i \in C_r$ for which this condition holds "good".

By Markov's inequality, for each good row, the fraction of $j \notin C_r$ such that $M(i,j) \geqslant \eta/k^2$ is at most η/k. Thus, for any good row in C_r, if we take the indices j corresponding to the largest n/k entries (i,j) in M, then, at most η fraction of such j are not in C_r. Thus, picking uniformly random row in C_r and taking the largest n/k entries in that row gives a subset that intersects with C_r in $(1-\eta)$ fraction of the points.

Thus, each iteration of our rounding algorithm succeeds with probability at least $1 - 1/k^2$. By union bound, all iterations succeed with probability at least $1 - 1/k$. The running time is dominated by the first step and the sample complexity follows from Lemma 2.3.2. □

Proving Lemma 2.3.3 In what follows, we focus attention on proving Lemma 2.3.3. Before describing the analysis, we set some notation/shorthand and simplifying assumptions that we will use throughout this section.

1. First, Lemma 2.3.2 guarantees us that C_r has mean and Covariance close to the true $\mu(r), \Sigma(r)$. We abuse the notation a little bit and use $\mu(r), \Sigma(r)$ to denote the mean and covariance of C_r too. This allows us the luxury of dropping an extra piece of notation and doesn't change the guarantees we obtain.
2. In the following, we will use $\mathcal{D}_r = \mathcal{D}(\mu(r), \Sigma(r))$ to denote the uniform distribution on C_r. We will use $\mathcal{D}_w$ to informally (in the context of non low-degree SoS reasoning) refer to the uniform distribution on the subset indicated by w.

Depending on whether $C_r, C_{r'}$ are mean separated, spectrally separated or separated in relative Frobenius distance, our proof of Lemma 2.3.3 breaks into three natural cases. The key part

of the analysis is dealing with the case of spectral separation which then plugs into the other two cases. So we begin with it.

2.3.2 Intersection Bounds from Spectral Separation

In this subsection, we give a sum-of-squares proof of an upper bound on $w(C_r)w(C_{r'})$ whenever $\mathcal{D}_r, \mathcal{D}_{r'}$ are samples chosen from *spectrally* separated distributions. Note that we do not have any control of the means of $\mathcal{D}_r$, $\mathcal{D}_{r'}$ in this subsection and our arguments must work regardless of the means (or their separation, whether large or small) of $\mathcal{D}_r, \mathcal{D}_{r'}$.

Formally, we will prove the following upper bound on $w(C_r)w(C_{r'})$ where the degree of the sum-of-squares proof grows logarithmically in the spread κ of the mixture.

Lemma 2.3.5 (Intersection Bounds from Spectral Separation)**.** *Let $X = C_1 \cup C_2 \cup \dots C_r$ be a good sample of size n. Suppose there exists a vector v such that $\Delta_{\mathsf{spectral}} v^\top \Sigma(r) v \leqslant v^\top \Sigma(r') v$ for $\Delta_{\mathsf{spectral}} \gg Cs/\delta^2$, where $s \geqslant 1$. Then, $\mathcal{A} \left|\frac{w}{O(\log(\kappa)/\delta^4)}\right. \left\{w(C_r)w(C_{r'}) \leqslant O(\sqrt{\delta})\right\}$ where $\kappa = \max_{i \leqslant k} \frac{v^\top \Sigma(i) v}{v^\top \Sigma(r') v}$.*

Observe that for $k = 2$, $\kappa = 1$ and thus, the lemma above results in a bound of $O(s/\delta^2)$ on the degree of the SoS proof. The proofs of both the statements above follow by using anti-concentration of $\mathcal{D}_r$ and $\mathcal{D}_{r'}$ to first show a lower-bound on the variance of $\Sigma(w)$ in terms of the $v^\top \Sigma(r) v$ and $v^\top \Sigma(r') v$ and then combine it with an upper bound on $v^\top \Sigma(w) v$ using anti-concentration of $\mathcal{D}_w$.

Lemma 2.3.6 (Large Intersection Implies High Variance, Spectral Separation)**.**

$$\mathcal{A} \left|\frac{w, \Sigma(w)}{4s}\right. \left\{ w(C_{r'})w(C_r) \left(v^\top \left(\Sigma(r) + \Sigma(r')\right) v\right)^s \leqslant \left(\frac{2}{\delta^2}\right)^s \left(v^\top \Sigma(w) v\right)^s + C\delta \left(v^\top \left(\Sigma(r) + \Sigma(r')\right) v\right)^s \right\} \tag{2.7}$$

Proof. We know from Lemma 2.3.2 that two-sample-centered points from both C_r and $C_{r'}$ are $2s$-certifiably $(\delta, C\delta)$-anti-concentrated. Using Definition 3.2.28, thus yields:

$$\mathcal{A} \vdash_{4s} \left\{ \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w_{i_1} w_{i_2} w_{j_1} w_{j_2} \langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \rangle^{2s} \right.$$
$$\geqslant \delta^{2s} w(C_r)^2 w(C_{r'})^2 \left(v^\top 2(\Sigma(r) + \Sigma(r'))v^\top \right)^s$$
$$\left. - \delta^{2s} \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w_{i_1} w_{i_2} w_{j_1} w_{j_2} q^2_{\delta, 2(\Sigma(r)+\Sigma(r'))}(x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v) \right\} \quad (2.8)$$

Using that $\mathcal{A} \vdash \{w_{i_1} w_{i_2} w_{j_1} w_{j_2} \leqslant 1\}$ for every i_1, i_2, j_1, j_2 and using $2s$-certifiable $(\delta, C\delta)$-anti-concentration of $x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}$ and invoking Definition 3.2.28, we have:

$$\mathcal{A} \vdash_{4s} \left\{ \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w_{i_1} w_{i_2} w_{j_1} w_{j_2} q^2_{\delta, 2(\Sigma(r)+\Sigma(r'))}(x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v) \right.$$
$$\left. \leqslant \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} q^2_{\delta, 2(\Sigma(r)+\Sigma(r'))}(x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v) \leqslant C\delta \left(v^\top 2(\Sigma(r) + \Sigma(r'))v \right)^s \right\} \quad (2.9)$$

Plugging in the above bound in (2.8) gives:

$$\mathcal{A} \vdash_{4s} \left\{ \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w_{i_1} w_{i_2} w_{j_1} w_{j_2} \langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \rangle^{2s} \right.$$
$$\left. \geqslant \delta^{2s} \left(w(C_r)^2 w(C_{r'})^2 - C\delta \right) \left(v^\top 2(\Sigma(r) + \Sigma(r'))v^\top \right)^s \right\} \quad (2.10)$$

Rearranging thus yields:

$$\mathcal{A} \vdash_{4s} \left\{ \frac{1}{\delta^{2s}} \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w_{i_1} w_{i_2} w_{j_1} w_{j_2} \langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \rangle^{2s} \right.$$
$$+ C\delta \left(v^\top 2(\Sigma(r) + \Sigma(r'))v^\top \right)^s$$
$$\left. \geqslant w(C_r)^2 w(C_{r'})^2 \left(v^\top 2(\Sigma(r) + \Sigma(r'))v^\top \right)^s \right\} \quad (2.11)$$

To finish the proof, we note that:

$$\mathcal{A} \vdash_{4s} \left\{ \left(\frac{4cs}{\delta^2}\right)^s \left(v^\top \Sigma(w) v\right)^s \geqslant \frac{1}{\delta^{2s}} \frac{k^4}{n^4} \sum_{i_1,i_2,j_1,j_2 \in [n]} w_{i_1} w_{i_2} w_{j_1} w_{j_2} \left\langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \right\rangle^{2s} \right.$$
$$\left. \geqslant \frac{1}{\delta^{2s}} \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w_{i_1} w_{i_2} w_{j_1} w_{j_2} \left\langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \right\rangle^{2s} \right\} \quad (2.12)$$

Plugging in the upper bound above in (2.11) and canceling out a copy of 2^s from both sides gives the lemma.

□

Moving forward with our proof plan, we can clearly complete the proof by giving an *upper* bound on $\left(v^\top \Sigma(w) v\right)$ that scales as the variance of the *smaller* variance component (i.e. r above). We make this happen by invoking certifiable anti-concentration again - this time, however, applying it to the w-samples instead of C_r and $C_{r'}$.

Lemma 2.3.7 (Spectral Upper Bound via Anti-Concentration)**.**

$$\mathcal{A} \vdash_{4s} \left\{ \left(w(C_r)^2 - C\delta\right) \left(v^\top \Sigma(w) v^\top\right)^s \leqslant \left(\frac{Cs}{\delta^2}\right)^s \left(v^\top \Sigma(r) v\right)^s \right\} \quad (2.13)$$

Proof. Our constraint system $\mathcal{A}$ allows us to derive that two-sample-centered points indicated by w are $2s$-certifiably $(\delta, C\delta)$-anti-concentrated with witnessing polynomial $p_{\mathcal{D}}$. Using Definition 3.2.28, thus yields:

$$\mathcal{A} \vdash_{4s} \left\{ \delta^{2s} w(C_r)^2 \left(v^\top \Sigma(w) v^\top\right)^s \right.$$
$$\left. \leqslant \frac{k^2}{n^2} \sum_{i,j \in C_r} w_i w_j \left\langle \frac{1}{\sqrt{2}} (x_i - x_j), v \right\rangle^{2s} + \delta^{2s} \frac{k^2}{n^2} \sum_{i \neq j \in C_r} w_i w_j q^2_{\delta, \Sigma(w)} \left(\frac{1}{\sqrt{2}} (x_i - x_j), v \right) \right\} \quad (2.14)$$

Using that $\mathcal{A} \vdash^{\Sigma, w}_{4s} \{w_i w_j \leqslant 1\}$ for every i, j, using that $\mathcal{A}$ derives $2s$-certifiable $(\delta, C\delta)$-anti-concentration of w-samples and invoking Definition 3.2.28, we have:

$$\mathcal{A} \vdash_{4s} \left\{ \frac{k^2}{n^2} \sum_{i \neq j \in C_r} w_i w_j q^2_{\delta,\Sigma(w)} (\frac{1}{\sqrt{2}} (x_i - x_j), v) \leqslant \frac{k^2}{n^2} \sum_{i \neq j \in [n]} w_i w_j q^2_{\delta,\Sigma(w)} \left(\frac{1}{\sqrt{2}} (x_i - x_j), v \right) \right.$$
$$\left. \leqslant C\delta \left(v^\top \Sigma(w) v \right)^s \right\} \tag{2.15}$$

Further, using that $\mathcal{A} \vdash^{\Sigma,w}_{4s} \{w_i w_j \leqslant 1\}$ for all i, j and relying on the certifiable Sub-gaussianity of C_r, we have:

$$\mathcal{A} \vdash_{4s} \left\{ \frac{k^2}{n^2} \sum_{i,j \in C_r} w_i w_j \left\langle \frac{1}{\sqrt{2}} (x_i - x_j), v \right\rangle^{2s} \leqslant \frac{k^2}{n^2} \sum_{i,j \in C_r} \left\langle \frac{1}{\sqrt{2}} (x_i - x_j), v \right\rangle^{2s} \right.$$
$$\left. = (Cs)^s \left(v^\top \Sigma(r) v \right)^s \right\} \tag{2.16}$$

Combining the last two bounds with (2.14) thus yields:

$$\mathcal{A} \vdash_{4s} \left\{ w(C_r)^2 \left(v^\top \Sigma(w) v^\top \right)^s \leqslant \frac{1}{\delta^{2s}} (Cs)^s \left(v^\top \Sigma(r) v \right)^s + C\delta \left(v^\top \Sigma(w) v^\top \right)^s \right\} \tag{2.17}$$

□

Digression: "Real-World" Proof We'd now like to combine the upper and lower bounds on $v^\top \Sigma(w) v$ obtained in the two previous lemmas in order to conclude a bound on the intersection size $w^2(C_r) w^2(C_{r'})$. To aid the intuition, observe that this is easy to do in "usual math" (in contrast to low-degree sum-of-squares proof system). If the reader prefers to skip this digression, they can skip to the paragraph titled *Upper Bounds via SoSizing Conditional Argument*.

Lemma 2.3.8 (Low Intersection Size from Spectral Separation (*not* a low-degree SoS Proof))**.** *Let $v \in \mathcal{R}^d$ be a unit vector such that $\Delta v^\top \Sigma(r) v \leqslant v^\top \Sigma(r') v$ for some $\Delta \gg 2Cs/\delta^3$. Then, $w^3(C_r) w^3(C_{r'}) \leqslant \delta$.*

Proof. We split into two cases: 1) $w^2(C_r) \leqslant \delta$ and 2) $w(C_r)^2 > \delta$. In the first, case $w^3(C_r)w^3(C_{r'})$ is clearly at most δ. So we are done!

In the second case, we invoke Lemma 2.3.6 to write:

$$w(C_{r'})w(C_r)\left(v^\top\left(\Sigma(r)+\Sigma(r')\right)v^\top\right)^s \leqslant \frac{2^s}{\delta^{2s}}\left(v^\top\Sigma(w)v\right)^s + C\delta\left(v^\top\left(\Sigma(r)+\Sigma(r')\right)v^\top\right)^s.$$

Since $(w^2(C_r)-\delta) \geqslant 0$, we can multiply both sides of above by $(w^2(C_r)-\delta)$ without changing the inequality. By Lemma 2.3.7:

$$\left(w(C_r)^2 - C\delta\right)\left(v^\top\Sigma(w)v^\top\right)^s \leqslant \frac{1}{\delta^{2s}}(Cs)^s\left(v^\top\Sigma(r)v\right)^s.$$

Using the above bound, using that $w(C_r)w(C_{r'}) \leqslant 1$ and rearranging, we have:

$$\begin{aligned} w(C_r)^2w(C_{r'})w(C_r)\left(v^\top\left(\Sigma(r)+\Sigma(r')\right)v^\top\right)^s \leqslant (C+1)\delta\left(v^\top\left(\Sigma(r)+\Sigma(r')\right)v^\top\right)^s \\ + \left(\frac{2}{\delta}\right)^s\frac{1}{\delta^{2s}}(Cs)^s\left(v^\top\Sigma(r)v\right)^s. \end{aligned} \tag{2.18}$$

Using the above bound with the spectrally separating direction v, we know that

$$v^\top\left(\Sigma(r)+\Sigma(r')\right)v^\top \geqslant \Delta v^\top\Sigma(r)v.$$

Thus rearranging the above inequality gives:

$$w(C_r)^3w(C_{r'})^3 \leqslant w^3(C_r)w(C_{r'}) \leqslant (C+1)\delta + \left(\frac{2}{\delta^3}\right)^s(Cs)^s\Delta^{-s},$$

which is at most $2C\delta$ whenever $\Delta \gg Cs/\delta^3$ as desired. □

Crucial to the above "real world" argument is the second step where we use the non-negativity of $w(C_r)^2 - \delta$ so as to multiply the starting inequality on both sides with it while preserving the direction of the inequality. This step relies on an "if-then" case analysis which, unfortunately, cannot, in general, be implemented *as is* in low-degree sum-of-squares proof system.

Upper Bounds via SoSizing Conditional Argument In order to implement an argument similar to the one above, within the low-degree SoS system, we will introduce a polynomial $\mathcal{J}$ which approximates the thresholding operation withing SoS. We prove the existence of such a polynomial in Appendix 2.11. This will, however, lose us a $\log(\kappa)$ factor in the SoS degree required (and thus cause an exponential dependence on $\log(\kappa)$ in the running time of our clustering

algorithm).

Lemma 2.3.9 (Polynomial Approximator for Thresholds, See Section 2.11 for a proof)**.** *Let* $1/2 \geqslant \rho \geqslant 0$ *and* $c \in [0,1]$. *There exists a square polynomial* $\mathcal{J}$ *satisfying:*

1. $\mathcal{J}(x) \in [1,2]$ *for all* $x \in [2c,1]$.
2. $\mathcal{J}(x) \leqslant \rho$ *for all* $x \in [0,c]$.
3. $\deg(\mathcal{J}) \leqslant O(\log(1/\rho)/c)$.

Lemma 2.3.10. *For any* $0 < \rho < 1$,

$$\{0 \leqslant w(C_r) \leqslant 1\} \left|\frac{w}{O(\log(1/\rho)/\delta^2)}\right. \{\mathcal{J}(w(C_r))(w(C_r)-\delta) \geqslant -\delta\rho\} \ ,$$

and,

$$\{0 \leqslant w(C_r) \leqslant 1\} \left|\frac{w}{O(\log(1/\rho)/\delta^2)}\right. \{\mathcal{J}(w(C_r))w(C_r) \geqslant (w(C_r)-2\delta)\} \ .$$

Proof. Observe that the conclusion is a polynomial inequality in single variable $w(C_r)$. Thus, it is enough to give *any* proof of $\mathcal{J}(w(C_r))(w(C_r)-\delta) \geqslant -\delta\rho$.

To see why the inequality holds, observe that if $w(C_r) \geqslant \delta$, $\mathcal{J}(w(C_r))(w(C_r)-\delta) \geqslant 0 > -\delta\rho$. On the other hand, if $w(C_r) \leqslant \delta$, then, $\mathcal{J}(w(C_r)) \leqslant \rho$ while $|w(C_r)-\delta| \leqslant \delta$. On the other hand, observe that $\mathcal{J}(w(C_r)(w(C_r)-\delta) \leqslant \mathcal{J}(w(C_r)w(C_r) \leqslant 2w(C_r)$. This completes the proof of the first inequality.

For the second claim, notice that if $w(C_r) < 2\delta$, the inequality trivially holds since $\mathcal{J}(w(C_r)) \geqslant 0$. If on the other hand, $w(C_r) > 2\delta$, then, $\mathcal{J}(w(C_r)) \geqslant 1 \geqslant w(C_r) \geqslant w(C_r)-\delta$. □

We can now implement the above real-world "conditional" argument within SoS using the polynomial $\mathcal{J}$ above. To do this, we will need a rough upper bound on $v^\top \Sigma(w) v$ in terms of $v^\top \Sigma(r) v$ for $r \leqslant k$. We will prove this via another application of certifiable anti-concentration of $\mathcal{D}_w$ - this time, invoked with the slightly different parameter τ.

Lemma 2.3.11 (Rough Spectral Upper bound on $\Sigma(w)$)**.**

$$\mathcal{A} \left|\!- \left\{ \left(v^\top \Sigma(w) v^\top\right)^s \leqslant (2Ck)^{s+1} (Cs)^s \sum_{r \leqslant k} \left(v^\top \Sigma(r) v\right)^s \right\} \right. \tag{2.19}$$

Proof. Our proof is similar to the proof of Lemma 2.3.7 with a key additional step. As in the

proof of Lemma 2.3.7, we start by invoking our constraints to conclude (note that we sum over all samples this time instead of those just in C_r as in the previous lemma:

$$\mathcal{A}\vdash \left\{\tau^{2s}\sum_{r\leqslant k} w'(C_r)^2\left(v^\top\Sigma(w)v^\top\right)^s \right.$$
$$\left.\leqslant \frac{k^2}{n^2}\sum_{r\leqslant k}\sum_{i,j\in C_r} w_iw_j\left\langle\frac{1}{\sqrt{2}}\left(x_i-x_j\right),v\right\rangle^{2s}+\tau^{2s}\frac{k^2}{n^2}\sum_{r\leqslant k}\sum_{i\neq j\in C_r} w_iw_jq^2_{\tau,\Sigma(w)}\left(\frac{1}{\sqrt{2}}\left(x_i-x_j\right),v\right)\right\} \tag{2.20}$$

The second term on the RHS can be upper bounded just as in the proof of Lemma 2.3.7 to yield:

$$\begin{aligned}\mathcal{A}\vdash \left\{\frac{k^2}{n^2}\sum_{r\leqslant k}\sum_{i\neq j\in C_r} w_iw_jq^2_{\tau,\Sigma(w)}\left(\left\langle\frac{1}{\sqrt{2}}\left(x_i-x_j\right),v\right\rangle\right)\right.\\ \leqslant \frac{k^2}{n^2}\sum_{i\neq j\in[n]} w_iw_jq^2_{\tau,\Sigma(w)}\left(\left\langle\frac{1}{\sqrt{2}}\left(x_i-x_j\right),v\right\rangle\right)\\ \left.\leqslant C\tau\left(v^\top\Sigma(w)v\right)^s\right\}\end{aligned} \tag{2.21}$$

The first term can be also be upper bounded - this time in terms of the Covariances of all the k components.

$$\begin{aligned}\mathcal{A}\vdash \left\{\frac{k^2}{n^2}\sum_{r\leqslant k}\sum_{i,j\in C_r} w_iw_j\left\langle\frac{1}{\sqrt{2}}\left(x_i-x_j\right),v\right\rangle^{2s}\right. &\leqslant \sum_{r\leqslant k}\frac{k^2}{n^2}\sum_{i,j\in C_r}\left\langle\frac{1}{\sqrt{2}}\left(x_i-x_j\right),v\right\rangle^{2s}\\ &\left.=(Cs)^s\sum_{r\leqslant k}\left(v^\top\Sigma(r)v\right)^s\right\}\end{aligned} \tag{2.22}$$

We can now combine the two estimates above to yield:

$$\mathcal{A}\vdash \left\{\left(\sum_{r\leqslant k} w(C_r)^2-C\tau\right)\left(v^\top\Sigma(w)v^\top\right)^s\leqslant\frac{1}{\tau^{2s}}(Cs)^s\sum_{r\leqslant k}\left(v^\top\Sigma(r)v\right)^s\right\} \tag{2.23}$$

So far the argument closely follows the proof of Lemma 2.3.7. The key departure we make is with the following simple observation:

$$\mathcal{A} \vdash \left\{ \sum_{r \leqslant k} w(C_r)^2 \geqslant \frac{1}{k} \left(\sum_{r \leqslant k} w(C_r) \right)^2 = \frac{1}{k} \right\}.$$

Thus, as long as $\tau < \frac{1}{2Ck}$, we can derive:

$$\mathcal{A} \vdash \left\{ \left(v^\top \Sigma(w) v \right)^s \leqslant k^{s+1} (Cs)^s \sum_{r \leqslant k} \left(v^\top \Sigma(r) v \right)^s \right\} \tag{2.24}$$

This is the "rough" upper bound on $\Sigma(w)$ we were after. □

We can use the above lemma to get an "upgraded" version of Lemma 2.3.7.

Lemma 2.3.12 (Upper Bound on Variance of $\mathcal{D}_w$)**.** *Let $\lambda_{\max}(v) \|v\|_2^2$ be the maximum of $v^\top \Sigma(r) v$ over all $r \leqslant k$. Then,*

$$\mathcal{A} \vdash \left\{ (\mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho) \left(v^\top \Sigma(w) v \right)^s \leqslant 2 \frac{1}{\delta^{2s}} (Cs)^s \left(v^\top \Sigma(r) v \right)^s \right.$$
$$\left. + \delta\rho s^{2s} (Cs)^s k \lambda_{\max}(v)^s \|v\|_2^{2s} \right\}. \tag{2.25}$$

Proof. From Lemma 2.3.11, we have:

$$\mathcal{A} \vdash \left\{ \left(v^\top \Sigma(w) v \right)^s \leqslant (s)^{s+1} (Cs)^s \sum_{r \leqslant k} \left(v^\top \Sigma(r) v \right)^s \right\} \tag{2.26}$$

Then, the above bound implies:

$$\mathcal{A} \vdash \left\{ \left(v^\top \Sigma(w) v \right)^s \leqslant (s^{s+1} (Cs)^s k \lambda_{\max}(v)^s \right\}. \tag{2.27}$$

From Lemma 2.3.10, we have: $\mathcal{A} \vdash \{\mathcal{J}(w(C_r)) \leqslant 2\}$. Thus, using Lemma 2.3.7 and applying (2.27) on the RHS, we can conclude:

$$\mathcal{A} \vdash \Big\{ (\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)\left(v^\top \Sigma(w) v\right)^s \leqslant \delta\rho \left(v^\top \Sigma(w) v\right)^s + 2\left(\frac{Cs}{\delta^2}\right)^s \left(v^\top \Sigma(r) v\right)^s$$
$$\leqslant \delta\rho s^{2s} (Cs)^s k\lambda_{\max}(v)^s \|v\|_2^{2s} + 2\left(\frac{Cs}{\delta^2}\right)^s \left(v^\top \Sigma(r) v\right)^s \Big\}.$$

□

We are now ready to complete the proof of Lemma 2.3.5.

Proof of Lemma 2.3.5. Observe that $\mathcal{A} \vdash \{0 \leqslant w(C_r) \leqslant 1\}$. Thus,

$$\mathcal{A} \vdash \{\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho \geqslant 0\}\ . \tag{2.28}$$

From Lemma 2.3.6, we have:

$$\mathcal{A} \vdash \left\{ w(C_{r'})w(C_r)\left(v^\top\left(\Sigma(r)+\Sigma(r')\right)v\right)^s \leqslant \frac{2^s}{\delta^{2s}}\left(v^\top\Sigma(w)v\right)^s + C\delta\left(v^\top\left(\Sigma(r)+\Sigma(r')\right)v\right)^s \right\}.$$

Using (2.28) along with (3.5) with $\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho$ gives:

$$\mathcal{A} \vdash \Big\{ \left(\mathcal{J}(w(C_r))(w(C_r)-\delta)\right) w(C_{r'})w(C_r)\left(v^\top\left(\Sigma(r)+\Sigma(r')\right)v\right)^s$$
$$\leqslant \delta\rho\left(v^\top\left(\Sigma(r)+\Sigma(r')\right)v\right)^s + \left(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho\right)\frac{2^s}{\delta^{2s}}\left(v^\top\Sigma(w)v\right)^s$$
$$+ \left(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho\right)\frac{1}{\delta^{2s}}(Cs)^s\left(v^\top\Sigma(r)v\right)^s$$
$$+ \left(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho\right)2C\delta\left(v^\top\Sigma(r')v\right)^s \Big\}. \tag{2.29}$$

Rearranging yields:

$$\mathcal{A} \vdash \Bigg\{ \mathcal{J}(w(C_r))(w(C_r)w(C_{r'})w(C_r) \left(v^\top \left(\Sigma(r) + \Sigma(r')\right) v\right)^s \\ \leqslant 2\delta\rho \left(v^\top \left(\Sigma(r) + \Sigma(r')\right) v\right)^s + \left(\mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho\right) \frac{2^s}{\delta^{2s}} \left(v^\top \Sigma(w) v\right)^s \\ + \left(\mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho\right) \frac{1}{\delta^{2s}} (Cs)^s \left(v^\top \Sigma(r) v\right)^s \\ + \left(\mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho\right) 2C\delta \left(v^\top \Sigma(r') v\right)^s \Bigg\}. \quad (2.30)$$

Using Lemma 2.3.10, we have that $\mathcal{J}(w(C_r))w(C_r) \geqslant (w(C_r) - \delta)$. Multiplying the above inequality (using (3.5)) by the SoS (and thus non-negative) polynomial $w(C_r)w(C_{r'}) \left(v^\top \left(\Sigma(r) + \Sigma(r')\right) v^\top\right)^s$ yields:

$$\mathcal{A} \vdash \Bigg\{ \mathcal{J}(w(C_r))w(C_{r'})w^2(C_r) \left(v^\top \left(\Sigma(r) + \Sigma(r')\right) v\right)^s \\ \geqslant (w(C_r) - \delta)w(C_{r'})w(C_r) \left(v^\top \left(\Sigma(r) + \Sigma(r')\right) v\right)^s \Bigg\}.$$

Thus, the LHS above is lower bounded by $(w(C_r) - \delta)w(C_{r'})w(C_r) \left(v^\top \left(\Sigma(r) + \Sigma(r')\right) v\right)^s$. Let's analyze the terms in the RHS one by one. The first term can be upper bounded directly by applying Lemma 2.3.12. The remaining two terms in the RHS can be upper bounded by relying on:

$$\mathcal{A} \vdash \left\{ \mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho \leqslant 2 \right\}.$$

Thus, using the above bounds we have:

$$\mathcal{A} \vdash \Bigg\{ w(C_r)^2 w(C_{r'}) \left(v^\top \left(\Sigma(r) + \Sigma(r')\right) v\right)^s \leqslant 3\delta \left(v^\top \left(\Sigma(r) + \Sigma(r')\right) v\right)^s \\ + \frac{2}{\delta^{2s}} (Cs)^s \left(v^\top \Sigma(r) v\right)^s + \delta\rho s^{2s} (Cs)^s k\lambda_{\max}(v)^s \|v\|_2^{2s} \\ + \frac{2}{\delta^{2s}} (Cs)^s \left(v^\top \Sigma(r) v\right)^s + 4C\delta \left(v^\top \Sigma(r') v\right)^s \Bigg\} \quad (2.31)$$

Next, observe that since $C_r, C_{r'}$ are spectrally separated and $0 \leqslant v^\top \Sigma(r) v < v^\top \Sigma(r') v$. Thus, $v^\top \Sigma(r') v = \lambda_{r'}(v)\|v\|_2^2 > 0$.

We now set $\rho \leqslant s^{-2s}\,(Cs)^{-s}\,k^{-1}\lambda_{\max}(v)^{-s}\lambda_{r'}(v)^s \leqslant s^{-O(s)}k^{-1}\kappa^{-s}$ and use that $\Delta \geqslant Cs/\delta^2$ to conclude:

$$\mathcal{A} \vdash_{O(\log(\kappa)/\delta^4)} \left\{ w(C_r)^2 w^2(C_{r'}) \leqslant w(C_r)^2 w(C_{r'}) \leqslant O(\delta) \right\} \tag{2.32}$$

Applying Lemma 2.8.2 completes the proof. □

Simpler Proof for Two Components

As an aside, we consider the case where the input mixture only has two components. For this special case where $k = 2$, we show that can bypass the use of the threshold approximator above to get a simpler proof.

Special case of $k = 2$. We proceed exactly as in the proof of Lemma 2.3.5 until equation (2.31) where we invoke the uniform eigenvalue upper bound. Instead of using the uniform eigenvalue upper bound on $\Sigma(w)$, we use Lemma 2.3.11, setting $t = s(1/2Ck) \leqslant 1/k^{\Theta(1)} = O(1)$ for $k = 2$ to derive:

$$\mathcal{A} \vdash_{4t} \left\{ \left(v^\top \Sigma(w) v^\top\right)^t \leqslant 2^{O(t)} \left(\left(v^\top \Sigma(1) v\right)^t + \left(v^\top \Sigma(2) v\right)^t \right) \right\} \tag{2.33}$$

With this sharper upper bound, we can complete the proof as in Lemma 2.3.5 by setting $\rho = 2^{-\Theta(s)}k^{-1}\delta$ instead of $1/\text{poly}(\kappa)$. Since $\log(1/\tau) = \Theta(s)/\delta = \text{poly}(1/\delta)$, the degree of the SoS proof does not grow with κ anymore. . Since $\log(1/\rho) = \Theta(s)/\delta = \text{poly}(1/\delta)$, the degree of the SoS proof does not grow with the spread parameters κ anymore.

□

Remark 58 (Difficulty in extending the simpler argument to $k > 2$)**.** For mixtures with larger number of components, the upper bound from Lemma 2.3.11 is not enough. This is because the upper bound in the Lemma 2.3.11 scales with the largest variance of any of the k component distributions which could be a lot larger than the variance of $\mathcal{D}_r$ and $\mathcal{D}_{r'}$ in the direction v.

2.3.3 Intersection Bounds from Mean Separation

In this section, we give a low-degree sum-of-squares proof that if $C_r, C_{r'}$ are mean separated then $w(C_r)w(C_{r'})$ must be small. Formally, we will show:

Lemma 2.3.13 (Intersection Bounds from Mean Separation)**.** *Let $X = C_1 \cup C_2 \cup \ldots C_r$ be a good sample of size n. Suppose there exists a vector $v \in \mathcal{R}^d$ such that $\langle \mu_r - \mu_{r'}, v\rangle_2^2 \geqslant \Delta_m^2 v^\top (\Sigma(r) + \Sigma(r'))\, v$.*

Then, whenever $\Delta_m \gg Cs/\delta$,

$$\mathcal{A} \left|\frac{w}{O(1/\delta^4 \log(\kappa))}\right. \left\{w(C_r)w(C_{r'}) \leqslant O(\sqrt{\delta})\right\} .$$

As in the previous subsection, we can get a sum-of-squares proof of absolute constant degree for the special case of $k = 2$ components.

Lemma 2.3.14 (Intersection Bounds from Mean Separation)**.** *Let $X = C_1 \cup C_2$ be a good sample of size n. Suppose there exists a vector $v \in \mathcal{R}^d$ such that $\langle \mu(1) - \mu(2), v\rangle_2^2 \geqslant \Delta_m^2 v^\top (\Sigma(1) + \Sigma(2))\, v$.*

Then, whenever $\Delta_m \gg \Theta(1)$,

$$\mathcal{A} \left|\frac{w}{O(1/\delta^4)}\right. \left\{w(C_1)w(C_2) \leqslant O(\sqrt{\delta})\right\} .$$

We will need the following technical fact in our proof.

Lemma 2.3.15 (Lower Bounding Sums)**.** *Let A, B, C, D be scalar-valued indeterminates. Then, for any $\tau > 0$,*

$$\{0 \leqslant A, B \leqslant A + B \leqslant 1\} \cup \{0 \leqslant C, D\} \cup \{C + D \geqslant \tau\} \left|\frac{A,B,C}{2}\right. \{AC + BD \geqslant \tau AB\} .$$

Proof. We have:

$$\{0 \leqslant A, B \leqslant A + B \leqslant 1\} \cup \{0 \leqslant C, D\} \cup \{C + D \geqslant F\} \vdash \Big\{ AC+BD \geqslant (A+B)(AC+BD) \geqslant A^2C + AB(C + D) + B^2D \geqslant AB(C + D) \geqslant \tau AB \Big\} \quad (2.34)$$

□

Proof of Lemma 2.3.13. Let v be the direction in which the means of C_r and $C_{r'}$ are separated. Then, we have:

$$\langle \mu_r - \mu_{r'}, v \rangle_2^{2s} \geqslant \Delta_m^{2s} \left(v^\top \left(\Sigma(r) + \Sigma(r') \right) v \right)^s . \tag{2.35}$$

Assume, WLOG, that $v^\top \Sigma(r) v \leqslant v^\top \Sigma(r') v$.

Applying Lemma 2.3.15 with $A = w(C_r)$, $B = w(C_{r'})$, $C = \langle \mu_r - \mu(w), v \rangle^{2s}$ and $D = \langle \mu_{r'} - \mu(w), v \rangle^{2s}$ along with the SoS Almost Triangle Inequality (Fact 2.2.8) and certifiable Subgaussianity constraints ($\mathcal{A}_5$) yields:

$$\begin{aligned}
\mathcal{A} \left|\frac{\mu,w}{4s}\right. \Bigg\{ & (Cs)^s \left(v^\top \Sigma(w) v \right)^s \geqslant \frac{1}{n} \sum_{i \leqslant n} w_i \langle x_i - \mu(w), v \rangle^{2s} \geqslant \frac{1}{n} \sum_{i \in C_r \cup C_{r'}} w_i \langle x_i - \mu(w), v \rangle^{2s} \\
& \geqslant \frac{1}{2^s} \left(w(C_r) \langle \mu_r - \mu(w), v \rangle^{2s} - \frac{1}{n} \sum_{i \in C_r} w_i \langle x_i - \mu_r, v \rangle^{2s} \right) \\
& + \frac{1}{2^s} \left(w(C_{r'}) w_i \langle \mu_{r'} - \mu(w), v \rangle^{2s} - \frac{1}{n} \sum_{i \in C_{r'}} w_i \langle x_i - \mu_{r'}, v \rangle^{2s} \right) \\
& \geqslant \frac{1}{2^s} \left(w(C_r) \langle \mu_r - \mu(w), v \rangle^{2s} + w(C_{r'}) \langle \mu_{r'} - \mu(w), v \rangle^{2s} \right) - \frac{1}{2^s} \left(v^\top \Sigma(r) v \right)^s - \frac{1}{2^s} \left(v^\top \Sigma(r') v \right)^s \\
& \geqslant \frac{1}{2^{s+1}} \left(w(C_r) w(C_{r'}) \left(\langle \mu_r - \mu(w), v \rangle^{2s} + \langle \mu_{r'} - \mu(w), v \rangle^{2s} \right) \right) - \frac{1}{2^s} \left(v^\top \Sigma(r) v \right)^s - \frac{1}{2^s} \left(v^\top \Sigma(r') v \right)^s \\
& \geqslant \left(\frac{\Delta_m}{4} \right)^{2s} \left(w(C_r) w(C_{r'}) \left(\left(v^\top \Sigma(r) v \right)^s + \left(v^\top \Sigma(r') v \right)^s \right) \right) - \frac{1}{2^s} \left(v^\top \Sigma(r) v \right)^s - \frac{1}{2^s} \left(v^\top \Sigma(r') v \right)^s \Bigg\},
\end{aligned}$$

where the last inequality follows from (2.35). Rearranging the chain of reasoning above thus yields:

$$\begin{aligned}
\mathcal{A} \left|\frac{}{4s}\right. \Bigg\{ & 2^s \left((Cs)^s \left(v^\top \Sigma(w) v \right)^s + \left(v^\top \Sigma(r) v \right)^s + \left(v^\top \Sigma(r') v \right)^s \right) \\
& \geqslant \Delta_m^{2s} w(C_r) w(C_{r'}) \left(\left(v^\top \Sigma(r) v \right)^s + \left(v^\top \Sigma(r') v \right)^s \right) \Bigg\}.
\end{aligned} \tag{2.36}$$

Lemma 2.3.10 shows a low-degree SoS proof of non-negativity of $\mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho$ in variables w:

$$\mathcal{A} \left|\frac{w}{O(\log(1/\rho)/\delta^2)}\right. \left\{ \mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho \geqslant 0 \right\} .$$

Thus, we can multiply (2.36) by $(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)$ throughout to obtain:

$$\mathcal{A}\left|\overset{\mu,w}{\vdash}\right.\Bigg\{(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)\left((2Cs)^s\left(v^\top\Sigma(w)v\right)^s+2^s\left(v^\top\Sigma(r)v\right)^s+2^s\left(v^\top\Sigma(r')v\right)^s\right)$$
$$\geqslant\Delta_m^{2s}\left(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho\right)\left(w(C_r)w(C_{r'})\right)\left(\left(v^\top\Sigma(r)v\right)^s+\left(v^\top\Sigma(r')v\right)^s\right)\Bigg\},\quad(2.37)$$

where the degree of the inequality above is $\ell=O(\log(1/\rho)s/\delta^2)$.

Applying Lemma 2.3.12 for the first term on the LHS and using that

$$\mathcal{A}\left|\vdash\right.\{(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)\leqslant 2\}$$

and rearranging the above inequality gives:

$$\mathcal{A}\left|\overset{\mu,w}{\underset{\ell}{\vdash}}\right.\Bigg\{(2Cs)^s\left(\delta\rho s^{2s}\left(Cs\right)^s k\lambda_{\max}(v)^s+\frac{2}{\delta^{2s}}\left(Cs\right)^s\left(v^\top\Sigma(r)v\right)^s\right)+2^s\left(v^\top\Sigma(r)v\right)^s+2^s\left(v^\top\Sigma(r')v\right)^s$$
$$+2\Delta_m^{2s}\delta\left(\left(v^\top\Sigma(r)v\right)^s+\left(v^\top\Sigma(r')v\right)^s\right)$$
$$\geqslant\Delta_m^{2s}\mathcal{J}(w(C_r))\left(w^2(C_r)w(C_{r'})\right)\left(\left(v^\top\Sigma(r)v\right)^s+\left(v^\top\Sigma(r')v\right)^s\right)\Bigg\}.\quad(2.38)$$

Using Lemma 2.3.10, we also have:

$$\mathcal{A}\left|\underset{O(\log(1/\rho)/\delta^2)}{\overset{w}{\vdash}}\right.\{\mathcal{J}(w(C_r))w(C_r)\geqslant(w(C_r)-\delta)\}\ .$$

Using this bound on the RHS of (2.38) and rearranging yields:

$$\mathcal{A}\left|\overset{\mu,w}{\underset{\ell}{\vdash}}\right.\Bigg\{(2Cs)^s\left(\delta\rho\lambda_{\max}^s+2\frac{1}{\delta^{2s}}\left(Cs\right)^s\left(v^\top\Sigma(r)v\right)^s\right)+2^s\left(v^\top\Sigma(r)v\right)^s+2^s\left(v^\top\Sigma(r')v\right)^s$$
$$+2\Delta_m^{2s}\delta\left(\left(v^\top\Sigma(r)v\right)^s+\left(v^\top\Sigma(r')v\right)^s\right)$$
$$\geqslant\Delta_m^{2s}\left(w^2(C_r)w(C_{r'})\right)\left(\left(v^\top\Sigma(r)v\right)^s+\left(v^\top\Sigma(r')v\right)^s\right)\Bigg\}.\quad(2.39)$$

Dividing throughout by $\Delta_m^{2s}\left(\left(v^\top\Sigma(r)v\right)^s+\left(v^\top\Sigma(r')v\right)^s\right)$ and recalling that $v^\top\Sigma(r)v\leqslant v^\top\Sigma(r')v$ yields:

$$\mathcal{A} \left|\frac{\mu,w}{\ell}\right. \left\{ \left(w^2(C_r)w(C_{r'})\right) \leqslant \Delta_m^{-2s}(2Cs)^s \left(\delta\rho\kappa^s\right) + 2\left(\frac{C\sqrt{s}}{\Delta_m\delta}\right)^{2s} + 2\delta \right\}. \tag{2.40}$$

Thus, choosing $\rho = \kappa^{-s}$ and using that $\Delta_m \gg Cs/\delta$ and $s = 1/\delta^2$ ensures that we obtain:

$$\mathcal{A} \left|\frac{w}{O(\log(1/\kappa)/\delta^4)}\right. \left\{ \left(w^2(C_r)w^2(C_{r'})\right) \leqslant \left(w^2(C_r)w(C_{r'})\right) \leqslant O(\delta) \right\}. \tag{2.41}$$

□

Improved SoS Degree Bounds for $k = 2$

Proof of Lemma 2.3.14. We proceed exactly as in the above proof of Lemma 2.3.13 up until (2.38) where we invoke a rough eigenvalue upper bound on $\Sigma(w)$. We replace this bound by the sharper bound for the $k = 2$ case given by Lemma 2.3.11 analogous to the case of spectral separation and get to choose $\log(1/\rho) = O(1/\delta^2)$. We can then finish the argument as in the proof of Lemma 2.3.13 above.

□

2.3.4 Intersection Bounds from Relative Frobenius Separation of Covariances

In this section, we show that if C_r and $C_{r'}$ are generated by Gaussians with covariances that are separated in relative Frobenius distance, then $w(C_r)w(C_{r'}) = O(\delta)$.

Recall that in this case, $\Sigma(r)$ and $\Sigma(r')$ have the same range (as linear operators). Thus, WLOG, we can assume them to be full rank.

Lemma 2.3.16 (Intersection Bounds from Relative Frobenius Separation)**.** *Suppose*

$$\left\|\Sigma(r')^{-1/2}\Sigma(r)\Sigma(r')^{-1/2} - I\right\|_F^2 \geqslant \Delta_{cov}^2 \left(\left\|\Sigma(r')^{-1/2}\Sigma(r)^{1/2}\right\|_{op}^4\right)$$

for $\Delta_{cov} \gg Cs/\delta^2$, where $s = O(1/\delta^2)$. Then,

$$\mathcal{A} \left|\frac{w}{O(\log(\kappa)/\delta^4)}\right. \left\{w(C_r)w(C_{r'}) \leqslant O(\delta^{1/3})\right\} .$$

As in the previous two subsections, we can get a constant degree sum-of-squares proof for the special case of $k = 2$ components.

Lemma 2.3.17 (Intersection Bounds from Relative Frobenius Separation, Two Components)**.** *Suppose $\left\|\Sigma(2)^{-1/2}\Sigma(1)\Sigma(2)^{-1/2} - I\right\|_F^2 \geqslant \Delta_{cov}^2 \left(\left\|\Sigma(2)^{-1/2}\Sigma(1)^{1/2}\right\|_{op}^4\right)$. Then,*

$$\mathcal{A} \left|\frac{w}{O(1/\delta^4)}\right. \left\{w(C_1)w(C_2) \leqslant O(\delta^{1/3})\right\} .$$

Let Q be a $d \times d$ matrix-valued indeterminate. In the following, we write $Q(z)$ for $z^\top Q z$ (the quadratic form associated with Q). We also use the notation $\mathbb{E}_w[Q] = \frac{k}{n}\sum_{i,j} w_i w_j Q(x_i - x_j)$ - the polynomial computing the mean of Q with respect to the subsample indicated by w. We also write $\mathbb{E}_{C_r}[Q] = \frac{k}{n}\sum_{i,j\in C_r} Q(x_i - x_j)$ and $\mathbb{E}_{C_{r'}}[Q] = \frac{k}{n}\sum_{i,j\in C_r} Q(x_i - x_j)$. We note that for any distribution $\mathcal{D}$ with covariance Σ, $\mathbb{E}_{x,y\sim\mathcal{D}}\left[(x-y)^\top Q(x-y)\right] = 2\operatorname{tr}(\Sigma Q)$.

Proof of Lemma 2.3.16 We can now proceed with the proof of Lemma 2.3.16. As in the previous two subsections, the idea is to show a lower bound on the variance of some polynomial in terms of the intersection size $w(C_r)w(C_{r'})$ and couple it with an upper bound on the variance that follows from certifiable hypercontractivity to obtain an upper bound on $w(C_r)w(C_{r'})$.

Observe that the relative Frobenius separation condition is invariant under linear transformations. Thus, we can assume that $\Sigma(r') = I$ WLOG. This simplifies notation quite a bit in this argument. With this simplification, we now have: $\|\Sigma(r) - I\|_F^2 \geqslant \Delta_{cov}^2 \|\Sigma(r)\|_{op}^2$. Further, the covariance of C_r is now $\Sigma(r')^{-1/2}\Sigma(r)\Sigma(r')^{-1/2}$ and that of $C_{r'}$ is now I after this linear transformation. It's also easy to verify that $\frac{k^2}{n^2}\sum_{i,j} w_i w_j \Sigma(r')^{-1/2}(x_i - x_j)(x_i - x_j)^\top \Sigma(r')^{-1/2} = 2\Sigma(r')^{-1/2}\Sigma(w)\Sigma(r')^{-1/2}$.

In order to simplify notation, we will simply treat $\Sigma(r') = I$ and $\Sigma(r) \to \Sigma(r')^{-1/2}\Sigma(r)\Sigma(r')^{-1/2}$ in the analysis below. We start with the lower-bound first.

Lemma 2.3.18 (Large Intersection Implies High Variance)**.** *Let* $Q = \Sigma(r')^{-1/2}\Sigma(r)\Sigma(r')^{-1/2} - I$.

$$\mathcal{A}^{\frac{w}{4}}\left\{4\mathbf{E}_w(Q-\mathbf{E}_wQ)^2 + 2\mathbf{E}_{C_r}(Q-\mathbf{E}_{C_r}Q)^2 + 2\mathbf{E}_{C_{r'}}(Q-\mathbf{E}_{C_{r'}}Q)^2 \right.$$
$$\left. \geqslant w(C_r)^2 w^2(C_{r'}) \left\|\Sigma(r')^{-1/2}\Sigma(r)\Sigma(r')^{-1/2} - I\right\|_F^4 \right\}.$$

Proof. Observe that $\mathbf{E}_{C_r}Q = \mathrm{tr}(\Sigma(r)(\Sigma(r)-I)) = \|\Sigma(r)-I\|_F^2 + \mathrm{tr}(\Sigma(r)-I)$ while, $\mathbf{E}_{Cr'}Q = \mathrm{tr}(\Sigma(r)-I)$. In particular, $\mathbf{E}_{C_r}Q - \mathbf{E}_{C_{r'}}Q = \|\Sigma(r)-I\|_F^2 \geqslant \Delta_{cov}^2 \|\Sigma(r)\|_{op}^2$. Thus, the mean of the polynomial $Q(x)$ is starkly different on the two components. By observing that the standard deviation of Q on each of C_r and $C_{r'}$ is much smaller than the mean, we will be able to derive a lower-bound on variance of Q under w-samples.

Thus, applying Lemma 2.3.15, with $A = w(C_r)^2$, $C = (\mathbf{E}_{C_r}Q - \mathbf{E}_wQ)^2$, $B = w(C_{r'})^2$, $D = \left(\mathbf{E}_{C_{r'}}Q - \mathbf{E}_wQ\right)^2$ and $\tau = \frac{1}{4}\|\Sigma(r)-I\|_F^4$ we have:

$$\mathcal{A}^{\frac{w}{4}}\left\{w(C_r)^2\left(\mathbf{E}_{C_r}Q - \mathbf{E}_wQ\right)^2 + w(C_{r'})^2\left(\mathbf{E}_{C_{r'}}Q - \mathbf{E}_wQ\right)^2 \geqslant \frac{1}{4}w(C_r)^2 w(C_{r'})^2 \|\Sigma(r)-I\|_F^4\right\} \tag{2.42}$$

Let's now lower bound $\mathbf{E}_w(Q-\mathbf{E}_wQ)^2$. We have:

$$\begin{aligned}
\mathcal{A}^{\frac{w}{4}}\Bigg\{&\mathbf{E}_w(Q-\mathbf{E}_wQ)^2 = \frac{k^2}{n^2}\sum_{i,j\leqslant n} w_i w_j \left(Q(x_i-x_j) - \mathbf{E}_wQ\right)^2 \\
&\geqslant \frac{k^2}{n^2}\sum_{i,j\in C_r \text{ or } i,j\in C_{r'}} w_i w_j \left(Q(x_i-x_j) - \mathbf{E}_wQ\right)^2 \\
&\geqslant \frac{k^2}{2n^2}\sum_{i,j\in C_r} w_i w_j \left(\mathbf{E}_{C_r}Q - \mathbf{E}_wQ\right)^2 - \frac{1}{2}\frac{k^2}{n^2}\sum_{i,j\in C_r} w_i w_j \left(Q(x_i-x_j) - \mathbf{E}_{C_r}Q\right)^2 \\
&+ \frac{k^2}{2n^2}\sum_{i,j\in C_r} w_i w_j \left(\mathbf{E}_{C_{r'}}Q - \mathbf{E}_wQ\right)^2 - \frac{1}{2}\frac{k^2}{n^2}\sum_{i,j\in C_{r'}} w_i w_j \left(Q(x_i-x_j) - \mathbf{E}_{C_{r'}}Q\right)^2 \\
&\geqslant \frac{1}{2}w(C_r)^2\left(\mathbf{E}_{C_r}Q - \mathbf{E}_wQ\right)^2 - \frac{1}{2}\frac{k^2}{n^2}\sum_{i,j\in C_r}\left(Q(x_i-x_j) - \mathbf{E}_{C_r}Q\right)^2 \\
&+ \frac{1}{2}w(C_{r'})^2\left(\mathbf{E}_{C_{r'}}Q - \mathbf{E}_wQ\right)^2 - \frac{1}{2}\frac{k^2}{n^2}\sum_{i,j\in C_{r'}}\left(Q(x_i-x_j) - \mathbf{E}_{C_{r'}}Q\right)^2 \\
&\geqslant \frac{1}{4}w(C_r)^2 w^2(C_{r'})\|\Sigma(r)-I\|_F^4 - \frac{1}{2}\mathbf{E}_{C_r}(Q-\mathbf{E}_{C_r}Q)^2 - \frac{1}{2}\mathbf{E}_{C_{r'}}(Q-\mathbf{E}_{C_{r'}}Q)^2\Bigg\},
\end{aligned}$$

where, in the final inequality, we applied (2.42). Rearranging completes the proof. □

Onwards to the upper bound now. Observe that the first two terms on the LHS of Lemma 2.3.18 can be upper bounded easily using Lemma 2.3.2: $\mathbf{E}_{C_r}(Q-\mathbf{E}_{C_r}Q)^2 \leqslant (C-1)\left\|\Sigma(r)^{1/2}Q\Sigma(r)^{1/2}\right\|_F^2 \leqslant \left\|\Sigma(r)^{1/2}\right\|_{op}^2 \|Q\|_F^2$. Similarly, $\mathbf{E}_{C_{r'}}(Q-\mathbf{E}_{C_r}Q)^2 \leqslant \|Q\|_F^2$. Thus, to finish the proof of Lemma 2.3.16, we need an upper bound on $\mathbf{E}_w(Q-\mathbf{E}_wQ)^2$ which we accomplish by relying on the certifiable hypercontractivity constraints.

In the following, we will use the following observation: From our bounded-variance constraints in $\mathcal{A}$, we have:

$$\mathcal{A}\left|\frac{\Pi,Q,w}{4}\right. \left\{\mathbf{E}_w(Q-\mathbf{E}_wQ)^2 \leqslant C\left\|\Pi(w)Q\Pi(w)\right\|_F^2\right\}. \tag{2.43}$$

From Lemma 2.3.12, we have:

$$\mathcal{A}\left|\frac{v,w}{\frac{s\log(\frac{1}{\rho})}{\delta^2}}\right. \left\{(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)\left(v^\top\Sigma(w)v\right)^s \leqslant 2\frac{1}{\delta^{2s}}(Cs)^s\left(v^\top\Sigma(r)v\right)^s+\delta\rho\lambda_{\max}^s\|v\|_2^{2s}\right\}.$$

To implement the linear transformation $x_i \to \Sigma(r')^{-1/2}x_i$, we substitute $v=\Sigma(r')^{-1/2}v$ and use that $\Sigma(r')^{-1} \succeq 1/\lambda_{\max}I$:

$$\mathcal{A}\left|\frac{\Pi,v,w}{O(s\log(1/\rho)/\delta^2)}\right. \left\{(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)\left\|\Pi(w)\Sigma(r')^{\dagger/2}v\right\|_2^{2s}\right.$$
$$\left.\leqslant 2\frac{1}{\delta^{2s}}(Cs)^s\|v\|_2^{2s}+\delta\rho\lambda_{\max}^s\|\Sigma(r')^{\dagger/2}v\|_2^{2s} \leqslant \left(2\frac{1}{\delta^{2s}}(Cs)^s+\delta\rho\kappa^s\right)\|v\|_2^{2s}\right\}. \tag{2.44}$$

We are now ready for the upper bound proof.

Lemma 2.3.19 (Certifiable Hypercontractivity Implies Low Variance)**.** *Let* $Q=\Sigma(r)-I$.

$$\mathcal{A}\left|\frac{w}{O(s\log(\kappa)/\delta^2)}\right. \left\{(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\eta)^{2s}\left(\mathbf{E}_w(Q-\mathbf{E}_wQ)^2\right)^s\right.$$
$$\left.\leqslant\left(4\frac{1}{\delta^{2s}}(Cs)^s\left\|\Sigma(r)^{1/2}\right\|_{op}^{2s}\right)^2 s^{2s}\left\|\Sigma(r)-I\right\|_F^2\right\} \tag{2.45}$$

Proof. Lemma 2.3.10 implies that $\mathcal{A}\vdash\{(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)\geqslant 0\}$. Thus, we can use

the multiplication rule (Fact 3.5) and multiply both sides of (2.89) with $(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)$ repeatedly while preserving the inequality.

Thus, we have using the bounded-variance constraints in $\mathcal{A}$:

$$\begin{aligned}
&\mathcal{A}\Big|_{O(s\log(1/\rho)/\delta^2)}^{\Pi,w}\Bigg\{(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)^s\left(\mathbf{E}_w(Q-\mathbf{E}_wQ)^2\right)^s\\
&\quad\leqslant(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)^s(C-1)^s\left\|\Pi(w)Q\Pi(w)\right\|_F^{2s}\\
&\quad\leqslant 2^s(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)^2(C-1)^s\left\|\Pi(w)Q'\Pi(w)\right\|_F^{2s}\\
&\quad\leqslant 2^s\left(\left(\frac{1}{\delta^2}\right)^s(Cs)^s\left\|\Sigma(r)^{1/2}\right\|_{op}^{2s}+\delta\rho\kappa^s\right)s^s(\mathcal{J}(w(C_r))(w(C_r)-\delta)+2^s\delta\rho)\left\|Q\Pi(w)\right\|_F^{2s}\\
&\quad\leqslant 2^s\left(\left(\frac{1}{\delta^2}\right)^s(Cs)^s\left\|\Sigma(r)^{1/2}\right\|_{op}^{2s}+\delta\rho\kappa^s\right)^2s^{2s}\left\|Q\right\|_F^{2s}\\
&\quad=\left(\left(\frac{2}{\delta^2}\right)^s(Cs)^s\left\|\Sigma(r)^{1/2}\right\|_{op}^{2s}+\delta\rho\kappa^s\right)^2s^{2s}\left\|\Sigma(r)-I\right\|_F^{2s}\Bigg\},
\end{aligned}$$

where, in the last two inequalities, we twice invoked the contraction bound from Lemma 2.8.1 along with the bound on $\left\|\Pi(w)\Sigma(r')^{-1/2}v\right\|_2^s$ from (2.44). Setting $\rho=\kappa^{-s}$ completes the proof. □

As in the previous subsection, we can improve the sum-of-squares degree of the proof above to be a fixed constant (independent of κ) in the case when $k=2$ by using the sharper bound on $\Sigma(w)$ in (2.44).

Lemma 2.3.20 (Certifiable Hypercontractivity Implies Low Variance, Two Components)**.** *Let* $Q=\Sigma(2)^{-1/2}\Sigma(1)\Sigma(2)^{-1/2}-I$.

$$\begin{aligned}
&\mathcal{A}\Big|_{O(1/\delta^4)}^{Q,\Sigma,w}\Bigg\{(\mathcal{J}(w(C(1)))(w(C_1)-\delta)+\delta\rho)^{2s}\left(\mathbf{E}_w(Q-\mathbf{E}_wQ)^2\right)^s\\
&\qquad\leqslant\left(4\frac{1}{\delta^{2s}}(Cs)^s\left\|\Sigma(r)^{1/2}\Sigma(2)^{-1/2}\right\|_{op}^{2s}\right)^2s^{2s}\left\|\Sigma(2)^{-1/2}\Sigma(1)\Sigma(2)^{-1/2}-I\right\|_F^2\Bigg\}
\end{aligned}\tag{2.46}$$

Proof. We proceed similarly as in the proof above up until (2.44) where, instead of using the uniform eigenvalue bound, we instead use the sharper bound from Lemma 2.3.11. As in the previous two subsections, following through the rest of the proof in Lemma 2.3.19 as is, allows us to eventually set $\log(1/\rho)=O(1/\delta^2)$ yielding a $O(1/\delta^4)$-degree SoS proof as desired. □

Proof of Lemma 2.3.16. As in the previous two lemmas, we argue after performing the linear transformation $\Sigma(r')^{-1/2}$ on the samples in order to simplify notation.

From Lemma 2.3.18, we have:

$$\mathcal{A} \left|\frac{w}{4}\right. \Big\{ 4\mathbf{E}_w (Q - \mathbf{E}_w Q)^2 + 2\mathbf{E}_{C_r} (Q - \mathbf{E}_{C_r} Q)^2 + 2\mathbf{E}_{C_{r'}} (Q - \mathbf{E}_{C_{r'}} Q)^2 \\ \geqslant w(C_r)^2 w^2(C_{r'}) \left\|\Sigma(r) - I\right\|_F^4 \Big\}$$

Multiplying both sides of the and apply the SoS Almost Triangle Inequality (Fact 2.2.8) and obtain:

$$\mathcal{A} \left|\frac{w,Q}{4s}\right. \Big\{ 2^{3s} \left(\mathbf{E}_w (Q - \mathbf{E}_w Q)^{2s} + \mathbf{E}_{C_r} (Q - \mathbf{E}_{C_r} Q)^{2s} + \mathbf{E}_{C_{r'}} (Q - \mathbf{E}_{C_{r'}} Q)^{2s} \right) \\ \geqslant w(C_r)^{2s} w^{2s}(C_{r'}) \left\|\Sigma(r) - I\right\|_F^{4s} \Big\}$$

Multiplying by $(\mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho)^s$ on both sides, we get:

$$\mathcal{A} \left|\frac{Q,w}{O(s\log(1/\rho)/\delta^2)}\right. \Big\{ (\mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho)^s \, w(C_r)^{2s} w^{2s}(C_{r'}) \left\|\Sigma(r) - I\right\|_F^{4s} \\ \leqslant (\mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho)^s \, 2^{3s} \cdot \\ \left(\mathbf{E}_w (Q - \mathbf{E}_w Q)^{2s} + \mathbf{E}_{C_r} (Q - \mathbf{E}_{C_r} Q)^{2s} + \mathbf{E}_{C_{r'}} (Q - \mathbf{E}_{C_{r'}} Q)^{2s} \right) \Big\} . \quad (2.47)$$

Using the upper bounds proved above (Lemma 2.3.19 and the preceding discussion) on each of the three terms on the RHS, we get:

$$\mathcal{A} \left|\frac{w}{O(s\log(\kappa)/\delta^2)}\right. \Big\{ (\mathcal{J}(w(C_r))(w(C_r) - \delta) + \delta\rho)^s \, w(C_r)^{2s} w^{2s}(C_{r'}) \left\|\Sigma(r) - I\right\|_F^{4s} \\ \leqslant 2^{O(s)} \left(4\frac{1}{\delta^{2s}} (Cs)^s \left\|\Sigma(r)^{1/2} \Sigma(r')^{-1/2}\right\|_{op}^{2s} + 1 \right) \left\|\Sigma(r) - I\right\|_F^{2s} \Big\} . \quad (2.48)$$

Applying the SoS Cancellation lemma (Lemma 2.8.2), we have:

$$\mathcal{A} \left|\frac{w}{O(s\log(\kappa)/\delta^2)}\right. \Bigg\{ (\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho)\, w(C_r)^2 w^2(C_{r'}) \left\|\Sigma(r)-I\right\|_F^4 \\ \leqslant 2^{O(s)} \left(4\frac{1}{\delta^2}\,(Cs)\left\|\Sigma(r)^{1/2}\Sigma(r')^{-1/2}\right\|_{op}^2\right) \left\|\Sigma(r)-I\right\|_F^2 \Bigg\}. \quad (2.49)$$

Applying Lemma 2.3.10 to observe

$$\mathcal{A} \left|\frac{w}{O(\log(1/\rho)/\delta^2)}\right. \left\{(\mathcal{J}(w(C_r))(w(C_r)-\delta)+\delta\rho) \geqslant (w(C_r)-2\delta)\right\}.$$

Thus, using $\mathcal{A} \vdash \{w(C_r)^2 w(C_{r'})^2) \leqslant 1\}$, we get:

$$\mathcal{A} \left|\frac{w}{O(s\log(\kappa)/\delta^2)}\right. \Bigg\{ w(C_r)^3 w^2(C_{r'}) \left\|\Sigma(r)-I\right\|_F^4 \\ \leqslant 2\delta \left\|\Sigma(r)-I\right\|_F^4 + 2^{O(s)} \left(4\frac{1}{\delta^2}\,(Cs)\left\|\Sigma(r)^{1/2}\Sigma(r')^{-1/2}\right\|_{op}^2\right) \left\|\Sigma(r)-I\right\|_F^2 \Bigg\}. \quad (2.50)$$

Dividing throughout by $\left\|\Sigma(r)-I\right\|_F^4$, and using that and that $\left\|\Sigma(r)-I\right\|_F^2 \geqslant \Delta_{cov}^2 \left\|\Sigma(r)^{1/2}\Sigma(r')^{-1/2}\right\|_{op}^2$ yields:

$$\mathcal{A} \left|\frac{w}{O(s\log(\kappa)/\delta^2)}\right. \left\{ w(C_r)^3 w(C_{r'})^3 \leqslant 2\delta + \left(4\frac{1}{\delta^2}\,(Cs)\,\Delta_{cov}^{-2s}\right) \left\|\Sigma(r)-I\right\|_F^{2s} \right\}. \quad (2.51)$$

Using that $\Delta_{cov} \gg Cs/\delta^2$ and $s = O(1/\delta^2)$ yields:

$$\mathcal{A} \left|\frac{w}{O(\log(\kappa)/\delta^4)}\right. \left\{ w(C_r)^3 w(C_{r'})^3 \leqslant O(\delta) \right\}. \quad (2.52)$$

Using SoS cancellation (Lemma 2.8.2) again yields:

$$\mathcal{A} \left|\frac{w}{O(\log(\kappa)/\delta^4)}\right. \left\{ w(C_r) w(C_{r'}) \leqslant O(\delta^{1/3}) \right\}. \quad (2.53)$$

□

Improved SoS Degree Bounds for $k = 2$ By using Lemma 2.3.20 instead of Lemma 2.3.19 in the above argument immediately yields Lemma 2.3.17.

2.4 Outlier-Robust Clustering of Reasonable Distributions

In this section, we augment the algorithm from the previous section to tolerate an $\epsilon \leqslant O(1/k)$ fraction of fully adversarial outliers. Recall that in this setting, the input sample Y is obtained by first generating a sample X from the underlying mixture model and adversarially corrupting an ϵ-fraction of X.

The following is the main result of this section:

Theorem 59 (Outlier-Robust Clustering of Mixture of Reasonable Distributions)**.** *Fix $\epsilon > 0$. Let $\mathcal{D}$ be a nice distribution that is $s(\delta)$-certifiably $(\delta, C\delta)$-anti-concentrated for all $\delta > 0$ and has h-certifiably C-hypercontractive degree 2 polynomials for every h. There exists an algorithm that takes input an ϵ corruption Y of X of size n generated according equi-weighted Δ-separated mixture of $\mathcal{D}(\mu(r), \Sigma(r))$ for $r \leqslant k$ with true clusters $C_1, C_2, \ldots, C_k$ and outputs $\hat{C}_1, \hat{C}_2, \ldots \hat{C}_k$ such that there exists a permutation $\pi : [k] \to [k]$ satisfying*

$$\min_{i \leqslant k} \frac{|C_i \cap \hat{C}_{\pi(i)}|}{|C_i|} \geqslant 1 - \eta - O(k\epsilon)\,.$$

The algorithm succeeds with probability at least $1 - 1/k$ whenever $\Delta \geqslant \Delta_{rob} = \Omega(\mathit{poly}(k/\eta))$, need $n \geqslant d^{O(\mathit{poly}(k/\eta))}$ samples and runs in time $n^{O(\log(\kappa)\mathit{poly}(k/\eta))}$ where κ is spread of the mixture.

For the special case of $k = 2$, the algorithm runs in time $n^{O(\mathit{poly}(k/\eta))}$ and uses $d^{O(\mathit{poly}(k/\eta))}$ samples (with no dependence on the spread κ.)

Recall that the spread $\kappa = \sup_{v \in \mathcal{R}^d} \max_{i,j \leqslant k} \frac{v^\top \Sigma(i) v}{v^\top \Sigma(j) v}$. In Section 2.5, we will use the algorithm above as a subroutine to get a fully-polynomial algorithm with no dependence on the spread κ of the mixture in the running time.

2.4.1 Algorithm

Constraint System. Our constraint system $\mathcal{A}_{rob}$ is similar to the one from the previous section with one key difference introduced in order to handle the adversarial outliers. In the uncorrupted setting, we are given the original uncorrupted sample $X = C_1 \cup C_2 \cup \ldots C_k$ and our program encodes constraints on a subset $\hat{C}$ of samples with the intended solutions to be the true clusters C_is.

In the outlier-robust setting, we only get to observe the ϵ-corruption Y of X. Thus, the points in the indices corresponding to C_i need not satisfy the constraints from the previous section.

We handle this by introducing an extra set of d-dimensional vector-valued indeterminates $X' = \{x'_1, x'_2, \ldots, x'_n\}$ that are intended to be the original uncorrupted sample X that generated Y. Since X' is (supposed to be) a uncorrupted sample, we can now encode finding a subset $\hat{C}$ of X' (instead of X) with the intended solutions to be the true clusters C_is of the original X. In order to force X' to be close to X, we force constraints intersection constraints (via the new matching variables m_is) that ask X' to intersect Y in $(1-\epsilon)$-fraction of points (just like the true X does). This implies that X' intersects X in $\geqslant (1-2\epsilon)$-fraction of the points and as we will soon see, this is enough for us to execute the arguments from the previous section with relatively little change.

Covariance constraints introduce a matrix valued indeterminate intended to be the square root of Σ.

$$\text{Covariance Constraints: } \mathcal{A}_1 = \left\{ \begin{array}{c} \Pi = UU^\top \\ \Pi^2 = \Sigma\,. \end{array} \right\} \tag{2.54}$$

The intersection constraints force that X' be close to X.

$$\text{Intersection Constraints: } \mathcal{A}_2 = \begin{Bmatrix} \forall i \in [n], & m_i^2 = m_i \\ & \sum_{i\in[n]} m_i = (1-\epsilon)n \\ \forall i \in [n], & m_i(y_i - x'_i) = 0\,. \end{Bmatrix} \tag{2.55}$$

The subset constraints introduce w, which indicates the subset $\hat{C}$ intended to be the true clusters of X'.

$$\text{Subset Constraints: } \mathcal{A}_3 = \begin{Bmatrix} \forall i \in [n]. & w_i^2 = w_i \\ & \sum_{i\in[n]} w_i = \dfrac{n}{k}\,. \end{Bmatrix} \tag{2.56}$$

Parameter constraints create indeterminates to stand for the covariance Σ and mean μ of $\hat{C}$

(indicated by w).

$$\text{Parameter Constraints: } \mathcal{A}_4 = \left\{ \begin{array}{r} \frac{1}{n}\sum_{i=1}^{n} w_i \left(x_i' - \mu\right)\left(x_i' - \mu\right)^\top = \Sigma \\ \frac{1}{n}\sum_{i=1}^{n} w_i x_i' = \mu\,. \end{array} \right\} \tag{2.57}$$

Finally, we enforce certifiable anti-concentration and hypercontractivity of $\hat{C}$.

$$\text{Certifiable Anti-Concentration : } \mathcal{A}_4 = \left\{ \begin{array}{l} \frac{k^2}{n^2}\sum_{i,j=1}^{n} w_i w_j q_{\delta,\Sigma}^2\left(\left(x_i' - x_j'\right), v\right) \leqslant 2^{s(\delta)} C\delta \left(v^\top \Sigma v\right)^{s(\delta)} \\ \frac{k^2}{n^2}\sum_{i,j=1}^{n} w_i w_j q_{\tau,\Sigma}^2\left(\left(x_i' - x_j'\right), v\right) \leqslant 2^{s(\tau)} C\tau \left(v^\top \Sigma v\right)^{s(\eta)} \end{array} \right\}, \tag{2.58}$$

where $s(x) = O(1/x^2)$. Certifiable Hypercontractivity: $\mathcal{A}_5 =$

$$\left\{ \begin{array}{l} \forall h \leqslant 2s, \quad \frac{k^2}{n^2}\sum_{i,j\leqslant n} w_i w_j \left(Q(x_i' - x_\ell') - \frac{k^2}{n^2}\sum_{i,\ell\leqslant n} w_i w_\ell Q(x_i' - x_\ell')\right)^{2h} \\ \qquad\qquad \leqslant (Ch)^{2h}\left(\frac{k^2}{n^2}\sum_{i,\ell\leqslant n} w_i w_\ell \left(Q(x_i' - x_\ell') - \frac{k^2}{n^2}\sum_{i,\ell\leqslant n} w_i w_\ell Q(x_i' - x_\ell')\right)^2\right)^h . \end{array} \right\} \tag{2.59}$$

Certifiable Bounded Variance: $\mathcal{A}_6$ =

$$\left\{ \forall j \leqslant 2s, \quad \frac{k^2}{n^2}\sum_{i,\ell\leqslant n} w_i w_\ell \left(Q(x_i' - x_\ell') - \frac{k^2}{n^2}\sum_{i,\ell\leqslant n} w_i w_\ell Q(x_i' - x_\ell')\right)^2 \leqslant C\left\|\Pi Q \Pi\right\|_F^2 . \right\} \tag{2.60}$$

Our rounding algorithm is exactly the same as in the previous section giving us:

Algorithm 60 (Outlier-Robust Clustering General Mixtures)**.**

Given: *An ϵ-corruption Y of original uncorrupted sample $X = C_1 \cup C_2 \cup \ldots C_k$ with true clusters $C_1, C_2, \ldots, C_k$.*

Output: *A partition of Y into an approximately correct clustering $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_k$.*

Operation:

1. *Find a pseudo-distribution $\tilde{\zeta}$ satisfying $\mathcal{A}_{rob}$ with $s = \log(\kappa)$poly(k/η), $\delta =$*

η^6/k^{12}, *and* $\tau = 1/(C\text{poly}(k))$, *and minimizing* $\left\|\tilde{\mathbb{E}}[w]\right\|_2^2$.

2. *For* $M = \tilde{\mathbb{E}}_{w\sim\tilde{\zeta}}[ww^\top]$, *repeat for* $1 \leqslant \ell \leqslant k$*:*

   (a) *Choose a uniformly random row* i *of* M.

   (b) *Let* $\hat{C}_\ell$ *be the largest* $\frac{n}{k}$ *entries in the* i*th row of* M.

   (c) *Remove the rows and columns with indices in* $\hat{C}_\ell$.

Analysis of Algorithm An analog of Lemma 2.3.2 extends to this setting without any change.

Lemma 2.4.1 (Typical samples are good)**.** *Let* X *be an original uncorrupted sample of size* n *from a equi-weighted* Δ*-separated mixture* $\mathcal{D}(\mu(r), \Sigma(r))$ *for* $r \leqslant k$.

Then, for $n_0 = \Omega\left((sd)^{8s} k \log k\right)$ *and for all* $n \geqslant n_0$*, the original uncorrupted sample* X *of size* n *is good with probability at least* $1 - 1/d$.

As in the previous section, the heart of the analysis is proving the following lemma that bounds the pairwise products $w(C_r)w(C_{r'})$ for all $r \neq r'$.

Lemma 2.4.2 (Intersection Bounds from Separation)**.** *Let* Y *be an* ϵ*-corruption of a good sample* X *from a* $\Delta \geqslant \Delta_{rob}$*-separated mixture of reasonable distribution* $\mathcal{D}$ *with true clusters* $C_1, C_2, \ldots, C_k$ *of size* n/k*. Let* $w(C_r)$ *denote the linear polynomial* $\frac{k}{n}\sum_{i\in C_r} w_i$ *for every* $r \leqslant k$*. Then, for every* $r \neq r'$*,*

$$\mathcal{A} \left|\frac{w}{O(\log(\kappa)/\delta^4)}\right. \left\{ \sum_{r\neq r'} w(C_r)w(C_{r'}) \leqslant O(k\epsilon) + O(k^2\delta^{1/3}) \right\} .$$

For the special case when the number of components in the mixture is $k = 2$, we can improve on the lemma above and give a sum-of-squares proof of degree $O(s(\delta)^2)$ with no dependence on κ.

Lemma 2.4.3 (Intersection Bounds from Separation, Two Components)**.** *Let* Y *be an* ϵ*-corruption of a good sample* X *from a* $\Delta \geqslant \Delta_{rob}$*-separated mixture of reasonable distribution* $\mathcal{D}$ *with true clusters* C_1, C_2 *of size* $n/2$ *each. Let* $w(C_r)$ *denote the linear polynomial* $\frac{k}{n}\sum_{i\in C_r} w_i$ *for every* $r \leqslant 2$*. Then,*

$$\mathcal{A}_{rob} \left|\frac{w}{O(1/\delta^4)}\right. \left\{ w(C_1)w(C_2) \leqslant O(\epsilon + \delta^{1/3}) \right\} .$$

Given Lemma 2.3.3, the proof of Theorem 59 follows by the same argument as for Theorem 56.

2.4.2 Proof of Lemmas 2.4.2 and 2.4.3

As we show in this section, the proof of Lemma 2.4.2 follows from essentially the same argument as in the previous section with two additional observations.

The key idea in bringing the machinery from the previous section into play is to consider the following variables that satisfy constraints of being the indicator of the intersection between X' (indeterminates in our program) and X (original uncorrupted sample we do not have access to) - let $m'_i = m_i \cdot \mathbf{1}(y_i = x_i)$ for every i. We now make the following key definition/notation.

Definition 2.4.4 (Proxy Variables and Cluster Sizes)**.** *Let* $w'_i = w_i m'_i = w_i m_i \mathbf{1}(y_i = x_i)$ *and define* $w'(C_r) = \frac{k}{n} \sum_{i \in C_r} w'_i$ *for every* r.

We refer to w'_i variables as proxy variables (they allow us to talk about subsets of X by "proxy"). Observe that we do not have access to the w'_i variables through our program. They only appear in our analysis of the algorithm. They allow us to "go between" x_is (the originals sample that we do not have access to) and x'_i (the indeterminates that our constraints are defined over).

The result that formally allows us to do this is:

Lemma 2.4.5 (Matching with Original Uncorrupted Samples)**.** *Let* $m'_i = m_i \cdot \mathbf{1}(y_i = x_i)$ *for every* i. *Let* $w'_i = w_i m'_i = w_i m_i \mathbf{1}(y_i = x_i)$. *Then,*

$$\mathcal{A}_{rob} \left|\frac{w'}{2}\right. \left\{w_i'^2 = w'_i \, \forall i\right\} \cup \{w'_i(x'_i - x_i) = 0\} \ .$$

Proof. For the first conclusion,

$$\mathcal{A}_{rob} \left|\frac{w'}{2}\right. \left\{w_i'^2 = w_i^2 m_i^2 \cdot \mathbf{1}(y_i = x_i)^2 = w_i m_i \mathbf{1}(y_i = x_i) = w'_i\right\} \ .$$

For the second conclusion,

$$\mathcal{A}_{rob} \left|\frac{w'}{2}\right. \Big\{ w_i'(x_i' - x_i) = w_i'(x_i' - y_i) + w_i'(y_i - x_i)$$
$$= \mathbf{1}(y_i = x_i) w_i m_i (x_i' - y_i) + m_i w_i \mathbf{1}(y_i = x_i)(x_i - y_i)$$
$$= 0 \Big\}.$$

□

Using this simple lemma, as we will soon discuss in some more detail, we get to apply our previous arguments to the original sample X by simply shifting to the "proxy" w_i' variables. As a result, we will be able to prove the following intersection bounds for the proxy cluster sizes.

Lemma 2.4.6 (Proxy Intersection Bounds from Separation)**.** *Let Y be an ϵ-corruption of a good sample X. Let $w'(C_r)$ denote the linear polynomial $\frac{k}{n}\sum_{i\in C_r} w_i'$ for every $r \leqslant k$. Then, for every $r \neq r'$,*

$$\mathcal{A}_{rob} \left|\frac{w}{O(\log(\kappa)/\delta^4)}\right. \left\{ w'(C_r) w'(C_{r'}) \leqslant O(\delta^{1/3}) \right\}.$$

For the special case when the number of components in the mixture is $k = 2$, we can improve on the lemma above and give a sum-of-squares proof of degree $O(s(\delta)^2)$ with no dependence on κ.

Lemma 2.4.7 (Proxy Intersection Bounds from Separation, Two Components)**.** *Let Y be an ϵ-corruption of a good sample X. Let $w'(C_r)$ denote the linear polynomial $\frac{k}{n}\sum_{i\in C_r} w_i'$ for every $r \leqslant 2$. Then,*

$$\mathcal{A}_{rob} \left|\frac{}{O(1/\delta^4)}\right. \left\{ w'(C_1) w'(C_2) \leqslant O(\delta^{1/3}) \right\}.$$

It is easy to complete the proof of Lemmas 2.4.2 and 2.4.7 using the above two lemmas. We show the proof for Lemma 2.4.2. The proof for Lemma 2.4.7 is analogous.

We will use the following bound that (in low-degree SoS) shows that X and X' intersect in $(1 - 2\epsilon)n$ points.

Lemma 2.4.8 (Matching with Original Uncorrupted Samples)**.** *Let $m_i' = m_i \cdot \mathbf{1}(y_i = x_i)$ for*

every i. Then,

$$\mathcal{A}_{rob} \vdash_2 \left\{ \sum_{i \leqslant n} m'_i \geqslant (1-2\epsilon)n \right\}.$$

Proof. Observe that using $\{m_i^2 = m_i\} \vdash_2^m \{m_i \leqslant 1\}$, we have:

$$\mathcal{A}_{rob} \vdash_2 \left\{ \sum_{i \leqslant n} m_i \cdot \mathbf{1}(y_i \neq x_i) \leqslant \sum_{i \leqslant n} \mathbf{1}(y_i \neq x_i) = \epsilon n \right\}.$$

Similarly,

$$\mathcal{A}_{rob} \vdash_2 \left\{ \sum_{i \leqslant n} (1-m_i) \cdot \mathbf{1}(y_i = x_i) \leqslant \sum_{i \leqslant n} (1-m_i) = \epsilon n \right\}.$$

Thus,

$$\mathcal{A}_{rob} \vdash_2 \left\{ \sum_{i \leqslant n} m_i \cdot \mathbf{1}(y_i = x_i) \geqslant \sum_{i \leqslant n} (m_i + (1-m_i)) \left(\mathbf{1}(y_i = x_i) + \mathbf{1}(y_i \neq x_i)\right) \geqslant n - 2\epsilon n \right\}.$$

□

Proof of Lemma 2.4.2. Observe that using $\mathcal{A}_{rob} \vdash \{m'_i \leqslant 1\}$ for every i, and $\mathcal{A}_{rob} \vdash \{\sum_{r \leqslant k} w(C_r) = 1\}$ we have:

$$\begin{aligned}
\mathcal{A}_{rob} \vdash_{O(\log(\kappa)/\delta^4)}^{w,w',m'} \Bigg\{ \sum_{r \neq r'} w'(C_r) w'(C_{r'}) &= \frac{k^2}{n^2} \sum_{r \neq r'} \sum_{i \in C_r, j \in C_{r'}} w_i w_j m'_i m'_j \\
&\geqslant \frac{k^2}{n^2} \sum_{r \neq r'} \sum_{i \in C_r, j \in C_{r'}} w_i w_j - 2\frac{k^2}{n^2} \sum_{r \neq r'} \sum_{i \in C_r, j \in C_{r'}} w_i w_j (1-m_i) \\
&\geqslant \frac{k^2}{n^2} \sum_{r \neq r'} \sum_{i \in C_r, j \in C_{r'}} w_i w_j - 2\frac{k^2}{n^2} \sum_{r \neq r'} \sum_{i \in C_r, j \in C_{r'}} w_i (1-m_i) \\
&\geqslant \frac{k^2}{n^2} \sum_{r \neq r'} \sum_{i \in C_r, j \in C_{r'}} w_i w_j - 2\frac{k}{n} \sum_{r \neq r'} \sum_{i \in C_r, j \in C_{r'}} (1-m_i) \\
&\geqslant \frac{k^2}{n^2} \sum_{r \neq r'} \sum_{i \in C_r, j \in C_{r'}} w_i w_j - 2\frac{k}{n} \sum_{r \neq r'} \sum_{i \in C_r, j \in C_{r'}} (1-m_i) \\
&= \frac{k^2}{n^2} \sum_{r \neq r'} \sum_{i \in C_r, j \in C_{r'}} w_i w_j - 2k\epsilon \Bigg\}.
\end{aligned}$$

Rearranging yields:

$$\mathcal{A}_{rob} \vdash^{w}_{O(\log(\kappa)/\delta^4)} \left\{ \sum_{r \neq r'} w(C_r)w(C_{r'}) \leqslant \sum_{r \neq r'} w'(C_r)w'(C_{r'}) + 2k\epsilon \right\} .$$

Plugging in the bound from Lemma 2.4.6 completes the proof. □

2.4.3 Proof of the Simultaneous Proxy Intersection Bounds

We prove Lemma 2.4.6 with a proof strategy that is essentially same as the one employed in the proofs of Lemmas 2.3.5, 2.3.13 and 2.3.16. We will start with constraints stated in terms of the X' variables and use Lemma 2.4.5 at appropriate places to transition into X variables. At that point, we can plug in our argument from the previous section without change.

We will do the case of spectral separation in detail to illustrate why this strategy works essentially syntactically.

Lemma 2.4.9 (Simultaneous Proxy Intersection Bounds from Spectral Separation)**.** *Suppose there exists a v such that $v^\top \Sigma(r')v > \Delta_{\mathsf{spectral}} v^\top \Sigma(r')v$. Let $\kappa = \sup_{v \in \mathbb{R}^d} \max_{i \leqslant k} \frac{v^\top \Sigma(i) v}{v^\top \Sigma(r')v}$.*

Then, whenever $\Delta_{\mathsf{spectral}} \gg Cs/\delta$,

$$\mathcal{A}_{rob} \vdash^{w'}_{O(\log(\kappa)/\delta^4)} \left\{ w'(C_r)w'(C_{r'}) \leqslant O(\sqrt{\delta}) \right\} .$$

Observe, as in the previous section, that $B = 1$ when $k = 2$.

As in the previous section, we start by proving a lower-bound on the variance of $\mathcal{D}_w$ in the direction v where $\Sigma(r)$ and $\Sigma(r')$ are spectrally separated. This gives us:

Lemma 2.4.10 (Large Intersection Implies High Variance, Spectral Separation)**.**

$$\mathcal{A}_{rob} \vdash_{4s} \left\{ w'(C_{r'})w'(C_r) \left(v^\top \left(\Sigma(r) + \Sigma(r') \right) v^\top \right)^s \right.$$
$$\left. \leqslant \left(\frac{2}{\delta^2} \right)^s \left(v^\top \Sigma(w) v \right)^s + C\delta \left(v^\top \left(\Sigma(r) + \Sigma(r') \right) v^\top \right)^s \right\}$$

Proof. We know from Lemma 2.3.2 that two-sample-centered points from both C_r and $C_{r'}$ (note that these are subsets of the original uncorrupted sample X) are $2s$-certifiably $(\delta, C\delta)$-

anti-concentrated. Using Definition 3.2.28, thus yields:

$$\mathcal{A}_{rob} \vdash_{4s} \left\{ \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w'_{i_1} w'_{i_2} w'_{j_1} w'_{j_2} \langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \rangle^{2s} \right.$$
$$\geqslant \delta^{2s} w'(C_r)^2 w'(C_{r'})^2 \left(v^\top 2(\Sigma(r) + \Sigma(r'))v^\top \right)^s$$
$$\left. - \delta^{2s} \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w'_{i_1} w'_{i_2} w'_{j_1} w'_{j_2} q^2_{\delta, 2(\Sigma(r)+\Sigma(r'))}(x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v) \right\} \quad (2.61)$$

Using that $\mathcal{A}_{rob} \vdash_4 \left\{ w'_{i_1} w'_{i_2} w'_{j_1} w'_{j_2} \leqslant 1 \right\}$ for every i_1, i_2, j_1, j_2 and using $2s$-certifiable $(\delta, C\delta)$-anti-concentration of $x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}$ and invoking Definition 3.2.28, we have:

$$\mathcal{A}_{rob} \vdash^{w',\Sigma}_{4s} \left\{ \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w'_{i_1} w'_{i_2} w'_{j_1} w'_{j_2} q^2_{\delta, 2(\Sigma(r)+\Sigma(r'))}(x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v) \right.$$
$$\left. \leqslant \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} q^2_{\delta, 2(\Sigma(r)+\Sigma(r'))}(x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v) \leqslant C\delta \left(v^\top 2(\Sigma(r) + \Sigma(r'))v \right)^s \right\} \quad (2.62)$$

Plugging in the above bound in (2.61) gives:

$$\mathcal{A}_{rob} \vdash \left\{ \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w'_{i_1} w'_{i_2} w'_{j_1} w'_{j_2} \langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \rangle^{2s} \right.$$
$$\left. \geqslant \delta^{2s} \left(w'(C_r)^2 w'(C_{r'})^2 - C\delta \right) \left(v^\top 2(\Sigma(r) + \Sigma(r'))v^\top \right)^s \right\} \quad (2.63)$$

Rearranging thus yields:

$$\mathcal{A}_{rob} \vdash_{4s} \left\{ \frac{1}{\delta^{2s}} \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w'_{i_1} w'_{i_2} w'_{j_1} w'_{j_2} \langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \rangle^{2s} \right.$$
$$+ C\delta \left(v^\top 2(\Sigma(r) + \Sigma(r'))v^\top \right)^s$$
$$\left. \geqslant w'(C_r)^2 w'(C_{r'})^2 \left(v^\top 2(\Sigma(r) + \Sigma(r'))v^\top \right)^s \right\} \quad (2.64)$$

So far in the proof, the only change (compared to the proof of Lemma 2.3.6) in the proof has been that we work with the subset indicated by w'_i.

The key additional step we observe now is the following consequence of $\mathcal{A}_{rob} \vdash \{w'_i(x_i - x'_i) = 0\}$ (Lemma 2.4.5).

$$\mathcal{A}_{rob} \vdash_4 \left\{ w'_{i_1} w'_{i_2} w'_{j_1} w'_{j2} \left\langle x'_{i_1} - x'_{i_2} - x'_{j_1} + x'_{j_2}, v \right\rangle = w'_{i_1} w'_{i_2} w'_{j_1} w'_{j_2} \langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \rangle \right\} .$$

Using further that $w_i \geqslant w'_i$, we have:

$$\mathcal{A}_{rob} \vdash_{4s} \left\{ \left(\frac{4cs}{\delta^2} \right)^s \left(v^\top \Sigma(w) v \right)^s \geqslant \frac{1}{\delta^{2s}} \frac{k^4}{n^4} \sum_{i_1,i_2,j_1,j_2 \in [n]} w_{i_1} w_{i_2} w_{j_1} w_{j_2} \left\langle x'_{i_1} - x'_{i_2} - x'_{j_1} + x'_{j_2}, v \right\rangle^{2s} \right.$$
$$\geqslant \frac{1}{\delta^{2s}} \frac{k^4}{n^4} \sum_{i_1,i_2,j_1,j_2 \in [n]} w'_{i_1} w'_{i_2} w'_{j_1} w'_{j_2} \left\langle x'_{i_1} - x'_{i_2} - x'_{j_1} + x'_{j_2}, v \right\rangle^{2s}$$
$$\geqslant \frac{1}{\delta^{2s}} \frac{k^4}{n^4} \sum_{i_1,i_2,j_1,j_2 \in [n]} w'_{i_1} w'_{i_2} w'_{j_1} w'_{j_2} \langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \rangle^{2s}$$
$$\left. \geqslant \frac{1}{\delta^{2s}} \frac{k^4}{n^4} \sum_{i_1,i_2 \in C_r, j_1,j_2 \in C_{r'}} w'_{i_1} w'_{i_2} w'_{j_1} w'_{j_2} \langle x_{i_1} - x_{i_2} - x_{j_1} + x_{j_2}, v \rangle^{2s} \right\} .$$

Plugging in the upper bound above in (2.64) and canceling out a copy of 2^s from both sides gives the lemma.

□

The basic spectral upper bound also follows by simply shifting to the proxy variables w'_i. This yields us the following analog of Lemma 2.3.7:

Lemma 2.4.11 (Spectral Upper Bound via Anti-Concentration)**.**

$$\mathcal{A}_{rob} \vdash_{4s} \left\{ \left(w'(C_r)^2 - C\delta\right) \left(v^\top \Sigma(w) v^\top\right)^s \leqslant \left(\frac{Cs}{\delta^2}\right)^s \left(v^\top \Sigma(r) v\right)^s \right\} \tag{2.65}$$

Proof. Our constraint system $\mathcal{A}_{rob}$ allows us to derive that two-sample-centered points indicated by w are $2s$-certifiably $(\delta, C\delta)$-anti-concentrated with witnessing polynomial $p_{\mathcal{D}}$. Using Definition 3.2.28 and summing up over all n after multiplying throughout by $w'_i w'_j$ yields:

$$\begin{aligned} &\mathcal{A}_{rob} \vdash_{4s} \left\{ \delta^{2s} w'(C_r)^2 \left(v^\top \Sigma(w) v^\top\right)^s \right. \\ &\left. \leqslant \frac{k^2}{n^2} \sum_{i,j \in C_r} w'_i w'_j \left\langle \frac{1}{\sqrt{2}} \left(x'_i - x'_j\right), v \right\rangle^{2s} + \delta^{2s} \frac{k^2}{n^2} \sum_{i \neq j \in C_r} w'_i w'_j q^2_{\delta,\Sigma(w)} \left(\frac{1}{\sqrt{2}} \left(x'_i - x'_j\right), v \right) \right\} \end{aligned} \tag{2.66}$$

Using that $\mathcal{A}_{rob} \vdash_2 \left\{ w'_i w'_j \left((x'_i - x'_j) - (x_i - x_j) \right) = 0 \right\}$ (two applications of Lemma 2.4.5) yields:

$$\begin{aligned} &\mathcal{A}_{rob} \vdash^{\Sigma, w'}_{4s} \left\{ \delta^{2s} w'(C_r)^2 \left(v^\top \Sigma(w) v^\top\right)^s \right. \\ &\left. \leqslant \frac{k^2}{n^2} \sum_{i,j \in C_r} w'_i w'_j \left\langle \frac{1}{\sqrt{2}} \left(x_i - x_j\right), v \right\rangle^{2s} + \delta^{2s} \frac{k^2}{n^2} \sum_{i \neq j \in C_r} w'_i w'_j q^2_{\delta,\Sigma(w)} \left(\frac{1}{\sqrt{2}} \left(x_i - x_j\right), v \right) \right\} \end{aligned} \tag{2.67}$$

Using that $\mathcal{A}_{rob} \vdash_2 \left\{ w'_i w'_j \leqslant 1 \right\}$ for every i, j, using that $\mathcal{A}_{rob}$ derives $2s$-certifiable $(\delta, C\delta)$-anti-concentration of w-samples and invoking Definition 3.2.28, we have:

$$\begin{aligned} \mathcal{A}_{rob} \vdash_{4s} \left\{ \frac{k^2}{n^2} \sum_{i \neq j \in C_r} w'_i w'_j q^2_{\delta,\Sigma(w)} \left(\frac{1}{\sqrt{2}} \left(x_i - x_j\right), v \right) \leqslant \frac{k^2}{n^2} \sum_{i \neq j \in [n]} w'_i w'_j q^2_{\delta,\Sigma(w)} \left(\frac{1}{\sqrt{2}} \left(x_i - x_j\right), v \right) \right. \\ \left. \leqslant C\delta \left(v^\top \Sigma(w) v\right)^s \right\} \end{aligned} \tag{2.68}$$

Further, using that $\mathcal{A}_{rob} \vdash_{\overline{2}} \left\{ w_i' w_j' \leqslant 1 \right\}$ for all i, j and relying on the certifiable subgaussianity of C_r, we have:

$$\begin{aligned} \mathcal{A}_{rob} \vdash_{\frac{4s}{}}^{\Sigma, w'} \Bigg\{ \frac{k^2}{n^2} \sum_{i,j \in C_r} w_i' w_j' \left\langle \frac{1}{\sqrt{2}} (x_i - x_j), v \right\rangle^{2s} \\ \leqslant \frac{k^2}{n^2} \sum_{i,j \in C_r} \left\langle \frac{1}{\sqrt{2}} (x_i - x_j), v \right\rangle^{2s} \\ = (Cs)^s \left(v^\top \Sigma(r) v \right)^s \Bigg\} \end{aligned} \tag{2.69}$$

Combining the last two bounds with (2.78) thus yields:

$$\mathcal{A}_{rob} \vdash_{\overline{4s}} \left\{ w'(C_r)^2 \left(v^\top \Sigma(w) v^\top \right)^s \leqslant \frac{1}{\delta^{2s}} (Cs)^s \left(v^\top \Sigma(r) v \right)^s + C\delta \left(v^\top \Sigma(w) v^\top \right)^s \right\} \tag{2.70}$$

□

Finally, we must translate the rough spectral upper bounds we had in Lemma 2.3.11. Yet again, the proof goes through essentially with only syntactic changes.

Lemma 2.4.12 (Rough Spectral Upper bound on $\Sigma(w)$)**.**

$$\mathcal{A}_{rob} \vdash_{\overline{4s}} \left\{ \left(v^\top \Sigma(w) v^\top \right)^s \leqslant (2Ck)^{s+1} (Cs)^s \sum_{r \leqslant k} \left(v^\top \Sigma(r) v \right)^s \right\} \tag{2.71}$$

Proof. For ease of exposition, we drop the variable and degree specifications since they are clear from context. As before, we start by invoking our constraints to conclude:

$$\mathcal{A}_{rob} \vdash \Bigg\{ \tau^{2s} \sum_{r \leqslant k} w'(C_r)^2 \left(v^\top \Sigma(w) v^\top \right)^s$$
$$\leqslant \frac{k^2}{n^2} \sum_{r \leqslant k} \sum_{i,j \in C_r} w'_i w'_j \left\langle \frac{1}{\sqrt{2}} \left(x'_i - x'_j \right), v \right\rangle^{2s} + \tau^{2s} \frac{k^2}{n^2} \sum_{r \leqslant k} \sum_{i \neq j \in C_r} w'_i w'_j q^2_{\tau, \Sigma(w)} \left(\frac{1}{\sqrt{2}} \left(x'_i - x'_j \right), v \right) \Bigg\} \tag{2.72}$$

We invoke Lemma 2.4.5 to conclude:

$$\mathcal{A}_{rob} \vdash \Bigg\{ \tau^{2s} \sum_{r \leqslant k} w'(C_r)^2 \left(v^\top \Sigma(w) v^\top \right)^s$$
$$\leqslant \frac{k^2}{n^2} \sum_{r \leqslant k} \sum_{i,j \in C_r} w'_i w'_j \left\langle \frac{1}{\sqrt{2}} \left(x_i - x_j \right), v \right\rangle^{2s} + \tau^{2s} \frac{k^2}{n^2} \sum_{r \leqslant k} \sum_{i \neq j \in C_r} w'_i w'_j q^2_{\tau, \Sigma(w)} \left(\frac{1}{\sqrt{2}} \left(x_i - x_j \right), v \right) \Bigg\} \tag{2.73}$$

The second term on the RHS can be upper bounded just as in the proof of Lemma 2.3.7 to yield:

$$\begin{aligned} \mathcal{A}_{rob} \vdash \Bigg\{ & \frac{k^2}{n^2} \sum_{r \leqslant k} \sum_{i \neq j \in C_r} w'_i w'_j q^2_{\tau, \Sigma(w)} \left(\left\langle \frac{1}{\sqrt{2}} \left(x_i - x_j \right), v \right\rangle \right) \\ & \leqslant \frac{k^2}{n^2} \sum_{i \neq j \in [n]} w'_i w'_j q^2_{\tau, \Sigma(w)} \left(\left\langle \frac{1}{\sqrt{2}} \left(x_i - x_j \right), v \right\rangle \right) \\ & \leqslant C\tau \left(v^\top \Sigma(w) v \right)^s \Bigg\} \end{aligned} \tag{2.74}$$

The first term can be also be upper bounded - this time in terms of the Covariances of all the k components.

$$\begin{aligned} \mathcal{A}_{rob} \vdash \Bigg\{ \frac{k^2}{n^2} \sum_{r \leqslant k} \sum_{i,j \in C_r} w'_i w'_j \left\langle \frac{1}{\sqrt{2}} \left(x_i - x_j \right), v \right\rangle^{2s} & \leqslant \sum_{r \leqslant k} \frac{k^2}{n^2} \sum_{i,j \in C_r} \left\langle \frac{1}{\sqrt{2}} \left(x_i - x_j \right), v \right\rangle^{2s} \\ & = (Cs)^s \sum_{r \leqslant k} \left(v^\top \Sigma(r) v \right)^s \Bigg\} \end{aligned} \tag{2.75}$$

We can now combine the two estimates above to yield:

$$\mathcal{A}_{rob} \vdash \left\{ \left(\sum_{r \leqslant k} w'(C_r)^2 - C\tau \right) \left(v^\top \Sigma(w) v^\top \right)^s \leqslant \frac{1}{\tau^{2s}} (Cs)^s \sum_{r \leqslant k} \left(v^\top \Sigma(r) v \right)^s \right\} \tag{2.76}$$

So far the argument closely follows the proof of Lemma 2.3.7. We now observe (note the change in the bound compared to the proof of Lemma 2.3.11)

$$\mathcal{A}_{rob} \vdash \left\{ \sum_{r \leqslant k} w'(C_r)^2 \geqslant \frac{1}{k} \left(\sum_{r \leqslant k} w'(C_r) \right)^2 \right\}.$$

Now,

$$\begin{aligned} \mathcal{A}_{rob} \vdash \left\{ \left(\sum_{r \leqslant k} w'(C_r) \right)^2 = \left(\frac{k}{n} \sum_{i \leqslant n} w_i m'_i \right)^2 &= \left(\frac{k}{n} \sum_{i \leqslant n} w_i \right)^2 - \left(\frac{k}{n} \sum_{i \leqslant n} w_i (1 - m'_i) \right)^2 \\ &\geqslant \left(\frac{k}{n} \sum_{i \leqslant n} w_i \right)^2 - \left(\frac{k}{n} \sum_{i \leqslant n} (1 - m'_i) \right)^2 \\ &\geqslant \left(\frac{k}{n} \sum_{i \leqslant n} w_i \right)^2 - k^2 \epsilon^2 \\ &\geqslant 1 - k^2 \epsilon^2 \Bigg\}. \end{aligned} \tag{2.77}$$

Thus,

$$\mathcal{A}_{rob} \vdash \left\{ \sum_{r \leqslant k} w'(C_r)^2 \geqslant \frac{1}{k} \left(\sum_{r \leqslant k} w'(C_r) \right)^2 \geqslant 1/k - k\epsilon^2 \right\}.$$

Thus, as long as $\tau \ll \frac{1}{2Ck}$, and $\epsilon < 1/k$ we can derive :

$$\mathcal{A}_{rob} \vdash \left\{ \left(v^\top \Sigma(w) v^\top \right)^s \leqslant k^{s+1} (Cs)^s \sum_{r \leqslant k} \left(v^\top \Sigma(r) v \right)^s \right\} \tag{2.78}$$

This concludes the proof. □

The argument for combining the upper and lower-bounds above proceeds exactly the same as in Section 2.3.

Proxy Intersection Bounds from Mean and Relative Frobenius Separation. The proof of the other two intersection bounds follows via similar strategy yielding:

Lemma 2.4.13 (Simultaneous Proxy Intersection Bounds from Mean Separation)**.** *Suppose there exists a* $v \in \mathcal{R}^d$ *such that* $\langle \mu(r) - \mu(r'), v \rangle_2^2 \geqslant \Delta_m^2 v^\top (\Sigma(r) + \Sigma(r'))\, v$.

Then, whenever $\Delta_m \gg Cs/\delta$,

$$\mathcal{A}_{rob} \vdash^{w'}_{O(\log(\kappa)/\delta^4)} \left\{ w'(C_r) w'(C_{r'}) \leqslant O(\sqrt{\delta}) \right\} .$$

For the special case of $k = 2$, *whenever* $\Delta_m \gg \Theta(1)$,

$$\mathcal{A}_{rob} \vdash^{w'}_{O(s)} \left\{ w'(C_1) w'(C_2) \leqslant O(\sqrt{\delta}) \right\} .$$

Lemma 2.4.14 (Simultaneous Proxy Intersection Bounds from Relative Frobenius Separation)**.** *Suppose* $\left\| \Sigma(r')^{-1/2} \Sigma(r) \Sigma(r')^{-1/2} - I \right\|_F^2 \geqslant \Delta_{cov}^2 \left(\left\| \Sigma(r')^{-1/2} \Sigma(r)^{1/2} \right\|_{op}^4 \right)$ *for* $\Delta_{cov} \gg C/\delta^2$. *Then,*

$$\mathcal{A}_{rob} \vdash^{w'}_{O(\log(\kappa)/\delta^2)} \left\{ w'(C_r) w'(C_{r'}) \leqslant O(\delta^{1/3}) \right\} .$$

For the special case of $k = 2$, *we have:*

$$\mathcal{A}_{rob} \vdash^{w'}_{O(s)} \left\{ w'(C_1) w'(C_2) \leqslant O(\delta^{1/3}) \right\} .$$

Combining the above three bounds yields Lemma 2.4.2.

2.5 Fully Polynomial Algorithm via Recursive Partial Clustering

In this section, we describe our fully polynomial time algorithm and prove Theorem 53.

Theorem 61 (Precise form of Theorem 53)**.** *Let* $\eta, \epsilon \leqslant k^{-\Omega(k)}$. *Let* $\Delta \geqslant poly(\eta/2^k)^k$. *There exists an algorithm that takes input a set of* n *points* $Y \subseteq \mathbb{Q}^d$ *and runs in time* $n^{k^{O(k)} poly \log(1/\eta)/\eta^2}$ *with the following guarantees: Let* X *be an i.i.d. sample from* Δ*-separated mixture of* k *reasonable distributions* $\{\mathcal{D}_r\}_{r \leqslant k}$ *with parameters* $\{\mu(r), \Sigma(r)\}_{r \leqslant k}$ *with true clusters* $C_1, C_2, \ldots, C_k$ *of size* n/k *each. If* Y *is an* ϵ*-corruption of* X, *then with probability* $\geqslant 0.99$ *over the draw of* X

and its random choices, the algorithm outputs a clustering $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_k$ of Y such that there exists a permutation $\pi : [k] \to [k]$ satisfying:

$$\min_{i \leqslant k} \frac{k}{n} |\hat{C}_i \cap C_{\pi(i)}| \geqslant 1 - O(k^{O(k)}(\eta + \epsilon)) \,.$$

Discussion In Section 2.4, we proved that our simple rounding (Algorithm 60) of any pseudo-distribution $\tilde{\zeta}$ of degree $\geqslant O(\log(\kappa)\text{poly}(k/\eta))$ consistent with $\mathcal{A}_{rob}$ produces an approximately correct clustering of any ϵ-corruption Y of a good sample X. In this section, we will establish two somewhat curious technical facts about Algorithm 60 and the constraints $\mathcal{A}_{rob}$ to show Theorem 53.

1. *All is not lost in constant degree* (Lemma 2.5.2). When the rounding in Algorithm 60 is run on a pseudo-distribution $\tilde{\zeta}$ of degree $\text{poly}(k/\eta)$ consistent with $\mathcal{A}_{rob}$, it still contains non-trivial information about the true clusters and in particular can be used to construct a *partial clustering*.
2. *Verification can be done in constant degree* (Lemma 2.5.3). While we cannot show that degree $\text{poly}(k/\eta)$ is enough to *find* a clustering, we will prove that it is enough to *verify* a purported approximate clustering.

These facts let us use a slightly more complicated recursive clustering algorithm combined with a verification subroutine to obtain an outlier-robust clustering algorithm with no dependence on the spread κ in the running time.

Algorithm. Our algorithm is the following recursive clustering subroutine that we invoke with the input corrupted sample Y and outlier parameter ϵ. The base case of the recursion uses a verification subroutine that confirms if a subset of n/k samples is close to a true cluster. The main recursive step employs the exact same rounding of the pseudo-distribution that we used in Algorithm 60.

Algorithm 62 (Recursive Partial Clustering)**.**

Given: *A subsample $Y' \subseteq Y$ of size jn/k for $j \in [k]$. A outlier parameter $\tau > 0$ and an accuracy parameter $\eta > 0$.*

Output: *A partition of Y' into an approximately correct clustering $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_j$.*

Operation:

1. ***Base Case:*** *If $|Y'| = n/k$, accept if verification subroutine from Algorithm 63 when run on Y' with outlier parameter τ accepts. Otherwise output fail.*

2. ***SDP Solving:*** *Find a degree poly(j/η) pseudo-distribution $\tilde{\zeta}$ satisfying $\mathcal{A}_{rob}$, and minimizing $\left\|\tilde{\mathbb{E}}[w]\right\|_2^2$ with number of components set to j and outlier parameter set to τ. If no such pseudo-distribution exists, output fail.*

3. ***Rounding:*** *Let $M = \tilde{\mathbb{E}}_{w\sim\tilde{\zeta}}[ww^\top]$.*

   (a) *Choose a uniformly random row i of M.*

   (b) *Choose $\ell = O(k\log(k/\eta))$ rows of M uniformly at random and independently.*

   (c) *For each $i \leqslant \ell$, let $\hat{C}_i$ be the indices of the columns j such that $M(i,j) \geqslant \eta/poly(k)$.*

   (d) *Let $\hat{C}_{\ell+1} = [n] \setminus \cup_{i\leqslant\ell}\hat{C}_i$.*

4. ***Brute-Force Search Over Partial Clusterings:*** *For each subset $S \subseteq [\ell+1]$, recursively run two instances of Algorithm 62 with inputs $\cup_{i\in S}\hat{C}_i$, $\cup_{i\notin S}\hat{C}_i$ respectively with outlier parameters $\eta + O(k^3\tau)$ for both runs.*

5. *If either run fails, output fail and return. Otherwise output the union of clusters returned by the two runs of the algorithm.*

Analysis of Algorithm. The analysis of our algorithm is based on the following two key pieces. The first shows that Algorithm 60, when run with a pseudo-distribution $\tilde{\zeta}$ of degree poly(k/η) consistent with $\mathcal{A}_{rob}$ recovers a *partial* clustering of the input sample. An (approximate) partial clustering is a non-trivial split of Y into (approximate) unions of clusters.

Definition 2.5.1 (Partial Clustering)**.** *A τ-approximate partial clustering of $Y = C_1 \cup C_2 \cup \ldots C_k \subseteq \mathcal{R}^d$ described by a partition of Y into $P_1 \cup P_2$ such that there exists $S \subseteq [k]$, $0 < |S| < k$ satisfying $\frac{|P_1\cap\cup_{i\in S}C_i|}{|\cup_{i\in S}C_i|}, \frac{|P_2\cap\cup_{i\notin S}C_i|}{|\cup_{i\notin S}C_i|} \geqslant 1-\tau$.*

The following lemma analyzes the output of Algorithm 62 when run with a τ-corrupted mixture of $k' \leqslant k$ reasonable distributions. We will use it to analyze all instantiations of Algorithm 62.

Lemma 2.5.2 (Outlier-Robust Partial Cluster Recovery)**.** *Let X be a good sample from a Δ-separated mixture of reasonable distributions with parameters $\{\mu(r), \Sigma(r)\}_{r\leqslant k}$ and true clusters*

$C_1, C_2, \ldots, C_{k'}$ of size $\frac{n}{k}$ each. Let Y be a τ-corruption of X. Then, whenever $\Delta \geqslant \text{poly}(\eta/k')^{k'}$, Algorithm 62 with probability at least $1 - 2^{-\Omega(k)}$ recovers a clustering $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_{k'}$ such that there exists a partition $G_S \cup G_L = [k]$ such that for $P_1 = \cup_{j \in G_S} \hat{C}_j$ and $P_2 = \cup_{j \in G_L} \hat{C}_j$ form a $(\eta + O(k^4\tau))$-approximate partial clustering of Y.

The next step is a *verification subroutine* that, in polynomial (degree depending only on k, η) time verifies if a given subset of n/k samples intersects in a true cluster in $(1 - \tau)$ fraction of points.

Lemma 2.5.3 (Verification Subroutine)**.** *Let X be a good sample from a Δ-separated mixture of reasonable distribution with parameters $\{\mu(r), \Sigma(r)\}_{r \leqslant k}$ and equal-size true clusters $C_1, C_2, \ldots, C_k$. Let Y be a τ-corruption of X, for $\tau \ll 1/k^6$. Let $\hat{C} \subseteq Y$ be such that $\max_{j \leqslant k} \frac{k}{n}|\hat{C} \cap C_j| < 1 - 2k\sqrt{\tau}$. Then, Algorithm 63 rejects on input $\hat{C}$. On the other hand, if $\exists r \leqslant k$ such that $\frac{k}{n}|\hat{C} \cap C_r| \geqslant 1 - \tau$, Algorithm 63 accepts on input $\hat{C}$.*

We can complete the analysis of Algorithm 62 and prove Theorem 61 using the above results:

Proof of Theorem 61. We run Algorithm 62 with input Y and initial outlier parameter $\tau = \epsilon$. Let's track the outlier parameters in the recursive calls - in each recursive call, $\tau \to \eta + O(k\tau)$. Since the depth of our recursive calls is at most k, $\tau = O(k^k\eta + k^{3k}\epsilon)$ throughout the algorithm, and thus $\epsilon \ll 1/k^{3k}$.

Let's bound the running time of the algorithm. The base case requires running the verification algorithm that needs $n^{\text{poly}(1/\eta)}$ time, and in the worst case, the fraction of outliers is $\tau = k^{O(k)}(\eta + \epsilon)$. Each run of the algorithm makes at most 2^k recursive calls to instances with number of components reduced by at least 1 and needs to solve an SDP that needs $n^{\text{poly}(k/\eta)}$ time. Thus, the running time follows the recurrence: $T(j) \leqslant 2^k T(j-1) + n^{\text{poly}(j/\eta)}$ and we can conclude that $T(k) \leqslant 2^{k^2} T(1) = 2^{k^2} n^{\text{poly}(k/\eta)}$.

Finally, let's confirm the correctness of the procedure. First, we show that if the algorithm doesn't fail, then it outputs a correct approximate clustering $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_k$ of Y. It's immediate that the algorithm always produces a partitioning of Y into subsets of size n/k each. Further, each $\hat{C}_i$ must cause Algorithm 63 to accept (base case of Algorithm 62). From Lemma 2.5.3, it must hold for each i, $\hat{C}_i$ some cluster $C_{\pi(i)}$ in $1-\tau$ fraction of the n/k samples for $\tau = O(k^{2k}\eta + k^{2k}\epsilon)$. Finally, observe that if $\tau \ll 1/k$ then, $C_{\pi(i)} \neq C_{\pi(j)}$ for $i \neq j$. Thus, π must be a permutation of $[k]$. This finishes the proof.

What remains is to argue that when run with ϵ-corruption Y of a good sample X, Algo-

rithm 62 does not output fail with probability at least 0.99. For this, we need to exhibit a choice of $S \subset [k]$ for each recursive call for which the algorithm does not fail. Observe, our algorithm never outputs fail if the input Y' intersects $(1-\tau)$ fraction of samples in some union of true clusters. This is guaranteed by Lemma 2.5.2 with probability at least $1 - 2^{-\Omega(k)}$. By a union bound, this guarantee holds for the output of all rounding steps incurred by making the choices of S above with probability at least 0.99. Thus, we must arrive at subsets $\hat{C}_i$ that are $(1-\tau)$-intersecting with some true cluster for $\tau = k^{O(k)}(\eta + \epsilon) \ll 1/k^2$, where the last inequality holds when $\eta, \epsilon < 1/k^{O(k)}$. By the completeness of our verification subroutine (Lemma 2.5.3), all $\hat{C}_i$ produced via these choices cause the verification algorithm to accept. This completes the proof. □

2.5.1 Partial Cluster Recovery

In this section, we prove Lemma 2.5.2. The crux of the proof is the following intersection bound that finds a bipartition of clusters and proves that the simultaneous intersection of $\hat{C}$ (searched for in $\mathcal{A}_{rob}$ via w-variables) with the two pieces of the bipartition is small. Note that this gets us a *weaker* guarantee than the inter-cluster simultaneous intersection bounds proven in Sections 2.3 and 2.4 with the upshot that the degree of the SoS proof here does not depend on κ, the spread of the mixture.

Lemma 2.5.4 (Simultaneous Intersections Bounds Across Bipartition)**.** *Let X, Y be as in the setting of Lemma 2.5.2 with true clusters $C_1, C_2, \ldots, C_{k'}$ with $\eta = O(1/k')$ and $\Delta = \Delta_{rob}^{k'} = poly(k'/\eta)^{k'}$ where Δ_{rob} is the separation requirement in Lemma 2.4.2. There exists a partition $S \cup L = [k']$ such that $|S| < k'$ satisfying:*

$$\mathcal{A} \left|\frac{w}{poly(k/\eta)+poly(1/\delta)}\right. \left\{ \sum_{r \in S, r' \in L} w(C_r) w(C_{r'}) \leqslant O((k'^3)\delta^{1/3} + (k')^2\tau) \right\} .$$

Proof. We break the proof into two cases.

Case 1: No pair of clusters $C_r, C_{r'}$ is spectrally separated. In this case, for every direction v, either $v^\top \Sigma(i) v = 0$ for all $i \leqslant k'$ or $\frac{v^\top \Sigma(r) v}{v^\top \Sigma(r') v} \leqslant \Delta \leqslant (\text{poly}(k'/\eta))^{k'}$ for all r, r' . Thus, in particular, the spread $\kappa \leqslant \Delta$. Applying Lemma 2.4.2 and plugging in the upper bound on κ immediately yields that for every $1 \leqslant r < r' \leqslant k'$

$$\mathcal{A} \left|\frac{w}{O(k' \log(k'/\eta)/\delta^4)}\right. \left\{ \sum_{r \neq r'} w(C_r) w(C_{r'}) \leqslant O(k'\tau) + O((k')^2 \delta^{1/3}) \right\} .$$

Thus, in this case, we recover every cluster approximately and thus can set S and L to be any non-trivial partition (that is, both S and L are non-empty) and finish the proof.

Case 2: There exist r, r' such that C_r and $C_{r'}$ that are spectrally separated. Then there is a direction v such that $\Delta_{rob}^k v^\top \Sigma(r) v \leqslant v^\top \Sigma(r') v$. Consider an ordering of the true clusters along the direction v, renaming cluster indices if needed, such that $v^\top \Sigma(1) v \leqslant v^\top \Sigma(2) v \leqslant \ldots v^\top \Sigma(k') v$. Then, clearly, $v^\top \Sigma(k') v \geqslant \Delta_{rob} v^\top \Sigma(r) v$.

Let $j \leqslant k'$ be the largest integer such that $\Delta_{rob} v^\top \Sigma(j) v \leqslant v^\top \Sigma(j+1) v$. Observe that since we are in Case 2, such a j exists. Further, observe that since j is defined to be the largest index which incurs separation Δ_{rob}, all indices in $[j, k']$ have spectral bound at most Δ_{rob} and thus $\frac{v^\top \Sigma(k') v}{v^\top \Sigma(j) v} \leqslant \Delta_{rob}^{k'}$. Applying Lemma 2.4.2 with the above direction v to every $r < j$ and $r' \geqslant j$ and observing that the spread parameter κ in each case is at most $\frac{v^\top \Sigma(k') v}{v^\top \Sigma(j) v} \leqslant \Delta_{rob}^{k'}$ yields:

$$\mathcal{A} \vdash_{O(k'^2 \log(k'/\eta)/\delta^4)}^{w} \left\{ w(C_r) w(C_{r'}) \leqslant O(k'\tau + (k')^2 \delta^{1/3}) \right\} .$$

Adding up the above inequalities over all $r \leqslant j-1$ and $r' \geqslant j+1$ and taking $S = [j-1]$, $T = [k'] \setminus [j-1]$ yields the claim. □

We are now ready to prove Lemma 2.5.2.

Proof of Lemma 2.5.2. We will prove that whenever $\Delta \geqslant \Delta_{rob} = \mathrm{poly}(k/\eta)^k$, Algorithm 62, when run with input Y recovers a collection $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_\ell$ of subsets of indices such that there is a partition $S \cup L = [\ell]$, $0 < |S| < \ell$ satisfying:

$$\min \left\{ \frac{k}{n} |\hat{C}_i \cap \cup_{j \in S} C_j|, \frac{k}{n} |\hat{C}_i \cap \cup_{j \in L} C_j| \right\} \leqslant \eta + O(k^4 \tau) . \tag{2.79}$$

This suffices to complete the proof: Split $[\ell]$ into two groups G_S, G_L as follows. For each i, let $j = \arg\max_{r \in [\ell]} \frac{k}{n} |\hat{C}_i \cap C_r|$. If $j \in S$, add it to G_S, else add it to G_L. Observe that this process is well-defined. To see this, suppose $j \in S$. Let $j' \in L$. Then, using Equation (3.40) and that $\eta + O(k^4\tau) \ll 1/k$ and that $\frac{k}{n} |\hat{C}_i \cap \cup_{r \in S} C_r| \geqslant \frac{k}{n} |\hat{C}_i \cap C_j| \geqslant 1/k$, we have that: $\frac{k}{n} |\hat{C}_i \cap \cup_{j' \in L} C_{j'}| < 1/k$.

We are now ready to verify the first claim. The second follows immediately from the first. For each $i \in G_S$, we have that $\frac{k}{n} |\hat{C}_i \cap \cup_{j \in L} C_j| \leqslant \eta + O(k^4\tau)$. Adding up these inequalities for all $i \in S$ yields that $\frac{k}{n} |P_1 \cap \cup_{j \in L} C_j| \leqslant |S| (\eta + O(k^4\tau))$. Using that $|P_1| = |S| \frac{n}{k}$ and S, L form

a partition of $[k]$ completes the proof.

We now go ahead and establish (3.40). Let $\tilde{\zeta}$ be a pseudo-distribution satisfying $\mathcal{A}$ of degree $\mathrm{poly}(k/\eta)$. Let $M = \tilde{\mathbb{E}}_{\tilde{\zeta}}[ww^\top]$. Reasoning similarly as in the proof of Theorem 56, we have:

1. $1/k \geqslant M(i,j) \geqslant 0$ for all i, j,
2. $M(i,i) = 1/k$ for all i,
3. $\mathbf{E}_{j\sim[n]} M(i,j) = \frac{1}{k^2}$ for every i.

For an η' to be chosen later, call an entry of M large if it exceeds η'/k^2. For each i, let B_i be the set of large entries in row i of M. Then, using (3) and (1) above gives that $|B_i| \geqslant (1 - k\eta')n/k$ for each $1 \leqslant i \leqslant n$. Next, call a row i "good" if $\frac{k}{n}\min\{|\cup_{r\in L} C_r \cap B_i|, |\cup_{r'\in S} C_{r'} \cap B_i|\} \leqslant 100k^2\eta' + O(k^3\tau)$. Let us estimate the fraction of rows of M that are good.

Towards that goal, let's apply Lemma 2.5.4 with $\eta = \eta'/2k$ and $\delta = \eta'^3/8k^6$. Then, using Fact 3.2.18, we obtain

$$\begin{aligned}\sum_{r\in S, r'\in L} \mathbf{E}_{i\in C_r}\mathbf{E}_{j\in C_{r'}} M(i,j) &\leqslant \sum_{r'\neq r} \mathbf{E}_{i\in C_r}\mathbf{E}_{j\in C_{r'}} \tilde{\mathbb{E}}[w_i w_j] \\ &= \tilde{\mathbb{E}}[w(C_r)w(C_{r'})] \\ &\leqslant \eta' + O(k^2\tau)\end{aligned}$$

Using Markov's inequality $1 - 1/100k^2$ over the uniformly random choice of i, $\mathbf{E}_{j\in C_{r'}} M(i,j) \leqslant 100k^2\eta' + O(k^4\tau)$. Thus, $1 - 1/100k^2$ fraction of the rows of M are good.

Next, let R be the set of $100k\log k/\eta'$ rows sampled in the run of the algorithm and set $\hat{C}_i = B_i$ for every $i \in R$. The probability that all of them are good is then at least $(1 - 1/100k^2)^{k\log k/\eta'} \geqslant 1 - \eta'\log k/100k$. Let's estimate the probability that $|\cup_{i\in R} \hat{C}_i| \geqslant (1 - 1/k^{10})n$. The chance that a given point $t \in B_i$ for a uniformly random B_i is at least $(1 - k\eta')/k$. Thus, the chance that $t \notin \cup_{i\in R} B_i$ is at most $(1 - 1/2k)^{100k\log k/\eta'} \leqslant \eta'/k^{50}$. Thus, the expected number of t that are not covered by $\cup_{i\in R}\hat{C}_i$ is at most $n\eta'/k^{50}$. Thus, by Markov's inequality, with probability at least $1 - 1/k^{10}$, $1 - \eta'/k^{40}$ fraction of t are covered in $\cup_{i\in R}\hat{C}_i$.

Let's now condition on the events that 1) each of the $100k\log k/\eta'$ rows R sampled are good and 2) $|\cup_{i\in R} \hat{C}_i| \geqslant (1 - \eta'/k^{40})n$. By the above computations and a union bound, this event happens with probability at least $1 - \eta'/k^{10}$. Let $\hat{C}_{\ell+1} = [n] \setminus \cup_{i\leqslant\ell}\hat{C}_i$ be the set of indices that

are not covered in $\cup_{i\in R}\hat{C}_i$. Then, $\cup_{i\leqslant \ell+1}\hat{C}_i$ is a partition of $[n]$.

We will show that the following way of grouping this partition into two buckets: $R_L = R \cap \cup_{i\in L} C_i$ and $R_S = R \setminus R_L$ satisfies the requirements of the lemma. To see this, note that

$$|\cup_{i\in R_L}\hat{C}_i \cap \cup_{i\in S} C_i| \leqslant n/k 100k^3\eta' + O(k^4\tau).$$

Similarly,

$$\begin{aligned}|\cup_{i\in R_S}\hat{C}_i \cup P \cap \cup_{i\in L} C_i| &\leqslant n/k 100k^3\eta' + |P| \\ &\leqslant n/k(100k^3\eta' + \eta' k^{-40}) + O(k^4\tau).\end{aligned}$$

Setting $\eta' \leqslant \eta/k^{10}$ completes the proof.

□

2.5.2 Verification Algorithm

In this section, we prove Lemma 2.5.3. We first describe our verification algorithm that involve computing (if one exists) a pseudo-distribution consistent with a system of constraints that verifies the properties of being close to a reasonable distribution for a given input subset $\hat{C}$ of size n/k of Y.

We first describe the verification constraint system $\mathcal{V} = \mathcal{V}(\hat{C}) = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \mathcal{V}_3 \cup \mathcal{V}_4 \cup \mathcal{V}_5$ that is closely related to those used in Sections 2.3 and 2.4. Covariance constraints introduce a matrix valued indeterminate intended to be the square root of Σ.

$$\text{Covariance Constraints: } \mathcal{V}_1 = \left\{ \begin{array}{c} \Pi = UU^\top \\ \Pi^2 = \Sigma \end{array} \right\} \tag{2.80}$$

The intersection constraints force that X' be close to X.

$$\text{Intersection Constraints: } \mathcal{V}_2 = \begin{Bmatrix} \forall i \in [n'], & m_i^2 = m_i \\ & \sum_{i\in[n']} m_i = (1-\tau)n' \\ \forall i \in [n'], & m_i(y_i - x_i') = 0 \end{Bmatrix} \tag{2.81}$$

The parameter constraints create indeterminates to stand for the covariance Σ and mean μ of $\hat{C}$

(indicated by m).

$$\text{Parameter Constraints: } \mathcal{V}_3 = \left\{ \begin{array}{c} \frac{1}{n'} \sum_{i=1}^{n'} m_i \left(x'_i - \mu\right) \left(x'_i - \mu\right)^\top = \Sigma \\ \\ \frac{1}{n'} \sum_{i=1}^{n'} m_i x'_i = \mu \end{array} \right\} \tag{2.82}$$

Finally, we enforce certifiable anti-concentration and hypercontractivity of $\hat{C}$. Certifiable Anti-Concentration :

$$\mathcal{V}_4 = \left\{ \quad \frac{1}{n'^2} \sum_{i,j=1}^{n'} m_i m_j q^2_{\tau/C, 2\Sigma} \left(\left(x'_i - x'_j\right), v \right) \leqslant 2^s \tau \left(v^\top \Sigma v \right)^s \quad \right\} \tag{2.83}$$

where $s = O(C^2/\tau^2)$, and C is the certifiable hypercontractivity constant.

Certifiable Hypercontractivity:

$$\mathcal{V}_5 = \left\{ \begin{array}{l} \forall h \leqslant 2s, \quad \frac{k^2}{n^2} \sum_{i,j \leqslant n} m_i m_j \left(Q(x'_i - x'_\ell) - \frac{k^2}{n^2} \sum_{i,\ell \leqslant n} m_i m_\ell Q(x'_i - x'_\ell) \right)^{2h} \\ \qquad\qquad \leqslant (Ch)^{2h} \left(\frac{k^2}{n^2} \sum_{i,\ell \leqslant n} m_i m_\ell \left(Q(x'_i - x'_\ell) - \frac{k^2}{n^2} \sum_{i,\ell \leqslant n} m_i m_\ell Q(x'_i - x'_\ell) \right)^2 \right)^h \end{array} \right\}, \tag{2.84}$$

and

Certifiable Bounded Variance: $\mathcal{V}_6$ =

$$\left\{ \forall j \leqslant 2s, \quad \frac{k^2}{n^2} \sum_{i,\ell \leqslant n} m_i m_\ell \left(Q(x'_i - x'_\ell) - \frac{k^2}{n^2} \sum_{i,\ell \leqslant n} m_i m_\ell Q(x'_i - x'_\ell) \right)^2 \leqslant C \left\| \Pi Q \Pi \right\|_F^2 . \right\}, \tag{2.85}$$

where $s = O(C^2/\tau^2)$.

Algorithm 63 (Verification Subroutine)**.**

Given: *A purported cluster $Y = \hat{C}$ of size $n' = \frac{n}{k}$, outlier parameter τ.*

Output: *Accept or Reject.*

Operation: *Accept iff $\exists$ a pseudo-distribution $\tilde{\zeta}$ of degree $O(C^2/\tau^2)$ consistent with $\mathcal{V}(\hat{C})$.*

Analysis of Verification Subroutine Let $m'_i = m_i \cdot \mathbf{1}(y_i = x_i)$ for every i. Define $m'(C_i) = \frac{k}{n}\sum_{j \in C_i} m'_j$ for every i.

Our proof of Lemma 2.5.3 will rely on the following three lemmas that give a degree-$O(C^2/\tau^2)$ refutation of $\mathcal{V}(\hat{C})$ whenever $\hat{C}$ intersects at least two clusters appreciably. The proofs follow the same conceptual plan of combining an upper and lower bound on the variance of $v^\top \Sigma v$ as in Sections 2.3 and 2.4. The key difference, as we suggested earlier, is that the degree of the proof is a fixed constant (instead of growing with $\log \kappa$). The proof exploits the fact that in the verification setting, $\hat{C}$ is *not* a variable in our constraint system.

Lemma 2.5.5 (SoS Refutation from Simultaneous Intersection with Spectrally Separated Components)**.** *Let X be a good sample from a Δ-separated reasonable distribution with parameters $\{\mu(r), \Sigma(r)\}_{r \leqslant k'}$ and true clusters $C_1, C_2, \ldots, C_{k'}$ of size $\frac{n}{k}$ each. Let Y be a τ-corruption of X. Let $\hat{C} \subseteq Y$ be a subset of size $\frac{n}{k}$. Suppose $C_r, C_{r'}$ are Δ-spectrally separated and $\frac{k}{n}|\hat{C} \cap C_r|, \frac{k}{n}|\hat{C} \cap C_{r'}| \geqslant 2\sqrt{\tau}$. Then, whenever $\Delta \geqslant \frac{1}{\tau^6}$, Then,*

$$\left\{\mathcal{V}(\hat{C})\right\} \vdash_{O(C^2/\tau^2)} \{-1 \geqslant 0\} \ .$$

Lemma 2.5.6 (SoS Refutation from Simultaneous Intersection with Mean Separated Components)**.** *Let X be a good sample from a Δ-separated reasonable distribution with parameters $\{\mu(r), \Sigma(r)\}_{r \leqslant k'}$ and true clusters $C_1, C_2, \ldots, C_{k'}$ of size $\frac{n}{k}$ each. Let Y be a τ-corruption of X. Let $\hat{C} \subseteq Y$ be a subset of size $\frac{n}{k}$. Suppose $C_r, C_{r'}$ are Δ-mean separated and $\frac{k}{n}|\hat{C} \cap C_r|, \frac{k}{n}|\hat{C} \cap C_{r'}| \geqslant 2\sqrt{\tau}$. Then, whenever $\Delta \geqslant \frac{1}{\tau^6}$, Then,*

$$\left\{\mathcal{V}(\hat{C})\right\} \vdash_{O(C^2/\tau^2)} \{-1 \geqslant 0\} \ .$$

Lemma 2.5.7 (SoS Refutation from Simultaneous Intersection with Frobenius Separated Components)**.** *Let X be a good sample from a Δ-separated reasonable distribution with parameters $\{\mu(r), \Sigma(r)\}_{r \leqslant k'}$ and true clusters $C_1, C_2, \ldots, C_{k'}$ of size $\frac{n}{k}$ each. Let Y be a τ-corruption of X. Let $\hat{C} \subseteq Y$ be a subset of size $\frac{n}{k}$. Suppose $C_r, C_{r'}$ are Δ_{cov}-relative Frobenius separated and $\frac{k}{n}|\hat{C} \cap C_r|, \frac{k}{n}|\hat{C} \cap C_{r'}| \geqslant 2\sqrt{\tau}$. Then, whenever $\Delta_{cov} \geqslant \frac{1}{\tau^6}$, Then,*

$$\left\{\mathcal{V}(\hat{C})\right\} \vdash_{O(C^2/\tau^2)} \{-1 \geqslant 0\} \ .$$

Proof of Lemma 2.5.3. Let j be the maximizer of $|\hat{C} \cap C_r|$ over all $r \leqslant k'$. Then, $|\hat{C} \cap C_j| \geqslant 1/k$. Let j' be the maximizer of $|\hat{C} \cap C_r|$ over all $r \neq j$. Then, $|\hat{C} \cap C_{j'}| \leqslant |\hat{C} \cap C_j|$. Then, observe that $\frac{k}{n}|\hat{C} \cap C_{j'}| \geqslant 2k\sqrt{\tau}/k \geqslant 2\sqrt{\tau}$.

Applying Lemmas 2.3.6, 2.5.6 and 2.5.7 for each of the three possible ways that C_i and C_j could be separated, we obtain that:

$$\mathcal{V} \left|\frac{}{O(C^2/\tau^2)}\right. \{-1 \geqslant 0\} .$$

This immediately implies that there's no degree $\geqslant \Omega(C^2/\tau^2)$ pseudo-distribution $\tilde{\zeta}$ consistent with $\mathcal{V}(\hat{C})$ -for if there was one, then the above inequality yields a contradiction. This completes the proof of the first part. For the second part, observe that setting X' to be the cluster closest (and thus $1-\tau$-intersecting) to $\hat{C}$ immediately completes the proof.

□

Sum-of-Squares Refutation of Reasonableness of Bad Clusters. We now prove Lemmas 2.5.5, 2.5.6 and 2.5.7. The proof of these lemmas closely resembles our proofs of the simultaneous intersection bounds in Sections 2.3 and 2.4. So it may appear somewhat confusing as to how we can get the SoS proofs to work in degrees that do not depend on κ. The key difference is that, informally speaking, here we already "know" that two clusters have large intersection with a purported bad cluster $\hat{C}$ (which is given to us, not a variable) and our goal is to obtain a contradiction from the axioms that $\hat{C}$ satisfies $\mathcal{V}$ in low-degree SoS. Such a difference, while inconsequential in "ordinary math", is key to obtaining the stronger degree bounds that do not depend on κ in this section.

We will use the following result in all the three proofs.

Lemma 2.5.8 (Matching with Original Uncorrupted Samples)**.** *Suppose* $\frac{1}{n'}|\hat{C} \cap C_r|, \frac{1}{n'}|\hat{C} \cap C_{r'}| \geqslant 2\sqrt{\tau}$. *Let* $m'(C_r) = \frac{1}{n'}\sum_{i \leqslant C_r} m'_i$. *Then,*

$$\mathcal{V}(\hat{C}) \left|\frac{}{O(C^2/\tau^2)}\right. \left\{ m'(C_r)^2 \geqslant \frac{k}{n}|C_r \cap \hat{C}| - 2\tau \geqslant 2\tau \right\} .$$

Proof. Reasoning as in Lemma 2.4.8, we obtain that for any subset $\hat{C}' \subseteq \hat{C}$, we have:

$$\mathcal{V}(\hat{C}) \left|\frac{m'}{O(C^2/\tau^2)}\right. \left\{ \frac{1}{n'}\sum_{i \in \hat{C}'} m'_i \geqslant |\hat{C}'| - 2\tau \right\} .$$

Applying this to subsets $\hat{C}' = \hat{C} \cap C_r$ yields:

$$\mathcal{V} \Big|_{\frac{m'}{O(C^2/\tau^2)}} \left\{ m'(C_r)^2 \geqslant \frac{k}{n} |C_r \cap \hat{C}| - 2\tau \geqslant 2\tau \right\}.$$

□

Proof of Lemma 2.5.5. WLOG, assume $\Delta v^\top \Sigma(r) v \leqslant v^\top \Sigma(r') v$ for some $v \in \mathcal{R}^d$. The proof follows by from combining certifiable anti-concentration constraints $\mathcal{V}_4$, certifiable anti-concentration of C_r and Lemma 2.5.8. We will use $\mathcal{V}$ to denote $\mathcal{V}(\hat{C})$ in the proof below.

Using certifiable anti-concentration of $C_{r'}$:

$$\mathcal{V} \Big|_{\frac{m'}{O(C^2/\tau^2)}} \left\{ (C\tau)^{2s} \left(m'(C_{r'})^2 - \tau \right) \left(v^\top \Sigma(r') v^\top \right)^s \leqslant \frac{1}{n'^2} \sum_{i,j \leqslant C_{r'}} m'_i m'_j \left\langle x'_i - x'_j, v \right\rangle^{2s} \leqslant \left(v^\top \Sigma(m) v \right)^s \right\} \tag{2.86}$$

Similarly, using certifiable anti-concentration constraints $\mathcal{V}_4$:

$$\mathcal{V} \Big|_{\frac{m'}{O(C^2/\tau^2)}} \left\{ \left(m'(C_r)^2 - \tau \right) \left(v^\top \Sigma(m) v^\top \right)^s \leqslant \left(\frac{1}{\tau^2} \right)^s \left(v^\top \Sigma(r) v \right)^s \right\} \tag{2.87}$$

Plugging in the estimates from Lemma 2.5.8 in (2.86) and (2.87), and rearranging yields:

$$\mathcal{V} \Big|_{\frac{m'}{O(C^2/\tau^2)}} \left\{ \tau^2 (C\tau)^{2s} \left(v^\top \Sigma(r') v^\top \right)^s \leqslant \tau \left(v^\top \Sigma(m) v \right)^s \leqslant \left(\frac{1}{\tau^2} \right)^s \left(v^\top \Sigma(r) v \right)^s \right\}.$$

Dividing throughout by $\left(v^\top \Sigma(r) v \right)^s$ yields:

$$\mathcal{V} \Big|_{O(C^2/\tau^2)} \left\{ \tau^2 (C\tau)^{4s} \Delta^s \leqslant 1 \right\}.$$

Using that $\Delta^s \geqslant 2 \frac{1}{\tau^6}$ and subtracting out 1 from both sides above yields:

$$\mathcal{V} \Big|_{O(C^2/\tau^2)} \{ -1 \geqslant 0 \}.$$

□

The proof of Lemma 2.5.6 follows via a similar argument as above. We now proceed to the proof of Lemma 2.5.7.

Proof of Lemma 2.5.7. As in the proof of Lemma 2.3.18, for the sake of the analysis, we first apply the linear transformation $y_i \to \Sigma(r')^{-1/2} y_i$. Let $Q = \Sigma(r) - I$.

From an argument similar to Lemma 2.3.18, we can obtain:

$$\mathcal{V}\Big|_{8}^{m'} \Big\{ 2\mathbf{E}_{X'} \left(Q - \mathbf{E}_{X'} Q\right)^2 + 2\mathbf{E}_{C_r} (Q - \mathbf{E}_{C_r} Q)^2 + 2\mathbf{E}_{C_{r'}} (Q - \mathbf{E}_{C_{r'}} Q)^2 \\ \geqslant m'(C_r)^2 m'(C_{r'})^2 \left\| \Sigma(r')^{-1/2} \Sigma(r) \Sigma(r')^{-1/2} - I \right\|_F^4 \Big\} \tag{2.88}$$

Reasoning as in Section 2.3.4, and using Lemma 2.3.2:

$$\begin{aligned} \mathbf{E}_{C_r} (Q - \mathbf{E}_{C_r} Q)^2 &\leqslant (C-1) \left\| \Sigma(r')^{-1/2} \Sigma(r)^{1/2} Q \Sigma(r)^{1/2} \Sigma(r')^{-1/2} \right\|_F^2 \\ &\leqslant \left\| \Sigma(r')^{-1/2} \Sigma(r)^{1/2} \right\|_{op}^2 \|Q\|_F^2 . \end{aligned}$$

Similarly, $\mathbf{E}_{C_{r'}} (Q - \mathbf{E}_{C_r} Q)^2 \leqslant \|Q\|_F^2$.

For the upper bound on $\mathbf{E}_{X'} \left(Q - \mathbf{E}_{X'} Q\right)^2$, our proof is similar to that of Lemma 2.3.19 but leverages the argument in the proof of Lemma 2.5.5 to obtain a degree bound independent of κ (without relying on the uniform polynomial approximator for the threshold):

From our bounded-variance constraints, we have:

$$\mathcal{A}\Big|_{4}^{\Pi, m} \left\{ \mathbf{E}_{X'} (Q - \mathbf{E}_{X'} Q)^2 \leqslant C \left\| \Pi Q \Pi \right\|_F^2 \right\} . \tag{2.89}$$

We will now apply Lemma 2.8.1 in order to bound the RHS above. Towards that, reasoning as in Lemma 2.5.5, we have:

$$\mathcal{A}\Big|_{O(C^2/\tau^2)} \left\{ \left(v^\top \Sigma(X') v \right)^s \leqslant \frac{1}{\tau^{2s+2}} \left(v^\top \Sigma(r) v \right)^s \right\} .$$

Substituting $v \to \Sigma(r')^{\dagger/2} v$ yields:

$$\mathcal{A}\Big|_{O(C^2/\tau^2)} \left\{ \left(v^\top \Sigma(r')^{\dagger/2} \Sigma(X') \Sigma(r')^{\dagger/2} v \right)^s \leqslant \frac{1}{\tau^{2s+2}} \left(v^\top \Sigma(r')^{\dagger/2} \Sigma(r) \Sigma(r')^{\dagger/2} v \right)^s \right\} .$$

Proceeding as in the proof of Lemma 2.3.19, we can now obtain:

$$\mathcal{A} \vdash_{O(C^2/\tau^2)} \left\{ \mathbf{E}_{X'}(Q - \mathbf{E}_{X'}Q)^{2s} \leqslant \frac{1}{\tau^{2s+2}} \left\| \Sigma(r')^{-1/2}\Sigma(r)\Sigma(r')^{-1/2} - I \right\|_F^2 \right\} . \tag{2.90}$$

Combining (2.88) and (2.90) and the SoS almost triangle inequality (Fact 2.2.8) we obtain:

$$\mathcal{V} \vdash_{O(C^2/\tau^2)} \left\{ m'(C_r)^{2s} m'(C_{r'})^{2s} \left\| \Sigma(r')^{-1/2}\Sigma(r)\Sigma(r')^{-1/2} - I \right\|_F^{4s} \right.$$
$$\left. \leqslant 2^{3s} \frac{1}{\tau^{2s+2}} \left\| \Sigma(r')^{-1/2}\Sigma(r)\Sigma(r')^{-1/2} - I \right\|_F^2 \right\}$$

Using the separation condition with the fact that $\Delta \geqslant \frac{3}{\tau^6}$ yields via an argument similar to that in the proof of Lemma 2.5.5:

$$\mathcal{V} \vdash_{O(C^2/\tau^2)} \{-1 \geqslant 0\} .$$

□

2.6 Outlier-Robust Covariance Estimation in Frobenius Distance

In this section, we give an outlier-robust algorithm for estimating covariances in relative Frobenius distance (i.e. Frobenius distance after putting one of the distribution in isotropic position).

Our stronger error bounds hold for distributions with certifiable hypercontractive degree 2 polynomials. This is a strictly stronger assumption (and thus a smaller class of distributions) than certifiable subgaussianity considered in [KSS18]. As pointed out in [KSS18] (see discussion in the last paragraph of page 6 for a simple counter-example), certifiable subgaussianity is provably insufficient to obtain the stronger relative Frobenius errors guarantees.

Our proof approach is similar to that of [KSS18] - the key difference being that we rely on certifiable hypercontractivity (instead of the weaker certifiable subgaussianity) and rely on an appropriate application of the contraction lemma (Lemma 2.8.1).

Theorem 64 (Robust Parameter Estimation for Certifiably Hypercontractive and Bounded-Variance Distributions)**.** *Fix an even $t \in \mathbb{N}$ and $\epsilon > 0$ small enough so that $Ct\epsilon^{1-1/t} \ll 1$*[7]*. There's an algorithm that takes input a B-bit rational truncation of an ϵ-corruption Y of a sample X of size $n \geqslant n_0 = d^{O(t)}/\epsilon^2$ from a $2t$-certifiably C-hypercontractive distribution with certifiably C-bounded variance with unknown mean μ_* and covariance Σ_* with entries of bit complexity B and in time $(Bn)^{O(t)}$ outputs an estimate $\hat{\mu}$ and $\hat{\Sigma}$ satisfying:*

1. $\left\|\Sigma^{-1/2}(\mu_* - \hat{\mu})\right\|_2 \leqslant O(Ct)^{1/2}\epsilon^{1-1/t}$,
2. $(1-\eta)\Sigma_* \preceq \hat{\Sigma} \preceq (1+\eta)\Sigma_*$ *for* $\eta \leqslant O(Ct)\epsilon^{1-1/t}$*, and,*
3. $\left\|\Sigma_*^{-1/2}\hat{\Sigma}\Sigma_*^{-1/2} - I\right\|_F \leqslant O\left(Ct\epsilon^{1-1/t}\right)$.

In particular, by choosing $t = O(\log(1/\epsilon))$ results in the error bounds of $\tilde{O}(\epsilon)$ in the inequalities above.

We consider the following system $\mathcal{A} := \mathcal{A}_{Y,\epsilon}$ of quadratic equations in scalar-valued variables $w_1, \ldots, w_n$ and vector-valued variables $x'_1, \ldots, x'_n$, where $\mathcal{A}_{Y,\epsilon} =$

$$\left\{\begin{array}{ll} & \sum_{i=1}^n w_i = (1-\epsilon)\cdot n \\ \forall i \in [n]. & w_i^2 = w_i \\ & \Pi = UU^\top \\ & \Pi^2 = \Sigma \\ \forall i \in [n]. & w_i \cdot (y_i - x'_i) = 0 \\ & \frac{1}{n}\sum_{i\leqslant n} x'_i = \mu \\ & \frac{1}{n}\sum_{i\leqslant n}(x'_i - \mu)(x'_i - \mu)^\top = \Sigma \\ & \frac{1}{n}\sum_{i\leqslant n}\left(Q(x'_i - \mu) - \frac{1}{n}\sum_{i\leqslant n} Q(x'_i - \mu)\right)^{2t} \leqslant (Ct)^{2t}\mathbf{Var}(Q) \\ & \frac{1}{n}\sum_{i\leqslant n}\left(Q(x'_i - \mu) - \frac{1}{n}\sum_{i\leqslant n} Q(x'_i - \mu)\right)^{2} \leqslant C\left\|\Pi Q \Pi\right\|_F^2 . \end{array}\right\} \tag{2.91}$$

where $\mathbf{Var}(Q) = \left(\frac{1}{n}\sum_{i\leqslant n}\left(Q(x'_i - \mu) - \frac{1}{n}\sum_{i\leqslant n} Q(x'_i - \mu)\right)^2\right)^t$.

[7]This notation means that we needed $Ct\epsilon^{1-1/t}$ to be at most c_0 for some absolute constant $c_0 > 0$

Algorithm 65 (Parameter Estimation Algorithm)**.**

Input: *B-bit truncation of an ϵ-corrupted sample $Y = \{y_1, \ldots, y_n\} \subseteq \mathbb{R}^d$ of a t-certifiably hypercontractive distribution D_0 over $\mathcal{R}^d$.*

Output: *Estimates $\hat{\mu}$ and $\hat{\Sigma}$.*

Operation:

1. *Find a level-$O(t)$ pseudo-distribution $\tilde{\zeta}$ that satisfies $\mathcal{A}_{Y,\epsilon}$.*
2. *Output estimates $\hat{\mu} = \tilde{\mathbb{E}}[\mu]$ and $\hat{\Sigma} = \tilde{\mathbb{E}}[\Sigma]$.*

Analysis of Algorithm Corollaries 4.6 and 4.7 in [KSS18] show the following low-degree sum-of-squares proofs of certifiability of mean and covariance under spectral distance. Let $\hat{\Sigma}_*$ be the covariance of the uncorrupted samples.

$$\mathcal{A}_{Y,\epsilon} \left|\frac{\Sigma,u}{O(t)}\right. \left\{(1-\eta)u^\top \Sigma_* u \leqslant \langle u, \Sigma u\rangle \leqslant (1+\eta)u^\top \Sigma_* u\right\}, \tag{2.92}$$

for some $\eta \leqslant O(Ct)\epsilon^{1-2/t}$, and,

$$\mathcal{A}_{Y,\epsilon} \left|\frac{\mu,u}{O(t)}\right. \left\{\langle u, \mu - \mu_*\rangle \leqslant \eta \langle u, \Sigma_* u\rangle^{1/2}\right\}, \tag{2.93}$$

for some $\eta = O(\sqrt{Ct}\epsilon^{1-1/t})$. The bit complexities of both these sum-of-squares proofs is poly$(Bn^t)$where B is the bit complexity of the entries of Σ_*.

We will rely on these to show:

Lemma 2.6.1 (Certifying Covariance Closeness in Frobenius Distance)**.** *For any $t \in \mathbb{N}$,*

$$\mathcal{A}_{Y,\epsilon} \left|\frac{\Sigma}{8t}\right. \left\{\left\|\hat{\Sigma}_*^{-1/2}\Sigma\hat{\Sigma}_*^{-1/2} - I\right\|_F^{8t} \leqslant \eta\right\}, \tag{2.94}$$

where $\eta = (Ct)^{8t}O\left(\epsilon^{8t-8}\right)$ and $\hat{\Sigma}_*$ is the covariance of the uncorrupted samples. Further, the bit complexity of the SoS proof above is poly(Bn^t).

We now conclude with proving the parameter proximity lemma:

Proof of Lemma 2.6.1. Let Σ_* be the covariance of the underlying distribution and $\hat{\Sigma}_*$ be the covariance of the uncorrupted samples.

In the following, we apply the the SoS Cauchy-Schwarz inequality (Fact 3.2.20) and guarantee for the mean estimation above (guarantee (2.93)), to obtain:

$$\begin{aligned}\mathcal{A}_{Y,\epsilon} \left|\frac{Q,\mu}{4t}\right. \Bigg\{ \left((\mu-\mu_*)^\top Q(\mu-\mu_*)\right)^{2t} &= \left\langle (\mu-\mu_*)(\mu-\mu_*)^\top, Q\right\rangle^{2t} \\ &= \left\langle \left(\hat{\Sigma}_*^{-1/2}(\mu-\mu_*)\right)\left(\hat{\Sigma}_*^{-1/2}(\mu-\mu_*)\right)^\top, \hat{\Sigma}_*^{1/2} Q \hat{\Sigma}_*^{1/2}\right\rangle^{2t} \\ &\leqslant \left\|\hat{\Sigma}_*^{-1/2}(\mu-\mu_*)\right\|_2^{2t} \left\|\Sigma_*^{1/2} Q \hat{\Sigma}_*^{1/2}\right\|_F^{2t} \\ &\leqslant (Ct)^{2t} O(\epsilon^{2t-2}) \left\|\hat{\Sigma}_*^{1/2} Q \hat{\Sigma}_*^{1/2}\right\|_F^{2t} \Bigg\}. \end{aligned} \tag{2.95}$$

where the last inequality follows from the mean closeness bound in (2.93). Next, observe

$$\begin{aligned}\mathcal{A}_{Y,\epsilon} \left|\frac{Q,\mu}{4}\right. \Bigg\{ &\left((x_i'-\mu_*)^\top Q(x_i'-\mu_*) - (x_i-\mu_*)^\top Q(x_i-\mu_*)\right) \\ &= \left\langle Q, (x_i'-\mu_*)(x_i'-\mu_*)^\top - (x_i-\mu_*)(x_i-\mu_*)^\top \right\rangle \\ &= (1-w_i)\left\langle Q, (x_i'-\mu_*)(x_i'-\mu_*)^\top - (x_i-\mu_*)(x_i-\mu_*)^\top \right\rangle \\ &\quad + \left\langle Q, w_i\left((x_i'-\mu_*)(x_i'-\mu_*)^\top - (x_i-\mu_*)(x_i-\mu_*)^\top\right)\right\rangle \Bigg\}. \end{aligned} \tag{2.96}$$

Let z_i be the indicator that x_i was not corrupted, i.e. $z_i(y_i - x_i) = 0$ and observe $\frac{1}{n}\sum_{i\in[n]} z_i = (1-\epsilon)$. Then,

$$\begin{aligned}\mathcal{A}_{Y,\epsilon} \left|\frac{\mu}{4}\right. \Bigg\{ & w_i\left((x_i'-\mu_*)(x_i'-\mu_*)^\top - (x_i-\mu_*)(x_i-\mu_*)^\top\right) \\ &= w_i(1-z_i)\left((x_i'-\mu_*)(x_i'-\mu_*)^\top - (x_i-\mu_*)(x_i-\mu_*)^\top\right) \\ &\quad + w_i z_i\left(x_i'x_i'^\top - x_i x_i^\top - (x_i'-x_i)\mu^\top - \mu(x_i-x_i)^\top\right) \\ &= w_i(1-z_i)\left((x_i'-\mu_*)(x_i'-\mu_*)^\top - (x_i-\mu_*)(x_i-\mu_*)^\top\right) \Bigg\}. \end{aligned} \tag{2.97}$$

Subsituting back into Equation (2.96) we can conclude

$$\mathcal{A}_{Y,\epsilon} \left| \frac{Q,\mu}{4} \right. \Bigg\{ (x_i' - \mu_*)^\top Q (x_i' - \mu_*) - (x_i - \mu_*)^\top Q (x_i - \mu_*) \\ = (1 - w_i + w_i - w_i z_i) \left((x_i' - \mu_*)^\top Q (x_i' - \mu_*) - (x_i - \mu_*)^\top Q (x_i - \mu_*) \right) \Bigg\} \tag{2.98}$$

Using that Σ is the covariance of X' and $\hat{\Sigma}_*$ is the covariance of the uncorrupted samples X, along with the SoS almost triangle inequality and the bound in (2.98) we have:

$$\begin{aligned} \mathcal{A} \left| \frac{\mu,w,Q}{4t} \right. \Bigg\{ & \left\langle \Sigma - \hat{\Sigma}_*, Q \right\rangle^{2t} \\ & = \left(\frac{1}{n} \sum_{i \leqslant n} \left((x_i' - \mu)^\top Q (x_i' - \mu) - (x_i - \mu_*)^\top Q (x_i - \mu_*) \right) \right)^{2t} \\ & \leqslant 2^{2t} \left(\frac{1}{n} \sum_{i \leqslant n} \left((x_i' - \mu_*)^\top Q (x_i' - \mu_*) - (x_i - \mu_*)^\top Q (x_i - \mu_*) \right) \right)^{2t} \\ & \quad + 2^{2t} (Ct)^{2t} O(\epsilon^{2t-2}) \left\| \Sigma_*^{1/2} Q \Sigma_*^{1/2} \right\|_F^{2t} \\ & \leqslant 2^{2t} \left(\frac{1}{n} \sum_{i \leqslant n} (1 - w_i z_i) \left((x_i' - \mu_*)^\top Q (x_i' - \mu_*) - (x_i - \mu_*)^\top Q (x_i - \mu_*) \right) \right)^{2t} \\ & \quad + 2^{2t} (Ct)^{2t} O(\epsilon^{2t-2}) \left\| \Sigma_*^{1/2} Q \Sigma_*^{1/2} \right\|_F^{2t} \\ & \leqslant 2^{4t} \left(\frac{1}{n} \sum_{i \leqslant n} (1 - w_i z_i) \left((x_i' - \mu_*)^\top Q (x_i' - \mu_*) - (x_i - \mu_*)^\top Q (x_i - \mu_*) - \left\langle Q, \Sigma - \hat{\Sigma}_* \right\rangle \right) \right)^{2t} \\ & \quad + 2^{4t} \left(\frac{1}{n} \sum_{i \leqslant n} (1 - w_i z_i) \left\langle Q, \Sigma - \hat{\Sigma}_* \right\rangle \right)^{2t} + 2^{2t} (Ct)^{2t} O(\epsilon^{2t-2}) \left\| \Sigma_*^{1/2} Q \Sigma_*^{1/2} \right\|_F^{2t} \Bigg\}. \end{aligned} \tag{2.99}$$

Observe, $\frac{1}{n} \sum_{i \leqslant n} (1 - w_i z_i) \left\langle Q, \Sigma - \hat{\Sigma}_* \right\rangle^{2t} = \epsilon^{2t} \left\langle Q, \Sigma - \hat{\Sigma}_* \right\rangle^{2t}$. Plugginng back into (2.99), and applying the SoS almost triangle inequality again,

$$\mathcal{A} \left|\frac{\mu,w,Q}{4t}\right. \Bigg\{ \left(1-(4\epsilon)^{2t}\right) \left\langle \Sigma - \hat{\Sigma}_*, Q \right\rangle^{2t}$$
$$\leqslant 2^{6t} \left(\frac{1}{n} \sum_{i\leqslant n} (1-w_i z_i) \left((x_i' - \mu_*)^\top Q (x_i' - \mu_*) - \langle Q, \Sigma \rangle \right) \right)^{2t}$$
$$+ (2Ct)^{2t} \epsilon^{2t-2} \left\| \Sigma_*^{1/2} Q \Sigma_*^{1/2} \right\|_F^{2t} + 2^{6t} \left(\frac{1}{n} \sum_{i\leqslant n} (1-w_i z_i) \left((x_i - \mu_*)^\top Q (x_i - \mu_*) - \left\langle Q, \hat{\Sigma}_* \right\rangle \right) \right)^{2t} \Bigg\}. \tag{2.100}$$

We bound each term above separately. Applying SoS Hölder's inequality to the first term, and using that $\mathcal{A}_{Y,\epsilon} \left|\frac{}{}\right. \{(1-w_i z_i)^2 = (1-w_i z_i)\}$, we obtain

$$\mathcal{A}_{Y,\epsilon} \left|\frac{\mu,w}{4t}\right. \Bigg\{ \left(\frac{1}{n} \sum_{i\leqslant n} (1-w_i z_i) \left((x_i' - \mu_*)^\top Q (x_i' - \mu_*) - \langle Q, \Sigma \rangle \right) \right)^{2t}$$
$$\leqslant \left(\frac{1}{n} \sum_{i\leqslant n} (1-w_i z_i)^{2t} \right)^{2t-1} \left(\frac{2^{2t}}{n} \sum_{i\leqslant n} \left((x_i' - \mu)^\top Q (x_i' - \mu) - \langle Q, \Sigma \rangle \right)^{2t} \right)$$
$$+ \left(\frac{1}{n} \sum_{i\leqslant n} (1-w_i z_i)^{2t} \right)^{2t-1} \left(2^{2t} \left((\mu - \mu_*)^\top Q (\mu - \mu_*) \right)^{2t} \right) \tag{2.101}$$
$$\leqslant \epsilon^{2t-1} 2^{2t} \left(\frac{1}{n} \sum_{i\leqslant n} \left((x_i' - \mu)^\top Q (x_i' - \mu) - \langle Q, \Sigma \rangle \right)^{2t} \right)$$
$$+ (2Ct)^{2t} O(\epsilon^{4t-3}) \left\| \Sigma_*^{1/2} Q \Sigma_*^{1/2} \right\|_F^{2t} \Bigg\},$$

where the last inequality follows from (2.95). Next,

$$\mathcal{A}_{Y,\epsilon} \left|\frac{\mu,w}{4t}\right. \Bigg\{ \frac{1}{n}\sum_{i\leqslant n}\Big((x_i'-\mu)^\top Q(x_i'-\mu) - \langle Q,\Sigma\rangle\Big)^{2t}$$
$$\leqslant (Ct)^{2t}\left(\frac{1}{n}\sum_{i\leqslant n}\left((x_i'-\mu)^\top Q(x_i'-\mu) - \frac{1}{n}\sum_{i\leqslant n}(x_i'-\mu)^\top Q(x_i'-\mu)\right)^2\right)^t \tag{2.102}$$
$$\leqslant (Ct)^{2t}\left\|\Pi Q\Pi\right\|_F^{2t}\Bigg\},$$

where in the second inenquality we use that $\mathcal{A}$ enforces the t-certifiable hypercontractivity of degree-2 polynomials of X' and in the third inequality, we invoked the bounded variance constraint. Combining the two equations above and substituting back into (2.101),

$$\mathcal{A}_{Y,\epsilon} \left|\frac{\mu,w}{4t}\right. \Bigg\{ \left(\frac{1}{n}\sum_{i\leqslant n}(1-w_iz_i)\Big((x_i'-\mu_*)^\top Q(x_i'-\mu_*) - \langle Q,\Sigma\rangle\Big)\right)^{2t}$$
$$\leqslant \epsilon^{2t-1}(2Ct)^{2t}\left\|\Pi Q\Pi\right\|_F^{2t} + (2Ct)^{2t}\epsilon^{4t-3}\left\|\Sigma_*^{1/2}Q\Sigma_*^{1/2}\right\|_F^{2t}\Bigg\}. \tag{2.103}$$

Similarly, we can bound $\left(\frac{1}{n}\sum_{i\leqslant n}(1-w_iz_i)(x_i-\mu_*)^\top Q(x_i-\mu_*) - \left\langle Q,\hat{\Sigma}_*\right\rangle\right)^{2t}$ using certifiable hypercontractivity and bounded variance of X (the samples from the true distribution) as follows:

$$\mathcal{A}_{Y,\epsilon} \left|\frac{\mu,w,Q}{4t}\right. \Bigg\{ \left(\frac{1}{n}\sum_{i\leqslant n}(1-w_iz_i)\Big((x_i-\mu_*)^\top Q(x_i-\mu_*) - \left\langle Q,\hat{\Sigma}_*\right\rangle\Big)\right)^{2t}$$
$$\leqslant \left(\frac{1}{n}\sum_{i\leqslant n}(1-w_iz_i)\right)^{2t-1}\left(\frac{1}{n}\sum_{i\leqslant n}\Big((x_i-\mu_*)^\top Q(x_i-\mu_*) - \left\langle Q,\hat{\Sigma}_*\right\rangle\Big)^{2t}\right)$$
$$\leqslant \epsilon^{2t-1}(2Ct)^{2t}\left\|\hat{\Sigma}_*^{1/2}Q\hat{\Sigma}_*^{1/2}\right\|_F^{2t}\Bigg\}. \tag{2.104}$$

Plugging (2.103) and (2.104) into (2.100) we get

$$\mathcal{A}_{Y,\epsilon} \vdash^{\Sigma,Q}_{4t} \left\{ \left\langle \Sigma - \hat{\Sigma}_*, Q \right\rangle^{2t} \leqslant \frac{\epsilon^{2t-1}(32Ct)^{2t}}{1-(4\epsilon)^{2t}} \left(\left\| \hat{\Sigma}_*^{1/2} Q \hat{\Sigma}_*^{1/2} \right\|_F^{2t} + \|\Pi Q \Pi\|_F^{2t} \right) + \frac{(2Ct)^{2t}\epsilon^{2t-2}}{1-(4\epsilon)^{2t}} \left\| \Sigma_*^{1/2} Q \Sigma_*^{1/2} \right\|_F^{2t} \right\}. \tag{2.105}$$

Next, observe for $n \geqslant d^2$, $0.99\Sigma_* \preceq \hat{\Sigma}_* \preceq 1.01\Sigma_*$, and using the sub-multiplicativity of the Frobenius norm, $\left\| \Sigma_*^{1/2} Q \Sigma_*^{1/2} \right\|_F \leqslant 2 \left\| \hat{\Sigma}_*^{1/2} Q \hat{\Sigma}_*^{1/2} \right\|_F$. Then, substituting $Q = \hat{\Sigma}_*^{-1/2} Q \hat{\Sigma}_*^{-1/2}$ and using cyclicity of trace,

$$\mathcal{A}_{Y,\epsilon} \vdash^{\Sigma,Q}_{4t} \left\{ \left\langle \hat{\Sigma}_*^{-1/2} \Sigma \hat{\Sigma}_*^{-1/2} - I, Q \right\rangle^{2t} \leqslant \frac{\epsilon^{2t-1}(32Ct)^{2t}}{1-(4\epsilon)^{2t}} \left(\|Q\|_F^{2t} + \left\| \Pi \hat{\Sigma}_*^{-1/2} Q \hat{\Sigma}_*^{-1/2} \Pi \right\|_F^{2t} \right) + \frac{(8Ct)^{2t}\epsilon^{2t-2}}{1-(4\epsilon)^{2t}} \|Q\|_F^{2t} \right\}. \tag{2.106}$$

Using the SoS Contraction of Frobenius norm, i.e. Lemma 2.8.1, along with the guarantee in (2.92), we have,

$$\left\{ \left(v^\top \Sigma_*^{-1/2} \Pi^2 \Sigma_*^{-1/2} v \right)^t \leqslant \left(1.01 \|v\|^2 \right)^t \right\} \vdash \left\{ \left\| \Pi \hat{\Sigma}_*^{-1/2} Q \hat{\Sigma}_*^{-1/2} \Pi \right\|_F^{2t} \leqslant (4t)^t \|Q\|_F^{2t} \right\}.$$

Substituting back into (2.106) and setting $Q = \hat{\Sigma}_*^{-1/2} \Sigma \hat{\Sigma}_*^{-1/2} - I$,

$$\mathcal{A}_{Y,\epsilon} \vdash^{\Sigma,Q}_{4t} \left\{ \left\| \hat{\Sigma}_*^{-1/2} \Sigma \hat{\Sigma}_*^{-1/2} - I, Q \right\|_F^{4t} \leqslant \left((64Ct)^{2t}\epsilon^{2t-1} + (32Ct)^{2t}\epsilon^{2t-2} \right) \left(\left\| \hat{\Sigma}_*^{-1/2} \Sigma \hat{\Sigma}_*^{-1/2} - I \right\|_F^{2t} \right) \right\}.$$

Applying Lemma 2.8.3 with $a = \left\| \hat{\Sigma}_*^{-1/2} \Sigma \hat{\Sigma}_*^{-1/2} - I \right\|_F^{2t}$,

$$\mathcal{A}_{Y,\epsilon} \vdash^{\Sigma}_{8t} \left\{ \left\| \Sigma_*^{-1/2} \Sigma \Sigma_*^{-1/2} - I \right\|_F^{8t} \leqslant \left((64Ct)^{2t}\epsilon^{2t-2} \right)^4 \right\},$$

which yields the lemma. □

It's easy to finish the proof of Theorem 64 from here.

Proof of Theorem 64. We prove Theorem 64 here under the additional assumption that $\Sigma_* \succeq$

$2^{-\mathrm{poly}(d)} I$. Then, by an argument similar to proof of Theorem 1.2 in [KSS18], $\tilde{\mathbb{E}}[\Sigma]$ satisfies the third guarantee in Theorem 64. Let $\tilde{\zeta}$ be the degree-$O(t)$ pseudo-distribution output by our algorithm above. Then, our estimator for the covariance is simply $\hat{\Sigma} = \mathbb{E}_{\tilde{\zeta}}[\Sigma]$. From Lemma 2.6.1 it follows that

$$\mathcal{A}_{Y,\epsilon} \left\{ \left\| \hat{\Sigma}_*^{-1/2} \Sigma \hat{\Sigma}_*^{-1/2} - I \right\|_F^{8t} \leqslant \eta \right\},$$

where $\eta = O\left((Ct)^{8t} \epsilon^{8t-8}\right)$. Therefore, for any Q, we have, $\mathbb{E}_{\tilde{\zeta}} \left[\left\| \hat{\Sigma}_*^{-1/2} \Sigma \hat{\Sigma}_*^{-1/2} - I \right\|_F^{8t} \right] \leqslant \eta$. Then, using Cauchy-Schwarz for pseudo-distributions we have

$$\begin{aligned} \left\| \hat{\Sigma}_*^{-1/2} \mathop{\mathbb{E}}_{\tilde{\zeta}} [\Sigma] \hat{\Sigma}_*^{-1/2} - I \right\|_F^{8t} &= \left\| \mathop{\mathbb{E}}_{\tilde{\zeta}} \left[\hat{\Sigma}_*^{-1/2} \Sigma \hat{\Sigma}_*^{-1/2} - I \right] \right\|_F^{8t} \\ &\leqslant \mathop{\mathbb{E}}_{\tilde{\zeta}} \left[\left\| \hat{\Sigma}_*^{-1/2} \Sigma \hat{\Sigma}_*^{-1/2} - I \right\|_F^{8t} \right] \\ &\leqslant \eta. \end{aligned} \tag{2.107}$$

Taking the $8t$-th root, $\left\| \hat{\Sigma}_*^{-1/2} \mathbb{E}_{\tilde{\zeta}} [\Sigma] \hat{\Sigma}_*^{-1/2} - I \right\|_F \leqslant O\left(Ct\epsilon^{1-1/t}\right)$. Recall, by standard convergence of emperical covariance, $\left(1 - O\left(\sqrt{d \log(d)/n}\right)\right) \Sigma_* \preceq \hat{\Sigma}_* \preceq \left(1 + O\left(\sqrt{d \log(d)/n}\right)\right) \Sigma_*$ and since $n \geqslant d^4/\epsilon^2$,

$$\left\| \Sigma_*^{-1/2} \hat{\Sigma}_* \Sigma_*^{-1/2} - I \right\|_F \leqslant \sqrt{d} \left\| \Sigma_*^{-1/2} \hat{\Sigma}_* \Sigma_*^{-1/2} - I \right\|_{\mathrm{op}} \leqslant \frac{\epsilon}{d} \tag{2.108}$$

Combining the above, and using triangle inequality,

$$\begin{aligned} \left\| \Sigma_*^{-1/2} \mathop{\mathbb{E}}_{\tilde{\zeta}} [\Sigma] \Sigma_*^{-1/2} - I \right\|_F &= \left\| \Sigma_*^{-1/2} \left(\mathop{\mathbb{E}}_{\tilde{\zeta}} [\Sigma] - \hat{\Sigma}_* + \hat{\Sigma}_* \right) \Sigma_*^{-1/2} - I \right\|_F \\ &\leqslant \left\| \Sigma_*^{-1/2} \left(\mathop{\mathbb{E}}_{\tilde{\zeta}} [\Sigma] - \hat{\Sigma}_* \right) \Sigma_*^{-1/2} \right\|_F + \left\| \Sigma_*^{-1/2} \hat{\Sigma}_* \Sigma_*^{-1/2} - I \right\|_F \\ &= \left\| \Sigma_*^{-1/2} \hat{\Sigma}_*^{1/2} \left(\hat{\Sigma}_*^{-1/2} \mathop{\mathbb{E}}_{\tilde{\zeta}} [\Sigma] \hat{\Sigma}_*^{-1/2} - I \right) \hat{\Sigma}_*^{1/2} \Sigma_*^{-1/2} \right\|_F \\ &\quad + \left\| \Sigma_*^{-1/2} \hat{\Sigma}_* \Sigma_*^{-1/2} - I \right\|_F \\ &\leqslant O\left(Ct\epsilon^{1-1/t}\right) + \frac{\epsilon}{d}, \end{aligned}$$

which concludes the proof. □

2.7 Reasonable Distributions

In this section, we recall known results that imply that Gaussian distributions and affine transforms of uniform distribution on the unit sphere are reasonable.

Certifiable Hypercontractivity of Degree 2 Polynomials

Definition 2.7.1 (Certifiable Hypercontractivity)**.** *Let $\mathcal{D}$ be a distribution on $\mathcal{R}^d$. For an even h, $\mathcal{D}$ is said to have h-certifiably C-hypercontractive degree 2 polynomials if for P - a $d \times d$ matrix-valued indeterminate,*

$$\mathbf{E}_{x\sim\mathcal{D}} \left\langle P, x^{\otimes 2} \right\rangle^h \leqslant (Ch)^h (\mathbf{E} x^\top P x^2)^{h/2} .$$

Observe that certifiable hypercontractivity is invariant under linear transformations of $\mathcal{D}$. This is because for any $d \times d$ symmetric matrix A, applying a linear transformation $x \to A$ is equivalent (for the purpose of the inequality above) to conjugating P by the matrix A. Specifically, $\mathbf{E}_{x\sim\mathcal{D}} \langle P, (Ax)^{\otimes 2}\rangle = \mathbf{E}_{x\sim\mathcal{D}} \langle APA, x^{\otimes 2}\rangle$. On the other hand, $\mathbf{E}_{x\sim\mathcal{D}}[(Ax)^\top P (Ax)^2] = \mathbf{E}_{x\sim\mathcal{D}}[x^\top APAx]$. Combined with certifiable hypercontractivity of degree 2 polynomials of standard Gaussians [KOTZ14], we obtain:

Fact 2.7.2 (Hypercontractivity of Degree-2 Polynomials of Gaussians)**.** *Gaussian distributions with mean 0 and arbitrary covariance Σ have h-certifiably 1-hypercontractive degree 2 polynomials.*

Lemma 2.7.3 (Certifiable Hypercontractivity Under Sampling)**.** *Let $\mathcal{D}$ be a 1-sub-gaussian, h-certifiably c-hypercontractive distribution over $\mathbb{R}^d$. Let $\mathcal{S}$ be a set of $n = \Omega((hd)^{8h})$ i.i.d. samples from $\mathcal{D}$. Then, with probability at least $1 - 1/poly(n)$, the uniform distribution on $\mathcal{S}$ has h-certifiably $(2c)$-hypercontractive degree 2 polynomials.*

Proof. Since $\mathcal{D}$ has h-certifiably c-hypercontractive degree 2 polynomials,

$$\left\vdash_{2h}^{P} \left\{ \mathbf{E}_{x\sim\mathcal{D}} \left[\left\langle P, x^{\otimes 2} \right\rangle^h \right] \leqslant (ch)^h \|P\|_F^h \right\}$$

Since for any matrices M and N, $\langle M, N\rangle^h = \langle M^{\otimes h}, N^{\otimes h}\rangle$ using the substitution rule,

$$\left\vdash_{2h}^{P} \left\{ \left\langle P^{\otimes h}, \mathbf{E}_{x\sim\mathcal{D}} \left[x^{\otimes 2h} \right] \right\rangle \leqslant (ch)^h \|P\|_F^h \right\} \tag{2.109}$$

Let $\mathcal{D}'$ be the uniform distribution over samples from $\mathcal{D}$. Then,

$$\mathbf{E}_{x\sim\mathcal{D}'}\left[\left\langle P, x^{\otimes 2}\right\rangle^h\right] = \left\langle P^{\otimes h}, \mathbf{E}_{x\sim\mathcal{D}'}\left[x^{\otimes 2h}\right]\right\rangle$$

Let $M = \mathbf{E}_{x\sim\mathcal{D}'}\left[x^{\otimes 2h}\right] - \mathbf{E}_{x\sim\mathcal{D}}\left[x^{\otimes 2h}\right]$. Therefore, assuming that $\|M\|_2 \leqslant (ch)^h$, using Fact 3.2.19 with the substitution rule, we can conclude

$$\left\{\frac{P}{2h}\right\}\left\{\left|\left\langle P^{\otimes h}, M\right\rangle\right| \leqslant (ch)^h \|P\|_F^h\right\} \tag{2.110}$$

Observe, we can then rewrite (2.109) as follows :

$$\left\{\frac{P}{2h}\right\}\left\{\left\langle P^{\otimes h}, \mathbf{E}_{x\sim\mathcal{D}'}\left[x^{\otimes 2h}\right] - M\right\rangle \leqslant (ch)^h \|P\|_F^h\right\}$$

Rearranging and using 2.110, we can conclude

$$\left\{\frac{P}{2h}\right\}\left\{\left\langle P^{\otimes h}, \mathbf{E}_{x\sim\mathcal{D}'}\left[x^{\otimes 2h}\right]\right\rangle \leqslant 2(ch)^h \|P\|_F^h\right\}$$

Therefore, it remains to show $\|M\|_2 \leqslant (ch)^h$. Let $x^{(1)}, x^{(2)}, \ldots x^{(n)}$ be n iid samples from $\mathcal{D}$. Then, observe

$$M_{i_1,\ldots,i_{2h}} = \left[\mathbf{E}_{x\sim\mathcal{D}'} x^{\otimes 2h}\right]_{i_1,\ldots i_{2h}} - \left[\mathbf{E}_{x\sim\mathcal{D}} x^{\otimes 2h}\right]_{i_1,\ldots i_{2h}} = \frac{1}{n}\sum_{\ell\in[n]}\left(x^{(\ell)}_{i_1} x^{(\ell)}_{i_2} \ldots x^{(\ell)}_{i_{2h}} - \mathbf{E}_{x\sim\mathcal{D}}\left[x_{i_1} x_{i_2} \ldots x_{i_{2h}}\right]\right).$$

Let $Z_\ell = \left(x^{(\ell)}_{i_1} x^{(\ell)}_{i_2} \ldots x^{(\ell)}_{i_{2h}}\right)$. Then, $M_{i_1,\ldots,i_{2h}}$ is an average of independent random variables $\bar{Z}_\ell = Z_\ell - \mathbf{E}[Z_\ell]$ for $\ell \in [n]$. We will estimate moments of $\sum_{\ell\leqslant n}\bar{Z}_\ell$ in order to order to obtain upper bounds on the deviation probabilities.

Towards that we observe the following: $\mathbf{E}\left[\left(\frac{1}{n}\sum_{\ell\in[n]}\bar{Z}_\ell\right)^{2t}\right] = \frac{1}{n^{2t}}\sum_{r_1,r_2,\ldots,r_{2t}}\mathbf{E}\left[\prod_{j\in[2t]}\bar{Z}_{r_j}\right]$. If $\mathbf{E}[\prod_{j\in[2t]}\bar{Z}_{r_j}] \neq 0$, then, each $\bar{Z}_{r_j}$ must appear even number of times in the product. Thus, the number of distinct $\bar{Z}_{r_j}$ in the product are at most t. Thus, the number of non-zero terms in the above sum is at most $n^t(2t)^{2t}$. Next, for any non-zero term in the above sum, using the AM-GM inequality,

$$\mathbf{E}\left[\prod_{i\in[2t]}\bar{Z}_{r_j}\right] \leqslant \frac{1}{(2t)^{2t}}\mathbf{E}\left[\left(\sum_{i\in[2t]}\bar{Z}_{r_j}\right)^{2t}\right] \leqslant \frac{1}{(2t)}\sum_{i\in[2t]}\mathbf{E}[\bar{Z}^{2t}_{r_j}] \tag{2.111}$$

By Jensen's inequality, $(\mathbf{E}[Z_{r_i}])^{2t} \leqslant \mathbf{E}[Z^{2t}_\ell]$ and thus $\mathbf{E}\left[\bar{Z}^{2t}_{r_j}\right] \leqslant 2^{2t}(\mathbf{E}[Z^{2t}_{r_j}] + (\mathbf{E}[Z_{r_j})^{2t}]) \leqslant$

$2^{2t+1}\mathbf{E}[Z_{r_j}^{2t}]$. Then,

$$\begin{aligned}\mathbf{E}\left[Z_{r_j}^{2t}\right] = \mathbf{E}\left[\left(x_{i_1}^{(r_j)}x_{i_2}^{(r_j)}\dots x_{i_{2h}}^{(r_j)}\right)^{2t}\right] &\leqslant \mathbf{E}\left[\left(\frac{1}{2h}\sum_{k\in[2h]}\left(x_{i_k}^{(r_i)}\right)^{2h}\right)^{2t}\right] \\ &\leqslant \frac{1}{2h}\sum_{k\in[2h]}\mathbf{E}\left[\left(x_{i_k}^{(r_i)}\right)^{4ht}\right] \\ &\leqslant (4ht)^{2ht}\end{aligned} \tag{2.112}$$

where the first inequality uses the AM-GM inequality, the second uses Jensen's inequality and the final inequality uses the 1-subgaussianity of $x_{i_j}^{(r_j)}$. Combining (2.111) and (2.112)

$$\mathbf{E}\left[\left(\frac{1}{n}\sum_{\ell\in[n]}\bar{Z}_\ell\right)^{2t}\right] \leqslant \frac{1}{2tn^{2t}}\cdot n^t(2t)^{2t}\cdot(4ht)^{2ht} \leqslant n^{-t}(2t)^{2t-1}(4ht)^{2ht}$$

Using Chebyshev's inequality,

$$\mathbf{Pr}\left[\left|\frac{1}{n}\sum_{\ell\in[n]}\bar{Z}_\ell\right| > \eta\right] \leqslant \frac{\mathbf{E}\left[\left(\frac{1}{n}\sum_{\ell\in[n]}\bar{Z}_\ell\right)^{2t}\right]}{\eta^{2t}} \leqslant \frac{(2t)^{2t-1}(4ht)^{2ht}}{\eta^{2t}n^t}$$

Setting $t > 2h\log d$ and $\eta = (ch/d^2)^h$ yields that whenever $n \geqslant n_0 = \Omega\left(\frac{d^{4h}}{c^{2h}}h^{9h}\log^{2h+2}(d)\right)$, $|M_{i_1,i_2,\dots i_{2h}}| \leqslant \eta$ with probability at least $1 - 1/d^{4h}$. By a union bound over the d^{2h} entries of M, we have that all entries of M are at most η with probability at least $1 - d^{-2h}$. We can then easily bound the operator norm of M by $d^{2h}\cdot(ch/d^2)^h = (ch)^h$, which completes the proof.

□

Certifiable Anti-Concentration

Lemma 2.7.4 (Certifiable Anti-Concentration of Gaussians, Theorem 5.5 [BK20a])**.** *Given $0 < \delta \leqslant 1/2$, there exists $s = O\left(\frac{\log^5(1/\delta)}{\delta^2}\right)$ such that the Gaussian distribution and the uniform distribution on the unit sphere is s-certifiably (C,δ)-anti-concentrated.*

Lemma 2.7.5 (Certifiable Anti-Concentration under Sampling, Lemma 5.8 [BK20a])**.** *Let $\mathcal{D}$ be s-certifiably (c,δ)-anti-concentrated Sub-Exponential distribution over $\mathbb{R}^d$. Let $\mathcal{S}$ be a set of $n = \Omega((sd\log(d))^s)$ i.i.d. samples from $\mathcal{D}$. Then, with probability at least $1 - 1/poly(n)$, the uniform distribution on $\mathcal{S}$ is s-certifiably $(2c,\delta)$-anti-concentrated.*

Bounded Variance of Degree-2 Polynomials. Recall that we say that a zero mean distribution $\mathcal{D}$ with covariance Σ has certifiably C-bounded variance degree 2 polynomials if $\vdash_2^Q$ $\left\{\mathbf{E}_{x\sim\mathcal{D}}(x^\top Qx - \mathbf{E}_{x\sim\mathcal{D}}x^\top Qx)^2 \leqslant C\left\|\Sigma^{1/2}Q\Sigma^{1/2}\right\|_F^2\right\}$.

Lemma 2.7.6 (Bounded Variance of Degree 2 Polynomials of 4-wise independent distributions)**.** *Let $\mathcal{D}$ be an isotropic, 4-wise independent distribution on $\mathcal{R}^d$. Then, $\mathcal{D}$ has certifiably 3-bounded variance degree 2 polynomials. That is,*

$$\vdash_2^Q \left\{\mathbf{E}_{\mathcal{D}}\left(x^\top Qx - \mathbf{E}_{\mathcal{D}}x^\top Qx\right)^2 \leqslant 3\left\|Q\right\|_F^2\right\}.$$

Proof. By viewing $xx^\top$ and $I \in \mathcal{R}^{d\times d}$ as d^2 dimensional vectors, and using that $\mathbf{E}_{y\sim\mathcal{D}}(yy^\top - I)(yy^\top - I)^\top \preceq 3I\otimes I$ for any 4-wise independent, isotropic distribution, we have:

$$\vdash_2^Q \Bigg\{\mathbf{E}_{\mathcal{D}}\left(x^\top Qx - \mathbf{E}_{\mathcal{D}}x^\top Qx\right)^2 = \mathbf{E}_{\mathcal{D}}\left\langle xx^\top - I, Q\right\rangle^2 \leqslant \left\|\mathbf{E}_{x\sim\mathcal{D}}(xx^\top - I)(xx^\top - I)^\top\right\|_2\left\|Q\right\|_F^2$$
$$\leqslant 3\left\|I\otimes I\right\|_2\left\|Q\right\|_F^2 = 3\left\|Q\right\|_F^2\Bigg\}. \quad (2.113)$$

□

The uniform distribution on $\sqrt{d}$-radius sphere in d dimensions is not 4-wise independent. However, the above proof only requires that $\mathbf{E}(y^{\otimes 2} - I)(y^{\otimes 2} - I)^\top \preceq CI\otimes I$. For the uniform distribution on the sphere, notice that i, j, k, ℓ-th entry of this matrix is non-zero iff the indices are in have two repeated indices and in that case, by negative correlation of the x_i^2 and x_j^2 on the sphere, it holds that $\mathbf{E}x_i^2x_j^2 \leqslant 1$. Thus, $\mathbf{E}(y^{\otimes 2} - I)(y^{\otimes 2} - I)^\top \preceq 3I\otimes I$ for y uniformly distribution on the $\sqrt{d}$-radius unit sphere. The above proof thus also yields:

Corollary 2.7.7. *Let y be uniform on $\sqrt{d}$-radius sphere in d dimensions. Then, y has certifiably 3-bounded variance degree 2 polynomials.*

Lemma 2.7.8 (Linear Invariance)**.** *Let x be a random variable with an isotropic distribution $\mathcal{D}$ on $\mathcal{R}^d$ with certifiably C-bounded variance degree 2 polynomials. Let $A \in \mathcal{R}^{d\times d}$ be an arbitrary $d\times d$ matrix. Then, the random variable $x' = Ax$ also has certifiably C-bounded variance degree 2 polynomials.*

Proof. The covariance of x' is $AA^\top = \Sigma$, say. Let $\Sigma^{1/2}$ be the PSD square root of Σ. The proof follows by noting that $x'^\top Qx' = (Ax)^\top Q(Ax) = x^\top(A^\top QA)x^\top$ and that $\left\|A^\top QA\right\|_F^2 =$

$$\mathrm{tr}(A^\top QAA^\top QA) = \mathrm{tr}(AA^\top QAA^\top Q) = \mathrm{tr}(\Sigma Q\Sigma Q) = \mathrm{tr}(\Sigma^{1/2}Q\Sigma^{1/2}\Sigma^{1/2}Q\Sigma^{1/2}) = \left\|\Sigma^{1/2}Q\Sigma^{1/2}\right\|_F^2.$$

□

Lemma 2.7.9 (Bounded Variance Under Sampling)**.** *Let $\mathcal{D}$ be have degree 2 polynomials with certifiably C-bounded variance and be 8-certifiably C-subgaussian. Let X be an i.i.d. sample from $\mathcal{D}$ of size $n \geqslant n_0 = O(C^4)d^{16}$. Then, with probability at least 0.99 over the draw of X, the uniform distribution on X has degree 2 polynomials with certifiable $2C$-bounded variance.*

Proof. Using Lemma 2.7.8, we can assume that $\mathcal{D}$ is isotropic. Arguing as in the proof of Lemma 2.7.3, it is enough to upper-bound the spectral norm

$$\left\|\frac{1}{n}\sum_i (x_i^{\otimes 2} - I)(x_i^{\otimes 2} - I)^\top - \mathbf{E}_{x\sim\mathcal{D}}(x^{\otimes 2} - I)(x^{\otimes 2} - I)^\top\right\|_2$$

by C (with probability 0.99 over the draw of X). We do this below:

By applying certifiable C-bounded variance property to $Q = vv^\top$ where e_i are standard basis vectors in $\mathcal{R}^d$, we have that $\mathbf{E}(\langle x_i, v\rangle^2 - \mathbf{E}\langle x_i, v\rangle^2)^2 \leqslant C\|v\|_2^4$ and thus, $\mathbf{E}\langle x_i, v\rangle^4 \leqslant (1+C)\|v\|_2^4$. By an application of the AM-GM inequality, we know that for every i, j, k, ℓ, $(\langle x, e_i\rangle^2\langle x, e_j\rangle^2\langle x, e_k\rangle\langle x, e_\ell\rangle)^2 \leqslant \langle x, e_i\rangle^8 + \langle x, e_j\rangle^8 + \langle x, e_k\rangle^8 + \langle x, e_\ell\rangle^8$. Thus, the variance of every entry of the matrix $\mathbf{E}x^{\otimes 4}$ is bounded above by $4(8C)^4 = O(C^4)$. Thus, by Chebyshev's inequality, any given entry of $\frac{1}{n}x_i^{\otimes 4} - \mathbf{E}_{x\sim\mathcal{D}}x^{\otimes 4}$ is upper-bounded by $O(C^2)d^4/\sqrt{n}$ with probability at least $1 - 1/(100d^4)$. By a union bound, all entries of this tensor are upper-bounded by $O(C^2)d^4/\sqrt{n}$ with probability at least 0.99. Thus, the Frobenius norm of this tensor is at most $d^8O(C^2)/\sqrt{n}$. Since $n \geqslant n_0 = O(C^4)d^{16}$, this bound is at most $C/2$. Thus, we obtain that with probability at least 0.99,

$$\left\|\frac{1}{n}\sum_i (x_i^{\otimes 2} - I)(x_i^{\otimes 2} - I)^\top - \mathbf{E}_{x\sim\mathcal{D}}(x^{\otimes 2} - I)(x^{\otimes 2} - I)^\top\right\|_2 \leqslant 2\left\|\frac{1}{n}\sum_i x_i^{\otimes 4} - \mathbf{E}_{x\sim\mathcal{D}}x^{\otimes 4}\right\|_F \leqslant C$$

.

□

The above three lemmas immediately yield that Gaussian distributions, linear transforms of uniform distribution on unit sphere, discrete product sets such as the Boolean hypercube and any 4-wise independent zero-mean distribution has certifiably C-bounded variance degree 2 polynomials.

2.8 Sum-of-Squares Toolkit

In this section, we give low-degree SoS proofs of some inequalities that we use repeatedly in our arguments.

The following is an SoS version of the following simple matrix analytic inequality: for any matrices A, B, $\|AB\|_F^2 \leqslant \|A\|_{op}^2 \|B\|_F^2$. We give a constant degree SoS proof of this inequality (with $O(1)$ factor loss) by relying on certifiable hypercontractivity of Gaussians.

Lemma 2.8.1 (Contraction and Frobenius Norms)**.** *Let A, B be $d \times d$ matrix valued indeterminates. Let β be a scalar-valued indeterminate. Then,*

$$\left\{\beta \left(v^\top A^\top A v\right)^t \preceq \Delta \|v\|_2^{2t}\right\} \vdash \left\{\beta \|AB\|_F^{2t} \leqslant \Delta t^t \|B\|_F^{2t}\right\},$$

and

$$\left\{\beta \left(v^\top A A^\top v\right)^t \preceq \Delta \|v\|_2^{2t}\right\} \vdash \left\{\beta \|BA\|_F^{2t} \leqslant \Delta t^t \|B\|_F^{2t}\right\},$$

Proof. We prove the first conclusion. The proof of the second one is similar.

We start by observing that for any matrix valued indeterminate M, $\left\{\frac{M}{2}\right\} \left\{\|M\|_F^2 = \mathbb{E}_{g\sim\mathcal{N}(0,I)} \left[\|Mg\|_2^2\right]\right\}$.

We thus have:

$$\begin{aligned}
\left\{\beta \left(v^\top A^\top A v\right)^t \leqslant \Delta \|v\|_2^{2t}\right\} \vdash \Bigg\{ \beta \left(\|AB\|_F^2\right)^t &= \beta \left(\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,I)} \left[\|ABg\|_2^2\right]\right)^t \\
&\leqslant \beta \mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,I)} \left[\|ABg\|_2^t\right] \\
&= \mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,I)} \left[\left((Bg)^\top \left(\beta A^\top A\right)(Bg)\right)^t\right] \\
&\leqslant \Delta \cdot \mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,I)} \left[\|Bg\|_2^{2t}\right] \\
&\leqslant t^t \Delta \left(\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,I)} \left[\|Bg\|_2^2\right]^t\right) \\
&= t^t \Delta \|B\|_F^{2t} \Bigg\}.
\end{aligned} \tag{2.114}$$

Here, the first inequality follows by using the SoS Hölder's inequality, the second one uses the constraint satisfied by $A^\top A$ with the substituting $v = Bg$ and finally, the last inequality relies on

certifiable hypercontractivity of quadratic forms of Gaussians. This completes the proof.

□

The following two lemmas allow us to "cancel out" common factors from both sides of an inequality in low-degree SoS.

Lemma 2.8.2 (Cancellation within SoS, Constant RHS)**.** *Let a be an indeterminate. Then,*

$$\left\{a^{2t} \leqslant 1\right\} \left|\frac{a}{2t}\right. \left\{a^2 \leqslant 1\right\} .$$

Proof. Applying the SoS AM-GM inequality (Fact 3.2.22) with $f_1 = a^2$, $f_2 = \ldots = f_t = 1$, we get:

$$\left|\frac{a}{2t}\right. \left\{a^2 \leqslant a^{2t}/t + 1 - 1/t\right\} .$$

Thus,

$$\left\{a^{2t} \leqslant 1\right\} \left|\frac{a}{2t}\right. \left\{a^2 \leqslant 1/t + 1 - 1/t = 1\right\} .$$

□

Lemma 2.8.3 (Cancellation Within SoS)**.** *Let a, C be indeterminates. Then,*

$$\{a \geqslant 0\} \cup \left\{a^t \leqslant Ca^{t-1}\right\} \left|\frac{a,C}{2t}\right. \left\{a^{2t} \leqslant C^{2t}\right\} .$$

Proof. We first prove the case of $t = 2$. We have:

$$\left|\frac{a,C}{2}\right. \left\{a^2 = (a - C/2 + C/2)^2 \leqslant 2(a - C/2)^2 + 2(C/2)^2\right\} .$$

And,

$$\left\{a^2 \leqslant Ca\right\} \left|\frac{a,C}{2}\right. \left\{(a - C/2)^2 \leqslant C^2/4\right\} .$$

Thus,

$$\left\{a^2 \leqslant Ca\right\} \left|\frac{a,C}{2}\right. \left\{a^2 \leqslant C^2\right\} .$$

Consider now the general case. Iteratively using $\{a^t \leqslant Ca^{t-1}\}$ yields:

$$\{a \geqslant 0\} \cup \left\{a^t \leqslant Ca^{t-1}\right\} \left|\frac{a,C}{2t}\right. \left\{a^{2t} \leqslant a^{t-2}a^tC^2 \leqslant a^{t-3}a^tC^3 \ldots \leqslant a^tC^t\right\} .$$

Applying the special case of $t = 2$ above to the indeterminate a^t now yields:

$$\{a \geqslant 0\} \left\{a^t \leqslant C a^{t-1}\right\} \left|\frac{a,C}{2t}\right. \left\{a^{2t} \leqslant C^{2t}\right\} .$$

□

2.9 Total Variation vs Parameter Distance for Gaussian Distributions

Proposition 2.9.1 (Parameter Closeness Implies TV Closeness for Gaussian Base Model)**.** *Fix* $\Delta > 0$ *and let* μ, μ' *and* $\Sigma, \Sigma' \succ 0$ *satisfy:*

1. ***Mean Closeness:*** *for all* $v \in \mathcal{R}^d$, $\left\|\left(\mu - \mu'\right), v\right\|_2^2 \leqslant \Delta^2 v^\top (\Sigma + \Sigma') v$.
2. ***Spectral Closeness:*** *for all* $v \in \mathcal{R}^d$ $\frac{1}{\Delta^2} v^\top \Sigma v \leqslant v^\top \Sigma' v \leqslant \Delta^2 v^\top \Sigma(r') v$.
3. ***Relative Frobenius Closeness:*** $\left\|\Sigma^{\dagger/2} \Sigma' \Sigma^{\dagger/2} - I\right\|_F^2 \leqslant \Delta^2 \cdot \left\|\Sigma^\dagger \Sigma'\right\|_2^2$.

Then, $d_{\mathsf{TV}}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')) \leqslant 1 - \exp(-O(\Delta^2 \log \Delta))$.

Proof of Lemma 2.9.1. We will work with the distributions after applying the transformation $x \to \Sigma^{-1/2} x$ to the associated random variables. Since d_{TV} is invariant under affine transformations, this is WLOG. The transformation produces distributions $\mathcal{N}(\mu_1, I)$ and $\mathcal{N}(\Sigma^{-1/2}\mu', \Sigma^{-1/2}\Sigma'\Sigma^{-1/2})$ for $\mu_1 = \Sigma^{-1/2}\mu$, $\mu_2 = \Sigma^{-1/2}\mu'$ and $\Sigma_2 = \Sigma^{-1/2}\Sigma'\Sigma^{-1/2}$.

We will first bound the Hellinger distance between the two distributions above. Recall that $h = h(\mathcal{N}(\Sigma^{-1/2}\mu, I), \mathcal{N}(\Sigma^{-1/2}\mu', \Sigma^{-1/2}\Sigma'\Sigma^{-1/2}))$ satisfies:

$$h(\mathcal{N}(\mu_1, I), \mathcal{N}(\mu_2, \Sigma_2))^2 = 1 - \frac{\det(\Sigma_2)^{1/4}}{\det\left(\frac{I+\Sigma_2}{2}\right)^{\frac{1}{2}}} \exp\left(-\frac{1}{8}(\mu_1 - \mu_2)^\top \left(\frac{I + \Sigma_2}{2}\right)^{-1} (\mu_1 - \mu_2)\right) .$$

We will estimate the RHS of the expression above to bound the Hellinger distance.

From the mean closeness condition, we have:

$$\langle \mu_1 - \mu_2, v \rangle = \langle \mu - \mu', \Sigma^{-1/2} v \rangle \leqslant \sqrt{\log 1/\eta} \sqrt{v^\top (I + \Sigma_2) v} .$$

Plugging in $v = \left(\frac{I+\Sigma_2}{2}\right)^{-1}(\mu_1 - \mu_2)$ gives:

$$\left\langle \mu_1 - \mu_2, \frac{I+\Sigma_2}{2}^{-1}(\mu_1 - \mu_2) \right\rangle \leqslant 2/\eta \sqrt{v^\top \left(\frac{I+\Sigma_2}{2}\right)^{-1} v},$$

or,

$$\left\langle \mu_1 - \mu_2, \left(\frac{I+\Sigma_2}{2}\right)^{-1}(\mu_1 - \mu_2) \right\rangle \leqslant 41/\eta^2 .$$

And thus,

$$\exp\left(-\frac{1}{8}(\mu_1 - \mu_2)^\top \left(\frac{I+\Sigma_2}{2}\right)^{-1}(\mu_1 - \mu_2)\right) \geqslant \exp\left(-1/2\eta^2\right).$$

Thus, we have:

$$h \leqslant 1 - \frac{\det(\Sigma_2)^{\frac{1}{4}}}{\det\left(\frac{\mathbf{I}+\Sigma_2}{2}\right)^{1/2}} \exp\left(-1/2\eta^2\right).$$

Let $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_d$ be eigenvalues of Σ_2. From the spectral closeness condition, observe that each $\frac{1}{\eta} \geqslant \lambda_1 \geqslant \cdots \lambda_d \geqslant \eta$.

Then,

$$\frac{\det(\Sigma_2)^{\frac{1}{4}}}{\det\left(\frac{\mathbf{I}+\Sigma_2}{2}\right)^{1/2}} = \frac{\Pi_{i\leqslant d}\lambda_i^{1/4}}{\Pi_{i\leqslant d}\left(\frac{1+\lambda_i}{2}\right)^{1/2}}.$$

Thus,

$$\log(1/(1-h)) \leqslant \frac{1}{2}\log(1/\eta) + \frac{1}{2}\sum_{i\in[d]} \log\left(\frac{1+\lambda_i}{2\sqrt{\lambda_i}}\right). \tag{2.115}$$

We break the second term in the RHS above based on the magnitude of the eigenvalues λ_is. Let's first bound the contribution to this term coming from eigenvalues $\lambda_i \geqslant 1.5$ - let's call these the *large* eigenvalues of Σ_2.

Next, observe that the Relative Frobenius Closeness condition gives us that $\|\mathbf{I} - \Sigma_2\|_F^2 \leqslant (1/\eta^2)$. Thus, $\sum_{i\in[d]}(1-\lambda_i)^2 = \|\mathbf{I} - \Sigma_2\|_F^2 \leqslant (1/\eta^2)$, the number of large eigenvalues is at most $4/\eta^2$. Further, for every large eigenvalue λ_i, $1 + \lambda_i \leqslant 2\lambda_i$. Thus,

$$\sum_{i:\lambda_i \text{ is large}} \log\left(\frac{1+\lambda_i}{2\sqrt{\lambda_i}}\right) \leqslant \sum_{i\in\mathcal{E}} \log\left(\sqrt{\lambda_i}\right) \leqslant \frac{2}{\eta}\cdot\log(1/\eta)$$

where the last step uses that $\lambda_i \leqslant 1/\eta$.

Let's now consider all the remaining *small* eigenvalues that satisfy $\eta \leqslant \lambda_i < 1.5$. Then, we can write $\lambda_i = 1 + \beta_i$ such that $-(1-\eta) \leqslant \beta_i \leqslant 0.5$. Then, we have

$$\begin{aligned}\sum_{i:\lambda_i \leqslant 1.5} \log\left(\frac{1+\lambda_i}{2}\right) + \frac{1}{2}\log\left(\frac{1}{\lambda_i}\right) &= \sum_{i \in \mathcal{E}'} \log\left(1 + \frac{\beta_i}{2}\right) - \frac{1}{2}\log\left(1+\beta_i\right) \\ &\leqslant \sum_{i:\lambda_i \leqslant 1.5} \frac{\beta_i}{2} - \frac{\beta_i}{2} + \frac{\beta_i^2}{4} \\ &= \sum_{i:\lambda_i \leqslant 1.5} \frac{(1-\lambda_i)^2}{4} \leqslant \frac{1}{4\eta^2}\end{aligned}$$

using the bound $\sum_i (1-\lambda_i)^2 \leqslant \frac{1}{\eta^2}$ in the last inequality. Plugging this estimate back in (2.115) yields $h \geqslant 1 - \exp(-O(1/\eta^2 \log(1/\eta))$.

To finish the proof, we observe that $d_{\mathsf{TV}}(p,q) \leqslant h(p,q)\sqrt{2 - h(p,q)} \leqslant 1 - \exp(-O(1/\eta^2 \log(1/\eta))$.

□

2.10 Typical Samples are Good with High Probability

Proof of Lemma 2.3.2. We begin with the empirical mean condition. For any fixed ℓ, C_ℓ contains samples from a 1-Sub-gaussian distributions and thus it follows from Fact 2.2.2 that with probability at least $1 - (1/\delta)$,

$$\left\langle \mu_\ell - \hat{\mu}_\ell, \Sigma_\ell^{\dagger/2} v \right\rangle^2 = v^\top \Sigma_\ell^{\dagger/2} (\mu_\ell - \hat{\mu}_\ell)(\mu_\ell - \hat{\mu}_\ell)^\top \Sigma_\ell^{\dagger/2} v \leqslant \left(\frac{kr + \log(1/\delta)k}{n}\right) v^T v$$

Since $n_0 = \Omega((k\log(rk) + kr))$, we can substitute $v \to \Sigma_\ell^{1/2} v$ to get

$$\left\langle \mu_\ell - \hat{\mu}_\ell, \Sigma_\ell^{\dagger/2} \Sigma_\ell^{1/2} v \right\rangle^2 \leqslant 1.01 v^T \Sigma_\ell v$$

Observe, $\left\langle \mu_\ell - \hat{\mu}_\ell, \Sigma_\ell^{\dagger/2}\Sigma_\ell^{1/2} v \right\rangle = \left\langle \Sigma_\ell^{\dagger/2}\Sigma_\ell^{1/2}(\mu_\ell - \hat{\mu}_\ell), v \right\rangle = \langle \mu_\ell - \hat{\mu}_\ell, v\rangle$, where the last equality follows from observing that $\mu_\ell - \hat{\mu}_\ell$ lies in the subspace spanned by Σ_ℓ. Union bound over failure events for all $\ell \in [k]$ and thus with probability at least $1 - 1/\mathrm{poly}(k)$, for all $\ell \in [k]$, $\langle \mu_\ell - \hat{\mu}_\ell, v\rangle^2 \leqslant 1.01 v^T \Sigma_\ell v$.

Similarly, using Fact 2.2.3 for i.i.d. samples from a 1-Sub-gaussian distribution, it follows

that for a fixed $\ell \in [k]$, with probability at least $1 - 1/d^{10}$,

$$\left(1 - c\sqrt{\frac{rk\log(k)}{n}}\right)\Sigma_\ell \preceq \hat{\Sigma}_\ell \preceq \left(1 + c\sqrt{\frac{rk\log(k)}{n}}\right)\Sigma_\ell$$

for fixed constants c. Union bounding over $\ell \in [k]$, and observing that $n_0 = \Omega(rk\log(k)/2^{2s})$ with probability at least $1 - 1/k^8$ for all $\ell \in [k]$,

$$\left(1 - \frac{1}{2^{2s}}\right)\Sigma_\ell \preceq \hat{\Sigma}_\ell \preceq \left(1 + \frac{1}{2^{2s}}\right)\Sigma_\ell \tag{2.116}$$

for any $s > 2$, which concludes the empirical covariance condition. By definition of a "nice" distribution, we know that the points in C_ℓ are drawn i.i.d. from a s-certifiably (C, δ)-anti-concentrated distribution denoted by $\mathcal{D}(\mu_\ell, \Sigma_\ell)$ and thus for all η,

$$\left\lfloor \frac{v}{2s} \right\rfloor \left\{ \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)} \left[q^2_{\eta,\Sigma_\ell}\left(\langle x - y, v\rangle\right)\right] \leqslant C\eta\left(v^\top \Sigma_\ell v\right)^s \right\}$$

Consider the substitution $v \to \Sigma^{\dagger/2}v$. Then,

$$\left\lfloor \frac{v}{2s} \right\rfloor \left\{ \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)} \left[q^2_{\eta,\Sigma_\ell}\left(\left\langle \Sigma_\ell^{\dagger/2}(x - y), v\right\rangle\right)\right] \leqslant C\eta\left\|v\right\|_2^{2s} \right\}$$

Since $q_{\eta,\hat{\Sigma}}$ is a degree-s even polynomial, $q^2_{\eta,\hat{\Sigma}}(z) = \sum_{i\in[s]} c_i z^{2i}$ and thus using the substitution rule,

$$\left\lfloor \frac{v}{2s} \right\rfloor \left\{ \sum_{j\in[s]} c_i \left\langle \mathbf{E}_{x,y\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)} \left(\Sigma_\ell^{\dagger/2}(x - y)\right)^{\otimes 2j}, v^{\otimes 2j}\right\rangle \leqslant C\eta\left\|v\right\|_2^{2s} \right\} \tag{2.117}$$

Let $\mathcal{D}$ be the true distribution and $\mathcal{D}'$ be the uniform distribution over n samples from $\mathcal{D}$. We can rewrite the above expression by adding and subtracting $\mathbf{E}_{x,y\sim\mathcal{D}'}\left(\Sigma_\ell^{\dagger/2}(x - y)\right)^{\otimes 2j}$ as follows:

$$\begin{aligned}\left\lfloor \frac{v}{2s} \right\rfloor \Bigg\{ & \frac{k^2}{n^2} \sum_{i\neq j\in C_\ell} q^2_{\eta,\hat{\Sigma}(r)}\left(x_i - x_j, \Sigma^{\dagger/2}v\right) \\ & \leqslant \sum_{j\in[s]} c_i \left\langle \mathbf{E}_{x,y\sim\mathcal{D}} \left(\Sigma_\ell^{\dagger/2}(x - y)\right)^{\otimes 2j} - \mathbf{E}_{x,y\sim\mathcal{D}'} \left(\Sigma_\ell^{\dagger/2}(x - y)\right)^{\otimes 2j}, v^{\otimes 2j}\right\rangle \\ & \quad + C\eta\left\|v\right\|_2^{2s} \Bigg\}\end{aligned} \tag{2.118}$$

By definition of a reasonable distribution, we know that $\Sigma^{\dagger/2}(x-y)$ is certifiably hypercontractive (and thus subgaussian with covariance bounded by identity). Then, using concentration of polynomials of sub-exponential random variables, for all $i_1, i_2 \in [d^j]$,

$$\mathbf{Pr}_{x\sim\mathcal{D}}\left[\left|\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)}\left[((x-y)^{\otimes j})_{i_1}((x-y)^{\otimes j})_{i_2}\right]-\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}(\mu_\ell,\hat{\Sigma}_\ell)}\left[((x-y)^{\otimes j})_{i_1}((x-y)^{\otimes j})_{i_2}\right]\right|>\epsilon\right] \leqslant \exp\left(-\left(\frac{\epsilon n}{\mathbb{E}_{x,y}[((x-y)^{\otimes j})_{i_1}((x-y)^{\otimes j})_{i_2}]^2}\right)^{\frac{1}{2s}}\right)$$

Setting $\epsilon = \mathbb{E}_{x,y\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)}\left[((x-y)^{\otimes j})_{i_1}((x-y)^{\otimes j})_{i_2}\right]/2^{2s}$, and union bounding over d^s entries, we can bound error probability by $d^{2s}\exp\left(-\left(\frac{n}{(2d)^{O(s)}}\right)^{\frac{1}{2s}}\right)$. Therefore, setting $n=\Omega((sd\log(d))^s)$ suffices and substituting $v\to\Sigma^{1/2}v$, we have with probability $1-1/\mathrm{poly}(d)$,

$$\begin{aligned}\left|_{\frac{v}{2s}}\left\{\frac{k^2}{n^2}\sum_{i\neq j\in C_\ell} q^2_{\eta,\hat{\Sigma}(r)}(x_i-x_j,v)\right.\right. \\ \left.\left.\leqslant\left(1+\frac{1}{2^{2s}}\right)^s\sum_{j\in[s]}c_i\left\langle\mathbf{E}_{x,y\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)}(x-y)^{\otimes 2j},v^{\otimes 2j}\right\rangle+C\eta\left(v^\top\Sigma_\ell v\right)_2^{2s}\right\}\right.\end{aligned} \tag{2.119}$$

Applying the definition of certifiable anti-concentration again, and using the spectral closeness from Eqn (2.116), we can conclude

$$\left|_{\frac{v}{2s}}\left\{\frac{k^2}{n^2}\sum_{i\neq j\in C_\ell} q^2_{\eta,\hat{\Sigma}(r)}(x_i-x_j,v)\leqslant 10C\eta\left(v^\top\hat{\Sigma}_\ell v\right)_2^{2s}\right\}\right. \tag{2.120}$$

A similar proof applies to 4-tuples and yields the second property for anti-concentration.

Since for all $\ell\in[k]$, $\mathcal{D}(\mu_\ell,\Sigma_\ell)$ is also s-certifiably C-hypercontractive,

$$\left|_{\frac{Q}{2s}}\left\{\mathbf{E}_{x,y\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)}\left[((x-y)^\top Q(x-y))^s\right]\leqslant(Cs)^s\mathbf{E}_{x\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)}\left[((x-y)^\top Q(x-y))^2\right]^{s/2}\right\}\right. \tag{2.121}$$

Substituting $Q=\Sigma^{\dagger/2}Q\Sigma^{\dagger/2}$ and observing

$$(x-y)^\top\Sigma^{\dagger/2}Q\Sigma^{\dagger/2}(x-y)=\left\langle\Sigma^{\dagger/2}(x-y)(x-y)^\top\Sigma^{\dagger/2},Q\right\rangle=\left\langle\left(\Sigma^{\dagger/2}(x-y)\right)^{\otimes 2},Q\right\rangle,$$

we have

$$\vdash^{Q}_{2s} \Big\{ \mathbf{E}_{x,y\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)} \Big[\Big(\Big\langle (\Sigma^{\dagger/2}(x-y))^{\otimes 2}, Q \Big\rangle \Big)^s \Big] \\ \leqslant (Cs)^s \mathbf{E}_{x\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)} \Big[((x-y)^\top \Sigma^{\dagger/2} Q \Sigma^{\dagger/2}(x-y))^2 \Big]^{s/2} \Big\} \tag{2.122}$$

Observing that $\mathbb{E}_{x,y\sim\mathcal{D}}[(x-y)] = 0$, we have

$$\vdash^{Q}_{2s} \Big\{ \Big(\Big\langle \mathbf{E}_{x,y\sim\mathcal{D}(\mu_\ell,\Sigma_\ell)} (\Sigma^{\dagger/2}(x-y))^{\otimes 2s}, Q^{\otimes s} \Big\rangle \Big) \leqslant (Cs)^{2s} \|Q\|_F^2 \Big\} \tag{2.123}$$

Let $\mathcal{D}$ represent the true distribution and $\mathcal{D}'$ represent the uniform distribution over pairs (x_i, x_j) sampled from $\mathcal{D}$. Then, adding and subtracting $\left\langle \mathbf{E}_{x,y\sim\mathcal{D}'} (\Sigma^{\dagger/2}(x-y))^{\otimes 2s}, Q^{\otimes s} \right\rangle$, we have

$$\vdash^{Q}_{2s} \left\{ \frac{k^2}{n^2} \sum_{i\neq j\in C_\ell} \Big((x-y)^\top \Sigma^{\dagger/2} Q \Sigma^{\dagger/2}(x-y) \Big)^s \leqslant |\Delta| + (Cs)^{2s} \|Q\|_F^2 \right\} \tag{2.124}$$

where $\Delta = \left\langle \mathbf{E}_{x,y\sim\mathcal{D}'} (\Sigma^{\dagger/2}(x-y))^{\otimes 2s}, Q^{\otimes s} \right\rangle - \left\langle \mathbf{E}_{x,y\sim\mathcal{D}} (\Sigma^{\dagger/2}(x-y))^{\otimes 2s}, Q^{\otimes s} \right\rangle$. Using Lemma 2.7.3, we can bound Δ by $C^s \|Q\|_F^{2s}$, to obtain

$$\vdash^{Q}_{2s} \left\{ \frac{k^2}{n^2} \sum_{i\neq j\in C_\ell} \Big((x-y)^\top \Sigma^{\dagger/2} Q \Sigma^{\dagger/2}(x-y) \Big)^s \leqslant (2Cs)^{2s} \|Q\|_F^2 \right\} \tag{2.125}$$

Substituting $Q \to \Sigma_\ell^{1/2} Q \Sigma_\ell^{1/2}$, and observing that $\Sigma_\ell^{1/2}\Sigma_\ell^{\dagger/2}(x_i - x_j) = (x_i - x_j)$, we can conclude

$$\vdash^{Q}_{2s} \left\{ \frac{k^2}{n^2} \sum_{i\neq j\in C_\ell} \Big((x-y)^\top Q (x-y) \Big)^s \leqslant (2Cs)^{2s} \|\Sigma_\ell^{1/2} Q \Sigma_\ell^{1/2}\|_F^2 \right\} \tag{2.126}$$

A similar argument holds for 4-tuples of samples. The final claim about certifiably bounded variance property follows by a similar bound on the empirical moments of the distribution along with Lemma 2.7.9. This concludes the proof. □

2.11 Polynomial Approximators for Thresholds

We will use elementary approximation theory to construct the polynomial.

Fact 2.11.1 (Jackson's Theorem)**.** *Let $f : [-1,1] \to \mathcal{R}$ be continuous. Let the modulus of continuity of f be defined as $\omega(\delta) = \sup_{x,y\in[-1,1]} \{|f(x) - f(y)| \leqslant \delta\}$ for every $\delta > 0$. Then,*

for every b, there's a degree b polynomial p such that for every $x \in [-1,1]$,

$$|p(x) - f(x)| \leqslant 6\omega(1/b)\,.$$

The following lemma gives an "amplifying polynomial" as in [DRST09] and is an easy consequence of Chernoff bounds.

Fact 2.11.2 (Claim 4.3 in [DRST09])**.** *Let $A_q(u) = \sum_{j \geqslant q/2} \binom{q}{j} \left(\frac{1+u}{2}\right)^j \left(\frac{1-u}{2}\right)^{q-j}$. Then, A_q is a degree q polynomial that satisfies:*

1. $A_q(u) \in [1 - e^{q/6}, 1]$ *for all* $u \in [3/5, 1]$,
2. $A_q(u) \in [0, e^{-q/6}]$ *for all* $u \in [-1, -3/5]$,
3. $A_q(u) \in [0, 1]$ *for all* $u \in [-1, 1]$.

Proof of Lemma 2.3.9. Let $\mathsf{thr} : [0,1] \to [0,1]$ be any function that is 0 on $[0,c]$, 1 on $[2c,1]$

Consider the piecewise linear function $f : [0,1] \to [0,1]$ such that $f(x) = 0$ whenever $|x| \leqslant c$, $f(x) = 1$ for $|x| \geqslant 2c$ and $f(x) = \frac{(x-c)}{c}$ otherwise. Then, f is continuous. Further, the modulus of continuity, $\omega(\delta)$ for f is at most $\frac{1}{c\delta}$.

Taking $q = 25/c$ and applying Fact 2.11.1 yields a polynomial $J(t)$ of degree at most q such that:

$$\max_{t \in [-1,1]} |J(t) - f(t)| \leqslant 1/4\,.$$

We now "amplify" this polynomial to get the final construction.

Let $p(t) = \left(A_r(8/5 J(t) - 4/5)\right)^2$ for $r = 15\log(1/\eta)$. Then, the argument of A_r in $p(t)$ lies in $[3/5, 1]$ whenever $t \geqslant 2c$ and in $[-1, -3/5]$ whenever $t \in [0, c]$. Thus, applying Fact 2.11.2 yields that:

$$\sup_{t \in [0,c] \cup [2c,1]} |p(t) - \mathsf{thr}(t)| \leqslant 2e^{-r/6} \leqslant \eta\,.$$

□

2.12 TV-Close Subgaussian Distributions with Arbitrarily Far Parameters

We give a simple example of a pair of (one-dimensional) subgaussian distributions that are $(1-\eta)$-close in TV-distance for some $\eta < 1/2$ while have an arbitrarily separated variances.

For $i = 1, 2$, let $\mathcal{D}_i$ be the distribution on $\mathcal{R}$ that outputs 0 with probability $\eta < 1/2$ and a sample from Gaussian $\mathcal{N}(0, \sigma_i^2)$ otherwise. Observe that $\mathcal{D}_1, \mathcal{D}_2$ are clearly 2-subgaussian: $\mathbf{E}_{\mathcal{D}_i} x^2 = (1-\eta)\sigma_i^2$ while for every t, $\mathbf{E}_{\mathcal{D}_i} x^{2t} \leqslant \left(\frac{1}{(1-\eta)}\right)^t (\mathbf{E}_{\mathcal{D}_i} x^2)^t$. Thus, both $\mathcal{D}_1, \mathcal{D}_2$ are $C = \frac{1}{(1-\eta)} \leqslant 2$-subgaussian. Further, since $\mathbf{Pr}_{\mathcal{D}_i}[x = 0] \geqslant \eta$, it's immediate that $d_{\mathsf{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leqslant (1-\eta)$. However, since we can choose σ_1, σ_2 arbitrary, the variances of $\mathcal{D}_1, \mathcal{D}_2$ are arbitrarily far.

Observe, however, that both $\mathcal{D}_1, \mathcal{D}_2$ are *not* anti-concentrated in the construction above. Observe, further that when η gets close to 1 (instead of $\leqslant 1/2$), the constant C in Sub-gaussianity blows-up. Thus, if we fix C before-hand and look at all C-subgaussian distributions, then we can hope to prove TV-closeness implies parameter closeness when TV distance is small enough but not when it's close to 1.

Chapter 3

Robustly Learning a Mixture of k Arbitrary Gaussians

3.1 Introduction

Given a collection of observations and a class of models, the objective of a typical learning algorithm is to find the model in the class that best fits the data. The classical assumption is that the input data are i.i.d. samples generated by a statistical model in the given class. This is a simplifying assumption that is, at best, only approximately valid, as real datasets are typically exposed to some source of systematic noise. Robust statistics challenges this assumption by focusing on the design of *outlier-robust* estimators — algorithms that can tolerate a *constant fraction* of corrupted datapoints, independent of the dimension. Despite significant effort over several decades starting with important early works of Tukey and Huber in the 60s, even for the most basic high-dimensional estimation tasks, all known computationally efficient estimators were until fairly recently highly sensitive to outliers.

This state of affairs changed with two independent works from the TCS community [DKK+19, LRV16], which gave the first computationally efficient and outlier-robust learning algorithms for a range of "simple" high-dimensional probabilistic models. In particular, these works developed efficient robust estimators for a single high-dimensional Gaussian distribution with unknown mean and covariance. Since these initial algorithmic works [DKK+19, LRV16], we have witnessed substantial research progress on algorithmic aspects of robust high-dimensional estimation by several communities of researchers, including TCS, machine learning, and mathematical statistics. The reader is referred to [DK19] for a recent survey on the topic.

One of the main original motivations for the development of algorithmic robust statistics within the TCS community was the problem of learning high-dimensional Gaussian mixture models. A *Gaussian mixture model (GMM)* is a convex combination of Gaussian distributions, i.e., a distribution on $\mathcal{R}^d$ of the form $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, where the weights w_i, mean vectors μ_i, and covariance matrices Σ_i are unknown. GMMs are *the* most extensively studied latent variable model in the statistics and machine learning literatures, starting with the pioneering work of Karl Pearson in 1894 [Pea94], which introduced the method of moments in this context.

In the absence of outliers, a long line of work initiated by Dasgupta [Das99, AK05, VW04, AM05, BV08] gave efficient clustering algorithms for GMMs under various separation assumptions. Subsequently, efficient learning algorithms were obtained [KMV10, MV10, BS15, HP15] under minimal information-theoretic conditions. Specifically, Moitra and Valiant [MV10] and Belkin and Sinha [BS15] designed the first polynomial-time learning algorithms for arbitrary Gaussian mixtures with any fixed number of components. These works qualitatively characterized the complexity of this fundamental learning problem in the noiseless setting. Alas, all aforementioned algorithms are very fragile in the presence of corrupted data. Specifically, a *single* outlier can completely compromise their performance.

Developing efficient learning algorithms for high-dimensional GMMs in the more realistic *outlier-robust* setting — the focus of the current paper — has turned out to be significantly more challenging. This was both one of the original motivations and the main open problem in the initial robust statistics works [DKK+19, LRV16]. We note that [DKK+19] developed a robust density estimation algorithm for mixtures of *spherical* Gaussians — a very special case of our problem where the covariance of each component is a multiple of the identity — and highlighted a number of key technical obstacles that need to be overcome in order to handle the general case. Since then, a number of works have made algorithmic progress on important special cases of the general problem. These include faster robust clustering for the spherical case under minimal separation conditions [HL18, KSS18, DKS18], robust clustering for separated (and potentially non-spherical) Gaussian mixtures [BK20b, DHKK20], and robustly learning *uniform* mixtures of two arbitrary Gaussian components [Kan20].

This progress notwithstanding, the algorithmic task of robustly learning a mixture of a constant number (or even two) arbitrary Gaussians (with arbitrary weights) has remained a central open problem in this field, as highlighted recently [DVW19].

This discussion motivates the following question, whose resolution is the main result of this work:

Question 2. *Is there a poly$(d, 1/\varepsilon)$-time robust GMM learning algorithm, in the presence of an ε-fraction of outliers, that has a dimension-independent error guarantee, for an arbitrary mixture of any constant number of arbitrary Gaussians on $\mathcal{R}^d$?*

3.1.1 Our Results

To formally state our main result, we define the model of robustness we study. We focus on the following standard data corruption model that generalizes Huber's contamination model [Hub64].

Definition 3.1.1 (Total Variation Contamination Model)**.** *Given a parameter $0 < \epsilon < 1/2$ and a class of distributions $\mathcal{F}$ on $\mathbb{R}^d$, the adversary operates as follows: The algorithm specifies the number of samples n. The adversary knows the true target distribution $X \in \mathcal{F}$ and selects a distribution F such that $d_{\mathsf{TV}}(F, X) \leqslant \varepsilon$. Then n i.i.d. samples are drawn from F and are given as input to the algorithm.*

Intuitively, the parameter ϵ in Definition 3.1.1 quantifies the power of the adversary. The total variation contamination model is strictly stronger than Huber's contamination model. Recall that in Huber's model [Hub64], the adversary generates samples from a mixture distribution F of the form $F = (1 - \epsilon)X + \epsilon N$, where X is the unknown target distribution and N is an adversarially chosen noise distribution. That is, in Huber's model the adversary is only allowed to add outliers.

Remark 66. The *strong contamination model* [DKK+19] is a strengthening of the total variation contamination, where an adversary can see the clean samples and then arbitrarily replace an ε-fraction of these points to obtain an ε-corrupted set of samples. Our robust learning algorithm succeeds in this strong contamination model, with the additional requirement that we can obtain two sets of independent ε-corrupted samples from the unknown mixture.

In the context of robustly learning GMMs, we want to design an efficient algorithm with the following performance: Given a sufficiently large set of samples from a distribution that is ε-close in total variation distance to an unknown GMM $\mathcal{M}$ on $\mathcal{R}^d$, the algorithm outputs a hypothesis GMM $\widehat{\mathcal{M}}$ such that with high probability the total variation distance $d_{\mathsf{TV}}(\widehat{\mathcal{M}}, \mathcal{M})$ is small. Specifically, we want $d_{\mathsf{TV}}(\widehat{\mathcal{M}}, \mathcal{M})$ to be only a function of ε and independent of the underlying dimension d.

The main result of this paper is the following:

Theorem 67 (Main Result, See Corollary 3.6.1)**.** *There is an algorithm with the following be-*

havior: Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)}\mathrm{poly}(\log(1/\varepsilon))$ samples from a distribution F on $\mathcal{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown target k-GMM $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time $\mathrm{poly}(n)\mathrm{poly}_k(1/\varepsilon)$ and outputs a k-GMM hypothesis $\widehat{\mathcal{M}} = \sum_{i=1}^{k} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ such that with high probability we have that $d_{\mathsf{TV}}(\widehat{\mathcal{M}}, \mathcal{M}) \leqslant g(\varepsilon, k)$. Here $g : \mathcal{R}_+ \times \mathbb{Z}_+ \to \mathcal{R}_+$ is a function such that $\lim_{\varepsilon \to 0} g(\varepsilon, k) = 0$.

Theorem 67 gives the first polynomial-time *robust proper learning* algorithm, with dimension-independent error guarantee, for *arbitrary* k-GMMs, for any fixed k. This is the first polynomial-time algorithm for this problem, even for $k = 2$.

Discussion Before proceeding, we make a few important remarks about Theorem 67.

1. *Sample Complexity and Runtime*: Our algorithm succeeds whenever the sample size n satisfies $n \geqslant n_0 = d^{O(k)}/\mathrm{poly}(\epsilon)$. The running time of our algorithm is $\mathrm{poly}(n)\mathrm{poly}_k(1/\varepsilon)$. Statistical query lower bounds [DKS17] suggest that $d^{\Omega(k)}$ samples are necessary for efficiently learning GMMs, even for approximation to constant accuracy in the simpler setting without outliers and under the more restrictive *clustering* setting (where components are pairwise well-separated in total variation distance). This provides some evidence that the sample-time tradeoff achieved by Theorem 67 is qualitatively optimal (within absolute constant factors in the exponent). We note that the algorithm establishing Theorem 67 works in the standard bit-complexity model of computation and its running time is polynomial in the bit-complexity of the input parameters.

   In the noiseless case, the first polynomial-time learning algorithm for k-GMMs on $\mathcal{R}^d$ was given in [MV10, BS15]. In particular, the sample complexity and running time of the [MV10] algorithm is $(d/\varepsilon)^{q(k)}$, for some function $q(k) = k^{\Omega(k)}$. We observe that our running time and sample complexity are exponentially better than the guarantees for the noiseless case in [MV10, BS15]. Moreover, the [MV10, BS15] algorithms are very sensitive to outliers and an entirely new approach is required to obtain an efficient robust learning algorithm.

2. *Handling Arbitrary Weights*: The algorithm of Theorem 67 succeeds *without any assumptions* on the weights of the mixture components. We emphasize that this is an important feature and not a technicality. Prior work [BK20b, DHKK20, Kan20], as well as the concurrent work [LM21], cannot handle the case of general weights — even for the case of $k = 2$ components. In fact, for the special case of uniform weights, we give a simpler algorithm for robustly learning GMMs (presented in Theorem 78). This algorithm naturally generalizes to give a sample complexity and running time that grows *exponentially* in

$1/w_{\min}$, where $w_{\min}$ is the minimum weight of any component in the mixture. Handling the general case (i.e., obtaining a fully polynomial-time algorithm, not incurring an exponential cost in $1/w_{\min}$) requires genuinely new algorithmic ideas and is one of the key technical innovations in the proof of Theorem 67.

3. *Handling Arbitrary Covariances*: The algorithm of Theorem 67 does not require assumptions on the variances of the component covariances, modulo basic limitations posed by numerical computation issues. Specifically, our algorithm works even if some of the component covariances are rank-deficient (i.e., have directions of 0 variance) with running time scaling polynomially in the bit-complexity of the unknown component means and covariances. Such a dependence on the bit complexity of the input parameters is unavoidable – there exist[1] examples of rank-deficient covariances with irrational entries such that the total variation distance between the corresponding Gaussian and every Gaussian with covariance matrix of rational entries is the maximum possible value of one.

4. *Error Guarantee*: The function g quantifying the final error guarantee of our basic algorithm is $g(\varepsilon, k) = 1/(\log(1/\varepsilon))^{C_k}$, for some function C_k that goes to 0 when k increases. Importantly, for any fixed k, the final error guarantee of our algorithm depends only on ε, tends to 0 as $\epsilon \to 0$ and is independent of the dimension d. In Theorem 68, we show that, by modifying our algorithm, we can obtain improved error – scaling as a fixed polynomial in ϵ. This turns out to be quantitatively close to best possible for any robust proper learning algorithm.

Our work is most closely related to the recent paper by Kane [Kan20], which gave a polynomial-time robust learning algorithm for the *uniform* $k = 2$ case, i.e., the case of two *equal weight* components, and the polynomial time algorithms [BK20b, DHKK20] for the problem under the (strong) assumption that the component Gaussians are pairwise well-separated in total variation distance.

Our algorithm builds on the ideas in the works [BK20b, DHKK20] that gave efficient clustering algorithms for any fixed number k of components, under the crucial assumption that the components have pairwise total variation distance close to 1. In this case, the above works actually succeed in efficiently *clustering* the input sample into k groups, such that each group contains the samples generated from one of the Gaussians, up to some small misclassification error. In contrast, the main challenge in this work is the information-theoretic impossibility of clustering in our setting where there are no separation assumptions. As we will explain in the

[1]For e.g., for unit vector $v = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, 0, \ldots, 0)$ and for every choice of rational covariance Σ, the total variation distance between $\mathcal{N}(0, I - vv^\top)$ and $\mathcal{N}(0, \Sigma)$ is the maximum possible 1.

proceeding discussion, while we draw ideas from Chapter 2, a number of significant conceptual and technical challenges need to be overcome in the non-clusterable setting.

Improvements to Theorem 67. We now describe refinements of our main theorem.

Improving Error to a Fixed Polynomial in ϵ. It turns out that the inverse poly-logarithmic accuracy (in $1/\epsilon$) in the final error guarantee of Theorem 67 can be traced to an exhaustive search subroutine in our novel tensor decomposition subroutine and probability of success of our rounding algorithm in our partial clustering routine. Via natural (and conceptually simple) quantitative improvements to these two ingredients, we obtain an algorithm achieving the qualitatively nearly best possible error of $\mathrm{poly}_k(\epsilon)$. Specifically, we show:

Theorem 68 (Robustly Learning k-Mixtures with $\mathrm{poly}(\epsilon)$-error, Informal, see Corollary 85)**.** *There is an algorithm with the following behavior: Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)}\mathrm{poly}_k(1/\epsilon)$ samples from a distribution F on $\mathcal{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown target k-GMM $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time $\mathrm{poly}(n)\mathrm{poly}_k(1/\varepsilon)$ and outputs a k-GMM hypothesis $\widehat{\mathcal{M}} = \sum_{i=1}^k \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ such that with high probability we have that $d_{\mathsf{TV}}(\widehat{\mathcal{M}}, \mathcal{M}) \leqslant \mathcal{O}(\varepsilon^{c_k})$, where c_k depends only on k.*

Robust Parameter Recovery. Finally, we show that the *same* algorithm as in Theorems 67 and 85 actually implies that the recovered mixture of Gaussians is close in *parameter* distance to the unknown target mixture. Such parameter estimation results are usually stated under the assumption that every pair of components of the unknown mixture are separated in total variation distance. In this work, we provide a stronger version of this parameter estimation guarantee.

More specifically, in the theorem below, we prove that whenever the components of the input mixture can be clustered together into some groups such that all mixtures in a group are close (and thus, indistinguishable), there exists a similar clustering of the output mixture such that all parameters (weight, mean, and covariances) of each cluster are close within $\mathrm{poly}_k(\epsilon)$ in total variation distance. In particular this means that for each significant component of the input mixture, there is a component of the output mixture with very close parameters.

We note that [LM21] gave a parameter estimation guarantee (under additional assumptions on the mixture weights and component variances) whenever every pair of components in the unknown mixture are $f(k)$-far in total variation distance, where f can be any function of k, but the choice of f affects the exponent in the running time and error guarantee of the [LM21] algorithm.)

By strengthening one of the structural results in their argument, we establish the following:

Theorem 69 (Parameter Recovery, See Theorem 3.9.1)**.** *Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)}\mathrm{poly}_k(1/\varepsilon)$ samples from a distribution F on $\mathcal{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown target k-GMM $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time $\mathrm{poly}(n)\mathrm{poly}_k(1/\varepsilon)$ and outputs a k'-GMM hypothesis $\widehat{\mathcal{M}} = \sum_{i=1}^{k'} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ with $k' \leqslant k$ such that with high probability we have that there exists a partition of $[k]$ into $k'+1$ sets $R_0, R_1, \ldots, R_{k'}$ such that*

1. *Let $W_i = \sum_{j \in R_i} w_j$, $i \in \{0, 1, \ldots, k'\}$. Then, for all $i \in [k']$, we have that*

$$|W_i - \widehat{w}_i| \leqslant \mathrm{poly}_k(\epsilon), \text{ and}$$
$$d_{\mathsf{TV}}(\mathcal{N}(\mu_j, \Sigma_j), \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)) \leqslant \mathrm{poly}_k(\epsilon) \quad \forall j \in R_i \ .$$

2. *The total weight of exceptional components in R_0 is $W_0 \leqslant \mathrm{poly}_k(\epsilon)$.*

If we assume additionally that any pair of components in the unknown mixture has total variation distance at least $\mathrm{poly}_k(\varepsilon)$, then the following result follows directly from Theorem 69.

Corollary 3.1.2. *Let $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$ be an unknown target k-GMM satisfying the following conditions: (i) $d_{\mathsf{TV}}(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j)) \geqslant \varepsilon^{f_1(k)}$ for all $i \neq j$, and (ii) $S = \{i \in [k] : w_i \geqslant \varepsilon^{f_2(k)}\}$ is a subset of $[k]$, where $f_1(k), f_2(k)$ are sufficiently small functions of k. Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)}\mathrm{poly}_k(1/\varepsilon)$ samples from a distribution F on $\mathcal{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, there exists an algorithm that runs in time $\mathrm{poly}(n)\mathrm{poly}_k(1/\varepsilon)$ and outputs a k'-GMM hypothesis $\widehat{\mathcal{M}} = \sum_{i=1}^{k'} \widehat{w}_i \mathcal{N}(\widehat{\mu}_i, \widehat{\Sigma}_i)$ with $k' \leqslant k$ such that with high probability there exists a bijection $\pi : S \to [k']$ satisfying the following: For all $i \in S$, it holds that*

$$|w_i - \widehat{w}_{\pi(i)}| \leqslant \mathrm{poly}_k(\epsilon)$$
$$d_{\mathsf{TV}}(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\widehat{\mu}_{\pi(i)}, \widehat{\Sigma}_{\pi(i)})) \leqslant \mathrm{poly}_k(\epsilon).$$

We note that both the pairwise separation between the components and the lower bounds on the weights in Corollary 3.1.2 scale as a fixed polynomial in ε (for fixed k), which is qualitatively information-theoretically necessary.

3.1.2 Organization

The structure of this chapter is as follows: In Section 4.2, we provide relevant background and technical facts. In Section 3.3, we describe and analyze our new tensor decomposition algorithm. In Section 3.4, we use a sum-of-squares based approach to partially cluster a mixture. In Section 3.5, we give a spectral separation algorithm to identify thin components. In Section 3.6, we put all these pieces together to prove Theorem 67. In Section 3.7, we present a refinement of our partial clustering procedure that improves the probability of success to a constant independent of the minimum weight of any component in the input mixture. In Section 3.8, we present an efficient algorithm that replaces an exhaustive search subroutine in the tensor decomposition algorithm and combines it with the improved partial clustering subroutine to get a $\mathrm{poly}_k(\epsilon)$-error guarantee for robust proper learning of Gaussian mixtures and prove Theorem 68. Finally, in Section 3.9, we show that our algorithm in fact achieves the stronger parameter estimation guarantees and prove Theorem 69.

3.2 Preliminaries

Basic Notation. For a vector v, we use $\|v\|_2$ to denote its Euclidean norm. For an $n \times m$ matrix M, we use $\|M\|_{\mathrm{op}} = \max_{\|x\|_2=1} \|Mx\|_2$ to denote the operator norm of M and $\|M\|_F = \sqrt{\sum_{i,j} M_{i,j}^2}$ to denote the Frobenius norm of M. We sometimes use the notation $M(i,j)$ to index the corresponding entries in M. For an $n \times n$ symmetric matrix M, we use $\succeq$ to denote the PSD/Loewner ordering over eigenvalues of M and $\mathrm{tr}\,(M) = \sum_{i \in [n]} M_{i,i}$ to denote the trace of M. We use $U\Lambda U^\top$ to denote the eigenvalue decomposition, where U is an $n \times n$ matrix with orthonormal columns and Λ is the $n \times n$ diagonal matrix of the eigenvalues. We use $M^\dagger = U\Lambda^\dagger U^\top$ to denote the Moore-Penrose pseudoinverse, where $\Lambda^\dagger$ inverts the non-zero eigenvalues of M. If $M \succeq 0$, we use $M^{\dagger/2} = U\Lambda^{\dagger/2}U^\top$ to denote taking the square-root of the inverted non-zero eigenvalues.

For $d \times d$ matrices A, B, the Kronecker product of A, B, denoted by $A \otimes B$, is indexed by $(i,j),(k,\ell) \in [d] \times [d]$ and has entries $(A \otimes B)((i,j),(k,\ell)) = A(i,k)B(j,\ell)$. We will equip every tensor T with the norm $\|\cdot\|_F$ that simply corresponds to the ℓ_2-norm of any flattening of T to a vector. The notation $T(\cdot,\cdot,x,y)$ is used to denote collapsing two modes of the tensor by plugging in x, y. For a positive integer ℓ and vector v, we also use $v^{\otimes \ell} = \underbrace{v \otimes v \ldots \otimes v}_{\ell \text{times}}$.

We use the notation $\mathcal{M} = \sum_{i \in [k]} w_i \mathcal{N}(\mu_i, \Sigma_i)$ to represent a k-mixture of Gaussians. The to-

tal variation distance between two probability distributions on $\mathcal{R}^d$ with densities p, q is defined as $d_{\mathsf{TV}}(p,q) = \frac{1}{2}\int_{\mathcal{R}^d} |p(x) - q(x)|dx$. We also use $\mathbb{E}[\cdot]$, $\mathbf{Var}\cdot$ and $\mathsf{Cov}(\cdot)$ to denote the expectation, variance and covariance of a random variable.

For a finite dataset X, we will use $Z \in_u X$ to denote that Z is the uniform distribution on X. We will sometimes use the term mean (resp. covariance) of X to refer to $\mathbb{E}_{Z\in_u X}[Z]$ (resp. $\mathsf{Cov}_{Z\in_u X}(Z)$).

3.2.1 Gaussian Background

The first few facts in this subsection can be found in Kane [Kan20].

Fact 3.2.1. *The total variation distance between two Gaussians $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ can be bounded above as follows:*

$$d_{\mathsf{TV}}\left(\mathcal{N}(\mu_1,\Sigma_1), \mathcal{N}(\mu_2,\Sigma_2)\right) = \mathcal{O}\Big((\mu_1-\mu_2)^\top \Sigma_1^\dagger (\mu_1-\mu_2) + \|\Sigma_1^{\dagger/2}(\Sigma_2-\Sigma_1)\Sigma_1^{\dagger/2}\|_F\Big) \ .$$

Fact 3.2.2 (Theorem 2.4 in [Kan20])**.** *Let $\mathcal{D}$ be a distribution on $\mathcal{R}^{d\times d}$, where $\mathcal{D}$ is supported on the subset of $\mathcal{R}^{d\times d}$ corresponding to the set of symmetric PSD matrices. Suppose that $\mathbf{E}[\mathcal{D}] = \Sigma$ and that for any symmetric matrix A we have that $\mathbf{Var}\mathrm{tr}(AX) = \mathcal{O}\Big(\sigma^2\|\Sigma^{1/2}A\Sigma^{1/2}\|_F^2\Big)$. Then, for $\epsilon \ll \sigma^{-2}$, there exists a polynomial-time algorithm that given sample access to an ε-corrupted set of samples from $\mathcal{D}$ returns a matrix $\hat{\Sigma}$ such that with high probability $\|\Sigma^{-1/2}(\Sigma - \hat{\Sigma})\Sigma^{-1/2}\|_F = O(\sigma\sqrt{\varepsilon})$.*

Fact 3.2.3 (Proposition 2.5 in [Kan20])**.** *Let $G \sim \mathcal{N}(\mu, \Sigma)$ be a Gaussian in $\mathcal{R}^d$. Then, we have that*

$$\mathbf{E}[G^{\otimes m}](i_1,\ldots,i_m) = \sum_{\substack{\text{Partitions } P \text{ of } [m] \\ \text{into sets of size 1 and 2}}} \bigotimes_{\{a,b\}\in P} \Sigma(i_a,i_b) \bigotimes_{\{c\}\in P} \mu(i_c) \ .$$

We will work with the coefficient tensors of d-dimensional Hermite polynomials:

Definition 3.2.4 (Hermite Tensors)**.** *Define the degree-m Hermite polynomial tensor as*

$$h_m(x) := \sum_{\substack{\text{Partitions } P \text{ of } [m] \\ \text{into sets of size 1 and 2}}} \bigotimes_{\{a,b\}\in P} -I(i_a,i_b) \bigotimes_{\{c\}\in P} x(i_c) \ .$$

We will use the following fact that relates Hermite moments to the raw moments of any

distribution.

Fact 3.2.5 (Hermite vs Raw Moments)**.** *For any real-valued random variable u, and $m \in \mathbb{N}$, $\max_{i \leqslant m} |\mathbf{E} u^i - \mathbf{E}_{z \sim \mathcal{N}(0,1)} z^i| \leqslant 2^{O(m)} \max_{i \leqslant m} |\mathbf{E} h_m(u)|$. Similarly, $\max_{i \leqslant m} |\mathbf{E} h_m(u)| \leqslant 2^{O(m)} \max_{i \leqslant m} |\mathbf{E} u^i - \mathbf{E}_{z \sim \mathcal{N}(0,1)} z^i|$.*

Fact 3.2.6 (Lemma 2.7 in [Kan20])**.** *If $G \sim \mathcal{N}(\mu, I + \Sigma)$, then we have that*

$$\mathbf{E}[h_m(G)] = \sum_{\substack{\text{Partitions } P \text{ of } [m] \\ \text{into sets of size 1 and 2}}} \bigotimes_{\{a,b\} \in P} \Sigma(i_a, i_b) \bigotimes_{\{c\} \in P} \mu(i_c) \ .$$

Fact 3.2.7 (Lemma 2.8 in [Kan20])**.** *If $G \sim \mathcal{N}(\mu, I + \Sigma)$, then $\mathbf{E}[h_m(G) \otimes h_m(G)]$ is equal to*

$$\sum_{\substack{\text{Partitions } P \text{ of } [2m] \\ \text{into sets of size 1 and 2}}} \bigotimes_{\substack{\{a,b\} \in P \\ a,b \text{ in same half of } [2m]}} \Sigma(i_a, i_b) \bigotimes_{\substack{\{a,b\} \in P \\ a,b \text{ in different halves of } [2m]}} (I + \Sigma)(i_a, i_b) \bigotimes_{\{c\} \in P} \mu(i_c) \ .$$

Lemma 3.2.8 (Slight Strengthening of Lemma 5.2 in [Kan20])**.** *For $G \sim \mathcal{N}(\mu, \Sigma)$, the covariance matrix of $h_m(G)$ satisfies:*

$$\|\mathbf{Cov}(h_m(G))\|_{op} \leqslant \|\mathbf{E}[h_m(G) \otimes h_m(G)]\|_{op} = \mathcal{O}\left(m(1 + \|\Sigma\|_F + \|\mu\|_2)\right)^{2m} \ .$$

This follows from the proof of Lemma 5.2 in [Kan20] by noting that the number of terms in the sum is at most 2^m times the number of partitions of $[2m]$ into sets of size 1 and 2, which is at most $O(m)^{2m}$.

Next, we use upper and lower bounds on low-degree polynomials of Gaussian random variables. We defer the proof of the subsequent Lemma to Appendix 3.10.

Lemma 3.2.9 (Concentration of low-degree polynomials)**.** *Let T be a d-dimensional, degree-4 tensor such that $\|T\|_F \leqslant \Delta$ for some $\Delta > 0$ and let $x, y \sim \mathcal{N}(0, I)$. Then, with probability at least $1 - 1/\mathrm{poly}(d)$, the following holds:*

$$\|T(\cdot, \cdot, x, y)\|_F^2 \leqslant \mathcal{O}\Big(\log(d)\Delta^2\Big) \ .$$

Note that for any matrix M, $\langle M, x \otimes y \rangle$, where $x, y \sim \mathcal{N}(0, I)$, is a degree-2 polynomial in Gaussian random variables. As a result, we have the following anti-concentration inequality.

Lemma 3.2.10 (Anti-concentration of bi-linear forms, [CW01])**.** *Let M be a $d \times d$ matrix and*

let $x, y \sim \mathcal{N}(0, I)$. Then, for any $\zeta \in (0, 1)$, the following holds:

$$\mathbf{Pr}\left[\langle M, x \otimes y\rangle^2 \leqslant \zeta\, \mathbb{E}\left[\langle M, x \otimes y\rangle^2\right]\right] \leqslant \mathcal{O}\Big(\sqrt{\zeta}\Big) \ .$$

3.2.2 Sum-of-Squares Proofs and Pseudo-distributions

We refer the reader to the monograph [FKP+19] and the lecture notes [Bar] for a detailed exposition of the sum-of-squares method and its usage in average-case algorithm design.

Let $x = (x_1, x_2, \ldots, x_n)$ be a tuple of n indeterminates and let $\mathcal{R}[x]$ be the set of polynomials with real coefficients and indeterminates $x_1, \ldots, x_n$. We say that a polynomial $p \in \mathcal{R}[x]$ is a *sum-of-squares (sos)* if there exist polynomials $q_1, \ldots, q_r$ such that $p = q_1^2 + \cdots + q_r^2$.

Pseudo-distributions

Pseudo-distributions are generalizations of probability distributions. We can represent a discrete (i.e., finitely supported) probability distribution over $\mathcal{R}^n$ by its probability mass function $D\colon \mathcal{R}^n \to \mathcal{R}$ such that $D \geqslant 0$ and $\sum_{x \in \mathrm{supp}(D)} D(x) = 1$. Similarly, we can describe a pseudo-distribution by its mass function by relaxing the constraint $D \geqslant 0$ to passing certain low-degree non-negativity tests.

Concretely, a *level-ℓ pseudo-distribution* is a finitely-supported function $D : \mathcal{R}^n \to \mathcal{R}$ such that $\sum_x D(x) = 1$ and $\sum_x D(x) f(x)^2 \geqslant 0$ for every polynomial f of degree at most $\ell/2$. (Here, the summations are over the support of D.) A straightforward polynomial-interpolation argument shows that every level-∞-pseudo distribution satisfies $D \geqslant 0$ and is thus an actual probability distribution. We define the *pseudo-expectation* of a function f on $\mathcal{R}^n$ with respect to a pseudo-distribution D, denoted $\tilde{\mathbb{E}}_{D(x)} f(x)$, as

$$\tilde{\mathbb{E}}_{D(x)} f(x) = \sum_x D(x) f(x) \ . \tag{3.1}$$

The degree-ℓ moment tensor of a pseudo-distribution D is the tensor $\tilde{\mathbb{E}}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes \ell}$. In particular, the moment tensor has an entry corresponding to the pseudo-expectation of all monomials of degree at most ℓ in x. The set of all degree-ℓ moment tensors of probability distribution is a convex set. Similarly, the set of all degree-ℓ moment tensors of degree-d pseudo-distributions is also convex. Unlike moments of distributions, there is an efficient separation oracle for moment tensors of pseudo-distributions.

Fact 3.2.11 ([Sho87, Nes00, Las01, Par00]). *fact]fact:sos-separation-efficient For any $n, \ell \in \mathbb{N}$, the following set has an $n^{O(\ell)}$-time weak separation oracle (in the sense of [GLS81]):*

$$\left\{ \tilde{\mathbb{E}}_{D(x)}(1, x_1, x_2, \ldots, x_n)^{\otimes d} \mid \textit{degree-d pseudo-distribution } D \textit{ over } \mathcal{R}^n \right\} . \tag{3.2}$$

This fact, together with the equivalence of weak separation and optimization [GLS81], allows us to efficiently optimize over pseudo-distributions (approximately) — this algorithm is referred to as the sum-of-squares algorithm. The *level-ℓ sum-of-squares algorithm* optimizes over the space of all level-ℓ pseudo-distributions that satisfy a given set of polynomial constraints (defined below).

Definition 3.2.12 (Constrained pseudo-distributions). *Let D be a level-ℓ pseudo-distribution over $\mathcal{R}^n$. Let $\mathcal{A} = \{f_1 \geqslant 0, f_2 \geqslant 0, \ldots, f_m \geqslant 0\}$ be a system of m polynomial inequality constraints. We say that D satisfies the system of constraints $\mathcal{A}$ at degree r, denoted $D \sststile{r}{} \mathcal{A}$, if for every $S \subseteq [m]$ and every sum-of-squares polynomial h with $\deg h + \sum_{i \in S} \max\{\deg f_i, r\}$, we have that $\tilde{\mathbb{E}}_D h \cdot \prod_{i \in S} f_i \geqslant 0$.*

We write $D \models \mathcal{A}$ (without specifying the degree) if $D \sststile{0}{} \mathcal{A}$ holds. Furthermore, we say that $D \sststile{r}{} \mathcal{A}$ holds approximately if the above inequalities are satisfied up to an error of $2^{-n^\ell} \cdot \|h\| \cdot \prod_{i \in S} \|f_i\|$, where $\|\cdot\|$ denotes the Euclidean norm[2] of the coefficients of a polynomial in the monomial basis.

We remark that if D is an actual (discrete) probability distribution, then we have that $D \models \mathcal{A}$ if and only if D is supported on solutions to the constraints $\mathcal{A}$. We say that a system $\mathcal{A}$ of polynomial constraints is *explicitly bounded* if it contains a constraint of the form $\{\|x\|^2 \leqslant M\}$. The following fact is a consequence of [GLS81]:

Fact 3.2.13 (Efficient Optimization over Pseudo-distributions). *There exists an $(n+m)^{O(\ell)}$-time algorithm that, given any explicitly bounded and satisfiable system[3] $\mathcal{A}$ of m polynomial constraints in n variables, outputs a level-ℓ pseudo-distribution that satisfies $\mathcal{A}$ approximately.*

Basic Facts about Pseudo-Distributions. We will use the following Cauchy-Schwarz inequality for pseudo-distributions:

Fact 3.2.14 (Cauchy-Schwarz for Pseudo-distributions). *Let f, g be polynomials of degree at*

[2]The choice of norm is not important here because the factor 2^{-n^ℓ} swamps the effects of choosing another norm.
[3]Here, we assume that the bit complexity of the constraints in $\mathcal{A}$ is $(n+m)^{O(1)}$.

most d in indeterminate $x \in \mathcal{R}^d$. Then, for any degree-d pseudo-distribution $\tilde{\zeta}$, we have that $\tilde{\mathbb{E}}_{\tilde{\zeta}}[fg] \leqslant \sqrt{\tilde{\mathbb{E}}_{\tilde{\zeta}}[f^2]}\sqrt{\tilde{\mathbb{E}}_{\tilde{\zeta}}[g^2]}$.

Fact 3.2.15 (Hölder's Inequality for Pseudo-Distributions)**.** *Let f, g be polynomials of degree at most d in indeterminate $x \in \mathcal{R}^d$. Fix $t \in \mathbb{N}$. Then, for any degree-dt pseudo-distribution $\tilde{\zeta}$, we have that $\tilde{\mathbb{E}}_{\tilde{\zeta}}[f^{t-1}g] \leqslant \left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[f^t]\right)^{\frac{t-1}{t}} \left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[g^t]\right)^{1/t}$.*

Corollary 3.2.16 (Comparison of Norms)**.** *Let $\tilde{\zeta}$ be a degree-t^2 pseudo-distribution over a scalar indeterminate x. Then, we have that $\tilde{\mathbb{E}}[x^t]^{1/t} \geqslant \tilde{\mathbb{E}}[x^{t'}]^{1/t'}$ for every $t' \leqslant t$.*

Sum-of-squares proofs

Let $f_1, f_2, \ldots, f_r$ and g be multivariate polynomials in x. A *sum-of-squares proof* that the constraints $\{f_1 \geqslant 0, \ldots, f_m \geqslant 0\}$ imply the constraint $\{g \geqslant 0\}$ consists of polynomials $(p_S)_{S \subseteq [m]}$ such that

$$g = \sum_{S \subseteq [m]} p_S \cdot \Pi_{i \in S} f_i \,. \tag{3.3}$$

We say that this proof has *degree* ℓ if for every set $S \subseteq [m]$ the polynomial $p_S \Pi_{i \in S} f_i$ has degree at most ℓ. If there is a degree ℓ SoS proof that $\{f_i \geqslant 0 \mid i \leqslant r\}$ implies $\{g \geqslant 0\}$, we write:

$$\{f_i \geqslant 0 \mid i \leqslant r\} \vdash_{\ell} \{g \geqslant 0\} \,. \tag{3.4}$$

For all polynomials $f, g \colon \mathcal{R}^n \to \mathcal{R}$ and for all functions $F \colon \mathcal{R}^n \to \mathcal{R}^m$, $G \colon \mathcal{R}^n \to \mathcal{R}^k$, $H \colon \mathcal{R}^p \to \mathcal{R}^n$ such that each of the coordinates of the outputs are polynomials of the inputs, we have the following inference rules.

The first one derives new inequalities by addition or multiplication:

$$\frac{\mathcal{A} \vdash_{\ell} \{f \geqslant 0, g \geqslant 0\}}{\mathcal{A} \vdash_{\ell} \{f + g \geqslant 0\}}, \frac{\mathcal{A} \vdash_{\ell} \{f \geqslant 0\}, \mathcal{A} \vdash_{\ell'} \{g \geqslant 0\}}{\mathcal{A} \vdash_{\ell+\ell'} \{f \cdot g \geqslant 0\}} \,. \tag{3.5}$$

The next one derives new inequalities by transitivity:

$$\frac{\mathcal{A} \vdash_{\ell} \mathcal{B}, \mathcal{B} \vdash_{\ell'} C}{\mathcal{A} \vdash_{\ell \cdot \ell'} C} \,. \tag{3.6}$$

Finally, the last rule derives new inequalities via substitution:

$$\frac{\{F \geqslant 0\} \sststile{}{\ell} \{G \geqslant 0\}}{\{F(H) \geqslant 0\} \sststile{}{\ell \cdot \deg(H)} \{G(H) \geqslant 0\}} \,. \tag{substitution}$$

Low-degree sum-of-squares proofs are sound and complete if we take low-level pseudo-distributions as models. Concretely, sum-of-squares proofs allow us to deduce properties of pseudo-distributions that satisfy some constraints.

Fact 3.2.17 (Soundness)**.** *If $D \sdtstile{}{r} \mathcal{A}$ for a level-ℓ pseudo-distribution D and there exists a sum-of-squares proof $\mathcal{A} \sststile{}{r'} \mathcal{B}$, then $D \sdtstile{}{r \cdot r' + r'} \mathcal{B}$.*

If the pseudo-distribution D satisfies $\mathcal{A}$ only approximately, soundness continues to hold if we require an upper bound on the bit-complexity of the sum-of-squares $\mathcal{A} \sststile{}{r'} B$ (i.e., the number of bits required to write down the proof). In our applications, the bit complexity of all sum-of-squares proofs will be $n^{O(\ell)}$ (assuming that all numbers in the input have bit complexity $n^{O(1)}$). This bound suffices in order to argue about pseudo-distributions that satisfy polynomial constraints approximately.

The following fact shows that every property of low-level pseudo-distributions can be derived by low-degree sum-of-squares proofs.

Fact 3.2.18 (Completeness)**.** *Suppose that $d \geqslant r' \geqslant r$ and $\mathcal{A}$ is a collection of polynomial constraints with degree at most r, and $\mathcal{A} \sststile{}{} \{\sum_{i=1}^n x_i^2 \leqslant B\}$ for some finite B. Let $\{g \geqslant 0\}$ be a polynomial constraint. If every degree-d pseudo-distribution that satisfies $D \sdtstile{}{r} \mathcal{A}$ also satisfies $D \sdtstile{}{r'} \{g \geqslant 0\}$, then for every $\epsilon > 0$, there is a sum-of-squares proof $\mathcal{A} \sststile{}{d} \{g \geqslant -\epsilon\}$.*

Basic Sum-of-Squares Proofs. We will require the following basic SoS proofs.

Fact 3.2.19 (Operator norm Bound)**.** *Let A be a symmetric $d \times d$ matrix and v be a vector in $\mathbb{R}^d$. Then, we have that*

$$\sststile{v}{2} \left\{ v^\top A v \leqslant \|A\|_2 \|v\|_2^2 \right\} \,.$$

Fact 3.2.20 (SoS Hölder's Inequality)**.** *Let f_i, g_i, for $1 \leqslant i \leqslant s$, be scalar-valued indeterminates. Let p be an even positive integer. Then,*

$$\sststile{f,g}{p^2} \left\{ \left(\frac{1}{s} \sum_{i=1}^s f_i g_i^{p-1} \right)^p \leqslant \left(\frac{1}{s} \sum_{i=1}^s f_i^p \right) \left(\frac{1}{s} \sum_{i=1}^s g_i^p \right)^{p-1} \right\} \,.$$

Observe that using $p = 2$ above yields the SoS Cauchy-Schwarz inequality.

Fact 3.2.21 (SoS Almost Triangle Inequality)**.** *Let $f_1, f_2, \ldots, f_r$ be indeterminates. Then, we have that*

$$\left|\frac{f_1, f_2, \ldots, f_r}{2t}\right. \left\{ \left(\sum_{i \leqslant r} f_i \right)^{2t} \leqslant r^{2t-1} \left(\sum_{i=1}^{r} f_i^{2t} \right) \right\} .$$

Fact 3.2.22 (SoS AM-GM Inequality, see Appendix A of [BKS15])**.** *Let $f_1, f_2, \ldots, f_m$ be indeterminates. Then, we have that*

$$\left|\frac{f_1, f_2, \ldots, f_m}{m}\right. \left\{ \left(\frac{1}{m} \sum_{i=1}^{n} f_i \right)^{m} \geqslant \Pi_{i \leqslant m} f_i \right\} .$$

We defer the proofs of the two subsequent lemmas to Appendix 3.10.

Lemma 3.2.23 (Spectral SoS Proofs)**.** *Let A be a $d \times d$ matrix. Then for d-dimensional vector-valued indeterminate v, we have:*

$$\left|\frac{v}{2}\right. \left\{ v^\top A v \leqslant \|A\|_2 \|v\|_2^2 \right\} .$$

Fact 3.2.24 (Cancellation within SoS, Lemma 9.2 [BK20b])**.** *Let a, C be scalar-valued indeterminates. Then,*

$$\{a \geqslant 0\} \cup \left\{ a^t \leqslant C a^{t-1} \right\} \left|\frac{a, C}{2t}\right. \left\{ a^{2t} \leqslant C^{2t} \right\} .$$

Lemma 3.2.25 (Frobenius Norms of Products of Matrices)**.** *Let B be a $d \times d$ matrix valued indeterminate for some $d \in \mathbb{N}$. Then, for any $0 \preceq A \preceq I$,*

$$\left|\frac{B}{2}\right. \left\{ \|AB\|_F^2 \leqslant \|B\|_F^2 \right\} ,$$

and,

$$\left|\frac{B}{2}\right. \left\{ \|BA\|_F^2 \leqslant \|B\|_F^2 \right\} ,$$

3.2.3 Analytic Properties of Gaussian Distributions

The following definitions and results describe the analytic properties of Gaussian distributoins that we will use. We also state the guarantees of known robust estimation algorithms for estimat-

ing the mean, covariance and moment tensors of Gaussian mixtures here.

Certifiable Subgaussianity. We will make essential use of the following definition.

Definition 3.2.26 (Certifiable Subgaussianity (Definition 5.1 in [KS17]))**.** *For $t \in \mathbb{N}$ and an absolute constant $C > 0$, a distribution $\mathcal{D}$ on $\mathcal{R}^d$ is said to be t-certifiably C-subgaussian if for every even $t' \leqslant t$, we have that*

$$\left|\frac{v}{t'}\right\} \left\{ \mathop{\mathbb{E}}_{\mathcal{D}} \left[\langle x, v \rangle^{t'} \right] \leqslant (Ct')^{t'/2} \left(\mathop{\mathbb{E}}_{\mathcal{D}} \left[\langle x, v \rangle^2 \right]^{t'/2} \right) \right\} .$$

Fact 3.2.27 (Mixtures of Certifiably Subgaussian Distributions, Analogous to Lemma 5.4 in [KS17])**.** *Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_q$ be t-certifiably C-subgaussian distributions on $\mathcal{R}^d$. Let $p_1, p_2, \ldots, p_q$ be non-negative weights such that $\sum_i p_i = 1$ and $p = \min_{i \leqslant q} p_i$. Then, the mixture $\sum_i p_i \mathcal{D}_i$ is t-certifiably C/p-subgaussian.*

Certifiable Anti-Concentration. The first is *certifiable anti-concentration* — an SoS formulation of classical anti-concentration inequalities — that was introduced in [KKK19, RY20a].

In order to formulate certifiable anti-concentration, we start with a univariate even polynomial p that serves as a uniform approximation to the delta function at 0 in an interval around 0. Such polynomials are constructed in [KKK19, RY20a]. Let $q_{\delta,\Sigma}(x, v)$ be a multivariate (in v) polynomial defined by $q_{\delta,\Sigma}(x, v) = \left(v^\top \Sigma v\right)^{2s} p_{\delta,\Sigma}\left(\frac{\langle x, v \rangle}{\sqrt{v^\top \Sigma v}}\right)$. Since $p_{\delta,\Sigma}$ is an even polynomial, $q_{\delta,\Sigma}$ is a polynomial in v.

Definition 3.2.28 (Certifiable Anti-Concentration)**.** *A mean-0 distribution D with covariance Σ is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated if for $q_{\delta,\Sigma}(x, v)$ defined above, there exists a degree-$2s$ sum-of-squares proof of the following two unconstrained polynomial inequalities in indeterminate v:*

$$\left\{ \langle x, v \rangle^{2s} + \delta^{2s} q_{\delta,\Sigma}(x, v)^2 \geqslant \delta^{2s} \left(v^\top \Sigma v\right)^{2s} \right\} , \left\{ \mathop{\mathbb{E}}_{x \sim D} \left[q_{\delta,\Sigma}(x, v)^2 \right] \leqslant C\delta \left(v^\top \Sigma v\right)^{2s} \right\} .$$

An isotropic subset $X \subseteq \mathcal{R}^d$ is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated if the uniform distribution on X is $2s$-certifiably $(\delta, C\delta)$-anti-concentrated.

Remark 70. The function $s(\delta)$ can be taken to be $O(\frac{1}{\delta^2})$ for standard Gaussian distribution and the uniform distribution on the unit sphere (see [KKK19] and [BK20a]).

Certifiable Hypercontractivity. Next, we define *certifiable hypercontractivity* of degree-2 polynomials that formulates (within SoS) the fact that higher moments of degree-2 polynomials of distributions (such as Gaussians) can be bounded in terms of appropriate powers of their 2nd moment.

Definition 3.2.29 (Certifiable Hypercontractivity)**.** *An isotropic distribution $\mathcal{D}$ on $\mathcal{R}^d$ is said to be h-certifiably C-hypercontractive if there is a degree-h sum-of-squares proof of the following unconstrained polynomial inequality in $d \times d$ matrix-valued indeterminate Q:*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Q x\right)^h\right] \leqslant (Ch)^h \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Q x\right)^2\right]\right)^{h/2}.$$

A set of points $X \subseteq \mathcal{R}^d$ is said to be C-certifiably hypercontractive if the uniform distribution on X is h-certifiably C-hypercontractive.

Hypercontractivity is an important notion in high-dimensional probability and analysis on product spaces [O'D14]. Kauers, O'Donnell, Tan and Zhou [KOTZ14] showed certifiable hypercontractivity of Gaussians and more generally product distributions with subgaussian marginals. Certifiable hypercontractivity strictly generalizes the better known *certifiable subgaussianity* property (studied first in [KS17]) that controls higher moments of linear polynomials.

Observe that the definition above is affine invariant. In particular, we immediately obtain:

Fact 3.2.30. *Given $t \in \mathbb{N}$, if a random variable x on $\mathcal{R}^d$ has t-certifiable C-hypercontractive degree-2 polynomials, then so does Ax for any $A \in \mathcal{R}^{d\times d}$.*

As observed in [KS17], the Gaussian distribution is t-certifiably 1-subgaussian and t-certifiable 1-hypercontractive for every t. Next, we establish certifiable hypercontractivity for mixtures of Gaussians. We defer the proofs to Appendix 3.10.

Lemma 3.2.31 (Shifts Cannot Decrease Variance)**.** *Let $\mathcal{D}$ be a distribution on $\mathcal{R}^d$, Q be a $d \times d$ matrix-valued indeterminate, and C be a scalar-valued indeterminate. Then, we have that*

$$\left\vdash^{Q,C}_{2} \left\{\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)^2\right] \leqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Q(x) - C)^2\right]\right\}.$$

Lemma 3.2.32 (Shifts of Certifiably Hypercontractive Distributions)**.** *Let x be a mean-0 random variable with distribution $\mathcal{D}$ on $\mathcal{R}^d$ with t-certifiably C-hypercontractive degree-2 polynomials. Then, for any fixed constant vector $c \in \mathcal{R}^d$, the random variable $x + c$ also has t-certifiable*

$4C$*-hypercontractive degree-*2 *polynomials.*

Lemma 3.2.33 (Mixtures of Certifiably Hypercontractive Distributions)**.** *Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ have t-certifiable C-hypercontractive degree-2 polynomials on $\mathcal{R}^d$, for some fixed constant C. Then, any mixture $\mathcal{D} = \sum_i w_i \mathcal{D}_i$ also has t-certifiably (C/α)-hypercontractive degree-2 polynomials for $\alpha = \min_{i \leqslant k, w_i > 0} w_i$.*

Corollary 3.2.34 (Certifiable Hypercontractivity of Mixtures of k Gaussians)**.** *Let $\mathcal{M}$ be a k-mixture of Gaussians $\sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ with weights $w_i \geqslant \alpha$ for every $i \in [k]$. Then, for all $t \in \mathbb{N}$, $\mathcal{D}$ has t-certifiably $4/\alpha$-hypercontractive degree-2 polynomials.*

We will use the following robust mean estimation algorithm for bounded covariance distributions [DKK+19]:

Fact 3.2.35 (Robust Mean Estimation for Bounded Covariance Distributions)**.** *There is a poly(n) time algorithm that takes input an ϵ-corruption Y of a collection of n points $X \subseteq \mathcal{R}^d$, and outputs an estimate $\hat{\mu}$ satisfying $\|\mathbf{E}_{x \sim_u X} x - \hat{\mu}\|_2 \leqslant O(\sqrt{\epsilon}) \|\mathbb{E}_{x \sim_u X} \left[(x - \mathbf{E}_{x \sim_u X} x)(x - \mathbf{E}_{x \sim_u X} x)^\top\right]\|_2$.*

We will use the following robust covariance estimation algorithm from [KS17]:

Fact 3.2.36 (Robust Covariance Estimation, [KS17])**.** *For every $C > 0, \epsilon > 0$ and even $k \in \mathbb{N}$ such that $Ck\epsilon^{1-2/k} \leqslant c$ for some small enough absolute constant c, there exists a polynomial-time algorithm that given an (corrupted) sample S outputs an estimate of the covariance $\hat{\Sigma} \in \mathcal{R}^{d \times d}$ with the following guarantee: there exists $n_0 \geqslant (C + d)^{O(k)}/\epsilon$ such that if S is an ϵ-corrupted sample with size $|S| \geqslant n_0$ of a k-certifiably C-subgaussian distribution D over $\mathcal{R}^d$ with mean $\mu \in \mathcal{R}^d$ and covariance $\Sigma \in \mathcal{R}^{d \times d}$, then with high probability:*

$$(1 - \delta)\Sigma \succeq \hat{\Sigma} \succeq (1 + \delta)\Sigma$$

for $\delta \leqslant O(Ck)\epsilon^{1-2/k}$.

We will also require the following robust estimation algorithm with Frobenius distance guarantees proven for certifiably hypercontractive distributions in [BK20b]. Since we obtain estimates to the true covariance in Lowner ordering, we can obtain the subspace spanned by the inliers exactly, project on to this subspace and apply Theorem 7.1 in [BK20b].

Fact 3.2.37 (Robust Mean and Covariance Estimation for Certifiably Hypercontractive Distribu-

tions, Theorem 7.1 in [BK20b]). *Given $t \in \mathbb{N}$, and $\epsilon > 0$ sufficiently small so that $Ct\epsilon^{1-4/t} \ll 1$*[4]*, for some absolute constant $C > 0$. Then, there is an algorithm that takes input Y, an ϵ-corruption of a sample X of size n with mean μ_*, covariance Σ_*, and $2t$-certifiably C-hypercontractive degree-2 polynomials, runs in time $n^{O(t)}$, and outputs an estimate $\hat{\mu}$ and $\hat{\Sigma}$ satisfying:*

1. $\|\Sigma_*^{\dagger/2}(\mu_* - \hat{\mu})\|_2 \leqslant O(Ct)^{1/2}\epsilon^{1-1/t}$,
2. $(1-\eta)\Sigma_* \preceq \hat{\Sigma} \preceq (1+\eta)\Sigma_*$ *for* $\eta \leqslant O(Ck)\epsilon^{1-2/t}$, *and,*
3. $\|\Sigma_*^{\dagger/2}\left(\hat{\Sigma} - \Sigma_*\right)\Sigma_*^{\dagger/2}\|_F \leqslant (C't)O(\epsilon^{1-1/t})$,

where $C' = \max\{C, B\}$ for the smallest possible $B > 0$ such that for $d \times d$-matrix-valued indeterminate Q, $\vdash_2^Q \left\{\mathbb{E}_{\mathcal{D}}\left[\left(x^\top Qx - \mathbf{E}_{\mathcal{D}} x^\top Qx\right)^2\right] \leqslant B\|\Sigma_*^{1/2} Q \Sigma_*^{1/2}\|_F^2\right\}$.[5]

The last line in the above fact asserts a bound (along with a degree 2 SoS proof) on the variance of degree 2 polynomials in terms of the Frobenius norm of its coefficient matrix. In the next few claims, we verify this property via elementary arguments for the two classes of distributions relevant to this paper. We note that whenever a distribution satisfies the bounded variance property (without an SoS proof), it also satisfies the property via a degree 2 SoS proof using Lemma 3.2.23. Thus, asking for an SoS proof of degree 2 in this context poses no additional restrictions on the distribution. Nevertheless, we provide explicit and direct SoS proofs in the following.

We first note that this property of having *certifiable bounded variance* is closed under linear transformations.

Lemma 3.2.38 (Linear Transformations of Certifiably Bounded-Variance Distributions)**.** *For $d \in \mathbb{N}$, let x be a random variable with distribution $\mathcal{D}$ on $\mathcal{R}^d$ such that for $d \times d$ matrix-valued indeterminate Q, $\vdash_2^Q \left\{\mathbf{E}_{x\sim\mathcal{D}}(x^\top Qx - \mathbf{E}_{\mathcal{D}} x^\top Qx)^2 \leqslant \|\Sigma^{1/2} Q \Sigma^{1/2}\|_F^2\right\}$. Let A be an arbitrary $d \times d$ matrix and let $x' = Ax$ be the random variable with covariance $\Sigma' = AA^\top$. Then, we have that*

$$\vdash_2^Q \left\{\mathbf{E}_{x'\sim\mathcal{D}'}(x'^\top Qx' - \mathbf{E}_{\mathcal{D}'} x'^\top Qx')^2 \leqslant \|\Sigma'^{1/2} Q \Sigma'^{1/2}\|_F^2\right\}.$$

[4]This notation means that we needed $Ct\epsilon^{1-2/t}$ to be at most c_0 for some absolute constant $c_0 > 0$.

[5]The first two guarantees here hold for the larger class of certifiably subgaussian distributions and were proven in [KS17] (see Theorem 1.2). Gaussian distribution (with arbitrary mean and covariance) are t-certifiably 1-subgaussian for all t and their mixtures (similar to Lemma 3.2.33 and explicitly proven in Lemma 5.4 of [KS17]) are t-certifiably $O(1/\alpha)$-subgaussian where α is the minimum mixing weight.

Lemma 3.2.39 (Variance of Degree-2 Polynomials of Standard Gaussians)**.** *We have that*

$$\vdash_2^Q \left\{ \mathbf{E}_{\mathcal{N}(0,I)} \left(x^\top Q x - \mathbf{E}_{\mathcal{N}(0,I)} x^\top Q x \right)^2 \leqslant 3\|Q\|_F^2 \right\} .$$

Remark 71. As is easy to verify, the same proof more generally holds for any distribution that has the same first four moments as the zero-mean Gaussian distribution.

As an immediate corollary of the previous two lemmas, we have:

Corollary 3.2.40 (Variance of Degree-2 Polynomials of Zero-Mean, Arbitrary Covariance Gaussians)**.** *For any $0 \preceq \Sigma$, we have that*

$$\vdash_2^Q \left\{ \mathbf{E}_{\mathcal{N}(0,\Sigma)} \left(x^\top Q x - \mathbf{E}_{\mathcal{N}(0,\Sigma)} x^\top Q x \right)^2 \leqslant 3\|\Sigma^{1/2} Q \Sigma^{1/2}\|_F^2 \right\} .$$

We next prove that the same property holds for mixtures of Gaussians satisfying certain conditions.

Lemma 3.2.41 (Variance of Degree-2 Polynomials of Mixtures)**.** *Let $\mathcal{M} = \sum_i w_i \mathcal{D}_i$ be a k-mixture of distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ with means μ_i and covariances Σ_i. Let $\mu = \sum_i w_i \mu_i$ be the mean of $\mathcal{M}$. Suppose that each of $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ have certifiably C-bounded-variance i.e. for Q: a symmetric $d \times d$ matrix-valued indeterminate.*

$$\vdash_2^Q \left\{ \mathbf{E}_{x' \sim \mathcal{D}_i} (x'^\top Q x' - \mathbf{E}_{\mathcal{D}_i} x'^\top Q x')^2 \leqslant C\|\Sigma'^{1/2} Q \Sigma'^{1/2}\|_F^2 \right\} .$$

Further, suppose that for some $H > 1$, $\|\mu_i - \mu\|_2^2, \|\Sigma_i - I\|_F \leqslant H$ for every $1 \leqslant i \leqslant k$. Then, we have that

$$\vdash_2^Q \left\{ \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left[x^\top Q x \right] \right)^2 \right] \leqslant 100 C H^2 \|Q\|_F^2 \right\} .$$

As an immediate corollary of Lemma 3.2.38 and Lemma 3.2.41, we obtain:

Lemma 3.2.42 (Variance of Degree-2 Polynomials of Mixtures of Gaussians)**.** *Let $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with $w_i \geqslant \alpha$, mean $\mu = \sum_i w_i \mu_i$ and covariance $\Sigma = \sum_i w_i ((\mu_i - \mu)(\mu_i - \mu)^\top + \Sigma_i)$. Suppose that for some $H > 1$, $\|\Sigma^{\dagger/2}(\Sigma_i - \Sigma)\Sigma^{\dagger/2}\|_F \leqslant H$ for every $1 \leqslant i \leqslant k$. Let Q be a symmetric $d \times d$ matrix-valued indeterminate. Then for $H' = \max\{H, 1/\alpha\}$,*

$$\vdash_2^Q \left\{ \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x \sim \mathcal{M}} \left[x^\top Q x \right] \right)^2 \right] \leqslant 100 H'^2 \|\Sigma^{1/2} Q \Sigma^{1/2}\|_F^2 \right\} .$$

Analytic Properties are Inherited by Samples. The following lemma can be proven via similar, standard techniques as in several prior works [KS17, KKK19, BK20a, BK20b].

Fact 3.2.43. *Let D be a distribution on $\mathcal{R}^d$ with mean μ and covariance Σ. Let $t \in \mathbb{N}$. Let X be a sample from D such that, $\|\frac{1}{|X|}\sum_{x\in X}(1,\bar{x})^{\otimes t} - \mathbf{E}_{x\sim D}(1,\bar{x})^{\otimes t}\|_F \leqslant d^{-O(t)}$. Here, $\bar{x} = \Sigma^{\dagger/2}(x-\mu_i)$. Then,*

1. *If D is $2t$-certifiably C-subgaussian, then the uniform distribution on X is t-certifiably $2C$-subgaussian.*
2. *If D has $2t$-certifiably C-hypercontractive degree 2 polynomials, then the uniform distribution on X has t-certifiably $2C$-hypercontractive degree 2 polynomials.*
3. *If D is $2t$-certifiably $C\delta$-anti-concentrated, then the uniform distribution on X is t-certifiably $2C\delta$-anti-concentrated.*
4. *If $\left|\frac{Q}{2}\right. \left\{\mathbf{E}_{x\sim\mathcal{D}}(x^\top Qx - \mathbf{E}_{x\sim\mathcal{D}}x^\top Qx)^2 \leqslant C\|Q\|_F^2\right\}$, then, for the uniform distribution $\mathcal{D}_X$ on X, $\left|\frac{Q}{2}\right. \left\{\mathbf{E}_{x\sim\mathcal{D}_X}(x^\top Qx - \mathbf{E}_{x\sim\mathcal{D}_X}x^\top Qx)^2 \leqslant 2C\|Q\|_F^2\right\}$.*

3.2.4 Deterministic Conditions on the Uncorrupted Samples

In this section, we describe the set of deterministic conditions on the set of uncorrupted samples, under which our algorithms succeed. We will require the following definition.

Definition 3.2.44. *Fix $0 < \varepsilon < 1/2$. We say that a multiset Y of points in $\mathcal{R}^d$ is an ε-corrupted version (or an ε-corruption) of a multiset X of points in $\mathcal{R}^d$ if $|X \cap Y| \geqslant \max\{(1-\varepsilon)|X|, (1-\varepsilon)|Y|\}$.*

Throughout this paper and unless otherwise specified, we will use X to denote a multiset of i.i.d. samples from the target k-mixture $\mathcal{M} = \sum_{i=1}^k w_i G_i$, where $G_i = \mathcal{N}(\mu_i, \Sigma_i)$. We will use X_i for the subset of points in X drawn from G_i, i.e., $X = \cup_{i=1}^k X_i$.

We will use Y to denote an ε-corrupted version of X, as per Definition 3.2.44. In this *strong contamination model*, the adversary can see the clean samples from X before they decide on the ε-corruption Y. The strong contamination model is known to subsume the total variation contamination of Definition 3.1.1 (see, e.g., Section 2 of [DKK+19]). We note that our robust learning algorithm succeeds in this stronger contamination model, with the additional requirement that we can obtain two sets of independent ε-corrupted samples from $\mathcal{M}$. (The second set is

needed to run a hypothesis testing routine after we obtain a small list of candidate hypotheses.)

Our algorithm works for any finite set of points in $\mathcal{R}^d$ that satisfies a natural set of deterministic conditions. As we will show later in this section, these deterministic conditions are satisfied with high probability by a sufficiently large set of i.i.d. samples from any k-mixture of Gaussians.

Condition 3.2.45 (Good Samples)**.** *Let $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians in $\mathcal{R}^d$. Let X be a set of n points in $\mathcal{R}^d$. We say that X satisfies Condition 3.2.45 with respect to $\mathcal{M}$ with parameters (γ, t) if there is a partition of X as $X_1 \cup X_2 \cup \ldots \cup X_k$ such that:*

1. *For all $i \in [k]$ with $w_i \geqslant \gamma$, any positive integer $m \leqslant t$, and any $v \in \mathcal{R}^d$,*

$$\left| \frac{1}{n} \sum_{x \in X_i} \langle v, x - \mu_i \rangle^m - w_i \mathbf{E}_{x \sim \mathcal{N}(\mu_i, \Sigma_i)} [\langle v, x - \mu_i \rangle^m] \right| \leqslant w_i \, \gamma \, m! \, (v^T \Sigma_i v)^{m/2} \, .$$

2. *For all $i \in [k]$ and any halfspace $H \subset \mathcal{R}^d$, we have that $\left| |X_i \cap H|/n - w_i \mathbf{Pr}_{x \sim \mathcal{N}(\mu_i, \Sigma_i)} [x \in H] \right| \leqslant \gamma$.*

We will also need the following consequences of Condition 3.2.45. The first one is immediate.

Lemma 3.2.46. *Condition 3.2.45 is invariant under affine transformations. In particular, if $A(x) : \mathcal{R}^d \to \mathcal{R}^{d'}$ is an affine transformation, and if X satisfies Condition 3.2.45 with respect to $\mathcal{M}$ with parameters (γ, t), then $A(X)$ satisfies Condition 3.2.45 with respect to $A(\mathcal{M})$ with parameters (γ, t).*

We note that the first part of Condition 3.2.45 implies that higher moment tensors are close in Frobenius distance.

Lemma 3.2.47. *If X satisfies Condition 3.2.45 with respect to $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ with parameters (γ, t), then if $w_i \geqslant \gamma$ for all $i \in [k]$, and if for some $B \geqslant 0$ we have that $\|\mu_i\|_2^2, \|\Sigma_i\|_{\mathrm{op}} \leqslant B$ for all $i \in [k]$, then for all $m \leqslant t$, we have that:*

$$\|\mathbf{E}_{x \in_u X} [x^{\otimes m}] - \mathbf{E}_{x \sim M} [x^{\otimes m}]\|_F^2 \leqslant \gamma^2 m^{O(m)} B^m d^m \, .$$

We note that Condition 3.2.45 also behaves well with respect to taking submixtures.

Lemma 3.2.48. *Let $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$. Let $S \subset [k]$ with $\sum_{i \in S} w_i = w$, and let $\mathcal{M}' = \sum_{i \in S} (w_i/w) \mathcal{N}(\mu_i, \Sigma_i)$. Then if X satisfies Condition 3.2.45 with respect to $\mathcal{M}$ with parameters (γ, t) for some $\gamma < 1/(2k)$ with the corresponding partition being $X = X_1 \cup X_2 \cup \ldots \cup X_k$, then $X' = \bigcup_{i \in S} X_i$ satisfies Condition 3.2.45 with respect to $\mathcal{M}'$ with parameters $(O(k\gamma/w), t)$.*

Finally, we show that given sufficiently many i.i.d. samples from a k-mixture of Gaussians, Condition 3.2.45 holds with high probability.

Lemma 3.2.49. *Let $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$ and let n be an integer at least $kt^{Ct} d^t / \gamma^3$, for a sufficiently large universal constant $C > 0$, some $\gamma > 0$, and some $t \in \mathbb{N}$. If X consists of n i.i.d. samples from $\mathcal{M}$, then X satisfies Condition 3.2.45 with respect to $\mathcal{M}$ with parameters (γ, t) with high probability.*

The proofs of the preceding lemmas can be found in Appendix 3.10.

3.2.5 Hypothesis Selection

Our algorithm will require a procedure to select a hypothesis from a list of candidates that contains an accurate hypothesis. A number of such procedures are known in the literature. Here we will use the following variant from [Kan20], showing that we can efficiently perform a hypothesis selection (tournament) step with access to ϵ-corrupted samples.

Fact 3.2.50 (Robust Tournament, [Kan20])**.** *Let X be an unknown distribution, $\eta \in (0, 1)$, and let $H_1, \ldots, H_n$ be distributions with explicitly computable probability density functions that can be efficiently sampled from. Assume furthermore than $\min_{1 \leqslant i \leqslant n} (d_{\mathsf{TV}}(X, H_i)) \leqslant \eta$. Then there exists an efficient algorithm that given access to $O(\log(n)/\eta^2)$ ϵ-corrupted samples from X, where $\varepsilon \leqslant \eta$, along with $H_1, \ldots, H_n$, computes an $m \in [n]$ such that with high probability we have that $d_{\mathsf{TV}}(X, H_m) = O(\eta)$.*

3.3 List-Recovery of Parameters via Tensor Decomposition

In this section, we give an algorithm that takes samples from a k-mixture of Gaussians, whose component means and covariances are not too far from each other in natural norms, and outputs a dimension-independent size list of candidate k-tuples of parameters (i.e., means and covariances) one of which is guaranteed to be close to the true target k-tuple of parameters. Our approach

involves a new tensor decomposition procedure that works in the absence of any non-degeneracy conditions on the components.

The goal of this section is to prove the following theorem:

Theorem 72 (Recovering Candidate Parameters when Component Covariances are close in Frobenius Distance)**.** *Fix any $\alpha > \epsilon > 0, \Delta > 0$. There is an algorithm that takes input X, a sample from a k-mixture of Gaussians $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ satisfying Condition 3.2.45 with parameters $\gamma = \varepsilon d^{-8k} k^{-Ck}$, for C a sufficiently large universal constant, and $t = 8k$, and let Y be an ε-corruption of X. If $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant \frac{2}{\sqrt{\alpha}}$ and $\|\Sigma_i - I\|_F \leqslant \Delta$ for every $i \in [k]$, then, given k, Y and ε, the algorithm outputs a list L of at most $\exp\left(\log(1/\epsilon)\,(k + 1/\alpha + \Delta)^{O(k)}/\eta^2\right)$ candidate hypotheses (component means and covariances), such that with probability at least 0.99 there exist $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]} \subseteq L$ satisfying $\|\mu_i - \hat{\mu}_i\|_2 \leqslant \mathcal{O}\left(\frac{\Delta^{1/2}}{\alpha}\right)\eta^{G(k)}$ and $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant \mathcal{O}(k^4)\,\frac{\Delta^{1/2}}{\alpha}\eta^{G(k)}$ for all $i \in [k]$. Here, $\eta = (2k)^{4k}\mathcal{O}(1/\alpha + \Delta)^{4k}\sqrt{\epsilon}$, $G(k) = \frac{1}{C^{k+1}(k+1)!}$. The running time of the algorithm is poly$(|L|, |Y|, d^k)$.*

In the body of this section, we establish Theorem 72. The structure of this section is as follows: In Section 3.3.1, we describe our algorithm, which is then analyzed in Sections 3.3.2-3.3.6.

3.3.1 List-Decodable Tensor Decomposition Algorithm

In this section, we describe our tensor decomposition algorithm, which is given in pseudocode below (Algorithm 73).

Algorithm 73 (List-Recovery of Candidate Parameters via Tensor Decomposition)**.**

Input: *An ϵ-corruption Y of a sample X from a k-mixture of Gaussians $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$.*

Requirements: *The guarantees of the algorithm hold if the mixture parameters and the sample X satisfy:*

1. *$w_i \geqslant \alpha$ for all $i \in [k]$,*
2. *$\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$ for all $i \in [k]$,*
3. *$\|\Sigma_i - I\|_F \leqslant \Delta$ for all $i \in [k]$.*
4. *X satisfies Condition 3.2.45 with parameters (γ, t), where $\gamma = \varepsilon d^{-8k} k^{-Ck}$, for*

C a sufficiently large universal constant, and $t = 8k$.

Parameters: $\eta = (2k)^{4k}(Ck(1/\alpha + \Delta))^{4k}\sqrt{\epsilon}$, $D = C(k^4/(\alpha\sqrt{\eta}))$, $\delta = 2\eta^{1/(C^{k+1}(k+1)!)}$, $\ell' = 100\log k\,(\eta/\,(k^5\,(\Delta^4 + 1/\alpha^4)))^{-4k}$, *for some sufficiently large absolute constant* $C > 0$, $\lambda = 4\eta$, $\phi = 10(1+\Delta^2)/(\sqrt{\eta}\alpha^5)$.

Output: *A list L of hypotheses such that there exists at least one, $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\leqslant k} \in L$, satisfying: $\|\mu_i - \hat{\mu}_i\|_2 \leqslant \mathcal{O}\big(\frac{\Delta^{1/2}}{\alpha}\big)\,\eta^{G(k)}$ and $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant \mathcal{O}(k^4)\,\frac{\Delta^{1/2}}{\alpha}\eta^{G(k)}$, where $G(k) = \frac{1}{C^{k+1}(k+1)!}$.*

Operation:

1. ***Robust Estimation of Hermite Tensors**: For $m \in [4k]$, compute $\hat{T}_m$ such that $\max_{m\in[4k]} \left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F \leqslant \eta$ using the robust mean estimation algorithm in Fact 3.2.35.*

2. ***Random Collapsing of Two Modes of $\hat{T}_4$**: Let L' be an empty list. Repeat ℓ' times: For $j \in [4k]$, choose independent standard Gaussians in $\mathbb{R}^d$, denoted by $x^{(j)}, y^{(j)} \sim \mathcal{N}(0, I)$, and uniform draws $a_1, a_2, \ldots, a_t$ from $[-D, D]$. Let $\hat{S}$ be a $d \times d$ matrix such that for all $r, s \in [d]$, $\hat{S}(r,s) = \sum_{j\in[4k]} a_j \hat{T}_4(r,s,x^{(j)},y^{(j)}) = \sum_{j\in[4k]} a_j \sum_{g,h\in[d]} \hat{T}_4(r,s,g,h)x^{(j)}(g)y^{(j)}(h)$. Add $\hat{S}$ to the list L'.*

3. ***Construct Low-Dimensional Subspace for Exhaustive Search**: Let V be the span of all singular vectors of the natural $d \times d^{m-1}$ flattening of $\hat{T}_m$ with singular values $\geqslant \lambda$ for $m \leqslant 4k$. For each $\hat{S} \in L'$, let $V'_{\hat{S}}$ be the span of V plus all the singular vectors of $\hat{S}$ with singular value larger than $\delta^{1/4}$.*

4. ***Enumerating Candidates in $V'_{\hat{S}}$**: Initialize L to be the empty list. For each $\hat{S} \in L'$, let $V_{\delta^{1/4}}$ be a $\delta^{1/4}$-cover of vectors in $V'_{\hat{S}}$ with ℓ_2-norm at most $2/\sqrt{\alpha}$. Enumerate over vectors $\hat{\mu}$ in $V_{\delta^{1/4}}$. Let $k' = Ck^2$ and let $\mathcal{C}_{\delta^{1/4}}$ be a $\delta^{1/4}$-cover of the interval $[-\phi, \phi]^{k'}$. For $\{\tau_j\}_{j\in[k']} \in \mathcal{C}_{\delta^{1/4}}$ and for all $\{v_j\}_{j\in[k']} \in V_{\delta^{1/4}}$, let $\hat{Q} = \sum_{j\in[k']} \tau_j v_j v_j^\top$. Add $\{\hat{\mu}, I + \hat{S} + \hat{Q}\}$ to L.*

3.3.2 Analysis of Algorithm

We analyze the three main steps of Algorithm 73 in the following lemmas. We will prove the following three propositions in the subsequent subsections that analyze Steps 1, 2 and 3 of Algorithm 73. For Step 1, we show that when X satisfies Condition 3.2.45, the empirical estimates of the moment tensors obtained by applying the robust mean estimation algorithm to X are suf-

ficiently close to the moment tensors of the input mixture $\mathcal{M}$.

Proposition 3.3.1 (Robustly Estimating Hermite Polynomial Tensors)**.** *For any integer $m \leqslant 4k$, and $\Delta \in \mathcal{R}_+$, there exists an algorithm with running time $\text{poly}_m(d/\varepsilon)$ that takes an ϵ-corruption Y of X, a set satisfying Condition 3.2.45 with respect to $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$ with parameters $\gamma = \varepsilon d^{-m} m^{-Cm}$, for C a sufficiently large constant, and $t = 2m$. If $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and $\|\Sigma_i - I\|_F \leqslant \Delta$ for each $i \in [k]$, then the algorithm outputs a tensor $\hat{T}_m$ such that $\left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F \leqslant \eta$, for $\eta = \mathcal{O}\left(m(1 + 1/\alpha + \Delta)\right)^m \sqrt{\varepsilon}$.*

The proof of Proposition 3.3.1 is deferred to Section 3.3.3.

Next, we analyze Step 2 of the algorithm and prove that, with non-negligible probability, randomly collapsing two modes of $\hat{T}_4$ yields a matrix $\hat{S}$ such that $\hat{S} - (\Sigma_i - I) = P_i + Q_i$, where P_i has small Frobenius norm and Q_i is a rank-$\mathcal{O}(k^2)$ matrix.

Proposition 3.3.2 (Tensor Decomposition up to Low-Rank Error)**.** *Let $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians satisfying $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and $\|\Sigma_i - I\|_F \leqslant \Delta$ for each $i \in [k]$. For $0 < \eta < 1$, let $\hat{T}_4$ be a tensor such that $\left\|\mathbf{E}[h_4(\mathcal{M})] - \hat{T}_4\right\|_F \leqslant \eta$, and let D be a sufficiently large constant multiple of $k^4/(\alpha\sqrt{\eta})$. For all $j \in [4k]$, let $x^{(j)}, y^{(j)} \sim \mathcal{N}(0, I)$ be independent and $a_j \sim \mathcal{U}[-D, D]$, where $\mathcal{U}[-D, D]$ is the uniform distribution over the interval $[-D, D]$, and let $\hat{S} = \sum_{j \in [4k]} a_j \hat{T}_4\left(\cdot, \cdot, x^{(j)}, y^{(j)}\right)$. Then, for each $i \in [k]$, with probability at least $\left(\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{4k}$, over the choice of $x^{(j)}, y^{(j)}$ and a_j, we have that $\hat{S} - (\Sigma_i - I) = P_i + Q_i$, where $\|P_i\|_F = \mathcal{O}\left(\sqrt{\eta/\alpha}\right)$, $\|Q_i\|_F = \mathcal{O}\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3}\right)$ and $\operatorname{rank}(Q_i) = \mathcal{O}(k^2)$.*

The proof of Proposition 3.3.2 is given in Section 3.3.4.

Finally, in Step 3, for any $\hat{S}$ such that $\hat{S} - (\Sigma_i - I) = P_i + Q_i$, where P_i has small Frobenius norm and Q_i is a rank $\mathcal{O}(k^2)$ matrix, we find a low-dimensional subspace V' such that the range space of Q_i is approximately contained in V'. We will use V' to exhaustively search for $\mathcal{O}(k^2)$ rank matrices to find candidates for Q_i.

Proposition 3.3.3 (Low-Dimensional Subspace V' for Exhaustive Search)**.** *Let $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians satisfying $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and $\|\Sigma_i - I\|_F \leqslant \Delta$ for each $i \in [k]$. Let $\left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F \leqslant \eta$, for each $1 \leqslant m \leqslant 4k$, and some $\eta > 0$. Let V be the span of all the left singular vectors of the $d \times d^{m-1}$ matrix obtained by the natural flattening of $\hat{T}_m$ with singular values at least 2η. For each $1 \leqslant i \leqslant k$, let $S_i = \Sigma_i - I$ and $\hat{S}_i$ be a $d \times d$ matrix such that $\hat{S}_i - S_i = P_i + Q_i$, where $\|P_i\|_F \leqslant \mathcal{O}\left(\sqrt{\eta/\alpha}\right)$, Q_i has rank $\mathcal{O}(k^2)$, and $\|Q_i\|_F \leqslant \mathcal{O}\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3}\right)$. Let V' be the span of V plus all singular vectors of $\hat{S}_i$ of singular values at least δ for all i. Then,*

for $\delta = 2\eta^{1/(C^{k+1}(k+1)!)}$ with a sufficiently large constant $C > 0$, we have that:

1. $\dim V' \leqslant \left(\mathcal{O}\left(k(1+1/\alpha+\Delta)\right)^{4k+5}\right)/\eta^2$.
2. *There is a vector $\mu_i' \in V'$ such that $\|\mu_i - \mu_i'\|_2^2 \leqslant \frac{20}{\alpha^2}\sqrt{\delta}\Delta$.*
3. *There are $q = \mathcal{O}(k^2)$ unit vectors $v_1, v_2, \ldots, v_q \in V'$ and scalars $\tau_1, \tau_2, \ldots, \tau_q \in \left[-10(1+\Delta^2)/(\sqrt{\eta}\alpha^5), 1\right.$ such that $\left\|Q_i - \sum_{i=1}^q \tau_i v_i v_i^\top\right\|_F \leqslant \mathcal{O}\left(\frac{k^2}{\alpha}\delta^{1/4}\Delta^{1/2}\right)$.*

The proof of Proposition 3.3.3 is given in Section 3.3.5.

We can now use these propositions to complete the proof of Theorem 72.

Proof of Theorem 72. Using Proposition 3.3.1, Step 1 of the algorithm outputs estimates $\hat{T}_i$ for $i \in [4k]$ such that $\max_{m\in[4k]} \left\|\hat{T}_m - \mathbf{E}h_m(\mathcal{M})\right\|_F \leqslant \eta$. Next, by the standard coupon collector analysis, using Proposition 3.3.2 and repeating Step 2 of the algorithm $\ell' = 100\log k\left(\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4k}$ times, guarantees that with probability at least $1 - 1/(100k)^{100}$, for every $1 \leqslant i \leqslant k$, there are $\hat{S}_i \in L$ such that $\hat{S}_i - (\Sigma_i - I) = P_i + Q_i$ for P_i, Q_i satisfying $\|P_i\|_F \leqslant \sqrt{\eta/\alpha}$, $\|Q_i\|_F \leqslant \frac{1+\Delta^2}{\sqrt{\eta}\alpha^5}$ and Q_i has rank $\mathcal{O}(k^2)$.

Next, Proposition 3.3.3 implies that for every such $\hat{S}_i \in L'$, we can construct a subspace $V' = V'_{\hat{S}_i}$ of dimension $\mathcal{O}\left((k(1+1/\alpha+\Delta))^{4k+5}/\eta^2\right)$ such that V' contains μ_i' that satisfies $\|\mu_i - \mu_i'\|_2^2 \leqslant \frac{\Delta}{\alpha^2}\cdot\sqrt{\delta}$, and there is a rank $O(k^2)$ matrix $\hat{Q}_i$ with range space contained in V' such that $\left\|Q_i - \hat{Q}_i\right\|_F \leqslant \mathcal{O}(\frac{k^2}{\alpha}\delta^{1/4}\Delta^{1/2})$.

Now, let $V_\tau \subseteq V'$ be a $\tau = \delta^{1/4}$-cover, in ℓ_2-norm, of vectors with ℓ_2 norm at most $2/\sqrt{\alpha}$ in V'. Then, since $\|\mu_i\|_2 \leqslant \frac{2}{\sqrt{\alpha}}$, there is a vector $\hat{\mu}_i \in V_\tau$ such that $\|\mu_i - \hat{\mu}_i\|_2^2 \leqslant \tau + \frac{20}{\alpha^2}\sqrt{\delta}\Delta \leqslant \frac{40}{\alpha^2}\sqrt{\delta}\Delta$.

Further, there exist $\tau_1, \tau_2, \ldots \tau_{\mathcal{O}(k^2)}$ in a τ-cover of $[-10(1+\Delta^2)/(\sqrt{\eta}\alpha^5), 10(1+\Delta^2)/(\sqrt{\eta}\alpha^5)]$ and vectors $v_1, v_2, \ldots, v_{\mathcal{O}(k^2)} \in V_\tau$ such that $\left\|\sum_{i=1}^{\mathcal{O}(k^2)} \tau_i v_i v_i^\top - Q_i\right\|_F \leqslant \mathcal{O}(k^4\delta^{1/4}\Delta^{1/2}/\alpha)$. In particular, $\hat{\Sigma}_i = I + \hat{S} - \sum_{i=1}^{\mathcal{O}(k^2)} \tau_i v_i v_i^\top$ satisfies

$$\|\hat{\Sigma}_i - \Sigma_i\|_F = \mathcal{O}(\sqrt{\eta}) + \mathcal{O}\left(\frac{k^4\delta^{1/4}\Delta^{1/2}}{\alpha}\right) = \mathcal{O}\left(\frac{k^4\delta^{1/4}\Delta^{1/2}}{\alpha}\right) . \tag{3.7}$$

The size of this search space for every fixed $\hat{S} \in L'$ can be bounded above by $\left(\frac{1+\Delta^2}{\delta\alpha^5}\right)^{\mathcal{O}(k^5 \dim V')}$.

Thus, the size of L can be bounded from above by

$$k^5\left(\frac{\Delta^4}{\eta}+\frac{1}{\alpha^4\eta}\right)^{4k}\cdot\left(\frac{1+\Delta^2}{\delta\sqrt{\eta}\alpha^5}\right)^{\mathcal{O}(k^5\dim V')}\leqslant\exp\left(\log(1/\epsilon)\left(k+1/\alpha+\Delta\right)^{O(k)}/\eta^2\right)\ .$$

This completes the proof. □

3.3.3 Robust Estimation of Hermite Tensors

In this section, we will prove Proposition 3.3.1.

Proof of Proposition 3.3.1. Consider the uniform distribution on the uncorrupted sample X. We want to analyze the effect of applying the robust mean estimation algorithm (Fact 3.2.35) to the points $h_m(x)$, for $x\in X$. In order for us to apply Fact 3.2.35, we need to ensure that the uniform distribution on $\{h_m(x)\}_{x\in X}$ has bounded covariance. This step gives us a good approximation to $\mathbf{E}_{x\sim_u X}h_m(x)$. In order for us to obtain an approximation to $\mathbf{E}h_m(\mathcal{M})$, we need to bound the difference between $\mathbf{E}h_m(\mathcal{M})$ and $\mathbf{E}_{x\sim_u X}h_m(x)$. We will do both these steps below.

The second part is immediate. By the definition of $h_m(X)$, we have that

$$\left\|\frac{1}{|X|}\sum_{x\in X}h_m(x)-\mathbf{E}h_m(\mathcal{M})\right\|_F\leqslant\sum_{j\leqslant m/2}m^{2j}d^j\left\|\frac{1}{|X|}\sum_{x\in X}x^{\otimes(m-2j)}-\mathbf{E}\mathcal{M}^{\otimes(m-2j)}\right\|_F.$$

By Lemma 3.2.47, this is at most $O(1+\Delta+1/\alpha)^m m^{O(m)}d^{m/2}\gamma\leqslant\eta/2$. We note that a similar argument bounds

$$\left\|\frac{1}{|X|}\sum_{x\in X}h_m(x)\otimes h_m(x)-\mathbf{E}h_m(\mathcal{M})\otimes h_m(\mathcal{M})\right\|_F\leqslant\eta^2.$$

Let us now verify the first part. We proceed via bounding the operator norm of the covariance of $h_m(\mathcal{M})$. We can then use the bound on the Frobenius norm

$$\left\|\frac{1}{|X|}\sum_{x\in X}h_m(x)\otimes h_m(X)-\mathbf{E}h_m(\mathcal{M})\otimes h_m(\mathcal{M})\right\|_F$$

to get a bound on $\left\|\frac{1}{|X|}\sum_{x\in X}h_m(x)h_m(x)^\top\right\|_{\mathrm{op}}$ (the operator norm of the canonical square flattening of the of the $2m$-th empirical Hermite moment tensor of X). This will complete the proof.

Let $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be the components of $\mathcal{M}$. We have that

$$\begin{aligned}\mathbf{Cov}(h_m(\mathcal{M})) = &\sum_{i\in[k]} w_i \mathbf{Cov}(h_m(G_i)) \\ &+ \frac{1}{2}\sum_{i,j\in[k]} w_i w_j \left(\mathbf{E}[h_m(G_i)] - \mathbf{E}[h_m(G_j)]\right)\left(\mathbf{E}[h_m(G_i)] - \mathbf{E}[h_m(G_j)]\right)^\top .\end{aligned} \tag{3.8}$$

By Lemma 3.2.8, we have that for all $i \in [k]$, it holds

$$\|\mathbf{Cov}(h_m(G_i))\|_{\text{op}} = \mathcal{O}\left(m(1 + \|\mu_i\|_2 + \|\Sigma_i - I\|_F)\right)^{2m} = \mathcal{O}\left(m(1 + 2/\sqrt{\alpha} + \Delta)\right)^{2m} ,$$

where for any matrix M, $\|M\|_{\text{op}} = \max_{\|u\|_2=1} \|Mu\|_2$ is the operator norm of the matrix. Further, for any $i, j \in [k]$,

$$\begin{aligned}\left\|\left(\mathbf{E}[h_m(G_i)] - \mathbf{E}[h_m(G_j)]\right)\left(\mathbf{E}[h_m(G_i)] - \mathbf{E}[h_m(G_j)]\right)^\top\right\|_{\text{op}} &= \|\mathbf{E}[h_m(G_i)] - \mathbf{E}[h_m(G_j)]\|_2 \\ &= \mathcal{O}\left(m(1 + 1/\alpha + \Delta)\right)^{2m} .\end{aligned} \tag{3.9}$$

This claim follows from the triangle inequality of the operator norm. □

3.3.4 List-Recovery of Covariances up to Low-Rank Error

In this section, we prove Proposition 3.3.2. We first set some useful notation. We will write $S_i \stackrel{\text{def}}{=} \Sigma_i - I$ throughout this section. We will also use S_i' to denote $S_i + \mu_i \otimes \mu_i$.

We first show that for every i, there exists a matrix P such that $\left(\sum_{i\in[k]} w_i S_i' \otimes S_i'\right)(\cdot, \cdot, P)$ is close to S_i'.

Lemma 3.3.4 (Existence of a 2-Tensor)**.** *Under the hypothesis of Proposition 3.3.2, for each $i \in [k]$, there exists a matrix P such that $\|P\|_F = \mathcal{O}\left(1/(\sqrt{\eta}\alpha)\right)$ and $\|T_4'(\cdot, \cdot, P) - S_i'\|_F = \mathcal{O}\left(\sqrt{\eta/\alpha}\right)$, where $T_4' = \left(\sum_{i\in[k]} w_i S_i' \otimes S_i'\right)$.*

Note that throughout this section it will be useful to think of T_4' as a $d^2 \times d^2$ matrix rather than as a tensor. In this case, we can think of T_4' as $\sum_{i=1}^k w_i (S_i')(S_i')^T$. From standard facts about positive semidefinite matrices it follows that S_i' is in the image of T_4', and Lemma 3.3.4 is just a slightly robustified version of this (saying that we can find an approximate preimage that it not

itself too large).

The proof of this Lemma 3.3.4 will involve linear programming duality with an infinite system of constraints. As the application of duality with infinitely many constraints has some technical issues, we state below an appropriate version of duality.

Fact 3.3.5 (Linear Programming Duality for Compact, Convex Constraint Sets)**.** *Let $K \subset \mathcal{R}^{n+1}$ be a compact convex set. There exists an $x \in \mathcal{R}^n$ so that $(x, 1) \cdot z > 0$ for all $z \in K$ if and only if there is no element $(0, 0, \ldots, 0, a) \in K$ for any $a \leqslant 0$.*

This fact can be proved by noting that if no such a exists, there must be a hyperplane separating K from the set of such points $(0, a)$. This separating hyperplane will be of the form $(z, y) \in H$ if and only if $y = x \cdot z$ for some x and this x will provide the solution to the linear system.

Proof of Lemma 3.3.4. To show that such a P exists for each i, we apply linear programming duality. In particular, the conditions imposed on P define a linear program, which has a feasible solution unless there is a solution to the dual linear program. For sufficiently large constants c_1 and c_2, consider the following primal in the variable P:

$$\langle v, P \rangle \leqslant \frac{c_1}{\sqrt{\eta}\alpha} \|v\|_F \qquad \forall \ \ v \in \mathbb{R}^{d \times d} \tag{3.10}$$

$$\langle u, T_4'(\cdot, \cdot, P) - S_i' \rangle \leqslant c_2 \sqrt{\eta} \|u\|_F \qquad \forall \ \ u \in \mathbb{R}^{d \times d}. \tag{3.11}$$

It is not hard to see that $\|P\|_F \leqslant \frac{c_1}{\sqrt{\eta}\alpha}$ if and only if (3.10) holds for all v and $\|T_4'(\cdot, \cdot, P) - S_i'\|_F \leqslant c_2\sqrt{\eta/\alpha}$ if and only if (3.11) holds for all u. Throughout the proof, we suggest that the reader think of u and v as vectors in d^2-dimensional vector space.

Our goal is to show that there exists a feasible solution P such that (3.10) and (3.11) hold simultaneously for all $u, v \in \mathcal{R}^{d \times d}$. We first note that this is equivalent to saying that

$$\langle v, P \rangle + \langle u, T_4'(\cdot, \cdot, P) \rangle - \langle u, S_i' \rangle \leqslant \frac{c_1}{\sqrt{\eta}\alpha} \|v\|_F + c_2 \sqrt{\eta} \|u\|_F \ , \tag{3.12}$$

for all $u, v \in \mathcal{R}^{d \times d}$. This is not quite in the form necessary to apply Fact 3.3.5, so we note that this is in turn equivalent to saying that

$$\langle v, P \rangle + \langle u, T_4'(\cdot, \cdot, P) \rangle - \langle u, S_i' \rangle \leqslant 1 \ , \tag{3.13}$$

for all $u, v \in \mathcal{R}^{d \times d}$ so that $\frac{c_1}{\sqrt{\eta}\alpha} \|v\|_F + c_2\sqrt{\eta} \|u\|_F \leqslant 1$, and $u \in \operatorname{span}\{S_i'\}$. As this is a convex set of linear equations, we have by Fact 3.3.5 that there exists such a P unless there exists such a pair of u and v so that the coefficient of P in Equation (3.13) is 0 and so that the resulting inequality of constants is either false or holds with equality. In particular, the coefficient of P vanishes if and only if $v = -T_4'(u, \cdot, \cdot)$. We then get a contradiction only if for some $u \in \operatorname{span}\{S_i'\}$

$$-\langle u, S_i' \rangle \geqslant 1 \geqslant \frac{c_1}{\sqrt{\eta}\alpha} \|T_4'(u, \cdot, \cdot)\|_F + c_2\sqrt{\eta} \|u\|_F . \tag{3.14}$$

We claim that this is impossible.

In particular, squaring Equation (3.14) would give

$$\begin{aligned} \langle u, S_i' \rangle^2 &\geqslant \left(\frac{c_1}{\sqrt{\eta}\alpha} \|T_4'(u, \cdot, \cdot)\|_F + c_2\sqrt{\eta} \|u\|_F \right)^2 \\ &\geqslant \frac{c}{\alpha} \|T_4'(u, \cdot, \cdot)\|_F \cdot \|u\|_F , \end{aligned} \tag{3.15}$$

for some large enough constant $c > 1$, where the last inequality follows from the AM-GM inequality. However, using the dual characterization of the Frobenius norm, we have

$$\|T_4'(u, \cdot, \cdot)\|_F \geqslant \frac{\langle u, T_4'(u, \cdot, \cdot) \rangle}{\|u\|_F} \geqslant \frac{w_i}{\|u\|_F} \langle u, S_i' \rangle^2 , \tag{3.16}$$

where the last inequality follows from T_4' containing a $w_i S_i \otimes S_i$ term, and the other terms contributing non-negatively. Rearranging Equation (3.16), we have

$$\langle u, S_i' \rangle^2 \leqslant \frac{1}{w_i} \|T_4'(u, \cdot, \cdot)\|_F \|u\|_F \leqslant \frac{1}{\alpha} \|T_4'(u, \cdot, \cdot)\|_F \|u\|_F .$$

This contradicts Equation (3.15) unless $T_4'(u, \cdot, \cdot) = 0$. This therefore suffices to prove the feasibility of the primal. □

We have thus shown that there is some matrix P so that $T_4'(P, \cdot, \cdot)$ suffices for our purposes. We need to show that our appropriate random linear combination of $x^{(j)} \otimes y^{(j)}$ suffices. In fact, we will show that with reasonably high probability over our choice of $x^{(j)}, y^{(j)}$ that there is some linear combination of the $x^{(j)} \otimes y^{(j)}$ (with coefficients that are not too large) so that their projection onto the space spanned by the S_i' (which is all that matters when applying T_4') equal to P.

For the sake of intuition, we note that if we removed the bound on the coefficients, we would

need that the projections of the $x^{(j)} \otimes y^{(j)}$ spanned $\mathrm{span}\{S'_i\}$. Since there are at least k of them, this will hold unless there is some $v \in \mathrm{span}\{S'_i\}$ so that v is orthogonal to all of the $x^{(j)} \otimes y^{(j)}$. This shouldn't happen because each $x^{(j)} \otimes y^{(j)}$ is very unlikely to be orthogonal to v.

To deal with the constraint that the coefficients are not too large, we use linear programming duality to show that there will be a solution unless there is some v that is *nearly* orthogonal to all of the $x^{(j)} \otimes y^{(j)}$. Again, this is unlikely to happen for any individual term, and thus, by independence, highly unlikely to happen for all j simultaneously. Combining this with a cover argument will give our proof.

Lemma 3.3.6 (Existence of a Bi-Linear Form)**.** *Given the preconditions in Proposition 3.3.2, with probability at least* $99/100$ *over the choice of* $x^{(j)}, y^{(j)}$*, there exist* $b_j \in [-D, D]$ *for* $j \in [4k]$*, where* $D = \mathcal{O}\Big(k^4/(\sqrt{\eta}\alpha)\Big)$*, such that the projection of* $\sum_{j=1}^{t} b_j x^{(j)} \otimes y^{(j)}$ *onto the space spanned by the* S'_i *is* P*, where* P *satisfies the conclusion of Proposition 3.3.4.*

Proof. To prove this lemma, we again use a linear programming based argument. Consider the following (primal) linear program in the variables b_j, for $j \in [4k]$:

$$\sum_{j \in [4k]} b_j \langle S'_i, x^{(j)} \otimes y^{(j)} \rangle = \langle S'_i, P \rangle \qquad \forall \;\; i \in [k] \tag{3.17}$$

$$-D \leqslant b_j \leqslant D \qquad \forall \;\; j \in [4k] \tag{3.18}$$

We note that a set of b_j satisfying Equation (3.17) will have the projection of $\sum_{j \in [4k]} b_j x^{(j)} \otimes y^{(j)}$ onto the span of the S'_i be the same as the projection of P, and that if the b_j's satisfy Equation (3.18) then we will have $|b_j| \leqslant D$ for all j. Thus, it suffices to show that with high probability over our choice of $x^{(j)}$ and $y^{(j)}$ that the above system is feasible.

We will show this by linear programming duality (since this is now a finite system of equations, we can use standard results rather than Fact 3.3.5). In particular, we have that Equations (3.17) and (3.18) are simultaneously satisfiable unless there are real numbers c_i and non-negative real numbers z_j, z'_j so that

$$\sum_{i=1}^{k} c_i \sum_{j \in [4k]} b_j \langle S'_i, x^{(j)} \otimes y^{(j)} \rangle + \sum_{j \in [4k]} (z_j - z'_j) b_j \leqslant \sum_{i=1}^{k} c_i \langle S'_i, P \rangle + \sum_{j \in [4k]} (z_j + z'_j) D$$

yields a contradiction. Setting $v = \sum_{i=1}^{k} c_i S_i'$, the above simplifies to

$$\sum_{j\in[4k]} b_j \left(\langle v, x^{(j)} \otimes y^{(j)}\rangle + z_j - z_j'\right) \leqslant \langle v, P\rangle + \sum_{j\in[4k]} (z_j + z_j')D \tag{3.19}$$

We note that in order for Equation (3.19) to be a contradiction, it must be the case that the coefficients of b_j are all 0. In particular, we must have

$$z_j' - z_j = \langle v, x^{(j)} \otimes y^{(j)}\rangle$$

for all j. In particular, this means that

$$z_j + z_j' \geqslant \left|\langle v, x^{(j)} \otimes y^{(j)}\rangle\right|.$$

In such a case, the right hand side of Equation (3.19) will be at least

$$\langle v, P\rangle + \sum_{j\in[4k]} \left|\langle v, x^{(j)} \otimes y^{(j)}\rangle\right| D$$

Therefore, Equation (3.19) can only yield a contradiction if there exists a $v \in \operatorname{span}\{S_i'\}$ so that

$$\langle v, P\rangle < -\sum_{j\in[4k]} \left|\langle v, x^{(j)} \otimes y^{(j)}\rangle\right| D. \tag{3.20}$$

We want to show that with high probability over our choice of $x^{(j)}, y^{(j)}$ that there is no $v \in \operatorname{span}\{S_i'\}$ satisfying Equation (3.20). In fact, we will show that for every such v that

$$\sum_{j\in[4k]} \left|\langle v, x^{(j)} \otimes y^{(j)}\rangle\right| \geqslant \frac{c_1}{\sqrt{\eta}\alpha} \|v\|_F.$$

We can scale v so that $\|v\|_F = 1$, and it suffices to show that

$$\sum_{j\in[4k]} \left|\left\langle \tilde{v}, x^{(j)} \otimes y^{(j)}\right\rangle\right| \geqslant \left(\frac{c_1}{\sqrt{\eta}\alpha D}\right) \tag{3.21}$$

holds for all unit vectors v in $\operatorname{span}\{S_i'\}$ with high probability.

Since we need to show that infinitely many equations all hold with high probability, we will use a cover argument. In particular, we can construct $\mathcal{C}$, a τ-cover for all unit vectors v in the span of the S_i', where we take $\tau = \left(\frac{c_1'}{k^2\sqrt{\eta}\alpha D}\right)$. Since this is a cover of a unit sphere in a k-dimensional

subspace, we can construct such a cover so that $|\mathcal{C}| = \mathcal{O}(1/\tau)^k$. Replacing v with the closest point in $\mathcal{C}$, denoted by v', it suffices to show that with high probability for all v that

$$\sum_{j\in[4k]} \left|\left\langle v, x^{(j)} \otimes y^{(j)} \right\rangle\right| \geqslant \sum_{j\in[4k]} \left|\left\langle v', x^{(j)} \otimes y^{(j)} \right\rangle\right| - \sum_{j\in[4k]} \left|\left\langle v - v', x^{(j)} \otimes y^{(j)} \right\rangle\right| \geqslant \left(\frac{2c_1}{\sqrt{\eta}\alpha D}\right). \tag{3.22}$$

We begin by bounding the terms

$$\sum_{j\in[4k]} \left|\left\langle v - v', x^{(j)} \otimes y^{(j)} \right\rangle\right|.$$

For this we notice by Cauchy-Schwartz that each term is at most $\|v - v'\|_F$ times the Frobenius norm of the projection of $x^{(j)} \otimes y^{(j)}$ onto the span of the S_i'. We note that for any k-dimensional subspace W with orthonormal basis $w_1, \ldots, w_k$ we have that

$$\begin{aligned}\mathbf{E}\left[\left\|\mathrm{Proj}_W(x^{(j)} \otimes y^{(j)})\right\|_F^2\right] &= \sum_{i=1}^{k} \left|\left\langle w_i, x^{(j)} \otimes y^{(j)} \right\rangle\right|^2 \\ &= k.\end{aligned}$$

Therefore, with high probability over the choice of $x^{(j)}, y^{(j)}$ each of the projections of $x^{(j)} \otimes y^{(j)}$ onto the span of the S_i' has Frobenius norm $\tilde{O}(\sqrt{k})$. Therefore, if this condition holds over our choice of $x^{(j)}$ and $y^{(j)}$, we can show Equation (3.22) if we can show that

$$\sum_{j\in[4k]} \left|\left\langle v', x^{(j)} \otimes y^{(j)} \right\rangle\right| \geqslant \left(\frac{c_1}{\sqrt{\eta}\alpha D}\right) \geqslant \left(\frac{2c_1}{\sqrt{\eta}\alpha D}\right) - \tau\tilde{O}(k^{3/2}) \tag{3.23}$$

for all $v' \in \mathcal{C}$.

Each term in $\sum_{j\in[4k]} \left\langle v', x^{(j)} \otimes y^{(j)} \right\rangle$ is a random bi-linear form given by $z_j = \sum_{\ell,p\in[d]} v'_{\ell,p} x_\ell^{(j)} y_p^{(j)}$. Then, we have that $\mathbb{E}[z_j] = 0$ and

$$\begin{aligned}\mathbb{E}\left[z_j^2\right] = \mathbb{E}\left[\left(\sum_{\ell,p\in[d]} v'_{\ell,p} x_\ell^{(j)} y_p^{(j)}\right)^2\right] &= \sum_{\ell,\ell',p,p'} \mathbb{E}\left[v'_{\ell,p} v'_{\ell',p'} x_\ell^{(j)} x_{\ell'}^{(j)} y_p^{(j)} y_{p'}^{(j)}\right] \\ &= \sum_{\ell,p\in[d]} \left(v'_{\ell,p}\right)^2 \cdot \mathbb{E}\left[\left(x_\ell^{(j)}\right)^2\right] \cdot \mathbb{E}\left[\left(y_p^{(j)}\right)^2\right] \\ &= 1\,,\end{aligned}$$

where the last equality follows from $\tilde{v}'_F = 1$.

Using Lemma 3.2.10 with $\zeta = \frac{2c_1}{\sqrt{\eta}\alpha D}$,

$$\mathbf{Pr}\left[|z_j| \leqslant \frac{c_1}{\sqrt{\eta}\alpha D}\right] \leqslant c_5 \left(\frac{2c_1}{\sqrt{\eta}\alpha D}\right)^{1/2} . \tag{3.24}$$

However, we note that Equation (3.23) will hold unless $|z_j| \leqslant \frac{c_1}{\sqrt{\eta}\alpha D}$ for all $j \in [4k]$. Since the z_j's are independent, we conclude that

$$\mathbf{Pr}\left[\sum_{j\in[4k]} \left|\left\langle v', x^{(j)} \otimes y^{(j)}\right\rangle\right| \leqslant \frac{c_1}{\sqrt{\eta}\alpha D}\right] \leqslant O\left(\frac{c_1}{\sqrt{\eta}\alpha D}\right)^{2k} . \tag{3.25}$$

Since the above argument holds for any $v' \in \mathcal{C}$, we can union bound over all elements in the cover $\mathcal{C}$, and the probability that there exists a $\tilde{v}'$ in the cover that does not satisfy Equation (3.23) is at most $O\left(k^2\sqrt{\eta}\alpha D\right)^k \cdot O\left(\frac{c_1}{\sqrt{\eta}\alpha D}\right)^{2k}$. Setting D to be a sufficiently large multiple of $(k^4/(\sqrt{\eta}\alpha))$ suffices to conclude that with probability at least $1 - 1/\mathrm{poly}(k)$, the primal is feasible. □

Proof of Proposition 3.3.2. We begin by bounding the Frobenius norm of $\hat{T}_4$. Let $T_4 = \mathbf{E}[h_4(\mathcal{X})]$. It then follows from Lemma 3.2.6 that

$$T_4 = \mathrm{Sym}\left(\sum_{i=1}^{k} w_i \left(3S_i \otimes S_i + 6S_i \otimes \mu_i^{\otimes 2} + \mu_i^{\otimes 4}\right)\right) . \tag{3.26}$$

Further, $\|S_i \otimes S_i\|_F \leqslant \|S_i\|_F^2 \leqslant \Delta^2$, $\left\|S_i \otimes \mu_i^{\otimes 2}\right\|_F \leqslant \|S_i\|_F \|\mu_i\|_2^2 \leqslant 4\Delta/\alpha$, and $\left\|\mu_i^{\otimes 4}\right\|_F \leqslant \|\mu_i\|_2^4 \leqslant 16/\alpha^2$. Since T_4 is an average of terms of the form $S_i^{\otimes 2}, S_i \otimes \mu_i^{\otimes 2}$ and $\mu_i^{\otimes 4}$, and each such term is upper bounded, we can conclude that $\|T_4\|_F = \mathcal{O}(\Delta^2 + 1/\alpha^2)$, and by the triangle inequality that $\left\|\hat{T}_4\right\|_F \leqslant \mathcal{O}(\Delta^2 + 1/\alpha^2 + \eta)$. Let $S_i' = S_i + \mu_i^{\otimes 2}$ and let $T_4' := \sum_{i=1}^k w_i \left(S_i' \otimes S_i'\right)$. We can then rewrite Equation (3.26) as follows:

$$T_4 = \mathrm{Sym}\left(\sum_{i=1}^{k} w_i \left(3S_i' \otimes S_i' - 2\mu_i^{\otimes 4}\right)\right) . \tag{3.27}$$

For $j \in [4k]$, let $x^{(j)}, y^{(j)} \sim \mathcal{N}(0, I)$. Collapsing two modes of $\hat{T}_4$, it follows from Equation

(3.27) that for any fixed j,

$$\begin{aligned}\hat{T}_4\left(\cdot,\cdot,x^{(j)},y^{(j)}\right) &= \left(\hat{T}_4 - T_4\right)\left(\cdot,\cdot,x^{(j)},y^{(j)}\right) + T_4\left(\cdot,\cdot,x^{(j)},y^{(j)}\right) \\ &= \left(\hat{T}_4 - T_4\right)\left(\cdot,\cdot,x^{(j)},y^{(j)}\right) + \mathrm{Sym}\left(\sum_{i=1}^{k} w_i\left(3S_i'\otimes S_i' - 2\mu_i^{\otimes 4}\right)\right)\left(\cdot,\cdot,x^{(j)},y^{(j)}\right) \\ &= \left(\hat{T}_4 - T_4 + T_4'\right)\left(\cdot,\cdot,x^{(j)},y^{(j)}\right) + \sum_{i\in[k]} w_i\left(S_i'x^{(j)}\right)\otimes\left(S_i'y^{(j)}\right) \\ &\quad + \sum_{i\in[k]} w_i\left(S_i'y^{(j)}\right)\otimes\left(S_i'x^{(j)}\right) + \sum_{i\in[k]} w_i\left(-2\mu_i^{\otimes 2}\left\langle\mu_i,x^{(j)}\right\rangle\left\langle\mu_i,y^{(j)}\right\rangle\right)\ ,\end{aligned} \tag{3.28}$$

where we use that $\mathrm{Sym}(\cdot)$ is a linear operator satisfying $\mathrm{Sym}\left(\mu_i^{\otimes 4}\right) = \mu_i^{\otimes 4}$, and

$$\mathrm{Sym}\left(S_i'\otimes S_i'\right) = \frac{1}{3}S_i'\otimes S_i' + \frac{1}{3}S_i'\oplus S_i' + \frac{1}{3}S_i'\ominus S_i'$$

where for indices (i_1,i_2,i_3,i_4),

$$\left(S_i'\oplus S_i'\right)(i_1,i_2,i_3,i_4) = \left(S_i'\otimes S_i'\right)(i_1,i_3,i_2,i_4)$$

and $\left(S_i'\ominus S_i'\right)(i_1,i_2,i_3,i_4) = \left(S_i'\otimes S_i'\right)(i_1,i_4,i_2,i_3)$.

Next, it follows from Lemma 3.3.4 that there exists a matrix $\tilde{P}_i$ such that $\left\|\tilde{P}_i\right\|_F = \mathcal{O}\left(1/(\sqrt{\eta}\alpha)\right)$ and $\left\|T_4'\left(\cdot,\cdot,\tilde{P}_i\right) - S_i'\right\|_F = \mathcal{O}\left(\sqrt{\eta/\alpha}\right)$. Furthermore, with probability at least 0.99, there exists a sequence of $b_j\in[-D,D]$, for $j\in[4k]$, such that $T_4'\left(\cdot,\cdot,\sum_{j\in[4k]} b_j x^{(j)}\otimes y^{(j)}\right) = T_4'\left(\cdot,\cdot,\tilde{P}_i\right)$.

Consider a cover, $\mathcal{C}$, of the interval $[-D,D]$ with points spaced at intervals of length $\tau = \mathcal{O}\left(\frac{\sqrt{\eta}}{\alpha k(\Delta^4+1/\alpha^4)}\right)$. Since we uniformly sample a_j's, with probability at least $(\tau/D)^{\mathcal{O}(k)}$, for all $j\in[4k]$, $|b_j - a_j|\leqslant\tau$, and we condition on this event. Thus,

$$\begin{aligned}\left\|T_4'\left(\cdot,\cdot,\sum_{j\in[4k]} a_j x^{(j)}y^{(j)}\right) - S_i'\right\|_F &\leqslant \left\|T_4'\left(\cdot,\cdot,\sum_{j\in[4k]} b_j x^{(j)}\otimes y^{(j)}\right) - S_i'\right\|_F \\ &\quad + \left\|T_4'\left(\cdot,\cdot,\sum_{j\in[4k]}(b_j-a_j)x^{(j)}\otimes y^{(j)}\right)\right\|_F \\ &\leqslant \mathcal{O}\left(\sqrt{\eta/\alpha}\right) + \mathcal{O}\left(\tau\Delta^2\right) \leqslant \mathcal{O}\left(\sqrt{\eta/\alpha}\right)\ .\end{aligned} \tag{3.29}$$

Taking the linear combinations with coefficients a_j in Equation (3.28), we have

$$\begin{aligned}\hat{T}_4\left(\cdot,\cdot,\sum_{j\in[4k]}a_jx^{(j)}\otimes y^{(j)}\right)-S_i&=\left(\hat{T}_4-T_4+T_4'\right)\left(\cdot,\cdot,\sum_{j\in[4k]}a_jx^{(j)}\otimes y^{(j)}\right)-S_i'-\mu_i\otimes\mu_i\\&+\sum_{j\in[4k]}a_j\sum_{i\in[k]}w_i\left(S_i'x^{(j)}\right)\otimes\left(S_i'y^{(j)}\right)+\sum_{j\in[4k]}a_j\sum_{i\in[k]}w_i\left(S_i'y^{(j)}\right)\otimes\left(S_i'x^{(j)}\right)\\&+\sum_{j\in[4k]}a_j\sum_{i\in[k]}w_i\left(-2\mu_i^{\otimes 2}\left\langle\mu_i,x^{(j)}\right\rangle\left\langle\mu_i,y^{(j)}\right\rangle\right)\ .\end{aligned}\tag{3.30}$$

Setting $P_i=\left(\hat{T}_4-T_4+T_4'\right)\left(\cdot,\cdot,\sum_{j\in[4k]}a_jx^{(j)}y^{(j)}\right)-S_i'$, it follows from Lemma 3.2.9 that with probability at least 0.99, $\left(\hat{T}_4-T_4\right)\left(\cdot,\cdot,\sum_{j\in[4k]}a_jx^{(j)}y^{(j)}\right)$ has Frobenius norm $\mathcal{O}(kD\eta)$ and it follows from Equation (3.29) that with probability at least 0.99, $T_4'\left(\cdot,\cdot,\sum_{j\in[4k]}a_jx^{(j)}y^{(j)}\right)-S_i'$ has Frobenius norm $\mathcal{O}\left(\sqrt{\eta/\alpha}\right)$. Setting the remaining terms to Q_i, with probability at least 0.99 we can bound their Frobenius norm as follows:

$$\begin{aligned}\|Q_i\|_F&\leqslant\|\mu_i\otimes\mu_i\|_F+\left\|\left(\sum_{i\in[k]}w_iS_i'\oplus S_i'+w_iS_i'\ominus S_i'-2w_i\mu_i^{\otimes 4}\right)\left(\cdot,\cdot,\sum_{j\in[4k]}a_jx^{(j)}y^{(j)}\right)\right\|_F\\&\leqslant\frac{4}{\alpha}+\left(2\max_{i\in k}\|S_i'\|_F^2+\frac{32}{\alpha^2}+k\tau\right)\cdot\left\|\tilde{P}\right\|_F\\&\leqslant\frac{4}{\alpha}+\mathcal{O}\left(\frac{1}{\sqrt{\eta}\alpha}\left(\Delta+\frac{1}{\alpha}\right)^2\right)\\&\leqslant\mathcal{O}\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3}\right)\ ,\end{aligned}\tag{3.31}$$

where the first inequality follows from the triangle inequality, the second follows from our assumptions that $\|\mu_i\|_2\leqslant 2/\sqrt{\alpha}$, $\sum_{j\in[4k]}b_jx^{(j)}y^{(j)}=\tilde{P}_i$ in the span of the S_i', and $|a_j-b_j|\leqslant\tau$ for all $j\in[4k]$, and the third inequality follows from the definition of S_i', the bound on $\left\|\tilde{P}\right\|_F$ and the bound on $\|S_i-I\|_F$. □

3.3.5 Finding a Low-dimensional Subspace for Exhaustive Search

In this subsection, we will prove Proposition 3.3.3.

We start by extending Theorem 4 of [MV10], which shows that large parameter distance between pairs of univariate Gaussian mixtures implies large distance between their low-degree

moments. In the following, we use $M_j(F) = \mathbb{E}\left[F^j\right]$ to denote the j-th moment of a distribution F. We show:

Lemma 3.3.7. *There exists a constant $C > 0$ such that the following holds: Fix any $D > 0$ and $0 \leqslant \beta \leqslant 1/(2(2k-1)!D^{2k-3})$. Suppose that $F = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \sigma_i^2)$ is a univariate k-mixture of Gaussians with $w_i \geqslant \beta$, and $|\mu_i|, \sigma_i \leqslant D$, for all $i \in [k]$. If $|\mu_i| + |\sigma_i^2 - 1| \geqslant \beta$ for some $i \leqslant k$, then*

$$\max_{j \in [2k]} |M_j(F) - M_j\left(\mathcal{N}(0,1)\right)| \geqslant \beta^{C^{k+1}(k+1)!-1} .$$

We give the proof of Lemma 3.3.7 in Section 3.3.6.

Lemma 3.3.8 (Bounding μ_i's and S_i's in non-influential directions for $\mathbb{E}\left[h_m(\mathcal{M})\right]$)**.** *Let $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians on $\mathcal{R}^d$ satisfying $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and $\|\Sigma_i - I\|_F \leqslant \Delta$ for every $i \in [k]$. For some $B \in \mathcal{R}$, let $u \in \mathcal{R}^d$ be a unit vector such that $|\mathbb{E}\left[h_m\left(\langle \mathcal{M}, u\rangle\right)\right]| \leqslant B$ for all $m \in [2k]$. Then, for $\delta = 2^{O(k)} B^{1/(C^{k+1}(k+1)!)}$ and $S_i = \Sigma_i - I$, we have that:*

1. *for all $i \leqslant k$, $|\langle u, \mu_i\rangle|, |u^\top (I - \Sigma_i) u| \leqslant \delta$,*
2. *$\|S_i u\|_2^2 \leqslant 20\delta\Delta/\alpha^2 + B/\alpha$,*

where $C > 0$ is a fixed universal constant.

Proof. The 1-D random variable $\langle u, \mathcal{M}\rangle$ is a mixture of Gaussians described by $\sum_{i=1}^{k} w_i \mathcal{N}(\langle \mu_i, u\rangle, u^\top \Sigma_i u)$. Towards a contradiction, assume that there is an $i \in [k]$ such that $|\langle u, \mu_i\rangle| + |u^\top (I - \Sigma_i) u| \geqslant \delta$. Then, applying Lemma 3.3.7, yields that there is a $j \in [2k]$ such that $|M_j(\langle u, \mathcal{M}\rangle) - M_j(\mathcal{N}(0,1))| \geqslant \delta^{C^{k+1}(k+1)!-1}$. Applying Fact 3.2.5 implies that there exists an $m \in [2k]$ such that

$$|\mathbb{E}\left[h_m(\langle u, \mathcal{M}\rangle)\right]| > 2^{-O(k)} \delta^{C^{k+1}(k+1)!-1} \gg B ,$$

yielding a contradiction.

We can now prove the second part. Recall that for $S_i = \Sigma_i - I$ for every i, we have that

$$\mathbb{E}\left[h_4(\mathcal{M})\right] = \sum_{i=1}^{k} w_i \mathrm{Sym}\left(3\left(S_i \otimes S_i\right) + 6\left(S_i \otimes \mu_i^{\otimes 2}\right) + \mu_i^{\otimes 4}\right) .$$

We consider the $d \times d$ matrix obtained by the natural flattening of the $d \times d$ tensor $u^{\otimes 2} \cdot \mathbb{E}\left[h_4(\mathcal{M})\right]$.

Then, we can write:

$$u^{\otimes 2} \cdot \mathbb{E}\left[h_4(\mathcal{M})\right] = \sum_{i=1}^{k} w_i \Big((u^\top S_i u) S_i + 2(S_i u)(S_i u)^\top + \langle u, \mu_i \rangle^2 S_i + 2 \langle u, \mu_i \rangle \mu_i (S_i u)^\top + 2 \langle u, \mu_i \rangle (S_i u) \mu_i^\top + (u^\top S_i u) \mu_i \mu_i^\top + \langle u, \mu_i \rangle^2 \mu_i \mu_i^\top \Big) . \quad (3.32)$$

Now, from the first part, we know that for all $i \in [k]$, $|u^\top S_i u| \leqslant \delta$ and the hypothesis of the lemma gives us that $\|S_i\|_F = \|\Sigma_i - I\|_F \leqslant \Delta$. Thus, for each i, the first term in the summation above has Frobenius norm at most $\Delta\delta$. Using that $\langle u, \mu_i \rangle_2^2 \leqslant \delta^2$ from the first part of the lemma, yields that, for each i, the Frobenius norm of the third term is at most $\Delta\delta^2$.

Next, using in addition that $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$ yields that, for each i, the Frobenius norm of the 4th and 5th terms are at most $2\delta\Delta/\sqrt{\alpha}$ and the Frobenius norm of the 6th and 7th terms are at most δ/α. Thus, for each i and all but the 2nd term in the summation above, we have an upper bound on the Frobenius norm of $4\delta\Delta/\alpha$.

Now, since $|\mathbb{E}\left[h_4\left(\langle \mathcal{M}, u \rangle\right)\right]| \leqslant B$, and u is a unit vector, we have that $\|u^{\otimes 2}\, \mathbb{E}\left[h_4(\mathcal{M})\right]\|_F \leqslant B$. Thus, combining the aforementioned argument with the triangle inequality, we have for each i,

$$\begin{aligned} \|S_i u\|_2^2 = \left\| S_i u \left(S_i u\right)^\top \right\|_F &\leqslant \frac{1}{\alpha} \left\| u^{\otimes 2} \cdot \mathbb{E}\left[h_4(\mathcal{M})\right] \right\|_F + \sum_{i \in [k]} w_i \Big(\left(u^\top S_i u + \langle u, \mu_i \rangle^2 \right) \|S_i\|_F \Big) \\ &\quad + \sum_{i \in [k]} 4 w_i \Big(\left(\langle u, \mu_i \rangle \right) \|\mu_i\|_2 \|S_i u\|_2 \Big) + \sum_{i \in [k]} 4 w_i \Big(\left(\langle u, \mu_i \rangle^2 + u^\top S_i u \right) \|\mu_i\|_2^2 \Big) \\ &\leqslant B/\alpha + 15\delta\Delta/\alpha \, , \end{aligned}$$

and the claim follows. □

Lemma 3.3.9 (Subspace covering all the means and large singular vectors of $S_i = \Sigma_i - I$)**.** *Let $\mathcal{M} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians on $\mathcal{R}^d$ satisfying $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$, and $\|\Sigma_i - I\|_F \leqslant \Delta$ for all $i \in [k]$. Given $0 < \eta < 1$, let $\hat{T}_m$ satisfy $\left\| \hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right] \right\|_F \leqslant \eta$ for every $m \in [4k]$ and let $\lambda \geqslant 2\eta$. Let V be the span of all the left singular vectors of the $d \times d^{m-1}$ matrix obtained by the natural flattening of $\hat{T}_m$ with singular values at least λ. Then, for $\delta = 2\lambda^{1/(2C^{k+1}(k+1)!)}$, we have that:*

1. $\dim V \leqslant \left(4k\eta^2 + k^{O(k)}\right) \mathcal{O}(1 + 1/\alpha + \Delta)^{4k} / \lambda^2$,

2. *Let*

$$V_{\text{inf}} = \{\mu_i\}_{i\in[k]} \cup \Big\{ v \mid \exists i \in [k],\ \textit{s.t. } \|v\|_2 = 1 \textit{ and } v \textit{ is an eigenvector of } S_i \textit{ and } \|S_i v\|_2 \geqslant \sqrt{\delta} \Big\}_{i\leqslant k} .$$

Then, for every unit vector $v \in V_{\text{inf}}$, $\|v - \Pi_V v\|_2^2 \leqslant 20\delta^{1/4}\Delta/\alpha^2$, *where* $\Pi_V v$ *is the projection of* v *onto* V.

Proof. From Fact 3.2.6, we have that $\mathbb{E}\left[h_m(\mathcal{M})\right] = \sum_{i\in[k]} w_i \,\mathbb{E}\left[h_m(G_i)\right]$, where $G_i = \mathcal{N}(\mu_i, \Sigma_i)$, and since $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$ and $\|\Sigma_i - I\|_F \leqslant \Delta$, it follows that $\|\mathbb{E}\left[h_m(\mathcal{M})\right]\|_F^2 \leqslant \mathcal{O}\left(m(1+1/\alpha+\Delta)\right)^{4m}$. From Proposition 3.3.1, we know that

$$\left\|\hat{T}_m\right\|_F^2 \leqslant 2\left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F^2 + 2\left\|\mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F^2 \leqslant \eta^2 + \mathcal{O}\left(m(1+1/\alpha+\Delta)\right)^{4m} .$$

Thus, the number of singular vectors of the $d \times d^{m-1}$ flattening of $\hat{T}_m$ with a singular value $\geqslant \lambda$ is at most $(\eta^2 + \mathcal{O}(m(1+1/\alpha+\Delta))^{4m})/\lambda^2$. Summing up this bound for all $m \in [4k]$, yields the claimed upper bound on $\dim V$.

For the second part, we will first bound $\langle u, v\rangle$ for any unit vector u orthogonal to the subspace V. Towards this, observe that since u is orthogonal to V and $\|u\|_2 = 1$, we have

$$\left\|u \cdot \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F \leqslant \|u\hat{T}_m\|_F + \|\hat{T}_m - \mathbf{E}h_m(\mathcal{M})\|_F \leqslant \lambda + \eta \leqslant 2\lambda ,$$

where $u \cdot \mathbb{E}\left[h_m(\mathcal{M})\right]$ is a matrix-vector product of u with a $d \times d^{m-1}$ flattening of $\mathbb{E}\left[h_m(\mathcal{M})\right]$. For $\delta = 2\lambda^{1/(C^{k+1}(k+1)!)}$, applying Lemma 3.3.8 yields that

$$\langle \mu_i, u\rangle^2 + \|S_i u\|_2^2 \leqslant \delta^2 + 20\delta\Delta/\alpha \leqslant 20\delta\Delta/\alpha^2 . \tag{3.33}$$

Now, if v is one of the μ_i's, then we immediately get from Equation 3.33 that $\langle v, u\rangle^2 \leqslant 20\delta\Delta/\alpha^2$. Similarly, note that if v is a unit length eigenvector of S_i satisfying $\|S_i v\|_2^2 \geqslant \sqrt{\delta}$, then,

$$\langle u, v\rangle^2 = \frac{1}{\|S_i v\|_2^2}\langle u, S_i v\rangle^2 = \frac{1}{\|S_i v\|_2^2}\langle S_i u, v\rangle^2 \leqslant \frac{\|S_i u\|_2^2}{\|S_i v\|_2^2} .$$

In both cases, setting $u = (v - \Pi_V v)/\|v - \Pi_V v\|_2$ completes the proof. □

We can now complete the proof of Proposition 3.3.3:

Proof of Proposition 3.3.3. We know that $\hat{S}_i - P_i - S_i$ is a symmetric, rank-k' matrix such that

$k' = \mathcal{O}(k^2)$, described by the eigenvalue decomposition $\sum_{i=1}^{k'} \tau_i v_i v_i^\top$, where v_i's are the eigenvectors and τ_i's are the corresponding eigenvalues. Since $\|S_i\|_F \leqslant \Delta$ and

$$\left\|\hat{S}_i\right\|_F \leqslant \|P_i\|_F + \|Q_i\|_F + \|S_i\|_F \leqslant \mathcal{O}\left(\sqrt{\eta/\alpha}\right) + \mathcal{O}\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3}\right) + \Delta = \mathcal{O}\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3}\right) ,$$

we have that the number of singular values of $\hat{S}_i$ that exceed $\delta^{1/4}$ is at most $\mathcal{O}\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3\sqrt{\delta}}\right)$. Recall that from Lemma 3.3.9 it follows that the dimension of the subspace V is at most $k^{\mathcal{O}(k)}\mathcal{O}(1 + 1/\alpha + \Delta)^{4k}/\lambda^2$. Thus, the dimension of V' is at most

$$k^{\mathcal{O}(k)}\mathcal{O}\left(\frac{(1+1/\alpha+\Delta)^{4k}}{\lambda^2}\right) + \mathcal{O}\left(\frac{1+\Delta^2}{\sqrt{\eta}\alpha^3\sqrt{\delta}}\right) = \mathcal{O}\left(\frac{k^{\mathcal{O}(k)}\left(1+\frac{1}{\alpha}+\Delta\right)^{4k+5}}{\eta^2}\right) .$$

Since V' contains V constructed in Lemma 3.3.9, we immediately obtain that for every μ_i, $\|\mu_i - \Pi_{V'}\mu_i\|_2^2 \leqslant \frac{20}{\alpha^2}\sqrt{\delta}\Delta$.

Next, let u be a unit vector orthogonal to V'. Then, since V' contains the V described in Lemma 3.3.9, we know that $\|S_i u\|_2^2 \leqslant \frac{20}{\alpha^2}\sqrt{\delta}\Delta$. Similarly, since V' contains all eigenvectors of $\hat{S}_i$ with singular values exceeding $\delta^{1/4}$, we know that $\left\|\hat{S}_i u\right\|_2^2 \leqslant \delta^{1/2}$. Thus, we can conclude that $\left\|(\hat{S}_i - S_i)u\right\|_2^2 \leqslant \frac{100}{\alpha^2}\sqrt{\delta}\Delta$. Let $Q_i = \sum_{j=1}^{k'} \tau_j v_j v_j^\top$ with orthonormal $v_j \in \mathcal{R}^d$. We know such τ_j's and v_j's exist because of the upper bound on $\operatorname{rank}(Q_i)$. Therefore, for any j, $|v_j^\top(\hat{S}_i - S_i)u| \leqslant \frac{10}{\alpha}\delta^{1/4}\Delta^{1/2}$. On the other hand, for any j, we have that

$$v_j^\top(\hat{S}_i - S_i)u \geqslant \langle v_j, u\rangle \tau_j - \|P_i\|_F = \langle v_j, u\rangle \tau_j - \mathcal{O}(\sqrt{\eta}) .$$

Combining the two bounds above, yields that whenever $\tau_j \geqslant \delta^{1/4}$,

$$|\langle v_j, u\rangle| \leqslant \mathcal{O}(\sqrt{\eta}/\tau_j) + \frac{10}{\alpha\tau_j}\delta^{1/4}\Delta^{1/2} \leqslant \frac{10}{\alpha}\delta^{1/2}\Delta^{1/2} .$$

Thus, the matrix $\hat{Q}_i = \sum_{j=1}^{k'} \tau_j \Pi_{V'} v_j (\Pi_{V'} v_j)^\top$ has its range space in V' and satisfies

$$\left\|\hat{Q}_i - Q_i\right\|_F \leqslant \mathcal{O}\left(k^2\delta^{1/4}\right) + \mathcal{O}\left(\frac{k^2}{\alpha}\delta^{1/2}\Delta^{1/2}\right) = \mathcal{O}\left(\frac{k^2}{\alpha}\delta^{1/2}\Delta^{1/2}\right) .$$

□

3.3.6 Parameter vs Moment Distance for Gaussian Mixtures

In this subsection, we prove Lemma 3.3.7. To that end, we will use the following two results; the second one is from [MV10].

Lemma 3.3.10. *Suppose $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ are univariate Gaussians with $|\mu_i|, |\sigma_i| \leqslant D$, for some $D \in \mathcal{R}_+$. If $|\mu_1 - \mu_2| + |\sigma_1^2 - \sigma_2^2| \leqslant \beta$, then the distance between raw moments of two Gaussians is*

$$\left|M_j(\mathcal{N}(\mu_1, \sigma_1^2)) - M_j(\mathcal{N}(\mu_2, \sigma_2^2))\right| \leqslant (j+1)! D^{j-1}\beta.$$

Proof. By Proposition 3.2.3, the j-th raw moment of a Gaussian $\mathcal{N}(\mu, \sigma^2)$ is a sum of monomials in μ and σ^2 of degree j. There are at most $(j+1)!$ terms in the polynomial. Thus, changing the mean or the variance by at most β will change the j-th moment by at most $(j+1)!D^{j-1}\beta$. □

Theorem 74. *([MV10]) Let F, F' be two univariate mixtures of Gaussians: $F = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \sigma_i^2)$ and $F' = \sum_{i=1}^{k'} w_i' \mathcal{N}(\mu_i', \sigma_i'^2)$. There is a constant $c > 0$ such that, for any $\beta < c$, if F, F' satisfy:*

1. *$w_i, w_i' \in [\beta, 1]$*
2. *$|\mu_i|, |\mu_i'| \leqslant 1/\beta$*
3. *$|\mu_i - \mu_{i'}| + |\sigma_i^2 - \sigma_{i'}^2| \geqslant \beta$ and $|\mu_i' - \mu_{i'}'| + |\sigma_i'^2 - \sigma_{i'}'^2| \geqslant \beta$ for all $i \neq i'$*
4. *$\beta \leqslant \min_\pi \sum_i \left(|w_i - w_{\pi(i)}'| + |\mu_i - \mu_{\pi(i)}'| + |\sigma^2 - \sigma_{\pi(i)}'^2|\right)$, where the minimization is taken over all mappings $\pi : \{1, \ldots, k\} \to \{1, \ldots, k'\}$,*

then

$$\max_{j \in [2(k+k'-1)]} |M_j(F) - M_j(F')| \geqslant \beta^{O(k)} .$$

We are now ready to complete the proof of Lemma 3.3.7.

Proof of Lemma 3.3.7. We proceed via induction on k. Consider the base case, i.e., $k = 1$. Then, either $|\mu_1| \geqslant \beta/2$ or $|\sigma_1 - 1| \geqslant \beta/2$, and thus the first or second moment differ by at least $\beta^2/4$. Let the inductive hypothesis be that Lemma 3.3.7 holds for at most k components.

Consider the case where $|\mu_i - \mu_{i'}| + |\sigma_i^2 - \sigma_{i'}^2| \geqslant \beta^{C^k k!}$ for all pairs of components $i, i' \in [k]$. Then, by Theorem 74, we have that

$$\max_{j \in [2k]} |M_j(F) - M_j(\mathcal{N}(0,1))| \geqslant \beta^{C^{k+1} k!} \geqslant \beta^{C^{k+1}(k+1)!-1} ,$$

and the lemma follows.

Otherwise, we know that there exists a pair of components with parameter distance less than $\beta^{C^k k!}$. In this case, we merge these two components and get a $(k-1)$-mixture F'. By Lemma 3.3.10, the distance between the j-th moments of F' and F is at most $(j+1)!D^{j-1}\beta^{C^k k!}$. Since we still have $|\mu_i'| + |\sigma_i'^2 - 1| \geqslant \beta - 3\beta^{C^k k!}$ for all components i in F', the inductive hypothesis implies that

$$\max_{j\in[2k-2]} |M_j(F') - M_j(\mathcal{N}(0,1))| \geqslant \left(\beta - 3\beta^{C^k k!}\right)^{C^k (k)!-1} .$$

By the triangle inequality, we can write

$$\begin{aligned}\max_{j\in[2k]} |M_j(F) - M_j(\mathcal{N}(0,1))| &\geqslant \max_{j\in[2k-2]} |M_j(F') - M_j(\mathcal{N}(0,1))| - \max_{j\in[2k-2]} |M_j(F) - M_j(F')| \\ &\geqslant \left(\beta - 3\beta^{C^k k!}\right)^{C^k k!-1} - (2k-1)!D^{2k-3}\beta^{C^k k!} \\ &\geqslant \beta^{C^{k+1}(k+1)!-1} .\end{aligned}$$

The last inequality follows from the assumption that $\beta \leqslant 1/(2(2k-1)!D^{2k-3})$. This completes the proof of Lemma 3.3.7. □

3.4 Robust Partial Cluster Recovery

In this section, we give two robust *partial clustering* algorithms. A partial clustering algorithm takes a set of points $X = \cup_{i\leqslant k} X_i$ with true clusters $X_1, X_2, \ldots, X_k$ and outputs a partition of the sample $X = X_1' \cup X_2'$ such that $X_1' = \cup_{i\in S} X_i$ and $X_2' = \cup_{i\notin S} X_i$, for some subset $S \subseteq [k]$ of size $1 \leqslant |S| < k$. That is, a partial clustering algorithm partitions the sample into two non-empty parts so that each part is a sample from a "sub-mixture". This is a weaker guarantee than clustering the entire mixture, which must find each of the original X_i's. We show that the relaxed guarantee is feasible even when the mixture as a whole is not clusterable. In our setting, we will get an approximate (that is, a small fraction of points are misclassified) partial clustering that works for ϵ-corruptions Y of any i.i.d. sample X from a mixture of k Gaussians, as long as there is a pair of components in the original mixture that have large total variation distance between them.

A partial clustering algorithm such as above was one of the innovations in [BK20b] that allowed for a polynomial-time algorithm for clustering all fully clusterable Gaussian mixtures.

In this section, we build on the ideas in [BK20b] to derive two new partial clustering algorithms that work even when the original mixture is not fully clusterable. Both upgrade the results of [BK20b] by handling mixtures with arbitrary weights w_is instead of uniform weights and handling mixtures where not all pairs of components are well-separated in TV distance. The first algorithm succeeds under the information-theoretically minimal separation assumption (i.e. separation in total variation distance) but runs in time exponential in the inverse mixing weight. The second algorithm is a key innovation of this paper – it gives an algorithm that runs in polynomial time in the inverse mixing weight at the cost of handling separation only in relative Frobenius distance. This improved running time guarantee (at the cost of strong separation requirement that we mitigate through a novel standalone spectral separation step in Section 3.5) is crucial to obtaining the fully polynomial running time in our algorithm.

In order to state the guarantees of our algorithms, we first formulate a notion of parameter separation as the next definition.

Definition 3.4.1 (Δ-Parameter Separation)**.** *We say that two Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ are Δ-parameter separated if at least one of the following three conditions hold:*

1. ***Mean-Separation:*** *$\exists v \in \mathcal{R}^d$ such that $\langle \mu_1 - \mu_2, v\rangle^2 > \Delta^2 v^\top(\Sigma_1 + \Sigma_2)v$,*
2. ***Spectral-Separation:*** *$\exists v \in \mathcal{R}^d$ such that $v^\top \Sigma_1 v > \Delta\, v^\top \Sigma_2 v$,*
3. ***Relative-Frobenius Separation:*** *Σ_i and Σ_j have the same range space and $\left\|\Sigma_1^{\dagger/2}(\Sigma_2 - \Sigma_1)\Sigma_1^{\dagger/2}\right\|_F^2 > \Delta^2 \left\|\Sigma_1^\dagger \Sigma_2\right\|_{\mathrm{op}}^2$.*

As shown in [BK20b, DHKK20], if a pair of Gaussians is $(1-\exp(-O(\Delta \log \Delta))$-separated in total variation distance, then, they are Δ-parameter separated.

Our first algorithm succeeds in robust partial clustering whenever there is a pair of component Gaussians that are Δ-parameter separated. The running time of this algorithm grows exponentially in the reciprocal of the minimum weight in the mixture.

Theorem 75 (Robust Partial Clustering in TV Distance)**.** *Let $0 \leqslant \epsilon < \alpha \leqslant 1$, and $\eta > 0$. There is an algorithm with the following guarantees: Let $\{\mu_i, \Sigma_i\}_{i \leqslant k}$ be means and covariances of k unknown Gaussians. Let Y be an ϵ-corruption of a sample X of size $n \geqslant (dk)^{Ct}/\epsilon$ for a large enough constant $C > 0$, from $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ satisfying Condition 3.2.45 with parameters $t = (k/\eta)^{O(k)}$ and $\gamma \leqslant \epsilon d^{-8t} k^{-Ct}$, for a sufficiently large constant $C > 0$. Suppose further that $w_i \geqslant \alpha > 2\epsilon$ for every i and that there are i, j such that $\mathcal{N}(\mu_i, \Sigma_i)$ and $\mathcal{N}(\mu_j, \Sigma_j)$ are Δ-parameter separated for $\Delta = (k/\eta)^{O(k)}$.*

Then, the algorithm on input Y, runs in time $n^{(k/\eta)^{O(k)}}$, and with probability at least $2^{-O\left(\frac{1}{\alpha}\log\left(\frac{k}{\eta\alpha}\right)\right)}$ over the draw of X and the algorithm's random choices, the algorithm outputs a partition of Y into Y_1, Y_2 satisfying:

1. ***Partition respects clustering:*** *for each i, $\max\{\frac{k}{n}|Y_1 \cap X_i|, \frac{k}{n}|Y_2 \cap X_i|\} \geqslant 1-\eta-O(\varepsilon/\alpha^4)$, and,*

2. ***Partition is non-trivial:*** $\max_i \frac{k}{n}|X_i \cap Y_1|, \max_i \frac{k}{n}|X_i \cap Y_2| \geqslant 1-\eta-O(\varepsilon/\alpha^4)$.

Our proof of the above theorem is based on a relatively straightforward extension of the ideas of [BK20b], albeit with two key upgrades 1) allowing the input mixtures to have arbitrary mixing weights (at an exponential cost in the inverse of the minimum weight) and 2) handling mixtures where some pair of components may not be well-separated in TV distance.

In order to get our main result that gives a fully polynomial algorithm (including in the inverse mixing weights), we will use a incomparable variant of the above partial clustering method that only handles a weaker notion of parameter separation, but runs in fixed polynomial time.

Theorem 76 (Robust Partial Clustering in Relative Frobenius Distance)**.** *Let $0 \leqslant \epsilon < \alpha/k \leqslant 1$ and $t \in \mathbb{N}$. There is an algorithm with the following guarantees: Let $\{\mu_i, \Sigma_i\}_{i\leqslant k}$ be means and covariances of k unknown Gaussians. Let Y be an ϵ-corruption of a sample X of size $n \geqslant (dk)^{Ct}/\epsilon$ for a large enough constant $C > 0$, from $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ that satisfies Condition 3.2.45 with parameters $2t$ and $\gamma \leqslant \epsilon d^{-8t}k^{-Ck}$, for a large enough constant $C > 0$. Suppose further that $w_i \geqslant \alpha > 2\epsilon$ for each $i \in [k]$, and that for some $t \in \mathbb{N}$, $\beta > 0$ there exist $i, j \leqslant k$ such that $\left\|\Sigma^{\dagger/2}(\Sigma_i - \Sigma_j)\Sigma^{\dagger/2}\right\|_F^2 = \Omega\left((k^2t^4)/(\beta^{2/t}\alpha^4)\right)$, where Σ is the covariance of the mixture $\mathcal{M}$. Then, the algorithm runs in time $n^{O(t)}$, and with probability at least $2^{-O\left(\frac{1}{\alpha}\log\left(\frac{k}{\eta}\right)\right)}$ over the random choices of the algorithm, outputs a partition $Y = Y_1 \cup Y_2$ satisfying:*

1. ***Partition respects clustering:*** *for each i, $\max\left\{\frac{1}{w_i n}|Y_1 \cap X_i|, \frac{1}{w_i n}|Y_2 \cap X_i|\right\} \geqslant 1-\beta-O(\epsilon/\alpha^4)$, and,*

2. ***Partition is non-trivial:*** $\max_i \frac{1}{w_i n}|X_i \cap Y_1|, \max_i \frac{1}{w_i n}|X_i \cap Y_2| \geqslant 1-\beta-O(\epsilon/\alpha^4)$.

The starting point for the proof of the above theorem is the observation that the running time of our first algorithm above is exponential in the inverse mixing weight almost entirely because of dealing with spectral separation (which requires the use of "certifiable anti-concentration" that we define in the next subsection). We formulate a variant of relative Frobenius separation (that is directly useful to us) and prove that whenever the original mixture has a pair of components separated in this notion, we can in fact obtain a fully polynomial partial clustering algorithm

building on the ideas in [BK20b].

3.4.1 Algorithm

Our algorithm will solve SoS relaxations of a polynomial inequality system. The constraints here use the input Y to encode finding a sample X' (the intended setting being $X' = X$, the original uncorrupted sample) and a cluster $\hat{C}$ in X' of size $= \alpha n$, indicated by z_is (the intended setting is simply the indicator for any of the k true clusters) satisfying properties of Gaussian distribution (certifiable hypercontractivity and anti-concentration).

Covariance constraints introduce a matrix valued indeterminate Π intended to be the square root of $\hat{\Sigma}$, the sos variable for the covariance of a single component.

$$\text{Covariance Constraints: } \mathcal{A}_1 = \left\{ \begin{array}{c} \Pi = UU^\top \\ \Pi^2 = \hat{\Sigma} \end{array} \right\} \tag{3.34}$$

The intersection constraints force that X' be ϵ-close to Y (and thus, 2ϵ-close to unknown sample X).

$$\text{Intersection Constraints: } \mathcal{A}_2 = \begin{Bmatrix} \forall i \in [n], & m_i^2 = m_i \\ & \sum_{i\in[n]} m_i = (1-\epsilon)n \\ \forall i \in [n], & m_i(y_i - x'_i) = 0 \end{Bmatrix} \tag{3.35}$$

The subset constraints introduce z, which indicates the subset $\hat{C}$ intended to be the true clusters of X'.

$$\text{Subset Constraints: } \mathcal{A}_3 = \begin{Bmatrix} \forall i \in [n]. & z_i^2 = z_i \\ & \sum_{i\in[n]} z_i = \alpha n \end{Bmatrix} \tag{3.36}$$

Parameter constraints create indeterminates to stand for the covariance $\hat{\Sigma}$ and mean $\hat{\mu}$ of $\hat{C}$ (indicated by z).

$$\text{Parameter Constraints: } \mathcal{A}_4 = \left\{ \begin{array}{c} \frac{1}{\alpha n}\sum_{i=1}^{n} z_i \left(x'_i - \hat{\mu}\right)\left(x'_i - \hat{\mu}\right)^\top = \hat{\Sigma} \\ \frac{1}{\alpha n}\sum_{i=1}^{n} z_i x'_i = \hat{\mu} \end{array} \right\} \tag{3.37}$$

Certifiable Hypercontractivity : $\mathcal{A}_4$=

$$\left\{\begin{array}{ll} \forall t \leqslant 2s & \frac{1}{\alpha^2 n^2}\sum_{i,j\leqslant n} z_i z_j (Q(x'_i - x'_j) - \mathbf{E}_z Q)^{2t} \leqslant \left(\frac{1}{\alpha^2 n^2}\sum_{i,j\leqslant n} z_i z_j (Q(x'_i - x'_j) - \mathbf{E}_z Q)^2\right)^t \\ & \frac{1}{\alpha^2 n^2}\sum_{i,j\leqslant n} z_i z_j \left(Q(x'_i - x'_j) - \mathbf{E}_z Q\right)^2 \leqslant \frac{6}{\alpha^2}\|Q\|_F^2 \end{array}\right\} \tag{3.38}$$

Here, we used the shorthand $\mathbf{E}_z Q = \frac{1}{\alpha^2 n^2}\sum_{i,j\leqslant n} z_i z_j Q(x'_i - x'_j)$.

In the constraint system for our first algorithm, we will use the following certifiable anti-concentration constraints on $\hat{C}$ for $\delta = \alpha^{-\mathrm{poly}(k)}$ and $\tau = \alpha/\mathrm{poly}(k)$ and $s(u) = 1/u^2$ for every u.

$$\text{Anti-Concentration} : \mathcal{A}_5 = \left\{\begin{array}{l} \frac{1}{\alpha^2 n^2}\sum_{i,j=1}^{n} z_i z_j q^2_{\delta,\Sigma}\left(\left(x'_i - x'_j\right), v\right) \leqslant 2^{s(\delta)} C\delta \left(v^\top \Sigma v\right)^{s(\delta)} \\ \frac{1}{\alpha^2 n^2}\sum_{i,j=1}^{n} z_i z_j q^2_{\tau,\Sigma}\left(\left(x'_i - x'_j\right), v\right) \leqslant 2^{s(\tau)} C\tau \left(v^\top \Sigma v\right)^{s(\eta)} \end{array}\right\} \tag{3.39}$$

We note that the constraint system for our second algorithm (running in fixed polynomial time), we will not use $\mathcal{A}_5$. Towards proving Theorems 76 and 75 we use the following algorithm that differs only in the degree of the pseudo-distribution computed and the constraint system that the pseudo-distribution satisfies.

Algorithm 77 (Partial Clustering)**.**

Given: *A sample Y of size n. An outlier parameter $\epsilon > 0$ and an accuracy parameter $\eta > 0$.*

Output: *A partition of Y into partial clustering $Y_1 \cup Y_2$.*

Operation:

1. ***SDP Solving:*** *Find a pseudo-distribution $\tilde{\zeta}$ satisfying $\cup_{i=1}^5 \mathcal{A}_i$ ($\cup_{i=1}^4 \mathcal{A}_i$ for Theorem 76) such that $\tilde{\mathbb{E}}_{\tilde{\zeta}} z_i \leqslant \alpha + o_d(1)$ for every i. If no such pseudo-distribution exists, output fail.*
2. ***Rounding:*** *Let $M = \tilde{\mathbb{E}}_{z\sim\tilde{\zeta}}[zz^\top]$.*
   (a) *Choose $\ell = \mathcal{O}\left(\frac{1}{\alpha}\log(k/\eta)\right)$ rows of M uniformly at random and independently.*
   (b) *For each $i \leqslant \ell$, let $\hat{C}_i$ be the indices of the columns j such that $M(i,j) \geqslant$*

$\eta^2\alpha^5/k$.

(c) Choose a uniformly random $S \subseteq [\ell]$ *and output* $Y_1 = \cup_{i\in S}\hat{C}_i$ *and* $Y_2 = Y \setminus Y_1$.

3.4.2 Analysis

Simultaneous Intersection Bounds. The key observation for proving the first theorem is the following lemma that gives a sum-of-squares proof that no z that satisfies the constraints $\cup_{i=1}^5 \mathcal{A}_i$ can have simultaneously large intersections with the Δ-parameter separated component Gaussians.

Lemma 3.4.2 (Simultaneous Intersection Bounds for TV-separated case)**.** *Let* Y *be an* ϵ*-corruption of a sample* X *of size* $n \geqslant (dk)^{Ct}/\epsilon$ *for a large enough constant* $C > 0$*, from* $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ *satisfying Condition 3.2.45 with parameters* $t = (k/\eta)^{O(k)}$ *and* $\gamma \leqslant \epsilon d^{-8t}k^{-Ct}$*, for a sufficiently large constant* $C > 0$*. Suppose further that* $w_i \geqslant \alpha > 2\epsilon$ *for every* i *and that there are* i, j *such that* $\mathcal{N}(\mu_i, \Sigma_i)$ *and* $\mathcal{N}(\mu_j, \Sigma_j)$ *are* Δ*-parameter separated for* $\Delta = (k/\eta)^{O(k)}$*. Then, there exists a partition of* $[k]$ *into* $S \cup L$ *such that,* $|S|, |L| < k$ *and for* $z(X_r) = \frac{1}{w_r n}\sum_{i\in X_r} z_i$*,*

$$\left\{\cup_{i=1}^5 \mathcal{A}_i\right\} \left|\frac{z}{(k/\eta\alpha)^{poly(k)}}\right. \left\{\sum_{i\in S, j\in L} z(X_i)z(X_j) \leqslant O(k^2\epsilon/\alpha) + \eta/\alpha\right\}.$$

The proof of Lemma 3.4.2 is given in Section 3.4.3.

For the second theorem, we use the following version that strengthens the separation assumption and lowers the degree of the sum-of-squares proof (and consequently the running time of the algorithm) as a result.

Lemma 3.4.3 (Simultaneous Intersection Bounds for Frobenius Separated Case)**.** *Let* X *be a sample of size* $n \geqslant (dk)^{Ct}/\epsilon$ *for a large enough constant* $C > 0$*, from* $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ *that satisfies Condition 3.2.45 with parameters* $2t$ *and* $\gamma \leqslant \epsilon d^{-8t}k^{-Ck}$*, for a large enough constant* $C > 0$*. Suppose further that* $w_i \geqslant \alpha > 2\epsilon$ *for each* $i \in [k]$*, and that for some* $t \in \mathbb{N}$*,* $\beta > 0$ *there exist* $i, j \leqslant k$ *such that* $\left\|\Sigma^{\dagger/2}(\Sigma_i - \Sigma_j)\Sigma^{\dagger/2}\right\|_F^2 = \Omega\left((k^2t^4)/(\beta^{2/t}\alpha^2)\right)$*, where* Σ *is the covariance of the mixture* $\mathcal{M}$*. Then, for any* ϵ*-corruption* Y *of* X*, there exists a partition of*

$[k] = S \cup T$ *such that*

$$\left\{\cup_{i=1}^{4}\mathcal{A}_i\right\} \vdash_{2t}^{z} \left\{\sum_{i\in S}\sum_{j\in T} z(X_i)z(X_j) \leqslant O(k^2)\beta + O(k^2)\epsilon/\alpha\right\}.$$

Here, $z(X_r) = \frac{1}{w_r n}\sum_{i\in X_r} z_i$ *for every* r.

The proof of Lemma 3.4.3 is given in Section 3.4.4.

Notice that the main difference between the above two lemmas is the constraint systems they use. Specifically, the second lemma does *not* enforce certifiable anti-concentration constraints. As a result, there is a difference in the degree of the sum-of-squares proofs they claim; the degree of the SoS proof in the second lemma does not depend on the inverse minimum mixture weight.

First, we complete the proof of the Theorem 75. The proof of Theorem 76 is exactly the same except for the use of Lemma 3.4.3 (and thus has the exponent in the running time independent of $1/\alpha$) instead of Lemma 3.4.2.

Proof of Theorem 75. Let $\eta' = O(\eta^2\alpha^3/k)$. We will prove that whenever $\Delta \geqslant \mathrm{poly}(k/\eta')^k = \mathrm{poly}\left(\frac{k}{\eta\alpha}\right)^k$, Algorithm 77, when run with input Y, with probability at least 0.99, recovers a collection $\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_\ell$ of $\ell = \mathcal{O}\left(\frac{1}{\alpha}\log k/\eta\right)$ subsets of indices satisfying $|\cup_{i\leqslant\ell}\hat{C}_i| \geqslant (1 - \eta'/k^{40})n$ such that there is a partition $S \cup L = [\ell]$, $0 < |S| < \ell$ satisfying:

$$\min\left\{\frac{1}{\alpha n}|\hat{C}_i \cap \cup_{j\in S}X_j|, \frac{1}{\alpha n}|\hat{C}_i \cap \cup_{j\in L}X_j|\right\} \leqslant 100\eta'/\alpha^3 + O(\epsilon/\alpha^4)\,. \tag{3.40}$$

We first argue that this suffices to complete the proof. Split $[\ell]$ into two groups G_S, G_L as follows. For each i, let $j = \arg\max_{r\in[\ell]}\frac{1}{\alpha n}|\hat{C}_i \cap X_r|$. If $j \in S$, add it to G_S, else add it to G_L. Observe that this process is well-defined - i.e, there cannot be $j \in S$ and $j' \in L$ that both maximize $\frac{1}{\alpha n}|\hat{C}_i \cap X_r|$ as r varies over $[k]$. To see this, WLOG, assume $j \in S$. Note that $\frac{1}{\alpha n}|\hat{C}_i \cap X_j| \geqslant 1/k$. Then, we immediately obtain: $\frac{1}{\alpha n}|\cup_{j\in S} X_j \cap \hat{C}_i| \geqslant 1/k$. Now, if we ensure that $\eta' \leqslant \alpha^3/k^2$ and $\epsilon \leqslant O(\alpha^4/k)$, then, $\frac{1}{\alpha n}|\hat{C}_i \cap \cup_{j'\in L}X_{j'}|$ is at most the RHS of (3.40) which is $\ll 1/k$. This completes the proof of well-definedness. Next, adding up (3.40) for each $i \in S$ yields that

$$\frac{1}{|\hat{C}_i|}|\left(\cup_{i\in G_S}X_i\right) \cap \cup_{j\in L}X_j| \leqslant O(\log(k/\eta')/\alpha)\left(\eta' + \mathcal{O}(\epsilon/\alpha)\right),$$

where we used that $|G_S| \leqslant \ell$. Combined with $|\cup_{i \leqslant \ell} \hat{C}_i| \geqslant (1 - \eta'/k^{40})n$, we obtain that

$$|\cup_{i \in G_S} X_i| \geqslant 1 - \eta'/k^{40} - O(\log(k/\eta')/\alpha)\,(\eta' + O(\epsilon/\alpha)) = \eta + \mathcal{O}\Big(\log(k/\eta\alpha)\epsilon/\alpha^2\Big)$$

for $\eta' \leqslant \mathcal{O}(\eta^2\alpha^3/k)$.

We now go ahead and establish (3.40). Let $\tilde{\zeta}$ be a pseudo-distribution satisfying $\mathcal{A}$ of degree $(k/\eta)^{\mathrm{poly}(k)}$ satisfying $\tilde{\mathbb{E}}_{\tilde{\zeta}} z_i = \alpha$ for every i. Such a pseudo-distribution exists. To see why, let $\tilde{\zeta}$ be the actual distribution that always sets $X' = X$, chooses an i with probability w_i and outputs a uniformly subset $\hat{C}$ of size αn of X_i conditioned on $\hat{C}$ satisfying $\mathcal{A}$. Then, notice that since X satisfies Condition 3.2.45, by Fact 3.2.43, the uniform distribution on each X_i has t-certifiably C-hypercontractive degree 2 polynomials and is t-certifiably $C\delta$-anti-concentrated. By an concentration argument using high-order Chebyshev inequality, similar to the proof of Lemma 3.2.49 (applied to uniform distribution on X_i of size $n \geqslant (dk)^{O(t)}$, $\hat{C}$ chosen above satisfies the constraints $\mathcal{A}$ with probability at least $1 - o_d(1)$. Observe that the probabililty that z_i is set to 1 under this distribution is then at most $\alpha + o_d(1)$. Thus, such a distribution satisfies all the constraints in $\mathcal{A}$.

Next, let $M = \tilde{\mathbb{E}}_{\tilde{\zeta}}[zz^\top]$. Then, we claim that:

1. $o_d(1) + \alpha \geqslant M(i,j) \geqslant 0$ for all i, j,
2. $M(i,i) \in \alpha \pm o_d(1)$ for all i,
3. $\mathbf{E}_{j \sim [n]} M(i,j) \geqslant \alpha^2 - o_d(1)$ for every i.

The proofs of these basic observations are similar to those presented in Chapter 4.3 of [FKP+19] (see also the proof of Theorem 5.1 in [BK20b]): Observe that $\mathcal{A} \vdash_4 \Big\{z_i z_j = z_i^2 z_j^2 \geqslant 0\Big\}$ for every i. Thus, by Fact 3.2.18, $\tilde{\mathbb{E}}[z_i z_j] \geqslant 0$ for every i, j. Next, observe that $\mathcal{A} \vdash_2 \{(1 - z_i) = (1 - z_i)^2 \geqslant 0\}$ for every i and thus, $\mathcal{A} \vdash_2 \{z_i(1 - z_j) \geqslant 0\}$. Thus, by Fact 3.2.18 again, we must have $\tilde{\mathbb{E}}[z_i z_j] \leqslant \tilde{\mathbb{E}}[z_i] \leqslant \alpha + o_d(1)$. Finally, $\mathcal{A} \vdash_2 \Big\{\sum_j z_i z_j = z_i \sum_j z_j = \alpha n z_i\Big\}$. Thus, by Fact 3.2.18 again, we must have $\sum_j M(i,j) = \sum_j \tilde{\mathbb{E}}[z_i z_j] = \alpha n \sum_j \tilde{\mathbb{E}}[z_i] \in (\alpha^2 \pm o_d(1))n$. Let B_i be the entries in the i-th row M_i that are larger than $\alpha^2/2$. Then, by (1) and (2), we immediately derive that B_i must have at least $\alpha n/2$ elements. Call an entry of M large if it exceeds $\alpha^2\eta'$. For each i, let B_i be the set of large entries in row i of M. Then, using (3) and (1) above gives that $|B_i| \geqslant \alpha(1 - \alpha\eta')n$ for each $1 \leqslant i \leqslant n$. Next, call a row i "good" if $\frac{1}{\alpha n} \min\{|\cup_{r \in L} X_r \cap B_i|, |\cup_{r' \in S} X_{r'} \cap B_i|\} \leqslant 100\eta'/\alpha^3 + O(\epsilon/\alpha^4)$. Let us estimate the fraction of rows of M that are good.

Towards that goal, let us apply Lemma 3.4.2 with η set to η' and use Fact 3.2.18 (SoS

Completeness), to obtain $\sum_{r\in S,r'\in L}\mathbf{E}_{i\in X_r}\mathbf{E}_{j\in X_{r'}}M(i,j) \leqslant \eta' + O(\epsilon/\alpha)$. Using Markov's inequality, with probability $1-\alpha^3/100$ over the uniformly random choice of i, $\mathbf{E}_{j\in X_{r'}}M(i,j) \leqslant 100\frac{1}{\alpha^3}\eta' + O(\epsilon/\alpha^4)$. Thus, $1-\alpha^3/100$ fraction of the rows of M are good.

Next, let R be the set of $\frac{100}{\alpha}\log\left(\frac{k^{50}}{\eta'}\right)$ rows sampled in the run of the algorithm and set $\hat{C}_i = B_i$ for every $i \in R$. The probability that all of them are good is then at least $(1-\alpha^3/100)^{\frac{100}{\alpha}\log\left(\frac{k^{50}}{\alpha\eta'}\right)} \geqslant 1-\alpha$. Let us estimate the probability that $|\cup_{i\in R}\hat{C}_i| \geqslant (1-\eta'/k^{40})n$. For a uniformly random i, the chance that a given point $t \in B_i$ is at least $\alpha(1-\alpha\eta')$. Thus, the chance that $t \notin \cup_{i\in R}B_i$ is at most $(1-\alpha/2)^{100/\alpha\log\left(k^{50}/(\alpha\eta')\right)} \leqslant \eta'/k^{50}$. Thus, the expected number of t that are not covered by $\cup_{i\in R}\hat{C}_i$ is at most $n\eta'/k^{50}$. Thus, by Markov's inequality, with probability at least $1-1/k^{10}$, $1-\eta'/k^{40}$ fraction of t are covered in $\cup_{i\in R}\hat{C}_i$. By the above computations and a union bound, with probability at least $1-\eta'/k^{10}$ both the conditions below hold simultaneously: 1) each of the $\frac{100}{\alpha}\log(k^{50}/\eta')$ rows R sampled are good and 2) $|\cup_{i\in R}\hat{C}_i| \geqslant (1-\eta'/k^{40})n$. This completes the proof. □

3.4.3 Proof of Lemma 3.4.2

Our proof is based on the following simultaneous intersection bounds from [BK20b]. We will use the following lemma that forms the crux of the analysis of the clustering algorithm in [BK20b]:

Lemma 3.4.4 (Simultaneous Intersection Bounds, Lemma 5.4 in [BK20b])**.** *Fix $\delta > 0, k \in \mathbb{N}$. Let $X = X_1 \cup X_2 \cup \ldots X_k$ be a good sample of size n from a k-mixture $\sum_i w_i\mathcal{N}(\mu_i,\Sigma_i)$ of Gaussians. Let Y be any ϵ-corruption of X. Suppose there are $r, r' \leqslant k$ such that one of the following three conditions hold for some $\Delta \geqslant (k/\delta)^{O(k)}$:*

1. *there exists a v such that $v^\top\Sigma(r')v > \Delta v^\top\Sigma(r')v$ and $B = \max_{i\leqslant k}\frac{v^\top\Sigma(i)v}{v^\top\Sigma(r')v}$, or*
2. *there exists a $v \in \mathcal{R}^d$ such that $\langle\mu(r)-\mu(r'),v\rangle_2^2 \geqslant \Delta^2 v^\top\left(\Sigma(r)+\Sigma(r')\right)v$, or,*
3. *$\left\|\Sigma(r')^{-1/2}\Sigma(r)\Sigma(r')^{-1/2} - I\right\|_F^2 \geqslant \Delta^2\left(\left\|\Sigma(r')^{-1/2}\Sigma(r)^{1/2}\right\|_{\mathrm{op}}^4\right)$.*

Then, for the linear polynomial $z(X_r) = \frac{1}{\alpha n}\sum_{i\in X_r} z_i$ in indeterminates z_is satisfies:

$$\{\cup_{i\leqslant 5}\mathcal{A}_i\} \vdash_{\frac{z}{(k/\delta)^{O(k)}\log(2B)}} \left\{z(X_r)z(X_{r'}) \leqslant O(\sqrt{\delta}) + O(\epsilon/\alpha)\right\}.$$

Proof of Lemma 3.4.2. Without loss of generality, assume that the pair of separated components are $\mathcal{N}(\mu_1,\Sigma_1)$ and $\mathcal{N}(\mu_2,\Sigma_2)$. Let us start with the case when the pair is spectrally separated. Then, there is a $v \in \mathcal{R}^d$ such that $\Delta v^\top\Sigma_1 v \leqslant v^\top\Sigma_2 v$.

Consider an ordering of the true clusters along the direction v, renaming cluster indices if needed, such that $v^\top \Sigma_1 v \leqslant v^\top \Sigma_2 v \leqslant \ldots v^\top \Sigma_k v$. Let $j \leqslant k'$ be the largest integer such that $\mathrm{poly}(k/\eta) v^\top \Sigma_j v \leqslant v^\top \Sigma_{j+1} v$. Further, observe that since j is defined to be the largest index which incurs separation $\mathrm{poly}(k/\eta)$, all indices in $[j,k]$ have spectral bound at most $\mathrm{poly}(k/\eta)$ and thus $\frac{v^\top \Sigma_k v}{v^\top \Sigma_j v} \leqslant \mathrm{poly}(k/\eta)^k$.

Applying Lemma 3.4.4 with the above direction v to every $r < j$ and $r' \geqslant j$ and observing that the parameter B in each case is at most $\frac{v^\top \Sigma_k v}{v^\top \Sigma_j v} \leqslant \Delta^k$ yields:

$$\mathcal{A}\Big|_{\frac{z}{O(k^2 s^2 \mathrm{poly}\log(\Delta))}} \left\{ z(X_r) z(X_{r'}) \leqslant O(\epsilon/\alpha) + \sqrt{\delta} \right\} .$$

Adding up the above inequalities over all $r \leqslant j-1$ and $r' \geqslant j+1$ and taking $S = [j-1]$, $T = [k] \setminus [j-1]$ completes the proof in this case.

Next, let us take the case when $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ are mean-separated. WLOG, suppose $\langle \mu_1, v \rangle \leqslant \langle \mu_2, v \rangle \ldots \leqslant \langle \mu_k, v \rangle$. Then, we know that $\langle \mu_k - \mu_1, v \rangle \geqslant \Delta v^\top \Sigma_i v$. Thus, there must exist an i such that $\langle \mu_i - \mu_{i+1}, v \rangle \geqslant \Delta v^\top \Sigma_i v / k$. Let $S = [i]$ and $L = [k] \setminus S$. Applying Lemma 3.4.4 and arguing as in the previous case (and noting that $\kappa = \mathrm{poly}(k)$) completes the proof.

Finally, let us work with the case of relative Frobenius separation. Since $\|\Sigma_1^{-1/2} \Sigma_k^{1/2}\| \leqslant \mathrm{poly}(k)$, the hypothesis implies that $\|\Sigma_1 - \Sigma_2\|_F \geqslant \Delta/\mathrm{poly}(k)$. Let $B = \Sigma_1 - \Sigma_2$ and let $A = B/\|B\|_F$. WLOG, suppose $\langle \Sigma_1, A \rangle \leqslant \ldots \langle \Sigma_k, A \rangle$. Then, since $\langle \Sigma_k, A \rangle - \langle \Sigma_1, A \rangle \geqslant \Delta/\mathrm{poly}(k)$, there must exist an i such that $\langle \Sigma_{i+1}, A \rangle - \langle \Sigma_i, A \rangle \geqslant \Delta/\mathrm{poly}(k)$. Let us now set $S = [i]$ and $L = [k] \setminus S$.

Then, for every $i \in S$ and $j \in L$, we must have: $\langle \Sigma_j, A \rangle - \langle \Sigma_i, A \rangle \geqslant \Delta/\mathrm{poly}(k)$. Thus, $\|\Sigma_j - \Sigma_i\|_F \geqslant \Delta/\mathrm{poly}(k)$. And thus, $\Delta/\mathrm{poly}(k) \leqslant \|\Sigma_j - \Sigma_i\|_F \leqslant \left\|\Sigma_i^{-1/2} \Sigma_j^{1/2}\right\|_2^2 \left\|\Sigma_i^{-1/2} \Sigma_j \Sigma_i^{-1/2} - I\right\|_F$. Rearranging and using the bound on $\left\|\Sigma_i^{-1/2} \Sigma_j^{1/2}\right\|_2^2$ yields that $\left\|\Sigma_i^{-1/2} \Sigma_j \Sigma_i^{-1/2} - I\right\|_F \geqslant \Delta/\mathrm{poly}(k)$.

A similar argument as in the two cases above now completes the proof.

□

3.4.4 Proof of Lemma 3.4.3

We use $\mathbf{E}_z$ as a shorthand for $\frac{1}{\alpha n} \sum_{i=1}^n z_i$. We will write $\frac{1}{w_r n} \sum_{j \in X_r} z_j = z(X_r)$. Note that $z(X_r) \in [0,1]$. And finally, we will write $z'(X_r) = \frac{1}{w_r n} \sum_{j \in X_r} z_j \mathbf{1}(x_j = y_j)$ – the version of

$z(X_r)$ that only sums over non-outliers.

We will use the following technical facts in the proof:

Fact 3.4.5 (Lower Bounding Sums, Fact 4.19 [BK20b])**.** *Let A, B, C, D be scalar-valued indeterminates. Then, for any $\tau > 0$,*

$$\{0 \leqslant A, B \leqslant A + B \leqslant 1\} \cup \{0 \leqslant C, D\} \cup \{C + D \geqslant \tau\} \left|\frac{A,B,C,D}{2}\right. \{AC + BD \geqslant \tau AB\} \ .$$

Fact 3.4.6 (Cancellation within SoS, Lemma 9.2 in [BK20b])**.** *For indeterminate a and any $t \in \mathbb{N}$,*

$$\left\{a^{2t} \leqslant 1\right\} \left|\frac{a}{2t}\right. \{a \leqslant 1\} \ .$$

Lemma 3.4.7 (Lower-Bound on Variance of Degree 2 Polynomials)**.** *Let $Q \in \mathcal{R}^{d\times d}$. Then, for any $i, j \leqslant k$, and $z'(X_r) = \frac{1}{w_r n} \sum_{i\in X_r} z_i \mathbf{1}(x_i = y_i)$, we have:*

$$\mathcal{A} \left|\frac{z}{4}\right. \left\{ z'(X_r)^2 z'(X_r')^2 \leqslant \frac{(32Ct)^{2t}}{(\mathbf{E}_{X_r} Q - \mathbf{E}_{X_{r'}} Q)^{2t}} \left(\frac{\alpha^4}{w_r^2 w_{r'}^2} \left(\mathbf{E}_z (Q - \mathbf{E}_z Q)^2\right)^t \right.\right.$$
$$\left.\left. + \frac{\alpha^2}{w_r^2} \left(\mathbf{E}_{X_{r'}} (Q - \mathbf{E}_{X_{r'}} Q)^2\right)^t + \frac{\alpha^2}{w_{r'}^2} \left(\mathbf{E}_{X_r} (Q - \mathbf{E}_{X_r} Q)^2\right)^t \right) \right\} .$$

Proof. Let $z_i' = z_i \mathbf{1}(x_i = y_i)$ for every i. Using the substitution rule and non-negativity constraints of the z_i's, we have

$$\begin{aligned} \mathcal{A} \left|\frac{z}{4}\right. \Bigg\{ \mathbf{E}_z (Q - \mathbf{E}_z Q)^{2t} &= \frac{1}{\alpha^2 n^2} \sum_{i,j\leqslant n} z_i' z_j' \left(Q(x_i - x_j) - \mathbf{E}_z Q\right)^{2t} \\ &\geqslant \frac{1}{\alpha^2 n^2} \sum_{i,j\in X_r \text{ or } i,j\in X_{r'}} z_i' z_j' \left(Q(x_i - x_j) - \mathbf{E}_z Q\right)^{2t} \Bigg\} \end{aligned} \tag{3.41}$$

Using the SoS almost triangle inequality, we have

$$\mathcal{A}\left|\tfrac{z}{4}\right.\Bigg\{\frac{1}{\alpha^2n^2}\sum_{i,j\in X_r \text{ or } i,j\in X_{r'}} z'_iz'_j\left(Q(x_i-x_j)-\mathbf{E}_zQ\right)^{2t}$$
$$\geqslant\left(\frac{1}{2^{2t}}\right)\left(\frac{1}{\alpha^2n^2}\sum_{i,j\in X_r} z'_iz'_j\left(\mathbf{E}_{X_r}Q-\mathbf{E}_zQ\right)^{2t}-\frac{1}{\alpha^2n^2}\sum_{i,j\in X_r} z'_iz'_j\left(Q(x_i-x_j)-\mathbf{E}_{X_r}Q\right)^{2t}\right)$$
$$+\left(\frac{1}{2^{2t}}\right)\left(\frac{1}{\alpha^2n^2}\sum_{i,j\in X_r} z'_iz'_j\left(\mathbf{E}_{X_{r'}}Q-\mathbf{E}_zQ\right)^{2t}-\frac{1}{\alpha^2n^2}\sum_{i,j\in X_{r'}} z'_iz'_j\left(Q(x_i-x_j)-\mathbf{E}_{X_{r'}}Q\right)^{2t}\right)$$
$$=2^{-2t}\left((w_r/\alpha)^2z(X_r)^2\left(\mathbf{E}_{X_r}Q-\mathbf{E}_zQ\right)^{2t}-\frac{1}{\alpha^2n^2}\sum_{i,j\in X_r}\left(Q(x_i-x_j)-\mathbf{E}_{X_r}Q\right)^{2t}\right)$$
$$+2^{-2t}\left((w_{r'}/\alpha)^2z(X_{r'})^2\left(\mathbf{E}_{X_{r'}}Q-\mathbf{E}_zQ\right)^{2t}-\frac{1}{\alpha^2n^2}\sum_{i,j\in X_{r'}}\left(Q(x_i-x_j)-\mathbf{E}_{X_{r'}}Q\right)^{2t}\right)\Bigg\} \tag{3.42}$$

Using Fact 3.4.5, we can further simplify the above as follows:

$$\mathcal{A}\left|\tfrac{z}{4}\right.\Bigg\{\mathbf{E}_z(Q-\mathbf{E}_zQ)^{2t}\geqslant 2^{-6t}\frac{w_r^2w_{r'}^2}{\alpha^4}z'(X_r)^2z'(X_{r'})^2(\mathbf{E}_{X_r}Q-\mathbf{E}_{X_{r'}}Q)^{2t}$$
$$-2^{-6t}(w_r/\alpha)^2\mathbf{E}_{X_r}(Q-\mathbf{E}_{X_r}Q)^{2t}-2^{-6t}(w_{r'}/\alpha)^2\mathbf{E}_{X_{r'}}(Q-\mathbf{E}_{X_{r'}}Q)^{2t}$$
$$\geqslant 2^{-6t}\frac{w_r^2w_{r'}^2}{\alpha^4}z'(X_r)^2z'(X_{r'})^2(\mathbf{E}_{X_r}Q-\mathbf{E}_{X_{r'}}Q)^{2t}$$
$$-(w_r/\alpha)^2(Ct)^{2t}\left(\left(\mathbf{E}_{X_r}(Q-\mathbf{E}_{X_r}Q)^2\right)^t+\left(\mathbf{E}_{X_{r'}}(Q-\mathbf{E}_{X_{r'}}Q)^2\right)^t\right)\Bigg\} \tag{3.43}$$

where the last inequality follows from the Certifiable Hypercontractivity constraint ($\mathcal{A}_4$). Rearranging completes the proof.

□

We can use the lemma above to obtain a simultaneous intersection bound guarantee when there are relative Frobenius separated components in the mixture.

Lemma 3.4.8. *Suppose* $\left\|\Sigma^{-1/2}(\Sigma_r-\Sigma_{r'})\Sigma^{-1/2}\right\|_F^2\geqslant 10^8\frac{C^6t^4}{\beta^{2/t}\alpha^2}$. *Then, for* $z'(X_r)=\frac{1}{\alpha n}\sum_{i\in X_r}z_i$.

$\mathbf{1}(y_i = x_i)$ *for every* r,

$$\mathcal{A} \left|\frac{w}{2t}\right. \{z'(X_r)z'(X'_r) \leqslant \beta\} \ .$$

Proof. We work with the transformation $x_i \to \Sigma^{-1/2}x_i$. Let $\Sigma'_z = \Sigma^{-1/2}\Sigma_z\Sigma^{-1/2}$, $\Sigma'_r = \Sigma^{-1/2}\Sigma_r\Sigma^{-1/2}$ and $\Sigma'_{r'} = \Sigma^{-1/2}\Sigma_{r'}\Sigma^{-1/2}$ be the transformed covariances. Note that transformation is only for the purpose of the argument - our constraint system does not depend on Σ.

Notice that $\|\Sigma'_{r'}\|_2 \leqslant \frac{1}{w_r}$ and $\|\Sigma'_{r'}\|_2 \leqslant \frac{1}{w_{r'}}$.

We now apply Lemma 3.4.7 with $Q = \Sigma'_r - \Sigma'_{r'}$. Then, notice that $\mathbf{E}_{X_r}Q - \mathbf{E}_{X_{r'}}Q = \|\Sigma'_r - \Sigma'_{r'}\|_F^2 = \|Q\|_F^2$. Then, we obtain:

$$\mathcal{A} \left|\frac{z}{2t}\right. \left\{ z'(X_r)^2 z'(X_{r'})^2 \leqslant \left(\frac{32Ct}{\mathbf{E}_{X_r}Q - \mathbf{E}_{X_{r'}}Q}\right)^{2t} \cdot \left(\frac{\alpha^4}{w_r^2 w_{r'}^2}\left(\mathbf{E}_z(Q - \mathbf{E}_z Q)^2\right)^t + \frac{\alpha^2}{w_r^2}\left(\mathbf{E}_{X_{r'}}(Q - \mathbf{E}_{X_{r'}}Q)^2\right)^t + \frac{\alpha^2}{w_{r'}^2}\left(\mathbf{E}_{X_r}(Q - \mathbf{E}_{X_r}Q)^2\right)\right)^t \right\}. \tag{3.44}$$

Since X_r and $X_{r'}$ have certifiably C-bounded variance polynomials for $C = 4$ (as a consequence of Condition 3.2.45 and Fact 3.2.43 followed by an application of Lemma 3.2.25), we have:

$$\mathcal{A} \left|\frac{Q}{2}\right. \left\{ \mathbf{E}_{X_{r'}}(Q - \mathbf{E}_{X_{r'}}Q)^2 \leqslant 6 \left\|{\Sigma'_{r'}}^{1/2} Q {\Sigma'_{r'}}^{1/2}\right\|_F^2 \leqslant \frac{6}{w_{r'}^2}\|Q\|_F^2 \right\} ,$$

and

$$\mathcal{A} \left|\frac{Q}{2}\right. \left\{ \mathbf{E}_{X_r}(Q - \mathbf{E}_{X_r}Q)^2 \leqslant 6 \left\|{\Sigma'_r}^{1/2} Q {\Sigma'_r}^{1/2}\right\|_F^2 \leqslant \frac{6}{w_r^2}\|Q\|_F^2 \right\} .$$

Finally, using the bounded-variance constraints in $\mathcal{A}$, we have:

$$\mathcal{A} \left|\frac{Q,z}{4}\right. \mathbf{E}(Q - \mathbf{E}_z Q)^2 \leqslant \frac{6}{\alpha^2}\|Q\|_F^2 \ .$$

Plugging these estimates back in (3.44) yields:

$$\mathcal{A} \left|\frac{z}{4}\right. \left\{ z'(X_r)^2 z'(X_{r'})^2 \leqslant \frac{(1000Ct)^{2t}}{\alpha^{2t}\|Q\|_F^{2t}} \right\} . \tag{3.45}$$

Plugging in the lower bound on $\|Q\|_F^{2t}$ and applying Fact 3.4.6 completes the proof. □

We can use the above lemma to complete the proof of Lemma 3.4.3:

Proof of Lemma 3.4.3. WLOG, assume that $\Sigma = I$. Let $Q = \Sigma_r - \Sigma_{r'}$ and let $\bar{Q} = Q/\|Q\|_F$. Consider the numbers $v_i = \mathrm{tr}(\Sigma_r \cdot Q)$. Then, we know that $\max_{i,j} |v_i - v_j| \geqslant \|Q\|_F$. Thus, there must exist a partition of $[k] = S \cup T$ such that $|v_i - v_j| \geqslant \|Q\|_F / k$ whenever $i \in S$ and $j \in T$.

Thus, for every $i \in S$ and $j \in T$, $\|\Sigma_i - \Sigma_j\|_F^2 \geqslant \|Q\|_F^2 / k^2 = 10^8 \frac{C^6 t^4}{(\beta^{2/t}\alpha^2)}$. We can now apply Lemma above to every $i \in S, j \in T$, observe that $\mathcal{A} \vdash_4 \{z(X_r)z(X_{r'}) \leqslant z'(X_r)z'(X_{r'}) + 2\epsilon/\alpha\}$, and add up the resulting inequalities to finish the proof.

□

3.4.5 Special Case: Algorithm for Uniform and Bounded Mixing Weights

In this subsection, we obtain a polynomial time algorithm when the input mixture has weights that are bounded from below. This includes the case of uniform weights and when the minimum mixing weight is at least some function of k. At a high level, our algorithm partitions the sample into clusters as long as there is a pair of components separated in TV distance and given samples that are not clusterable, runs the tensor decomposition algorithm to list decode. We then use standard robust tournament results to pick a hypothesis from the list.

Theorem 78 (Robustly Learning Mixtures of Gaussians with Bounded Weights)**.** *Given* $0 < \epsilon < \mathcal{O}_k(1)$*, let* $Y = \{y_1, y_2, \ldots, y_n\}$ *be a multiset of* $n \geqslant n_0 = \mathit{poly}_k(d, 1/\varepsilon)$ ϵ*-corrupted samples from a* k*-mixture of Gaussians* $\mathcal{M} = \sum_{i \leqslant k} w_i \mathcal{N}(\mu_i, \Sigma_i)$*, such that* $w_i \geqslant \alpha$*. Then, there exists an algorithm with running time* $\mathit{poly}_k(n^{1/\alpha}) \cdot \exp(\mathit{poly}_k(1/\alpha, 1/\epsilon))$ *such that with probability at least* $9/10$ *it outputs a hypothesis* k*-mixture of Gaussians* $\widehat{\mathcal{M}} = \sum_{i \leqslant k} \hat{w}_i \mathcal{N}\left(\hat{\mu}_i, \hat{\Sigma}_i\right)$ *such that* $d_{TV}\left(\mathcal{M}, \widehat{\mathcal{M}}\right) = \mathcal{O}_k(\varepsilon)$.

Briefly, our algorithm simply does the following:

1. **Clustering via SoS**: Guess a partition of the mixture such that each component in the partition is not clusterable. Let the resulting partition have $t \leqslant k$ components. In parallel, try all possible ways to run Algorithm 77 repeatedly to obtain a partition of the samples, $\{\tilde{Y}_j\}_{j \in [t]}$ into exactly t components. For each such partition repeat the following.
2. **Robust Isotropic Transformation**: Run the algorithm corresponding to Lemma 3.6.4 on each set $\tilde{Y}_j$ to make the sample approximately isotropic. Grid search for weights over $[\alpha, 1/k]^k$ with precision α.

3. **List-Decoding via Tensor Decomposition**: Run Algorithm 73 on each $\tilde{Y}_j$. Concatenate the lists to obtain $\mathcal{L}$.
4. **Robust Tournament**: Run the tournament from Fact 3.2.50 over all the hypotheses in $\mathcal{L}$, and output the winning hypothesis.

Proof Sketch. Setting $\Delta = (k^{k^{O(k)}})$, it follows from Theorem 75 that we obtain a partition of Y into $\{\tilde{Y}_j\}_{j\in[t]}$, for some $t \in [k]$ such that $\tilde{Y}_j$ has at most $\mathcal{O}(k\epsilon/\alpha)$ outliers, $(1-\mathcal{O}(k\epsilon/\alpha))$-fraction of samples from at least one component of the input mixture and the resulting samples are not Δ-separated (see Definition 3.4.1). It then follows from Lemma 3.6.4 that the mean μ_j and covariance Σ_j of $\tilde{Y}_j$ satisfy : a) $\|\mu_j\|_2 \leqslant \mathcal{O}(\sqrt{\epsilon}k^{1.5}/\alpha^{1.5})$, b) $(1-\sqrt{\epsilon}k^{1.5}/\alpha^{1.5})\, I \preceq \Sigma_j \preceq (1-\sqrt{\epsilon}k^{1.5}/\alpha^{1.5})\, I$, and c) $\|\Sigma_j - I\|_F \leqslant \mathcal{O}(\sqrt{\epsilon}k^{1.5}/\alpha^{1.5})$.

Each component, $\tilde{Y}_j$, of the partition can have at most k components. Assuming these correspond to $\{w_i^{(j)}, \mu_i^{(j)}\Sigma_i^{(j)}\}_{i\in[k]}$, observe, $\sum_{i\in[k]} w_i^{(j)}\Sigma_i^{(j)} + w_i^{(j)}\mu_i^{(j)}\left(\mu_i^{(j)}\right)^\top \preceq (1+\sqrt{\epsilon}k^{1.5}/\alpha^{1.5})\, I$. Thus, we have that $\left\|\mu_i^{(j)}\right\|_2^2 \leqslant (1+\sqrt{\epsilon}k^{1.5})/\alpha^{2.5}$ and combined with not being Δ-separated, it follows that for all $i' \in [k]$,

$$\begin{aligned}
\left\|\Sigma_{i'}^{(j)} - I\right\|_F &= \left\|\Sigma_{i'}^{(j)} - \Sigma_j + (\Sigma_j - I)\right\|_F \\
&\leqslant \left\|\Sigma_{i'}^{(j)} - \sum_{i\in[k]} w_i^{(j)}\Sigma_i^{(j)} + \sum_{i\in[k]} w_i^{(j)}\mu_i^{(j)}\left(\mu_i^{(j)}\right)^\top\right\|_F + \|\Sigma_j - I\|_F \\
&\leqslant \left\|\sum_{i\in[k]} w_i^{(j)}\left(\Sigma_{i'}^{(j)} - \Sigma_i^{(j)}\right)\right\|_F + \mathcal{O}\left(k^{1.5}/\alpha^{2.5}\right) \\
&\leqslant \mathcal{O}(\Delta/\alpha)\ .
\end{aligned}$$

There are at most $\mathcal{O}\left(k^k\right)$ ways in which we can partition the set of input points such that each resulting component is not partially clusterable. We run the algorithm in parallel for each one. Then, for the correct iteration, we apply Theorem 72 to get a list $\mathcal{L}$ of size $\exp(\text{poly}_k(1/\alpha, 1/\epsilon))$ such that it contains a hypothesis $\{\hat{w}_i^{(j)}, \hat{\mu}_i^{(j)}, \hat{\Sigma}_i^{(j)}\}_{i\in[k]}$ such that $|\hat{w}_i^{(j)} - w_i^{(j)}| \leqslant \alpha$, $\left\|\hat{\mu}_i^{(j)} - \mu_i^{(j)}\right\|_2 \leqslant \mathcal{O}_k(\epsilon)$ and $\left\|\hat{\Sigma}_i^{(j)} - \Sigma_i^{(j)}\right\|_F \leqslant \mathcal{O}_k(\epsilon)$. Since $(1-1/\Delta)\, I \preceq \Sigma_i^{(j)}$, it then follows from Lemma 3.6.2 that the hypothesis is $\mathcal{O}_k(\epsilon)$-close to the input in total variation distance.

Algorithm 77 is called at most $\mathcal{O}\left(k^k\right)$ times, and along with the robust isotropic transformation, this requires $\text{poly}_k\left(n^{1/\alpha}, 1/\epsilon\right)$. The grid search contributes a multiplicative factor of $(1/\alpha)^k$. The tensor decomposition algorithm and robust hypothesis section $\text{poly}_k(n^{1/\alpha}) \cdot \exp\left(\text{poly}_k(1/\alpha, 1/\epsilon)\right)$. □

3.5 Spectral Separation of Thin Components

In this section, we show how to efficiently separate a thin component, if such a component exists, given sufficiently accurate approximations to the component means and covariances. This is an important step in our overall algorithm and is required to obtain total variation distance guarantees.

Specifically, the main algorithmic result of this section is described in the following lemma:

Lemma 3.5.1. *There is a polynomial-time algorithm with the following properties: Let $\mathcal{M} = \sum_{i=1}^{k} w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians on $\mathcal{R}^d$, and let X be a set of points in $\mathcal{R}^d$ satisfying Condition 3.2.45 with respect to $\mathcal{M}$ for some parameters (γ, t). The algorithm takes input parameters η, δ, satisfying $0 < \delta < \eta < 1/(100k)$, and Y, an ε-corrupted version of X, as well as candidate parameters $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i \leqslant k}$. Then as long as*

1. $\mathbf{Cov}(\mathcal{M}) \succeq I/2$,
2. $\|\mu_i - \hat{\mu}_i\|_2 < \delta$ *and* $\|\Sigma_i - \hat{\Sigma}_i\|_F < \delta$, *for all* $i \in [k]$, *and*
3. *there exists an* $s \in [k]$ *such that* Σ_s *has an eigenvalue* $< \eta$,

the algorithm outputs a partition of Y into $Y_1 \cup Y_2$ such that there is a non-trivial partition of $[k]$ into $Q_1 \cup Q_2$, so that letting $\mathcal{M}_j$, $j \in \{1, 2\}$, be proportional to $\sum_{i \in Q_j} w_i G_i$ and $W_j = \sum_{i \in Q_j} w_i$, then Y_j is an $((O(k^2\gamma) + \tilde{O}(\eta^{1/2k}))/W_j)$-corruption of a set satisfying Condition 3.2.45 with respect to $\mathcal{M}_j$ with parameters $(O(k\gamma/W_j), t)$.

The key component in the proof of Lemma 3.5.1 is the following lemma:

Lemma 3.5.2. *Let $\mathcal{M} = \sum_{i=1}^{k} w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians in $\mathcal{R}^d$ with $\mathbf{Cov}(\mathcal{M}) \succeq I/2$. Suppose that, for some $0 < \delta < 1/(100k)$, we are given $\hat{\mu}_i$ and $\hat{\Sigma}_i$ satisfying $\|\mu_i - \hat{\mu}_i\|_2 < \delta$ and $\|\Sigma_i - \hat{\Sigma}_i\|_F < \delta$, for all $i \in [k]$. Suppose furthermore that for some $\eta > \delta$, there is a Σ_s, $s \in [k]$, with an eigenvalue less than η. There exists a computationally efficient algorithm that takes inputs η, δ, $\hat{\mu}_i$, $\hat{\Sigma}_i$, and computes a function $F : \mathcal{R}^d \to \{0, 1\}$ such that:*

1. *For each $i \in [k]$, $F(G_i)$ returns the same value in $\{0, 1\}$ with probability at least $1 - \tilde{O}_k(\eta^{1/(2k)})$. We define the most likely value of $F(G_i)$ to be this value.*
2. *There exist $i, j \in [k]$ such that the most likely values of $F(G_i)$ and $F(G_j)$ are different.*

Furthermore, $F(x)$ can be chosen to be of the form $f(v \cdot x)$, for some $v \in \mathcal{R}^d$, and $f : \mathcal{R} \to \{0,1\}$ is an $O(k)$-piecewise constant function.

Given Lemma 3.5.2, it is easy to finish the proof of Lemma 3.5.1.

Proof of Lemma 3.5.1. We simply take the candidate parameters, obtain F from Lemma 3.5.2, and partition $Y = Y_1 \cup Y_2$, so that F is constant on both Y_1 and Y_2. We let Q_j be the set of i so that $F(G_i)$ returns the value $j-1$ with large probability. Letting the partition of X for Condition 3.2.45 be $X = X_1 \cup \ldots \cup X_k$, we let $X^j = \bigcup_{i \in Q_j} X_i$. Lemma 3.2.48 shows that the X^j satisfy the appropriate conditions for $\mathcal{M}_j$. It remains to prove that Y_j equals X^j with a sufficiently small rate of corruptions. The fraction of points misclassified by F equals ϵ (the fraction of outliers in the sample Y) plus the misclassification error of F. We note that given the form of F and the fact that the uncorrupted samples in Y satisfy Condition 3.2.45, the fraction of misclassified samples from each component i is at most the probability that a random sample from G_i gets misclassified (at most $\tilde{O}_k(\eta^{1/(2k)})$ by Lemma 3.5.2) plus $O(k\gamma)$. Summing this over components, gives Lemma 3.5.1. □

Let us now describe the algorithm to prove Lemma 3.5.2 (and evaluate F), which is given in pseudocode below (Algorithm 79).

Algorithm 79 (Algorithm for Spectrally Separating Thin Components)**.**

Input: *Estimated parameters $\left\{\hat{\mu}_i, \hat{\Sigma}_i\right\}_{i \leqslant k}$, parameters η, δ.*

Output: *A function $F : \mathcal{R}^d \to \{0,1\}$.*

Operation:

1. *Find a unit-norm direction v such that there exists $s \in [k]$, $v^T \hat{\Sigma}_s v < 2\eta$.*
2. *Compute $(v^T \hat{\Sigma}_i v)$ for all $i \in [k]$.*
   (a) *If there exists $j \in [k]$ such that $(v^T \hat{\Sigma}_j v) > \sqrt{\eta}$, find a t such that $\sqrt{\eta} > t > 2\eta$ and there is no $j \in [k]$ with $t < v^T \hat{\Sigma}_j v < t\,\Omega(\eta^{-1/(2k)})$. Set $F(x) = 1$ if there is an i such that $|v \cdot (x - \hat{\mu}_i)| < \sqrt{t}\log(1/\eta)$ and 0 otherwise.*
   (b) *Otherwise, compute $v \cdot \hat{\mu}_i$ for all $i \in [k]$. Find a t between the minimum and the maximum of $v \cdot \hat{\mu}_i$ such that there is no $v \cdot \hat{\mu}_i$ within $1/(20k)$ of t. Set $F(x) = 1$ if $v \cdot x > t$ and 0 otherwise.*

Proof of Lemma 3.5.2. Let v be a unit vector and $s \in [k]$ such that $v^T \hat{\Sigma}_s v < 2\eta$. By assumption, we have that $\mathbf{Var} v \cdot \mathcal{M} \geqslant 1/2$. Furthermore,

$$\mathbf{Var} v \cdot \mathcal{M} = \sum_i w_i (v^T \Sigma_i v) + \sum_i w_i (v \cdot (\mu_i - \mu))^2 \leqslant \sum_i w_i (v^T \Sigma_i v) + \sum w_i (v \cdot (\mu_i - \mu_s))^2 \ ,$$

where μ is the mean of $\mathcal{M}$. This means that either there exists $j \in [k]$ such that $(v^T \Sigma_j v) > 1/4$, or there exists $j \in [k]$ such that $|v \cdot (\mu_j - \mu_s)| > 1/4$. Since we have approximations of these quantities to order δ, we have that there is $j \in [k]$ such that $(v^T \hat{\Sigma}_j v) > 1/10$ or that there is $j \in [k]$ with $|v \cdot (\hat{\mu}_j - \hat{\mu}_s)| > 1/10$.

We first consider the case that there is a $j \in [k]$ such that $(v^T \hat{\Sigma}_j v) > \sqrt{\eta}$. Since there is a $j \in [k]$ with $(v^T \hat{\Sigma}_j v) > \sqrt{\eta}$ and another $s \in [k]$ with $(v^T \hat{\Sigma}_s v) < 2\eta$, there must be some $\sqrt{\eta} > t > 2\eta$ such that there is no $j \in [k]$ with $t < v^T \hat{\Sigma}_j v < t\,\Omega(\eta^{-1/(2k)})$. Otherwise, there must be at least one $\hat{\Sigma}_i$ in each $2\eta \leqslant \Omega(\eta^{-1/(2k)})^i \leqslant \sqrt{\eta}$, where we need more than k components.

For a given x, we define $F(x)$ to be 1 if there exists i such that $|v \cdot (x - \hat{\mu}_i)| < \sqrt{t}\log(1/\eta)$, and $F(x) = 0$ otherwise.

To show that this works, we note that for all $i \in [k]$, if $v^T \hat{\Sigma}_i v \leqslant t$, then $\mathbf{Var} v \cdot G_i \leqslant t + \delta$, and since $|v \cdot (\mu_i - \hat{\mu}_i)| < \delta$, by the Gaussian tail bound, we have that

$$\mathbf{Pr}_{x \sim G_i}\left(|x - \mu_i| \geqslant (\sqrt{t}\log(1/\eta) - \delta)\right) \leqslant \exp\left(-\frac{(\sqrt{t}\log(1/\eta) - \delta)^2}{2(t+\delta)}\right) = O(\eta) \ .$$

Thus, all but an $O(\eta)$-fraction of the samples of G_i have $F(x) = 1$.

On the other hand, for components i with $v^T \hat{\Sigma}_i v \gg t\eta^{-1/(2k)}$, we have that $\mathbf{Var} v \cdot G_i \gg t\eta^{-1/(2k)}$. Then, the density of G_i is at most $1/\sqrt{2\pi t\eta^{-1/(2k)}}$. So, the probability that a sample from $v \cdot G_i$ lies in any interval of length $2\sqrt{t}\log(1/\eta)$ is at most

$$\frac{1}{\sqrt{2\pi t\eta^{-1/(2k)}}} 2\sqrt{t}\log(1/\eta) = \tilde{O}(\eta^{1/(4k)}) \ .$$

Since there are k such intervals, the probability that $F(x)$ is 1 when x is drawn from G_i is at most $\tilde{O}_k(\eta^{1/(4k)})$. This completes our proof of point (1), and point (2) follows from the fact that we know of component G_j in one class and G_s in the other class.

We next consider the case where $(v^T \hat{\Sigma}_j v) \leqslant \sqrt{\eta}$ for all $j \in [k]$, and where $|v \cdot (\hat{\mu}_j - \hat{\mu}_s)| > 1/10$ for some $j \in [k]$. Then we can find some t between $v \cdot \hat{\mu}_j$ and $v \cdot \hat{\mu}_s$ such that no $v \cdot \hat{\mu}_i$

is within $1/(20k)$ of t. In this case, we define $F(x)$ be 1 if $v \cdot x > t$ and 0 otherwise. To show part (1), first consider $i \in [k]$ such that $v \cdot \hat{\mu}_i < t - 1/(20k)$. Then we have that $v \cdot \mu_i < t - 1/(30k)$. Furthermore, $\mathbf{Var} v \cdot G_j \leqslant \delta + \sqrt{\eta}$. Therefore, the probability that $v \cdot G_i > t$ is at most $\exp(-\Omega_k((\delta + \sqrt{\eta})^{-2}))$, which is sufficient.

A similar argument holds in the other direction for $i \in [k]$ such that $v \cdot \hat{\mu}_i > t + 1/(20k)$, and statement (2) holds because we know that there are both kinds of components. This completes the proof. □

3.6 Robust Proper Learning: Proof of Theorem 67

In this section, we show how to combine the partial clustering, tensor decomposition, and recursive clustering algorithms to establish our main result. The main theorem we prove is as follows:

Theorem 80 (Robustly Learning k-Mixtures of Arbitrary Gaussians)**.** *Given $0 < \epsilon < 1/k^{k^{O(k^2)}}$ and a multiset $Y = \{y_1, y_2, \ldots, y_n\}$ of n i.i.d. samples from a distribution F such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown k-mixture of Gaussians $\mathcal{M} = \sum_{i \leqslant k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, where $n \geqslant n_0 = d^{O(k)}/\mathrm{poly}(\varepsilon)$, Algorithm 81 runs in time $n^{O(1)} \exp(O(k)/\varepsilon^2)$ and with probability at least 0.99 outputs a hypothesis k-mixture of Gaussians $\widehat{\mathcal{M}} = \sum_{i \leqslant k} \hat{w}_i \mathcal{N}\left(\hat{\mu}_i, \hat{\Sigma}_i\right)$ such that $d_{TV}\left(\mathcal{M}, \widehat{\mathcal{M}}\right) = \mathcal{O}(\varepsilon^{c_k})$, with $c_k = 1/(100^k C^{(k+1)!} k! \mathrm{sf}(k+1))$, where $C > 0$ is a universal constant and $\mathrm{sf}(k) = \Pi_{i \in [k]}(k-i)!$ is the super-factorial function.*

As an immediate corollary, we obtain the following:

Corollary 3.6.1 (Robustly Learning k-Mixtures of Gaussians in Polynomial Time)**.** *Given $0 < \epsilon < 1/\exp\left(k^{k^{O(k^2)}}\right)$, and a multiset $Y = \{y_1, y_2, \ldots, y_n\}$ of n i.i.d. samples from a distribution F such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown k-mixture of Gaussians $\mathcal{M} = \sum_{i \leqslant k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, where $n \geqslant n_0 = d^{O(k)} \log^{O(1)}(1/\epsilon)$, there exists an algorithm that runs in time $\mathrm{poly}_k(n, 1/\varepsilon)$ and with probability at least 0.99 outputs a k-mixture of Gaussians $\widehat{\mathcal{M}} = \sum_{i \leqslant k} \hat{w}_i \mathcal{N}\left(\hat{\mu}_i, \hat{\Sigma}_i\right)$ such that $d_{\mathsf{TV}}\left(\mathcal{M}, \widehat{\mathcal{M}}\right) = \mathcal{O}\left((1/\log(1/\epsilon))^{1/\left(k^{O(k^2)}\right)}\right)$.*

The corollary follows by running Algorithm 81 with $\epsilon \leftarrow \sqrt{1/\log(1/\epsilon)}$ and applying Theorem 80.

The algorithm establishing Theorem 80 is given in pseudocode below. Algorithm 82 takes as input a corrupted sample from a k-mixture of Gaussians and outputs a set of k mixing weights, means, and covariances, such that the resulting mixture is close to the input mixture in total variation distance with non-negligible probability. Algorithm 81 simply runs Algorithm 82 many times to create a small list of candidate hypotheses (consisting of mixing weights, means, and covariances), and finally runs a robust tournament to outputs a winner. This boosts the probability of success to at least 0.99.

Algorithm 81 (Algorithm for Robustly Learning Arbitrary GMMs)**.**

Input: *An outlier parameter $\epsilon > 0$ and a component-number parameter $k \in \mathbb{N}$. An ϵ-corrupted sample $Y = \{y_1, y_2, \ldots, y_n\}$ from a k-mixture of Gaussians $\mathcal{M} = \sum_{i\in[k]} w_i \mathcal{N}(\mu_i, \Sigma_i)$.*

Parameters: *Let $c_k = 1/(100^k C^{(k+1)!} \mathrm{sf}(k+1)k!)$ be a scalar function of k, where sf$(k) = \prod_{i\in[k]}(k-1)!$ and C is a sufficiently large constant.*

Output: *A set of parameters $\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]}$, such that with probability at least 0.99 the mixture $\widehat{\mathcal{M}} = \sum_{i\in[k]} \hat{w}_i \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i)$ is $\mathcal{O}(\epsilon^{c_k})$-close in total variation distance to $\mathcal{M}$.*

Operation:

1. *Let $\mathcal{L} = \{\phi\}$ be an empty list. Repeat the following $\exp(O(k)/\varepsilon^2)$ times :*
   (a) *Run Algorithm 82 with input Y, fraction of outliers ε, and number of components k. Let the resulting output be a set of k mixing weights, means and covariances, denoted by $\left\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\right\}_{i\in[k]}$. Add $\left\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\right\}_{i\in[k]}$ to $\mathcal{L}$.*
2. *Run the robust tournament from Fact 3.2.50 over all the hypotheses in $\mathcal{L}$. Output the winning hypothesis, denoted by $\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]}$.*

Algorithm 82 (Cluster or List-Decode)**.**

Input: *An outlier parameter $0 < \epsilon < 1$ and a component-number parameter $k \in \mathbb{N}$. An ϵ-corrupted version $Y = \{y_1, y_2, \ldots, y_n\}$ of X, where X is a set of n samples from a k-mixture of Gaussians $\mathcal{M} = \sum_{i\in[k]} w_i \mathcal{N}(\mu_i, \Sigma_i)$ such that X satisfies Condition 3.2.45 with respect to $\mathcal{M}$ with parameters $(\varepsilon d^{-8k} k^{-C'k}, 8k+48)$, where $C' > 0$ is a sufficiently large constant.*

Parameters: *Let $c_k = 1/(100^k C^{(k+1)!} \mathrm{sf}(k+1)k!)$ be a scalar function of k, where sf$(k) = \prod_{i\in[k]}(k-1)!$ and C is a sufficiently large constant.*

Output: *A set of parameters* $\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]}$ *such that with probability at least* $\exp(-O(k)/\varepsilon^2)$, $d_{TV}\left(\sum_{i\in[k]} \hat{w}_i \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i), \mathcal{M}\right) \leqslant \mathcal{O}(\epsilon^{c_k})$.

Operation:

1. ***Treat Light Component as Noise****: If* $k = 0$*, ABORT. With probability* $1/2$*, run Algorithm 82 on samples* Y*, with fraction of outliers* $\epsilon + \epsilon^{1/(10C^{k+1}(k+1)!)}$ *and number of components* $k-1$*. Return the resulting set of estimated parameters,* $\{(\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k-1]}$*, appended with* $(0, 0, I)$*. Else, do the following:*
   // We guess whether the event that the minimum mixing weight α *is at least* $\epsilon^{1/(10C^{k+1}(k+1)!)}$
   // holds. If it does not, we proceed with the algorithm. Else, we treat the smallest weight
   // component as noise and recurse with $k-1$ *components.*

2. ***Robust Isotropic Transformation****: With probability* 0.5*, run the algorithm corresponding to Lemma 3.6.4 on the samples* Y*, and let* $\hat{\mu}, \hat{\Sigma}$ *be the robust estimates of the mean and covariance. If* $k = 1$*, return* $\left(\hat{w} = 1, \hat{\mu}, \hat{\Sigma}\right)$*. Else, compute* $\hat{U}\hat{\Lambda}\hat{U}^\top$*, the eigendecomposition of* Σ*, and for all* $i \in [n]$*, apply the affine transformation* $y_i \to \hat{U}^\top \hat{\Sigma}^{\dagger/2}(y_i - \hat{\mu})$.
   // The resulting estimates $\hat{\mu}, \hat{\Sigma}$ *satisfy Lemma 3.6.4, and the uncorrupted samples are*
   // effectively drawn from a nearly isotropic k*-mixture.*

3. *With probability* $1/2$*, run either* (a) *or* (b) *in the following:*

   (a) ***Partial Clustering via SoS****: Run Algorithm 77 with outlier parameter* ϵ *and accuracy parameter* $\epsilon^{1/(5C^{k+1}(k+1)!)}$*. Let* Y_1*,* Y_2 *be the partition returned. Guess the number of components in* Y_1 *to be some* $k_1 \in [k-1]$ *uniformly at random. Run Algorithm 82 with input* Y_1*, fraction of outliers* $\epsilon^{1/(10c^{k+1}(k+1)!)}$*, and number of components* k_1*, and let* $\left\{(\hat{w}_i^{(1)}, \hat{\mu}_i^{(1)}, \hat{\Sigma}_i^{(1)})\right\}_{i\in[k_1]}$ *be the resulting output. Similarly, run Algorithm 82 with input* Y_2*, fraction of outliers* $\epsilon^{1/(10c^{k+1}(k+1)!)}$*, and number of components* $k-k_1$*, and let* $\left\{(\hat{w}_i^{(2)}, \hat{\mu}_i^{(2)}, \hat{\Sigma}_i^{(2)})\right\}_{i\in[k-k_1]}$ *be the resulting output. Output the set* $\left\{(\hat{w}_i^{(1)}|Y_1|/|Y|, \hat{\mu}_i^{(1)}, \hat{\Sigma}_i^{(1)})\right\}_{i\in[k_1]} \cup \left\{(\hat{w}_i^{(2)}|Y_2|/|Y|, \mu_i^{(2)}, \hat{\Sigma}_i^{(2)})\right\}_{i\in[k-k_1]}$.
   // When the mixture is covariance separated, the preconditions of Theorem

76 are
// satisfied (see Lemma 3.6.5). The partition is non-trivial, and the fraction of outliers
// increases from $\epsilon \to \epsilon^{1/(10c^{k+1}(k+1)!)}$.

(b) ***List-Decoding via Tensor Decomposition****: Run Algorithm 73 and let* L *be the resulting list of hypotheses such that each hypothesis is a set of parameters* $\{(\hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]}$. *Let* $\tau = \Theta\left(\epsilon^{1/(40C^{k+1}(k+1)!)}\right)$ *be an eigenvalue threshold. Select a hypothesis,* $\{(\hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]} \in L$ *uniformly at random.*
// Conditioned on not being covariance separated, we satisfy the preconditions of
// Theorem 72 (see Lemma 3.6.6). The output is a list that contains $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$
// such that for all $i \in [k]$, $\|\hat{\mu}_i - \mu_i\|_2 = \mathcal{O}\left(\epsilon^{1/(20C^{k+1}(k+1)!)}\right)$ *and*
// $\left\|\hat{\Sigma}_i - \Sigma_i\right\|_F = \mathcal{O}\left(\epsilon^{1/(20C^{k+1}(k+1)!)}\right)$.

i. ***Large Eigenvalues****: If for all* $i \in [k]$, $\hat{\Sigma}_i \succeq \tau I$, *sample* $\hat{w}_i$ *from* $[0, 1]$ *uniformly at random such that* $\sum_i \hat{w}_i = (1 \pm k\epsilon)$. *Return*

$$\left\{\left(\hat{w}_i, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\mu}_i + \hat{\mu}, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\Sigma}_i\hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\right)\right\}_{i\in[k]}.$$

// If all estimated covariances have all eigenvalues larger than τ, *the recursion*
// bottoms out and the hypothesis is returned.

ii. ***Spectral Separation of Thin Components****: Else,* $\exists v, i$ *s.t.* $v^\top\hat{\Sigma}_i v \leqslant \tau$. *Run the algorithm corresponding to Lemma 3.5.1 with input* Y, *parameter estimates* $\{(\hat{\mu}_i, \hat{\Sigma}_i)\}_{i\in[k]}$ *and threshold* τ. *Let* Y_1 *and* Y_2 *be the resulting partition.*
// Use small eigenvalue directions to partition the points.

A. *If* $\min(|Y_1|, |Y_2|) < \epsilon^{1/(400kC^{k+1}(k+1)!)}n$, *run Algorithm 82 with input* Y, *fraction of outliers* $2\epsilon^{1/(400kC^{k+1}(k+1)!)}$ *and number of components being* $k - 1$, *and let* $\left\{(\hat{w}_i^{(1)}, \hat{\mu}_i^{(1)}, \hat{\Sigma}_i^{(1)})\right\}_{i\in[k_1]}$ *be the resulting output.*

Output the resulting hypothesis

$$\left\{(\hat{w}_i, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\mu}_i + \hat{\mu}, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\Sigma}_i\hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top)\right\}_{i\in[k-1]} \cup (0, 0, I).$$

B. Else, select $k_1 \in [k-1]$ *uniformly at random. Run Algorithm 82 with input* Y_1*, fraction of outliers* $\epsilon^{1/(100kC^{k+1}(k+1)!)}$ *and number of components being* k_1*. Similarly, run Algorithm 82 with input* Y_2*, fraction of outliers* $\epsilon^{1/(100kC^{k+1}(k+1)!)}$ *and number of components* $k - k_1$*, and let* $\left\{(\hat{w}_i^{(2)}, \hat{\mu}_i^{(2)}, \hat{\Sigma}_i^{(2)})\right\}_{i\in[k-k_1]}$ *be the resulting output. Output the set*

$$\left\{(\hat{w}_i^{(1)}|Y_1|/|Y|, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\mu}_i^{(1)} + \hat{\mu}, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\Sigma}_i^{(1)}\hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top)\right\}_{i\in[k_1]} \cup$$
$$\left\{(\hat{w}_i^{(2)}|Y_2|/|Y|, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\mu_i^{(2)} + \hat{\mu}, \hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top\hat{\Sigma}_i^{(2)}\hat{U}\hat{\Lambda}^{1/2}\hat{U}^\top)\right\}_{i\in[k-k_1]}.$$

3.6.1 Analysis of Algorithm 81

To prove Theorem 80, we will require the following intermediate results. We defer some proofs in this subsection to Appendix 3.10.

We use the following lemma to relate the Frobenius distance of covariances to the total variation distance between two Gaussians, when the eigenvalues of the covariances are bounded below.

Lemma 3.6.2 (Frobenius Distance to TV Distance)**.** *Suppose* $\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)$ *are Gaussians with* $\|\mu_1 - \mu_2\|_2 \leqslant \delta$ *and* $\|\Sigma_1 - \Sigma_2\|_F \leqslant \delta$*. If the eigenvalues of* Σ_1 *and* Σ_2 *are at least* $\lambda > 0$*, then* $d_{\mathsf{TV}}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = O(\delta/\lambda)$.

We start by showing that when Condition 3.2.45 holds, the uniform distribution on a $(1-\epsilon)$-fraction of the points is certifiably hypercontractive.

Lemma 3.6.3 (Component Moments to Mixture Moments)**.** *Let* $\mathcal{M} = \sum_{i\in[k]} w_i\mathcal{N}(\mu_i, \Sigma_i)$ *be a* k*-mixture with mean* μ *and covariance* Σ *such that* $w_i \geqslant \alpha$*, for some* $0 < \alpha < 1$*, and for all* $i, j \in [k]$*,* $\left\|\Sigma^{\dagger/2}(\Sigma_i - \Sigma_j)\Sigma^{\dagger/2}\right\|_F \leqslant 1/\sqrt{\alpha}$*. Let* X *be a multiset of* n *samples satisfying Condition 3.2.45 with respect to* $\mathcal{M}$ *with parameters* (γ, t)*, for* $0 < \gamma < (dk/\alpha)^{-ct}$*, for a sufficiently large constant* c*, and* $t \in \mathbb{N}$*. Let* $\mathcal{D}$ *be the uniform distribution over* X*. Then,* $\mathcal{D}$ *is* $2t$*-certifiably* (c/α)*-hypercontractive and for* $d \times d$*-matrix-valued indeterminate* Q*,* $\left\{\frac{Q}{2}\right\}\left\{\mathbf{E}_{\mathcal{M}}\left(x^\top Qx - \mathbf{E}_{\mathcal{M}}x^\top Qx\right)^2 \leqslant \mathcal{O}(1/\alpha)\left\|\Sigma^{1/2}Q\Sigma^{1/2}\right\|_F^2\right\}$.

Next, we show how to robustly estimate the mean and covariance of an ϵ-corrupted set of samples satisfying Condition 3.2.45 when the mixture is not partially clusterable, and make the inliers nearly isotropic.

Lemma 3.6.4 (Robust Isotropic Transformation)**.** *Given $0 < \epsilon < 1$, and $k \in \mathbb{N}$, let $\alpha = \epsilon^{1/(10C^{k+1}(k+1)!)}$. Let $\mathcal{M} = \sum_{i=1}^{k} w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with $w_i \geqslant \alpha$ for all $i \in [k]$, and let μ and Σ be the mean and covariance of $\mathcal{M}$ such that $r = \mathrm{rank}(\Sigma)$ and for all $i, j \in [k]$, $\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F \leqslant 1/\sqrt{\alpha}$. Let X be a set of points satisfying Condition 3.2.45 with respect to $\mathcal{M}$ for some parameters (γ, t). Given a set Y, an ϵ-corrupted version of X, of size $n \geqslant n_0 = d^{O(1)}$, there exists an algorithm that takes Y as input and in time $n^{\mathcal{O}(1)}$ outputs estimators $\hat{\mu}$ and $\hat{\Sigma}$ such that $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^\top$ is the eigenvalue decomposition, where $\hat{U} \in \mathcal{R}^{n \times r}$ has orthonormal columns and $\Lambda \in \mathcal{R}^{r \times r}$ is a diagonal matrix. Further, we can obtain n samples Y' by applying the affine transformation $y_i \to \hat{U}^\top \hat{\Sigma}^{\dagger/2}\left(y_i - \hat{\mu}\right)$ to each sample, such that a $(1-\epsilon)$-fraction have mean μ' and covariance Σ' satisfying*

1. $\|\mu'\|_2 \leqslant \mathcal{O}\Big(\Big(1 + \frac{\sqrt{\epsilon}k}{\alpha}\Big)\sqrt{\epsilon/\alpha}\Big)$,
2. $\Big(\frac{1}{1+(k\sqrt{\epsilon}/\alpha)}\Big) I_r \preceq \Sigma' \preceq \Big(\frac{1}{1-(k\sqrt{\epsilon}/\alpha)}\Big) I_r$,
3. $\|\Sigma' - I_r\|_F \leqslant \mathcal{O}(\sqrt{\epsilon}k/\alpha)$,

where I_r is the r-dimensional Identity matrix, and the remaining points are arbitrary. Let X' be the set obtained by $\hat{U}^\top \hat{\Sigma}^{\dagger/2}\left(x_i - \hat{\mu}\right)$. Then, X' satisfies Condition 3.2.45 with respect to $\sum_{i=1}^{k} w_i \mathcal{N}\left(\hat{U}^\top \hat{\Sigma}^{\dagger/2}(\mu_i - \hat{\mu}), \hat{U}^\top \hat{\Sigma}^{\dagger/2}\Sigma_i \hat{\Sigma}^{\dagger/2}\hat{U}\right)$ and parameters (γ, t), and Y' is an ϵ-corruption of X'.

Proof. For any $t' \in \mathbb{N}$, it follows from Corollary 3.2.34 that $\mathcal{M}$ has $2t'$-certifiably $(4/\alpha)$-hypercontractive degree-2 polynomials, since $w_i \geqslant \alpha$ for all i. Next, Lemma 3.6.3 implies that the uniform distribution over X also has $2t'$-certifiably $(8/\alpha)$-hypercontractive degree-2 polynomials and for $d \times d$-matrix-valued indeterminate Q,

$$\Big|\frac{Q}{2}\Big\{\mathbf{E}_{\mathcal{M}}\left(x^\top Q x - \mathbf{E}_{\mathcal{M}} x^\top Q x\right)^2 \leqslant \mathcal{O}(1/\alpha)\left\|\Sigma^{1/2} Q \Sigma^{1/2}\right\|_F^2\Big\}.$$

Then, it follows from Fact 3.2.37 that if $\frac{16}{\alpha}t'\epsilon^{1-4/t'} \ll 1$, there exists an algorithm that runs in time $n^{O(t')}$ and outputs estimates $\hat{\mu}$ and $\hat{\Sigma}$ satisfying:

1. $\left\|\Sigma^{\dagger/2}(\mu - \hat{\mu})\right\|_2 \leqslant O(t'/\alpha)^{1/2}\epsilon^{1-1/t'}$,
2. $\left(1 - (k/\alpha)\epsilon^{1-2/t'}\right)\Sigma \preceq \hat{\Sigma} \preceq \left(1 + (k/\alpha)\epsilon^{1-2/t'}\right)\Sigma$ and,

3. $\left\|\Sigma^{\dagger/2}\left(\hat{\Sigma}-\Sigma\right)\Sigma^{\dagger/2}\right\|_F \leqslant (t'/\alpha)O(\epsilon^{1-1/t'})$.

Setting $t'=2$, compute $\hat{\Sigma}=\hat{U}\hat{\Lambda}\hat{U}^\top$, the eigendecomposition of $\hat{\Sigma}$, such that $\hat{U}\in\mathcal{R}^{n\times r}$ has orthonormal columns, where $r\leqslant d$ is the rank of $\hat{\Sigma}$ and $\hat{\Lambda}\in\mathcal{R}^{r\times r}$ is a diagonal matrix. Similarly, let $\Sigma=U\Lambda U^\top$ be the eigendecomposition of Σ. We apply the affine transformation $y_i\to\hat{U}^\top\hat{\Sigma}^{\dagger/2}\left(y_i-\hat{\mu}\right)$ to each sample and thus we can assume throughout the rest of our argument that we have access to ϵ-corrupted samples from a k-mixture of Gaussians with mean $\mu'=\hat{U}^\top\hat{\Sigma}^{\dagger/2}(\mu-\hat{\mu})$ and covariance $\Sigma'=\hat{U}^\top\hat{\Sigma}^{\dagger/2}\Sigma\hat{\Sigma}^{\dagger/2}\hat{U}$. Then, we have that

$$\begin{aligned}\|\mu'\|_2=\left\|\hat{U}^\top\hat{\Sigma}^{\dagger/2}(\mu-\hat{\mu})\right\|_2 &\leqslant \left\|\hat{U}^\top\right\|_{\text{op}}\left\|\hat{\Sigma}^{\dagger/2}(\mu-\hat{\mu})\right\|_2\\ &\leqslant \mathcal{O}\left(\left(1+\frac{\sqrt{\epsilon}k}{\alpha}\right)\sqrt{\epsilon/\alpha}\right)\ ,\end{aligned}$$

where the last inequality follows from (1) and (2). It also follows from (2) that

$$\left(\frac{1}{1+(k\sqrt{\epsilon}/\alpha)}\right)\hat{\Sigma}\preceq\Sigma\preceq\left(\frac{1}{1-(k\sqrt{\epsilon}/\alpha)}\right)\hat{\Sigma}\ . \tag{3.46}$$

Multiplying out (3.46) with $\hat{U}^\top\hat{\Sigma}^{\dagger/2}$ on the left and $\hat{\Sigma}^{\dagger/2}\hat{U}$ on the right, we have

$$\left(\frac{1}{1+(k\sqrt{\epsilon}/\alpha)}\right)\hat{U}^\top\hat{\Sigma}^{\dagger/2}\hat{\Sigma}\hat{\Sigma}^{\dagger/2}\hat{U}\preceq\Sigma'\preceq\left(\frac{1}{1-(k\sqrt{\epsilon}/\alpha)}\right)\hat{U}^\top\hat{\Sigma}^{\dagger/2}\hat{\Sigma}\hat{\Sigma}^{\dagger/2}\hat{U}\ .$$

Observe that (2) implies that the rank of $\hat{\Sigma}$ and Σ is the same, and thus $\hat{U}^\top\hat{\Sigma}^{\dagger/2}\hat{\Sigma}\hat{\Sigma}^{\dagger/2}\hat{U}=I_r$, where I_r is the r-dimensional Identity matrix. Finally, we have that

$$\begin{aligned}\|\Sigma'-I_r\|_F=\left\|\hat{U}^\top\hat{\Sigma}^{\dagger/2}\Sigma\hat{\Sigma}^{\dagger/2}\hat{U}-\hat{U}^\top\hat{\Sigma}^{\dagger/2}\hat{\Sigma}\hat{\Sigma}^{\dagger/2}\hat{U}\right\|_F &\leqslant \left\|\hat{U}\hat{\Lambda}^{-1/2}\hat{U}^\top\left(\Sigma-\hat{\Sigma}\right)\hat{U}\hat{\Lambda}^{-1/2}\hat{U}^\top\right\|_F\\ &= \left\|\hat{U}\hat{\Lambda}^{-1/2}\Lambda^{1/2}\Lambda^{-1/2}\hat{U}^\top\left(\Sigma-\hat{\Sigma}\right)\hat{U}\Lambda^{-1/2}\Lambda^{1/2}\hat{\Lambda}^{-1/2}\hat{U}^\top\right\|\\ &\leqslant \left\|\hat{\Lambda}^{-1/2}\Lambda^{1/2}\right\|_{\text{op}}^2\left\|\Sigma^{\dagger/2}\left(\hat{\Sigma}-\Sigma\right)\Sigma^{\dagger/2}\right\|_F\\ &\leqslant \mathcal{O}\left(\sqrt{\epsilon}k/\alpha\right)\ ,\end{aligned}$$

where we use that $\hat{\Lambda}^{-1/2}=\hat{\Lambda}^{-1/2}\Lambda^{1/2}\Lambda^{-1/2}$, the sub-multiplicative property of the Frobenius norm, the column span U and $\hat{U}$ is identical (see (2)), and the Frobenius recovery guarantee in (3).

Finally, it follows from Lemma 3.2.46 that Condition 3.2.45 is affine invariant and is thus preserved under $x_i\to\hat{U}^\top\hat{\Sigma}^{-1/2}\left(x_i-\hat{\mu}\right)$, for $i\in[n]$, with parameters (γ,t). □

The above robust isotropic transformation lemma allows us to obtain a covariance that is close to the identity matrix in a full-dimensional subspace (potentially smaller than the input dimension). Therefore, we will subsequently drop the subscript for the dimension, wherever it is clear from the context.

Next, we show that whenever the minimum mixing weight is sufficiently larger than the fraction of outliers, and a pair of components is covariance separated, we can partially cluster the samples.

Lemma 3.6.5 (Non-negligible Weight and Covariance Separation)**.** *Given $0 < \epsilon < 1/k^{k^{O(k^2)}}$ and $k \in \mathbb{N}$, let $\alpha = \epsilon^{1/(10C^{k+1}(k+1)!)}$. Let $\mathcal{M} = \sum_{i=1}^{k} w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with mixture covariance Σ such that $w_i \geqslant \alpha$ for all $i \in [k]$ and there exist $i, j \in [k]$ such that $\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F > 1/\sqrt{\alpha}$. Further, let X be a set of points satisfying Condition 3.2.45 with respect to $\mathcal{M}$ for some parameters $\gamma \leqslant \varepsilon d^{-8k} k^{-Ck}$, for a sufficiently large constant C, and $t \geqslant 8k$. Let Y be an ϵ-corrupted version of X of size $n \geqslant n_0 = (dk)^{\Omega(1)}/\epsilon$, Algorithm 77 partitions Y into Y_1, Y_2 in time $n^{O(1)}$ such that with probability at least $\alpha^{k\log(k/\alpha)}$ there is a non-trivial partition of $[k]$ into $Q_1 \cup Q_2$ so that letting $\mathcal{M}_j$ be a distribution proportional to $\sum_{i\in Q_j} w_i G_i$ and $W_j = \sum_{i\in Q_j} w_i$, then Y_j is an $\mathcal{O}\left(\epsilon^{1/(10C^{k+1}(k+1)!)}\right)$-corrupted version of $\bigcup_{i\in Q_j} X_i$ satisfying Condition 3.2.45 with respect to $\mathcal{M}$ with parameters $\left(\mathcal{O}(k\gamma/W_j), t\right)$.*

Proof. We run Algorithm 77 with sample set Y, number of components k, the fraction of outliers ϵ and the accuracy parameter β. Since X satisfies Condition 3.2.45, we can set $t' \geqslant 24$, $\beta = \alpha^{t'/4-4} k^{t'} (t')^{2t'} \leqslant \alpha$ in Theorem 76. Then, by assumption, there exist i, j such that

$$\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F > \frac{1}{\sqrt{\alpha}} = \Omega\left(\frac{k^2 (t')^4}{(\beta\alpha^4)^{2/t'}}\right) .$$

We observe that we also satisfy the other preconditions for Theorem 76, since $n \geqslant (dk/)^{\Omega(1)}/\epsilon$.

Then, Theorem 76 implies that with probability at least $\alpha^{k\log(k/\alpha)}$, the set Y is partitioned in two sets Y_1 and Y_2 such that there is a non-trivial partition of $[k]$ into $Q_1 \cup Q_2$ so that letting $\mathcal{M}_j$ be a distribution proportional to $\sum_{i\in Q_j} w_i G_i$ and $W_j = \sum_{i\in Q_j} w_i$, then Y_j is an $\mathcal{O}\left(\epsilon^{1/(10C^{k+1}(k+1)!)}\right)$-corrupted version of $\bigcup_{i\in Q_j} X_i$. By Lemma 3.2.48, $\bigcup_{i\in Q_j} X_i$ satisfies Condition 3.2.45 with respect to $\mathcal{M}$ with parameters $\left(\mathcal{O}(k\gamma/W_j), t\right)$. □

When the mixture is not covariance separated and nearly isotropic, we can obtain a small list of hypotheses such that one of them is close to the true parameters, via tensor decomposition.

Lemma 3.6.6 (Mixture is List-decodable). *Given $0 < \epsilon < 1/k^{k^{O(k^2)}}$ let $\alpha = \epsilon^{1/(10C^{k+1}(k+1)!)}$. Let $\mathcal{M} = \sum_{i=1}^{k} w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with mixture mean μ and mixture covariance Σ, such that $\|\mu\|_2 \leqslant \mathcal{O}\Big(\sqrt{\epsilon/\alpha}\Big)$, $\|\Sigma - I\|_F \leqslant \mathcal{O}(\sqrt{\varepsilon}/\alpha)$, $w_i \geqslant \alpha$ for all $i \in [k]$, and $\|\Sigma_i - \Sigma_j\|_F \leqslant 1/\sqrt{\alpha}$ for any pair of components, and let X be a set of points satisfying Condition 3.2.45 with respect to $\mathcal{M}$ for some parameters $\gamma = \varepsilon d^{-8k} k^{-Ck}$, for a sufficiently large constant C, and $t = 8k$. Let Y be an ϵ-corrupted version of X of size n, Algorithm 73 outputs a list L of hypotheses of size $\exp(1/\varepsilon^2)$ in time poly$(|L|, n)$ such that if we choose a hypothesis $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$ uniformly at random, $\|\mu_i - \hat{\mu}_i\|_2 \leqslant \mathcal{O}\Big(\epsilon^{1/(20C^{k+1}(k+1)!)}\Big)$ and $\Big\|\Sigma_i - \hat{\Sigma}_i\Big\|_F \leqslant \mathcal{O}\Big(\epsilon^{1/(20C^{k+1}(k+1)!)}\Big)$ for all i with probability at least $\exp(-1/\varepsilon^2)$.*

Proof. Recall we run Algorithm 73 on the samples Y, the number of clusters k, the fraction of outliers ε and the minimum weight $\alpha = \epsilon^{1/(10C^{k+1}(k+1)!)}$. Next, we show that the preconditions of Theorem 72 are satisfied. First, the upper bounds on $\|\mu\|_2$ and $\|\Sigma - I\|_F$ imply $\sum_{i\in k} w_i \left(\Sigma_i + \mu_i\mu_i^\top\right) = \Sigma + \mu\mu^\top \preceq (1 + \mathcal{O}(\sqrt{\varepsilon}/\alpha))I$. Since the LHS is a conic combination of PSD matrices, it follows that for all $i \in [k]$, $\mu_i\mu_i^\top \preceq \frac{1}{\alpha}\left(1 + \mathcal{O}(\sqrt{\varepsilon}/\alpha)\right) I$, and thus $\left\|\mu_i\mu_i^\top\right\|_F \leqslant \frac{2}{\alpha}$. Next, we can write:

$$\begin{aligned}
\|\Sigma_i - I\|_F &\leqslant \|\Sigma_i - (\Sigma + \mu\mu^\top)\|_F + \|\Sigma - I\|_F + \|\mu\mu^\top\|_F \\
&= \left\|\Sigma_i - \sum_{j\in[k]} w_j \left(\Sigma_j + \mu_j\mu_j^\top\right)\right\|_F + \frac{\sqrt{\varepsilon}k}{\alpha} + \frac{\varepsilon}{\alpha} \\
&\leqslant \left\|\sum_{j\in[k]} w_j \left(\Sigma_i - \Sigma_j\right)\right\|_F + \frac{2}{\alpha} + \frac{\sqrt{\varepsilon}k}{\alpha} + \frac{\varepsilon}{\alpha} \\
&\leqslant \frac{4}{\alpha}\,,
\end{aligned}$$

where the first and the third inequalities follow from the triangle inequality and the upper bound on $\left\|\mu_i\mu_i^\top\right\|_F$, and the last inequality follows from the assumption that $\|\Sigma_i - \Sigma_j\|_F \leqslant 1/\sqrt{\alpha}$ for every pair of covariances Σ_i, Σ_j. So, we can set $\Delta = 4/\alpha$ in Theorem 72. Then, given the definition of α, we have that

$$\eta = 2k^{4k}\mathcal{O}(1 + \Delta/\alpha)^{4k}\sqrt{\varepsilon} = \mathcal{O}\Big(\epsilon^{2/5}\Big)$$

and $1/\varepsilon^2 \geqslant \log(1/\eta)(k + 1/\alpha + \Delta)^{4k+5}/\eta^2$. Therefore, Algorithm 73 outputs a list L of hypotheses such that $|L| = \exp(1/\varepsilon^2)$, and with probability at least 0.99, L contains a hypothesis

that satisfies the following: for all $i \in [k]$,

$$\|\hat{\mu}_i - \mu_i\|_2 = \mathcal{O}\left(\frac{\Delta^{1/2}}{\alpha}\right)\eta^{G(k)} = \mathcal{O}\left(\epsilon^{-1/(20C^{k+1}(k+1)!)} \cdot \varepsilon^{1/(10C^{k+1}(k+1)!)}\right) = \mathcal{O}\left(\epsilon^{1/(20C^{k+1}(k+1)!)}\right) \text{ and}$$
$$\left\|\hat{\Sigma}_i - \Sigma_i\right\|_F = \mathcal{O}\left(k^4\right)\frac{\Delta^{1/2}}{\alpha}\eta^{G(k)} = \mathcal{O}\left(\epsilon^{1/(20C^{k+1}(k+1)!)}\right) . \tag{3.47}$$

Then if we choose a hypothesis in L uniformly at random, the probability that we choose the hypothesis satisfying (3.47) is at least $1/|L| = \exp\left(-1/\varepsilon^2\right)$. □

Finally, if the mixture has a covariance matrix with small variance along any direction, we can further cluster the points by projecting the mixture along that direction.

Lemma 3.6.7 (Spectral Separation of Thin Components)**.** *Given $0 < \epsilon < 1/k^{k^{O(k^2)}}$, let $\alpha = \epsilon^{1/(10C^{k+1}(k+1)!)}$. Let $\mathcal{M} = \sum_{i=1}^k w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with mixture covariance Σ such that $\|\Sigma - I\|_F \leqslant \mathcal{O}(\sqrt{\epsilon}k/\alpha)$, and let X be a set of points satisfying Condition 3.2.45 with respect to $\mathcal{M}$ for some parameters (γ, t). Given a set Y being an ϵ-corrupted version of X of size n, and estimates $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$, such that $\|\mu_i - \hat{\mu}_i\|_2 \leqslant \mathcal{O}\left(\epsilon^{1/(20C^{k+1}(k+1)!)}\right)$, $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant \mathcal{O}\left(\epsilon^{1/(20C^{k+1}(k+1)!)}\right)$, suppose there exists a unit vector $v \in \mathcal{R}^d$ such that $v^\top \hat{\Sigma}_s v \leqslant \mathcal{O}\left(\epsilon^{1/(40C^{k+1}(k+1)!)}\right)$, for some $s \in [k]$. Then, there is an algorithm that efficiently partitions Y into Y_1 and Y_2 such that there is a non-trivial partition of $[k]$ into $Q_1 \cup Q_2$ so that letting $\mathcal{M}_j$ be a distribution proportional to $\sum_{i\in Q_j} w_i G_i$ and $W_j = \sum_{i\in Q_j} w_i$, then Y_j is an $\left(\mathcal{O}(k^2\gamma) + \mathcal{O}\left(\epsilon^{1/(80kC^{k+1}(k+1)!)}/W_j\right)\right)$-corrupted version of $\bigcup_{i\in Q_j} X_i$ satisfying Condition 3.2.45 with respect to $\mathcal{M}_j$ with parameter $\left(\mathcal{O}(k\gamma/W_j), t\right)$.*

Proof. We run the algorithm from Lemma 3.5.1 with the input being the samples Y, the current hypothesis $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$, and the minimum eigenvalue $\eta = \mathcal{O}\left(\epsilon^{1/(40C^{k+1}(k+1)!)}\right)$. Observe that the mixture covariance satisfies $\Sigma \succeq \left(1 - \mathcal{O}(\sqrt{\varepsilon}k/\alpha)\right) I \succeq I/2$ and the upper bound on means and covariance is $\delta = \mathcal{O}\left(\epsilon^{1/(20kC^{k+1}(k+1)!)}n\right)$ by assumption. Therefore, we satisfy the preconditions of Lemma 3.5.1. Thus, we obtain a partition Y_1, Y_2 such that there is a non-trivial partition of $[k]$ into $Q_1 \cup Q_2$ so that letting $\mathcal{M}_j$ be a distribution proportional to $\sum_{i\in Q_j} w_i G_i$ and $W_j = \sum_{i\in Q_j} w_i$, then it follows from Lemma 3.2.48 that Y_j is an $\left(\mathcal{O}(k^2\gamma) + \mathcal{O}\left(\epsilon^{1/(80kC^{k+1}(k+1)!)}/W_j\right)\right)$-corrupted version of $\bigcup_{i\in Q_j} X_i$ satisfying Condition 3.2.45 with respect to $\mathcal{M}_j$ with parameter $\left(\mathcal{O}(k\gamma/W_j), t\right)$. □

3.6.2 Proof of the Main Theorem

We are now ready to complete the proof of Theorem 80.

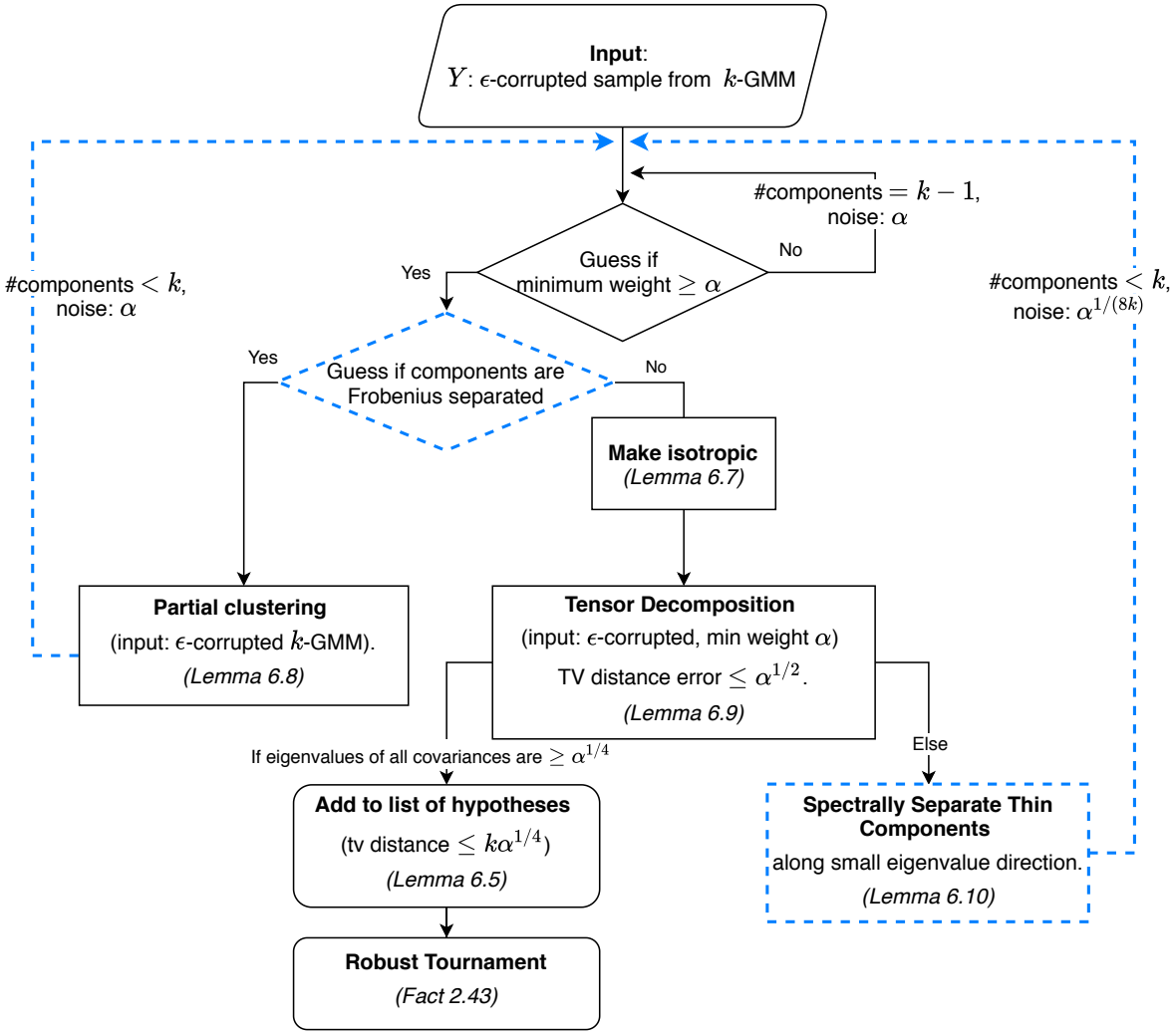

Figure 3.1: If we assume a $1/\mathrm{poly}(k)$ lower bound on minimum weight, then we can skip all blue steps above; the partial clustering is carried out till it can no longer be done within a cluster and then followed by the tensor decomposition step.

Proof of Theorem 80. We divide the proof into two parts: first we show that Algorithm 82 outputs a hypothesis $\widehat{\mathcal{M}} = \sum_{i\in[k]} \hat{w}_i \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i)$ such that $\widehat{\mathcal{M}}$ and $\mathcal{M}$ are $\mathcal{O}(\epsilon^{c_k})$-close in total variation distance with probability at least $\exp\left(-O(k)/\varepsilon^2\right)$; then we show that Algorithm 81 outputs a k-mixture of Gaussians $\widehat{\mathcal{M}}$ such that $\widehat{\mathcal{M}}$ and $\mathcal{M}$ are $\mathcal{O}_k(\epsilon^{c_k})$-close in total variation distance with probability 0.99.

We proceed the first part by induction on k. Let $c_k = \frac{1}{(100)^k C^{(k+1)!} \mathrm{sf}(k+1)k!}$ be a scalar that only depends on k, where $C > 0$ is a sufficiently large universal constant.

Induction Hypothesis: Let X be a set of points satisfying Condition 3.2.45 with respect to a k-mixture of Gaussians $\mathcal{M}$ for some parameters $\gamma = \varepsilon d^{-8k} k^{-C'k}$, where C' is a sufficiently large constant and $t = 8k + 48$. Given a set Y being an ϵ-corrupted version of X of size n, the outlier parameter ε and the component-number parameter k, Algorithm 81 returns a k-mixture of Gaussians $\widehat{\mathcal{M}}$ such that $\widehat{\mathcal{M}}$ and $\mathcal{M}$ are $\mathcal{O}_k(\epsilon^{c_k})$-close in total variation distance with probability $\exp\left(-(3k-2)/\varepsilon^2\right)$.

Base Case: For $k = 1$, the algorithm returns the single Gaussian with mean $\hat{\mu}$ and $\hat{\Sigma}$ at Step 2. Suppose the true Gaussian is $N(\mu, \Sigma)$. It follows from the proof of Lemma 3.6.4,

$$\left\|\Sigma^{\dagger/2}\left(\hat{\mu} - \mu\right)\right\|_2 = \left\|\Sigma^{\dagger/2}\left(\hat{\mu} - \mu\right)\right\|_2 \leqslant \mathcal{O}\left(\sqrt{\epsilon}\right)$$

and

$$\left\|\Sigma^{\dagger/2}\left(\hat{\Sigma} - \Sigma\right)\Sigma^{\dagger/2}\right\|_F \leqslant \mathcal{O}\left(\sqrt{\epsilon}\right) \ ,$$

and thus it follows from Fact 3.2.1 that the total variation distance between the hypothesis Gaussian and the true Gaussian is at most $\mathcal{O}(\sqrt{\epsilon})$. We can then conclude that the base case is true.

Inductive Step: We assume that our induction hypothesis holds for any $m < k$ and then prove that the induction hypothesis holds for k.

Small Clusters Can be Treated as Noise. Conditioning on the base case being true, we begin by guessing whether the minimum weight is less than $\epsilon^{1/(10C^{k+1}(k+1)!)}$ with equal probability.

Let $w_{\min} = \min_i w_i$. If $w_{\min} \leqslant \epsilon^{1/(10C^{k+1}(k+1)!)}$, our algorithm takes step 1 with probability 0.5. In this case, we treat the smallest component as noise and recurse on the set of samples Y. We set the number of components to be $k-1$ and the fraction of outliers being $\epsilon + \epsilon^{1/(10C^{k+1}(k+1)!)} \leqslant 2\epsilon^{1/(10C^{k+1}(k+1)!)}$. By Lemma 3.2.48, Y is an $2\epsilon^{1/(10C^{k+1}(k+1)!)}$-corrupted version of a set satisfying Condition 3.2.45 with respect to a $(k-1)$-mixture for parameters $\gamma = \mathcal{O}\left(k\varepsilon d^{-8k} k^{-C'k}/(1-w_{\min})\right) \leqslant \varepsilon d^{-8(k-1)}(k-1)^{-C'(k-1)}$ and $t = 8k + 48$. Thus applying the inductive hypothesis to Y, we learn the mixture up to total variation distance $\mathcal{O}_k\left(\left(2\epsilon^{1/(10C^{k+1}(k+1)!)}\right)^{c_{k-1}}\right) \leqslant \mathcal{O}_k(\epsilon^{c_k})$ with probability $0.5\exp\left(-(3(k-1)-2)/\varepsilon^2\right) \geqslant \exp\left(-(3k-2)/\varepsilon^2\right)$. Now we may assume for all $i \in [k]$, $w_i \geqslant \epsilon^{1/(10C^{k+1}(k+1)!)}$.

Mixture is Covariance Separated. Let $\alpha = \epsilon^{1/(10C^{k+1}(k+1)!)}$ and $\psi_1 = \{\exists\, \mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j) \mid \|\Sigma_i - \Sigma_j\|_F > \alpha^{-1/2}\}$ be the event that the samples were drawn from a mixture that is *covari-*

ance separated. First, consider the case where ψ_1 is true. We will run 3(a) with probability 0.5. Then it follows from Lemma 3.6.5 that Y can be partitioned into Y_1 and Y_2 in time $d^{O(1)}$, such that they both have at least one component and the fraction of outliers in each set Y_1, Y_2 is at most $\epsilon^{1/(10C^{k+1}(k+1)!)}$ with probability $\alpha^{O(k\log(k/\alpha))}$. Then, we can guess the number of components in Y_1 and we will be correct with probability $1/k$. Conditioned on our guess being correct, let Y_1 consist of k_1 components and Y_2 consist of k_2 components and $k_1 + k_2 = k$.

Let $Q_1 \cup Q_2$ be the non-trivial partition of $[k]$ in Lemma 3.6.5, $\mathcal{M}_j$ be a distribution proportional to $\sum_{i\in Q_j} w_i G_i$ and $W_j = \sum_{i\in Q_j} w_i$, then By Lemma 3.2.48, Y_j is an $\mathcal{O}\Big(\epsilon^{1/(10C^{k+1}(k+1)!)}\Big)$-corrupted version of $\bigcup_{i\in Q_j} X_i$ satisfying Condition 3.2.45 with respect to $\mathcal{M}$ with parameters $\gamma = \mathcal{O}\Big(k\varepsilon d^{-8k}k^{-C'k}/\alpha\Big) \leqslant \varepsilon d^{-8k_j}(k_j)^{-C'k_j}$ and $t = 8k + 48$. Then, applying the inductive hypothesis on Y_j for $j = 1, 2$, with number of components k_j, we can learn the mixtures $\mathcal{M}_j$ up to total variation distance error $\mathcal{O}_k\Big(\epsilon^{c_{k_j}/(10C^{k+1}(k+1)!)}\Big)$ with probability $\exp\left(-(3k_j - 2)/\varepsilon^2\right)$. Finally if this is the case, we combine the two hypotheses on Y_1, Y_2 by multiplying each weight in the hypothesis of Y_j by $|Y_j|/|Y|$ and then taking union of two hypotheses. Then our combining method gives a final output that learns our full hypothesis to total variation distance error $\mathcal{O}_k\Big(\epsilon^{c_{k_1}/(10C^{k+1}(k+1)!)}\Big) + \mathcal{O}_k\Big(\epsilon^{c_{k_2}/(10C^{k+1}(k+1)!)}\Big) \leqslant \mathcal{O}_k(\epsilon^{c_k})$ with probability at least $0.5 \cdot 0.5 \cdot \frac{1}{k} \cdot \alpha^{O(k\log(k/\alpha))} \exp\left(-(3k_1 - 2)/\varepsilon^2\right) \exp\left(-(3k_2 - 2)/\varepsilon^2\right) \geqslant \exp\left(-(3k - 2)/\varepsilon^2\right)$.

Mixture is not Covariance Separated. Next, consider the case where ψ_1 is false. With probability 0.5, the algorithm guesses correctly and executes Step 2. Since the mixture is not covariance separated, we satisfy the preconditions of Lemma 3.6.4, and after applying the transformation in Step 2, Σ, the covariance of the mixture $\mathcal{M}$, is $\sqrt{\varepsilon}k/\alpha$-close to the r-dimensional identity, where r is the rank of Σ. However, since we obtain the subspace exactly, we can simply project all samples on the subspace and we drop the r in the subsequent exposition.

Let X' be the set of points obtained by applying the Affine transformation from Step 2 as defined in Lemma 3.6.4. Then, X' satisfies Condition 3.2.45 with respect to a nearly isotropic mixture and parameters $\gamma = \varepsilon d^{-8k}k^{-C'k}$ and $t = 8k + 48$ so that we can continue the algorithm with X'. Whenever we return a hypothesis in the following steps, we will first apply the inverse of the transformation on our estimates $\hat{\mu}_i$ and $\hat{\Sigma}_i$. Since total variation distance is affine invariant, we have the same error guarantee in total variation distance after applying the transformation. From now on, we reduce to the case where Σ is $\sqrt{\varepsilon}k/\alpha$-close to the Identity.

There is a 50% chance our algorithm runs Step 3(b) and we will analyze the remainder of this case under that assumption. It follows from Lemma 3.6.6 that we obtain a hypothesis $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$ such that $\|\mu_i - \hat{\mu}_i\|_2 \leqslant \mathcal{O}\Big(\epsilon^{1/(20C^{k+1}(k+1)!)}\Big)$ and $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant \mathcal{O}\Big(\epsilon^{1/(20C^{k+1}(k+1)!)}\Big)$

with probability $\exp(-1/\varepsilon^2)$. Conditioned on the hypothesis being correct, we now split into two cases: either all eigenvalues of all the estimated covariances are large (in which case we obtain total variation distance guarantees), or there is a direction along which we can project and cluster further.

Covariance Estimates have Large Eigenvalues. For the hypothesis $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$ from the last step, we compute all the eigenvalues of the estimated covariance matrices, $\hat{\Sigma}_i$, for all $i \in [k]$. If, for all $i \in [k]$, $\lambda_{\min}\left(\hat{\Sigma}_i\right) \geqslant c\epsilon^{1/(40C^{k+1}(k+1)!)}$, we land in Step 3(b).i that we guess the mixing weights $\hat{w}_i$ uniformly in the range $[0, 1]$ and then we output the corresponding hypothesis $\{\hat{w}_i, \hat{\mu}_i, \hat{\Sigma}_i\}_{i\in[k]}$. With probability at least ε^k, $\hat{w}_i$ are within ε of the true mixing weights. Under this condition, by Lemma 3.6.2, the mixture $\widehat{\mathcal{M}} = \sum_{i\in[k]} \hat{w}_i \mathcal{N}\left(\hat{\mu}_i, \hat{\Sigma}_i\right)$ is $\mathcal{O}_k\left(\epsilon^{1/(40C^{k+1}(k+1)!)}\right) \leqslant \mathcal{O}_k(\epsilon^{c_k})$-close to $\mathcal{M}$ in total variation distance with probability $0.5 \cdot 0.5 \cdot \varepsilon^k \cdot \exp(-1/\varepsilon^2) \geqslant \exp(-(3k-2)/\varepsilon^2)$.

One Covariance Has a Small Eigenvalue. Consider the case (Step 3(b).ii) where there exists a unit-norm direction v and an estimate $\hat{\Sigma}_i$ such that $v^\top \hat{\Sigma}_i v \leqslant c\epsilon^{1/(40C^{k+1}(k+1)!)}$. It then follows from Lemma 3.6.7 that we can partition Y into Y_1 and Y_2 such that each has at least one cluster and the total number of outliers in both Y_1 and Y_2 is at most $\mathcal{O}\left(\epsilon^{1/(80kC^{k+1}(k+1)!)}\right) n$. If Y_1 or Y_2 has size less than $\epsilon^{1/(400kC^{k+1}(k+1)!)}n$, then we can treat it as noise and get an additive $O(\epsilon^{1/(400kC^{k+1}(k+1)!)})$-error in total variation distance. Otherwise, the fraction of outliers in both sets is at most $\mathcal{O}\left((\epsilon^{1/(80kC^{k+1}(k+1)!)}n)/(\epsilon^{1/(400kC^{k+1}(k+1)!)}n)\right) = \mathcal{O}\left(\epsilon^{1/(100kC^{k+1}(k+1)!)}\right)$. We then guess the number of components, k_1, in Y_1 with success probability $1/k$. Let $k_2 = k - k_1$ be the number of components in Y_2. Then, conditioned on this event holding, Y_j is an $\mathcal{O}\left(\epsilon^{1/(100kC^{k+1}(k+1)!)}\right)$-corrupted version of a set satisfying Condition 3.2.45 with respect to a mixture of k_j components with parameter $\gamma = k\varepsilon d^{-8k}k^{-C'k}/\alpha \leqslant \varepsilon d^{-8(k_j)}(k_j)^{-C'(k_j)}$ and $t = 8k + 48$. We can apply the inductive hypothesis to Y_1 with number of components k_1 and fraction of outliers $\mathcal{O}\left(\epsilon^{1/(100kC^{k+1}(k+1)!)}\right)$, and conclude that we learn the components of Y_1 to total variation distance $\mathcal{O}_k\left(\epsilon^{c_{k_1}/(100kC^{k+1}(k+1)!)}\right)$ with probability $\exp(-(3k_1-2)/\varepsilon^2)$. A similar argument holds for Y_2. Finally if this is the case, we combine the two hypotheses on Y_1, Y_2 by multiplying each weight by $|Y_j|/|Y|$ and then taking union of two hypotheses. Then our combining method gives a final output that learns our full hypothesis to total variation distance error $\mathcal{O}_k\left(\epsilon^{c_{k_1}/(100C^{k+1}(k+1)!)}\right) + \mathcal{O}_k\left(\epsilon^{c_{k_2}/(100C^{k+1}(k+1)!)}\right) + \mathcal{O}\left(\epsilon^{1/(400kC^{k+1}(k+1)!)}\right) \leqslant \mathcal{O}_k(\epsilon^{c_k})$ with probability at least $0.5 \cdot 0.5 \cdot \frac{1}{k} \cdot \exp(-1/\varepsilon^2 - (3k_1-2)/\varepsilon^2 - (3k_2-2)/\varepsilon^2) \geqslant \exp(-(3k-2)/\varepsilon^2)$.

Sample Size and Running Time of Algorithm 82 By Lemma 3.2.49, we need $n \geqslant kt^{C't}d^t/\gamma^3$ samples to generate X satisfying Condition 3.2.45 with parameters (γ, t). We set $\gamma = \varepsilon d^{-8k}k^{-C'k}$

and $t = 8k + 48$. Then $n \geqslant n_0 = (8k)^{O(k)} d^{O(k)}/\varepsilon^3$. The running time in each sub-routine we invoke is dominated by the running time of the tensor decomposition algorithm, and by Lemma 3.6.6 in the worst case this is $\text{poly}(|L|, n) = \text{poly}\left(\exp\left(1/\varepsilon^2\right), d^{O(k)}/\varepsilon^3\right) = d^{O(k)} \exp\left(1/\varepsilon^2\right)$.

This completes the first part of the proof.

Aggregating Hypotheses. We run Algorithm 82 repeatedly on set Y and add the return hypothesis into a list $\mathcal{L}$ until with probability 0.99, there exists a hypothesis $\widehat{\mathcal{M}} \in \mathcal{L}$ such that $\widehat{\mathcal{M}}$ and $\mathcal{M}$ are $\mathcal{O}_k(\varepsilon^{c_k})$-close in total variation distance. Since Algorithm 82 outputs a correct mixture with probability $\exp\left(-(3k-2)/\varepsilon^2\right)$, we will run Algorithm 82 for $\exp\left(O(k)/\varepsilon^2\right)$ times. Then the total running time is $\exp\left(O(k)/\varepsilon^2\right) \cdot d^{O(k)} \exp\left(1/\varepsilon^2\right) = d^{O(k)} \exp\left(O(k)/\varepsilon^2\right)$.

Robust Tournament. Then we need to run a robust tournament in order to find a hypothesis that is close to the true mixture in total variation distance. Fact 3.2.50 shows that we can do this efficiently only with access to an ϵ-corrupted set of samples of size $\mathcal{O}_k(\log(1/\varepsilon)/\varepsilon^{2c_k})$.

This completes the proof. □

3.7 More Efficient Robust Partial Cluster Recovery

In this section, we prove the following upgraded partial clustering theorem. In contrast to Theorem 76, here we obtain a probability of success that is inverse exponential in k instead of $1/\alpha$.

Theorem 83 (Robust Partial Clustering in Relative Frobenius Distance)**.** *Let $0 \leqslant \epsilon < \alpha/k \leqslant 1$ and $t \in \mathbb{N}$. There is an algorithm with the following guarantees: Let Y be an ϵ-corruption of a sample X of size $n \geqslant (dk)^{Ct}/\epsilon$ for a large enough constant $C > 0$, from $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ that satisfies Condition 3.2.45 with parameters $2t$ and $\gamma \leqslant \epsilon d^{-8t} k^{-Ck}$, for a large enough constant $C > 0$. Suppose further that $w_i \geqslant \alpha > 2\epsilon$ for each $i \in [k]$, and that for some $t \in \mathbb{N}$, $\beta > 0$ there exist $i, j \leqslant k$ such that $\left\|\Sigma^{\dagger/2}(\Sigma_i - \Sigma_j)\Sigma^{\dagger/2}\right\|_F^2 = \Omega\left((k^2t^4)/\beta^{2/t}\alpha^4\right)$, where Σ is the covariance of the mixture $\mathcal{M}$. Then, for any $\eta \gg \sqrt{\epsilon/\alpha}$, the algorithm runs in time $n^{O(t)}$, and with probability at least $2^{-O(k)}(1 - O(\eta/\alpha - \sqrt{\eta}))$ over the random choices of the algorithm, outputs a partition $Y = Y_1 \cup Y_2$ satisfying:*

1. ***Partition respects clustering:*** *for each i, $\max\left\{\frac{1}{w_i n}|Y_1 \cap X_i|, \frac{1}{w_i n}|Y_2 \cap X_i|\right\} \geqslant 1 - O(\sqrt{\eta}) - O(\frac{\beta}{\eta\alpha^2})$, where $X_i \subset X$ corresponding to the points drawn from $\mathcal{N}(\mu_i, \Sigma_i)$.*
2. ***Partition is non-trivial:*** $\max_i \frac{1}{w_i n}|X_i \cap Y_1|, \max_i \frac{1}{w_i n}|X_i \cap Y_2| \geqslant 1 - O(\sqrt{\eta}) - O(\frac{\beta}{\eta\alpha^2})$.

3.7.1 Algorithm

Our algorithm will solve SoS relaxations of a polynomial inequality system. The indeterminates in this system are X' (that is intended to be the guess for the original uncorrupted sample), a cluster of size αn within X' (indicated by z_is) with mean $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$ and Π (intended to be the square root of $\hat{\Sigma}$). The input corrupted sample Y is a constant in this inequality system. Let $U \in \mathcal{R}^{d\times d}$ and $m, z \in \mathcal{R}^d$ also be indeterminates of the proof system. The system can be thought of as encoding the task of finding clusters $\hat{C}$ within Y that satisfies certifiable hypercontractivity of degree 2 polynomials.

We present the constraints grouped together into meaningful categories below: The first set of constraints enforce that $\hat{\Sigma}$ is the square of Π.

$$\text{Covariance Constraints: } \mathcal{A}_1 = \left\{ \begin{array}{c} \Pi = UU^\top \\ \Pi^2 = \hat{\Sigma} \end{array} \right\} \tag{3.48}$$

The intersection constraints force that X' intersects Y in all but an ϵn points (and thus, 2ϵ-close to unknown sample X).

$$\text{Intersection Constraints: } \mathcal{A}_2 = \begin{Bmatrix} \forall i \in [n], & m_i^2 = m_i \\ & \sum_{i\in[n]} m_i = (1-\epsilon)n \\ \forall i \in [n], & z_i(\tilde{y}_i - x'_i) = 0 \end{Bmatrix} \tag{3.49}$$

The subset constraints enforce that z indicate a subset of size αn of X'.

$$\text{Subset Constraints: } \mathcal{A}_3 = \begin{Bmatrix} \forall i \in [n]. & z_i^2 = z_i \\ & \sum_{i\in[n]} z_i = \alpha n \end{Bmatrix} \tag{3.50}$$

Parameter constraints create indeterminates to stand for the covariance $\hat{\Sigma}$ and mean $\hat{\mu}$ of $\hat{C}$ (indicated by z).

$$\text{Parameter Constraints: } \mathcal{A}_4 = \left\{ \begin{array}{c} \frac{1}{\alpha n}\sum_{i=1}^{n} z_i \left(x'_i - \hat{\mu}\right)\left(x'_i - \hat{\mu}\right)^\top = \hat{\Sigma} \\ \frac{1}{\alpha n}\sum_{i=1}^{n} z_i x'_i = \hat{\mu} \end{array} \right\} \tag{3.51}$$

Certifiable Hypercontractivity : $\mathcal{A}_4$=

$$\left\{ \begin{array}{c} \forall t \leqslant 2s \quad \mathbf{E}_z (Q - \mathbf{E}_z Q)^{2t} \leqslant (Ct/\alpha)^t 2^{2t} \left(\mathbf{E}_z (Q - \mathbf{E}_z Q)^2 \right)^t \\ \mathbf{E}_z (Q - \mathbf{E}_z Q)^2 \leqslant 10 \left(\frac{1}{\alpha} \right)^2 \|Q\|_F^2 \end{array} \right\} \tag{3.52}$$

where we write $\mathbf{E}_z Q$ as a shorthand for the polynomial $\frac{1}{\alpha^n} \sum_i z_i Q(x_i)$ and $\mathbf{E}_z (Q - \mathbf{E}_{X_r} Q)^{2j}$ for the polynomial $\frac{1}{\alpha n} \sum_i z_i \left(Q(x'_i) - \frac{1}{\alpha n} \sum_{i \leqslant n} z_i Q(x'_i) \right)^{2j}$ for any j. Note that Q is a $d \times d$-matrix valued indeterminate. Observe that Q itself can be eliminated from the system as is standard in several applications of SoS proofs in obtaining a succinct set of polynomial constraints (see Section 4.3 on "Succinct Representation of Constraints" in [FKP+19] for an exposition).

Algorithm 84 (Polynomial Time Partial Clustering)**.**

Given: *A sample Y of size n. An outlier parameter $\epsilon > 0$ and an accuracy parameter $\eta > 0$.*

Output: *A partition of Y into partial clustering $Y_1 \cup Y_2$.*

Operation:

1. ***Mean and Covariance Estimation**: Apply Robust Mean and Covariance Estimation (Fact 3.2.36) to estimate $\hat{\mu}$ and $\tilde{\Sigma}$ such that $\frac{1}{2}\Sigma \preceq \tilde{\Sigma} \preceq 1.5\Sigma$ where Σ is the covariance of the uncorrupted input mixture.*
2. ***Approximate Isotropic Transformation**: For each $y_i \in Y$, let $\tilde{y}_i = \tilde{\Sigma}^{\dagger/2}(y_i - \hat{\mu})$. Let $\tilde{Y} = \cup_{i \leqslant n} \tilde{y}_i$.*
3. ***SDP Solving:** Find a pseudo-distribution $\tilde{\zeta}$ satisfying $\cup_{i=1}^4 \mathcal{A}_i$ such that $\tilde{\mathbb{E}}_{\tilde{\zeta}} z_i \in \alpha \pm o_d(1)$ for every i. If no such pseudo-distribution exists, output fail.*
4. ***Rounding:** Let $M = \tilde{\mathbb{E}}_{z \sim \tilde{\zeta}} [zz^\top]$.*
   (a) ***Generate candidate clusters**: For $\ell = O(1/\alpha \log \eta/\alpha)$ times, draw a uniformly random $i \in [n]$ and let $\hat{C}_i = \{j \mid M(i,j) \geqslant \alpha^2/2\}$. Let $\mathcal{L} = \cup_{i \leqslant \ell} \hat{C}_i$.*
   (b) ***Candidate 2nd Moment Estimation:** For each $\hat{C}_i \in \mathcal{L}$, let S_i be the output of running robust 2nd moment estimation with Frobenius error (Lemma 3.7.5) on $\hat{C}_i$ with outlier parameter $\eta'_i = O(\frac{\epsilon}{\alpha} + \frac{\beta}{\alpha^2 \eta})$.*
   (c) ***Merge candidate clusters**: For each $i \leqslant \ell$, find $\mathcal{L}_i$ to be the collection of all j such that $\|S_i - S_j\|_F \leqslant 2C\tau$ for a large enough constant $C > 0$. Set*

$\hat{C}_i \cup \mathcal{L}_i = \hat{B}_i$. *Repeat on* $\mathcal{L} \setminus \{\mathcal{L}_i \cup i\}$.

(d) ***Output a union of a random subset of candidates***: *For* $\mathcal{L}' = \cup_i \hat{B}_i$, *choose a uniformly random subset* S *of* $\mathcal{L}'$, *set* $Y_1 = \cup_{j \in S} \hat{B}_j$ *and set* $Y_2 = Y \setminus Y_1$.

Analysis of Algorithm

Lemma 3.7.1 (Success of Step 1)**.** *Let* $\tilde{\Sigma}$ *be the output of the robust covariance estimation algorithm (Fact 3.2.36) applied to the input sample* Y *with outlier parameter* ϵ. *If* Y *is an* ϵ*-corruption of a sample* X *from a GMM with minimum weight* $\geqslant \alpha \geqslant \Omega(\sqrt{\epsilon})$, *mixture mean* μ *and covariance* Σ *satisfying Condition 3.2.45, then,*

$$0.5\Sigma \preceq \tilde{\Sigma} \preceq 1.5\Sigma\,,$$
$$\left\|\tilde{\Sigma}^{-1/2}(\mu - \hat{\mu})\right\|_2 \leqslant O(\sqrt{\epsilon}/\alpha)\,.$$

Proof. The lemma immediately follows by noting that GMMs with minimum weight α are 4-certifiably $1/\alpha$-subgaussian (Fact 3.2.27) and $\alpha \geqslant \Omega(\sqrt{\epsilon})$.

□

Lemma 3.7.2 (Simultaneous Intersection Bounds for Frobenius Separated Case)**.** *Let* $X = X_1 \cup X_2 \cup \ldots X_k$ *be a sample of size* $n \geqslant (dk)^{Ct}/\epsilon$ *for a large enough constant* $C > 0$, *from* $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ *that satisfies Condition 3.2.45 with parameters* $2t$ *and* $\gamma \leqslant \epsilon d^{-8t} k^{-Ck}$. *Suppose further that* $\|\mu_i\|_2 \leqslant \frac{2}{\alpha}$ *for every* i, $\|\Sigma_i\|_2 \leqslant \frac{1}{\alpha}$ *for every* i *and the mixture mean* μ, *covariance* Σ *satisfy* $\|\mu\|_2 \leqslant 1$ *and* $0.5I \preceq \Sigma \preceq 1.5I$. *Let* $\tau = 10^8 \frac{C^6 t^4}{\beta^{2/t}\alpha^2}$, *for any* $\beta > 0$. *Then, given any* ϵ*-corruption* Y *of* X, *for every* i, j *such that* $\|\Sigma_i - \Sigma_j\|_F^2 \geqslant \Omega(\tau)$,

$$\left\{\cup_{i=1}^4 \mathcal{A}_i\right\} \Big|_{2t}^{z} \{z'(X_i)z'(X_j) \leqslant \beta\}\,,$$

where $z'(X_i) = \frac{1}{w_i n} \sum_{j \in X_i} z_j \mathbf{1}(x_j = y_j)$ *for every* i.

Proof of Theorem 83. First, since Y is an ϵ-corruption of a sample X from a GMM such that X satisfies Condition 3.2.45, our robust mean and covariance estimation procedure (Step 1) applied to the mixture succeeds and recovers an estimate of the covariance that is multiplicative 1 ± 0.5-factor approximation in Löwner order. Thus, for the rest of the analysis, we can assume that the smallest and largest eigenvalue of the mixture covariance are at least 0.5 and at most 1.5. Since each component has weight at least α, this means that each of the constitute component

covariance can now be assumed to have a spectral norm at most $1.5/\alpha$.

Next, by an argument similar to the one presented in the proof of Theorem 75, the convex program we wrote is approximately solvable in polynomial time and is feasible whenever the uncorrupted sample X satisfies Condition 3.2.45. The only change here is in the certifiable hypercontractivity constraints where instead of the RHS of the bounded variance constraint is stated in terms of $\|Q\|_F^2$ instead of $\|\Pi Q \Pi\|_F^2$ with an additional slack of $O(1/\alpha^2)$. This modified constraint is satisfied by all true clusters by an application of Lemma 3.2.25 since each of their covariance has spectral norm at most $1.5/\alpha$.

Rounding Let $M = \tilde{\mathbb{E}}_{\tilde{\xi}} zz^\top$. Then, by an argument similar to the proof of Theorem 75, we can conclude:

1. $o_d(1) + \alpha \geqslant M(i,j) \geqslant 0$.
2. $\sum_{j=1}^n M(i,j) \geqslant (\alpha^2 - o_d(1))n$ for every i.
3. For every i, let B_i be the set of "large entries": i.e. j such that $M(i,j) \geqslant \alpha^2/2$. Then, $|B_i| \geqslant \alpha n/2$.

In the following, let M_i denote the i-th row of M and $\|M_i\|_1$ for the sum of the non-negative entries of the vector M_i.

Candidate Clusters For every i, let $F_i \subseteq [k]$ be the set of all $i' \in [k]$ such that $\|\Sigma_i - \Sigma_{i'}\|_F^2 \geqslant \tau$ (i.e., F_i is the set of indices of true clusters whose covariances are far from that of the i-th cluster in Frobenius norm). For every row $j \in [n]$, let $C(j) \in [k]$ be such that $j \in X_{C(j)}$ Let's call j-th row of M "good" if $x_j = y_j$ (i.e j-th sample is not an outlier) and the following condition holds:

$$\sum_{r \in F_{C(j)}} \sum_{\ell \in X_r : x_\ell = y_\ell} M(j,\ell) \leqslant \|M_j\|_1 \left(\frac{\beta}{\eta}\right) .$$

Thus, by Markov's inequality, the fraction of non-outlier entries in B_j that come from $X_{r'}$ such that $r' \in F_r$ is at most $2\left(\frac{\beta}{\eta\alpha^2}\right)$.

Let us estimate the fraction of good rows now. From Lemma 3.4.3 and Fact 3.2.18, we have that for every r and $r' \in F_r$:

$$\tilde{\mathbb{E}}[z'(X_r)z'(X_{r'})] \leqslant \beta .$$

Here, recall that $z'(X_r) = \frac{1}{w_r n} \sum_{i \leqslant n} z_i \mathbf{1}(y_i = x_i)$ for every r. Summing up over $r' \in F_r$ yields:

$$\frac{1}{w_r n} \sum_{r' \in F_r} \sum_{i \in X_r : x_i = y_i} \sum_{j \in X_{r'} : x_j = y_j} \tilde{\mathbb{E}}[z_i z_j] \leqslant n\beta \,.$$

Thus, by Markov's inequality, with probability at least $1 - \eta$ over the choice of $i \in X_r$ such that $x_i = y_i$, it must hold that:

$$\sum_{r' \in F_r} \sum_{j \in X_{r'} : x_j = y_j} \tilde{\mathbb{E}}[z_i z_j] \leqslant n \left(\frac{\beta}{\eta} \right) .$$

Using that $(1 - \epsilon/\alpha)$-fraction of $i \in X_r$ satisfy $x_i = y_i$, for every r, we conclude that $1 - \eta - \epsilon/\alpha$-fraction of the rows X_r are good.

Thus, with probability at least $(1 - \eta - \epsilon/\alpha)^\ell \geqslant (1 - O(\ell(\eta + \epsilon/\alpha)))$, every candidate cluster picked in Step 1 of our rounding algorithm corresponds to the large entries from a good row of M.

We next claim that we cover most of the points in the input in the union of the candidate clusters:

$$|\cup_{i \leqslant \ell} \hat{C}_i| \geqslant \left(1 - 2\sqrt{\eta} - \frac{\epsilon}{\sqrt{\eta}\alpha} \right) n \tag{3.53}$$

with probability at least $1 - \sqrt{\eta}$. To see why, let's estimate the chance that an element $j \in [n]$ does not appear in any of the $\hat{C}_i$s. First, we can assume that j-th row of M is good (this loses us $\eta + \epsilon/\alpha$-fraction js). For each such j, there are at least $\alpha n/2$ large entries. Since M is symmetric, the j-th column of M also has $\alpha n/2$ large entries. Further, j appears in $\cup_{i \leqslant \ell} \hat{C}_i$ if at least one of the $\alpha n/2$ large entries are chosen in our rounding. The chance that this does not happen in any of the ℓ picks is at most $(1 - \alpha/2)^\ell$. Since $\ell = \Theta\left(\frac{1}{\alpha} \log(1/\eta)\right)$, this chance is at most $O(\eta)$. Thus, in expectation $|[n] \setminus \cup_{i \leqslant \ell} \hat{C}_i| \leqslant O\left(\eta + \epsilon/\alpha\right) n$. By Markov's inequality, with probability at least $1 - \sqrt{\eta}$, $|[n] \setminus \cup_{i \leqslant \ell} \hat{C}_i| \leqslant O\left(\sqrt{\eta} + \frac{\epsilon}{\sqrt{\eta}\alpha}\right) n$.

By a union bound, with probability at least $1 - O\left(\eta\ell - \epsilon\ell/\alpha\right) - \sqrt{\eta} \geqslant 1 - O\left(\eta \log(1/\eta)/\alpha - \sqrt{\eta}\right)$, we must thus have both the following events hold simultaneously:

$$|\cup_{i \leqslant \ell} \hat{C}_i| \geqslant \left(1 - 2\sqrt{\eta} - \frac{\epsilon}{\sqrt{\eta}\alpha} \right) n \geqslant (1 - 3\sqrt{\eta}) n \tag{3.54}$$

and, for every $1 \leqslant i \leqslant \ell$,

$$|\hat{C}_i \cap (\cup_{r' \in F_{C(r)}} X_{r'})| \leqslant 2\left(\frac{\beta}{\eta\alpha} + \epsilon/\alpha\right) \cdot |\hat{C}_i|\,. \tag{3.55}$$

Merging Candidate Clusters Observe, following the proof of Theorem 76, we know that there exists a partition of Y into sets Y_1 and Y_2 such that for all i,

$$\max\left\{\frac{1}{w_i n}|Y_1 \cap X_i|, \frac{1}{w_i n}|Y_2 \cap X_i|\right\} \geqslant 1 - O(\sqrt{\eta}) - O(\frac{\beta}{\eta\alpha^2}),$$

and

$$\max_i \frac{1}{w_i n}|X_i \cap Y_1|, \max_i \frac{1}{w_i n}|X_i \cap Y_2| \geqslant 1 - O(\sqrt{\eta}) - O(\frac{\beta}{\eta\alpha^2}).$$

Next, we show that the merging step preserves this partition. For each $\hat{C}_i$, let $\hat{C}_i' = \hat{C}_i \cap \cup_{j \notin F_{C(i)}} X_j$. That is, $\hat{C}_i'$ is the subset of $\hat{C}_i$ obtained by removing points from "far-off" clusters and the outliers. Then, since we know that $|\hat{C}_i| \geqslant \alpha n/2$ and $|X \cap Y| \geqslant (1-\epsilon)n$, we must have $|\hat{C}_i| - |\hat{C}_i'| = \eta_i'|\hat{C}_i| \leqslant \left(\frac{3\epsilon}{\alpha} + \frac{2\beta}{\eta\alpha^2}\right)|\hat{C}_i|$, where we note that $\eta_i' \leqslant \left(\frac{3\epsilon}{\alpha} + \frac{2\beta}{\eta\alpha^2}\right)$.

Thus, $\hat{C}_i'$ is a collection of $\geqslant (1-\eta_i')\alpha n/2$ points from the submixture $\cup_{j \notin F_{C(i)}} X_j$. We know that each μ_i is of ℓ_2 norm at most $1/\alpha$, each Σ_i has spectral norm at most $1/\alpha$ and that for every $r, r' \notin F_{C(i)}$, $\|\Sigma_r - \Sigma_{r'}\|_F^2 \leqslant \tau$. Further, Σ_r is at most $\tau + 1/\alpha = O(\tau)$-different in Frobenius norm from the covariance of the sub-mixture. By an argument similar to the proof of Lemma 3.2.42, we can establish that the submixture with components r such that $r \notin F_{C(i)}$ is $O(\tau)$-certifiably bounded variance. Since $\hat{C}_i'$ is a subset of this sub-mixture of size $\alpha n/2$, we immediately obtain that $\hat{C}_i'$ is $O(\tau/\alpha)$-certifiably bounded variance. Thus, applying Lemma 3.7.5 with outlier parameter η_i' to input $\hat{C}_i$ yields an estimate S_i of the 2nd moment of $\hat{C}_i'$ within a Frobenius error of at most $O(\tau/\alpha)$ From Lemma 3.7.6, this is an additional $O(1/\alpha)$ different in Frobenius norm from the 2nd moment of the sub-mixture which, as argued above, is itself at most $O(\tau)$ different in Frobenius norm from Σ_i. Chaining together yields that $\|\Sigma_i - S_i\|_F^2 \leqslant O(\tau/\alpha)$ for some constant C.

Since for every $r \in S, r' \in T$ it holds that $\|\Sigma_r - \Sigma_{r'}\|_F^2 \gg \Omega(\tau/\alpha)$, conditioned on the good event above, our algorithm never merges $\hat{C}_i$ and $\hat{C}_j$ whenever i, j are non-outliers and i is in some cluster in S and j is in some cluster in T. On the other hand, if i, j belong to the same cluster, then, the corresponding estimate $\|S_i - S_j\|_F^2 \leqslant 2C\tau$. Thus, our merging process always merges together any such candidates.

As a result, the output of the merging process can have at most one i from any true cluster

– thus, the number of distinct members of $\mathcal{L}'$ is at most k. We note that the running time is dominated by computing a pseudo distribution satisfying the union of all the constraints (Step 3 in Algorithm 84) and requires $n^{O(t)}$ time. Step 4 computes a degree $O(1)$ sos relaxation for at most $O(\ell)$ components and the merging only requires a fixed polynomial in d and k time.

□

3.7.2 Proof of Lemma 3.7.2

In the following lemma, we show that the constraint system $\mathcal{A}$, via a low-degree sum-of-squares proof, implies that a lower bound on the variance of any degree 2 polynomial on X' whenever the cluster $\hat{C}$ (indicated by z) appreciably intersects two well-separated true clusters.

Lemma 3.7.3 (Lower-Bound on Variance of Degree 2 Polynomials)**.** *Let $Q \in \mathcal{R}^{d\times d}$ be any fixed matrix. Then, for any $i, j \leqslant k$, and $z'(X_r) = \frac{1}{w_r n}\sum_{i\in X_r} z_i \cdot \mathbf{1}(y_i = x_i)$, we have for any $r \neq r' \in [k]$,*

$$\mathcal{A}\Big|_{\frac{z}{4t}} \left\{ z'(X_r)z'(X_r') \leqslant \frac{(32Ct/\alpha)^{2t}}{(\mathbf{E}_{X_r}Q - \mathbf{E}_{X_{r'}}Q)^{2t}} \left(\frac{\alpha^4}{w_r^2 w_{r'}^2} \left(\mathbf{E}_z(Q-\mathbf{E}_z Q)^2\right)^t \right.\right.$$
$$\left.\left. + \frac{\alpha^2}{w_r^2}\left(\mathbf{E}_{X_{r'}}(Q-\mathbf{E}_{X_{r'}}Q)^2\right)^t + \frac{\alpha^2}{w_{r'}^2}\left(\mathbf{E}_{X_r}(Q-\mathbf{E}_{X_r}Q)^2\right)^t \right) \right\}.$$

Proof. Let $z_i' = z_i \mathbf{1}(y_i = x_i)$ for every i. For every $1 \leqslant r \leqslant k$, let $\mathbf{E}_{X_r}Q$ denote the expectation of the homogenous degree 2 polynomial defined by Q: $\mathbf{E}_{X_r}Q = \frac{1}{w_r n}\sum_{i,j\in X_r} Q(x_i)$ for every r where $Q(x_i) = x_i^\top Q x_i$. Similarly, let $\mathbf{E}_z Q$ be the quadratic polynomial in z defined by $\mathbf{E}_z Q = \frac{1}{\alpha n}\sum_{i\leqslant n} z_i Q(x_i)$. Using the substitution rule and non-negativity of the z_i's, we have for any $r, r' \in [k]$:

$$\mathcal{A}\Big|_{\frac{z}{4t}} \left\{ \mathbf{E}_z(Q-\mathbf{E}_z Q)^{2t} = \frac{1}{\alpha n}\sum_{i\in[n]} z_i \left(Q(x_i) - \mathbf{E}_z Q\right)^{2t} \right.$$
$$\left. \geqslant \frac{1}{\alpha n}\sum_{i\in X_r\cup X_{r'}: x_i = y_j} z_i' \left(Q(x_i) - \mathbf{E}_z Q\right)^{2t} \right\} \tag{3.56}$$

Then, using the SoS almost triangle inequality (Fact 3.2.21), we have:

$$
\begin{aligned}
\mathcal{A} \Big|_{\frac{4t}{}}^{\frac{z}{}} \Bigg\{ & \frac{1}{\alpha n} \sum_{i \in X_r \cup X_{r'}} z_i' \left(Q(x_i) - \mathbf{E}_z Q\right)^{2t} \\
& \geqslant 2^{-2t} \left(\frac{1}{\alpha n} \sum_{i \in X_r : i} z_i' \left(\mathbf{E}_{X_r} Q - \mathbf{E}_z Q\right)^{2t} - \frac{1}{\alpha n} \sum_{i \in X_r} z_i' \left(Q(x_i) - \mathbf{E}_{X_r} Q\right)^{2t} \right) \\
& \quad + 2^{-2t} \left(\frac{1}{\alpha n} \sum_{i \in X_r : i : x_i = y_i} z_i' \left(\mathbf{E}_{X_{r'}} Q - \mathbf{E}_z Q\right)^{2t} - \frac{1}{\alpha n} \sum_{i \in X_{r'}, x_i = y_i} z_i' \left(Q(x_i) - \mathbf{E}_{X_{r'}} Q\right)^{2t} \right) \\
& = 2^{-2t} \left(\frac{w_r}{\alpha} z'(X_r) \left(\mathbf{E}_{X_r} Q - \mathbf{E}_z Q\right)^{2t} - \frac{1}{\alpha n} \sum_{i \in X_r} \left(Q(x_i) - \mathbf{E}_{X_r} Q\right)^{2t} \right) \\
& \quad + 2^{-2t} \left(\frac{w_{r'}}{\alpha} z'(X_{r'}) \left(\mathbf{E}_{X_{r'}} Q - \mathbf{E}_z Q\right)^{2t} - \frac{1}{\alpha n} \sum_{i \in X_{r'}} \left(Q(x_i) - \mathbf{E}_{X_{r'}} Q\right)^{2t} \right) \Bigg\}
\end{aligned} \tag{3.57}
$$

Next, observe that by the SoS almost triangle inequality (Fact 3.2.21), we must have:

$$
\mathcal{A} \Big|_{\frac{4t}{}} \left\{ \left(\mathbf{E}_{X_r} Q - \mathbf{E}_z Q\right)^{2t} + \left(\mathbf{E}_{X_{r'}} Q - \mathbf{E}_z Q\right)^{2t} \geqslant 2^{-2t} \left(\mathbf{E}_{X_r} Q - \mathbf{E}_{X_{r'}} Q\right)^{2t} \right\} .
$$

Further, note that $\mathcal{A} \Big|_{\frac{}{O(1)}} \left\{ \frac{w_r}{\alpha} z'(X_r) + \frac{w_{r'}}{\alpha} z'(X_{r'}) \leqslant \frac{1}{\alpha n} \sum_i z_i \leqslant 1 \right\}$. Thus, using Fact 3.4.5 with $A = \frac{w_r}{\alpha} z'(X_r)$, $B = \frac{w_{r'}}{\alpha} z'(X_{r'})$, $C = \left(\mathbf{E}_{X_r} Q - \mathbf{E}_z Q\right)^{2t}$, and $D = \left(\mathbf{E}_{X_{r'}} Q - \mathbf{E}_z Q\right)^{2t}$ and $\tau = 2^{-2t} \left(\mathbf{E}_{X_r} Q - \mathbf{E}_{X_{r'}} Q\right)^{2t}$, we can derive:

$$
\begin{aligned}
\mathcal{A} \Big|_{\frac{4}{}}^{\frac{z}{}} \Bigg\{ & (Ct/\alpha)^{2t} \left(\mathbf{E}_z (Q - \mathbf{E}_z Q)^2\right)^t \\
& \geqslant \mathbf{E}_z (Q - \mathbf{E}_z Q)^{2t} \\
& \geqslant 2^{-6t} \frac{w_r w_{r'}}{\alpha^2} z'(X_r) z'(X_{r'}) \left(\mathbf{E}_{X_r} Q - \mathbf{E}_{X_{r'}} Q\right)^{2t} \\
& - 2^{-6t} \frac{w_r}{\alpha} \mathbf{E}_{X_r} \left(Q - \mathbf{E}_{X_r} Q\right)^{2t} - 2^{-6t} \frac{w_{r'}}{\alpha} \mathbf{E}_{X_{r'}} \left(Q - \mathbf{E}_{X_{r'}} Q\right)^{2t} \\
& \geqslant 2^{-6t} \frac{w_r w_{r'}}{\alpha^2} z'(X_r) z'(X_{r'}) (\mathbf{E}_{X_r} Q - \mathbf{E}_{X_{r'}} Q)^{2t} \\
& - \frac{w_r}{\alpha} (Ct/\alpha)^{2t} \left(\mathbf{E}_{X_r} \left(Q - \mathbf{E}_{X_r} Q\right)^2\right)^t - \frac{w_{r'}}{\alpha} (Ct/\alpha)^{2t} \left(\mathbf{E}_{X_{r'}} \left(Q - \mathbf{E}_{X_{r'}} Q\right)^2\right)^t \Bigg\}
\end{aligned} \tag{3.58}
$$

where the first inequality uses the Certifiable Hypercontractivity constraints ($\mathcal{A}_4$) and the last inequality follows from the Certifiable Hypercontractivity of X_r and $X_{r'}$ (Condition 3.2.45). Rearranging completes the proof.

□

We can use the lemma above to obtain a simultaneous intersection bound guarantee when there are relative Frobenius separated components in the mixture.

Lemma 3.7.4 (Lemma 3.4.3, restated)**.** *Suppose* $\left\|\Sigma^{-1/2}(\Sigma_r - \Sigma_{r'})\Sigma^{-1/2}\right\|_F^2 \geqslant 10^8 \frac{C^6 t^4}{\beta^{2/t}\alpha^4}$. *Then, for* $z(X_r) = \frac{1}{w_r n}\sum_{i\in X_r} z_i \cdot \mathbf{1}(y_i = x_i)$,

$$\mathcal{A} \vdash_{2t} \{z(X_r)z(X_{r'}) \leqslant \beta\} \ .$$

Proof. WLOG, we will work with the transformed points $x_i \to \Sigma^{-1/2}x_i$ where Σ is the covariance of the mixture. Note that our algorithm does not need to know Σ – this transformation is only for simplifying notation in the analysis that follows.

Let $\tilde{\Sigma}_z = \Sigma^{-1/2}\Sigma_z\Sigma^{-1/2}$, $\tilde{\Sigma}_r = \Sigma^{-1/2}\Sigma_r\Sigma^{-1/2}$ and $\tilde{\Sigma}_{r'} = \Sigma^{-1/2}\Sigma_{r'}\Sigma^{-1/2}$ be the transformed covariances. Then, notice that $\left\|\tilde{\Sigma}_r\right\|_2 \leqslant \frac{1}{w_r}\|\Sigma\|_2 \leqslant \frac{1.5}{w_r}$ and $\left\|\tilde{\Sigma}_{r'}\right\|_2 \leqslant \frac{1}{w_{r'}}\|\Sigma\|_2 \leqslant \frac{1.5}{w_{r'}}$.

We now apply Lemma 3.4.7 with $Q = \tilde{\Sigma}_r - \tilde{\Sigma}_{r'}$. Then, notice that $\mathbf{E}_{X_r}Q - \mathbf{E}_{X_{r'}}Q = \left\|\tilde{\Sigma}_r - \tilde{\Sigma}_{r'}\right\|_F^2 + \mu_r^\top(\tilde{\Sigma}_r - \tilde{\Sigma_{r'}})\mu_r - \mu_{r'}^\top(\tilde{\Sigma}_r - \tilde{\Sigma_{r'}})\mu_{r'} \geqslant \|Q\|_F^2 - \frac{4}{\alpha}$. Then, we obtain:

$$\mathcal{A} \vdash_{2t}^{z} \left\{ z(X_r)z(X_{r'}) \leqslant \left(\frac{32Ct/\alpha}{\mathbf{E}_{X_r}Q - \mathbf{E}_{X_{r'}}Q}\right)^{2t} \cdot \left(\frac{\alpha^2}{w_r w_{r'}}\left(\mathbf{E}_z(Q - \mathbf{E}_z Q)^2\right)^t + \frac{\alpha}{w_r}\left(\mathbf{E}_{X_{r'}}(Q - \mathbf{E}_{X_{r'}}Q)^2\right)^t + \frac{\alpha}{w_{r'}}\left(\mathbf{E}_{X_r}(Q - \mathbf{E}_{X_r}Q)^2\right)\right)^t \right\}. \tag{3.59}$$

Since X_r and $X_{r'}$ have certifiably C-bounded variance polynomials for $C = 4$ (as a consequence of Condition 3.2.45 and Fact 3.2.43 followed by an application of Lemma 3.2.25), we have:

$$\mathbf{E}_{X_{r'}}(Q - \mathbf{E}_{X_{r'}}Q)^2 \leqslant 6\left\|\tilde{\Sigma}_{r'}^{1/2}Q\tilde{\Sigma}_{r'}^{1/2}\right\|_F^2 \leqslant \frac{10}{w_{r'}^2}\|Q\|_F^2 \leqslant \frac{10}{\alpha^2}\|Q\|_F^2 \ ,$$

and

$$\mathbf{E}_{X_r}(Q - \mathbf{E}_{X_r}Q)^2 \leqslant 6\left\|\tilde{\Sigma}_r^{1/2}Q\tilde{\Sigma}_r^{1/2}\right\|_F^2 \leqslant \frac{10}{w_r^2}\|Q\|_F^2 \leqslant \frac{10}{\alpha^2}\|Q\|_F^2 \ .$$

Finally, using the bounded-variance constraints in $\mathcal{A}$, we have:

$$\mathcal{A} \left|\frac{Q,z}{4}\right. \mathbf{E}(Q - \mathbf{E}_z Q)^2 \leqslant \frac{10}{\alpha^2} \|Q\|_F^2 .$$

Plugging these estimates back in (3.59) yields:

$$\begin{aligned} \mathcal{A} \left|\frac{z}{4}\right. \Bigg\{ z(X_r)z(X_{r'}) &\leqslant \frac{(1000Ct/\alpha)^{2t}}{\|Q\|_F^{2t}\,\alpha^{2t}} \left(\frac{\alpha^2}{w_r} + \frac{\alpha}{w_{r'}} + \frac{\alpha}{w_r w_{r'}}\right) \\ &\leqslant \frac{3}{w_r w_{r'}} \frac{(1000Ct)^{2t}}{\alpha^{2t}\,\|Q\|_F^{2t}} \leqslant \frac{3(1000Ct)^{2t}}{\alpha^{2t}\,\|Q\|_F^{2t}} \Bigg\}. \end{aligned} \tag{3.60}$$

Plugging in the lower bound on $\|Q\|_F^{2t}$ and applying cancellation within SoS (Fact 3.4.6) completes the proof. □

3.7.3 2nd Moment Estimation Subroutine

The following lemma gives a 2nd moment estimation algorithm with error in Frobenius norm for distributions that have a certifiably bounded covariance. The proof is very similar to the SoS based mean and covariance estimation algorithms but we provide it in full for completeness here.

Lemma 3.7.5 (2nd Moment Estimation in Frobenius Norm)**.** *Let $1/100 \geqslant \eta > 0$. There is an $n^{O(1)}$ time algorithm that takes input an η-corruption Y of an sample X of size n and outputs an estimate M_2 of the 2nd moment of X with the following properties: Let $X \subseteq \mathcal{R}^d$ be a collection of n points satisfying $\left|\frac{Q}{2}\right. \left\{\frac{1}{|X|}\sum_{x\in X}\left(Q(x) - \frac{1}{|X|}Q(x)\right)^2 \leqslant C\|Q\|_F^2\right\}$ for a matrix-valued indeterminate Q. Let $M_2 = \frac{1}{n}\sum_{x\in X} xx^\top$. Then, the estimate $\hat{M}_2$ output by the algorithm satisfies:*

$$\left\|\hat{M}_2 - M_2\right\|_F^2 \leqslant 80C\eta\,.$$

Proof. Consider the constraint system with scalar-valued indeterminates z_i for $1 \leqslant i \leqslant n$ and d-dimensional vector-valued indeterminates $x'_1, x'_2, \ldots, x'_n$ with the following set of constraints:

$\mathcal{A} =$

$$\left\{\begin{array}{ll} \forall i \leqslant n & z_i^2 = z_i \\ & \sum_{i=1}^n z_i = (1-\eta)n \\ & \tilde{M}_2 = \frac{1}{n}\sum_{i=1}^n x'_i x_i'^\top \\ \forall i \leqslant n & z_i x'_i = z_i y_i \\ & \frac{1}{n}\sum_{i=1}^n \left(x'^\top_i Q x'_i - \frac{1}{n}\sum_{i=1}^n x'^\top_i Q x'_i\right)^2 \leqslant C\|Q\|_F^2 \end{array}\right\} \tag{3.61}$$

Observe that $X' = X$ and z_i set to the 0-1 indicator of non-outliers satisfies the constraint system. Thus, the constraints are feasible.

Our algorithm finds a pseudo-distribution $\tilde{\zeta}$ of degree 10 satisfying the above constraints and output $\tilde{\mathbb{E}}[\tilde{M}_2]$. Let us now analyze this algorithm. The key is the following statement that gives a sum-of-squares proof of closeness of $\tilde{M}_2$ and M_2 in Frobenius norm. We use the notation $\mathbf{E}_X Q$ and $\mathbf{E}_{X'} Q$ to abbreviate $\frac{1}{n}\sum_{i=1}^n x_i^\top Q x_i$ and $\frac{1}{n}\sum_{i=1}^n x'^\top_i Q x'_i$ respectively.

$$\begin{aligned} \mathcal{A} \vdash_{\frac{Q}{2}} \Bigg\{ & \left(\frac{1}{n}\sum_{i=1}^n x_i^\top Q x_i - \frac{1}{n}\sum_{i=1}^n x'^\top_i Q x'_i\right)^2 \\ & = \left(\frac{1}{n}\sum_{i=1}^n (1 - z_i \mathbf{1}(x_i = y_i)) x_i^\top Q x_i - x'^\top_i Q x'_i\right)^2 \\ & \leqslant \left(\frac{1}{n}(1 - z_i \mathbf{1}(x_i = y_i))^2\right)\left(\frac{1}{n}\sum_{i=1}^n \left(x_i^\top Q x_i - x'^\top_i Q x'_i\right)^2\right) \\ & \leqslant 20\eta \cdot \left(\frac{1}{n}\sum_{i=1}^n \left(x_i^\top Q x_i - \mathbf{E}_X Q\right)^2 + \frac{1}{n}\sum_{i=1}^n \left(x'^\top_i Q x'_i - \mathbf{E}_{X'} Q\right)^2 + (\mathbf{E}_X Q - \mathbf{E}_{X'} Q)^2\right) \\ & \leqslant 20\eta(2C\|Q\|_F^2) + 20\eta\,(\mathbf{E}_X Q - \mathbf{E}_{X'} Q)^2 \Bigg\} \end{aligned}$$

where the first inequality follows by the SoS version of the Cauchy-Schwarz inequality and the 2nd by the SoS version of the Almost Triangle inequality.

Rearranging and using that $1 - 20\eta > 1/2$ now yields that:

$$\mathcal{A} \vdash_{\frac{Q}{2}} \left\{ \left(\frac{1}{n}\sum_{i=1}^n x_i^\top Q x_i - \frac{1}{n}\sum_{i=1}^n x'^\top_i Q x'_i\right)^2 \leqslant 80C\eta\|Q\|_F^2 \right\}$$

Notice that the LHS above equals the linear polynomial $\langle \tilde{M}_2 - M_2, Q \rangle$. We now plug in $Q = \tilde{M}_2 - M_2$ to obtain:

$$\mathcal{A} \left| \frac{Q}{2} \right. \left\{ \left\| \tilde{M}_2 - M_2 \right\|_F^4 \leqslant 80C\eta \left\| \tilde{M}_2 - M_2 \right\|_F^2 \right\}$$

Applying Fact 3.2.24 yields:

$$\mathcal{A} \left| \frac{Q}{2} \right. \left\{ \left\| \tilde{M}_2 - M_2 \right\|_F^8 \leqslant 80^4 C^4 \eta^4 \right\}$$

Taking pseudo-expectations with respect to $\tilde{\zeta}$ and using Hölder's inequality for pseudo-distributions yields that

$$\left\| \tilde{\mathbb{E}}_{\tilde{\zeta}} \tilde{M}_2 - M_2 \right\|_F^8 \leqslant \tilde{\mathbb{E}}_{\tilde{\zeta}} \left\| \tilde{M}_2 - M_2 \right\|_F^8 \leqslant 80^4 C^4 \eta^4 \,.$$

Taking the 4-th root, we can conlcude our rounded value $\hat{M}_2 = \tilde{\mathbb{E}}_{\tilde{\zeta}} \tilde{M}_2$ satisfies:

$$\left\| \hat{M}_2 - M_2 \right\|_F^2 \leqslant 80C\eta \,.$$

This completes the proof.

□

We also note the following simple consequence of the certifiable bounded variance property that follows via an argument similar to the one employed in the proof of the previous lemma.

Lemma 3.7.6 (Subsamples of Bounded-Variance Distributions)**.** *Let $X \subseteq \mathcal{R}^d$ be a collection of n points satisfying $\left| \frac{Q}{2} \right. \left\{ \frac{1}{|X|} \sum_{x \in X} (Q(x) - \frac{1}{|X|} Q(x))^2 \leqslant C \left\| Q \right\|_F^2 \right\}$ for a matrix-valued indeterminate Q. Let $M_2 = \frac{1}{|X|} \sum_{x \in X} xx^\top$ be the 2nd moment of X. Let $S \subseteq X$ be a subset of size at least $\beta |X|$. Then,*

$$\left|_{\overline{4}} \right. \left\{ \left\| \frac{1}{|S|} \sum_{x \in S} xx^\top - M_2 \right\|_F^2 \leqslant \frac{1}{\beta} \right\} .$$

Proof. We have by the Cauchy-Schwarz inequality:

$$\left|\frac{Q}{2}\right\{\left(\frac{1}{|X|}\sum_{x\in X}\mathbf{1}(x\in S)(x^\top Qx-M_2),Q\rangle\right)^2\leqslant\left(\frac{1}{|X|}\mathbf{1}(x\in S)^2\right)\left(\sum_{x\in X}(x^\top Qx-M_2),Q\rangle)^2\right)$$
$$\leqslant\left(\frac{|S|}{|X|}\right)\|Q\|_F^2\,.$$

We now substitute in $Q=\frac{1}{|X|}\sum_{x\in X}\mathbf{1}(x\in S)(x^\top Qx-M_2$ to obtain:

$$\left|\frac{Q}{2}\right\{\left\|\frac{1}{|X|}\sum_{x\in X}\mathbf{1}(x\in S)(x^\top Qx-M_2)\right\|_F^4\leqslant\left(\frac{|S|}{|X|}\right)\left\|\frac{1}{|X|}\sum_{x\in X}\mathbf{1}(x\in S)(x^\top Qx-M_2)\right\|_F^2\,.$$

We now apply Fact 3.2.24 to yield:

$$\left|\frac{Q}{2}\right\{\left\|\frac{1}{|X|}\sum_{x\in X}\mathbf{1}(x\in S)(x^\top Qx-M_2)\right\|_F^8\leqslant\left(\frac{|S|}{|X|}\right)^4\,.$$

We finally apply Fact 3.4.6 to conclude that:

$$\left|\frac{Q}{2}\right\{\left\|\frac{1}{|X|}\sum_{x\in X}\mathbf{1}(x\in S)(x^\top Qx-M_2)\right\|_F^2\leqslant\left(\frac{|S|}{|X|}\right)\,.$$

Rescaling gives the claim. □

3.8 Getting poly(ϵ)-close in TV Distance: Proof of Theorem 68

Theorem 85 (Robustly Learning k-Mixtures with small error)**.** *Given $0<\epsilon<1/k^{k^{O(k^2)}}$ and a multiset $Y=\{y_1,y_2,\ldots,y_n\}$ of n i.i.d. samples from a distribution F such that $d_{\mathsf{TV}}(F,\mathcal{M})\leqslant\varepsilon$, for an unknown k-mixture of Gaussians $\mathcal{M}=\sum_{i\leqslant k}w_i\mathcal{N}(\mu_i,\Sigma_i)$, where $n\geqslant n_0=d^{O(k)}\mathrm{poly}_k(1/\varepsilon)$, there exists an algorithm that runs in time $n^{O(1)}\mathrm{poly}_k(1/\varepsilon)$ and with probability at least 0.99 outputs a hypothesis k-mixture of Gaussians $\widehat{\mathcal{M}}=\sum_{i\leqslant k}\hat{w}_i\mathcal{N}\left(\hat{\mu}_i,\hat{\Sigma}_i\right)$ such that $d_{TV}\left(\mathcal{M},\widehat{\mathcal{M}}\right)=\mathcal{O}(\varepsilon^{c_k})$, with $c_k=1/(100^kC^{(k+1)!}k!\mathrm{sf}(k+1))$, where $C>0$ is a universal constant and $\mathrm{sf}(k)=\Pi_{i\in[k]}(k-i)!$ is the super-factorial function.*

In order to obtain the above theorem, we require recovering a polynomial sized list of candi-

date parameters, in addition to the efficient partial clustering result we obtained in the previous section. To this end, we show the following list-recovery theorem which is similar to Theorem 72, but the algorithm outputs a polynomial-size list instead.

Theorem 86 (Recovering a small list of candidate parameters)**.** *Fix any $\alpha > \epsilon > 0, \Delta > 0$. Let X, a sample from a k-mixture of Gaussians $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ satisfying Condition 3.2.45 with parameters $\gamma = \varepsilon d^{-8k} k^{-Ck}$, for C a sufficiently large universal constant, and $t = 8k$, and let Y be an ε-corruption of X. Let X' be a set of $n' = O\left(\epsilon\eta / \left(k^5 \left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4ck}$ fresh samples from $\mathcal{M}$ and Z be an ϵ-corruptionn of X'. If $w_i \geqslant \alpha$, $\|\mu_i\|_2 \leqslant \frac{2}{\sqrt{\alpha}}$ and $\|\Sigma_i - I\|_F \leqslant \Delta$ for every $i \in [k]$, then, given k, Y, Z and ε, the algorithm outputs a list L of at most $\ell' = O\left(\left(k^5 \left(\Delta^4 + 1/\alpha^4\right)\right)^{4k} / \eta^{4k}\right)$ candidate hypotheses (component means and covariances), such that with probability at least $99/100$ there exist $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i \in [k]} \subseteq L$ satisfying $\|\mu_i - \hat{\mu}_i\|_2 \leqslant \mathcal{O}\left(\frac{\Delta^{1/2}}{\alpha}\right) \eta^{G(k)}$ and $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant \mathcal{O}(k^4) \frac{\Delta^{1/2}}{\alpha} \eta^{G(k)}$, for all $i \in [k]$. Here, $\eta = (2k)^{4k} \mathcal{O}(1/\alpha + \Delta)^{4k} \cdot \varepsilon^{1/(k^{O(k^2)})}$ and $G(k) = \frac{1}{C^{k+1}(k+1)!}$. The running time of the algorithm is poly$(|L|, |Y|, d^k) \cdot$ poly$_k(1/\epsilon)$.*

3.8.1 Proof of Theorem 86

We use the following notation and background from Moitra-Valiant [MV10]:

Definition 3.8.1 (Statistically Learnable)**.** *Given $\varepsilon > 0$, we call a mixture of Gaussians $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ ε-statistically learnable if $\min_i w_i \geqslant \varepsilon$ and $\min_{i \neq j} d_{TV}(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}(\mu_j, \Sigma_j)) \geqslant \varepsilon$.*

Definition 3.8.2 (Correct Subdivision)**.** *Given a Gaussian mixture of k Gaussians, $\mathcal{M} = \sum_i w_i N(\mu_i, \Sigma_i)$ and a mixture of $k' \leqslant k$ Gaussians $\hat{\mathcal{M}} = \sum_i \hat{w}_i N(\hat{\mu}_i, \hat{\Sigma}_i)$, we call $\hat{\mathcal{M}}$ an ε-correct subdivision of $\mathcal{M}$ if there is a function $\pi : [k] \to [k']$ that is onto and*

1. *$\forall j \in [k']$, $\left|\sum_{i:\pi(i)=j} w_i - \hat{w}_j\right| \leqslant \varepsilon$*
2. *$\forall i \in [k]$, $\|\mu_i - \hat{\mu}_{\pi(i)}\| + \|\Sigma_i - \hat{\Sigma}_{\pi(i)}\|_F \leqslant \varepsilon$.*

Theorem 87 (Theorem 8 in [MV10])**.** *Given an ε-statistically learnable Gaussian mixture $\mathcal{M}$ in isotropic position, for some $\varepsilon > 0$, there exists an algorithm that requires $n =$ poly(d/ε) samples and runs in time $O($poly$_k(n))$ and with probability at least $99/100$ recovers an ε-correct sub-division $\hat{\mathcal{M}}$. Let the corresponding algorithm be referred to as PARTITION PURSUIT.*

The algorithm has two steps: first run the first three steps of Algorithm 3.2 to get the list L' of $\hat{S}$ and $V'_{\hat{S}}$; then apply the following proposition to learn the mixture in the subspace $V'_{\hat{S}}$. This proposition is a generalization of Theorem 87 without the assumption that the total variation distance between each pair of components is at least ε. The sample and time complexities has a worse, but still polynomial dependence on ε. Note that although the algorithm in the proposition is non-robust, we can take a sample without noise with constant probability because the algorithm only requires a polynomial number of samples in ε.

Algorithm 88 (Efficient List-Recovery of Candidate Parameters)**.**

Input: *An ϵ-corruption Y of a sample X from a k-mixture of Gaussians $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$. Let Z be an additional ϵ-corrupted sample of size n' from $\mathcal{M}$.*

Requirements: *The guarantees of the algorithm hold if the mixture parameters and the sample X satisfy:*

1. $w_i \geqslant \alpha$ *for all* $i \in [k]$,
2. $\|\mu_i\|_2 \leqslant 2/\sqrt{\alpha}$ *for all* $i \in [k]$,
3. $\|\Sigma_i - I\|_F \leqslant \Delta$ *for all* $i \in [k]$.
4. *X satisfies Condition 3.2.45 with parameters (γ, t), where $\gamma = \varepsilon d^{-8k} k^{-Ck}$, for C a sufficiently large universal constant, and $t = 8k$.*
5. *The number of fresh samples $n' = O\left(\epsilon\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4ck}$, for a fixed constant c.*

Parameters: $\eta = (2k)^{4k}\mathcal{O}(1/\alpha + \Delta)^{4k}\,\varepsilon^{1/(k^{O(k^2)})}$, $D = C(k^4/(\alpha\sqrt{\eta}))$, $\delta = 2\eta^{1/(C^{k+1}(k+1)!)}$, $\ell' = 100\log k\left(\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4k}$, *for some sufficiently large absolute constant* $C > 0$, $\lambda = 4\eta$, $\phi = 10(1+\Delta^2)/(\sqrt{\eta}\alpha^5)$, $\varepsilon_1 = O\left(\sqrt{\Delta}\delta^{1/4}/\alpha\right)$.

Output: *A list L of hypotheses such that there exists at least one, $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i\leqslant k} \in L$, satisfying: $\|\mu_i - \hat{\mu}_i\|_2 \leqslant \mathcal{O}\left(\frac{\Delta^{1/2}}{\alpha}\right)\eta^{G(k)}$ and $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant \mathcal{O}(k^4)\,\frac{\Delta^{1/2}}{\alpha}\eta^{G(k)}$, where $G(k) = \frac{1}{C^{k+1}(k+1)!}$.*

Operation:

1. ***Robust Estimation of Hermite Tensors**: For $m \in [4k]$, compute $\hat{T}_m$ such that $\max_{m\in[4k]}\left\|\hat{T}_m - \mathbb{E}\left[h_m(\mathcal{M})\right]\right\|_F \leqslant \eta$ using the robust mean estimation algorithm in Fact 3.2.35.*

2. ***Random Collapsing of Two Modes of*** $\hat{T}_4$*: Let* L' *be an empty list. Repeat* ℓ' *times: For* $j \in [4k]$*, choose independent standard Gaussians in* $\mathbb{R}^d$*, denoted by* $x^{(j)}, y^{(j)} \sim \mathcal{N}(0, I)$*, and uniform draws* $a_1, a_2, \ldots, a_t$ *from* $[-D, D]$*. Let* $\hat{S}$ *be a* $d \times d$ *matrix such that for all* $r, s \in [d]$*,* $\hat{S}(r,s) = \sum_{j\in[4k]} a_j \hat{T}_4(r, s, x^{(j)}, y^{(j)}) = \sum_{j\in[4k]} a_j \sum_{g,h\in[d]} \hat{T}_4(r, s, g, h) x^{(j)}(g) y^{(j)}(h)$*. Add* $\hat{S}$ *to the list* L'*.*

3. ***Construct Low-Dimensional Subspace****: Let* V *be the span of all singular vectors of the natural* $d \times d^{m-1}$ *flattening of* $\hat{T}_m$ *with singular values* $\geqslant \lambda$ *for* $m \leqslant 4k$*. For each* $\hat{S} \in L'$*, let* $V'_{\hat{S}}$ *be the span of* V *plus all the singular vectors of* $\hat{S}$ *with singular value larger than* $\delta^{1/4}$*.*

4. ***Moitra-Valiant for Low-Dimensional Subspace****: Initialize* L *to be the empty list. For each* $\hat{S} \in L'$*, let* $\hat{P} = UU^\top$ *be the orthogonal projection matrix onto the span of* $V'_{\hat{S}}$*, where* $U \in \mathcal{R}^{d\times d}$ *has orthonormal columns. Let* $m = \dim V'_{\hat{S}}$ *and let* $\hat{Z} \subset Z$ *be a randomly chosen subset of size poly*(m/ε_1)*. Let* U_m *denote the first* m *columns of* U *and for all* $z \in \hat{Z}$*, compute* $U_m^\top z$*. Run PARTITION PURSUIT on the resulting set of points and let* $\{\hat{\mu}_i^{\hat{P}}, \hat{\Sigma}_i^{\hat{P}}\}_{i\in[k]}$ *be the parameters corresponding to the* ϵ*-correct subdivision output by PARTITION PURSUIT. Let* $\hat{\mu}_i^\top = [(\hat{\mu}_i^{\hat{P}})^\top, 0]$ *be a* d *dimensional vector padded with* 0*s and* $\hat{\Sigma}_i$ *be a* $d \times d$ *matrix with* $\hat{\Sigma}_i^{\hat{P}}$ *in the top left* $m \times m$ *sub-matrix and* 0*'s elsewhere. Add* $\{U\hat{\mu}_i, U\hat{\Sigma}_i U^\top + (\hat{S} + I) - \hat{P}\,(I + S)\,\hat{P}\}_{i\in[k]}$ *to* L*.*

Proposition 3.8.3. *Given* $\varepsilon > 0$ *and a sample* X *of size poly*$(d, 1/\varepsilon)$ *from a* k*-mixture of Gaussians* $\mathcal{M}$ *with mixture covariance* Σ *such that* $0.99I \preceq \Sigma \preceq 1.01I$ *and satisfies* $w_i \geqslant \varepsilon$*, the PARTITION PURSUIT algorithm runs in time poly*$(d, 1/\varepsilon)$ *and with probability at least* $9/10$ *returns an* $O(\varepsilon)$*-correct sub-division, denoted by* $\hat{\mathcal{M}}$*.*

Recall, the PARTITION PURSUIT algorithm satisfies Theorem 87 and we will prove that with an appropriately chosen parameter ε, PARTITION PURSUIT also satisfies Proposition 3.8.3. The main idea is that if any two components are actually close enough in total variation distance, then any algorithm with access to only a polynomial number of samples could never distinguish these two components from a single Gaussian. So if all pairwise distances are either sufficiently large or sufficiently small, the algorithm will behave as if it were given sample access to a mixture that meets the requirements of Theorem 87.

Lemma 3.8.4. *Given* $0 < \gamma, \delta < 1$ *and two distributions,* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *over* $\mathbb{R}^d$ *such that* $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) < \gamma$*, let* X_1 *be set of* n *i.i.d. samples from* $\mathcal{D}_1$ *and* X_2 *be* n *i.i.d. samples from*

$\mathcal{D}_2$. *Let $\mathcal{A}$ be any algorithm that takes as input X_1 and outputs a list of m real numbers, $Y_1 = \{y_i\}_{i\in[m]}$, such that $y_i \in [-1,1]$ with probability at least $1-\delta$. Then, for any $\tau > 0$, $\mathcal{A}$ on input X_2 outputs a list of m real numbers $Y_2 = \{y_i'\}_{i\in[m]}$ such that with probability at least $1-\delta-(4mn\gamma/\tau)$, for all $i \in [m]$, $|y_i - y_i'| \leqslant \tau$.*

Proof. Let $\mathcal{U}_1$ be the uniform distribution over X_1 and $\mathcal{U}_2$ be the uniform distribution over X_2. Then,

$$\begin{aligned} d_{TV}\left(\mathcal{U}_1, \mathcal{U}_2\right) \leqslant \sqrt{2} H^2\left(\mathcal{U}_1, \mathcal{U}_2\right) &= \sqrt{2} n H^2\left(\mathcal{D}_1, \mathcal{D}_2\right) \\ &\leqslant \sqrt{2} n d_{TV}\left(\mathcal{D}_1, \mathcal{D}_2\right) \\ &\leqslant \sqrt{2}\gamma n \end{aligned} \tag{3.62}$$

Consider the family of functions $\mathcal{F}$ that take as input n samples and output a single bit in $\{0,1\}$. We know that for any function $f \in \mathcal{F}$, the probability that $f(X_1) \neq f(X_2)$ is at most $\sqrt{2}\gamma n$. Recall, the algorithm outputs m real numbers in the range $[-1,1]$, which we can discretize into a grid Δ of length τ. There are at most $2/\tau$ distinct grid points and for any $y_i \in [-1,1]$, there exists a point $z_i \in \Delta$ such that $|y_i - z_i| \leqslant \tau$. Further, observe we can represent each y_i using $2/\tau$ functions $f \in \mathcal{F}$. Then, union bounding over the events that each of the $2/\tau$ functions output different bits, for each of the m parameters, we have that with probabiltiy at least $1 - \left(2\sqrt{2}\gamma nm/\tau\right)$, any algorithm outputs a list $\{y_i'\}_{i'\in[m]}$ such that $|y_i - y_i'| \leqslant \tau$. Finally, union bounding over the event that algorithm $\mathcal{A}$ fails with probability δ yields the claim. □

We then prove there is a gap $[f(d,\varepsilon_1),\varepsilon_1)$ between pairwise distances of components so that if we merge components within distance $f(d,\varepsilon_1)$, the resulting mixture is ε_1-statistically learnable.

Lemma 3.8.5. *Let $f(d)(\varepsilon) = f(d,\varepsilon)$. There exists $\ell \in [k^2]$ such that for every pair of components, either $d_{TV}(\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j)) < (f(d))^{\ell}(\varepsilon)$ or $d_{TV}(\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j)) \geqslant (f(d))^{\ell-1}(\varepsilon)$. Moreover, the set of Gaussians with total variation distance at most $(f(d))^{\ell}(\varepsilon)$ is an equivalence class.*

Proof. We can see that intervals $\left\{\left[(f(d))^{\ell}(\varepsilon),(f(d))^{\ell-1}(\varepsilon)\right)\right\}_{\ell\in[k^2]}$ are disjoint. There are at most k^2-1 distinct values of $d_{TV}(\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j))$. So there exists an interval $\left[(f(d))^{\ell}(\varepsilon),(f(d))^{\ell-1}(\varepsilon)\right)$ such that for every pair of components $\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j)$, either $d_{TV}(\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j)) < (f(d))^{\ell}(\varepsilon)$ or $d_{TV}(\mathcal{N}(\mu_i,\Sigma_i),\mathcal{N}(\mu_j,\Sigma_j)) \geqslant (f(d))^{\ell-1}(\varepsilon)$.

Next, we show for any ℓ, Gaussians with pair wise TV distance $(f(d)^{\ell})(\epsilon)$ form an equivalence class. Consider component Gaussians G_1, G_2 and G_3 such that G_1 and G_2 are at total

variation distance at most $(f(d))^{\ell}(\varepsilon)$ and G_2 and G_3 are also at total variation distance at most $(f(d))^{\ell}(\varepsilon)$.

$$\begin{aligned} d_{TV}(G_1, G_3) &\leqslant d_{TV}(G_1, G_2) + d_{TV}(G_2, G_3) \\ &\leqslant 2(f(d))^{\ell}(\varepsilon) \\ &\ll (f(d))^{\ell-1}(\varepsilon) \end{aligned}$$

and since there is no pair of Gaussians with total variation distance inside the interval $[(f(d))^{\ell}(\varepsilon), (f(d))^{\ell-1}(\varepsilon))$, this implies $d_{TV}(G_1, G_3) \leqslant (f(d))^{\ell}(\varepsilon)$. □

We can now complete the proof of Proposition 3.8.3 :

Proof of Proposition 3.8.3. By Lemma 3.8.5, there exists an interval $\left[(f(d))^{\ell}(\varepsilon), (f(d))^{\ell-1}(\varepsilon)\right)$ such that there is no pair of Gaussians with total variation distance inside the interval and $(f(d))^{\ell}(\varepsilon), (f(d))^{\ell-1}(\varepsilon)$ are polynomials in d and ε. Let $\varepsilon_1 = (f(d))^{\ell-1}(\varepsilon)$ and $f(d, \varepsilon_1) = (f(d))^{\ell}(\varepsilon)$. Let X be a set of $n = (d/\epsilon)^c$ samples from $\mathcal{M}$, where c is fixed universal constant. Let $\bar{\mathcal{M}}$ be the mixture obtained by merging all components in an equivalence class with total variation distance at most $f(d, \varepsilon_1)$ to a single Gaussian and observe $d_{TV}\left(\mathcal{M}, \bar{\mathcal{M}}\right) \leqslant k f(d, \varepsilon_1)$. Next, observe that PARTITION PURSUIT outputs at most k means and covariances, which can be represented as a list of at most $2kd^2$ real numbers. Further, since $\Sigma \preceq 1.01I$ and $w_i \geqslant \varepsilon$, the means of each component $\|\mu_i\|_2^2 \leqslant 2/\varepsilon$ and $\|\Sigma_i\|_F^2 \leqslant O(d^2/\varepsilon)$.

Then, rescaling the instance by $O(\varepsilon/d^2)$ and applying Lemma 3.8.4 with $\mathcal{D}_1 = \mathcal{M}, \mathcal{D}_2 = \bar{\mathcal{M}}$, input samples X and accuracy parameter $\tau = (\varepsilon/d)^{c_2}$, for a large enough constant c_2, it follows that with probability at least $1 - 0.99 - O(f(d, \varepsilon_1) \cdot (\varepsilon/d)^{c_3})$, for a fixed constant c_3, the resulting list of numbers is τ-close to that obtained by running PARTITION PURSUIT on a set of n samples from $\bar{\mathcal{M}}$. Since $\bar{\mathcal{M}}$ is ε_1-statistically learnable, it follows from Theorem 87 that with probability at least $9/10$, PARTITION PURSUIT will output an $O(\varepsilon_1)$-correct sub-division $\hat{\mathcal{M}}$.

□

Proof of Theorem 86. Recall, by part (1) of Proposition 3.3.3, the dimension of the subspace $V'_{\hat{S}}$ is $m = \dim V'_{\hat{S}} = O\left(\frac{(k(1+\Delta+1/\alpha))^{4k+5}}{\eta^2}\right)$ and let $\varepsilon_1 = \sqrt{\Delta}\delta^{1/4}/\alpha$. Let c_{mv} be a fixed constant such that $(m/\varepsilon_1)^{c_{mv}}$ samples suffice for applying Theorem 87. Further, obseve in the fresh sample Y, the probability that any given sample is corrupted is ϵ. Let ζ be the event that a random subset of $(m/\varepsilon_1)^{c_{mv}}$ samples from Z does not contain any corrupted points. Then, the event ζ holds with probability at least $(1-\epsilon)^{(m/\varepsilon_1)^{c_{mv}}}$. Conditioning on ζ and running step 4 of Algorithm

88, it follows from Proposition 3.8.3 that we recover $O(\varepsilon_1)$-accurate estimates to the parameters of $\mathcal{M}$ in the subspace, i.e. $\|U^\top \mu_i - \hat{\mu}_i\|_2 \leqslant O(\varepsilon_1)$ and $\|U^\top \Sigma_i U - \hat{\Sigma}_i\|_F \leqslant O(\varepsilon_1)$. Since we repeat the above for ℓ' candidate subspaces in L', the probability over all probability of success is $(1-\epsilon)^{(m/\varepsilon_1)^{cmv}\cdot \ell'}$.

By part (2) in Proposition 3.3.3, there is a vector $\mu_i' \in V_{\hat{S}}'$ such that $\|\mu_i - \mu_i'\| \leqslant \frac{20}{\alpha}\delta^{1/4}\Delta^{1/2}$ where $\delta = 2\eta^{1/(C^{k+1}(k+1)!)}$ and $\eta = O(4k(1+1/\alpha+\Delta)^{4k}\sqrt{\varepsilon_1})$. Let $\hat{P} = UU^\top$ be a projection matrix where the columns of Q span $V_{\hat{S}}'$ and let $Q^\top \mu_i$ be the projection of the true means to the corresponding subspace. Then,

$$\begin{aligned}
\|U\hat{\mu}_i - \mu_i\|_2 &\leqslant \|U\hat{\mu}_i - \mu_i'\|_2 + \|\mu_i' - \mu_i\|_2 \\
&\leqslant \|U\hat{\mu}_i - P(\mu_i' - \mu_i + \mu_i)\|_2 + O\left(\frac{\sqrt{\Delta}\delta^{1/4}}{\alpha}\right) \\
&\leqslant \|\hat{\mu}_i - U^\top \mu_i\|_2 + O\left(\frac{\sqrt{\Delta}\delta^{1/4}}{\alpha}\right) \\
&\leqslant O\left(\frac{\sqrt{\Delta}\delta^{1/4}}{\alpha}\right).
\end{aligned}$$

where the third inequality follows from observing that $U^\top \mu_i$ is the true mean in the low dimensional subspace and applying Proposition 3.8.3.

By Proposition 3.3.2, there exists $\hat{S} \in L'$ such that $\hat{S} - (\Sigma_i - I) = P_i + Q_i$ where $\|P_i\|_F = O(\sqrt{\eta/\alpha})$. Again by part (3) in Proposition 3.3.3, there exists a symmetric matrix $Q_i' \in V_{\hat{S}}' \times V_{\hat{S}}'$ such that $\|Q_i - Q_i'\|_F \leqslant O(\frac{k^2}{\alpha}\delta^{1/4}\Delta^{1/2})$. We also know that in the subspace spanned by $V_{\hat{S}}'$, $\|\hat{\Sigma}_i - U^\top \Sigma_i U\|_F^2 \leqslant \text{poly}(\varepsilon_2)$. Recall, Algorithm 88 outputs the following estimate: $\hat{M} = U\hat{\Sigma}U^\top + \left(I + \hat{S}\right) - \hat{P}\left(I + \hat{S}\right)\hat{P}$. Observe, for any matrix M and projection matrix P, $M = PMP + (I-P)M(I-P) + PM(I-P) + (I-P)MP$. Then,

$$\begin{aligned}
\|\Sigma_i - \hat{M}\|_F \leqslant &\underbrace{\|\hat{P}\left(\Sigma_i - \hat{M}\right)\hat{P}\|_F}_{(1)} + \underbrace{\|\left(I-\hat{P}\right)\left(\Sigma_i - \hat{M}\right)\left(I-\hat{P}\right)\|_F}_{(2)} \\
&+ \underbrace{\|\hat{P}\left(\Sigma_i - \hat{M}\right)\left(I-\hat{P}\right)\|_F}_{(3)} + \underbrace{\|\left(I-\hat{P}\right)\left(\Sigma_i - \hat{M}\right)\hat{P}\|_F}_{(4)}
\end{aligned} \tag{3.63}$$

We bound each of the terms above. Since $\hat{P}\left(I + S - P(I+S)P\right)\hat{P} = 0$, we can bound term

(1) as follows

$$\|\hat{P}\left(\Sigma_i - \hat{M}\right)\hat{P}\|_F = \|\hat{P}\Sigma_i\hat{P} - \hat{P}U\hat{\Sigma}_i U^\top \hat{P}\|_F = \|U^\top \Sigma_i U - \hat{\Sigma}_i\| \leqslant O\left(\frac{\sqrt{\Delta}\delta^{1/4}}{\alpha}\right) \tag{3.64}$$

Similarly, since $\left(I - \hat{P}\right)\left(U\hat{\Sigma}_i U^\top\right)\left(I - \hat{P}\right) = 0$ and $\Sigma_i = I + \hat{S} - P_i - Q_i$, we can bound term (2) as follows:

$$\begin{aligned}
\|\left(I - \hat{P}\right)\left(\Sigma_i - \hat{M}\right)\left(I - \hat{P}\right)\|_F &= \|\left(I - \hat{P}\right)\left(\Sigma_i - \left(I + \hat{S}\right)\right)\left(I - \hat{P}\right)\|_F \\
&\leqslant \|\left(I - \hat{P}\right)\left(\Sigma_i - \left(I + \hat{S} - Q_i\right)\right)\left(I - \hat{P}\right)\|_F + \|\left(I - \hat{P}\right)Q_i\left(I - \hat{P}\right)\|_F \\
&\leqslant \|P_i\|_F^2 + \|\left(I - \hat{P}\right)\left(Q_i - Q_i'\right)\left(I - \hat{P}\right)\|_F + \|\left(I - \hat{P}\right)Q_i'\left(I - \hat{P}\right)\|_F \\
&\leqslant O\left(\sqrt{\frac{\eta}{\alpha}} + \frac{k^2\delta^{1/4}\Delta^{1/2}}{\alpha}\right)
\end{aligned} \tag{3.65}$$

Next, we bound term (3). Observe, $\hat{P}\left(U\hat{\Sigma}_i U^\top\right)\left(I - \hat{P}\right) = 0$ and $\hat{P}(I + S)\hat{P}(I - \hat{P}) = 0$. Thus,

$$\begin{aligned}
\|\hat{P}\left(\Sigma_i - \hat{M}\right)\left(I - \hat{P}\right)\|_F &= \|\hat{P}\left(\Sigma_i - \left(I + \hat{S}\right)\right)\left(I - \hat{P}\right)\|_F \\
&= \|\hat{P}\left(P_i + Q_i\right)\left(I - \hat{P}\right)\|_F \\
&\leqslant \|\hat{P}P_i\left(I - \hat{P}\right)\|_F \\
&\leqslant O\left(\sqrt{\frac{\eta}{\alpha}}\right)
\end{aligned} \tag{3.66}$$

Obseve, term (4) follows from a similar argument. Combining equations (3.64), (3.65),(3.66) and substituting back into (3.63) we can conclude

$$\|\Sigma_i - \hat{M}\|_F \leqslant O\left(\sqrt{\frac{\eta}{\alpha}} + \frac{k^2\delta^{1/4}\Delta^{1/2}}{\alpha}\right)$$

The size of L' is $\ell' = O\left(\log k\left(\eta/\left(k^5\left(\Delta^4 + 1/\alpha^4\right)\right)\right)^{-4k}\right)$ and since we add a single tuple of k means and covariances for each subspace in L', the list L has the same size. The running time is poly $\left(|Y|, |L|, d^k, m, 1/\varepsilon_1\right)$ concluding the proof. □

3.8.2 Proof of Theorem 85

Since we have all the main ingredients: the tensor decomposition algorithm recovering a polynomial size of list (Theorem 86), the upgraded partial clustering algorithm with high probability of success (Theorem 83) and the spectral separation algorithm of thin components (Lemma 3.5.1), we can now complete the proof of Theorem 85.

The algorithm establishing Theorem 85 is almost the same as Algorithm 81. The only difference is we will replace Algorithm 77 by Algorithm 84 and replace Algorithm 73 by Algorithm 88. The following two lemmas show that by modifying the parameters slightly and applying the upgraded partial clustering and tensor decomposition algorithms, we can have the same conclusions as in Lemma 3.6.5 and Lemma 3.6.6 with a polynomial success probability. Then the proof of Theorem 85 is exactly the same as the proof of Theorem 80 in Section 3.6.2 except for the use of Lemma 3.8.6 and Lemma 3.8.7 instead of Lemma 3.6.5 and Lemma 3.6.6.

Lemma 3.8.6 (Non-negligible Weight and Covariance Separation)**.** *Given* $0 < \epsilon < 1/k^{k^{O(k^2)}}$ *and* $k \in \mathbb{N}$*, let* $\alpha = \epsilon^{1/(45C^{k+1}(k+1)!)}$.

Let $\mathcal{M} = \sum_{i=1}^{k} w_i G_i$ *with* $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ *be a* k*-mixture of Gaussians with mixture covariance* Σ *such that* $w_i \geqslant \alpha$ *for all* $i \in [k]$ *and there exist* $i, j \in [k]$ *such that* $\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F^2 > 1/\alpha^5$*. Further, let* X *be a set of points satisfying Condition 3.2.45 with respect to* $\mathcal{M}$ *for some parameters* $\gamma \leqslant \varepsilon d^{-8k}k^{-Ck}$*, for a sufficiently large constant* C*, and* $t \geqslant 8k$*. Let* Y *be an* ϵ*-corrupted version of* X *of size* $n \geqslant n_0 = (dk)^{\Omega(1)}/\epsilon$*, Algorithm 84 partitions* Y *into* Y_1*,* Y_2 *in time* $n^{O(1)}$ *such that with probability at least* $2^{-O(k)}(1 - O(\alpha))$ *there is a non-trivial partition of* $[k]$ *into* $Q_1 \cup Q_2$ *so that letting* $\mathcal{M}_j$ *be a distribution proportional to* $\sum_{i \in Q_j} w_i G_i$ *and* $W_j = \sum_{i \in Q_j} w_i$*, then* Y_j *is an* $\mathcal{O}\left(\epsilon^{1/(45C^{k+1}(k+1)!)}\right)$*-corrupted version of* $\bigcup_{i \in Q_j} X_i$ *satisfying Condition 3.2.45 with respect to* $\mathcal{M}$ *with parameters* $\left(\mathcal{O}(k\gamma/W_j), t\right)$.

Proof. We run Algorithm 84 with sample set Y, number of components k, the fraction of outliers ϵ and the accuracy parameter η. Since X satisfies Condition 3.2.45, we can set $t = 10$, $\beta = (k^2t^4\alpha)^{t/2} = O_k(\alpha^5)$ and $\eta = \alpha^2 \gg \sqrt{\varepsilon/\alpha}$ in Theorem 83. Then, by assumption, there exist i, j such that

$$\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F^2 > \frac{1}{\alpha^5} = \Omega\left(\frac{k^2t^4}{\beta^{2/t}\alpha^4}\right).$$

We observe that we also satisfy the other preconditions for Theorem 83, since $n \geqslant (dk/)^{\Omega(1)}/\epsilon$.

Then, Theorem 83 implies that with probability at least $2^{-O(k)}(1-O(\eta/\alpha-\sqrt{\eta})) = 2^{-O(k)}(1-O(\alpha))$, the set Y is partitioned in two sets Y_1 and Y_2 such that there is a non-trivial partition of $[k]$

into $Q_1 \cup Q_2$ so that letting $\mathcal{M}_j$ be a distribution proportional to $\sum_{i \in Q_j} w_i G_i$ and $W_j = \sum_{i \in Q_j} w_i$, then Y_j is an $\mathcal{O}\left(\epsilon^{1/(45C^{k+1}(k+1)!)}\right)$-corrupted version of $\bigcup_{i \in Q_j} X_i$. By Lemma 3.2.48, $\bigcup_{i \in Q_j} X_i$ satisfies Condition 3.2.45 with respect to $\mathcal{M}$ with parameters $(\mathcal{O}(k\gamma/W_j), t)$.

□

When the mixture is not covariance separated and nearly isotropic, we can obtain a small list of hypotheses such that one of them is close to the true parameters, via tensor decomposition.

Lemma 3.8.7 (Mixture is List-decodable)**.** *Given $0 < \epsilon < 1/k^{k^{O(k^2)}}$ let $\alpha = \epsilon^{1/(45C^{k+1}(k+1)!)}$. Let $\mathcal{M} = \sum_{i=1}^{k} w_i G_i$ with $G_i = \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with mixture mean μ and mixture covariance Σ, such that $\|\mu\|_2 \leqslant \mathcal{O}\left(\sqrt{\epsilon/\alpha}\right)$, $\|\Sigma - I\|_F \leqslant \mathcal{O}(\sqrt{\varepsilon}/\alpha)$, $w_i \geqslant \alpha$ for all $i \in [k]$, and $\|\Sigma_i - \Sigma_j\|_F^2 \leqslant 1/\alpha^5$ for any pair of components, and let X be a set of points satisfying Condition 3.2.45 with respect to $\mathcal{M}$ for some parameters $\gamma = \varepsilon d^{-8k} k^{-Ck}$, for a sufficiently large constant C, and $t = 8k$. Let Y be an ϵ-corrupted version of X of size n, Algorithm 88 outputs a list L of hypotheses of size $O((1/\varepsilon)^{4k^2})$ in time poly$(|L|, n)$ such that if we choose a hypothesis $\{\hat{\mu}_i, \hat{\Sigma}_i\}_{i \in [k]}$ uniformly at random, $\|\mu_i - \hat{\mu}_i\|_2 \leqslant \mathcal{O}\left(\epsilon^{1/(20C^{k+1}(k+1)!)}\right)$ and $\left\|\Sigma_i - \hat{\Sigma}_i\right\|_F \leqslant \mathcal{O}\left(\epsilon^{1/(20C^{k+1}(k+1)!)}\right)$ for all i with probability at least $O(\varepsilon^{4k^2})$.*

Proof. Recall we run Algorithm 88 on the samples Y, the number of clusters k, the fraction of outliers ε and the minimum weight $\alpha = \epsilon^{1/(20C^{k+1}(k+1)!)}$. Next, we show that the preconditions of Theorem 86 are satisfied. First, the upper bounds on $\|\mu\|_2$ and $\|\Sigma - I\|_F$ imply $\sum_{i \in k} w_i \left(\Sigma_i + \mu_i \mu_i^\top\right) = \Sigma + \mu\mu^\top \preceq (1 + \mathcal{O}(\sqrt{\varepsilon}/\alpha))I$. Since the LHS is a conic combination of PSD matrices, it follows that for all $i \in [k]$, $\mu_i \mu_i^\top \preceq \frac{1}{\alpha}\left(1 + \mathcal{O}(\sqrt{\varepsilon}/\alpha)\right) I$, and thus $\left\|\mu_i \mu_i^\top\right\|_F \leqslant \frac{2}{\alpha}$. Next, we can write:

$$\begin{aligned}
\|\Sigma_i - I\|_F &\leqslant \|\Sigma_i - (\Sigma + \mu\mu^\top)\|_F + \|\Sigma - I\|_F + \|\mu\mu^\top\|_F \\
&= \left\|\Sigma_i - \sum_{j \in [k]} w_j \left(\Sigma_j + \mu_j \mu_j^\top\right)\right\|_F + \frac{\sqrt{\varepsilon} k}{\alpha} + \frac{\varepsilon}{\alpha} \\
&\leqslant \left\|\sum_{j \in [k]} w_j \left(\Sigma_i - \Sigma_j\right)\right\|_F + \frac{2}{\alpha} + \frac{\sqrt{\varepsilon} k}{\alpha} + \frac{\varepsilon}{\alpha} \\
&\leqslant \frac{2}{\alpha^{5/2}} ,
\end{aligned}$$

where the first and the third inequalities follow from the triangle inequality and the upper bound on $\left\|\mu_i \mu_i^\top\right\|_F$, and the last inequality follows from the assumption that $\|\Sigma_i - \Sigma_j\|_F^2 \leqslant 1/\alpha^5$ for every pair of covariances Σ_i, Σ_j. So, we can set $\Delta = 2\alpha^{-5/2}$ in Theorem 86. Then, given the

definition of α, we have that

$$\eta = 2k^{4k}\mathcal{O}(1+\Delta/\alpha)^{4k}\sqrt{\varepsilon} = \mathcal{O}\Big(\epsilon^{2/5}\Big)$$

and $1/\varepsilon^2 \geqslant \log(1/\eta)(k+1/\alpha+\Delta)^{4k+5}/\eta^2$. Therefore, Algorithm 88 outputs a list L of hypotheses such that $|L| = \exp(1/\varepsilon^2)$, and with probability at least 0.99, L contains a hypothesis that satisfies the following: for all $i \in [k]$,

$$\|\hat{\mu}_i - \mu_i\|_2 = \mathcal{O}\Big(\frac{\Delta^{1/2}}{\alpha}\Big)\eta^{G(k)} = \mathcal{O}\Big(\epsilon^{-1/(20C^{k+1}(k+1)!)}\cdot\varepsilon^{1/(10C^{k+1}(k+1)!)}\Big) = \mathcal{O}\Big(\epsilon^{1/(20C^{k+1}(k+1)!)}\Big) \text{ and}$$
$$\big\|\hat{\Sigma}_i - \Sigma_i\big\|_F = \mathcal{O}\Big(k^4\Big)\frac{\Delta^{1/2}}{\alpha}\eta^{G(k)} = \mathcal{O}\Big(\epsilon^{1/(20C^{k+1}(k+1)!)}\Big)\ . \tag{3.67}$$

Then if we choose a hypothesis in L uniformly at random, the probability that we choose the hypothesis satisfying (3.47) is at least $1/|L| = \exp(-1/\varepsilon^2)$. □

3.9 Robust Parameter Recovery: Proof of Theorem 69

In order to show that our algorithm recovers the individual components and the parameters, we will prove the following identifiability theorem. Without any assumption on the mixtures, it is impossible to distinguish components within ε total variation distance with ε-fraction of noise. So given two mixtures of Gaussians with ε total variation distance, the theorem shows that there exist two partitions of components of the two mixtures respectively such that any two components in the matched pair is are $\mathrm{poly}(\varepsilon)$-close in total variation distance.

Theorem 89 (Identifiability)**.** *Let $\mathcal{M} = \sum_{i=1}^{k_1} w_i G_i, \mathcal{M}' = \sum_{i=1}^{k_2} w_i' G_i'$ be two mixtures of Gaussians such that $d_{\mathsf{TV}}(\mathcal{M},\mathcal{M}') \leqslant \epsilon$. Then there exists a partition of $[k_1]$ into sets $R_0, R_1, \ldots, R_\ell$ and a partition of $[k_2]$ into sets $S_0, S_1, \ldots, S_\ell$ such that*

1. *Let $W_i = \sum_{j\in R_i} w_j$ for $i = 0,1,\ldots,k_1$, $W_i' = \sum_{j\in S_i} w_j'$ for $i = 0,1,\ldots,k_2$. Then for all $i \in [\ell]$,*

$$|W_i - W_i'| \leqslant \mathit{poly}_k(\epsilon)$$
$$d_{\mathsf{TV}}(G_j, G_{j'}') \leqslant \mathit{poly}_k(\epsilon) \quad \forall j \in R_i, j' \in S_i$$

2. $W_0, W_0' \leqslant \mathit{poly}_k(\epsilon)$.

Corollary 3.9.1. *There is an algorithm with the following behavior: Given $\varepsilon > 0$ and a multiset of $n = d^{O(k)}\mathit{poly}(\varepsilon)$ samples from a distribution F on $\mathcal{R}^d$ such that $d_{\mathsf{TV}}(F, \mathcal{M}) \leqslant \varepsilon$, for an unknown target k-GMM $\mathcal{M} = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, the algorithm runs in time $d^{O(k)}\mathit{poly}_k(1/\varepsilon)$ and outputs a k'-GMM hypothesis $\widehat{\mathcal{M}} = \sum_{i=1}^{k'} \hat{w}_i \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i)$ with $k' \leqslant k$ such that with high probability there exists a partition of $[k]$ into $k'+1$ sets $R_0, R_1, \ldots, R_{k'}$ such that*

1. *Let $W_i = \sum_{j \in R_i} w_j$. Then for all $i \in [k']$,*

$$|W_i - \hat{w}_i| \leqslant \mathit{poly}_k(\epsilon)$$
$$d_{\mathsf{TV}}(\mathcal{N}(\mu_j, \Sigma_j), \mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i)) \leqslant \mathit{poly}_k(\epsilon) \quad \forall j \in R_i$$

2. *The sum of weights of exceptional components in R_0 is at most $\mathit{poly}_k(\epsilon)$.*

Parameter estimation is implied by TV distance for individual Gaussians (in relative Frobenius norm). The corollary follows immediately from the identifiability theorem.

Outline of Proof. The first step is to deal with the components in $\mathcal{M}$ and $\mathcal{M}'$ with small weights. We will construct $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$ by removing components with small weights. If we prove the statement on $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$, we can then deduce the theorem in the general case with worse, but still polynomial dependencies on ε. The second step is a partial clustering, after which the components within each cluster have TV distance bounded by $1 - \mathrm{poly}(\varepsilon)$. We prove this lemma in a separate section. After that we modify the parameters slightly so that the resulting parameters for different components are either identical or have a minimum separation. After this, we can use a lemma from [LM21] that provides a 1-1 mapping between the components of two such mixtures with small TV distance such that the mapped pairs have small TV distance.

Distance between Gaussians. We use the following facts for Gaussian distributions.

Lemma 3.9.2 (Frobenius Distance to TV Distance)**.** *Suppose $N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)$ are Gaussians with $\|\mu_1 - \mu_2\|_2 \leqslant \delta$ and $\|\Sigma_1 - \Sigma_2\|_F \leqslant \delta$. If the eigenvalues of Σ_1 and Σ_2 are at least $\lambda > 0$, then $d_{\mathsf{TV}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)) = O(\delta/\lambda)$.*

Lemma 3.9.3 (Lemma 5.4 in [LM21])**.** *Let $\mathcal{M}$ be a mixture of k Gaussians that is connected if we draw edges between all components i, j in $\mathcal{M}$ such that $d_{\mathsf{TV}}(G_i, G_j) \leqslant 1 - \delta$. Let Σ be the covariance matrix of $\mathcal{M}$. Then for any components Σ_i of the mixture*

1. $\Sigma_i \succeq \mathit{poly}_k(\delta)\Sigma$
2. $\|\Sigma^{-1/2}(\Sigma - \Sigma_i)\Sigma^{-1/2}\|_F \leqslant \mathit{poly}_k(\delta)^{-1}$.

The proof is identical to Lemma 5.4 in [LM21]. The only difference is that in [LM21] the authors assume that the minimal weight of $\mathcal{M}$ is at least δ and TV distance between any pair of components is at least δ but here we do not need these two assumptions, which does not affect the proof.

Fact 3.9.4 (Claim 3.9 in [LM21])**.** *Let ∂ denote the differential operator with respect to y. If*

$$f(y) = P(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where P is a polynomial in y of degree k (whose coefficients are polynomials in X) and $a(X), b(X)$ are polynomials in X then

$$(\partial - (a(X) + yb(X)))f(y) = Q(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where Q is a polynomial in y with degree exactly $k - 1$ whose leading coefficient is k times the leading coefficient of P.

Fact 3.9.5 (Corollary 3.10 in [LM21])**.** *Let ∂ denote the differential operator with respect to y. If*

$$f(y) = P(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where P is a polynomial in y of degree k then

$$(\partial - (a(X) + yb(X)))^{k+1} f(y) = 0.$$

Fact 3.9.6 (Claim 3.11 in [LM21])**.** *Let ∂ denote the differential operator with respect to y. If*

$$f(y) = P(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where P is a polynomial in y of degree k. Let the leading coefficient of P (viewed as a polynomial in y) be $L(X)$. Let $c(X)$ be a linear polynomial in X and $d(X)$ be a quadratic polynomial in X such that $\{a(X), b(X)\} \neq \{c(X), d(X)\}$. If $b(X) \neq d(X)$ then

$$(\partial - (c(X) + yd(X)))^{k'} f(y) = Q(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where Q is a polynomial of degree $k + k'$ in y with leading coefficient

$$L(X)(b(X) - d(X))^{k'}$$

and if $b(X) = d(X)$ then

$$(\partial - (c(X) + yd(X)))^{k'} f(y) = Q(y, X) \exp\left(a(X)y + \frac{1}{2}b(X)y^2\right)$$

where Q is a polynomial of degree k in y with leading coefficient

$$L(X)(a(X) - c(X))^{k'}.$$

Lemma 3.9.7. *Let $\mathcal{M} = \sum_{i=1}^{k_1} w_i G_i, \mathcal{M}' = \sum_{i=1}^{k_2} w'_i G'_i$ be two mixtures of Gaussians such that $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leqslant \epsilon$. For any constant $0 < c_1 < 1$, there exists $i \in [k_1 + k_2 + 1]$ such that $w_j, w'_{j'} \notin [\varepsilon^{c_1^{i-1}}, \varepsilon^{c_1^i})$ for any $j \in [k_1], j' \in [k_2]$. Moreover, if*

$$\tilde{\mathcal{M}} = \frac{\sum_{\{j:w_j \geqslant \varepsilon^{c_1^i}\}} w_j G_j}{\sum_{\{j:w_j \geqslant \varepsilon^{c_1^i}\}} w_j}$$
$$\tilde{\mathcal{M}}' = \frac{\sum_{\{j:w'_j \geqslant \varepsilon^{c_1^i}\}} w'_j G'_j}{\sum_{\{j:w'_j \geqslant \varepsilon^{c_1^i}\}} w'_j}$$

then $d_{\mathsf{TV}}(\tilde{\mathcal{M}}, \tilde{\mathcal{M}}') \leqslant O_k(\varepsilon^{c_1^{i-1}})$.

Proof. We can see that $[\varepsilon^{c_1^{i-1}}, \varepsilon^{c_1^i})$ with $i \in [k_1 + k_2 + 1]$ are $k_1 + k_2 + 1$ disjoint intervals and $w_j, w'_{j'}$ with $j \in [k_1], j' \in [k_2]$ have at most $k_1 + k_2$ distinct values. So there is one interval containing no weights.

We then construct $\tilde{\mathcal{M}}$ by removing the small components in $\mathcal{M}$. The sum of weights removed is at most $k\varepsilon^{c_1^{i-1}}$. So $d_{\mathsf{TV}}(\mathcal{M}, \tilde{\mathcal{M}}) \leqslant k\varepsilon^{c_1^{i-1}}$. Similarly, we have $d_{\mathsf{TV}}(\mathcal{M}', \tilde{\mathcal{M}}') \leqslant k\varepsilon^{c_1^{i-1}}$. By the triangle inequality,

$$d_{\mathsf{TV}}(\tilde{\mathcal{M}}, \tilde{\mathcal{M}}') \leqslant d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') + d_{\mathsf{TV}}(\mathcal{M}, \tilde{\mathcal{M}}) + d_{\mathsf{TV}}(\mathcal{M}', \tilde{\mathcal{M}}') \leqslant O_k(\varepsilon^{c_1^{i-1}}).$$

□

Lemma 3.9.7 shows that we can remove components with tiny weights in the mixtures. So in the following lemma, we will assume M and M' are Gaussian mixtures with minimal weights

at least poly(ε). We will show that we can partition the union of components of two mixtures so that if we prove Theorem 89 for each part of the partition, we can combine them to prove Theorem 89 on the full mixtures.

Lemma 3.9.8. *For any constant $0 < c_3 < 1$, there exist $c_1, c_2 > 0$ that depend on k and c_3, such that if $\mathcal{M} = \sum_{i=1}^{k_1} w_i G_i, \mathcal{M}' = \sum_{i=1}^{k_2} w_i' G_i'$ with $k_1, k_2 \leqslant k$, $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leqslant \varepsilon$ and $w_i, w_i' \geqslant \varepsilon^{c_1}$ for all i, then there exists a partition of $[k_1]$ into sets $R_1, \ldots, R_\ell$ and a partition of $[k_2]$ into sets $S_1, \ldots, S_\ell$ such that*

1. *For all $i \in [\ell]$, let $W_i = \sum_{j \in R_i} w_j, W_i' = \sum_{j \in S_i} w_j'$ be the sum of weights in each piece. Let $\mathcal{M}_i = \frac{1}{W_i} \sum_{j \in R_i} w_j G_j, \mathcal{M}_i' = \frac{1}{W_i'} \sum_{j \in S_i} w_j' G_j'$ be the submixtures of Gaussians after partition. Then for all $i \in [\ell]$,*

$$|W_i - W_i'| \leqslant \mathit{poly}_k(\epsilon)$$
$$d_{\mathsf{TV}}(\mathcal{M}_i, \mathcal{M}_i') \leqslant O_k(\varepsilon^{c_2})$$

2. *Consider the graph with vertices corresponding to components in $\mathcal{M}$ and $\mathcal{M}'$ and two components are adjacent if the total variation distance between them is at most $1 - \varepsilon^{c_2 c_3}$. Then the induced subgraph of vertices with indices $R_i \cup S_i$ is connected for all $i \in [\ell]$.*

The proof of Lemma 3.9.8 is deferred to Section 3.9.1. In the following two lemmas, we then prove Theorem 89 for each pair $\mathcal{M}_i, \mathcal{M}_i'$ defined in Lemma 3.9.8. In Lemma 3.9.9, we construct two mixtures of which pairs of parameters are identical or separated. We also shows it suffices to work under this simplification.

Lemma 3.9.9. *For any constant $0 < c_4 < 1$, there exist c_3, c_5 that depend on k and c_4, such that if $\mathcal{M} = \sum_{i=1}^{k_1} w_i G_i, \mathcal{M}' = \sum_{i=1}^{k_2} w_i' G_i'$ with $k_1, k_2 \leqslant k$ and*

1. *$\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$ is isotropic,*
2. *$d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leqslant \varepsilon$,*
3. *$w_i, w_i' \geqslant \varepsilon^{c_3}$ for all i,*
4. *Let $\mathcal{G}$ be a graph with components G_i, G_i' in $\mathcal{M}$ and $\mathcal{M}'$ as vertex set and two components are adjacent if the total variation distance between them is at most $1 - \varepsilon^{c_3}$. Then $\mathcal{G}$ is connected*

then there exist two mixtures of Gaussians $\tilde{\mathcal{M}} = \sum_{i=1}^{\tilde{k}_1} \tilde{w}_i \tilde{G}_i, \tilde{\mathcal{M}}' = \sum_{i=1}^{\tilde{k}_2} \tilde{w}_i' \tilde{G}_i'$ such that

1. *Any pair in* $\{\tilde{\mu}_i\} \cup \{\tilde{\mu}'_i\}$ *is either identical or separated by at least* $\varepsilon^{c_4 c_5}$
2. *Any pair in* $\{\tilde{\Sigma}_i\} \cup \{\tilde{\Sigma}'_i\}$ *is either identical or separated by at least* $\varepsilon^{c_4 c_5}$ *in Frobenius norm.*
3. $\|\mathbf{E}(h_m(\tilde{\mathcal{M}})) - \mathbf{E}(h_m(\tilde{\mathcal{M}}'))\|_F \leqslant O_k(\varepsilon^{c_5})$ *for any* $m \leqslant O(k)$
4. *There exist* $\pi_1 : [k_1] \to [\tilde{k}_1]$ *and* $\pi_2 : [k_2] \to [\tilde{k}_2]$ *such that*

$$\sum_{i:\pi_1(i)=j} w_i = \tilde{w}_j, \sum_{i:\pi_2(i)=j} w'_i = \tilde{w}'_j,$$
$$d_{\mathsf{TV}}(G_i, \tilde{G}_{\pi_1(i)}) \leqslant \mathit{poly}_k(\varepsilon), \quad \textit{for all } i \in [k_1]$$
$$d_{\mathsf{TV}}(G'_i, \tilde{G}'_{\pi_2(i)}) \leqslant \mathit{poly}_k(\varepsilon), \quad \textit{for all } i \in [k_2].$$

Proof. For any $0 < c_4 < 1$, there is $\ell \in [k^2]$ such that the distance between any pair of parameters in $\{\mu_i\} \cup \{\mu'_i\}$ or the Frobenius distance between any pair in $\{\Sigma_i\} \cup \{\Sigma'_i\}$ is not in the interval $[\varepsilon^{(c_4/2)^{\ell-1}}, \varepsilon^{(c_4/2)^{\ell}})$.

Now consider a graph $\mathcal{G}$ on $k_1 + k_2$ nodes where each node represents a vector in $\{\mu_i\} \cup \{\mu'_i\}$ and two vectors a, b are adjacent if

$$\|a - b\| \leqslant \varepsilon^{(c_4/2)^{\ell-1}}.$$

We now construct new mixtures $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$. For each connected component in $\mathcal{G}$ say $\{\mu_{i_1}, \ldots, \mu'_{j_1}, \ldots\}$, pick a representative say μ_{i_1} and set $\tilde{\mu}_{i_1} = \cdots = \tilde{\mu}'_{j_1} = \cdots = \mu_{i_1}$. Do this for all connected components and similar in the graph on covariance matrices with edges (i, j) if

$$\|\Sigma_i - \Sigma_j\|_F \leqslant \varepsilon^{(c_4/2)^{\ell-1}}.$$

After replacing close parameters with a representative, we may get some exactly same components in each new mixture. We then merge components with same means and covariances by adding their weights. Since all representatives of means and covariances are in different connected components of the graphs, they are separated by at least $\varepsilon^{(c_4/2)^{\ell}}$. Setting $c_5 = 1/2(c_4/2)^{\ell-1}$ gives a separation of $\varepsilon^{c_4 c_5}$.

Next we prove 3. There is a natural mapping $\pi_1 : [k_1] \to [\tilde{k}_1]$ that maps any component in $\mathcal{M}$ to the merged component in $\tilde{\mathcal{M}}$ and a similar mapping $\pi_2 : [k_2] \to [\tilde{k}_2]$ for $\mathcal{M}', \tilde{\mathcal{M}}'$. For all i, we have

$$\|\tilde{\mu}_{\pi_1(i)} - \mu_i\|, \|\tilde{\mu}'_{\pi_2(i)} - \mu'_i\|, \|\tilde{\Sigma}_{\pi_1(i)} - \Sigma_i\|_F, \|\tilde{\Sigma}'_{\pi_2(i)} - \Sigma'_i\|_F \leqslant O_k(1)\varepsilon^{(c_4/2)^{\ell-1}} \tag{3.68}$$

because for any pair of parameters above say $\tilde{\mu}_{\pi_1(i)}$ and μ_i, there is a path of length at most $2k$ connecting μ_i to the representative of the connected component, and each edge connects a pair with TV distance at most ε. Suppose $\|\mu_i\|, \|\Sigma_i - I\|_F \leqslant \Delta$. Then by Definition 3.2.4, we have for any integer m,

$$\|\mathbf{E}(h_m(\mathcal{M})) - \mathbf{E}(h_m(\tilde{\mathcal{M}}))\|_F \leqslant O_k(m)\Delta^m \varepsilon^{(c_4/2)^{\ell-1}}.$$

Since $\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$ is isotropic and the minimum weight in $\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$ is at least $\frac{1}{2}\varepsilon^{c_3}$, we have $\|\mu_i\| \leqslant \sqrt{2/\varepsilon^{c_3}}$ for all i. Applying Lemma 3.9.3 to $\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$, we have $\|I - \Sigma_i\|_F \leqslant \mathrm{poly}_k(\varepsilon^{c_3})^{-1}$. So there is a constant a such that $\Delta \leqslant \varepsilon^{-ac_3}$. If we take $c_3 > 0$ so that $ac_3 O(k) \leqslant 1/2(c_4/2)^{\ell-1}$ and take $c_5 = 1/2(c_4/2)^{\ell-1}$, then

$$\|\mathbf{E}(h_m(\mathcal{M})) - \mathbf{E}(h_m(\tilde{\mathcal{M}}))\|_F \leqslant O_k(m)\varepsilon^{(c_4/2)^{\ell-1} - O(m)ac_3} = O_k(\varepsilon^{c_5})$$

for $m \leqslant O(k)$. By the same argument, we have the similar inequality for $\mathcal{M}'$ and $\tilde{\mathcal{M}}'$

$$\|\mathbf{E}(h_m(\mathcal{M}')) - \mathbf{E}(h_m(\tilde{\mathcal{M}}'))\|_F = O_k(\varepsilon^{c_5}).$$

Since we can use Proposition 3.3 to robustly estimate the Hermite tensors of a Gaussian mixture with ε-fraction of noise and $\mathrm{poly}(\varepsilon)$ error guarantee, we must have

$$\|\mathbf{E}(h_m(\mathcal{M})) - \mathbf{E}(h_m(\mathcal{M}'))\|_F \leqslant \mathrm{poly}_k(\varepsilon).$$

Then by the triangle inequality,

$$\begin{aligned}\|\mathbf{E}(h_m(\tilde{\mathcal{M}})) - \mathbf{E}(h_m(\tilde{\mathcal{M}}'))\|_F \leqslant \|\mathbf{E}(h_m(\mathcal{M})) - \mathbf{E}(h_m(\tilde{\mathcal{M}}))\|_F + \\ \|\mathbf{E}(h_m(\mathcal{M})) - \mathbf{E}(h_m(\mathcal{M}'))\|_F + \|\mathbf{E}(h_m(\mathcal{M}')) - \mathbf{E}(h_m(\tilde{\mathcal{M}}'))\|_F = O(\varepsilon^{c_5}).\end{aligned}$$

For the last conclusion, from the definition of π_1 and π_2, we know that

$$\sum_{i:\pi_1(i)=j} w_i = \tilde{w}_j, \sum_{i:\pi_2(i)=j} w'_i = \tilde{w}'_j.$$

Applying Lemma 3.9.3 to $\frac{1}{2}\mathcal{M}+\frac{1}{2}\mathcal{M}'$, we have that eigenvalues of Σ_i and Σ'_i are at least $\mathrm{poly}(\varepsilon^{c_3})$ for all i. Then if c_3 is sufficiently small, by Lemma 3.9.2, (3.68) implies $d_{\mathsf{TV}}(G_i, \tilde{G}_{\pi(i)}) \leqslant \mathrm{poly}_k(\varepsilon)$ and $d_{\mathsf{TV}}(G'_i, \tilde{G}'_{\pi(i)}) \leqslant \mathrm{poly}_k(\varepsilon)$ for all i. □

The following lemma shows the identifiability under the simplification of Lemma 3.9.9. It is proved in the proof of Lemma 8.2 in [LM21].

Lemma 3.9.10. *Suppose $\mathcal{M} = \sum_{i=1}^{k_1} w_i G_i, \mathcal{M}' = \sum_{i=1}^{k_2} w_i' G_i'$ satisfies 1,2,3 in the conclusion of Lemma 3.9.9 with constants c_4, c_5 and the minimal weights are at least ε^{c_3}. There exists a sufficiently small function $f(k) > 0$ depending only on k such that if $c_4 \leqslant f(k)$, then $k_1 = k_2$ and there exists a permutation π such that $|w_i - w'_{\pi(i)}| \leqslant \mathrm{poly}_k(\varepsilon)$ and $G_i = G'_{\pi(i)}$.*

Proof. Consider the component $G'_{k_2} = N(\mu'_{k_2}, \Sigma'_{k_2})$ in $\mathcal{M}'$. We claim that there must be some $i \in [k_1]$ such that

$$(\mu_i, \Sigma_i) = (\mu'_{k_2}, \Sigma'_{k_2}).$$

Assume for the sake of contradiction that this is not the case. Let $S_1 = \{i \in [k_1] : \Sigma_i = \Sigma'_{k_2}\}$ and $S_2 = \{i \in [k_2 - 1] : \Sigma'_i = \Sigma'_{k_2}\}$. Suppose F, F' are the generating functions of $\mathcal{M}$ and $\mathcal{M}'$

$$F = \sum_{i=1}^{k_1} w_i \exp\left(\mu_i^T X + \frac{1}{2} X^T \Sigma_i X y^2\right) = \sum_{m=0}^{\infty} \frac{1}{m!} h_m(\mathcal{M}) y^n$$
$$F' = \sum_{i=1}^{k_2} w_i' \exp\left(\mu_i'^T X + \frac{1}{2} X^T \Sigma_i' X y^2\right) = \sum_{m=0}^{\infty} \frac{1}{m!} h_m(\mathcal{M}') y^n.$$

Then define the differential operators

$$\mathcal{D}_i = \partial - \mu_i^T X - X^T \Sigma_i X y$$
$$\mathcal{D}'_i = \partial - \mu_i'^T X - X^T \Sigma'_i X y$$

where partial derivatives are taken with respect to y. Now consider the differential operator

$$\mathcal{D} = (\mathcal{D}'_{k_2-1})^{2^{k_1+k_2-2}} \cdots (\mathcal{D}'_1)^{2^{k_1}} \mathcal{D}_{k_1}^{2^{k_1-1}} \cdots \mathcal{D}_1$$

By Fact 3.9.5, $\mathcal{D}(F) = 0$. By Fact 3.9.5 and Fact 3.9.6, we have

$$\mathcal{D}(F') = P(y, X) \exp\left(\mu_{k_2}'^T X + \frac{1}{2} X^T \Sigma'_{k_2} X y^2\right)$$

where P is a polynomial of degree

$$\deg(P) = 2^{k_1+k_2-1} - 1 - \sum_{i \in S_1} 2^{i-1} - \sum_{i \in S_2} 2^{k_1+i-2}$$

with leading coefficient

$$C_0 = w'_{k_2} \prod_{i\in[k_1]\setminus S_1} (X^T(\Sigma'_{k_2} - \Sigma_i)X)^{2^{i-1}} \prod_{i\in S_1} ((\mu'_{k_2} - \mu_i)^T X)^{2^{i-1}} \\ \prod_{i\in[k_2-1]\setminus S_2} (X^T(\Sigma'_{k_2} - \Sigma'_i)X)^{2^{k_1+i-2}} \prod_{i\in S_2} ((\mu'_{k_2} - \mu'_i)^T X)^{2^{k_1+i-2}}.$$

We now compare the following differentials evaluated at $y = 0$

$$(\mathcal{D}'_{k_2})^{\deg(P)}\mathcal{D}(F)$$
$$(\mathcal{D}'_{k_2})^{\deg(P)}\mathcal{D}(F')$$

The first quantity is 0 because $\mathcal{D}(F)$ is identically 0 as a formal power series. The second one is $\Omega_k(1)C_0$. Since for any i $(\mu_i, \Sigma_i) \neq (\mu'_{k_2}, \Sigma'_{k_2})$, our assumptions imply that the separation between μ_i, μ'_{k_2} or Σ_i, Σ'_{k_2} is at least $\varepsilon^{c_4c_5}$. Then we have $C_0 \geqslant \varepsilon^{c_4c_5O_k(1)}$ for some X. On the other hand, the coefficients of the formal power series F, F' are the Hermite polynomials $h_m(\mathcal{M})$ and $h_m(\mathcal{M}')$. This is a contradiction with our assumption that

$$\|\mathbf{E}(h_m(\mathcal{M}) - \mathbf{E}(h_m(\mathcal{M})\|_F \leqslant O_k(\varepsilon^{c_5})$$

as long as c_4 is smaller than some sufficiently small function $f(k)$ depending only on k. Thus there must be some component of $\mathcal{M}$ that matches $G'_{k_2} = N(\mu'_{k_2}, \Sigma'_{k_2})$. We can repeat the argument for each component in $\mathcal{M}'$ and in $\mathcal{M}$ to conclude that $\mathcal{M}$ and $\mathcal{M}'$ have the same components.

Next we will show that the weights of the same components in $\mathcal{M}$ and $\mathcal{M}'$ are close. We can assume that $\mathcal{M} = \sum_{i=1}^k w_i G_i, \mathcal{M}' = \sum_{i=1}^k w'_i G_i$ are two mixtures on the same set of components. Without loss of generality,

$$w_1 - w'_1 \leqslant \cdots w_\ell - w'_\ell \leqslant 0 \leqslant w_{\ell+1} - w'_{\ell+1} \leqslant \cdots \leqslant w_k - w'_k.$$

Then we can consider the following two mixtures

$$(w_1 - w'_1)G_1 + \cdots + (w_\ell - w'_\ell)G_\ell$$
$$(w_{\ell+1} - w'_{\ell+1})G_{\ell+1} + \cdots + (w_k - w'_k)G_k.$$

If

$$\sum_{i=1}^{k} |w_i - w_i'| > \varepsilon^{\zeta}$$

for some sufficiently small ζ depending only on k, we can then normalize each of the above into a distribution and repeat the same argument, using the fact that pairs of components cannot be too close, to obtain a contradiction. Thus, the mixing weights of $\mathcal{M}$ and $\mathcal{M}'$ are $\mathrm{poly}_k(\varepsilon)$-close and this completes the proof. □

Proof of Theorem 89. We first set $c_4 = f(k)$ as in Lemma 3.9.10, and then c_3, c_5 according to c_4 as in Lemma 3.9.9, and c_1', c_2 according to c_3 as in Lemma 3.9.8. Let $c_1 = \min\{c_1', c_2 c_3\}$.

By Lemma 3.9.7, we can find i such that there is no w_j, w_j' in $[\varepsilon^{c_1^{i-1}}, \varepsilon^{c_1^i})$. Let $\tilde{\mathcal{M}} = \sum_{\{j: w_j \geqslant \varepsilon^{c_1^i}\}} w_j G_j$ and $\tilde{\mathcal{M}}' = \sum_{\{j: w_j' \geqslant \varepsilon^{c_1^i}\}} w_j' G_j'$. Then $d_{\mathsf{TV}}(\tilde{\mathcal{M}}, \tilde{\mathcal{M}}') \leqslant O(\varepsilon^{c_1^{i-1}})$. Let $\varepsilon_1 = \varepsilon^{c_1^{i-1}}$. We have $d_{\mathsf{TV}}(\tilde{\mathcal{M}}, \tilde{\mathcal{M}}') \leqslant O(\varepsilon_1)$ and the minimum weights of $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$ are at least $\varepsilon_1^{c_1}$.

Now we can apply Lemma 3.9.8 on $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$ and get partitions of components of $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$. For $i \in [\ell]$, let $\mathcal{M}_i$ and $\mathcal{M}_i'$ be the mixtures defined in Lemma 3.9.8. We can apply a linear transformation to make $\frac{1}{2}\mathcal{M}_i + \frac{1}{2}\mathcal{M}_i'$ in isotropic position. Since the total variation distance is invariant under linear transformations, so we still have both conclusions in Lemma 3.9.8. Let $\varepsilon_2 = \varepsilon_1^{c_2}$. Then $d_{\mathsf{TV}}(\mathcal{M}_i, \mathcal{M}_{\pi(i)}) \leqslant O(\varepsilon_2)$ and $\frac{1}{2}\mathcal{M} + \frac{1}{2}\mathcal{M}'$ satisfies Lemma 3.9.3 with $\delta = \varepsilon_2^{c_3}$. Weights of both mixtures increase when we do the partition. So minimum weights are at least $\varepsilon_1^{c_1} \geqslant \varepsilon_1^{c_2 c_3} = \varepsilon_2^{c_3}$.

We now prove the statement on these smaller mixtures. First we can use Lemma 3.9.9 to merge close parameters of $\mathcal{M}_i, \mathcal{M}_i'$ so that all pairs of parameters are either equal or separated by $\varepsilon_2^{c_4 c_5}$. Under this simplification, Lemma 3.9.10 shows that there is a perfect matching between the same components in two mixtures and their weights are almost the same. By the last statement in Lemma 3.9.9, it is also a matching between components of $\mathcal{M}_i$ and $\mathcal{M}_i'$ by combining π and π_1, π_2. Moreover, if $\tilde{G}_j = \tilde{G}'_{\pi(j)}$, then $d_{\mathsf{TV}}(G_\ell, G'_\ell) \leqslant \mathrm{poly}(\varepsilon_2)$ for all ℓ, ℓ' such that $\pi_1(\ell) = j, \pi_2(\ell') = \pi(j)$. Repeating the argument for all pieces in $\tilde{\mathcal{M}}, \tilde{\mathcal{M}}'$ completes the proof. □

3.9.1 Proof of Lemma 3.9.8

In this section, we will prove Lemma 3.9.8. The following fact in [Liu-Moitra] shows that a good set of clusters of one mixture exists.

Fact 3.9.11 (Claim 7.6 in [LM'20]). *Let $\mathcal{M} = \sum_{i=1}^{k} w_i G_i$ be a mixture of Gaussians. For any*

constants $0 < \delta < 1$ and $\varepsilon > 0$, there exists $t \in [k^2]$ such that there exists a partition (possibly trivial) of $[k]$ into sets $R_1, \ldots, R_\ell$ such that

1. *If we draw edges between all pairs i, j such that $d_{\mathsf{TV}}(G_i, G_j) \leqslant 1 - \varepsilon^{\delta^t}$, then each piece of the partition is connected*
2. *For any i, j in different pieces of the partition, $d_{\mathsf{TV}}(G_i, G_j) \geqslant 1 - \varepsilon^{\delta^{t-1}}$.*

Remark 90. Fact 3.9.11 can be applied to a set of Gaussians instead of a mixture of Gaussians by randomly assigning positive weights for all Gaussians.

Lemma 3.9.12. *For any constant $0 < c < 1$, suppose $\mathcal{M} = \sum_{i=1}^{k_1} w_i A_i, \mathcal{M}' = \sum_{i=1}^{k_2} w_i' B_i$ are two mixtures of arbitrary distributions with $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \leqslant \varepsilon$ and $w_i, w_i' \geqslant \varepsilon^c$. If for any $i \neq j$, $d_{\mathsf{TV}}(A_i, B_j) \geqslant 1 - \varepsilon$, then $k_1 = k_2$ and $d_{\mathsf{TV}}(A_i, B_i) \leqslant \mathrm{poly}_{k_1}(\varepsilon)$ for all $i \in [k_1]$.*

Proof. Suppose π is any coupling of $\mathcal{M}$ and $\mathcal{M}'$ and X, Y are random variables with distributions $\mathcal{M}$ and $\mathcal{M}'$. Then $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') = \min_\pi\{\mathbf{Pr}_\pi(X \neq Y)\}$. We define π to be the optimal coupling such that $d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') = \mathbf{Pr}_\pi(X \neq Y)$. Then we can define $\hat{\pi}$ on variables i, j, X, Y such that $\sum_{i \in [k_1], j \in [k_2]} \hat{\pi}(i, j, X, Y) = \pi(X, Y)$ and the marginal distribution $\hat{\pi}_X$ with fixed i of X is $w_i A_i$ for all $i \in [k_1]$ and the marginal distribution $\hat{\pi}_Y$ with fixed j is $w_j' B_j$ for all $j \in [k_2]$. Let $P_{ij} = \int_{X,Y} \hat{\pi}(i, j, X, Y) dX dY$ and $A_{ij} = \frac{1}{P_{ij}} \int_Y \hat{\pi}(i, j, X, Y) dY$ be distributions on X, $B_{ij} = \frac{1}{P_{ij}} \int_X \hat{\pi}(i, j, X, Y) dX$ be distributions on Y. Then we have

$$\begin{aligned} d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') &= \mathbf{Pr}_\pi(X \neq Y) = \mathbf{Pr}_{\hat{\pi}}(X \neq Y) \\ &= \sum_{i,j} P_{ij} \mathbf{Pr}_{\hat{\pi}}(X \neq Y \mid i, j) \\ &\geqslant \sum_{i,j} P_{ij} \cdot d_{\mathsf{TV}}(A_{ij}, B_{ij}). \end{aligned} \tag{3.69}$$

By the definition of P_{ij}, A_{ij}, B_{ij},

$$\begin{aligned} w_i A_i &= P_{ij} A_{ij} + \sum_{j' \neq j} P_{ij'} A_{ij'} \\ w_i' B_i &= P_{ij} B_{ij} + \sum_{i' \neq i} P_{i'j} B_{i'j} \end{aligned}$$

Dividing both sides by $\max\{w_i, w'_j\}$, we get

$$A_i = \frac{P_{ij}}{\max\{w_i, w'_j\}} A_{ij} + \left(1 - \frac{w_i}{\max\{w_i, w'_j\}}\right) A_i + \sum_{j' \neq j} \frac{P_{ij'}}{\max\{w_i, w'_j\}} A_{ij'}$$
$$B_i = \frac{P_{ij}}{\max\{w_i, w'_j\}} B_{ij} + \left(1 - \frac{w'_j}{\max\{w_i, w'_j\}}\right) B_i + \sum_{i' \neq i} \frac{P_{i'j}}{\max\{w_i, w'_j\}} B_{i'j}$$

From the above two equations, we can write A_i, B_i as linear combinations of two distributions.

$$A_i = \frac{P_{ij}}{\max\{w_i, w'_j\}} A_{ij} + \left(1 - \frac{P_{ij}}{\max\{w_i, w'_j\}})\right) A'_i$$
$$B_i = \frac{P_{ij}}{\max\{w_i, w'_j\}} B_{ij} + \left(1 - \frac{P_{ij}}{\max\{w_i, w'_j\}}\right) B'_i$$

Then by the triangle inequality,

$$d_{\mathsf{TV}}(A_i, B_j) \leqslant \frac{P_{ij}}{\max\{w_i, w'_j\}} d_{\mathsf{TV}}(A_{ij}, B_{ij}) + \left(1 - \frac{P_{ij}}{\max\{w_i, w'_j\}}\right)$$

$$P_{ij} \cdot d_{\mathsf{TV}}(A_{ij}, B_{ij}) \geqslant P_{ij} - (1 - d_{\mathsf{TV}}(A_i, B_j)) \max\{w_i, w'_j\}. \tag{3.70}$$

Combining (3.69) and (3.70), we have the following inequality on the TV distance between mixtures and the TV distance between components

$$d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \geqslant \sum_{i,j} \left(P_{ij} - (1 - d_{\mathsf{TV}}(A_i, B_j)) \max\{w_i, w'_j\}\right). \tag{3.71}$$

By the lower bounds on $d_{\mathsf{TV}}(A_i, B_j)$, we have

$$\begin{aligned}
\epsilon \geqslant d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') &\geqslant \sum_{i,j} P_{ij} - \sum_{i \neq j} (1 - d_{\mathsf{TV}}(A_i, B_j)) \max\{w_i, w'_j\} - \sum_i (1 - d_{\mathsf{TV}}(A_i, B_i)) \max\{w_i, w'_i\} \\
&\geqslant 1 - \sum_{i \neq j} \varepsilon - \sum_i \max\{w_i, w'_i\} + \sum_i \max\{w_i, w'_i\} d_{\mathsf{TV}}(A_i, B_i) \\
&\geqslant 1 - \sum_{i \neq j} \varepsilon - \sum_i \max\{w_i, w'_i\} + w_{\min} d_{\mathsf{TV}}(A_1, B_1)
\end{aligned} \tag{3.72}$$

where A_1, B_1 can be replaced by any A_i, B_i pair. Let $k = \max\{k_1, k_2\}$. When $i \neq j$ and

$d_{\mathsf{TV}}(A_i, B_j) \geqslant 1 - \varepsilon$, we plug it into Equation (3.71) and get

$$\epsilon \geqslant d_{\mathsf{TV}}(\mathcal{M}, \mathcal{M}') \geqslant \sum_{i \neq j} P_{ij} - (1 - d_{\mathsf{TV}}(A_i, B_j)) \max\{w_i, w'_j\} \geqslant \sum_{i \neq j} (P_{ij} - \epsilon).$$

This implies

$$\sum_{i \neq j} P_{ij} \leqslant k^2 \varepsilon.$$

Then we can bound $\sum_i \max\{w_i, w'_i\} - 1$ in Equation (3.72)

$$\sum_i \max\{w_i, w'_i\} - 1 = \sum_i \max\{w_i, w'_i\} - w_i \leqslant \sum_{i \neq j} P_{ij} \leqslant k^2 \varepsilon.$$

Plugging this bound into Equation (3.72), for any i, we have

$$d_{\mathsf{TV}}(A_i, B_i) \leqslant \frac{1}{w_{\min}} \left(k^2 \varepsilon + \sum_i \max\{w_i, w'_i\} - 1 \right) \leqslant \frac{2k^2 \varepsilon}{\varepsilon^c}.$$

□

Proof of Lemma 3.9.8. We apply Fact 3.9.11 on the union set of components of $\mathcal{M}$ and $\mathcal{M}'$ with parameter δ to find a partition $R_1, \ldots, R_\ell$. Let

$$\mathcal{M}_i = \frac{\sum_{G_j \in R_i} w_j G_j}{\sum_{G_j \in R_i} w_j}$$
$$\mathcal{M}'_i = \frac{\sum_{G'_j \in R_i} w'_j G'_j}{\sum_{G'_j \in R_i} w'_j}.$$

Then for any $i \neq j$, we know $d_{\mathsf{TV}}(G_a, G'_b) \geqslant 1 - \varepsilon^{\delta^{t-1}}$ for $G_a \in R_i, G'_b \in R_j$. By (3.71) in the proof of Lemma 3.9.12, we have

$$d_{\mathsf{TV}}(M_i, M'_j) \geqslant 1 - 2k\varepsilon^{\delta^{t-1}}.$$

Then by Lemma 3.9.12, for any i, there exists a such that $d_{\mathsf{TV}}(M_i, M'_i) \leqslant \varepsilon^{a\delta^{t-1}}$. Let $c_2 = a\delta^{t-1}$. If we set $\delta = c_2 c_3 / \delta^{t-1} = a c_3$, the partition satisfies the second conclusion. □

3.10 Omitted Proofs

In this subsection, we provide the proofs that were omitted from Section 4.2 and Section 3.6.

3.10.1 Omitted Proofs from Section 3.2.1

Lemma 3.10.1 (Concentration of low-degree polynomials, Lemma 3.2.9 restated)**.** *Let T be a d-dimensional, degree-4 tensor such that $\|T\|_F \leqslant \Delta$ for some $\Delta > 0$ and let $x, y \sim \mathcal{N}(0, I)$. Then, with probability at least $1 - 1/poly(d)$, the following holds:*

$$\|T(\cdot, \cdot, x, y)\|_F^2 \leqslant \mathcal{O}\Big(\log(d)\Delta^2\Big) \ .$$

Proof. We note that

$$\begin{aligned}
\mathbf{E}\left[\|T(\cdot,\cdot,x,y)\|_F^2\right] &= \mathbf{E}\left[\sum_{i_1,i_2}\left(\sum_{i_3,i_4} T(i_1,i_2,i_3,i_4)\, x(i_3)\, y(i_4)\right)^2\right] \\
&= \mathbf{E}\left[\sum_{i_1,i_2}\left(\sum_{i_3,i_4} T(i_1,i_2,i_3,i_4)^2\, x(i_3)^2\, y(i_4)^2\right)\right] \\
&= \sum_{i_1,i_2,i_3,i_4} T(i_1,i_2,i_3,i_4)^2 \leqslant \Delta^2 \ .
\end{aligned}$$

The second equality follows from the fact that $x(i_3), y(i_4)$ are independent and have zero means. So the only non-zero terms are the squares. The third equality follows from the fact that $x(i_3), y(i_4)$ are independent with unit variances. Observe that $\|T(\cdot, \cdot, x, y)\|_F^2$ is a degree-2 polynomial in Gaussian random variables. Using standard concentration bounds for low-degree Gaussian polynomials, we obtain

$$\mathbf{Pr}\left[\|T(\cdot,\cdot,x,y)\|_F^2 \geqslant t^2 \mathbf{E}\left[\|T(\cdot,\cdot,x,y)\|_F^2\right]\right] \leqslant \exp(-ct) \ .$$

Setting $t = \Omega(\log(d))$ completes the proof. □

3.10.2 Omitted Proofs from Section 3.2.2

Lemma 3.10.2 (Spectral SoS Proofs, Lemma 3.2.23 restated)**.** *Let A be a $d \times d$ matrix. Then for d-dimensional vector-valued indeterminate v, we have:*

$$\left|\frac{v}{2}\right. \left\{v^\top A v \leqslant \|A\|_2 \|v\|_2^2\right\} .$$

Proof. Note that v is the only variable in the proof here (A is a matrix of constants). We note that $A \leqslant \|A\|_2 I$ or $\|A\|_2 I - A$ is PSD and thus $\|A\|_2 I - A = QQ^\top$ for some $d \times d$ matrix Q. Thus, $\|Qv\|_2^2 = v^\top(\|A\|_2 I - A)v = \|A\|_2 \|v\|_2^2 - v^\top A v$. Thus, $\|A\|_2 \|v\|_2^2 - v^\top A v$ is a sum of squares polynomial (namely $\|Qv\|_2^2$) in variable v. This completes the proof.

□

Lemma 3.10.3 (Frobenius Norms of Products of Matrices, Lemma 3.2.25 restated)**.** *Let B be a $d \times d$ matrix valued indeterminate for some $d \in \mathbb{N}$. Then, for any $0 \preceq A \preceq I$,*

$$\left|\frac{B}{2}\right. \left\{\|AB\|_F^2 \leqslant \|B\|_F^2\right\} ,$$

and,

$$\left|\frac{B}{2}\right. \left\{\|BA\|_F^2 \leqslant \|B\|_F^2\right\} ,$$

Proof. The proof of the second claim is similar so we prove only the first. We have:

$$\left|\frac{B}{2}\right. \left\{\|B\|_F^2 = \|(A + I - A)B\|_F^2 = \|AB\|_F^2 + \|(I - A)B\|_F^2 + 2\operatorname{tr}((I - A)BB^\top A)\right\}$$

Now, $A - A^2 \succeq 0$, thus, $A - A^2 = RR^\top$ for some $d \times d$ matrix R. Thus, $\operatorname{tr}((A - A^2)BB^\top) = \operatorname{tr}(RR^\top BB^\top) = \|BR\|_F^2$ - a sum of squares polynomial of degree 2 in indeterminate B. Thus, $\left|\frac{B}{2}\right. \left\{\operatorname{tr}((A - A^2)BB^\top) \geqslant 0\right\}$. □

3.10.3 Omitted Proofs from Section 3.2.3

Lemma 3.10.4 (Shifts Cannot Decrease Variance, Lemma 3.2.31 restated)**.** *Let $\mathcal{D}$ be a distribution on $\mathcal{R}^d$, Q be a $d \times d$ matrix-valued indeterminate, and C be a scalar-valued indeterminate. Then, we have that*

$$\left|\frac{Q,C}{2}\right. \left\{\mathbf{E}_{x\sim\mathcal{D}}\left[(Q(x) - \mathbf{E}_{x\sim\mathcal{D}}[Q(x)])^2\right] \leqslant \mathbf{E}_{x\sim\mathcal{D}}\left[(Q(x) - C)^2\right]\right\} .$$

Proof.

$$\left|\frac{Q,C}{2}\right\} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Q(x)-C)^2\right] = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x)-\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]+\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]-C\right)^2\right]$$
$$= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x)-\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)^2\right]+\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Q(x)-C)^2\right]$$
$$+2\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x)-\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]-C\right)\right]$$
$$= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x)-\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)^2\right]+\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(Q(x)-C)^2\right]$$
$$\geqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(Q(x)-\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[Q(x)]\right)^2\right] \quad \Big\}.$$

□

Lemma 3.10.5 (Shifts of Certifiably Hypercontractive Distributions, Lemma 3.2.32 restated)**.** *Let x be a mean-0 random variable with distribution $\mathcal{D}$ on $\mathcal{R}^d$ with t-certifiably C-hypercontractive degree-2 polynomials. Then, for any fixed constant vector $c \in \mathcal{R}^d$, the random variable $x+c$ also has t-certifiable $4C$-hypercontractive degree-2 polynomials.*

Proof. Observe that using that $\mathbb{E}_{x\sim\mathcal{D}}[x] = 0$, we have that

$$\left|\frac{Q}{2}\right\} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x+c)^\top Q(x+c)\right] = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx + c^\top Qc\right]\Big\}\ .$$

Next, by two applications of the SoS Triangle Inequality (Fact 3.2.21), an application of Lemma 3.2.31 followed by certifiable hypercontractivity of $\mathcal{D}$, we have:

$$\left|\frac{Q}{t'}\right\} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left((x+c)^\top Q(x+c)-\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x+c)^\top Q(x+c)\right]\right)^{t'}\right]$$
$$= \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(\left(x^\top Qx-\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx\right]\right)+x^\top Qc+c^\top Qx\right)^{t'}\right]$$
$$\leqslant 4^{t'}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx-\mathbf{E}_{\mathcal{D}}x^\top Qx\right)^{t'}\right]+\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x^\top Qc)^{t'}\right]+\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(c^\top Qx)^{t'}\right]\right)$$
$$\leqslant 4^{t'}(Ct')^{t'}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx-\mathbf{E}_{\mathcal{D}}x^\top Qx\right)^2\right]^{t'/2}+\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x^\top Qc)^2\right]^{t'/2}+\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(c^\top Qx)^2\right]^{t'/2}\right)\Big\}.$$

On the other hand, notice that

$$\left|\frac{Q}{2}\right\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left((x+c)^\top Q(x+c) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x+c)^\top Q(x+c)\right]\right)^2\right] \\ = \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx - \mathbf{E}_{\mathcal{D}} x^\top Qx\right)^2\right] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x^\top Qc)^2\right] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(c^\top Qx)^2\right]\right) \Big\}.$$

Thus,

$$\left|\frac{Q}{t'}\right\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left(x^\top Qx - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[x^\top Qx\right]\right)^2\right]^{t'/2} + \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x^\top Qc)^2\right]\right)^{t'/2} + \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(c^\top Qx)^2\right]\right)^{t'/2} \\ \leqslant 4^{t'}(Ct')^{t'}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left((x+c)^\top Q(x+c) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x+c)^\top Q(x+c)\right]\right)^2\right]\right)^{t'/2}\Big\}.$$

As a result, we obtain:

$$\left|\frac{Q}{t'}\right\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left((x+c)^\top Q(x+c) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x+c)^\top Q(x+c)\right]\right)^{t'}\right] \\ \leqslant (4Ct')^{t'}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left((x+c)^\top Q(x+c) - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(x+c)^\top Q(x+c)\right]\right)^2\right]\right)^{t'/2}\Big\},$$

which completes the proof. □

Lemma 3.10.6 (Mixtures of Certifiably Hypercontractive Distributions, Lemma 3.2.33 restated)**.** *Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ have t-certifiable C-hypercontractive degree-2 polynomials on $\mathcal{R}^d$, for some fixed constant C. Then, any mixture $\mathcal{D} = \sum_i w_i \mathcal{D}_i$ also has t-certifiably (C/α)-hypercontractive degree-2 polynomials for $\alpha = \min_{i\leqslant k, w_i>0} w_i$.*

Proof. Applying Lemma 3.2.21 followed by SoS Hölder's inequality on the second term and

followed by a final application of SoS Hölder's inequality (Fact 3.2.20), we obtain:

$$\begin{aligned}
\vdash_{t'}^{Q} \left\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[x^\top Q x \right] \right)^{t'} \right] \right. &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left(x^\top Q x - \sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[x^\top Q x \right] \right)^{t'} \right] \\
&= \sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[\left(x^\top Q x - \sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[x^\top Q x \right] \right)^{t'} \right] \\
&\leqslant 2^{t'} \left(\sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[x^\top Q x \right] \right)^{t'} \right] \right. \\
&\quad \left. + \left(\sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[x^\top Q x \right] \right]^{t'} \right) \right) \\
&\leqslant 2^{t'} \left((Ct')^{t'} \left(\sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[x^\top Q x \right] \right)^2 \right] \right)^{t'/2} \right. \\
&\quad \left. + \sum_i w_i \left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[x^\top Q x \right] \right)^2 \right] \right)^{t'} \right) \\
&\leqslant \left(\frac{4Ct'}{\alpha} \right)^{t'} \left(\sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[x^\top Q x \right] \right)^2 \right] \right)^{t'/2} \Bigg\}.
\end{aligned}$$

On the other hand, note that by Lemma 3.2.31, we know that

$$\begin{aligned}
\vdash_{2}^{Q} \left\{ \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} [x]^\top Q x \right)^2 \right] = \sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[x^\top Q x \right] \right)^2 \right] \right. \\
\left. \geqslant \sum_i w_i \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_i} \left[x^\top Q x \right] \right)^2 \right] \right\}.
\end{aligned}$$

Combining the two equations above completes the proof. □

Corollary 3.10.7 (Certifiable Hypercontractivity of k-Mixtures of Gaussians, Corollary 3.2.34 restated)**.** *Let $\mathcal{D}$ be a k-mixture of Gaussians $\sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ with weights $w_i \geqslant \alpha$ for every $i \in [k]$. Then, $\mathcal{D}$ has t-certifiably $4/\alpha$-hypercontractive degree-2 polynomials.*

Proof. From [KOTZ14], we know that the standard Gaussian random variable has t-certifiably 1-hypercontractive degree-2 polynomials. From Fact 3.2.30, we immediately obtain that for any PSD matrix Σ, the Gaussian $\mathcal{N}(0, \Sigma)$ also has t-certifiable 1-hypercontractive degree-2 polynomials. From Lemma 3.2.32, we obtain that for any μ, the Gaussian $\mathcal{N}(\mu, \Sigma)$ has t-certifiable 4-hypercontractive degree-2 polynomials. Finally, applying Lemma 3.2.33 to $\mathcal{D}_i = \mathcal{N}(\mu_i, \Sigma_i)$

and mixture weights $w_1, w_2, \dots w_k$, yields that $\mathcal{D} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ has t-certifiably $4/\alpha$-hypercontractive degree-2 polynomials. This completes the proof. □

Lemma 3.10.8 (Linear Transformations of Certifiably Bounded-Variance Distributions, Lemma 3.2.38 restated)**.** *For $d \in \mathbb{N}$, let x be a random variable with distribution $\mathcal{D}$ on $\mathcal{R}^d$ such that for $d \times d$ matrix-valued indeterminate Q, $\vdash_2^Q \left\{ \mathbf{E}_{x\sim\mathcal{D}}(x^\top Q x - \mathbf{E}_{\mathcal{D}} x^\top Q x)^2 \leqslant \left\|\Sigma^{1/2} Q \Sigma^{1/2}\right\|_F^2 \right\}$. Let A be an arbitrary $d \times d$ matrix and let $x' = Ax$ be the random variable with covariance $\Sigma' = AA^\top$. Then, we have that*

$$\vdash_2^Q \left\{ \mathbf{E}_{x'\sim\mathcal{D}'}(x'^\top Q x' - \mathbf{E}_{\mathcal{D}'} x'^\top Q x')^2 \leqslant \left\|\Sigma'^{1/2} Q \Sigma'^{1/2}\right\|_F^2 \right\} .$$

Proof. The covariance of x' is $AA^\top = \Sigma'$, say. Let $\Sigma'^{1/2}$ be the PSD square root of Σ'. The proof follows by noting that $x'^\top Q x' = (Ax)^\top Q (Ax) = x^\top (A^\top Q A) x^\top$ and that

$$\begin{aligned}
\left\|A^\top Q A\right\|_F^2 = \mathrm{tr}(A^\top Q A A^\top Q A) &= \mathrm{tr}(A A^\top Q A A^\top Q) = \mathrm{tr}(\Sigma' Q \Sigma' Q) \\
&= \mathrm{tr}(\Sigma'^{1/2} Q \Sigma'^{1/2} \Sigma'^{1/2} Q \Sigma'^{1/2}) \\
&= \left\|\Sigma'^{1/2} Q \Sigma'^{1/2}\right\|_F^2 .
\end{aligned}$$

□

Lemma 3.10.9 (Variance of Degree-2 Polynomials of Standard Gaussians, Lemma 3.2.39 restated)**.** *We have that*

$$\vdash_2^Q \left\{ \mathbf{E}_{\mathcal{N}(0,I)} \left(x^\top Q x - \mathbf{E}_{\mathcal{N}(0,I)} x^\top Q x \right)^2 \leqslant 3 \left\|Q\right\|_F^2 \right\} .$$

Proof. We will view $xx^\top$ and $I \in \mathcal{R}^{d\times d}$ as d^2-dimensional vectors. Consider the matrix $\mathbf{E}_{x\sim\mathcal{N}(0,I)}(xx^\top - I)(xx^\top - I)^\top$. The diagonal of this matrix is $2I_{d^2}$. The off-diagonal part has exactly one non-zero entry in any row (which corresponds to entry indexed by (i,j) and (j,i) for $i \neq j$), and thus has spectral norm at most 1 by the Gershgorin circle theorem. Thus, $\mathbf{E}_{x\sim\mathcal{N}(0,I)}(xx^\top - I)(xx^\top - I)^\top \preceq 3I_{d^2}$.

We thus have:

$$\left|\frac{Q}{2}\right\{ \mathbf{E}_{\mathcal{N}(0,I)}\left(x^\top Qx - \mathbf{E}_{\mathcal{N}(0,I)} x^\top Qx\right)^2 = \mathbf{E}_{\mathcal{N}(0,I)}\left\langle xx^\top - I, Q\right\rangle^2 = \mathbf{E}_{\mathcal{N}(0,I)}\left\langle xx^\top - I, Q\right\rangle^2$$

$$\leqslant \left\|\mathbf{E}_{x\sim\mathcal{N}(0,I)}(xx^\top - I)(xx^\top - I')^\top\right\|_2 \|Q\|_F^2 \leqslant 3\left\|I_{d^2}\right\|_2 \|Q\|_F^2 = 3\|Q\|_F^2\Big\}. \quad (3.73)$$

□

Lemma 3.10.10 (Variance of Degree-2 Polynomials of Mixtures, Lemma 3.2.41 restated)**.** *Let $\mathcal{M} = \sum_i w_i \mathcal{D}_i$ be a k-mixture of distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ with means μ_i and covariances Σ_i. Let $\mu = \sum_i w_i \mu_i$ be the mean of $\mathcal{M}$. Suppose that each of $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ have certifiably C-bounded-variance i.e. for Q: a symmetric $d \times d$ matrix-valued indeterminate.*

$$\left|\frac{Q}{2}\right\{ \mathbf{E}_{x'\sim\mathcal{D}_i}(x'^\top Qx' - \mathbf{E}_{\mathcal{D}_i} x'^\top Qx')^2 \leqslant C\left\|\Sigma'^{1/2} Q \Sigma'^{1/2}\right\|_F^2\Big\}.$$

Further, suppose that for some $H > 1$, $\|\mu_i - \mu\|_2^2, \|\Sigma_i - I\|_F \leqslant H$ for every $1 \leqslant i \leqslant k$. Then, we have that

$$\left|\frac{Q}{2}\right\{ \mathop{\mathbb{E}}_{x\sim\mathcal{M}}\left[\left(x^\top Qx - \mathop{\mathbb{E}}_{x\sim\mathcal{M}}\left[x^\top Qx\right]\right)^2\right] \leqslant 100CH^2\|Q\|_F^2\Big\}.$$

Proof. We have the following sequence of (in-)equalities:

$$\left|\frac{Q}{2}\right\{ \mathbf{E}_{x\sim\mathcal{M}}\left(x^\top Qx - \mathbf{E}_{x\sim\mathcal{M}} x^\top Qx\right)^2 = \sum_{i\leqslant k} w_i \mathbf{E}_{x\sim\mathcal{D}_i}\left(x^\top Qx - \mathbf{E}_{x\sim\mathcal{M}} x^\top Qx\right)^2 \quad (3.74)$$

$$= \sum_{i\leqslant k} w_i \mathbf{E}_{x\sim\mathcal{D}_i}\left(x^\top Qx - \mathbf{E}_{x\sim\mathcal{D}_i} x^\top Qx + \mathbf{E}_{x\sim\mathcal{D}_i} x^\top Qx - \mathbf{E}_{x\sim\mathcal{M}} x^\top Qx\right)^2 \quad (3.75)$$

$$\leqslant 2\sum_{i\leqslant k} w_i \mathbf{E}_{x\sim\mathcal{D}_i}\left(x^\top Qx - \mathbf{E}_{x\sim\mathcal{D}_i} x^\top Qx\right)^2 + 2\sum_i w_i\left(\mathbf{E}_{x\sim\mathcal{D}_i} x^\top Qx - \mathbf{E}_{x\sim\mathcal{M}} x^\top Qx\right)^2\Big\}, \quad (3.76)$$

where the third line follows from Fact 3.2.21 (SoS Almost Triangle Inequality).

Let us first bound the 2nd term in the RHS above. Towards that, let $\Sigma = \sum_i w_i((\mu_i - \mu)(\mu_i - \mu)^\top + \Sigma_i)$ be the covariance of the mixture $\mathcal{M}$. Then, notice that $\Sigma = \sum_i w_i((\mu_i - \mu)(\mu_i - \mu)^\top +$

$\Sigma_i) = \sum_i w_i \mu_i \mu_i^\top + \sum_i w_i \Sigma_i - \mu\mu^\top$. Thus, we can write

$$\begin{aligned}
\mu_i\mu_i^\top + \Sigma_i - \Sigma - \mu\mu^\top &= \sum_{j\neq i} w_j(\mu_i\mu_i^\top - \mu_j\mu_j^\top) + \sum_{j\neq i} w_j(\Sigma_i - \Sigma_j) \\
&= \sum_{j\neq i} w_j(\mu_j - \mu)(\mu_j - \mu)^\top - \sum_{j\neq i}(\mu - \mu_j)(\mu - \mu_j)^\top + \sum_{j\neq i} w_j(\Sigma_i - \Sigma_j) \\
&= \sum_{j\neq i} w_j(\mu_j - \mu)(\mu_j - \mu)^\top - \sum_{j\neq i}(\mu - \mu_j)(\mu - \mu_j)^\top \\
&\quad + \sum_{j\neq i} w_j(\Sigma_i - I) - \sum_{j\neq i} w_j(\Sigma_j - I)\,.
\end{aligned}$$

Here, in the second to last step, we added and subtracted $\sum_{j\neq i} w_j \mu\mu^\top$ and used that $\sum_i w_i\mu_i = \mu$, and in the last step we added and subtracted $\sum_{j\neq i} w_j I$.

By application of the triangle inequality for Frobenius norm to the RHS of the above, we have that:

$$\begin{aligned}
\left\|\mu_i\mu_i^\top + \Sigma_i - \mu\mu^\top - \Sigma\right\|_F &\leqslant \sum_{j\neq i} w_j \left\|(\mu_i - \mu)(\mu_i - \mu)^\top\right\|_F + \sum_{j\neq i} w_j \left\|(\mu_j - \mu)(\mu_j - \mu)^\top\right\|_F \\
&\quad + \sum_{j\neq i} w_j \left\|(\Sigma_i - I)\right\|_F + \sum_{j\neq i} w_j \left\|(I - \Sigma_j)\right\|_F \leqslant H + H + H + H = 4H\,.
\end{aligned}$$

Using the SoS version of the Cauchy-Schwarz inequality (Fact 3.2.20) on indeterminate Q and constant $\mu\mu^\top - \mu_i\mu_i^\top + \Sigma_i - \Sigma$ and the above bound, we have:

$$\begin{aligned}
\vdash_2^Q \Bigg\{ \sum_i w_i \left(\mathbf{E}_{x\sim\mathcal{D}_i} x^\top Q x - \mathbf{E}_{x\sim\mathcal{M}} x^\top Q x\right)^2 &= \sum_i w_i \left(\left\langle \mu_i\mu_i^\top + \Sigma_i, Q\right\rangle - \left\langle \mu\mu^\top + \Sigma, Q\right\rangle\right)^2 \\
\leqslant \sum_i w_i \left\|\mu\mu^\top - \mu_i\mu_i^\top + \Sigma_i - \Sigma\right\|_F^2 \|Q\|_F^2 &\leqslant 16H^2 \sum_i w_i \|Q\|_F^2 = 16H^2 \|Q\|_F^2 \Bigg\}\,.
\end{aligned}$$

Let us now bound the first term in the RHS of (3.76) above. First, observe that $x^\top Q x - \mathbf{E}_{\mathcal{N}(\mu_i,\Sigma_i)} x^\top Q x = (x-\mu_i)^\top Q(x-\mu_i) - \mathbf{E}_{\mathcal{N}(\mu_i,\Sigma_i)}(x-\mu_i)^\top Q(x-\mu_i) + 2(x-\mu_i)^\top Q\mu_i$.

Thus, using Fact 3.2.21 and Lemma 3.2.40, we have:

$$\left|\frac{Q}{2}\right\}\left\{\sum_{i\leqslant k} w_i \mathbf{E}_{\mathcal{D}_i}\left(x^\top Q x - \mathbf{E}_{x\sim\mathcal{D}_i} x^\top Q x\right)^2\right. \tag{3.77}$$

$$\leqslant 2\sum_{i\leqslant k} w_i \mathbf{E}_{\mathcal{D}_i}\left((x-\mu_i)^\top Q(x-\mu_i) - \mathbf{E}_{x\sim\mathcal{D}_i}(x-\mu_i)^\top Q(x-\mu_i)\right)^2 + 8\sum_{i\leqslant k} w_i \mathbf{E}_{\mathcal{D}_i}\left((x-\mu_i)^\top Q\mu_i\right)^2 \tag{3.78}$$

$$\left.\leqslant 6\sum_i w_i C\left\|\Sigma_i^{1/2} Q \Sigma_i^{1/2}\right\|_F^2 + 8\sum_{i\leqslant k} w_i \mathbf{E}_{\mathcal{D}_i}\left((x-\mu_i)^\top Q\mu_i\right)^2\right\}. \tag{3.79}$$

For the first term, note that $\|\Sigma_i\|_2 \leqslant 1 + \|\Sigma_i - I\|_F \leqslant 1 + H$. Thus, $\left\|\Sigma_i^{1/2}\right\|_2 \leqslant \sqrt{1+H}$. Thus, we have that $\Sigma_i^{1/2} \preceq I + (\Sigma_i^{1/2} - I) \preceq \sqrt{1+H} I$. Using Lemma 3.2.25 with $A = (1+H)^{-1/2}\Sigma_i^{1/2}$ and $B = Q\Sigma_i^{1/2}$, we have: $\left|\frac{Q}{2}\right\}\left\{\left\|\Sigma_i^{1/2} Q \Sigma_i^{1/2}\right\|_F^2 \leqslant (1+H)\left\|Q\Sigma_i^{1/2}\right\|_F^2\right\}$. By another application of Lemma 3.2.25, we have: $\left|\frac{Q}{2}\right\}\left\{\left\|Q\Sigma_i^{1/2}\right\|_F^2 \leqslant (1+H)\|Q\|_F^2\right\}$. Thus, altogether, we have: $\left|\frac{Q}{2}\right\}$ $\left\{\left\|\Sigma_i^{1/2} Q \Sigma_i^{1/2}\right\|_F^2 \leqslant (1+H)^2\|Q\|_F^2\right\}$. Using our assumption that $1 < H$, we thus have:

$$\left|\frac{Q}{2}\right\}\left\{\sum_i w_i C\left\|\Sigma_i^{1/2} Q \Sigma_i^{1/2}\right\|_F^2 \leqslant C(1+H)^2\|Q\|_F^2 \leqslant 4CH^2\|Q\|_F^2\right\}.$$

For the second term, first observe that the following equality of quadratic polynomials in indeterminate Q: $\left((x-\mu_i)^\top Q\mu_i\right)^2 = \left((\Sigma_i^{\dagger/2}(x-\mu_i))^\top \Sigma_i^{1/2} Q\mu_i\right)^2$. Thus, $\mathbf{E}_{x\sim\mathcal{D}_i}\left((x-\mu_i)^\top Q\mu_i\right)^2 = \left\|\Sigma_i^{1/2} Q\mu_i\right\|_2^2$. Next, by the SoS Cauchy-Schwarz inequality (Fact 3.2.20), we have that

$$\left|\frac{Q}{2}\right\}\left\{\left\|\Sigma_i^{1/2} Q\mu_i\right\|_2^2 = \operatorname{tr}(\mu_i\mu_i^\top Q\Sigma_i Q) \leqslant H\operatorname{tr}(Q\Sigma_i Q) = H\left\|\Sigma_i^{1/2} Q\right\|_F^2\right\}.$$

Applying Lemma 3.2.25 with the observation above that $\Sigma_i^{1/2} \leqslant (1+H)^{1/2} I$ yields:

$$\left|\frac{Q}{2}\right\}\left\{\left\|\Sigma_i^{1/2} Q\right\|_F^2 \leqslant (1+H)\|Q\|_F^2\right\}.$$

Thus, altogether, we obtain: $\left|\frac{Q}{2}\right\}\left\{\mathbf{E}_{x\sim\mathcal{D}_i}(x-\mu_i)^\top Q\mu_i\right\}^2 \leqslant H(1+H)^2\|Q\|_F^2 \leqslant 4H^3\|Q\|_F^2$. We thus have:

$$\left|\frac{Q}{2}\right\}\left\{\sum_{i\leqslant k} w_i \mathbf{E}_{\mathcal{D}_i}\left((x-\mu_i)^\top Q\mu_i\right)^2 \leqslant 4H^3\|Q\|_F^2\right\}.$$

Plugging in these bounds into (3.79) completes the proof. □

As an immediate corollary of Lemma 3.2.38 and Lemma 3.2.41, we obtain:

Lemma 3.10.11 (Variance of Degree-2 Polynomials of Mixtures of Gaussians, Lemma 3.2.42 restated)**.** *Let $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture of Gaussians with $w_i \geqslant \alpha$, mean $\mu = \sum_i w_i \mu_i$ and covariance $\Sigma = \sum_i w_i((\mu_i - \mu)(\mu_i - \mu)^\top + \Sigma_i)$. Suppose that for some $H > 1$, $\left\|\Sigma^{\dagger/2}(\Sigma_i - I)\Sigma^{\dagger/2}\right\|_F \leqslant H$ for every $1 \leqslant i \leqslant k$. Let Q be a symmetric $d \times d$ matrix-valued indeterminate. Then for $H' = \max\{H, 1/\alpha\}$,*

$$\left|\frac{Q}{2}\right\} \left\{ \mathop{\mathbb{E}}_{x\sim\mathcal{M}} \left[\left(x^\top Q x - \mathop{\mathbb{E}}_{x\sim\mathcal{M}} \left[x^\top Q x \right] \right)^2 \right] \leqslant 100 H'^2 \left\| \Sigma^{1/2} Q \Sigma^{1/2} \right\|_F^2 \right\} .$$

Proof. Let $\Sigma = U \Lambda U^\top$ be the covariance of the mixture $\mathcal{M}$ along with its eigendecomposition. We want to apply Lemma 3.2.41 and Lemma 3.2.38 with the linear transformation $x \to Ax$ for $A = \Lambda^{\dagger/2} U^\top$. For this, we need to check that the conditions of the Lemma 3.2.41 are met after this linear transformation. The new component covariance is $\Sigma_i' = A \Sigma_i A^\top$ and the hypothesis implies that they are within H in Frobenius distance of the new mixture covariance $I' = A \Sigma A^\top$ (I in the range space of Σ). The new means of the components after the linear transformation are $\mu_i' = A\mu_i$ and the new mixture mean is $\mu' = A\mu$. Thus, noting that $I' = \sum_i w_i (\mu_i' - \mu')(\mu' - \mu')^\top + \sum_i w_i \Sigma_i'$, and since each of the terms in the RHS of the preceding equality are PSD, we must have that $I' \succeq w_i (\mu_i' - \mu')(\mu_i' - \mu')^\top$ for every i. Thus, $1 = \|I'\|_2 \geqslant w_i \left\| (\mu_i' - \mu')(\mu_i' - \mu')^\top \right\|_2 = \|\mu_i' - \mu'\|_2^2$. Rearranging yields that $\|\mu_i' - \mu'\|_2^2 \leqslant 1/w_i \leqslant 1/\alpha$. Thus, we can now apply Lemma 3.2.41 to the linearly transformed mixture and the conclusion follows. □

3.10.4 Omitted Proofs from Section 3.2.4

Lemma 3.10.12 (Lemma 3.2.47 restated)**.** *If X satisfies Condition 3.2.45 with respect to $\mathcal{M} = \sum_i w_i \mathcal{N}(\mu_i, \Sigma_i)$ with parameters (γ, t), then if $w_i \geqslant \gamma$ for all $i \in [k]$, and if for some $B \geqslant 0$ we have that $\|\mu_i\|_2^2, \|\Sigma_i\|_{\mathrm{op}} \leqslant B$ for all $i \in [k]$, then for all $m \leqslant t$, we have that:*

$$\left\| \mathbf{E}_{x\in_u X}[x^{\otimes m}] - \mathbf{E}_{x\sim M}[x^{\otimes m}] \right\|_F^2 \leqslant \gamma^2 m^{O(m)} B^m d^m .$$

Proof. We begin by noting that for any symmetric m-tensor A we have that

$$\|A\|_F^2 \leqslant m^{O(m)} (\mathbf{E}_{v\sim\mathcal{N}(0,I)}[\langle A, v^{\otimes m}\rangle^2]).$$

This is because the squared expectation of $\langle A, v^{\otimes m}\rangle$ is $\mathbf{E}[\mathrm{Hom}_A(v)^2] \geqslant m!\mathbf{E}[h_A(v)^2] = m!\,\|A\|_F^2$, where the first inequality holds because $\sqrt{m!}h_A(v)$ is the degree-m harmonic part of $\mathrm{Hom}_A(v)$, and the equality is by Claim 3.22. Therefore, to prove the lemma, it suffices to bound

$$
\begin{aligned}
&\mathbf{E}_{v\sim\mathcal{N}(0,I)}\left[(\mathbf{E}_{x\in_u X}[(v\cdot x)^m] - \mathbf{E}_{x\sim\mathcal{M}}[(v\cdot x)^m])^2\right]\\
=&\mathbf{E}_{v\sim\mathcal{N}(0,I)}\left[\left(\sum_{i=1}^{k}\left(\frac{1}{n}\sum_{x\in X_i}(v\cdot x)^m - w_i\mathbf{E}_{x\sim\mathcal{N}(\mu_i,\Sigma_i)}(v\cdot x)^m\right)\right)^2\right]\\
=&\mathbf{E}_{v\sim\mathcal{N}(0,I)}\left[\left(\sum_{i=1}^{k}\sum_{j=0}^{m}\binom{m}{j}\mu_i^{m-j}\left(\frac{1}{n}\sum_{x\in X_i}(v\cdot(x-\mu_i))^j - w_i\mathbf{E}_{x\sim\mathcal{N}(\mu_i,\Sigma_i)}(v\cdot(x-\mu_i))^j\right)\right)^2\right]\\
\leqslant&\mathbf{E}_{v\sim\mathcal{N}(0,I)}\left[\left(\sum_{i=1}^{k}\sum_{j=0}^{m}\binom{m}{j}|v\cdot\mu_i|^{m-j}\left(w_i\gamma m!(v^T\Sigma v)^{j/2}\right)\right)^2\right]\\
\leqslant&\gamma^2 m^{O(m)}\mathbf{E}_{v\sim\mathcal{N}(0,I)}\left[\sum_{i=1}^{k}\left(w_i(|v\cdot\mu_i| + (v^T\Sigma v)^{1/2})^{2m}\right)\right]\\
\leqslant&\gamma^2 m^{O(m)}\mathbf{E}_{v\sim\mathcal{N}(0,I)}\left[2B^m\|v\|_2^{2m}\right]\\
\leqslant&\gamma^2 m^{O(m)}B^m d^m.
\end{aligned}
$$

This completes the proof. □

Lemma 3.10.13 (Lemma 3.2.48 restated)**.** *Let $\mathcal{M} = \sum_i w_i\mathcal{N}(\mu_i,\Sigma_i)$. Let $S\subset[k]$ with $\sum_{i\in S}w_i = w$, and let $\mathcal{M}' = \sum_{i\in S}(w_i/w)\mathcal{N}(\mu_i,\Sigma_i)$. Then if X satisfies Condition 3.2.45 with respect to $\mathcal{M}$ with parameters (γ,t) for some $\gamma < 1/(2k)$ with the corresponding partition being $X = X_1\cup X_2\cup\ldots\cup X_k$, then $X' = \bigcup_{i\in S}X_i$ satisfies Condition 3.2.45 with respect to $\mathcal{M}'$ with parameters $(O(k\gamma/w),t)$.*

Proof. After noting that $|X'| = w|X|(1+O(k\gamma/w))$, the rest follows straightforwardly from the definitions using the partition $X' = \bigcup_{i\in S}X_i$. □

Lemma 3.10.14 (Lemma 3.2.49 restated)**.** *Let $\mathcal{M} = \sum_{i=1}^k w_i\mathcal{N}(\mu_i,\Sigma_i)$ and let n be an integer at least $kt^{Ct}d^t/\gamma^3$, for a sufficiently large universal constant $C>0$, some $\gamma>0$, and some $t\in\mathbb{N}$. If X consists of n i.i.d. samples from $\mathcal{M}$, then X satisfies Condition 3.2.45 with respect to $\mathcal{M}$ with parameters (γ,t) with high probability.*

Proof. We will show that Condition 3.2.45 holds with high probability using that partition where X_i is the set of samples drawn from the i-th component of $\mathcal{M}$. Note that the second part of

Condition 3.2.45 holds with high probability, so long as n is a sufficiently large multiple of d/γ^2 by the VC-Theorem. In particular, if we think of samples as being drawn from $\mathcal{R}^d \times [k]$, where the second coordinate denotes the component that the sample was drawn from, the second part of Condition 3.2.45 says that the empirical probability of any event $H \times \{i\}$ is correct to within additive error γ. It is easy to see and well-known that the class of such events has VC-dimension $O(d)$, from which the desired bound follows.

For the first part of Condition 3.2.45, we claim that it holds with high probability so long as $n \geqslant kt^{Ct}d^t/\gamma^3$. To prove this, we show it separately for each i with $w_i \geqslant \gamma$ (as otherwise there is nothing to prove) and take a union bound. As Condition 3.2.45 is invariant under affine transformations, we may perform an invertible affine transformation so that $\mu_i = 0$ and Σ_i is the projection onto the first d' coordinates, for some d'. It is clear that only the first d' coordinates of any element of X_i will be non-zero. We claim that the first part of our condition will follow for a given m, so long as $||X_i|/n - w_i| \leqslant \gamma w_i$ (which holds with high probability if $n \gg \log(k)/\gamma^3$), and

$$\left\|\mathbf{E}_{x\in_u X_i}[x^{\otimes m}] - \mathbf{E}_{x\sim\mathcal{N}(0,I_{d'})}[x^{\otimes m}]\right\|_F^2 \leqslant \gamma^2 \,, \tag{3.80}$$

as $\frac{1}{n}\sum_{x\in X_i}\langle v, x-\mu_i\rangle^m = w_i(1\pm\gamma)\langle \mathbf{E}_{x\in_u X_i}[x^{\otimes m}], v^{\otimes m}\rangle$. It is easy to see that each entry of the tensor on the left hand side of Equation (3.80) has mean 0 and variance $m^{O(m)}/|X_i|$, and thus the expected size of the left hand side is $m^{O(m)}d^m/|X_i|$. Then, when $n \geqslant k^{Ck}d^{4k}/\gamma^3$ for a sufficiently large constant C, all parts of our condition hold with high probability. This completes the proof. □

3.10.5 Omitted Proofs from Section 3.6

Lemma 3.10.15 (Frobenius Distance to TV Distance, Lemma 3.6.2, restated)**.** *Suppose $\mathcal{N}(\mu_1, \Sigma_1)$, $\mathcal{N}(\mu_2, \Sigma_2)$ are Gaussians with $\|\mu_1 - \mu_2\|_2 \leqslant \delta$ and $\|\Sigma_1 - \Sigma_2\|_F \leqslant \delta$. If the eigenvalues of Σ_1 and Σ_2 are at least $\lambda > 0$, then*

$$d_{\mathsf{TV}}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = O(\delta/\lambda) \,.$$

Proof. By Fact 3.2.1, we have

$$d_{\mathsf{TV}}\left(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)\right) = \mathcal{O}\Big(\Big((\mu_1-\mu_2)^\top\Sigma_1^{-1}(\mu_1-\mu_2)\Big)^{1/2} + \|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - I\|_F\Big) \,.$$

Then the first term is $\left\langle \mu_1-\mu_2, \Sigma_1^{-1}(\mu_1-\mu_2)\right\rangle^{1/2} \leqslant (\|\Sigma_1^{-1}\|_{\mathrm{op}}\|\mu_1-\mu_2\|_2^2)^{1/2} \leqslant \delta/\sqrt{\lambda}$. The

second term is

$$\begin{aligned}\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - I\|_F^2 &= \|\Sigma_1^{-1/2}(\Sigma_1 - \Sigma_2)\Sigma_1^{-1/2}\|_F^2 \\ &= \operatorname{tr}\left(\left(\Sigma_1^{-1/2}(\Sigma_1 - \Sigma_2)\Sigma_1^{-1/2}\right)^2\right) \\ &\leqslant \operatorname{tr}\left((\Sigma_1 - \Sigma_2)^2\right)(1/\lambda)^2 \\ &\leqslant (\delta/\lambda)^2.\end{aligned}$$

Thus,

$$d_{\mathsf{TV}}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = O(\delta/\sqrt{\lambda} + \delta/\lambda) = O(\delta/\lambda)\ .$$

□

Lemma 3.10.16 (Component Moments to Mixture Moments, Lemma 3.6.3 restated)**.** *Let $\mathcal{M} = \sum_{i\in[k]} w_i\mathcal{N}(\mu_i, \Sigma_i)$ be a k-mixture such that $w_i \geqslant \alpha$, for some $0 < \alpha < 1$, and $\mathcal{M}$ has mean μ and covariance Σ and for all $i \neq j \in [k]$, $\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F \leqslant 1/\sqrt{\alpha}$. Let X be a multiset of n samples satisfying Condition 3.2.45 with respect to $\mathcal{M}$ with parameters (γ, t), for $0 < \gamma < (dk/\alpha)^{-ct}$, for a sufficiently large constant c, and $t \in \mathbb{N}$. Let $\mathcal{D}$ be the uniform distribution over X. Then, $\mathcal{D}$ is $2t$-certifiably (c/α)-hypercontractive and for $d\times d$-matrix-valued indeterminate Q, $\vdash^{Q}_{2} \left\{\mathbf{E}_{\mathcal{M}}\left(x^\top Qx - \mathbf{E}_{\mathcal{M}} x^\top Qx\right)^2 \leqslant \mathcal{O}(1/\alpha)\left\|\Sigma^{1/2}Q\Sigma^{1/2}\right\|_F^2\right\}$.*

Proof. First, since $\mathcal{M}$ is a k-mixture of Gaussians with minimum mixing weight $w_{\min} \geqslant \alpha$, it follows from Corollary 3.2.34 that $\mathcal{M}$ is t-certifiably $(4/\alpha)$ hypercontractive. Further, since X satisfies Condition 3.2.45 with parameters (γ, t), it follows from Lemma 3.2.46 that the set $X' = \{\Sigma^{\dagger/2}(x_i - \mu)\}_{x_i\in X}$ also satisfies Condition 3.2.45 with parameters (γ, t) w.r.t. $\mathcal{M}' = \sum_{i\in[k]} w_i\mathcal{N}\left(\Sigma^{\dagger/2}(\mu_i - \mu), \Sigma^{\dagger/2}\Sigma_i\Sigma^{\dagger/2}\right)$. Since $\left\|\Sigma^{\dagger/2}\Sigma_i\Sigma^{\dagger/2}\right\|_{\text{op}} \leqslant \mathcal{O}(1/\alpha)$, it follows from Lemma 3.2.47 that for all $m \leqslant t$, $\|\mathbf{E}_{x\in_u X'}[x^{\otimes m}] - \mathbf{E}_{x\sim M'}[x^{\otimes m}]\|_F^2 \leqslant \gamma^2 m^{O(m)} d^m (1/\alpha)^m$. Since $\gamma < (dk/\alpha)^{-O(t)}$, it follows from Fact 3.2.43 that X is $2t$-certifiably (c/α)-hypercontractive.

By assumption, for all $i \neq j \in [k]$, we have that $\left\|\Sigma^{\dagger/2}\left(\Sigma_i - \Sigma_j\right)\Sigma^{\dagger/2}\right\|_F \leqslant 1/\sqrt{\alpha}$. We can now apply Lemma 3.2.42 to obtain

$$\vdash^{Q}_{2} \left\{\mathop{\mathbb{E}}_{x\sim\mathcal{M}}\left[\left(x^\top Qx - \mathop{\mathbb{E}}_{x\sim\mathcal{M}}\left[x^\top Qx\right]\right)^2\right] \leqslant \mathcal{O}(1/\alpha)\left\|\Sigma^{1/2}Q\Sigma^{1/2}\right\|_F^2\right\}\ .$$

Therefore, it follows from Fact 3.2.43 that since X satisfies Condition 3.2.45 with parameters

(γ, t), the uniform distribution $\mathcal{D}_X$ on X,

$$\left|\frac{Q}{2}\right\rangle \left\{ \mathbf{E}_{x\sim\mathcal{D}_X}(x^\top Qx - \mathbf{E}_{x\sim\mathcal{D}_X} x^\top Qx)^2 \leqslant \mathcal{O}(1/\alpha) \left\|\Sigma^{1/2} Q \Sigma^{1/2}\right\|_F^2 \right\}.$$

□

3.11 Bit Complexity Analysis

Here we address numerical issues related to our computation. We begin wth the assumption that the eigenvalues of our covariance matrices are bounded below.

Lemma 3.11.1. *If $\mathcal{M} = \sum_{i=1}^k w_i G_i$ is a mixture of Gaussians G_i where each G_i has mean and covariance of norm at most 2^b for some positive integer b and each G_i has covariance matrix whose eigenvalues are bounded below by some $\lambda > 0$. Let $\mathcal{M}'$ be an ε-corruption of $\mathcal{M}$ whose outputs are bounded by $2^{O(b)}$. Let N be a sufficiently large polynomial in d^k/ε and let η be λ divided by a sufficiently large polynomial in $2^b d/\varepsilon$ (where sufficiently large is degree $O(1)$). Then if our algorithm is given N i.i.d. samples from $\mathcal{M}'$ with each of their coordinates rounded to a nearby multiple of η (by which we mean one of the two closest), then our algorithm runs in time poly$(N, b, \log(1/\eta))$ and with high probability returns a list of mixtures of Gaussians X_i with at least one of the X_i poly$_k(\varepsilon)$-close to $\mathcal{M}$ in parameter distance.*

Proof. This follows from noting firstly that with high probability the any subset of the rounded samples will have moments $\lambda/\text{poly}(d/\varepsilon)$-close to their moments before rounding. This means that with high probability these rounded samples will satisfy Condition 3.2.45. This means that our algorithm satisfies the necessary correctness guarantees. Furthermore, given that our samples now all have bounded bit complexity, it is easy to see that the runtime of our algorithm is polynomial in N and the bit complexity. □

More generally, as long as the parameters of the components of our mixture can be expressed with bounded bit complexity, we can prove a similar result, without needing any lower bound on the covariances.

Theorem 91. *Let $\mathcal{M} = \sum_{i=1}^k w_i G_i$ be a mixture of Gaussians where the G_i are Gaussians whose means and covariance matrices can all be written with coefficients given by rational numbers with bit complexity at most b for some integer b. Let $\mathcal{M}'$ be an ε-corruption of $\mathcal{M}$*

so that with probability 1 *the returned points have size* $2^{O(b)}$*. Let* N *be a sufficiently large polynomial in* d^k/ε*. Then there exists an algorithm that given* b *bit-oracle access to these samples runs in time* $\text{poly}(N, b)$ *and with high probability returns a mixture of Gaussians* X *so that* $d_{\mathsf{TV}}(X, \mathcal{M}) < \text{poly}_k(\varepsilon)$*.*

Proof. We begin by showing that we can find a list of hypotheses at least one of which is close. It is then straightforward to show that we can run a tournament over these hypotheses to find a specific one that works. We also assume for simplicity that each w_i is at least 3ε.

We begin by setting λ to be $2^{-b \cdot d^{kC}}$ for a sufficiently large constant C. By adding each sample to a random sample from $N(0, \lambda I)$, we can produce samples from $\tilde{\mathcal{M}}'$, and ε-corruption of $\tilde{\mathcal{M}} = \sum_{i=1}^{k} w_i \tilde{G}_i$ where $\tilde{G}_i$ is the convolution of G_i with $N(0, \lambda I)$. Note that $\tilde{G}_i$ is a Gaussian whose covariance has eigenvalues at least λ. Furthermore, if the covariance matrix of G_i is non-singular, the smallest eigenvalue of the covariance matrix must be at least $2^{O(b \cdot d)}$, and therefore $d_{\mathsf{TV}}(G_i, \tilde{G}_i) < \varepsilon$.

Since the eigenvalues of the components of $\tilde{\mathcal{M}}$ are bounded below, we can apply Lemma 3.11.1 to our samples from $\tilde{\mathcal{M}}'$ rounded to an appropriate accuracy η, and in $\text{poly}(N, b)$-time obtain a list of hypothesis mixtures at least one of which is (with high probability) close to $\tilde{\mathcal{M}}$ in total variation distance.

If the covariances of all of the G_i with weights more than some sufficiently large $\text{poly}_k(\varepsilon)$ are all non-singular, then one of these hypotheses will be close to $\mathcal{M}$. Otherwise, there must be some i for which w_i is relatively large and for which G_i has singular covariance matrix. In particular, there must be an integer vector v with bit complexity $O(bd)$ in the kernel of the covariance matrix of G_i. The hypothesis mixture X that is close to $\tilde{\mathcal{M}}$ in parameter distance must contain some component close to $\tilde{G}_i$. Since $\tilde{G}_i$ has covariance matrix $\tilde{\Sigma}_i = \lambda I + \Sigma_i$ where Σ_i is the covariance matrix of G_i. We note that $\tilde{\Sigma}_i$ will have an eigenvalue of λ and that therefore, our close hypothesis will have an eigenvalue at most 2λ.

If any of our returned hypotheses have any component with a covariance matrix Σ which has any eigenvalue less than 2λ, we do the following. We consider the quadratic form on integer vectors v defined by

$$Q(v) = v^T \Sigma v + \sqrt{\lambda} |v|_2^2.$$

We note that if this Σ is close in parameter distance to a singular $\tilde{\Sigma}_i$ where Σ_i had a null-vector v of norm $2^{O(bd)}$, then for that same value of v we will have that $Q(v) < \lambda^{1/4}$. Using the Lovász local lemma, we can find a v so that $Q(v)$ is within a $2^{O(d)}$-factor of the minimum possible value

over all non-zero, integer vectors v. If for this v, $Q(v) > 2^{\Omega(d)}\lambda^{1/4}$, we know that the hypothesis in question is not close to $\tilde{\mathcal{M}}$ in parameter distance and can be ignored. On the other hand, any v with $Q(v)$ this small must have $|v| < 2^{O(d)}\lambda^{-1/2}$ and $v^T\Sigma v < 2^{O(d)}\lambda^{1/4}$. Note that the projection of v onto the $\ker(\Sigma_i)^\perp$ is either zero or has magnitude at least $2^{O(bd)}$. In the latter case, it would need to be the case that $Q(v)$ is substantially larger. Thus, if such a hypothesis is close to $\tilde{\mathcal{M}}$ in parameter distance, then v is in the kernel of some Σ_i.

If our algorithm finds some v for some hypothesis, we then compute $v \cdot x$ to error λ for each of our samples x. If $\mathcal{M}$ really has a component with v in the kernel of its covariance matrix, all of the x's taken from this component will have $v \cdot x$ the same. This means that at least a $(3/2)\varepsilon$ fraction of our samples x will have $v \cdot x$ within λ of each other. Note that if v is not in the kernel of any covariance matrix of any G_i than $\mathbf{Var}v \cdot G_i$ will be at least $2^{O(db)}$ for each i, and with high probability we will not find this many close samples.

To summarize, if our algorithm applies this procedure to every component of every hypothesis and does not find such a v, then it cannot be the case that $\mathcal{M}$ contains any components of weight more than $\text{poly}_k(\varepsilon)$ that are singular, and thus one of our original hypotheses must be close in total variational distance. We can then run a tournament to find a single one that is close. Otherwise, if we find such a v for which many points do have $v \cdot x$ close by, then v must be a null vector of the covariance matrix of some G_i. Furthermore, all of the samples within λ of this common value of $v \cdot x$, with high probability are either errors or come from components contained in some lower dimensional subspace. We can determine what this subspace is by noting that it is defined by $v \cdot x = q$ for some rational number q with bit-complexity at most $O(bd)$ and using continued fractions on a good numerical approximation of q in order to determine its true value. Our algorithm can then recurse on the points in this subspace (a mixture of Gaussians in a lower dimensional space) and on the remaining points (which are from a mixture of fewer Gaussians), and return an appropriate mixture of the results. □

Chapter 4

Robustly Linear Regression

4.1 Introduction

While classical statistical theory has focused on designing statistical estimators assuming access to i.i.d. samples from a nice distribution, estimation in the presence of adversarial outliers has been a challenging problem since it was formalized by Huber [Hub64].

Regression continues to be extensively studied under various models, including realizable regression (no noise), true linear models (independent noise), asymmetric noise, agnostic regression and generalized linear models (see [Wei05] and references therein). In each model, a variety of distributional assumptions are considered over the covariates and the noise. As a consequence, there exist innumerable estimators for regression achieving various trade-offs between sample complexity, running time and rate of convergence. The presence of adversarial outliers adds yet another dimension to design and compare estimators.

Seminal works on robust regression focused on designing non-convex loss functions, including M-estimators [Hub11], Theil-Sen estimators[The92, Sen68], R-estimators[Jae72], Least-Median-Squares [Rou84] and S-estimators[RY84]. These estimators have desirable statistical properties under disparate assumptions, yet remain computationally intractable in high dimensions. Further, recent works show that it is information-theoretically impossible to design robust estimators for linear regression without distributional assumptions [KKM18].

An influential recent line of work showed that when the data is drawn from the well studied and highly general class of *hypercontractive* distributions (see Definition 4.1.1), there exist robust and computationally efficient estimators for regression [KKM18, PSBR20, DKS19]. Several

families of natural distributions fall into this category, including Gaussians, strongly log-concave distributions and product distributions on the hypercube. However, both estimators converge to the the true hyperplane (in $\ell_,2$-norm) at a sub-optimal rate, as a function of the fraction of corrupted points.

Given the vast literature on ad-hoc and often incomparable estimators for high-dimensional robust regression, the central question we address in this work is as follows:

Does there exist a unified approach to design robust and computationally efficient estimators achieving optimal rates for all linear regression models under mild distributional assumptions?

We address the aforementioned question by introducing a framework to design robust estimators for linear regression when the input is drawn from a *hypercontractive* distribution. Our estimators converge to the true hyperplanes at the information-theoretically optimal rate (as a function of the fraction of corrupted data) under various well-studied noise models, including independent and agnostic noise. Further, we show that our estimators can be computed in polynomial time using the *sum-of-squares* convex hierarchy.

We note that, despite decades of progress, prior to our work, estimators achieving optimal convergence rate in terms of the fraction of corrupted points were not known, even with independent noise and access to unbounded computation.

4.1.1 Our Results

We begin by formalizing the regression model we work with. In classical regression, we assume $\mathcal{D}$ is a distribution over $\mathcal{R}^d \times \mathcal{R}$ and for a vector $\Theta \in \mathcal{R}^d$, the least-squares loss is given by $\mathsf{err}_{\mathcal{D}}(\Theta) = \mathbb{E}_{x,y\sim\mathcal{D}}\left[\left(y - x^\top\Theta\right)^2\right]$. The goal is to learn $\Theta^* = \arg\min_\Theta \mathsf{err}_{\mathcal{D}}(\Theta)$. We assume sample access to $\mathcal{D}$, and given n i.i.d. samples, we want to obtain a vector Θ that approximately achieves optimal error, $\mathsf{err}_{\mathcal{D}}(\Theta^*)$.

In contrast to the classical setting, we work in the *strong contamination model*. Here, an adversary has access to the input samples and is allowed to corrupt an ϵ-fraction arbitrarily. Note, the adversary has access to unbounded computation and has knowledge of the estimators we design. We note that this is the most stringent corrupt model and captures Huber contamination, additive corruption, label noise, agnostic learning etc (see [DK19]). Formally,

Model 92 (Robust Regression Model)**.** Let $\mathcal{D}$ be a distribution over $\mathcal{R}^d \times R$ such that the marginal

distribution over $\mathcal{R}^d$ is centered and has covariance Σ^* and let $\Theta^* = \arg\min_\Theta \mathbb{E}_{x,y\sim\mathcal{D}}\left[(y - \langle\Theta, x\rangle)^2\right]$ be the optimal hyperplane for $\mathcal{D}$. Let $\{(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots (x_n^*, y_n^*)\}$ be n i.i.d. random variables drawn from $\mathcal{D}$. Given $\epsilon > 0$, the robust regression model $\mathcal{R}_\mathcal{D}(\epsilon, \Sigma^*, \Theta^*)$ outputs a set of n samples $\{(x_1, y_1), \ldots (x_n, y_n)\}$ such that for at least $(1-\epsilon)n$ points $x_i = x_i^*$ and $y_i = y_i^*$. The remaining ϵn points are arbitrary, and potentially adversarial w.r.t. the input and estimator.

A natural starting point is to assume that the marginal distribution over the covariates (the x's above) is heavy-tailed and has bounded, finite covariance. However, we show that there is no robust estimator in this setting, even when the linear model has no noise and the uncorrupted points lie on a line.

Theorem 93 (Bounded Covariance does not suffice, Theorem 105 informal)**.** *For all $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^d \times \mathcal{R}$ such that $d_{\mathsf{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leqslant \epsilon$ and the marginal distribution over the covariates has bounded covariance, denoted by $\Sigma^2 = \Theta(1)$, yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega(1)$, where Θ_1 and Θ_2 are the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$.*

The aforementioned result precludes any statistical estimator that converges to the true hyperplane as the fraction of corrupted points tends to 0. Therefore, we strengthen the distributional assumption consider hypercontractive distributions instead.

Definition 4.1.1 ((C,k)-Hypercontractivity)**.** *A distribution $\mathcal{D}$ over $\mathbb{R}^d$ is (C,k)-hypercontractive for an even integer $k \geqslant 4$, if for all $r \in [k/2]$, for all $v \in \mathbb{R}^d$,*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\left\langle v, x - \mathbb{E}[x]\right\rangle^{2r}\right] \leqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[C\left\langle v, x - \mathbb{E}[x]\right\rangle^{2}\right]^r$$

Remark 94. Hypercontractivity captures a broad class of distributions, including Gaussian distributions, uniform distributions over the hypercube and sphere, affine transformations of isotropic distributions satisfying Poincare inequalities [KSS18] and strongly log-concave distributions. Further, hypercontractivity is preserved under natural closure properties like affine transformations, products and weighted mixtures [KS17]. Further, efficiently computable estimators appearing in this work require *certifiable*-hypercontractivity (Definition 4.2.5), a strengthening that continues to capture aforementioned distribution classes.

In this work we focus on the *rate of convergence* of our estimators to the true hyperplane, Θ^*, as a function of the fraction of corrupted points, denoted by ϵ. We measure convergence in both parameter distance (ℓ_2-distance between the hyperplanes) and least-squares error on the true distribution ($\mathsf{err}_\mathcal{D}$).

We introduce a simple analytic condition on the relationship between the noise (marginal distribution over $y - x^\top \Theta^*$) and covariates (marginal distribution over x) that can be considered as a proxy for independence of $y - x^\top \Theta^*$ and x :

Definition 4.1.2 (Negatively Correlated Moments)**.** *Given a distribution $\mathcal{D}$ over $\mathcal{R}^d \times \mathcal{R}$, such that the marginal distribution on $\mathcal{R}^d$ is (c_k, k)-hypercontractive, the corresponding regression instance has negatively correlated moments if for all $r \leqslant k$, and for all v,*

$$\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\langle v, x\rangle^r \left(y - x^\top\Theta^*\right)^r\right] \leqslant \mathcal{O}(1) \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\langle v, x\rangle^r\right] \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y - x^\top\Theta^*\right)^r\right]$$

Informally, the *negatively correlated moments* condition can be viewed as a polynomial relaxation of independence of random variables. Note, it is easy to see that when the noise is independent of the covariates, the above definition is satisfied.

Remark 95. We show that when this condition is satisfied by the true distribution, $\mathcal{D}$, we obtain rates that match the information theoretically optimal rate in a *true linear model*, where the noise (marginal distribution over $y - x^\top\Theta^*$) is independent of the covariates (marginal distribution over x). Further, when this condition is not satisfied, we show that there exist distributions for which obtaining rates matching the *true linear model* is impossible.

When the distribution over the input is hypercontractive and has negatively correlated moments, we obtain an estimator achieving *rate* proportional to $\epsilon^{1-1/k}$ for parameter recovery. Further, our estimator can be computed efficiently. Thus, our main algorithmic result is as follows:

Theorem 96 (Robust Regresssion with Negatively Correlated Noise, Theorem 101 informal)**.** *Given $\epsilon > 0, k \geqslant 4$, and $n \geqslant (d\log(d))^{\mathcal{O}(k)}$ samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, such that $\mathcal{D}$ is (c, k)-certifiably hypercontractive and has negatively correlated moments, there exists an algorithm that runs in $n^{\mathcal{O}(k)}$ time and outputs an estimator $\tilde{\Theta}$ such that with high probability,*

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \tilde{\Theta}\right)\right\|_2 \leqslant \mathcal{O}\left(\epsilon^{1-1/k}\right)\left(\mathit{err}_{\mathcal{D}}(\Theta^*)^{1/2}\right)$$

and,

$$\mathit{err}_{\mathcal{D}}(\tilde{\Theta}) \leqslant \left(1 + \mathcal{O}\left(\epsilon^{2-2/k}\right)\right)\mathit{err}_{\mathcal{D}}(\Theta^*)$$

Remark 97. We note that prior work does not draw a distinction between the independent and dependent noise models. In comparison (see Table 4.1), Klivans, Kothari and Meka [KKM18] obtained a sub-optimal least-squares error scales proportional to $\epsilon^{1-2/k}$. For the special case of

$k = 4$, Prasad et. al. [PSBR20] obtain least squares error proportional to $O(\epsilon\kappa^2(\Sigma))$, where κ is the condition number. In very recent independent work Zhu, Jiao and Steinhardt [ZJS20] obtained a sub-optimal least-squares error scales proportional to $\epsilon^{2-4/k}$.

Further, we show that the rate we obtained in Theorem 96 is information-theoretically optimal, even when the noise and covariates are independent:

Theorem 98 (Lower Bound for Independent Noise, Theorem 103 informal)**.** *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^2 \times \mathcal{R}$ such that the marginal distribution over $\mathcal{R}^2$ has covariance Σ and is (c,k)-hypercontractive for both distributions, and yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\epsilon^{1-1/k}\sigma\right)$, where Θ_1, Θ_2 are the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, $\sigma = \max($ $\mathsf{err}_{\mathcal{D}_1}(\Theta_1), \mathsf{err}_{\mathcal{D}_2}(\Theta_2))$ and the noise is uniform over $[-\sigma, \sigma]$. Further, $|\mathsf{err}_{\mathcal{D}_1}(\Theta_2) - \mathsf{err}_{\mathcal{D}_1}(\Theta_1)| = \Omega\left(\epsilon^{2-2/k}\sigma^2\right)$.*

Next, we consider the setting where the noise is allowed to arbitrary, and need not have negatively correlated moments with the covariates. A simple modification to our algorithm and analysis yields an efficient estimator that obtains rate proportional to $\epsilon^{1-2/k}$ for parameter recovery.

Corollary 4.1.3 (Robust Regresssion with Dependent Noise, Corollary 4.3.1 informal)**.** *Given $\epsilon > 0, k \geqslant 4$ and $n \geqslant (d\log(d))^{\mathcal{O}(k)}$ samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, such that $\mathcal{D}$ is (c,k)-certifiably hypercontractive, there exists an algorithm that runs in $n^{\mathcal{O}(k)}$ time and outputs an estimator $\tilde{\Theta}$ such that with probability 9/10,*

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \tilde{\Theta}\right)\right\|_2 \leqslant \mathcal{O}\left(\epsilon^{1-2/k}\right)\left(\mathsf{err}_{\mathcal{D}}(\Theta^*)^{1/2}\right)$$

and,

$$\mathsf{err}_{\mathcal{D}}(\tilde{\Theta}) \leqslant \left(1 + \mathcal{O}\left(\epsilon^{2-4/k}\right)\right)\mathsf{err}_{\mathcal{D}}(\Theta^*)$$

Further, we show that the dependence on ϵ is again information-theoretically optimal:

Theorem 99 (Lower Bound for Dependent Noise, Theorem 104 informal)**.** *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^2 \times \mathcal{R}$ such that the marginal distribution over $\mathcal{R}^2$ has covariance Σ and is (c,k)-hypercontractive for both distributions, and yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\epsilon^{1-2/k}\sigma\right)$, where Θ_1, Θ_2 be the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively and $\sigma = \max(\mathsf{err}_{\mathcal{D}_1}(\Theta_1), \mathsf{err}_{\mathcal{D}_2}(\Theta_2))$. Further, $|\mathsf{err}_{\mathcal{D}_1}(\Theta_2) - \mathsf{err}_{\mathcal{D}_1}(\Theta_1)| = \Omega\left(\epsilon^{2-4/k}\sigma^2\right)$.*

Estimator	Independent Noise	Arbitrary Noise
Prasad et. al. [PSBR20], Diakonikolas et. al. [DKK+18]	$\epsilon\,\kappa^2$ (only $k=4$)	$\epsilon\,\kappa^2$ (only $k=4$)
Klivans, Kothari and Meka [KKM18]	$\epsilon^{1-2/k}$	$\epsilon^{1-2/k}$
Zhu, Jiao and Steinhardt [ZJS20]	$\epsilon^{2-4/k}$	$\epsilon^{2-4/k}$
Our Work Thm 96, Cor 4.1.3	$\epsilon^{2-2/k}$	$\epsilon^{2-4/k}$
Lower Bounds Thm 98, Thm 99	$\epsilon^{2-2/k}$	$\epsilon^{2-4/k}$

Table 4.1: Comparison of convergence rate (for least-squares error) achieved by various computationally efficient estimators for Robust Regression, when the underlying distribution is (c_k, k)-hypercontractive.

Applications for Gaussian Covariates. The special case where the marginal distribution over x is Gaussian has received considerable interest recently [DKS19, DKK+18]. We note that our estimators extend to the setting of Gaussian covariates, since the uniform distribution over samples from $\mathcal{N}(0,\Sigma)$ are $(\mathcal{O}(k)\,,\mathcal{O}(k))$-certifiably hypercontractive for all k (see Section 5 in Kothari and Steurer [KS17]). As a consequence, instantiating Corollary 4.1.3 with $k=\log(1/\epsilon)$ yields the following:

Corollary 4.1.4 (Robust Regression with Gaussian Covariates)**.** *Given $\epsilon>0$ and $n \geqslant (d\log(d))^{\mathcal{O}(log(1/\epsilon))}$ samples from $\mathcal{R}_{\mathcal{N}}(\epsilon,\Sigma^*,\Theta^*)$, such that the marginal distribution over the x's is $\mathcal{N}(0,\Sigma^*)$, there exists an algorithm that runs in $n^{\mathcal{O}(\log(1/\epsilon)}$ time and outputs an estimator $\tilde{\Theta}$ such that with high probability,*

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^*-\tilde{\Theta}\right)\right\|_2 \leqslant \mathcal{O}(\epsilon\log(1/\epsilon))\left(\mathit{err}_{\mathcal{N}}(\Theta^*)\right)^{1/2}$$

and,

$$\mathit{err}_{\mathcal{N}}(\tilde{\Theta}) \leqslant \left(1+\mathcal{O}\Big((\epsilon\log(1/\epsilon))^2\Big)\right)\mathit{err}_{\mathcal{N}}(\Theta^*)$$

We note that our estimators obtain the rate matching recent work for Gaussians, albeit in quasi-polynomial time. In comparison, Diakonikolas, Kong and Stewart [DKS19] obtain the same rate in polynomial time, when the noise is independent of the covariates. We note that obtaining the optimal rate for Gaussian covariates (shaving the additional $\log(1/\epsilon)$ factor) remains an outstanding open question.

Concurrent Work. We note that a statistical estimator achieving rate proportional to $\epsilon^{1-1/k}$ can be obtained from combining ideas in [ZJS19] and [ZJS20][1]. However, this approach remains computationally intractable. Finally, Cherapanamjeri et al. [CAT+20] consider the special case of $k = 4$ and obtain nearly linear sample complexity and running time. However, their running time and rate incurs a condition number dependence. Further, their rate scales proportional to $\epsilon^{1/2}$, even when the noise is independent of the covariates (as opposed to $\epsilon^{3/4}$).

We emphasize that the bottleneck in all prior and concurrent work remains algorithmically exploiting the independence of the noise and covariates, which we achieve via the *negatively correlated moments* condition (Definition 4.1.2).

4.2 Preliminaries

Throughout this paper, for a vector v, we use $\|v\|_2$ to denote the Euclidean norm of v. For a $n \times m$ matrix M, we use $\|M\|_2 = \max_{\|x\|_2=1} \|Mx\|_2$ to denote the spectral norm of M and $\|M\|_F = \sqrt{\sum_{i,j} M_{i,j}^2}$ to denote the Frobenius norm of M. For symmetric matrices we use $\succeq$ to denote the PSD/Loewner ordering over eigenvalues of M. Recall, the definition of total variation distance between probability measures:

Definition 4.2.1 (Total Variation Distance)**.** *The TV distance between distributions with PDFs p, q is defined as $\frac{1}{2}\int_{-\infty}^{\infty} |p(x) - q(x)|dx$.*

Given a distribution $\mathcal{D}$ over $\mathcal{R}^d \times \mathcal{R}$, we consider the least squares error of a vector Θ w.r.t. $\mathcal{D}$ to be $\mathsf{err}_{\mathcal{D}}(\Theta) = \mathbb{E}_{x,y\sim\mathcal{D}}\left[(y - \langle x, \Theta\rangle)^2\right]$. The linear regression problem minimizes the error over all Θ. The minimizer, $\Theta_{\mathcal{D}}$ of the aformentioned error satisfies the following "gradient condition" : for all $v \in \mathcal{R}^d$,

$$\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left\langle v, xx^\top\Theta_{\mathcal{D}} - xy\right\rangle\right] = 0$$

Fact 4.2.2 (Convergence of Empirical Moments, implicit in Lemma 5.5 [KS17])**.** *Let $\mathcal{D}$ be a (c_k, k)-hypercontractive distribution with covariance Σ and let $\mathcal{X} = \{x_1, \ldots x_n\}$ be $n = \Omega((d\log(d)/\delta)^{k/2})$ i.i.d. samples from $\mathcal{D}$. Then, with probability at least $1 - \delta$,*

$$(1 - 0.1)\Sigma \preceq \frac{1}{n}\sum_{i}^{n} x_i x_i^\top \preceq (1 + 0.1)\Sigma$$

[1]We thank Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt for communicating their observation to us.

Fact 4.2.3 (TV Closeness to Covariance Closeness, Lemma 2.2 [KS17]). *Let $\mathcal{D}_1, \mathcal{D}_2$ be (c_k, k)-hypercontractive distributions over $\mathcal{R}^d$ such that $\|\mathcal{D} - \mathcal{D}'\|_{TV} \leqslant \epsilon$, where $0 < \epsilon < \mathcal{O}\Big((1/c_k)^{\frac{k}{k-1}}\Big)$. Let Σ_1, Σ_2 be the corresponding covariance matrices. Then, for $\delta \leqslant \mathcal{O}\Big(c_k\, \epsilon^{1-1/k}\Big) < 1$,*

$$(1-\delta)\Sigma_2 \preceq \Sigma_1 \preceq (1+\delta)\Sigma_2$$

Lemma 4.2.4 (Löwner Ordering for Hypercontractive Samples). *Let $\mathcal{D}$ be a (c_k, k)-hypercontractive distribution with covariance Σ and and let $\mathcal{U}$ be the uniform distribution over n samples. Then, with probability $1-\delta$,*

$$\left\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I\right\|_F \leqslant \frac{C_4 d^2}{\sqrt{n}\sqrt{\delta}},$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i\in[n]} x_i x_i^\top$.

Next, we define the technical conditions required for efficient estimators. Formally,

Definition 4.2.5 (Certifiable Hypercontractivity). *A distribution $\mathcal{D}$ on $\mathcal{R}^d$ is (c_k, k)-certifiably hypercontractive if for all $r \leqslant k/2$, there exists a degree $\mathcal{O}(k)$ sum-of-squares proof (defined below) of the following inequality in the variable v*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\langle x, v\rangle^{2r}\right] \leqslant \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[c_r \langle x, v\rangle^2\right]^r$$

such that $c_r \leqslant c_k$.

Next, we note that if a distribution $\mathcal{D}$ is certifiably hypercontractive, the uniform distribution over n i.i.d. samples from $\mathcal{D}$ is also certifiably hypercontractive.

Fact 4.2.6 (Sampling Preserves Certifiable Hypercontractivity, Lemma 5.5 [KS17]). *Let $\mathcal{D}$ be a (c_k, k)-certifiably hypercontractive distribution on $\mathcal{R}^d$. Let $\mathcal{X}$ be a set of $n = \Omega\left((d\log(d/\delta))^{k/2}/\gamma^2\right)$ i.i.d. samples from $\mathcal{D}$. Then, with probability $1-\delta$, the uniform distribution over $\mathcal{X}$ is $(c_k+\gamma, k)$-certifiably hypercontractive.*

We also note that certifiably hypercontractivity is preserved under Affine transformations of the distribution.

Fact 4.2.7 (Certifiable Hypercontractivity under Affine Transformations, Lemma 5.1, 5.2 [KS17]). *Let $x \in \mathcal{R}^d$ be a random variable drawn from a (c_k, k)-certifiably hypercontractive distribution. Then, for matrix A and vector b, the distribution over the random variable $Ax + b$ is also (c_k, k)-certifiably hypercontractive.*

Next, we formally define the condition on the moments and noise that we require to obtain efficient algorithms. We note that for technical reasons it is not simply a polynomial identity encoding Definition 4.1.2.

Definition 4.2.8 (Certifiable Negatively Correlated Moments)**.** *A distribution $\mathcal{D}$ on $\mathcal{R}^d \times \mathcal{R}$ has $\mathcal{O}(1)$-certifiable negatively correlated moments if for all $r \leqslant k/2$ there exists a degree $\mathcal{O}(k)$ sum-of-squares proof of the following inequality*

$$\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(v^\top x\left(y - x^\top\Theta\right)\right)^{2r}\right] \leqslant \mathcal{O}(\lambda_r^r)\left(\mathbb{E}\left[(v^\top x)^2\right]^r\right)\left(\mathbb{E}\left[\left(y - x^\top\Theta\right)^2\right]^r\right)$$

for a fixed vector Θ.

4.3 Robust Certifiability and Information Theoretic Estimators

In this section, we provide an estimator that obtains the information theoretically optimal rate for robust regression. We note that we consider the setting where both the covariates and the noise are hypercontractive and the are independent of each other. This setting displays all the key ideas of our estimator. Further, our estimator extends to the remaining settings, such as bounded dependent noise, by simple modifications to the subsequent analysis.

Theorem 100 (Robust Certifiability with Optimal Rate)**.** *Given $\epsilon > 0$, let $\mathcal{D}, \mathcal{D}'$ be distributions over $\mathcal{R}^d \times \mathcal{R}$ such that the respective marginal distributions over $\mathcal{R}^d$, denoted by $\mathcal{D}_\mathcal{X}, \mathcal{D}'_\mathcal{X}$, are (c_k, k)-hypercontractive and $\|\mathcal{D} - \mathcal{D}'\|_{TV} \leqslant \epsilon$. Let $\mathcal{R}_\mathcal{D}(\epsilon, \Sigma_\mathcal{D}, \Theta_\mathcal{D})$ and $\mathcal{R}_{\mathcal{D}'}(\epsilon, \Sigma_{\mathcal{D}'}, \Theta_{\mathcal{D}'})$ be the corresponding instances of robust regression such that $\mathcal{D}, \mathcal{D}'$ have negatively correlated moments. Further, for $(x, y) \sim \mathcal{D}, \mathcal{D}'$, let the marginal distribution over $y - \left\langle x, \mathbb{E}\left[xx^\top\right]^{-1}\mathbb{E}\left[xy\right]\right\rangle$ be (η_k, k)-hypercontractive Then,*

$$\left\|\Sigma_\mathcal{D}^{1/2}(\Theta_\mathcal{D} - \Theta_{\mathcal{D}'})\right\|_2 \leqslant \mathcal{O}\left(\sqrt{c_k\,\eta_k}\,\epsilon^{1-1/k}\right)\left(\mathit{err}_\mathcal{D}(\Theta_\mathcal{D})^{1/2} + \mathit{err}_{\mathcal{D}'}(\Theta_{\mathcal{D}'})^{1/2}\right)$$

Further,

$$\mathit{err}_\mathcal{D}(\Theta_{\mathcal{D}'}) \leqslant \left(1 + \mathcal{O}\left(c_k\,\eta_k\,\epsilon^{2-2/k}\right)\right)\mathit{err}_\mathcal{D}(\Theta_\mathcal{D}) + \mathcal{O}\left(c_k\,\eta_k\,\epsilon^{2-2/k}\right)\mathit{err}_{\mathcal{D}'}(\Theta_{\mathcal{D}'})$$

Proof. Consider a maximal coupling of $\mathcal{D}, \mathcal{D}'$ over $(x, y) \times (x', y')$, denoted by $\mathcal{G}$, such that the

marginal of $\mathcal{G}$ (x, y) is $\mathcal{D}$, the marginal on (x', y') is $\mathcal{D}'$ and $\mathbb{P}_{\mathcal{G}}[\mathbb{I}(x, y) = (x', y')] = 1 - \epsilon$. Then, for all v,

$$\begin{aligned}\langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}) + xy - xy\right\rangle\right] \\ &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(\langle x, \Theta_{\mathcal{D}}\rangle - y\right)\rangle\right] + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle\right]\end{aligned} \tag{4.1}$$

Since $\Theta_{\mathcal{D}}$ is the minimizer for the least squares loss, we have the following gradient condition : for all $v \in \mathbb{R}^d$,

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\langle v, (\langle x, \Theta_{\mathcal{D}}\rangle - y)x\rangle\right] = 0 \tag{4.2}$$

Since $\mathcal{G}$ is a coupling, using the gradient condition (4.2) and using that $1 = \mathbb{I}(x, y) = (x', y')$ $+\mathbb{I}(x, y) \neq (x', y')$, we can rewrite equation (4.1) as

$$\begin{aligned}\langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x, y) = (x', y')\right] \\ &\quad + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x, y) \neq (x', y')\right] \\ &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x, y) = (x', y')\right] \\ &\quad + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x, y) \neq (x', y')\right]\end{aligned} \tag{4.3}$$

Consider the first term in the last equality above. Using the gradient condition for $\Theta_{\mathcal{D}'}$ along with Hölder's Inequality, we have

$$\begin{aligned}&\left|\mathbb{E}_{\mathcal{G}}\left[\,\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x, y) = (x', y')\right]\right| \\ &\quad = \left|\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle\right] - \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x, y) \neq (x', y')\right]\right| \\ &\quad = \left|\mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x, y) \neq (x', y')\right]\right| \\ &\quad \leqslant \left|\mathop{\mathbb{E}}_{\mathcal{G}}\left[\mathbb{I}(x, y) \neq (x', y')^{k/(k-1)}\right]^{(k-1)/k}\right| \cdot \left|\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle^k\right]^{1/k}\right|\end{aligned} \tag{4.4}$$

Observe, since $\mathcal{G}$ is a maximal coupling $\mathbb{E}_{\mathcal{G}}\left[\mathbb{I}(x, y) \neq (x', y')\right]^{(k-1)/k} \leqslant \epsilon^{1-1/k}$. Further, since $\mathcal{D}'$ has negatively correlated moments,

$$\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\langle v, x'\rangle^k \cdot \left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^k\right] = \mathop{\mathbb{E}}_{\mathcal{D}'}\left[\langle v, x'\rangle^k\right] \mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^k\right]$$

By hypercontractivity of the covariates and the noise, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\langle v, x'\rangle^k\right]^{1/k}\mathop{\mathbb{E}}_{\mathcal{D}'}\left[(y'-\langle x', \Theta_{\mathcal{D}'}\rangle)^k\right]^{1/k} \leqslant \mathcal{O}(\sqrt{c_k\,\eta_k})\left(v^\top\Sigma_{\mathcal{D}'}v\right)^{1/2}\mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[(y'-\langle x', \Theta_{\mathcal{D}'}\rangle)^2\right]^{1/2}$$

Therefore, we can restate (4.4) as follows

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x'\,(y'-\langle x', \Theta_{\mathcal{D}'}\rangle)\rangle\,\mathbb{I}(x,y)=(x',y')\right]\right| \leqslant \mathcal{O}\Big(\sqrt{c_k\,\eta_k}\,\epsilon^{\frac{k-1}{k}}\Big)\left(v^\top\Sigma_{\mathcal{D}'}v\right)^{\frac{1}{2}} \\ \mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[(y'-\langle x', \Theta_{\mathcal{D}'}\rangle)^2\right]^{\frac{1}{2}}\end{aligned} \tag{4.5}$$

It remains to bound the second term in the last equality of equation (4.3), and we proceed as follows :

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\,(y-\langle x, \Theta_{\mathcal{D}'}\rangle)\rangle\,\mathbb{I}(x,y)\neq(x',y')\right] = \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle\mathbb{I}(x,y)\neq(x',y')\right] \\ + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\,(y-\langle x, \Theta_{\mathcal{D}}\rangle)\rangle\,\mathbb{I}(x,y)\neq(x',y')\right]\end{aligned} \tag{4.6}$$

We bound the two terms above separately. Observe, applying Hölder's Inequality to the first term, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle\mathbb{I}(x,y)\neq(x',y')\right] &\leqslant \mathop{\mathbb{E}}_{\mathcal{G}}\left[\mathbb{I}(x,y)\neq(x',y')\right]^{\frac{k-2}{k}}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{\frac{k}{2}}\right]^{\frac{2}{k}} \\ &\leqslant \epsilon^{\frac{k-2}{k}}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{\frac{k}{2}}\right]^{\frac{2}{k}}\end{aligned} \tag{4.7}$$

To bound the second term in equation 4.6, we again use Hölder's Inequality followed $\mathcal{D}$ having negatively correlated moments,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y-\langle x, \Theta_{\mathcal{D}}\rangle\right)\rangle \mathbb{I}(x, y) \neq (x', y')\right] &\leqslant \mathop{\mathbb{E}}_{\mathcal{G}}\left[\mathbb{I}(x, y) \neq (x', y')\right]^{\frac{k-1}{k}} \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y-\langle x, \Theta_{\mathcal{D}}\rangle\right)\rangle^{k}\right]^{\frac{1}{k}} \\ &\leqslant \epsilon^{\frac{k-1}{k}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\langle v, x\rangle^{k}\right]^{1/k} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x, \Theta_{\mathcal{D}}\rangle\right)^{k}\right]^{1/k} \\ &\leqslant \epsilon^{\frac{k-1}{k}} \sqrt{c_k\, \eta_k}\left(v^{\top}\Sigma_{\mathcal{D}} v\right)^{1/2} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x, \Theta_{\mathcal{D}}\rangle\right)^{2}\right]^{1/2}\end{aligned} \tag{4.8}$$

where the last inequality follows from hypercontractivity of the covariates and noise. Substituting the upper bounds obtained in Equations (4.7) and (4.8) back in to (4.6),

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y-\langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle \mathbb{I}(x, y) \neq (x', y')\right] &\leqslant \epsilon^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^{\top}\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{\frac{k}{2}}\right]^{\frac{2}{k}} \\ &\quad + \epsilon^{\frac{k-1}{k}} \sqrt{c_k\, \eta_k}\left(v^{\top}\Sigma_{\mathcal{D}} v\right)^{1/2} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x, \Theta_{\mathcal{D}}\rangle\right)^{2}\right]^{1/2}\end{aligned}$$

Therefore, we can now upper bound both terms in Equation (4.3) as follows:

$$\begin{aligned}\langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'})\rangle &\leqslant \mathcal{O}\Big(c_k\, \eta_k\, \epsilon^{\frac{k-1}{k}}\Big)\left(v^{\top}\Sigma_{\mathcal{D}'} v\right)^{1/2} \mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y'-\langle x', \Theta_{\mathcal{D}'}\rangle\right)^{2}\right]^{1/2} \\ &\quad + \mathcal{O}\Big(\epsilon^{\frac{k-2}{k}}\Big) \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^{\top}\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{k/2}\right]^{2/k} \\ &\quad + \mathcal{O}\Big(\epsilon^{\frac{k-1}{k}}\sqrt{c_k\, \eta_k}\Big)\left(v^{\top}\Sigma_{\mathcal{D}} v\right)^{1/2} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x, \Theta_{\mathcal{D}}\rangle\right)^{2}\right]^{1/2}\end{aligned} \tag{4.9}$$

Recall, since the marginals of $\mathcal{D}$ and $\mathcal{D}'$ on $\mathcal{R}^d$ are (c_k, k)-hypercontractive and $\|\mathcal{D}-\mathcal{D}'\|_{\mathsf{TV}} \leqslant \epsilon$, it follows from Fact 4.2.3 that

$$(1-0.1)\,\Sigma_{\mathcal{D}'} \preceq \Sigma_{\mathcal{D}} \preceq (1+0.1)\,\Sigma_{\mathcal{D}'} \tag{4.10}$$

when $\epsilon \leqslant \mathcal{O}\Big((1/c_k k)^{k/k-1}\Big)$. Now, consider the substitution $v = \Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}$. Observe,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^{\top}\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{k/2}\right]^{2/k} &= \mathop{\mathbb{E}}_{\mathcal{D}}\left[\langle x, \left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\rangle^{k}\right]^{2/k} \\ &\leqslant c_k \left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'})\right\|_2^2\end{aligned} \tag{4.11}$$

Then, using the bounds in (4.10) and (4.11) along with $v = \Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}$ in Equation 4.9, we have

$$\left(1 - \mathcal{O}\left(\epsilon^{\frac{k-2}{k}} c_k\right)\right) \left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2^2 \leqslant \mathcal{O}\left(\sqrt{c_k\, \eta_k}\, \epsilon^{\frac{k-1}{k}}\right) \left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2 \left(\mathop{\mathbb{E}}_{x',y' \sim \mathcal{D}'}\left[(y' - \langle x', \Theta_{\mathcal{D}'}\rangle)^2\right]^{\frac{1}{2}} + \mathop{\mathbb{E}}_{x,y \sim \mathcal{D}}\left[(y - \langle x, \Theta_{\mathcal{D}}\rangle)^2\right]^{\frac{1}{2}}\right) \tag{4.12}$$

Dividing out (4.12) by $\left(1 - \mathcal{O}\left(\epsilon^{\frac{k-2}{k}} c_k\right)\right) \left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2^2$ and observing that $\mathcal{O}\left(\epsilon^{\frac{k-2}{k}} c_k\right)$ is upper bounded by a fixed constant less than 1 yields the parameter recovery bound.

Given the parameter recovery result above, we bound the least-squares loss between the two hyperplanes on $\mathcal{D}$ as follows:

$$\begin{aligned}
\left|\mathsf{err}_{\mathcal{D}}(\Theta_{\mathcal{D}}) - \mathsf{err}_{\mathcal{D}}(\Theta_{\mathcal{D}'})\right| &= \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(y - x^\top \Theta_{\mathcal{D}}\right)^2 - \left(y - x^\top \Theta_{\mathcal{D}'} + x^\top \Theta_{\mathcal{D}} - x^\top \Theta_{\mathcal{D}}\right)^2\right] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\langle x, (\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle^2 + 2(y - x^\top \Theta_{\mathcal{D}}) x^\top (\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right] \right| \\
&\leqslant \mathcal{O}\left(c_k\, \eta_k\, \epsilon^{2-2/k}\right) \left(\mathop{\mathbb{E}}_{x',y' \sim \mathcal{D}'}\left[(y' - \langle x', \Theta_{\mathcal{D}'}\rangle)^2\right] + \mathop{\mathbb{E}}_{x,y \sim \mathcal{D}}\left[(y - \langle x, \Theta_{\mathcal{D}}\rangle)^2\right]\right)
\end{aligned} \tag{4.13}$$

where the last inequality follows from observing $\mathbb{E}\left[\left\langle \Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}, x(y - x^\top \Theta_{\mathcal{D}})\right\rangle\right] = 0$ (gradient condition) and squaring the parameter recovery bound. □

Next, we consider the setting where the noise is allowed to dependent arbitrarily on the covariates, which captures the well-studied agnostic model. With a slightly modification in our certifiability proof above (using Cauchy-Schwarz instead of independence), we obtain the optimal rate in this setting. We defer the details to Appendix 4.7.

Corollary 4.3.1 (Robust Regression with Dependent Noise)**.** *Let $\mathcal{D}, \mathcal{D}'$ be distributions over $\mathcal{R}^d \times \mathcal{R}$ and let $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma_{\mathcal{D}}, \Theta_{\mathcal{D}})$, $\mathcal{R}_{\mathcal{D}'}(\epsilon, \Sigma_{\mathcal{D}'}, \Theta_{\mathcal{D}'})$ be robust regression instances satisfying the hypothesis in Theorem 100 such that the negatively correlated moments condition is not satisfied. Then,*

$$\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\right\|_2 \leqslant \mathcal{O}\left(\sqrt{c_k\, \eta_k}\, \epsilon^{1-2/k}\right) \left(\mathsf{err}_{\mathcal{D}}(\Theta_{\mathcal{D}})^{1/2} + \mathsf{err}_{\mathcal{D}'}(\Theta_{\mathcal{D}'})^{1/2}\right)$$

Further,

$$\mathsf{err}_{\mathcal{D}}(\Theta_{\mathcal{D}'}) \leqslant \left(1 + \mathcal{O}\left(c_k\, \eta_k\, \epsilon^{2-4/k}\right)\right) \mathsf{err}_{\mathcal{D}}(\Theta_{\mathcal{D}}) + \mathcal{O}\left(c_k\, \eta_k\, \epsilon^{2-4/k}\right) \mathsf{err}_{\mathcal{D}'}(\Theta_{\mathcal{D}'})$$

4.4 Robust Regression in Polynomial Time

In this section, we describe an algorithm to compute our robust estimator for linear regression efficiently. We consider a polynomial system that encodes our robust estimator. We then consider a sum-of-squares relaxation of this program and compute an approximately optimal solution for our relaxation. To analyze our algorithm, we consider the dual of the sum-of-squares relaxation and show that the sum-of-squares proof system caputures a variant of our robust identifiability proof.

We begin by recalling notation: let $\mathcal{D}$ be a distribution over $\mathcal{R}^d \times \mathcal{R}$ such that it is (λ_k, k)-certifiably hypercontractive. Let $\mathcal{X} = \{(x_1^*, y_1^*), (x_2^*, y_2^*) \dots (x_n^*, y_n^*)\}$ denote n uncorrupted i.i.d samples from $\mathcal{D}$ and let $\mathcal{X}_\epsilon = .\{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$ be an ϵ-corruption of the samples $\mathcal{X}$, drawn from a Robust Regression model, $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$ (Model 92). We consider a polynomial system in the variables $\mathcal{X}' = \{(x_1', y_1'), (x_2', y_2') \dots (x_n', y_n')\}$ and $w_1, w_2, \dots w_n \in \{0,1\}^n$ as follows:

$$\mathcal{A}_{\epsilon,\lambda_k} : \left\{ \begin{array}{ll} & \sum_{i\in[n]} w_i = (1-\epsilon)n \\ \forall i \in [n]. & w_i^2 = w_i \\ \forall i \in [n] & w_i(x_i' - x_i) = 0 \\ \forall i \in [n] & w_i(y_i' - y_i) = 0 \\ & \left\langle v, \frac{1}{n}\sum_{i\in[n]} x_i' \left(\langle x_i', \Theta\rangle - y_i\right)\right\rangle^k = 0 \\ \forall r \leqslant k/2 & \frac{1}{n}\sum_{i\in[n]} \langle x_i', v\rangle^{2r} \leqslant \left(\frac{\lambda_r}{n}\sum_{i\in[n]} \langle x_i', v\rangle^2\right)^r \\ \forall r \leqslant k/2 & \frac{1}{n}\sum_{i\in[n]} (y_i' - \langle \Theta, x_i'\rangle)^{2r} \leqslant \left(\frac{\lambda_r}{n}\sum_{i\in[n]} (y_i' - \langle \Theta, x_i'\rangle)^2\right)^r \\ \forall r \leqslant k/2 & \mathbb{E}\left[\left(v^\top x_i' \left(y_i' - (x_i')^\top \Theta\right)\right)^{2r}\right] \leqslant \mathcal{O}\left(\lambda_r^{2r}\right) \mathbb{E}\left[\langle v, x_i'\rangle^2\right]^r \mathbb{E}\left[(y_i' - \langle x_i', \Theta\rangle)^2\right]^r \end{array} \right\}$$

We show that optimizing an appropriate convex function subject to the aforementioned constraint system results in an efficiently computable robust estimator for regression, achieving the information-theoretically optimal rate. Formally,

Theorem 101 (Robust Regression with Negatively Correlated Moments, Theorem 96 restated)**.**

Given $k \in \mathbb{N}$, $\epsilon > 0$ and $n \geqslant n_0$ samples $\mathcal{X}_\epsilon = \{(x_1, y_1), \dots (x_n, y_n)\}$ from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Sigma^*, \Theta^*)$, where $\mathcal{D}$ is a (λ_k, k)-certifiably hypercontractive distribution over $\mathcal{R}^d \times \mathcal{R}$. Further, $\mathcal{D}$ has certifiable negatively correlated moments. Then, Algorithm 102 runs in $n^{\mathcal{O}(k)}$ time and outputs an estimator $\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]$ such that when $n_0 = \Omega\left((d\log(d))^{\Omega(k)}/\gamma^2\right)$ with probability $1 - 1/poly(d)$ (over the draw of the input),

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]\right)\right\|_2 \leqslant \mathcal{O}\left(\lambda_k\, \epsilon^{1-1/k} + \lambda_k \gamma\right) \mathit{err}_{\mathcal{D}}(\Theta^*)^{1/2}$$

Further,

$$\mathit{err}_{\mathcal{D}}\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]\right) \leqslant \left(1 + \mathcal{O}\left(\lambda_k^2\, \epsilon^{2-2/k} + \lambda_k^2\, \gamma^2\right)\right) \mathit{err}_{\mathcal{D}}(\Theta^*).$$

Algorithm 102 (Optimal Robust Regression in Polynomial Time)**.**

Input: *n samples $\mathcal{X}_\epsilon$ from the robust regression model $\mathcal{R}_{\mathcal{D}}(\epsilon, \Theta^*, \Sigma^*)$.*

Operation:

1. *Find a degree-$\mathcal{O}(k)$ pseudo-distribution $\tilde{\zeta}$ satisfying $\mathcal{A}_{\epsilon,\lambda_k}$ and minimizing*

$$\min_{w,x',y',\Theta} \tilde{\mathbb{E}}_{\tilde{\zeta}} \left[\left(\frac{1}{n}\sum_{i\in[n]} w_i\,(y'_i - \langle\Theta, x'\rangle)^2\right)^k\right]$$

.

2. *Round the pseudo-distribution to obtain an estimator $\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]$.*

Output: *A vector $\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]$ such that the recovery guarantee in Theorem 101 is satisfied.*

Efficient Estimator for Arbitrary Noise. We note that an argument similar to the one presented for Theorem 101 results in a polynomial time estimator when the regression instance does not have negatively correlated moments (definition 4.1.2), albeit at a slightly worse rate. Formally,

Corollary 4.4.1 (Robust Regression with Arbitrary Noise)**.** *Consider the hypothesis of Theorem 101, without the negatively correlated moments assumption. Then, there exists an algorithm that runs in time $n^{\mathcal{O}(k)}$ outputs an estimator $\tilde{\Theta}$ such that when $n_0 = (d\log(d))^{\Omega(k)}/\gamma^2$, with probability $1 - 1/poly(d)$ (over the draw of the input),*

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \tilde{\Theta}\right)\right\|_2 \leqslant \mathcal{O}\left(\lambda_k\, \epsilon^{1-2/k} + c_2\, \eta_2\, \gamma\right) \mathit{err}_{\mathcal{D}}(\Theta^*)^{1/2}$$

Further,

$$\mathit{err}_{\mathcal{D}}\left(\tilde{\Theta}\right) \leqslant \left(1 + \mathcal{O}\Big(\lambda_k^2\, \epsilon^{2-4/k} + \lambda_2^2\, \gamma\Big)\right) \mathit{err}_{\mathcal{D}}(\Theta^*)$$

At a high level, we simply do not enforce the negatively correlated moments constraint in our polynomial system $\mathcal{A}_{\epsilon,\lambda_k}$ and instead use the SoS Cauchy-Schwarz inequality in our key technical lemma (Lemma 4.4.3). For completeness, we provide the proof of the SoS lemma in Appendix 4.8.

4.4.1 Analysis

We begin by observing that we can efficiently optimize the polynomial program above since it admits a compact representation. In particular, $\mathcal{A}_{\epsilon,\lambda_k}$ can be represented as a system of $\mathrm{poly}(n^k)$ constraints in $n^{O(k)}$ variables. We refer the reader to [FKP+19] for a detailed overview on how to efficiently implement the aforementioned constraints.

Lemma 4.4.2 (Soundness of the Constraint System)**.** *Given $n \geqslant n_0$ samples from $\mathcal{R}_{\mathcal{D}}(\epsilon, \Theta^*, \Sigma)$, with probability at least $1 - 1/\mathrm{poly}(d)$ over the draw of the samples, there exists an assignment for w, x', y' and Θ such that $\mathcal{A}_{\epsilon,\lambda_k}$ is feasible when $n_0 = \Big((d\log(d))^{\Omega(k)}\Big)$.*

Proof. Consider the following assignment: for all $i \in [n]$ the w_i's indicate the set of uncorrupted points in $\mathcal{X}_\epsilon$, i.e. $w_i = 1$ if $(x_i, y_i) = (x_i^*, y_i^*)$, $x_i' = x_i$ and $y_i' = y_i$. Further, $\Theta = \Theta^*$, the true hyperplane. It is easy to see that the first four constraints (intersection constraints) are satisfied.

We observe that the marginal distribution over the covariates and the noise are both (λ_k, k)-certifiably hypercontractive since they are Affine transformations of $\mathcal{D}$ (Fact 4.2.7). Next, it follows from Fact 4.2.6, that for $n_0 = \Omega\left(d\log(d)^{\mathcal{O}(k)}\right)$, the uniform distribution over the samples x_i, is $(2\,\lambda_k, k)$-certifiably hypercontractive with probability at least $1 - 1/\mathrm{poly}(d)$. Similarly, the uniform distribution on $y_i - \langle x_i, \Theta^*\rangle$ is $(2\,\lambda_k, k)$-certifiably hypercontractive.

It remains to show that sampling preserves certifiable negatively correlated moments. Recall, since the joint distribution is hypercontractive, by Fact 4.2.6 we know that there's a degree $\mathcal{O}(k)$

proof of

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]}\langle v,x_i\rangle^k\left(y_i-\langle x_i,\Theta^*\rangle\right)^k &\leqslant \mathcal{O}\Big(\lambda_k^k\Big)\left(\frac{1}{n}\sum_{i\in[n]}\langle v,x_i\rangle^2\left(y_i-\langle x_i,\Theta^*\rangle\right)^2\right)^{k/2}\\ &= \mathcal{O}\Big(\lambda_k^k\Big)\left(\frac{1}{n}\sum_{i\in[n]}v^\top x_i(x_i)^\top\left(y_i-\langle x_i,\Theta^*\rangle\right)^2 v\right)^{k/2}\end{aligned} \tag{4.14}$$

It thus suffices to bound the Operator norm of $\frac{1}{n}\sum_{i\in[n]}x_ix_i^\top\left(y_i-\langle x_i,\Theta^*\rangle\right)^2$. It follows from Lemma 4.2.4 that with probability at least $1-1/\mathrm{poly}(d)$,

$$\frac{1}{n}\sum_{i\in[n]}x_ix_i^\top\left(y_i-\langle x_i,\Theta^*\rangle\right)^2\preceq\mathcal{O}(1)\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[xx^\top\left(y-\langle x,\Theta^*\rangle\right)^2\right] \tag{4.15}$$

when $n\geqslant n_0$. Using that $\mathcal{D}$ has negatively correlated moments,

$$\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[xx^\top\left(y-\langle x,\Theta^*\rangle\right)^2\right]\preceq\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[xx^\top\right]\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x,\Theta^*\rangle\right)^2\right] \tag{4.16}$$

Using Lemma 4.2.4 on $xx^\top$ and $\left(y-\langle x,\Theta^*\rangle\right)^2$, we can bound (4.16) as follows:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[xx^\top\right]\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x,\Theta^*\rangle\right)^2\right]\preceq\mathcal{O}(1)\,\mathbb{E}\left[x_ix_i^\top\right]\left(y_i-\langle x_i,\Theta^*\rangle\right)^2 \tag{4.17}$$

Combining Equations (4.15), (4.16), and (4.17), and substituting in (4.14), we have

$$\frac{1}{n}\sum_{i\in[n]}\langle v,x_i\rangle^k\left(y_i-\langle x_i,\Theta^*\rangle\right)^k\leqslant\mathcal{O}\Big(\lambda_k^k\Big)\left(\frac{1}{n}\sum_{i\in[n]}\langle x_i,v\rangle^2\right)^{\frac{k}{2}}\left(\frac{1}{n}\sum_{i\in[n]}\left(y_i-\langle x_i,\Theta^*\rangle\right)^2\right)^{\frac{k}{2}}$$

which concludes the proof. □

Let $\hat{\Sigma}$ be the empirical covariance of the uncorrupted samples $\mathcal{X}$ and let $\hat{\Theta}$ be an optimizer for the empirical loss. Applying Theorem 100 with $\mathcal{D}$ being the uniform distribution on the uncorrupted samples $\mathcal{X}$ and $\mathcal{D}'$ being the uniform distribution on x_i', we get

$$\left\|\hat{\Sigma}^{1/2}\left(\Theta-\hat{\Theta}\right)\right\|_2\leqslant\mathcal{O}\Big(\lambda_k\,\epsilon^{1-1/k}\Big)\,\mathsf{err}_{\mathcal{D}}(\Theta^*)^{1/2}$$

Observe, the aforementioned bound is not a polynomial identity and thus cannot be expressed in the SoS framework. Therefore, we provide a low-degree SoS proof of a slightly modified version of the inequality above, that is inspired by our information theoretic identifiability proof

in Theorem 100.

Lemma 4.4.3 (Robust Identifiability in SoS)**.** *Consider the hypothesis of Theorem 101. Let w, x', y' and Θ be feasible solutions for the polynomial constraint system $\mathcal{A}$. Let*

$$\hat{\Theta} = \arg\min_{\Theta} \frac{1}{n} \sum_{i \in [n]} (y_i^* - \langle x_i^*, \Theta \rangle)^2$$

be the empirical loss minimizer on the uncorrupted samples and let $\hat{\Sigma} = \mathbb{E}\left[x_i^*(x_i^*)^\top\right]$ be the covariance of the uncorrupted samples. Then,

$$\mathcal{A} \left|\frac{w,x',y',\Theta}{4k}\right. \left\{ \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta} - \Theta\right)\right\|_2^{2k} \leqslant 2^{3k}(2\epsilon)^{k-1}\lambda_k^k\, \sigma^{k/2} \left\|\mathbb{E}\left[x_i'(x_i')^\top\right]^{1/2}\left(\hat{\Theta} - \Theta\right)\right\|_2^k \right.$$
$$+ 2^{3k}(2\epsilon)^{k-2}\lambda_k^{2k} \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta} - \Theta\right)\right\|_2^{2k}$$
$$\left. + 2^{3k}(2\epsilon)^{k-1}\lambda_k^k\, \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^2\right]^{k/2} \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta} - \Theta\right)\right\|_2^k \right\}$$

Proof. Consider the empirical covariance of the uncorrupted set given by $\hat{\Sigma} = \mathbb{E}\left[x_i^*(x_i^*)^\top\right]$. Then, using the substitution, along with SoS Almost Triangle Inequality (Fact 2.2.8),

$$\left|\frac{\Theta}{2k}\right. \left\{ \left\langle v, \hat{\Sigma}\left(\hat{\Theta} - \Theta\right)\right\rangle^k = \left\langle v, \mathbb{E}\left[x_i^*(x_i^*)^\top\left(\hat{\Theta} - \Theta\right) + x_i^* y_i^* - x_i^* y_i^*\right]\right\rangle^k \right.$$
$$= \left\langle v, \mathbb{E}\left[x_i^*\left(\left\langle x_i^*, \hat{\Theta}\right\rangle - y_i^*\right)\right] + \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k$$
$$\left. \leqslant 2^k \left\langle v, \mathbb{E}\left[x_i^*\left(\left\langle x_i^*, \hat{\Theta}\right\rangle - y_i^*\right)\right]\right\rangle^k + 2^k \left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \right\} \tag{4.18}$$

Observe, the first term in (4.18) only consists of constants of the proof system. Since $\hat{\Theta}$ is the minimizer of $\mathbb{E}\left[\left(\langle x_i^*, \Theta\rangle - y_i^*\right)^2\right]$, the gradient condition on the samples (appearing in Equation (4.2) of the indentifiability proof) implies this term is 0. Therefore, applying the substitution it suffices to bound the second term.

To this end, we introduce the following auxiliary variables : for all $i \in [n]$, let $w_i' = w_i$ iff the i-th sample is uncorrupted in $\mathcal{X}_\epsilon$, i.e. $x_i = x_i^*$. Then, it is easy to see that $\sum_i w_i' \geqslant (1 - 2\epsilon)n$.

Further, since $\mathcal{A} \vdash_{\frac{w}{2}} \{(1 - w_i' w_i)^2 = (1 - w_i' w_i)\}$,

$$\mathcal{A} \vdash_{\frac{w}{2}} \left\{ \frac{1}{n} \sum_{i \in [n]} (1 - w_i' w_i)^2 = \frac{1}{n} \sum_{i \in [n]} (1 - w_i' w_i) \leqslant 2\epsilon \right\} \tag{4.19}$$

The above equation bounds the uncorrupted points in $\mathcal{X}_\epsilon$ that are not indicated by w. Then, using the substitution, along with the SoS Almost Triangle Inequality (Fact 2.2.8),

$$\begin{aligned}
\mathcal{A} \vdash_{\frac{\Theta, w'}{2k}} \left\{ \left\langle v, \mathbb{E}\left[x_i^* \left(y_i^* - \langle x_i^*, \Theta \rangle\right)\right] \right\rangle^k &= \left\langle v, \mathbb{E}\left[x_i^* \left(y_i^* - \langle x_i^*, \Theta \rangle \left(w_i' + 1 - w_i'\right)\right)\right] \right\rangle^k \\
&= \left\langle v, \mathbb{E}\left[w_i' x_i^* \left(y_i^* - \langle x_i^*, \Theta \rangle\right)\right] + \mathbb{E}\left[(1 - w_i') x_i^* \left(y_i^* - \langle x_i^*, \Theta \rangle\right)\right] \right\rangle^k \\
&\leqslant 2^k \left\langle v, \mathbb{E}\left[w_i' x_i^* \left(y_i^* - \langle x_i^*, \Theta \rangle\right)\right] \right\rangle^k \\
&\quad + 2^k \left\langle v, \mathbb{E}\left[(1 - w_i') x_i^* \left(y_i^* - \langle x_i^*, \Theta \rangle\right)\right] \right\rangle^k \Bigg\}
\end{aligned} \tag{4.20}$$

Consider the first term of the last inequality in (4.20). Observe, since $w_i' x_i^* = w_i w_i' x_i'$ and similarly, $w_i' y_i^* = w_i w_i' y_i'$,

$$\mathcal{A} \vdash_{\frac{\Theta, w'}{4}} \left\{ \mathbb{E}\left[w_i' x_i^* \left(y_i^* - \langle x_i^*, \Theta \rangle\right)\right] = \mathbb{E}\left[w_i' w_i x_i' \left(y_i' - \langle x_i', \Theta \rangle\right)\right] \right\}$$

For the sake of brevity, the subsequent statements hold for relevant SoS variables and have degree $O(k)$ proofs. Using the substitution,

$$\begin{aligned}
\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[w_i' x_i^* \left(y_i^* - \langle x_i^*, \Theta \rangle\right)\right] \right\rangle^k &= \left\langle v, \mathbb{E}\left[w_i' w_i x_i' \left(y_i' - \langle x_i', \Theta \rangle\right)\right] \right\rangle^k \\
&= \left\langle v, \mathbb{E}\left[x_i' \left(y_i' - \langle x_i', \Theta \rangle\right)\right] + \mathbb{E}\left[(1 - w_i' w_i) x_i' \left(y_i' - \langle x_i', \Theta \rangle\right)\right] \right\rangle^k \\
&\leqslant 2^k \left\langle v, \mathbb{E}\left[x_i' \left(y_i' - \langle x_i', \Theta \rangle\right)\right] \right\rangle^k \\
&\quad + 2^k \left\langle v, \mathbb{E}\left[(1 - w_i' w_i) x_i' \left(y_i' - \langle x_i', \Theta \rangle\right)\right] \right\rangle^k \Bigg\}
\end{aligned} \tag{4.21}$$

Observe, the first term in the last inequality above is identically 0, since we enforce the gradient

condition on the SoS variables x', y' and Θ. We can then rewrite the second term using linearity of expectation, followed by applying SoS Hölder's Inequality (Fact 3.2.20) combined with $\mathcal{A}\left|\frac{w}{2}\right. \{(1-w_i'w_i)^2 = 1 - w_i'w_i\}$ to get

$$\begin{aligned}\mathcal{A}\vdash \left\{\left\langle v, \mathbb{E}\left[(1-w_i'w_i)x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k &= \mathbb{E}\left[\langle v, (1-w_i')w_i x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\rangle\right]^k \\ &= \mathbb{E}\left[(1-w_i'w_i)\left\langle v, x_i'\right\rangle\left(y_i' - \langle x_i', \Theta\rangle\right)\right]^k \\ &\leqslant \mathbb{E}\left[(1-w_i'w_i)\right]^{k-1}\mathbb{E}\left[\left\langle v, x_i'\right\rangle^k\left(y_i' - \langle x_i', \Theta\rangle\right)^k\right] \\ &\leqslant (2\epsilon)^{k-1}\mathbb{E}\left[\left\langle v, x_i'\right\rangle^k\left(y_i' - \langle x_i', \Theta\rangle\right)^k\right]\Bigg\}\end{aligned} \tag{4.22}$$

where the last inequality follows from Equation (4.19). Next, we use the certifiable negatively correlated moments constraint with the substitution,

$$\mathcal{A}\vdash \left\{\mathbb{E}\left[\left\langle v, x_i'\right\rangle^k\left(y_i' - \langle x_i', \Theta\rangle\right)^k\right] \leqslant \mathcal{O}\left(\lambda_k^k\right)\mathbb{E}\left[\left\langle v, x_i'\right\rangle^2\right]^{\frac{k}{2}}\mathbb{E}\left[\left(y_i' - \langle x_i', \Theta\rangle\right)^2\right]^{\frac{k}{2}}\right\} \tag{4.23}$$

For brevity, let $\sigma = \mathbb{E}\left[\left(y_i' - \langle x_i', \Theta\rangle\right)^2\right]$. Using the substitution, plugging Equation (4.23) back into (4.22), we get

$$\mathcal{A}\vdash \left\{\left\langle v, \mathbb{E}\left[(1-w_i')x_i'\left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \leqslant (2\epsilon)^{k-1}\lambda_k^k\,\sigma^{k/2}\left\langle v, \mathbb{E}\left[x_i'(x_i')^\top\right]v\right\rangle^{k/2}\right\} \tag{4.24}$$

Recall, we have now bounded the first term of the last inequality in (4.20). Therefore, it remains to bound the second term of the last inequality in (4.20). Using the substitution, we have

$$\begin{aligned}\mathcal{A}\vdash \left\{\left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k &= \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \left\langle x_i^*, \Theta - \hat{\Theta} + \hat{\Theta}\right\rangle\right)\right]\right\rangle^k \\ &\leqslant 2^k\left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)\right]\right\rangle^k \\ &\quad + 2^k\left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(\left\langle x_i^*, \Theta - \hat{\Theta}\right\rangle\right)\right]\right\rangle^k\Bigg\}\end{aligned} \tag{4.25}$$

We again handle each term separately. Observe, the first term when decoupled is a statement

about the uncorrupted samples. Therefore, using the SoS Hölder's Inequality (Fact 3.2.20),

$$\begin{aligned}\mathcal{A}\vdash \Bigg\{\Big\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)\right]\Big\rangle^k &= \mathbb{E}\left[(1-w_i')\left\langle v, x_i^*\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)\right\rangle\right]^k \\ &\leqslant \mathbb{E}\left[(1-w_i')\right]^{k-1}\mathbb{E}\left[\left\langle v, x_i^*\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)\right\rangle^k\right] \\ &\leqslant (2\epsilon)^{k-1}\,\mathbb{E}\left[\langle v, x_i^*\rangle^k\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^k\right]\Bigg\}\end{aligned} \tag{4.26}$$

Observe, the uncorrupted samples have negatively correlated moments, and thus

$$\mathbb{E}\left[\langle v, x_i^*\rangle^k\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^k\right] \leqslant \mathcal{O}\big(\lambda_k^k\big)\,\mathbb{E}\left[\langle v, x_i^*\rangle^2\right]^{k/2}\mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^2\right]^{k/2}$$

Then, by the substitution, we can bound (4.26) as follows:

$$\mathcal{A}\vdash \left\{\Big\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)\right]\Big\rangle^k \leqslant (2\epsilon)^{k-1}\lambda_k^k\,\mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^2\right]^{k/2}\left\langle v, \hat{\Sigma}v\right\rangle^{k/2}\right\} \tag{4.27}$$

In order to bound the second term in (4.25), we use the SoS Hölder's Inequality,

$$\begin{aligned}\mathcal{A}\vdash \Bigg\{\Big\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(\left\langle x_i^*, \Theta - \hat{\Theta}\right\rangle\right)\right]\Big\rangle^k &= \mathbb{E}\left[(1-w_i')^{k-2}\left\langle v, x_i^*\left(\left\langle x_i^*, \Theta - \hat{\Theta}\right\rangle\right)\right\rangle\right] \\ &\leqslant \mathbb{E}\left[1-w_i'\right]^{k-2}\mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \\ &\leqslant (2\epsilon)^{k-2}\,\mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2\Bigg\}\end{aligned} \tag{4.28}$$

Combining the bounds obtained in (4.27) and (4.28), we can restate Equation (4.25) as follows

$$\begin{aligned}\mathcal{A}\vdash \Bigg\{\Big\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\Big\rangle^k \leqslant\ & 2^k(2\epsilon)^{k-1}\lambda_k^k\,\mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^2\right]^{k/2}\left\langle v, \hat{\Sigma}v\right\rangle^{k/2} \\ &+ 2^k(2\epsilon)^{k-2}\,\mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta - \hat{\Theta})\right)^{k/2}\right]^2\Bigg\}\end{aligned} \tag{4.29}$$

Combining (4.29) with (4.24), we obtain an upper bound for the last inequality in Equation (4.20). Therefore, using the substitution, we obtain

$$\begin{aligned}\mathcal{A}\vdash\Bigg\{\Big\langle v,\mathbb{E}\left[x_i^*\left(y_i^*-\langle x_i^*,\Theta\rangle\right)\right]\Big\rangle^k &\leqslant 2^k(2\epsilon)^{k-1}\lambda_k^k\,\sigma^{k/2}\Big\langle v,\mathbb{E}\left[x_i'(x_i')^\top\right]v\Big\rangle^{k/2}\\ &\quad+2^{2k}(2\epsilon)^{k-2}\,\mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta-\hat{\Theta})\right)^{\frac{k}{2}}\right]^2\\ &\quad+2^{2k}(2\epsilon)^{k-1}\lambda_k^k\,\mathbb{E}\left[\left(y_i^*-\left\langle x_i^*,\hat{\Theta}\right\rangle\right)^2\right]^{k/2}\Big\langle v,\hat{\Sigma}v\Big\rangle^{k/2}\Bigg\}\end{aligned}\tag{4.30}$$

Recall, an upper bound on Equation (4.18) suffices to obtain an upper bound on $\left\langle v,\hat{\Sigma}\left(\hat{\Theta}-\Theta\right)\right\rangle$ as follows:

$$\begin{aligned}\mathcal{A}\vdash\Bigg\{\Big\langle v,\hat{\Sigma}\left(\hat{\Theta}-\Theta\right)\Big\rangle^k &\leqslant 2^{2k}(2\epsilon)^{k-1}\lambda_k^k\,\sigma^{k/2}\Big\langle v,\mathbb{E}\left[x_i'(x_i')^\top\right]v\Big\rangle^{k/2}\\ &\quad+2^{3k}(2\epsilon)^{k-2}\,\mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta-\hat{\Theta})\right)^{\frac{k}{2}}\right]^2\\ &\quad+2^{3k}(2\epsilon)^{k-1}\lambda_k^k\,\mathbb{E}\left[\left(y_i^*-\left\langle x_i^*,\hat{\Theta}\right\rangle\right)^2\right]^{k/2}\Big\langle v,\hat{\Sigma}v\Big\rangle^{k/2}\Bigg\}\end{aligned}\tag{4.31}$$

Consider the substitution $v\mapsto\left(\hat{\Theta}-\Theta\right)$. Then,

$$\begin{aligned}\Big\langle v,\hat{\Sigma}\left(\hat{\Theta}-\Theta\right)\Big\rangle^k &= \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k}\\ \Big\langle v,\mathbb{E}\left[x_i'(x_i')^\top\right]v\Big\rangle^{k/2} &= \left\|\mathbb{E}\left[x_i'(x_i')^\top\right]^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{k}\\ \mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta-\hat{\Theta})\right)^{\frac{k}{2}}\right]^2 &= \mathbb{E}\left[\langle x_i^*,\hat{\Theta}-\Theta\rangle^k\right]^2\leqslant\lambda_k^{2k}\left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k}\\ \Big\langle v,\hat{\Sigma}v\Big\rangle^{k/2} &= \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta}-\Theta\right)\right\|_2^{k}\end{aligned}$$

Combining the above with (4.31), we conclude

$$\mathcal{A} \vdash \left\{ \left\| \hat{\Sigma}^{1/2} \left(\hat{\Theta} - \Theta \right) \right\|_2^{2k} \leqslant 2^{3k} (2\epsilon)^{k-1} \lambda_k^k \, \sigma^{k/2} \left\| \mathbb{E}\left[x_i'(x_i')^\top \right]^{1/2} \left(\hat{\Theta} - \Theta \right) \right\|_2^k \right.$$
$$+ 2^{3k} (2\epsilon)^{k-2} \lambda_k^{2k} \left\| \hat{\Sigma}^{1/2} \left(\hat{\Theta} - \Theta \right) \right\|_2^{2k}$$
$$\left. + 2^{3k} (2\epsilon)^{k-1} \lambda_k^k \, \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right)^2 \right]^{k/2} \left\| \hat{\Sigma}^{1/2} \left(\hat{\Theta} - \Theta \right) \right\|_2^k \right\} \tag{4.32}$$

□

Next, we relate the covariance of the samples indicated by w to the covariance on the uncorrupted points. Observe, a real world proof of this follows simply from Fact 4.2.3.

Lemma 4.4.4 (Bounding Sample Covariance)**.** *Consider the hypothesis of Theorem 101. Let w, x', y' and Θ be feasible solutions for the polynomial constraint system $\mathcal{A}$. Then, for $\delta \leqslant \mathcal{O}\left(\lambda_k \epsilon^{1-1/k}\right) < 1$,*

$$\mathcal{A} \left|\frac{w,x'}{2k}\right. \left\{ \left\langle v, \mathbb{E}\left[x_i'(x_i')^\top \right] v \right\rangle^{k/2} \leqslant \left(1 + \mathcal{O}\left(\delta^{k/2}\right)\right) \left\langle v, \hat{\Sigma} v \right\rangle^{k/2} \right\}$$

Proof. Our proof closely follows Lemma 4.5 in [KS17]. For $i \in [n]$, let z_i be an indicator variable such $z_i(x_i^* - x_i') = 0$. Observe, there exists an assignment to z_i such that $\sum_{i \in [n]} z_i = (1-\epsilon)n$, since at most ϵn points were corrupted. Further, $z_i^2 = z_i$ and $\frac{1}{n} z_i = \epsilon$. Then, using the substitution,

$$\begin{aligned}
\mathcal{A} \left|\frac{w,x'}{2k}\right. \left\{ \left\langle v, \left(\mathbb{E}\left[x_i'(x_i')^\top \right] - \hat{\Sigma} \right) v \right\rangle^k \right. &= \left\langle v, \mathbb{E}\left[(1 + z_i - z_i) \left(x_i'(x_i')^\top - x_i^*(x_i^*)^\top \right) \right] v \right\rangle^k \\
&= \mathbb{E}\left[(1 - z_i) \left\langle v, \left(x_i'(x_i')^\top - x_i^*(x_i^*)^\top \right), v \right\rangle \right]^k \\
&\leqslant \epsilon^{k-2} \cdot \mathbb{E}\left[\left(\langle v, x_i' \rangle^2 - \langle v, x_i^* \rangle^2 \right)^{k/2} \right]^2 \\
&\leqslant \epsilon^{k-2} \, \mathbb{E}\left[2^{k/2} \langle v, x_i' \rangle^k + 2^{k/2} \langle v, x_i^* \rangle^k \right]^2 \\
&\left. \leqslant \epsilon^{k-2} 2^k \left(c_k^k \, \mathbb{E}\left[\langle v, x_i' \rangle^2 \right]^{k/2} + \lambda_k^k \, \mathbb{E}\left[\langle v, x_i^* \rangle^2 \right]^{k/2} \right)^2 \right\}
\end{aligned} \tag{4.33}$$

where the first inequality follows from applying the SoS Hölder's Inequality, the second follows from the SoS Almost Triangle Inequality and the third inequality follows from certifiable hypercontractivity of the SoS variables and the uncorrupted samples. Using the SoS Almost Triangle Inequality again, we have

$$\mathcal{A} \vdash \left\{ \left(c_k^k \ \mathbb{E}\left[\langle v, x_i' \rangle^2\right]^{k/2} + \lambda_k^k \ \mathbb{E}\left[\langle v, x_i^* \rangle^2\right]^{k/2} \right)^2 \leqslant \lambda_k^{2k} \ 2^2 \ \left(\left\langle v, \mathbb{E}\left[x_i'(x_i')^\top v\right] \right\rangle^k + \left\langle v, \hat{\Sigma} v \right\rangle^k \right) \right\} \tag{4.34}$$

Combining Equations 4.33, 4.34, we obtain

$$\mathcal{A} \vdash \left\{ \left\langle v, \left(\mathbb{E}\left[x_i'(x_i')^\top\right] - \hat{\Sigma} \right) v \right\rangle^k \leqslant \epsilon^{k-2} \ \lambda_k^{2k} \ 2^{k+2} \left\langle v, \left(\mathbb{E}\left[x_i'(x_i')^\top\right] + \hat{\Sigma} \right) v \right\rangle^k \right\} \tag{4.35}$$

Using Lemma A.4 from [KS17], rearranging and setting $k = k/2$ yields the claim. □

Lemma 4.4.5 (Rounding)**.** *Consider the hypothesis of Theorem 101. Let* $\hat{\Theta} = \arg\min_\Theta \frac{1}{n} \sum_{i \in [n]} (y_i^* - \langle x_i^*, \Theta \rangle)^2$ *be the empirical loss minimizer on the uncorrupted samples. Then,*

$$\left\| \hat{\Sigma}^{1/2} \left(\hat{\Theta} - \tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta] \right) \right\|_2 \leqslant \mathcal{O}\left(\epsilon^{1-\frac{1}{k}} \ \lambda_k \right) \left(\tilde{\mathbb{E}}_{\tilde{\zeta}} \left[\mathbb{E}\left[\left(y_i' - \langle x_i', \Theta \rangle \right)^2 \right]^k \right]^{\frac{1}{2k}} + \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right)^2 \right]^{\frac{1}{2}} \right)$$

Proof. Observe, combining Lemma 4.4.3 and Lemma 4.4.4, we obtain

$$\begin{aligned} \mathcal{A} \vdash \Bigg\{ \left\| \hat{\Sigma}^{1/2} \left(\hat{\Theta} - \Theta \right) \right\|_2^{2k} \leqslant \mathcal{O}\left(\frac{2^{3k} \epsilon^{k-1} \ \lambda_k^k}{1 + 2^{3k} (2\epsilon)^{k-2} \lambda_k^{2k}} \right) \left\| \hat{\Sigma}^{1/2} \left(\hat{\Theta} - \Theta \right) \right\|_2^k \\ \left(\mathbb{E}\left[\left(y_i' - \langle x_i', \Theta \rangle \right)^2 \right]^{\frac{k}{2}} + \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right)^2 \right]^{\frac{k}{2}} \right) \Bigg\} \end{aligned} \tag{4.36}$$

Using Cancellation within SoS (Fact 2.8.3) along with the SoS Almost Triangle Inequality, we can conclude

$$\begin{aligned} \mathcal{A} \vdash \Bigg\{ \left\| \hat{\Sigma}^{1/2} \left(\hat{\Theta} - \Theta \right) \right\|_2^{2k} \leqslant \mathcal{O}\left(2^{3k} \ \epsilon^{k-1} \ \lambda_k^k \right)^2 \\ \left(\mathbb{E}\left[\left(y_i' - \langle x_i', \Theta \rangle \right)^2 \right]^k + \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right)^2 \right]^k \right) \Bigg\} \end{aligned} \tag{4.37}$$

Recall, $\tilde{\zeta}$ is a degree-$\mathcal{O}(k)$ pseudo-expectation satisfying $\mathcal{A}$. Therefore, it follows from Fact

3.2.17 along with Equation 4.36,

$$\tilde{\mathbb{E}}_{\tilde{\zeta}}\left[\left\|\hat{\Sigma}^{\frac{1}{2}}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k}\right] \leqslant \mathcal{O}\Big(2^{4k}\,\epsilon^{k-1}\,\lambda_k^k\Big)^2 \left(\tilde{\mathbb{E}}_{\tilde{\zeta}}\left[\mathbb{E}\left[\left(y_i'-\langle x_i',\Theta\rangle\right)^2\right]^k\right]+\mathbb{E}\left[\left(y_i^*-\left\langle x_i^*,\hat{\Theta}\right\rangle\right)^2\right]^k\right) \tag{4.38}$$

Further, using Fact 3.2.15, we have $\left\|\hat{\Sigma}^{\frac{1}{2}}\left(\hat{\Theta}-\tilde{\mathbb{E}}_{\tilde{\zeta}}\,\Theta\right)\right\|_2^{2k} \leqslant \tilde{\mathbb{E}}_{\tilde{\zeta}}\left[\left\|\hat{\Sigma}^{\frac{1}{2}}\left(\hat{\Theta}-\Theta\right)\right\|_2^{2k}\right]$. Substituting above and taking the $(1/2k)$-th root,

$$\begin{aligned}\left\|\hat{\Sigma}^{\frac{1}{2}}\left(\hat{\Theta}-\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta]\right)\right\|_2 &\leqslant \mathcal{O}\Big(\epsilon^{1-\frac{1}{k}}\,\lambda_k\Big)\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}\left[\mathbb{E}\left[\left(y_i'-\langle x_i',\Theta\rangle\right)^2\right]^k\right]+\mathbb{E}\left[\left(y_i^*-\left\langle x_i^*,\hat{\Theta}\right\rangle\right)^2\right]^k\right)^{1/2k}\\ &\leqslant \mathcal{O}\Big(\epsilon^{1-\frac{1}{k}}\,\lambda_k\Big)\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}\left[\mathbb{E}\left[\left(y_i'-\langle x_i',\Theta\rangle\right)^2\right]^k\right]^{\frac{1}{2k}}+\mathbb{E}\left[\left(y_i^*-\left\langle x_i^*,\hat{\Theta}\right\rangle\right)^2\right]^{\frac{1}{2}}\right)\end{aligned} \tag{4.39}$$

which concludes the proof. □

Lemma 4.4.6 (Bounding Optimization and Generalization Error)**.** *Under the hypothesis of Theorem 101,*

1. $\tilde{\mathbb{E}}_{\tilde{\zeta}}\left[\mathbb{E}\left[\left(y_i'-\langle x_i',\Theta\rangle\right)^2\right]^k\right]^{\frac{1}{2k}} \leqslant \mathbb{E}\left[y_i^*-\left\langle x_i^*,\hat{\Theta}\right\rangle^2\right]^{\frac{1}{2}}$, *and*
2. *For any* $\zeta>0$, *if* $n \geqslant n_0$, *such that* $n_0=\Omega\left(\max\{c_4 d/\zeta^2, d^{\mathcal{O}(k)}\}\right)$, *with probability at least* $1-1/\mathit{poly}(d)$, $\mathbb{E}\left[y_i^*-\left\langle x_i^*,\hat{\Theta}\right\rangle^2\right]^{\frac{1}{2}} \leqslant (1+\zeta)\,\mathbb{E}_{x,y\sim\mathcal{D}}\left[y-\langle x,\Theta^*\rangle^2\right]^{\frac{1}{2}}$.

Proof. We exhibit a degree-$\mathcal{O}(k)$ pseudo-distribution $\hat{\zeta}$ such that it is supported on a point mass and attains objective value at most $\mathbb{E}\left[y_i^*-\left\langle x_i^*,\hat{\Theta}\right\rangle^2\right]^{\frac{1}{2}}$. Since our objective function minimizes over all degree-$\mathcal{O}(k)$ pseudo-distributions, the resulting objective value w.r.t. $\tilde{\zeta}$ can only be better. Let $\hat{\zeta}$ be the pseudo-distribution supported on $(w,x^*,y^*,\hat{\Theta})$ such that $w_i=1$ if $x_i=x_i^*$ (i.e. the i-th sample is not corrupted.) It follows from $n \geqslant n_0$ and Lemma 4.4.2 that this assignment satisfies the constraint system $\mathcal{A}_{\epsilon,\lambda_k}$. Then, the objective value satisfies

$$\tilde{\mathbb{E}}_{\tilde{\zeta}}\left[\mathbb{E}\left[\left(y_i'-\langle x_i',\Theta\rangle\right)^2\right]^k\right] \leqslant \tilde{\mathbb{E}}_{\hat{\zeta}}\left[\mathbb{E}\left[\left(y_i'-\langle x_i',\Theta\rangle\right)^2\right]^k\right] = \mathbb{E}\left[\left(y_i^*-\left\langle x_i^*,\hat{\Theta}\right\rangle\right)^2\right]^k \tag{4.40}$$

Taking $(1/2k)$-th roots yields the first claim.

To bound the second claim, let $\mathcal{U}$ be the uniform distribution on the uncorrupted samples,

x_i^*, y_i^*. Observe, by optimality of $\hat{\Theta}$ on the uncorrupted samples, $\mathsf{err}_{\mathcal{U}}(\hat{\Theta}) \leqslant \mathsf{err}_{\mathcal{U}}(\Theta^*)$. Consider the random variable $z_i = (y_i^* - \langle x_i^*, \Theta^* \rangle)^2 - \mathbb{E}_{x,y\sim\mathcal{D}}\left[(y - \langle x, \Theta^* \rangle)^2\right]$. Since $\mathbb{E}\left[z_i\right] = 0$, we apply Chebyschev's inequality to obtain

$$\begin{aligned}\mathbf{Pr}\left[\frac{1}{n}\sum_{i\in[n]} z_i \geqslant \zeta\right] = \frac{\mathbb{E}\left[z_1^2\right]}{\zeta^2 n} &\leqslant \frac{\mathbb{E}\left[(y - \langle x, \Theta \rangle)^4\right]}{\zeta^2 n} \\ &\leqslant c_4 \frac{\mathsf{err}_{\mathcal{D}}(\Theta^*)^2}{n\zeta^2}\end{aligned}$$

Therefore, with probability at least $1 - \delta$,

$$\mathsf{err}_{\mathcal{U}}(\hat{\Theta}) \leqslant \left(1 + \sqrt{\frac{c_4}{n\delta}}\right) \mathsf{err}_{\mathcal{D}}(\Theta^*)$$

Therefore, setting $n = \Omega(c_4 d/\zeta^2)$, it follows that with probability $1 - 1/\mathrm{poly}(d)$, for any $\zeta > 0$,

$$\mathsf{err}_{\mathcal{U}}(\hat{\Theta}) \leqslant (1 + \zeta)\, \mathsf{err}_{\mathcal{D}}(\Theta^*)$$

Taking square-roots concludes the proof. □

Proof of Theorem 101. Given $n \geqslant n_0$ samples, it follows from Lemma 4.4.2, that with probability $1 - 1/\mathrm{poly}(d)$, the constraint system $\mathcal{A}_{\epsilon,\lambda_k}$ is feasible. Let ξ_1 be the event that the system is feasible and condition on it. Then, it follows from Lemma 4.4.5 and Lemma 4.4.6, with probability $1 - 1/\mathrm{poly}(d)$,

$$\left\|\hat{\Sigma}^{1/2}\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta] - \hat{\Theta}\right)\right\|_2 \leqslant \mathcal{O}\left(\lambda_k\, \epsilon^{1-1/k}\right) \mathsf{err}_{\mathcal{D}}(\Theta^*)^{1/2} \tag{4.41}$$

Let ξ_2 be the event that (4.41) holds and condition on it. It then follows from Fact 4.2.2, with probability $1 - 1/\mathrm{poly}(d)$,

$$\left\|(\Sigma^*)^{1/2}\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta] - \hat{\Theta}\right)\right\|_2 \leqslant \mathcal{O}\left(\lambda_k\, \epsilon^{1-1/k}\right) \mathsf{err}_{\mathcal{D}}(\Theta^*)^{1/2} \tag{4.42}$$

Let ξ_2 be the event that (4.42) holds and condition on it. It remains to relate the hyperplanes $\hat{\Theta}$ and Θ^*. By reverse triangle inequality,

$$\left\|(\Sigma^*)^{1/2}\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta] - \Theta^*\right)\right\|_2 - \left\|(\Sigma^*)^{1/2}\left(\Theta^* - \hat{\Theta}\right)\right\|_2 \leqslant \left\|(\Sigma^*)^{1/2}\left(\tilde{\mathbb{E}}_{\tilde{\zeta}}[\Theta] - \hat{\Theta}\right)\right\|_2$$

Using normal equations, we have $\hat{\Theta} = \hat{\Sigma}^{-1}\,\mathbb{E}\left[x_i y_i\right]$ and $\Theta^* = (\Sigma^*)^{-1}\,\mathbb{E}\left[xy\right]$. Since $\hat{\Sigma} \preceq (1 +$

$0.01)\Sigma^*$,

$$\begin{aligned}\left\|(\Sigma^*)^{1/2}\left(\Theta^*-\hat{\Theta}\right)\right\|_2 &= \left\|(\Sigma^*)^{1/2}\left(\hat{\Sigma}^{-1}\hat{\Sigma}\Theta^*-\hat{\Sigma}^{-1}\,\mathbb{E}\left[x_i y_i\right]\right)\right\|_2 \\ &= \left\|(\Sigma^*)^{1/2}\,\hat{\Sigma}^{-1}\left(\mathbb{E}\left[x_i\left(y_i-x_i^\top\Theta^*\right)\right]\right)\right\|_2 \\ &\leqslant 1.01\left\|\mathbb{E}\left[(\Sigma^*)^{-1/2}\,x_i\left(y_i-x_i^\top\Theta^*\right)\right]\right\|_2\end{aligned} \tag{4.43}$$

By Jensen's inequality

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i\in[n]}(\Sigma^*)^{-1/2}\,x_i\left(y_i-x_i^\top\Theta^*\right)\right\|_2\right] \leqslant \sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i\in[n]}(\Sigma^*)^{-1/2}\,x_i\left(y_i-x_i^\top\Theta^*\right)\right\|_2^2\right]}$$

Let $z_i = \sum_{i\in[n]}(\Sigma^*)^{-1/2}\,x_i\left(y_i-x_i^\top\Theta^*\right)$. Let $\left(\sum_{i\in[n]} z_i\right)_1$ denote the first coordinate of the vector. We bound the expectation of this coordinate as follows:

$$\begin{aligned}\mathbb{E}\left[(\sum_{i\in[n]} z_i)_1^2\right] &= \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i,i'\in[n]}\left((\Sigma^*)^{-1}x_i x_{i'}\right)_1\left(y_i-x_i^\top\Theta^*\right)\left(y_{i'}-x_{i'}^\top\Theta^*\right)\right] \\ &= \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i\in[n]}\left((\Sigma^*)^{-1}x_i^2\right)_1\left(y_i-x_i^\top\Theta^*\right)^2\right] \\ &= \frac{1}{n}\,\mathbb{E}\left[(\Sigma^*)^{-1}(x)_1^2\left(y-x^\top\Theta^*\right)\right]\end{aligned} \tag{4.44}$$

where the second equality follows from independence of the samples. Using negatively correlated moments, we have

$$\mathbb{E}\left[(\Sigma^*)^{-1}(x)_1^2\left(y-x^\top\Theta^*\right)^2\right] \leqslant \mathbb{E}\left[(\Sigma^*)^{-1}(x)_1^2\right]\mathbb{E}\left[\left(y-x^\top\Theta^*\right)^2\right]$$

Setting $v = (\Sigma^*)^{1/2}e_1$ and using Hypercontractivity of the covariates and the noise in the above equation,

$$\mathbb{E}\left[\Sigma^{-1}(x)_1^2\right]\mathbb{E}\left[\left(y-x^\top\Theta^*\right)^2\right] \leqslant \mathcal{O}\left(c_2^2\,\eta_2^2\right)\mathsf{err}_{\mathcal{D}}(\Theta^*) \tag{4.45}$$

Summing over the coordinates, and combining (4.44), (4.45), we obtain

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i\in[n]}(\Sigma^*)^{-1/2}\,x_i\left(y_i - x_i^\top\Theta^*\right)\right\|_2\right] \leqslant \mathcal{O}(c_2\eta_2)\sqrt{\frac{d\,\mathsf{err}_{\mathcal{D}}(\Theta^*)}{n}} \tag{4.46}$$

Applying Chebyschev's Inequality , with probability $1-\delta$

$$\left\|(\Sigma^*)^{1/2}\left(\Theta^* - \mathbf{E}_{\tilde{\zeta}}[\Theta]\right)\right\|_2 \leqslant \mathcal{O}\left(\lambda_k\,\epsilon^{1-1/k} + c_2\,\eta_2\sqrt{\frac{d}{\delta n}}\right)\mathsf{err}_{\mathcal{D}}(\Theta^*)^{1/2}$$

Since $n \geqslant n_0$, we can simplify the above bound and obtain the claim.

The running time of our algorithm is clearly dominated by computing a degree-$\mathcal{O}(k)$ pseudo-distribution satisfying $\mathcal{A}_{\epsilon,\lambda_k}$. Given that our constraint system consists of $O(n)$ variables and poly(n) constraints, it follows from Fact 3.2.13 that the pseudo-distribution $\tilde{\zeta}$ can be computed in $n^{\mathcal{O}(k)}$ time.

□

4.5 Lower bounds

In this section, we present information-theoretic lower bounds on the rate of convergence of parameter estimation and least-squares error for robust regression. Our constructions proceed by demonstrating two distributions over regression instances that are ϵ-close in total variation distance and the marginal distribution over the covariates is hypercontractive, yet the true hyperplanes are $f(\epsilon)$-far in scaled ℓ_2 distance.

4.5.1 True Linear Model

Consider the setting where there exists an optimal hyperplane Θ^* that is used to generate the data, with the addition of independent noise added to each sample, i.e.

$$y = \langle x, \Theta^*\rangle + \omega,$$

where ω is independent of x. Further, we assume that covariates and noise are hypercontractive. In this setting, Theorem 100 implies that we can recover a hyperplane close to Θ^* at a rate proportional to $\epsilon^{1-1/k}$. We show that this dependence is tight for $k = 4$. We note that independent noise is a special case of the distribution having negatively correlated moments.

Theorem 103 (True Linear Model Lower Bound, Theorem 98 restated)**.** *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^2 \times \mathcal{R}$ such that the marginal distribution over $\mathcal{R}^2$ has covariance Σ and is (c_k, k)-hypercontractive yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\sqrt{c_k}\ \sigma\ \epsilon^{1-1/k}\right)$, where Θ_1, Θ_2 be the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, $\sigma = \max(\textsf{err}_{\mathcal{D}_1}(\Theta_1), \textsf{err}_{\mathcal{D}_2}(\Theta_2)) < 1/\epsilon^{1/k}$ and the noise ω is uniform over $[-\sigma, \sigma]$.*

Proof. We construct a 2-dimensional instance where the marginal distribution over covariates is identical for $\mathcal{D}_1$ and $\mathcal{D}_2$. The pdf is given as follows: for $q \in \{1, 2\}$ on the first coordinate, x_1,

$$\mathcal{D}_q(x_1) = \begin{cases} 1/2, & \text{if } x_1 \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

and on the second coordinate, x_2,

$$\mathcal{D}_q(x_2) = \begin{cases} \epsilon/2, & \text{if } x_2 \in \{-1/\epsilon^{1/k}, 1/\epsilon^{1/k}\} \\ \frac{1-\epsilon}{2\epsilon\sigma} & \text{if } x_2 \in [-\epsilon\sigma, \epsilon\sigma] \\ 0 & \text{otherwise} \end{cases}$$

Next, we set $\Theta_1 = (1, 1)$, $\Theta_2 = (1, -1)$ and ω to be uniform over $[-\sigma, \sigma]$. Therefore,

$$\begin{aligned} \mathcal{D}_1(y \mid (x_1, x_2)) &= x_1 + x_2 + \omega \quad \text{and} \\ \mathcal{D}_2(y \mid (x_1, x_2)) &= x_1 - x_2 + \omega \end{aligned} \tag{4.47}$$

Observe, $\mathbb{E}\left[x_1^k\right] = \int_{-1}^{1} x^k/2 = 1/(k+1)$ and $\mathbb{E}\left[x_1^2\right] = \int_{-1}^{1} x^2/2 = 1/3$. Further,

$$\begin{aligned} \mathbb{E}\left[x_2^k\right] &= \frac{(1-\epsilon)}{\epsilon\sigma}\frac{(\epsilon\sigma)^{k+1}}{k+1} + \epsilon \cdot \left(\frac{1}{\epsilon^{1/k}}\right)^k = 1 + \frac{(1-\epsilon)}{(k+1)}(\epsilon\sigma)^k \\ \mathbb{E}\left[x_2^2\right] &= \frac{(1-\epsilon)}{3\epsilon\sigma}(\epsilon\sigma)^3 + \epsilon \cdot \left(\frac{1}{\epsilon^{1/k}}\right)^2 = \epsilon^{1-2/k} + \frac{1-\epsilon}{3}(\epsilon\sigma)^2 \end{aligned}$$

Observe, $\mathbb{E}\left[x_2^k\right] \leqslant (1/(c\epsilon^{k/2-1}))\ \mathbb{E}\left[x_2^2\right]^{k/2}$, for a fixed constant c. Then, for any unit vector v,

$$\begin{aligned} \mathbb{E}\left[\langle x, v\rangle^k\right] \leqslant \mathbb{E}\left[(2x_1v_1)^k + (2x_2v_2)^k\right] &\leqslant c_k^{k/2}\ \left(\mathbb{E}\left[(x_1v)^2\right]^{k/2} + \mathbb{E}\left[(x_2v)^2\right]^{k/2}\right) \\ &\leqslant c_k^{k/2}\ \mathbb{E}\left[\langle x, v\rangle^2\right]^{k/2} \end{aligned}$$

where $c_k^{k/2} = 2^k/c\epsilon^{k/2-1}$. Therefore, $\mathcal{D}_1, \mathcal{D}_2$ are (c_k, k)-hypercontractive over $\mathcal{R}^2$. Next, we

compute the TV distance between the two distributions.

$$\begin{aligned} d_{\mathsf{TV}}\left(\mathcal{D}_1, \mathcal{D}_2\right) &= \frac{1}{2}\int_{\mathcal{R}^2\times\mathcal{R}} |\mathcal{D}_1(x_1, x_2, y) - \mathcal{D}_2(x_1, x_2, y)| \\ &= \frac{1}{2}\int_{\mathcal{R}^2} \mathcal{D}_1(x_1, x_2) \int_{\mathcal{R}} |\mathcal{D}_1(y \mid (x_1, x_2)) - \mathcal{D}_2(y \mid (x_1, x_2))| \end{aligned} \tag{4.48}$$

where the last equality follows from the definition of conditional probability. It follows from Equation (4.47) that $\mathcal{D}_1(y \mid (x_1, x_2)) = \mathcal{U}(x_1 + x_2 - \sigma, x_1 + x_2 + \sigma)$ and $\mathcal{D}_2(y \mid (x_1, x_2)) = \mathcal{U}(x_1 - x_2 - \sigma, x_1 - x_2 + \sigma)$. If $|x_2| \geqslant \sigma$ the intervals are disjoint and $|\mathcal{D}_1(y \mid (x_1, x_2)) - \mathcal{D}_2(y \mid (x_1, x_2))| = 2$. If $|x_2| < \sigma$, then two symmetric non-intersecting regions have mass $2|x_2|/2\sigma$ and the intersection region contributes 0. Therefore, $|\mathcal{D}_1(y \mid (x_1, x_2)) - \mathcal{D}_2(y \mid (x_1, x_2))| = 2|x_2|/\sigma$ and (4.48) can be evaluated as

$$\begin{aligned} d_{\mathsf{TV}}(\mathcal{D}_1, \mathcal{D}_2) &= \frac{1}{2}\int_{\mathcal{R}} 2\mathbb{I}|x_2| \geqslant \sigma + \frac{2|x_2|}{\sigma}\mathbb{I}|x_2| < \sigma \\ &= \mathbf{Pr}\left[|x_2| \geqslant \sigma\right] + \frac{1}{\sigma}\mathop{\mathbb{E}}_{x_2\sim\mathcal{D}_1}\left[|x_2|\mathbb{I}|x_2| < \sigma\right] \\ &= 2\epsilon \end{aligned}$$

Finally, we lower bound the parameter distance. Since the coordinates are independent, Σ is a diagonal matrix with $\Sigma_{1,1} = \mathbb{E}\left[x_1^2\right] = 1/3$ and $\Sigma_{2,2} = \mathbb{E}\left[x_2^2\right] = \epsilon^{1-2/k} + (\epsilon\sigma)^2/3$. Further, $\Theta_1 - \Theta_2 = (0, 2)$. Thus, $\left\|\Sigma^{1/2}\left(\Theta_1 - \Theta_2\right)\right\|_2 = 2\Sigma_{2,2}^{1/2} \geqslant 2\epsilon^{1/2-1/k}$. For any $\sigma < 1/\epsilon^{1/k}$,

$$\begin{aligned} \left\|\Sigma^{1/2}\left(\Theta_1 - \Theta_2\right)\right\|_2 \geqslant 2\,\epsilon^{1/2-1/k} &> 2\,\sigma\,\epsilon^{1/2} \\ &\geqslant 2\,\sqrt{c_k}\,\sigma\,\epsilon^{1-1/k} \end{aligned}$$

which concludes the proof.

□

4.5.2 Agnostic Model

Next, consider the setting where we simply observe samples from $(x, y) \sim \mathcal{D}$, and our goal is to return is to return the minimizer of the squared error, given by $\Theta^* = \mathbb{E}\left[xx^\top\right]^{-1}\mathbb{E}\left[xy\right]$. Here, the distribution of the noise is allowed to depend on the covariates arbitrarily. We further assume the noise is hypercontractive and obtain a lower bound proportional to $\epsilon^{1-2/k}$ for recovering an estimator close to Θ^*. This matches the upper boundd obtained in Corollary 4.3.1.

Theorem 104 (Agnostic Model Lower Bound, Theorem 99 restated)**.** *For any $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R}^2 \times \mathcal{R}$ such that the marginal distribution over $\mathcal{R}^2$ has covariance Σ and is (c_k, k)-hypercontractive yet $\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \Omega\left(\sqrt{c_k}\ \sigma\ \epsilon^{1-2/k}\right)$, where Θ_1, Θ_2 be the optimal hyperplanes for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, $\sigma = \max(\mathsf{err}_{\mathcal{D}_1}(\Theta_1), \mathsf{err}_{\mathcal{D}_2}(\Theta_2)) < 1/\epsilon^{1/k}$ and the noise is a function of the marginal distribution of $\mathcal{R}^2$.*

Proof. We provide a proof for the special case of $k = 4$. The same proof extends to general k. We again construct a 2-dimensional instance where the marginal distribution over covariates is identical for $\mathcal{D}_1$ and $\mathcal{D}_2$. The pdf is given as follows: for $q \in \{1, 2\}$ on the first coordinate, x_1,

$$\mathcal{D}_q(x_1) = \begin{cases} 1/2, & \text{if } x_1 \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

and on the second coordinate, x_2,

$$\mathcal{D}_q(x_2) = \begin{cases} \epsilon/2, & \text{if } x_2 \in \{-1/\epsilon^{1/4}, 1/\epsilon^{1/4}\} \\ \frac{1-\epsilon}{2} & \text{if } x_2 \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

Observe, $\mathbb{E}\left[x_1^4\right] = 1/5$ and $\mathbb{E}\left[x_1^2\right] = 1/3$. Similarly, $\mathbb{E}\left[x_2^4\right] = 1 + (1-\epsilon)/5$ and $\mathbb{E}\left[x_2^2\right] = \sqrt{\epsilon} + (1-\epsilon)/3$. Therefore, the marginal distribution over $\mathcal{R}^2$ is $(c, 4)$-hypercontractive for a fixed constant c. Next, let

$$\begin{aligned} &\mathcal{D}_1(y \mid (x_1, x_2)) = x_2 \qquad \text{and} \\ &\mathcal{D}_2(y \mid (x_1, x_2)) = \begin{cases} 0 & \text{if } |x_2| = 1/\epsilon^{1/4} \\ x_2 & \text{otherwise} \end{cases} \end{aligned} \tag{4.49}$$

Then,

$$\begin{aligned} d_{\mathsf{TV}}(\mathcal{D}_1, \mathcal{D}_2) &= \frac{1}{2}\int_{\mathcal{R}^2} \mathcal{D}_1(x_1, x_2) \int_{\mathcal{R}} |\mathcal{D}_1(y \mid (x_1, x_2)) - \mathcal{D}_2(y \mid (x_1, x_2))| \\ &= \frac{1}{2}\int_{\mathcal{R}} |x_2| \mathbb{I}|x_2| = 1/\epsilon^{1/4} \\ &= \epsilon \end{aligned}$$

Since the coordinates over $\mathcal{R}^2$ are independent the covariance matrix Σ is diagonal, such that $\Sigma_{1,1} = \mathbb{E}\left[x_1^2\right] = 1/3$ and $\Sigma_{2,2} = \mathbb{E}\left[x_2^2\right] = \sqrt{\epsilon} + (1-\epsilon)/3$. We can then compute the optimal

hyperplanes using normal equations:

$$\Theta_1 = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_1}\left[xx^\top\right]^{-1} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_1}[xy] = \Sigma^{-1} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_1}[xy]$$

Observe, using (4.49),

$$\mathbb{E}[x_1 y] = \int_{\mathcal{R}} x_1 y \mathcal{D}_1(x_1 y) = \int_{\mathcal{R}} x_1 y \mathcal{D}_1(x_1)\mathcal{D}_1(y) = 0$$

since x_1 and y are independent. Further,

$$\mathbb{E}[x_2 y] = \int_{\mathcal{R}} x_2 y D(x_2, y) = \int_{\mathcal{R}} x_2^2 D(x_2) = \sqrt{\epsilon} + (1-\epsilon)/3$$

Therefore, $\Theta_1 = (0, 1)$. Similarly,

$$\Theta_2 = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_2}\left[xx^\top\right]^{-1} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_2}[xy] = \Sigma^{-1} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}_2}[xy]$$

Further, $\mathbb{E}[x_1 y] = 0$. However,

$$\mathbb{E}[x_2 y] = \int_{\mathcal{R}} x_2 y \mathcal{D}_2(x_2, y) = \int_{\mathcal{R}} x_2^2 \mathbb{I}|x_2| \leqslant 1 \mathcal{D}_2(x_2) = 1 - \epsilon$$

Therefore, $\Theta_2 = \left(0, \frac{1-\epsilon}{1+\sqrt{\epsilon}}\right)$. Then,

$$\left\|\Sigma^{1/2}(\Theta_1 - \Theta_2)\right\|_2 = \sqrt{\sqrt{\epsilon} + (1-\epsilon)/3} \cdot \frac{\sqrt{\epsilon}+\epsilon}{1+\sqrt{\epsilon}} = \Omega(\sqrt{\epsilon})$$

which concludes the proof. □

4.6 Bounded Covariance Distributions

In the heavy-tailed setting, the minimal assumption is to consider a distribution over the covariates with bounded covariance. In this setting, we show that robust estimators for linear regression do not exist, even when the underlying linear model has no noise, i.e. the uncorrupted samples are drawn as follows: $y_i = \langle \Theta^*, x_i \rangle$.

Theorem 105 (Lower Bound for Bounded Covariance Distributions)**.** *For all $\epsilon > 0$, there exist two distributions $\mathcal{D}_1, \mathcal{D}_2$ over $\mathcal{R} \times \mathcal{R}$ corresponding to the linear model $y = \langle \Theta_1, x \rangle$ and $y = \langle \Theta_2, x \rangle$ respectively, such that the marginal distribution over $\mathcal{R}$ has variance σ and*

$d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \leqslant \epsilon$, *yet* $|\sigma(\Theta_1 - \Theta_2)| = \Omega(\sigma)$.

Our hard instance relies on the so called *Student's t-distribution*, which has heavy tails when the degrees of freedom are close to 2.

Definition 4.6.1 (Student's t-distribution)**.** *Given $\nu > 1$, Student's t-distribution has the following probability density function:*

$$f_\nu(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where $\Gamma(z) = \int_0^\infty x^{z-1}e^x dx$, for $z \in \mathcal{R}$, is the Gamma function.

We use the following facts about Student's t-distribution:

Fact 4.6.2 (Mean and Variance)**.** *The mean of Student's t-distribution is $\mathbb{E}_{x\sim f_\nu}[x] = 0$ for $\nu > 1$ and undefined otherwise. The variance of Student's t-distribution is*

$$\mathop{\mathbb{E}}_{x\sim f_\nu}\left[x^2\right] = \begin{cases} \infty & \text{if } 1 < \nu \leqslant 2 \\ \frac{\nu}{\nu-2} & \text{if } 2 < \nu \\ \text{undefined} & \text{otherwise} \end{cases}$$

The intuition behind our lower bound is to construct a regression instance where the covariates are non-zero only on an ϵ-measure support and are heavy tailed when non-zero. As a consequence, the adversary can introduce a distinct valid regression instance by changing a different ϵ-measure of the support. It is then information-theoretically impossible to distinguish between the true and the planted models.

Proof of Theorem 105. We construct a 1-dimensional instance where the marginal distribution over covariates is identical for $\mathcal{D}_1$ and $\mathcal{D}_2$. The pdf is given as follows: for $q \in \{1, 2\}$ the marginal distribution on the covariates is given as follows:

$$\mathcal{D}_q(x) = \begin{cases} 1 - \epsilon, & \text{if } x = 0 \\ \epsilon \cdot f_{2+\epsilon}(x) & \text{otherwise} \end{cases}$$

The distribution of the labels is gives as follows:

$$\mathcal{D}_1\left(y \mid x\right) = x \;\text{ and }\; \mathcal{D}_2\left(y \mid x\right) = -x$$

Next, we compute the total variation distance between $\mathcal{D}_1$ and $\mathcal{D}_2$. Recall,

$$\begin{aligned} d_{\mathsf{TV}}(\mathcal{D}_1, \mathcal{D}_2) &= \frac{1}{2}\int_{\mathcal{R}\times\mathcal{R}} |\mathcal{D}_1(x,y) - \mathcal{D}_2(x,y)| \\ &= \frac{1}{2}\int_{\mathcal{R}} \mathcal{D}_1(x) \int_{\mathcal{R}} |\mathcal{D}_1(y \mid x) - \mathcal{D}_2(y \mid x)| \\ &= \frac{1}{2}\int_{\mathcal{R}} |\mathcal{D}_1(y \mid x) - \mathcal{D}_2(y \mid x)| \left(\mathbb{I}x = 0 + \mathbb{I}x \neq 0\right) \\ &= \frac{1}{2}\int_{\mathcal{R}} |2x| \mathbb{I}x \neq 0 \leqslant \epsilon \end{aligned} \tag{4.50}$$

Observe, since the regression instances have no noise, we can obtain a perfect fit by setting $\Theta_1 = 1$ and $\Theta_2 = -1$. Further, for $q \in \{1, 2\}$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_q}[x] = (1-\epsilon)\cdot 0 + \epsilon \cdot \mathop{\mathbb{E}}_{x\sim f_{2+\epsilon}}[x] = 0 \tag{4.51}$$

and

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_q}\left[x^2\right] = (1-\epsilon)\cdot 0 + \epsilon \cdot \mathop{\mathbb{E}}_{x\sim f_{2+\epsilon}}\left[x^2\right] = \epsilon \cdot \frac{2+\epsilon}{\epsilon} \tag{4.52}$$

Thus,

$$\left| \mathop{\mathbb{E}}_{x\sim\mathcal{D}_q}\left[x^2\right]^{1/2} (\Theta_1 - \Theta_2) \right| = (2+\epsilon)^{1/2} \cdot 2 = 2\sigma \tag{4.53}$$

which completes the proof. We note that the 4-th moment of $f_{2+\epsilon}(t)$ is infinite and thus it is not hypercontractive, even for $k = 4$. □

4.7 Robust Identifiability for Arbitrary Noise

Proof of Corollary 4.3.1. Consider a maximal coupling of $\mathcal{D}, \mathcal{D}'$ over $(x, y)\times(x', y')$, denoted by $\mathcal{G}$, such that the marginal of $\mathcal{G}$ (x, y) is $\mathcal{D}$, the marginal on (x', y') is $\mathcal{D}'$ and $\mathbb{P}_{\mathcal{G}}[\mathbb{I}(x, y) = (x', y')] = 1 - \epsilon$. Then, for all v,

$$\begin{aligned} \langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}) + xy - xy\right\rangle\right] \\ &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(\langle x, \Theta_{\mathcal{D}}\rangle - y\right)\rangle\right] + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle\right] \end{aligned} \tag{4.54}$$

Since $\Theta_{\mathcal{D}}$ is the minimizer for the least squares loss, we have the following gradient condition : for all $v \in \mathbb{R}^d$,

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\langle v, (\langle x, \Theta_{\mathcal{D}}\rangle - y)x\rangle\right] = 0 \tag{4.55}$$

Since $\mathcal{G}$ is a coupling, using the gradient condition (4.55) and using that $1 = \mathbb{I}(x,y) = (x',y') + \mathbb{I}(x,y) \neq (x',y')$, we can rewrite equation (4.54) as

$$\begin{aligned}
\langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle &= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x,y) = (x',y')\right] \\
&\quad + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x,y) \neq (x',y')\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x,y) = (x',y')\right] \\
&\quad + \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y - \langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x,y) \neq (x',y')\right]
\end{aligned} \tag{4.56}$$

Consider the first term in the last equality above. Using the gradient condition for $\Theta_{\mathcal{D}'}$ along with Hölder's Inequality, we have

$$\begin{aligned}
&\left|\mathbb{E}_{\mathcal{G}}\left[\,\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x,y) = (x',y')\right]\right| \\
&\quad = \left|\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle\right] - \mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x,y) \neq (x',y')\right]\right| \\
&\quad = \left|\mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle\, \mathbb{I}(x,y) \neq (x',y')\right]\right| \\
&\quad \leqslant \left|\mathop{\mathbb{E}}_{\mathcal{G}}\left[\mathbb{I}(x,y) \neq (x',y')^{k/(k-2)}\right]^{(k-2)/k}\right| \cdot \left|\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\langle v, x'\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)\rangle^{k/2}\right]^{2/k}\right|
\end{aligned} \tag{4.57}$$

Observe, since $\mathcal{G}$ is a maximal coupling $\mathbb{E}_{\mathcal{G}}\left[\mathbb{I}(x,y) \neq (x',y')\right]^{(k-2)/k} \leqslant \epsilon^{1-2/k}$. Here, we no longer have independence of the noise and the covariates, therefore using Cauchy-Schwarz

$$\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\langle v, x'\rangle^{k/2} \cdot \left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^{k/2}\right] \leqslant \left(\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\langle v, x'\rangle^{k}\right] \mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^{k}\right]\right)^{1/2}$$

By hypercontractivity of the covariates and the noise, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}'}\left[\langle v, x'\rangle^{k}\right]^{1/k} \mathop{\mathbb{E}}_{\mathcal{D}'}\left[\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^{k}\right]^{1/k} \leqslant \mathcal{O}(c_k\, \eta_k)\left(v^\top \Sigma_{\mathcal{D}'} v\right)^{1/2} \mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y' - \langle x', \Theta_{\mathcal{D}'}\rangle\right)^2\right]^{1/2}$$

Therefore, we can restate (4.57) as follows

$$\left| \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x' \left(y' - \langle x', \Theta_{\mathcal{D}'} \rangle \right) \rangle \, \mathbb{I}(x, y) = (x', y') \right] \right| \leqslant \mathcal{O}\Big(c_k \, \eta_k \, \epsilon^{\frac{k-2}{k}} \Big) \left(v^\top \Sigma_{\mathcal{D}'} v \right)^{\frac{1}{2}} \mathop{\mathbb{E}}_{x', y' \sim \mathcal{D}'} \left[\left(y' - \langle x', \Theta_{\mathcal{D}'} \rangle \right)^2 \right]^{\frac{1}{2}} \tag{4.58}$$

It remains to bound the second term in the last equality of equation (4.56), and we proceed as follows :

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(y - \langle x, \Theta_{\mathcal{D}'} \rangle \right) \rangle \, \mathbb{I}(x, y) \neq (x', y') \right] &= \mathop{\mathbb{E}}_{\mathcal{G}} \left[\left\langle v, x x^\top \left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'} \right) \right\rangle \mathbb{I}(x, y) \neq (x', y') \right] \\ &\quad + \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right) \rangle \, \mathbb{I}(x, y) \neq (x', y') \right] \end{aligned} \tag{4.59}$$

We bound the two terms above separately. Observe, applying Hölder's Inequality to the first term, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\left\langle v, x x^\top \left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'} \right) \right\rangle \mathbb{I}(x, y) \neq (x', y') \right] &\leqslant \mathop{\mathbb{E}}_{\mathcal{G}} \left[\mathbb{I}(x, y) \neq (x', y') \right]^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\left\langle v, x x^\top \left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'} \right) \right\rangle^{\frac{k}{2}} \right]^{\frac{2}{k}} \\ &\leqslant \epsilon^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\left\langle v, x x^\top \left(\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'} \right) \right\rangle^{\frac{k}{2}} \right]^{\frac{2}{k}} \end{aligned} \tag{4.60}$$

To bound the second term in equation 4.59, we again use Hölder's Inequality followed by Cauchy-Schwarz noise and covariates.

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right) \rangle \, \mathbb{I}(x, y) \neq (x', y') \right] &\leqslant \mathop{\mathbb{E}}_{\mathcal{G}} \left[\mathbb{I}(x, y) \neq (x', y') \right]^{\frac{k-1}{k}} \mathop{\mathbb{E}}_{\mathcal{G}} \left[\langle v, x \left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right) \rangle^k \right]^{\frac{1}{k}} \\ &\leqslant \epsilon^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\langle v, x \rangle^{k/2} \right]^{2/k} \mathop{\mathbb{E}}_{x, y \sim \mathcal{D}} \left[\left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right)^{k/2} \right]^{2/k} \\ &\leqslant \epsilon^{\frac{k-2}{k}} \, c_k \, \eta_k \left(v^\top \Sigma_{\mathcal{D}} v \right)^{1/2} \mathop{\mathbb{E}}_{x, y \sim \mathcal{D}} \left[\left(y - \langle x, \Theta_{\mathcal{D}} \rangle \right)^2 \right]^{1/2} \end{aligned} \tag{4.61}$$

where the last inequality follows from hypercontractivity of the covariates and noise. Substituting

the upper bounds obtained in Equations (4.60) and (4.61) back in to (4.59),

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\langle v, x\left(y-\langle x, \Theta_{\mathcal{D}'}\rangle\right)\rangle \mathbb{I}(x,y) \neq (x',y')\right] \leqslant \epsilon^{\frac{k-2}{k}} \mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{\frac{k}{2}}\right]^{\frac{2}{k}} \\ + \epsilon^{\frac{k-2}{k}}\, c_k\, \eta_k \left(v^\top \Sigma_{\mathcal{D}} v\right)^{1/2} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x, \Theta_{\mathcal{D}}\rangle\right)^2\right]^{1/2}\end{aligned}$$

Therefore, we can now upper bound both terms in Equation (4.56) as follows:

$$\begin{aligned}\langle v, \Sigma_{\mathcal{D}}(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'})\rangle \leqslant \mathcal{O}\Big(c_k\, \eta_k\, \epsilon^{\frac{k-2}{k}}\Big)\left(v^\top \Sigma_{\mathcal{D}'} v\right)^{1/2} \mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y'-\langle x', \Theta_{\mathcal{D}'}\rangle\right)^2\right]^{1/2} \\ + \mathcal{O}\Big(\epsilon^{\frac{k-2}{k}}\Big)\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{k/2}\right]^{2/k} \\ + \mathcal{O}\Big(\epsilon^{\frac{k-2}{k}}\, c_k\, \eta_k\Big)\left(v^\top \Sigma_{\mathcal{D}} v\right)^{1/2} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x, \Theta_{\mathcal{D}}\rangle\right)^2\right]^{1/2}\end{aligned} \tag{4.62}$$

Recall, since the marginals of $\mathcal{D}$ and $\mathcal{D}'$ on $\mathcal{R}^d$ are (c_k, k)-hypercontractive and $\|\mathcal{D}-\mathcal{D}'\|_{\mathsf{TV}} \leqslant \epsilon$, it follows from Fact 4.2.3 that

$$(1-0.1)\,\Sigma_{\mathcal{D}'} \preceq \Sigma_{\mathcal{D}} \preceq (1+0.1)\,\Sigma_{\mathcal{D}'} \tag{4.63}$$

when $\epsilon \leqslant \mathcal{O}\Big((1/c_k k)^{k/(k-2)}\Big)$. Now, consider the substitution $v = \Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}$. Observe,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{G}}\left[\left\langle v, xx^\top\left(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'}\right)\right\rangle^{k/2}\right]^{2/k} &= \mathop{\mathbb{E}}_{\mathcal{D}}\left[\langle x, (\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'})\rangle^k\right]^{2/k} \\ &\leqslant c_k^2 \left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'})\right\|_2^2\end{aligned} \tag{4.64}$$

Then, using the bounds in (4.63) and (4.64) along with $v = \Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}$ in Equation 4.62, we have

$$\begin{aligned}\Big(1-\mathcal{O}\Big(\epsilon^{\frac{k-2}{k}} c_k^2\Big)\Big)\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'})\right\|_2^2 \leqslant \mathcal{O}\Big(c_k\, \eta_k\, \epsilon^{\frac{k-2}{k}}\Big)\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'})\right\|_2 \\ \left(\mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'}\left[\left(y'-\langle x', \Theta_{\mathcal{D}'}\rangle\right)^2\right]^{\frac{1}{2}} + \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}}\left[\left(y-\langle x, \Theta_{\mathcal{D}}\rangle\right)^2\right]^{\frac{1}{2}}\right)\end{aligned} \tag{4.65}$$

Dividing out (4.65) by $\Big(1-\mathcal{O}\Big(\epsilon^{\frac{k-2}{k}} c_k^2\Big)\Big)\left\|\Sigma_{\mathcal{D}}^{1/2}(\Theta_{\mathcal{D}}-\Theta_{\mathcal{D}'})\right\|_2^2$ and observing that $\mathcal{O}\Big(\epsilon^{\frac{k-2}{k}} c_k^2\Big)$ is upper bounded by a fixed constant less than 1 yields the parameter recovery bound.

Given the parameter recovery result above, we bound the least-squares loss between the two

hyperplanes on $\mathcal{D}$ as follows:

$$\begin{aligned}\left|\mathsf{err}_{\mathcal{D}}(\Theta_{\mathcal{D}}) - \mathsf{err}_{\mathcal{D}}(\Theta_{\mathcal{D}'})\right| &= \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\left(y - x^\top \Theta_{\mathcal{D}}\right)^2 - \left(y - x^\top \Theta_{\mathcal{D}'} + x^\top \Theta_{\mathcal{D}} - x^\top \Theta_{\mathcal{D}}\right)^2 \right] \right| \\ &= \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\langle x, (\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'})\rangle^2 + 2(y - x^\top \Theta_{\mathcal{D}}) x^\top (\Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}) \right] \right| \\ &\leqslant \mathcal{O}\Big(c_k^2\, \eta_k^2\, \epsilon^{2-4/k}\Big) \left(\mathop{\mathbb{E}}_{x',y'\sim\mathcal{D}'} \left[(y' - \langle x', \Theta_{\mathcal{D}'}\rangle)^2 \right] + \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}} \left[(y - \langle x, \Theta_{\mathcal{D}}\rangle)^2 \right] \right) \end{aligned} \tag{4.66}$$

where the last inequality follows from observing $\mathbb{E}\left[\left\langle \Theta_{\mathcal{D}} - \Theta_{\mathcal{D}'}, x(y - x^\top \Theta_{\mathcal{D}})\right\rangle\right] = 0$ (gradient condition) and squaring the parameter recovery bound. □

4.8 Efficient Estimator for Arbitrary Noise

In this section, we provide a proof of the key SoS lemma required to obtain a polynomial time estimator. The remainder of the proof, including the feasibility of the constraints and rounding is identical to the one presented in Section 4.4.

Lemma 4.8.1 (Robust Identifiability in SoS for Arbitrary Noise)**.** *Consider the hypothesis of Theorem 101. Let w, x', y' and Θ be feasible solutions for the polynomial constraint system $\mathcal{A}$. Let $\hat{\Theta} = \arg\min_\Theta \frac{1}{n}\sum_{i\in[n]}(y_i^* - \langle x_i^*, \Theta\rangle)^2$ be the empirical loss minimizer on the uncorrupted samples and let $\hat{\Sigma} = \mathbb{E}\left[x_i^*(x_i^*)^\top\right]$ be the covariance of the uncorrupted samples. Then,*

$$\begin{aligned}\mathcal{A} \vdash_{4k}^{w,x',y',\Theta} \Bigg\{ \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta} - \Theta\right)\right\|_2^{2k} \leqslant\; & 2^{3k}(2\epsilon)^{k-2} c_k^k\, \eta_k^k\, \sigma^{k/2} \left\|\mathbb{E}\left[x_i'(x_i')^\top\right]^{1/2}\left(\hat{\Theta} - \Theta\right)\right\|_2^k \\ & + 2^{3k}(2\epsilon)^{k-2} c_k^{2k} \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta} - \Theta\right)\right\|_2^{2k} \\ & + 2^{3k}(2\epsilon)^{k-2} c_k^k\, \eta_k^k\, \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^2\right]^{k/2} \left\|\hat{\Sigma}^{1/2}\left(\hat{\Theta} - \Theta\right)\right\|_2^k \Bigg\}\end{aligned}$$

Proof. Consider the empirical covariance of the uncorrupted set given by $\hat{\Sigma} = \mathbb{E}\left[x_i^*(x_i^*)^\top\right]$.

Then, using the substitution, along with Fact 2.2.8

$$\begin{aligned}
\left|\frac{\Theta}{2k}\right. \left\{ \left\langle v, \hat{\Sigma}\left(\hat{\Theta}-\Theta\right)\right\rangle^k &= \left\langle v, \mathbb{E}\left[x_i^*(x_i^*)^\top\left(\hat{\Theta}-\Theta\right) + x_i^* y_i^* - x_i^* y_i^*\right]\right\rangle^k \\
&= \left\langle v, \mathbb{E}\left[x_i^*\left(\left\langle x_i^*, \hat{\Theta}\right\rangle - y_i^*\right)\right] + \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \\
&\leqslant 2^k \left\langle v, \mathbb{E}\left[x_i^*\left(\left\langle x_i^*, \hat{\Theta}\right\rangle - y_i^*\right)\right]\right\rangle^k + 2^k \left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \Bigg\}
\end{aligned} \tag{4.67}$$

Since $\hat{\Theta}$ is the minimizer of $\mathbb{E}\left[\left(\langle x_i^*, \Theta\rangle - y_i^*\right)^2\right]$, the gradient condition (appearing in Equation (4.55) of the indentifiability proof) implies this term is 0. Therefore, it suffices to bound the second term.

For all $i \in [n]$, let $w_i' = w_i$ iff the i-th sample is uncorrupted in $\mathcal{X}_\epsilon$, i.e. $x_i = x_i^*$. Then, it is easy to see that $\sum_i w_i' \geqslant (1-2\epsilon)n$. Further, since $\mathcal{A}\left|\frac{w}{2}\right. \{(1-w_i'w_i)^2 = (1-w_i'w_i)\}$,

$$\mathcal{A}\left|\frac{w}{2}\right. \left\{ \frac{1}{n}\sum_{i\in[n]}(1-w_i'w_i)^2 = \frac{1}{n}\sum_{i\in[n]}(1-w_i'w_i) \leqslant 2\epsilon \right\} \tag{4.68}$$

The above equation bounds the uncorrupted points in $\mathcal{X}_\epsilon$ that are not indicated by w. Then, using the substitution, along with the SoS Almost Triangle Inequality (Fact 2.2.8),

$$\begin{aligned}
\mathcal{A}\left|\frac{\Theta, w'}{2k}\right. \left\{ \left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k &= \left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\left(w_i' + 1 - w_i'\right)\right)\right]\right\rangle^k \\
&= \left\langle v, \mathbb{E}\left[w_i' x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right] + \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \\
&\leqslant 2^k \left\langle v, \mathbb{E}\left[w_i' x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \\
&\quad + 2^k \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \Bigg\}
\end{aligned} \tag{4.69}$$

Consider the first term of the last inequality in (4.69). Observe, since $w_i' x_i^* = w_i w_i' x_i'$ and

similarly, $w_i' y_i^* = w_i w_i' y_i'$,

$$\mathcal{A} \left|\frac{\Theta, w'}{4}\right. \left\{ \mathbb{E}\left[w_i' x_i^* \left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right] = \mathbb{E}\left[w_i' w_i x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right] \right\}$$

For the sake of brevity, the subsequent statements hold for relevant SoS variables and have degree $O(k)$ proofs. Using the substitution,

$$\begin{aligned}
\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[w_i' x_i^* \left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k &= \left\langle v, \mathbb{E}\left[w_i' w_i x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \\
&= \left\langle v, \mathbb{E}\left[x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right] + \mathbb{E}\left[(1 - w_i' w_i) x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \\
&\leqslant 2^k \left\langle v, \mathbb{E}\left[x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \\
&\quad + 2^k \left\langle v, \mathbb{E}\left[(1 - w_i' w_i) x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k \Bigg\}
\end{aligned} \tag{4.70}$$

Observe, the first term in the last inequality above is identically 0, since we enforce the gradient condition on the SoS variables x', y' and Θ. We can then rewrite the second term using linearity of expectation, followed by applying SoS Hölder's Inequality (Fact 3.2.20) combined with $\mathcal{A} \left|\frac{w}{2}\right. \{(1 - w_i' w_i)^2 = 1 - w_i' w_i\}$ to get

$$\begin{aligned}
\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1 - w_i' w_i) x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\right]\right\rangle^k &= \mathbb{E}\left[\langle v, (1 - w_i') w_i x_i' \left(y_i' - \langle x_i', \Theta\rangle\right)\rangle\right]^k \\
&= \mathbb{E}\left[(1 - w_i' w_i) \langle v, x_i'\rangle \left(y_i' - \langle x_i', \Theta\rangle\right)\right]^k \\
&\leqslant \mathbb{E}\left[(1 - w_i' w_i)\right]^{k-2} \mathbb{E}\left[\langle v, x_i'\rangle^{k/2} \left(y_i' - \langle x_i', \Theta\rangle\right)^{k/2}\right] \\
&\leqslant (2\epsilon)^{k-2} \mathbb{E}\left[\langle v, x_i'\rangle^k\right] \mathbb{E}\left[\left(y_i' - \langle x_i', \Theta\rangle\right)^k\right] \Bigg\}
\end{aligned} \tag{4.71}$$

where the last inequality follows from (4.68) and the SoS Cauchy Schwarz Inequality. Using the certifiable-hypercontractivity of the covariates,

$$\mathcal{A} \left|\frac{w, x'}{2k}\right. \left\{ \mathbb{E}\left[\langle v, x_i'\rangle^k\right] \leqslant c_k^k \mathbb{E}\left[\langle v, x_i'\rangle^2\right]^{k/2} = c_k^k \left\langle v, \mathbb{E}\left[x_i' (x_i')^\top\right] v\right\rangle^{k/2} \right\} \tag{4.72}$$

Further, using certifiable hypercontractivity of the noise,

$$\mathcal{A} \vdash \left\{ \mathbb{E}\left[(y_i' - \langle w_i x_i', \Theta \rangle)^k \right] \leqslant \eta_k^k \, \mathbb{E}\left[(y_i' - \langle x_i', \Theta \rangle)^2) \right]^{k/2} \right\} \tag{4.73}$$

Recall, $\sigma = \mathbb{E}\left[(y_i' - \langle x_i', \Theta \rangle)^2) \right]$ Combining the upper bounds obtained in (4.72) and (4.73), and plugging this back into (4.71), we get

$$\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1 - w_i') x_i' \left(y_i' - \langle x_i', \Theta \rangle \right) \right] \right\rangle^k \leqslant (2\epsilon)^{k-2} c_k^k \, \eta_k^k \, \sigma^{k/2} \left\langle v, \mathbb{E}\left[x_i' (x_i')^\top \right] v \right\rangle^{k/2} \right\} \tag{4.74}$$

Recall, we have now bounded the first term of the last inequality in (4.69). Therefore, it remains to bound the second term of the last inequality in (4.69). Using the substitution, we have

$$\begin{aligned} \mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1 - w_i') x_i^* \left(y_i^* - \langle x_i^*, \Theta \rangle \right) \right] \right\rangle^k &= \left\langle v, \mathbb{E}\left[(1 - w_i') x_i^* \left(y_i^* - \left\langle x_i^*, \Theta - \hat{\Theta} + \hat{\Theta} \right\rangle \right) \right] \right\rangle^k \\ &\leqslant 2^k \left\langle v, \mathbb{E}\left[(1 - w_i') x_i^* \left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right) \right] \right\rangle^k \\ &\quad + 2^k \left\langle v, \mathbb{E}\left[(1 - w_i') x_i^* \left(\left\langle x_i^*, \Theta - \hat{\Theta} \right\rangle \right) \right] \right\rangle^k \Bigg\} \end{aligned} \tag{4.75}$$

We again handle each term separately. Observe, the first term when decoupled is a statement about the uncorrupted samples. Therefore, using the SoS Hölder's Inequality (Fact 3.2.20),

$$\begin{aligned} \mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1 - w_i') x_i^* \left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right) \right] \right\rangle^k &= \mathbb{E}\left[(1 - w_i') \left\langle v, x_i^* \left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right) \right\rangle \right]^k \\ &\leqslant \mathbb{E}\left[(1 - w_i') \right]^{k-2} \mathbb{E}\left[\left\langle v, x_i^* \left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right) \right\rangle^{k/2} \right] \\ &\leqslant (2\epsilon)^{k-2} \, \mathbb{E}\left[\langle v, x_i^* \rangle^k \right] \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta} \right\rangle \right)^k \right] \Bigg\} \end{aligned} \tag{4.76}$$

Using certifiable hypercontractivity of the x_i^*s,

$$\mathbb{E}\left[\langle v, x_i^* \rangle^k \right] \leqslant c_k^k \, \mathbb{E}\left[\langle v, x_i^* \rangle^2 \right]^{k/2} = c_k^k \left\langle v, \hat{\Sigma} v \right\rangle^{k/2}$$

where $\hat{\Sigma} = \mathbb{E}\left[x_i^*(x_i^*)^\top\right]$ and similarly using hypercontractivity of the noise,

$$\mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^k\right] \leqslant \eta_k^k \, \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^2\right]^{k/2}$$

Then, by the substitution, we can bound (4.76) as follows:

$$\mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)\right]\right\rangle^k \leqslant (2\epsilon)^{k-1} c_k^k \, \eta_k^k \, \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^2\right]^{k/2} \left\langle v, \hat{\Sigma} v\right\rangle^{k/2} \right\} \tag{4.77}$$

In order to bound the second term in (4.75), we use the SoS Hölder's Inequality,

$$\begin{aligned} \mathcal{A} \vdash \left\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(\left\langle x_i^*, \Theta - \hat{\Theta}\right\rangle\right)\right]\right\rangle^k &= \mathbb{E}\left[(1-w_i')^{k-2}\left\langle v, x_i^*\left(\left\langle x_i^*, \Theta - \hat{\Theta}\right\rangle\right)\right\rangle\right] \\ &\leqslant \mathbb{E}\left[1-w_i'\right]^{k-2} \mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \\ &\leqslant (2\epsilon)^{k-2} \, \mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \Bigg\} \end{aligned} \tag{4.78}$$

Combining the bounds obtained in (4.77) and (4.78), we can restate Equation (4.75) as follows

$$\begin{aligned} \mathcal{A} \vdash \Bigg\{ \left\langle v, \mathbb{E}\left[(1-w_i')x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \leqslant{}& 2^k (2\epsilon)^{k-1} c_k^k \, \eta_k^k \, \mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^2\right]^{k/2} \left\langle v, \hat{\Sigma} v\right\rangle^{k/2} \\ &+ 2^k (2\epsilon)^{k-2} \, \mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta - \hat{\Theta})\right)^{\frac{k}{2}}\right]^2 \Bigg\} \end{aligned} \tag{4.79}$$

Combining (4.79) with (4.74), we obtain an upper bound for the last inequality in Equation

(4.69). Therefore, using the substitution, we obtain

$$\mathcal{A}\vdash \left\{\left\langle v, \mathbb{E}\left[x_i^*\left(y_i^* - \langle x_i^*, \Theta\rangle\right)\right]\right\rangle^k \leqslant 2^k(2\epsilon)^{k-1}c_k^k\,\eta_k^k\,\sigma^{k/2}\left\langle v, \mathbb{E}\left[x_i'(x_i')^\top\right]v\right\rangle^{k/2} + 2^{2k}(2\epsilon)^{k-2}\,\mathbb{E}\left[\left(v^\top x_i^*(x_i^*)^\top(\Theta-\hat{\Theta})\right)^{\frac{k}{2}}\right]^2 + 2^{2k}(2\epsilon)^{k-1}c_k^k\,\eta_k^k\,\mathbb{E}\left[\left(y_i^* - \left\langle x_i^*, \hat{\Theta}\right\rangle\right)^2\right]^{k/2}\left\langle v, \hat{\Sigma}v\right\rangle^{k/2}\right\} \tag{4.80}$$

The remaining proof is identical to Lemma 4.4.3. □

4.9 Proof of Lemma 4.2.4

Lemma 4.9.1 (Löwner Ordering for Hypercontractive Samples (restated))**.** *Let $\mathcal{D}$ be a (c_k, k)-hypercontractive distribution with covariance Σ and and let $\mathcal{U}$ be the uniform distribution over n samples. Then, with probability $1-\delta$,*

$$\left\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I\right\|_F \leqslant \frac{C_4 d^2}{\sqrt{n}\sqrt{\delta}},$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i\in[n]} x_i x_i^\top$.

Proof. Let $\tilde{x_i} = \Sigma^{-1/2}x_i$ and observe that $\frac{1}{n}\sum_i \tilde{x_i}\tilde{x_i}^T = \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}$. Moreover, we know that $\mathbb{E}\left[\tilde{x}\tilde{x}^T\right] = I$. Let $z_{j,k}$ be the (j,k) entry of $\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I$ given by,

$$z_{j,k} = \frac{1}{n}\sum_{i\in[n]} \tilde{x_i}(j)\tilde{x_i}(k) - \mathbb{E}\left[\tilde{x}(i)\tilde{x}(k)\right]$$

Using Chebyshev's inequality, we get that with probability at least $1-\delta$,

$$|z_{jk}| \leqslant \frac{\mathbb{E}\left[\tilde{x}(j)^2\tilde{x}(k)^2\right]}{\sqrt{n}\sqrt{\delta}} \leqslant \frac{\mathbb{E}_{\tilde{x}(j)^4}[+]\,\mathbb{E}\left[\tilde{x}(k)^4\right]}{2\sqrt{n}\sqrt{\delta}},$$

where the inequality follows from AM-GM inequality. To bound $\mathbb{E}\left[\tilde{x}(j)^4\right]$, we use hypercontractivity.

$$\mathop{\mathbb{E}}_{\tilde{x}(j)^4}[=]\,\mathbb{E}\left[(v^Tx)^4\right] \leqslant C_4\,\mathbb{E}\left[(v^Tx)^2\right]^2,$$

where $v = \Sigma^{-1/2} e_j$. Plugging this above, we get that $\mathbb{E}\left[\tilde{x}(j)^4\right] \leqslant C_4$. which in turn implies that with probability at least $1 - \delta$,

$$|z_{jk}| \leqslant \frac{C_4}{\sqrt{n\delta}}.$$

Taking a union bound over d^2 entries of $\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I$, we get that with probability at least $1 - \delta$,

$$\left\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I\right\|_F \leqslant \frac{C_4 d^2}{\sqrt{n}\sqrt{\delta}}$$

□

Chapter 5

List-Decodable Subspace Recovery

5.1 Introduction

In this chapter, we focus on the harsher *list-decodable estimation* model where the fraction of inliers α is $\ll 1/2$ - i.e.,a majority of the input sample are outliers. First considered in [BBV08] in the context of clustering, this was proposed as a model for *untrusted* data in a recent influential recent work of Charikar, Steinhardt and Valiant [CSV17]. Since unique recovery is information-theoretically impossible in this setting, the goal is to recover a small (ideally $O(1/\alpha)$) size list of parameters one of which is guaranteed to be close to those of the inlier distribution. A recent series of works have resulted in a high-level blueprint based on the *sum-of-squares method* for list-decodable estimation yielding algorithms for list-decodable mean estimation [DKS18] and linear regression [KKK19, RY20a].

We extend this line of work by giving the first efficient algorithm for *list-decodable subspace recovery*. In this setting, we are given data with α fraction inliers generated i.i.d. according $\mathcal{N}(0, \Sigma_*)$[1] on $\mathcal{R}^d$ with a (possibly low-rank, say $r < d$) covariance matrix Σ_* and rest being arbitrary outliers. We give an algorithm that succeeds in returning a list of size $O(1/\alpha)$ that contains a $\hat{\Pi}$ satisfying $\left\|\hat{\Pi} - \Pi_*\right\|_F^2 \leqslant \log(r\kappa)\tilde{O}(\kappa^4/\alpha^2))$ where Π_* is the projector to the range space of Σ_* and κ is the ratio of the largest to smallest non-zero eigenvalues of Σ_*. Our Frobenius norm recovery guarantees are the strongest possible and imply guarantees in other well-studied norms such as spectral norm or principle angle distance between subspaces. Our algorithm runs in time $n^{\log(r\kappa)\tilde{O}(1/\alpha^4)}$ and requires $n = d^{\log(r\kappa)\tilde{O}(1/\alpha^2)}$ samples.

[1] Our techniques naturally extend to distributions with non-zero means but we will omit this generalization to not complicate the notation.

Our results work more generally for any distribution D that satisfies *certifiable anti-concentration* and mild concentration properties (concentration of PSD forms). Certifiable anti-concentration was first defined and studied in recent works on list-decodable regression [RY20a, KKK19]. Gaussian distribution and uniform distribution on sphere (restricted to a subspace) are natural examples of distributions satisfying this property. We note that Karmalkar et. al. [KKK19] proved that anti-concentration of D is *necessary* for list-decodable regression (and thus also subspace recovery) to be information theoretically possible.

Why List-Decodable Estimation? List-decodable estimation is a strict generalization of related and well-studied *clustering* problems (for e.g., list-decodable mean estimation generalizes clustering spherical mixture models, list-decodable regression generalizes mixed linear regression). In our case, list-decodable subspace recovery generalizes the well-studied problem of subspace clustering where given a mixtur of k distributions with covariances non-zero in different subspaces, the goal is to recover the underlying k subspaces [AGGR98, CFZ99, GNC99, PJAM02, AY00]. Algorithms in this model thus naturally yield robust algorithms for the related clustering formulations. In contrast to known results, such algorithms allow "partial recovery" (e.g. for example recovery $k-1$ or fewer clusters) even in the presence of outliers that garble up one or more clusters completely.

Another important implication of list-decodable estimation is algorithms for *unique recovery* that work all the way down to the information-theoretic threshold (i.e. fraction of inliers $\alpha > 1/2$). Thus, specifically in our case, we obtain an algorithm for (uniquely) estimating the subspace spanned by the inlier distribution D whenever the fraction of inliers satisfy $\alpha > 1/2$ - the information theoretically minimum possible value. We note that such a result will follow from outlier-robust covariance estimation algorithms [LRV16, DKK+19] whenever α is sufficiently close to 1. While prior works do not specify precise constants, all known works appear to require α at least > 0.75.

5.1.1 Our Results

We are ready to formally state our results. Our results apply to input samples generated according to the following model:

Model 106 (Robust Subspace Recovery with Large Outliers)**.** For $0 \leqslant \alpha < 1$ and $r < d$, let $\mu \in \mathbb{R}^d$, $\Sigma_* \in \mathbb{R}^{d\times d}$ be a rank r PSD matrix and let $\mathcal{D}$ be a distribution on $\mathcal{R}^d$ with mean μ_* and covariance Σ_*. Let $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ denote the following probabilistic process to generate n

samples, $x_1, x_2 \dots x_n$ with αn inliers $\mathcal{I}$ and $(1-\alpha)n$ outliers $\mathcal{O}$:

1. Construct $\mathcal{I}$ by choosing αn i.i.d. samples from $\mathcal{D}$.
2. Construct $\mathcal{O}$ by choosing the remaining $(1-\alpha)n$ points arbitrarily and potentially adversarially w.r.t. the inliers.

Remark 107. We will mainly focus on the case when $\mu_* = 0$. The case of non-zero μ_* can be easily reduced to the case of $\mu_* = 0$ by modifying samples by randomly pairing them up and subtracting off samples in each pair (this changes the fraction of inliers from α to α^2).

Remark 108. Our results naturally extend to the harsher strong contamination model (where one first chooses an i.i.d. sample from D and then corrupts an arbitrary $(1-\alpha)$ fraction of them) with no change in the algorithm.

An η-approximate *list-decodable subspace recovery* algorithm takes input a sample $\mathcal{S}$ drawn according to $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ and outputs a list L of *absolute constant* (depending only on α) such that there exists a $\Pi \in L$ satisfying $\|\Pi - \Pi_*\|_F^2 \leqslant \eta$, where Π_* is the projector to the range space of Σ_*.

Before stating our results we observe that since list-decodable subspace recovery strictly generalizes list-decodable regression (by viewing samples as $d+1$ dimensional points with a rank d covariance), we can import the result of Karamalkar, Klivans and Kothari [KKK19] that shows the information-theoretic necessity of anti-concentration of the distribution D.

Fact 5.1.1 (Theorem 6.1, Page 19 in [KKK19])**.** *There exists a distribution $\mathcal{D}$ that $(\alpha+\epsilon)$-anti-concentrated for every $\epsilon > 0$ but there is no algorithm for $\alpha/2$-approximate list-decodable subspace recovery for $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ that outputs a list of size $< d$.*

The distribution $\mathcal{D}$ is simply the uniform distribution on an affine subcube of dimension $n-1$ of $\{0,1\}^n$ (and more generally, q-ary discrete cube).

Our first main result shows that given any arbitrarily small $\eta > 0$, we can recover a polynomial (in the rank r) size list of subspaces that contains a $\hat{\Pi}$ satisfying $\left\|\hat{\Pi} - \Pi_*\right\|_F^2 \leqslant \eta$. The surprising aspect of this result is that we can get an error that can be made arbitrarily small (independent of the rank r or the dimension d) at the cost of increasing the list size from a fixed constant to polynomially large in the rank r of Σ_*. This result crucially relies on our new exponential error reduction method (see Lemma 5.4.3).

Theorem 109 (Large-List Subspace Recovery)**.** *Let $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ be such that Σ_* has rank r and condition number κ, and $\mathcal{D}$ is k-certifiably (c, δ)-anti-concentrated. For any $\eta > 0$ and $t \in \mathbb{N}$, there exists an algorithm that takes input $n \geqslant n_0 = (kd\log(d))^{O(k)}$ samples from $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ and outputs a list of matrices, $\mathcal{L}$, such that $|\mathcal{L}| = O(1/\alpha^t)$ and with probability at least 0.99 over the draw of the sample and the randomness of the algorithm, there is a matrix $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant O\left(r^{1/k}(2r\kappa)^{1/2}(\delta/\alpha)^{t/k}\right)$. The algorithm has time complexity at most $n^{O(t+k)}$.*

Remark 110. We note that our algorithm obtains a trade-off between list size, accuracy and running time as setting $t = 1$ results in a polynomial time algorithm with list size $O(1/\alpha)$ and accuracy $O(r^{1/k}(2r\kappa)^{1/2}(\delta/\alpha))$, which is comparable to the result obtained by Raghavendra-Yau [RY20b].

Remark 111. In general our algorithm for Large-List Subspace Recovery can obtain arbitrarily high accuracy at the expense of running time and list size. Formally, for any $\eta > 0$, we obtain $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant \eta$, with list size $O(1/\alpha^{k\log(r\kappa/\eta)})$ and runnning time $n^{O(k^2\log(r\kappa/\eta)/\log(\delta/\alpha))}$. Further, for $\eta < 0.1$, we can ensure that the resulting list only contains projection matrices.

Remark 112. We note that our large-list rounding algorithm only requires the inliers to be certifiably anti-concentrated. Our subsequent results require subgaussianity as well.

We use a new *pruning* procedure to get the optimal list size of $O(1/\alpha)$ at the cost of increasing the Frobenius error to $\widetilde{O}(\kappa^4\log(r)/\alpha^2)$.

Theorem 113 (List-Decodable Subspace Recovery)**.** *Let $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ be such that Σ_* has rank r and condition number κ, and $\mathcal{D}$ be k-certifiably $(\alpha/2)$-anti concentrated as well as subgaussian with covariance Σ_*. Then, there exists an algorithm that takes as input $n = n_0 \geqslant (d\log(d)/\alpha^2)^{\widetilde{O}(k)}$ samples from $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ and outputs a list $\mathcal{L}$ of $O(1/\alpha)$ projection matrices such that with probability at least 0.99 over the draw of the sample and the randomness of the algorithm, there is a $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant \widetilde{O}(\kappa^4\log(r)/\alpha^2)$. The algorithm has time complexity at most $n^{\widetilde{O}(\log(r\kappa)k^2)}$.*

Remark 114. Our algorithm can also achieve a smooth trade-off between list-size and accuracy. For any $\gamma \geqslant 1$, we can output a list of size $O(1/\alpha^\gamma)$ such that it contains a projection matrix satisfying $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant \widetilde{O}(\kappa^4\log(r)/\alpha^2\gamma)$ by obtaining $1/\alpha^\gamma$ fresh samples from $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$.

Remark 115. For Gaussian distributions with mean 0 and covariance Σ_*, it suffices to set $k = \widetilde{O}(1/\alpha^2)$ (see Theorem 121 for details).

As discussed above, our results immediately extends by means of a simple reduction to the case when μ_* is non-zero.

Corollary 5.1.2 (Large-List Affine Recovery)**.** *Let $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ be such that Σ_* has rank r and condition number κ, and $\mathcal{D}$ is k-certifiably $(\alpha/2)$-anti concentrated as well as subgaussian with covariance Σ_*. Then, there exists an algorithm that takes as input $n = n_0 \geqslant (d\log(d)/\alpha^4)^{\widetilde{O}(k)}$ samples from $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ and outputs a list $\mathcal{L}$ of $O(1/\alpha^2)$ projection matrices such that with probability at least 0.99 over the draw of the sample and the randomness of the algorithm, there is a $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant \widetilde{O}(\kappa^4 \log(r)/\alpha^4)$. The algorithm has time complexity at most $n^{\widetilde{O}(\log(r\kappa)k^2)}$.*

Remark 116. We note that our algorithm and subsequent analysis can be carried out using the projection onto the orthogonal complement of the subspace spanned by Σ_* as variables in the constraint system and therefore, our running time depends on $\min(r, d-r)$. Recall, in the List-Decodable Linear Regression problem, α-fraction of the input spans the $(d-1)$ dimension subspace represented by $\{\langle x_i, \ell_* \rangle = y_i\}$, where ℓ_* is the regressor we would like to recover. Combining these two observations, we obtain a faster algorithm for List-Decodable Linear Regression.

Corollary 5.1.3 (List-Decodable Regression)**.** *Let $\mathsf{Lin}_{\mathcal{D}}(\alpha, \ell_*)$ be such that an α-fraction of the input satisfies $\langle x_i, \ell_* \rangle = y_i$, where the x_i are drawn from $\mathcal{D}$, and the remaining fraction is chosen arbitrarily, potentially adversarially w.r.t the inliers. Let Σ_* be a rank-$(d-1)$ matrix with condition number κ. Let $\mathcal{D}$ be k-certifiably $(\alpha/2)$-anti concentrated as well as subgaussian with covariance Σ_*. Then, there exists an algorithm that takes as input $n = n_0 \geqslant (d\log(d)/\alpha^2)^{\widetilde{O}(k)}$ samples from $\mathsf{Lin}_{\mathcal{D}}(\alpha, \ell_*)$ and outputs a list $\mathcal{L}$ of $O(1/\alpha)$ projection matrices such that with probability at least 0.99 over the draw of the sample and the randomness of the algorithm, there is a $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant \widetilde{O}(\kappa^4/\alpha^2)$. The algorithm has time complexity at most $n^{\widetilde{O}(\log(\kappa)k^2)}$.*

5.1.2 Related Work

Subspace Clustering. Prior work on subspace recovery focused on the closely related problem of subspace clustering in high dimension, where to goal is to partition a set of points into k-clusters according to their underlying subspaces. Subspace clustering methods have found numerous applications computer vision tasks such as image compression [HWHM06], motion segmentation [CK98], data mining [PHL04], disease classfication [MM14], recommendaation

systems [ZFIM12] etc. Algorithms for subspace clustering include iterative methods, algebraic and statistical methods and spectral techniques. We refer the readers to the following surveys for a comprehensive overview [EV13, PHL04]. Elhamifar and Vidal [EV13] also introduced *sparse subspace clustering*, building on the compressed sensing and matrix completion literature. Soltanolkotabi et. al. [SEC14] extend *sparse subspace clustering* to work in the presence of noise and provide rigorous algorithmic guarantees. They assume the outliers contribute a small fraction of the input and are distributed uniformly distributed of the unit sphere.

Robust Subspace Recovery. A recent line of work on robust subspace recovery has focused on projection pursuit techniques, ℓ_1-PCA (robust PCA), exhaustive subspace search and robust covariance estimation. Here, the goal is to recover a set of inliers that span a single low-dimensional space. Projection pursuit algorithms iteratively find directions that maximize a scale function. The scale function often accounts on outliers and thus may be non-convex. McCoy and Tropp [MT$^+$11] consider one such function and develop a rounding which approximates the global optimizer. The ℓ_1 or Robust PCA objective replaces the Frobenius norm objective with a sum of absolute values objective, since it is less sensitive to outliers. While this formulation is non-convex and NP-hard in general, many special cases are tractable, as discussed here [VN18]. Hardt and Moitra [HM13] provide a worst-case exhaustive search algorithm, where both the inliers and outliers are required to be in general position and the inliers are generated deterministically. For a more comprehensive treatment of robust subspace recovery we refer the reader to [LM18a].

In a concurrent and independent work, Raghavendra and Yau proved related results for list-decodable subspace recovery [RY20a].

5.2 Technical Overview

In this section, we give a high level overview of our algorithm and the new ideas that go into making it work. At a high level, our algorithm generalizes the framework for list-decodable estimation recently used to obtain an efficient algorithm for list-decodable regression in the recent work of [KKK19].

In the list-decodable subspace recovery problem, our input is a collection of samples $x_1, x_2, \ldots, x_n \in \mathcal{R}^d$, an αn of which are drawn i.i.d. from distribution $\mathcal{D}$ with mean 0 and unknown covariance Σ_* of rank r. For the purpose of this overview, we will think of Σ_* itself being a projection matrix Π_*. Our algorithm starts from a polynomial feasibility program that simply tries to find a subset of sample that contains at least an αn points such that all of these points

lie in a subspace of dimension $r \leqslant d$. We can encode these two requirements as the following system $\mathcal{A}_{w,\Pi}$ of polynomial constraints as follows:

$$\mathcal{A}_{w,\Pi}\colon \left\{\begin{array}{rr} & \sum_{i\in[n]} w_i = \alpha n \\ \forall i \in [n]. & w_i(\mathbf{I}-\Pi)x_i = 0 \\ \forall i \in [n]. & w_i^2 = w_i \\ & \Pi^2 = \Pi \\ & \mathrm{Tr}(\Pi) = r \end{array}\right\} \tag{5.1}$$

In this system of constraints, $w_1, w_2, \ldots, w_n$ are indicators (due to the constraint $w_i^2 = w_i$) of the subset of sample we pick. Since $\sum_{i=1}^n w_i = \alpha n$, the constraints force w to indicate a subset of the sample of size αn. To force that all the points indicated by w lie in a subspace of dimension r, we define variable Π intended to be the projector to this unknown subspace. The constraint $\Pi^2 = \Pi$ forces Π to be a projection matrix and $\mathrm{tr}(\Pi) = r$ forces its rank to be r. Given these constraints, it's easy to verify the constraint $w_i(\mathbf{I} - \Pi)x_i = 0$ forces x_i to be in the subspace projected to by Π whenever $w_i = 1$.

5.2.1 Designing an Inefficient Algorithm

A feasible solution (w, Π) to the aforementioned constraint system (ignoring for now, the issue of efficiency), results in a subset of αn samples that span a subspace of dimension r. However, there can be multiple r dimensional subspaces that satisfy this requirement for various αn subsets chosen entirely out of the *outliers*[2]. Thus, even if we were to find a solution to this program, it's not immidiately clear how to recover a subspace close to the one spanned by the *inliers*.

High-Entropy Distributions. In order to force our solution to (5.1) to give us information about the true inliers, it seems beneficial to try to find not one but *multiple* solution pairs $\{(w^i, \Pi^i)\}$ such that at least one of the w^i indicates a subset that has a substantial intersection with the true inliers. An important conceptual insight in (see Overview section in [KKK19] for a longer discussion) is to thus ask for a probability distribution (which, at this point can be thought of as a method to ask for multiple solutions) μ over solutions (w, Π) satisfying (5.1). It turns out that we can ensure that there are solutions (w, Π) in the support of μ where w indicates a subset with a non-trivial intersection with the inliers by finding a distribution μ so that $\|\sum_{i=1}^n \mathbf{E} w_i\|_2^2$ is

[2]See Section 3 in [KKK19] for examples showing how outliers can generate $\exp(\Omega(d))$ many possible subspaces that can all be far from the ground truth subspace.

minimized. This constraint serves as a proxy for *high entropy distributions*. Formally, we can conclude the following useful result that shows that the expected (over μ) intersection of a subset indicated by w and the inliers is at least α fraction of the inliers.

Proposition 5.2.1. *Let μ be a distribution on (w, Π) satisfying $\mathcal{A}_{w,\Pi}$. Then, $\mathbf{E}_{\mu}\left[\sum_{i\in\mathcal{I}} w_i\right] \geqslant \alpha|\mathcal{I}|$.*

This result follows by a simple "weight-shifting" argument (if the distribution is over w that do not intersect enough with the inliers, we can shift probability mass on the inliers and decrease $\left\|\sum_{i=1}^{n} \tilde{\mathbb{E}}\, w_i\right\|_2^2$)).

Anti-Concentration. Our distribution over μ is guaranteed to contain w with at least α fraction of the points of $\mathcal{I}$ in the intersection. Our hopes of finding information about the true subspace are pinned on such "good" (w, Π) at this point. We would like that for such w, the corresponding projector Π matches the ground truth subspace corresponding to the projector Π_*. Let S be the "intersection indices", i.e., the set of indices of samples in $\mathcal{I}$ for which $w_i = 1$. Why should this be true? Since we have no control over S, it could, a priori, consist of the points in $\mathcal{I}$ that span only a proper subspace, say V of the ground truth subspace. In this case, Π may not equal Π_*.

The key observation is that in this "bad" case, there is a vector v that is in the orthogonal complement of Π_V inside the subspace spanned by Π_* such that $\langle x_i, v\rangle = 0$ for every $i \in S$. That is, there's a direction that inliers have a zero projection in α fraction of the times. Such an eventuality is ruled out if we force D, the distribution of the inliers to be *anti-concentrated*.

Definition 5.2.2 (Anti-Concentration)**.** *A $\mathcal{R}^d$-valued random variable Y with mean 0 and covariance $\mathbf{\Sigma}$ is δ-anti-concentrated if for all v satisfying $v^\top \mathbf{\Sigma} v > 0$, $\mathbf{Pr}[\langle Y, v\rangle = 0] < \delta$. A set $T \subseteq \mathcal{R}^d$ is δ-anti-concentrated if the uniform distribution on T is δ-anti-concentrated.*

The following proposition is now a simple corollary:

Proposition 5.2.3 (High Intersection Implies Same Subspace (TV Distance to Parameter Distance))**.** *Let $\mathcal{S}$ be a sample of size n from $Sub_{\mathcal{D}}(\alpha, \mathbf{\Sigma}^*, r)$ for a projection matrix $\Sigma_* = \Pi^*$ of rank r such that the inliers $\mathcal{I}$ are α-anti-concentrated. Let $T \subseteq \mathcal{S}$ be a subset of size αn such that $\Pi x = x$ for every $x \in T$ for some projection matrix Π of rank r. Suppose $|T \subseteq \mathcal{I}| \geqslant \alpha|\mathcal{I}|$. Then, $\Pi = \Pi^*$.*

Proof. Let $I - \Pi = \sum_{i=1}^{d-r} v_i v_i^\top$ for an orthonormal set of vectors v_is. Since $\Pi x = x$ for every $x \in T$, $\langle x, v_i\rangle = 0$ for every $x \in T$. Thus, $\mathbf{Pr}_{x\sim\mathcal{I}}[\langle x, v_i\rangle = 0] \geqslant |T \cap \mathcal{I}|/|\mathcal{I}| \geqslant \alpha$. Since $\mathcal{I}$ is α-anti-concentrated, this must mean that $v_i^\top \Pi^* v_i = 0$.

Thus, $\sum_i v_i^\top \Pi^* v_i = \mathrm{tr}(\Pi^* \sum_{i=1}^{d-r} v_i v_i^\top) = \mathrm{tr}(\Pi^*(I - \Pi)) = 0$. Or $\mathrm{tr}(\Pi^*) = \mathrm{tr}(\Pi \cdot \Pi^*)$. On the other hand, by Cauchy-Schwarz inequality, $\mathrm{tr}(\Pi \cdot \Pi^*) \leqslant \sqrt{\mathrm{tr}(\Pi^2)\,\mathrm{tr}((\Pi^*)^2)} = \mathrm{tr}(\Pi)$ with equality iff $\Pi = \Pi^*$. Here, we used the facts that $\Pi = \Pi^2, (\Pi^*)^2 = \Pi^*$ and that $\mathrm{tr}(\Pi) = \mathrm{tr}(\Pi^*) = r$. Thus, $\Pi = \Pi^*$. □

Inefficient Algorithm for Anti-Concentrated Distributions. We can use the lemma above to give an *inefficient* algorithm for list-decodable subspace recovery.

Lemma 5.2.4 (Identifiability for Anti-Concentrated inliers)**.** *Let $\mathcal{S}$ be a sample drawn according to $\mathrm{Sub}_{\mathcal{D}}(\alpha, \mathbf{\Sigma}^*, r)$ such that the inliers $\mathcal{I}$ are δ-anti-concentrated for $\delta < \alpha$. Then, there is an (inefficient) randomized algorithm that finds a list L of projectors of rank r of size $20/(\alpha - \delta)$ such that $\Pi^* \in L$ with probability at least 0.99.*

Proof. Let μ be any maximally uniform distribution over soluble subset-projection pairs (w, Π) where w indicates a set S of size at least αn. For $k = 20/(\alpha - \delta)$, let $(S_1, \Pi_1), (S_2, \Pi_2), \ldots, (S_k, \Pi_k)$ be i.i.d. samples from μ. Output $\{\Pi_1, \Pi_2, \ldots, \Pi_k\}$. To finish the proof, we will show that there is an i such that $|S_i \cap \mathcal{I}| \geqslant \frac{\alpha+\delta}{2}|\mathcal{I}| > \delta|\mathcal{I}|$. Then, we can then apply Proposition 5.2.3 to conclude that $\Pi_i = \mathbf{\Sigma}$.

By Proposition 5.2.1, $\mathbf{E}_{S\sim\mu}|S \cap \mathcal{I}| \geqslant \alpha|\mathcal{I}|$. Thus, by averaging, $\mathbf{Pr}_{S\sim\mu}[|S \cap \mathcal{I}| \geqslant \frac{\alpha+\delta}{2}|\mathcal{I}|] \geqslant \frac{\alpha-\delta}{2}|\mathcal{I}|$. Thus, the probability that at least one of $S_1, S_2, \ldots S_k$ satisfy $|S_i \cap \mathcal{I}| \geqslant \frac{\alpha+\delta}{2}|\mathcal{I}|$ is at least $1 - (1 - \frac{\alpha-\delta}{2})^k \geqslant 0.99$. □

5.2.2 Efficient Algorithm

Our key technical contributions are in making the above inefficient algorithm into an efficient algorithm via the sum-of-squares method. At a high level, we consider a low-degree sum-of-squares relaxation of the constraint system and design an efficient rounding algorithm. As in prior works, it is natural at this point to consider the algorithm that finds a *pseudo-distribution* minimizing $\|\sum_{i\leqslant n} w_i\|_2^2$ and satisfying $\mathcal{A}_{w,\Pi}$.

A precise discussion of pseudo-distributions and sum-of-squares proofs appears in Section 5.3 - at this point, it suffices to think of pseudo-distributions as objects similar to the distribution μ that appeared above for all "properties" that have a low-degree sum-of-squares proofs. Sum-of-squares proofs are a system of reasoning about polynomial inequalities under polynomial inequality constraints. It turns out that the analog of Proposition 5.2.1 holds even for pseudo-distributions. Our central goal is then to find a sum-of-squares proof of the "high-intersection

implies same subspace" property and use such a statement algorithmically to obtain a small list of projectors. To this end, we describe three novel technical contributions that go into achieving this goal.

Anti-Concentration as a Polynomial Identity. As we recall from our discussion above, such an argument relies on the distribution $\mathcal{D}$ being anti-concentrated. While as stated, anti-concentration does not have a natural formalization as a low-degree polynomial identity, recent works [KKK19, RY20a] made progress towards formalizing it within the SoS system in slightly different ways.

Our proofs are more attuned to the formalization in [KKK19]. But for technical reasons the precise formulation proposed in [KKK19] is not directly useful for us. Briefly and somewhat imprecisely put, anti-concentration formalizations posit that there be a low-degree SoS proof (in the variable v) for polynomial inequalities of the form $\mathbf{E}_D p^2(\langle x, v\rangle) \leqslant \delta$ for a univariate polynomial p that approximates a Dirac Delta function at 0. In the prior works, this requirement was formulated in a *constrained* manner ("$\|v\|_2^2 \leqslant 1$ implies $\mathbf{E}_{\mathcal{D}} p^2(\langle x, v\rangle) \leqslant \delta$"). For the application to subspace recovery, natural arguments require *unconstrained* versions of the above inequality (i.e. that hold without the norm bound constraint on v). Definition 2.2.12 formulates this condition precisely. We then show that our formalization of anti-concentration holds for natural distribution families such as Gaussians (see Section 5.5 for details).

Spectral Bound on Subsamples. Given our modified formalization of anti-concentration, we give a sum-of-squares proof of the analog of Proposition 5.2.3. In particular, we prove a polynomial identity that states if the samples indicated by w non-trivially intersect the true inliers, then the projector Π is close to Π_* in Frobenius norm. Further, we are able to achieve a trade-off between the degree of the polynomial identity and the closeness in Frobenius norm. This statement as well as the trade-off between the degree and error (see Lemma 5.4.1) is a key technical contribution of our work and we expect will find applications in future works.

Alternatively, we can view this statement as an SoS version of results relating total variation distance between anti-concentrated distributions to the Frobenius norm difference between their covariance. Here, the analog of closeness in total variation distance is the size of the intersection between the samples indicated by w and the true inliers and the closeness is between the corrresponding projectors.

An important technical component in our proof is to show that given a set of points $\mathcal{S}$ sampled from an anti-concentrated distribution, we can lower bound the eigenvalues of the empirical covariance of a significantly large subset of $\mathcal{S}$ (see Lemma 5.4.5 for a precise statement). For subspace recovery, this implies that non-zero directions in the empirical covariance for $\mathcal{S}$ remain

non-zero for a subset of $\mathcal{S}$. Intuitively, such a statement implies that we preserve subspace even when the samples indicated by w intersect a small fraction of the inliers.

Exponential Error Reduction and Large List Rounding. Recall, a high-entropy pseudo-distribution satisfying $\mathcal{A}_{w,\Pi}$ can be interpreted as a "distribution" over tuples (w, Π) satisfying the constraint system such that the samples indicated by w non-trivially intersect the inliers in expectation. Next, we design a rounding algorithm that takes the pseudo-distribution as input and outputs a list of projectors such that one is close to Π_*. Note, for the list-decodable regression problem, simply applying the rounding "by votes" strategy from sufficed to get a polynomial time algorithm for any fixed constant error [KKK19]. However, for subspace recovery, the same rounding strategy gives an error bound that depends polynomially on the rank (or co-rank) of the unknown subspace and the fraction of inliers. When the rank of the subspace is high (say $d/2$, where d is the dimension), such a bound may not even be meaningful. To reduce error down to something that is dimension independent ends up needing running time that is (super)-exponential in the dimension.

We extend the voting based *rounding* algorithm such that it allows for a trade-off between the list size and the closeness of Π to Π_* and our exponential error reduction mechanism allows us to obtain a dimension-independent error bound in quasi-polynomial time. We show that picking a sufficiently large subset of the points indicated by w proportional to the high-entropy pseudo-distribution results in an projector that is η close to Π_* in Frobenius norm, with probability $1/\mathrm{poly}(d)$. Further, the running time of our algorithm scales proportional to $n^{\log(1/\eta)}$ and the list size blows up by $1/\alpha^{\log(1/\eta)}$ [3]. Our powering and error reduction technique is quite general and will likely find new uses in list-decodable estimation.

Pruning Lists. In order to get optimal list size bounds, the last step in our algorithm is to introduce a "pruning method" that decreases the size of the large list obtained by rounding pseudo-distributions. Here, we obtain $O(1/\alpha)$ fresh samples from $\mathcal{D}$ and for each fresh sample x compute the projection on to the orthogonal complement for each projector in our large list. We then pick an arbitrary projector Π such that $\|(\mathbf{I} - \Pi)x\|_2$ is a small fraction of $\|x\|_2$. Our resulting list thus has at most $O(1/\alpha)$ projectors. Further, when x is drawn from the inliers, we show that we add a projector close to Π_* to our list using our aforementioned test.

[3] Here, we ignore the dependence on the remaining parameters.

5.3 Preliminaries

Throughout this paper, for a vector v, we use $\|v\|_2$ to denote the Euclidean norm of v. For a $n \times m$ matrix M, we use $\|M\|_2 = \max_{\|x\|_2=1} \|Mx\|_2$ to denote the spectral norm of M and $\|M\|_F = \sqrt{\sum_{i,j} M_{i,j}^2}$ to denote the Frobenius norm of M. For symmetric matrices we use $\succeq$ to denote the PSD/Loewner ordering over eigenvalues of M. For a $n \times n$, rank-r symmetric matrix M, we use $U\Lambda U^\top$ to denote the Eigenvalue Decomposition, where U is a $n \times r$ matrix with orthonormal columns and Λ is a $r \times r$ diagonal matrix denoting the eigenvalues. We use $M^\dagger = U\Lambda^\dagger U^\top$ to denote the Moore-Penrose Pseudoinverse, where $\Lambda^\dagger$ inverts the non-zero eigenvalues of M. If $M \succeq 0$, we use $M^{\dagger/2} = U\Lambda^{\dagger/2}U^\top$ to denote taking the square-root of the non-zero eigenvalues. We use $\Pi = UU^\top$ to denote the Projection matrix corresponding to the column/row span of M. Since $\Pi = \Pi^2$, the pseudo-inverse of Π is itself, i.e. $\Pi^\dagger = \Pi$.

Reweightings Pseudo-distributions. The following fact is easy to verify and has been used in several works (see [BKS17] for example).

Fact 5.3.1 (Reweighting)**.** *Let $\tilde{\mu}$ be a pseudo-distribution of degree k satisfying a set of polynomial constraints $\mathcal{A}$ in variable x. Let p be a sum-of-squares polynomial of degree t such that $\tilde{\mathbb{E}}[p(x)] \neq 0$. Let $\tilde{\mu}'$ be the pseudo-distribution defined so that for any polynomial f, $\tilde{\mathbb{E}}_{\tilde{\mu}'}[f(x)] = \tilde{\mathbb{E}}_{\tilde{\mu}}[f(x)p(x)] / \tilde{\mathbb{E}}_{\tilde{\mu}}[p(x)]$. Then, $\tilde{\mu}'$ is a pseudo-distribution of degree $k - t$ satisfying $\mathcal{A}$.*

5.4 Algorithm

In this section, we describe an efficient algorithm for list-decodable subspace recovery. Let $\mathcal{A}_{w,\Pi}$ be the following system of polynomial inequality constraints in indeterminates w, Π.

$$\mathcal{A}_{w,\Pi}\colon \left\{ \begin{aligned} \textstyle\sum_{i\in[n]} w_i &= \alpha n \\ \forall i \in [n]. \quad w_i(\mathbf{I} - \Pi)x_i &= 0 \\ \forall i \in [n]. \quad w_i^2 &= w_i \\ \Pi^2 &= \Pi \\ \mathrm{Tr}(\Pi) &= r \end{aligned} \right\} \tag{5.2}$$

Our algorithm finds a pseudo-distribution consistent with $\mathcal{A}_{w,\Pi}$. It then uses the large-list

rounding algorithm as a first step to get a polynomial (in d) size list that contains a subspace that is η-close in Frobenius norm to the range space of Σ_*. Finally, we apply a pruning procedure to obtain a $O(1/\alpha)$ size from the large list procedure.

Algorithm 117. *List-Decodable Subspace Recovery*

Given: *Sample $\mathcal{S} = \{x_1, x_2, \dots x_n\} = \mathcal{I} \cup \mathcal{O}$ of size n drawn according to Sub$_D(\alpha, \Sigma_*)$ such that the $\mathcal{D}$ is k-certifiably (c, δ)-anti-concentrated, has mean 0 and the condition number of Σ_* is κ.*

Operation:

1. *Let $t = \Delta \cdot \left(\frac{\log^5(1/\alpha)\log(r\kappa)}{\alpha^2}\right)$ for a large enough constant $\Delta > 0$.*
2. *Compute a $(t+2k)$-degree pseudo-distribution $\tilde{\mu}$ satisfying $\mathcal{A}_{w,\Pi}$ that minimizes $\left\|\sum_{i=1}^n \tilde{\mathbb{E}}[w_i]\right\|_2^2$.*
3. *Run Large-List Rounding with $\eta = 0.1$ (Algorithm 118) to output a $O(1/\alpha^t)$ sized list $\mathcal{L}'$.*
4. *Run pruning (Algorithm 119) on $\mathcal{L}'$ with $\gamma = 1$ and output the resulting list $\mathcal{L}$.*

Output: *A list $\mathcal{L}$ of $O(1/\alpha)$ projection matrices containing a $\tilde{\Pi} \in \mathcal{L}$ satisfying $\|\tilde{\Pi} - \Pi_*\|_F^2 \leqslant \widetilde{O}(\kappa^4 \log(r)/\alpha^2)$.*

Algorithm 118. *Large List Rounding*

Given: *A pseudo-distribution $\tilde{\mu}$ of degree $t+2k$ satisfying $\mathcal{A}_{w,\Pi}$ and minimizing $\left\|\sum_{i\leqslant n} \tilde{\mathbb{E}}\, w_i\right\|_2^2$ such that $t = \Delta \cdot \left(\frac{\log^5(1/\alpha)\log(r\kappa/\eta)}{\alpha^2}\right)$, for a large constant Δ, accuracy parameter $\eta > 0$.*

Operation: *Repeat $\ell = O(1/\alpha^t)$ times:*

1. *Let $S \subset [n]$ such that $|S| = \alpha n$. Draw S with probability proportional to $\binom{n}{S} \tilde{\mathbb{E}}_{\tilde{\mu}}[w_S]$.*
2. *Let $\tilde{\Pi} = \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_S \Pi]}{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_S]}$ be the corresponding matrix. Compute the Eigenvalue Decomposition of $\tilde{\Pi} = \tilde{U}\tilde{\Lambda}\tilde{U}^\top$ and let $\hat{\Pi} = \tilde{U}_r\tilde{U}_r^\top$, where $\tilde{U}_r$ are the eigenvectors corresponding to the top-r eigenvalues of $\tilde{\Pi}$.*
3. *Add $\hat{\Pi}$ to the list $\mathcal{L}'$.*

Output: *A list $\mathcal{L}' \subseteq \mathcal{R}^d$ of size $O(1/\alpha^t)$ containing a Projection matrix $\hat{\Pi} \in \mathcal{L}'$ satisfying $\|\hat{\Pi} - \Pi_*\|_F^2 < \eta$.*

Algorithm 119. *Pruning Lists*

Given: *A list $\mathcal{L}'$ of $d \times d$ projection matrices such that there exists $\hat{\Pi} \in \mathcal{L}'$ satisfying $\|\Pi - \Pi_*\|_F^2 \leqslant 0.1$, $O(1/\alpha^\gamma)$ fresh samples $\mathcal{S}$, drawn according to $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$, for some $\gamma \geqslant 1$, a threshold $\tau = \widetilde{O}(\kappa^4 \log(r)/\alpha^2\gamma)$.*

Operation:

For $i = 1, 2, \ldots, |\mathcal{S}|$:

1. *Let $\mathcal{L}'_i = \mathcal{L}'$. For $j = i, \ldots, i + \gamma - 1$*
   (a) *For each $\bar{\Pi} \in \mathcal{L}'_i$, compute $\|(\mathbf{I} - \bar{\Pi})x_j\|_2^2$.*
   (b) *If $\|(\mathbf{I} - \bar{\Pi})x_j\|_2^2 > \tau$, discard $\bar{\Pi}$ from $\mathcal{L}'_i$.*
2. *If $\mathcal{L}'_i$ is non-empty, pick an arbitrary matrix $\hat{\Pi}$ from this set and add it to $\mathcal{L}$.*

Output: *A $\mathcal{L} \subseteq \mathcal{R}^d$ of size $O(1/\alpha)$ such that there exists a Projection matrix $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant \tau$.*

5.4.1 Analysis of Algorithm 117.

The following theorem captures the guarantees we prove on Algorithm 117.

Theorem 120 (List-Decodable Subspace Recovery, restated)**.** *Let $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ be such that Σ_* has rank r and condition number κ, and $\mathcal{D}$ is k-certifiably $(c, \alpha/2)$-anti-concentrated and sub-gaussian with covariance Σ_*. Then, Algorithm 117 takes as input $n = n_0 \geqslant (d \log(d)/\alpha^2)^{\widetilde{O}(k)}$ samples from $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$ and outputs a list $\mathcal{L}$ of $O(1/\alpha)$ projection matrices such that with probability at least 0.9 over the draw of the sample and the randomness of the algorithm, there is a $\hat{\Pi} \in \mathcal{L}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant \widetilde{O}(\kappa^4 \log(r)/\alpha^2)$. Further, Algorithm 117 has time complexity at most $n^{\widetilde{O}(\log(r\kappa)k^2)}$.*

Our proof of Theorem 120 is based on the following four pieces. The key technical piece is the following consequence of the constraint system $\mathcal{A}_{w,\Pi}$ in the low-degree SoS proof system.

Lemma 5.4.1. *Given $\delta > 0$ and any $t \in \mathbb{N}$, and an instance of $\mathrm{Sub}_{\mathcal{D}}(\alpha, \Sigma_*)$, such that the inlier*

distribution $\mathcal{D}$ has mean 0 and is k-certifiably (C,δ)-anti-concentrated,

$$\mathcal{A}_{w,\Pi} \vdash^{\Pi,w}_{2k+t} \left\{ \left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i\right)^t \|\Pi-\Pi_*\|_F^k = \left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i\right)^t 2^{k/2}\operatorname{tr}(M\Pi_* M)^{k/2} \leqslant (2r\kappa)^{k/2}\delta^t \right\}.$$

where κ is the condition number of Σ_* and Π_* is the corresponding rank-r Projection matrix.

Next, we show that "high-entropy" pseudo-distributions must place a large enough weight on the inliers. This is similar to the usage of high-entropy pseudo-distributions in [KKK19].

Lemma 5.4.2 (Large weight on inliers from high-etropy constraints)**.** *Let $\tilde{\mu}$ pseudo-distribution of degree $\geqslant t$ that satisfies $\mathcal{A}_{w,\Pi}$ and minimizes $\left\|\tilde{\mathbb{E}}_{\tilde{\mu}'}\sum_{i\in[n]} w_i\right\|_2$. Then, $\frac{1}{|\mathcal{I}|^t}\tilde{\mathbb{E}}\left[\left(\sum_{i\in\mathcal{I}} w_i\right)^t\right] \geqslant \alpha^t$.*

The above two lemmas allow us to argue that our large-list rounding algorithm (Algorithm 118) succeeds.

Lemma 5.4.3 (Large-List Subspace Recovery, Theorem 109 restated)**.** *Let $\mathrm{Sub}_{\mathcal{D}}(\alpha,\Sigma_*)$ be such that Σ_* has rank r and condition number κ, and $\mathcal{D}$ is k-certifiably $(c,\alpha/2)$-anti-concentrated. For any $\eta>0$, there exists an algorithm that takes input $n\geqslant n_0=(kd\log(d))^{O(k)}$ samples from $\mathrm{Sub}_{\mathcal{D}}(\alpha,\Sigma_*)$ and outputs a list $\mathcal{L}$ of size $O(1/\alpha^{k\log(r\kappa/\eta)})$ of projection matrices such that with probability at least 0.99 over the draw of the sample and the randomness of the algorithm, there is a $\hat{\Pi}\in\mathcal{L}$ satisfying $\|\hat{\Pi}-\Pi_*\|_F^2\leqslant\eta$. The algorithm has time complexity at most $n^{O(k^2\log(r\kappa/\eta))}$.*

Finally, we show that we can prune the list output by Algorithm 118 to a list of size $O(1/\alpha)$ such that it still contains a Projection matrix close to Π_*. Here, we require that $\mathcal{D}$ is subgaussian. Formally,

Lemma 5.4.4 (Pruning Algorithm)**.** *Let $\gamma\geqslant 1$ and $\mathcal{L}'$ be the list output by Algorithm 118. Given $O(1/\alpha^\gamma)$ fresh samples from $\mathrm{Sub}_{\mathcal{D}}(\alpha,\Sigma_*)$, Algorithm 119 outputs a list $\mathcal{L}$ of size $O(1/\alpha^\gamma)$ such that with probability at least $99/100$, there exists a projection matrix $\hat{\Pi}\in\mathcal{L}$ satisfying $\|\hat{\Pi}-\Pi_*\|_F^2\leqslant\widetilde{O}\left(\frac{\kappa^4\log(r)}{\alpha^2\gamma}\right)$.*

Theorem 120 follows easily by combining the above claims :

Proof of Theorem 120. Recall, $\mathcal{D}$ is k-certifiably $(c,\alpha/2)$-anti-concentrated, and thus it follows from Lemma 5.5.6 that the uniform distribution on $\mathcal{I}$ is also $O(k)$-certifiably (c,α)-anti-concentrated if the number of samples are at least $n_0=(d\log(d)/\alpha^2)^{\widetilde{O}(k)}$.

We begin by observing that the system of constraints $\mathcal{A}_{w,\Pi}$ is feasible when we set w_i to

indicate the inliers, and $\Pi = \Pi_*$. Next, the hypothesis of Lemma 5.4.3 is now satisfies for $\eta = 0.1$, Algorithm 118 runs in time $n^{\widetilde{O}(\log(r\kappa)k)}$ and outputs a list $\mathcal{L}'$ of size $(1/\alpha)^{\widetilde{O}(\log(r\kappa)k)}$ such that with probability at least $99/100$, it contains a projector $\tilde{\Pi}$ satisfying $\|\tilde{\Pi} - \Pi_*\|_F^2 \leqslant 0.1$. Recall, Π_* is the projector corresponding to Σ_* and let ζ_1 be the event that $\mathcal{L}'$ contains Π_*.

Conditioning on ζ_1, we now have a list satifying the hypothesis for Lemma 5.4.4 and access to $O(1/\alpha)$ fresh samples ($\gamma = 1$) we can conlcude that with probability $99/100$ Algorithm 119 outputs a list of size $O(1/\alpha)$ which containts a projection matrix $\hat{\Pi}$ satisfying $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant \widetilde{O}(\kappa^4 \log(r)/\alpha^2)$, as desired. Let ζ_2 be the event that Algorithm 119 succeeds. Therefore, union bounding over ζ_1 and ζ_2 implies Algorithm 117 succeeds with probability at least $9/10$. The overall running time is dominated by Algorithm 118, which completes the proof.

□

5.4.2 Analyzing $\mathcal{A}_{w,\Pi}$: Proof of Lemma 5.4.1

We first show that covariance of all large enough subsamples of certifiably anti-concentrated samples have lower-bounded eigenvalues. Recall, for a PSD matrix Σ_*, $U\Lambda U^\top$ denotes the Eigenvalue Decomposition and $\Pi_* = UU^\top$ denotes the corresponding rank-r Projection matrix.

Lemma 5.4.5 (Covariance of Subsets of Certifiably Anti-Concentrated Distributions)**.** *Let $\mathcal{S} = \{x_1, x_2, \dots x_n\} \subseteq \mathcal{R}^d$ be k-certifiably (C, δ)-anti-concentrated with $\frac{1}{n}\sum_{x\in\mathcal{S}} xx^\top = \Sigma$. Then,*

$$\left\{w_i^2 = w_i \mid \forall i\right\} \left|\frac{w,v}{2k}\right. \left\{\frac{1}{n}\sum_{i=1}^n \|v\|_2^{k-2} w_i \left\langle \Sigma^{\dagger/2} x_i, v\right\rangle^2 \geqslant \delta^2 \left(\frac{1}{n}\sum_{i=1}^n w_i - C\delta\right)\|v\|_2^k\right\}, \tag{5.3}$$

Proof. Let p be the degree k polynomial provided by Definition 2.2.12 applied to $\mathcal{S}$. Thus, for each $1 \leqslant i \leqslant n$, we must have:

$$\left|\frac{v}{2k}\right. \left\{\|v\|_2^{k-2}\left\langle \Sigma^{\dagger/2}x_i, v\right\rangle^2 + \delta^2 p^2\left(\left\langle \Sigma^{\dagger/2}x_i, v\right\rangle\right) \geqslant \delta^2\|v\|_2^k\right\}.$$

Observe that

$$\left\{w_i^2 = w_i\right\} \left|\frac{w_i}{2}\right. \{w_i \geqslant 0\}.$$

Using the above along with (3.5) for manipulating SoS proofs, we must have:

$$\left\{w_i^2 = w_i \mid \forall i\right\} \left|\frac{w,v}{2k}\right. \left\{\frac{1}{n}\sum_{i=1}^n \|v\|_2^{k-2} w_i \left\langle \Sigma^{\dagger/2} x_i, v\right\rangle^2 + \delta^2 \frac{1}{n}\sum_{i=1}^n w_i p^2\left(\left\langle \Sigma^{\dagger/2}x_i, v\right\rangle\right) \geqslant \delta^2 \frac{1}{n}\sum_{i=1}^n w_i\|v\|_2^k\right\}.$$

Rearranging yields:

$$\left\{w_i^2 = w_i \mid \forall i\right\} \Big|\frac{w,v}{2k} \left\{\frac{1}{n}\sum_{i=1}^{n}\|v\|_2^{k-2} w_i \left\langle \Sigma^{\dagger/2}x_i, v\right\rangle^2 \geqslant \delta^2 \frac{1}{n}\sum_{i=1}^{n} w_i \|v\|_2^k - \delta^2 \frac{1}{n}\sum_{i=1}^{n} w_i p^2\left(\left\langle \Sigma^{\dagger/2}x_i, v\right\rangle\right)\right\}. \tag{5.4}$$

Next, observe that $\{w_i^2 = w_i\} \Big|\frac{w_i}{2} \{(1-w_i) = (1-w_i)^2 \geqslant 0\}$. Thus, $\{w_i^2 = w_i\} \Big|\frac{w_i}{2} \{w_i \leqslant 1\}$. As a consequence, $\{w_i^2 = w_i\} \Big|\frac{w_i,v}{k+2} \left\{w_i p^2(\left\langle \Sigma^{\dagger/2}x_i, v\right\rangle) \leqslant p^2(\left\langle \Sigma^{\dagger/2}x_i, v\right\rangle)\right\}$. Summing up over $1 \leqslant i \leqslant n$ yields:

$$\left\{w_i^2 = w_i \mid \forall i\right\} \Big|\frac{w,v}{2k} \left\{\frac{1}{n}\sum_{i=1}^{n} w_i p^2\left(\left\langle \Sigma^{\dagger/2}x_i, v\right\rangle\right) \leqslant \frac{1}{n}\sum_{i=1}^{n} p^2\left(\left\langle \Sigma^{\dagger/2}x_i, v\right\rangle\right) \leqslant C\delta \|v\|_2^k\right\},$$

where in the final inequality on the RHS above, we used the second condition from Definition 2.2.12 satisfied by $\mathcal{S}$. Plugging this back in (5.4), we thus have:

$$\left\{w_i^2 = w_i \mid \forall i\right\} \Big|\frac{w,v}{2k} \left\{\frac{1}{n}\sum_{i=1}^{n}\|v\|_2^{k-2} w_i \left\langle \Sigma^{\dagger/2}x_i, v\right\rangle^2 \geqslant \delta^2\left(\frac{1}{n}\sum_{i=1}^{n} w_i - C\delta\right)\|v\|_2^k\right\}, \tag{5.5}$$

as desired.

□

Lemma 5.4.6 (Technical SoS fact about Powering)**.** *For indeterminates a, b, Z and any $t \in \mathbb{N}$,*

$$\{a \geqslant 0, b \geqslant 0, (a-b)Z \leqslant 0\} \Big|\frac{a,b}{t} \left\{(a^t - b^t)Z \leqslant 0\right\} \tag{5.6}$$

Proof. We have:

$$\{a \geqslant 0, b \geqslant 0\} \Big|\frac{a}{t} \left\{\sum_{i=0}^{t-1} a^{t-1-i}b^i \geqslant 0\right\}.$$

Using the above identity with (3.5) yields:

$$\{a \geqslant 0, b \geqslant 0, (a-b)Z \leqslant 0\} \Big|\frac{a,b}{t} \left\{(a-b)\left(\sum_{i=0}^{t-1} a^{t-1-i}b^i\right)Z \leqslant 0\right\}.$$

Using the identity: $(a^2 - \delta)\left(\sum_{i=0}^{t-1} a^{t-1-i}b^i\right) = a^t - b^t$, we finally obtain:

$$\{a \geqslant 0, b \geqslant 0, (a-b)Z \leqslant 0\} \Big|\frac{a,b}{t} \left\{\left(a^t - b^t\right)Z \leqslant 0\right\}.$$

□

Proof of Lemma 5.4.1. We begin by applying Lemma 5.4.5 to the set $\mathcal{I}$. Observe, the uniform distribution on $\mathcal{I}$ is k-certifiably (C, δ)-anti-concentrated. Thus,

$$\left\{w_i^2 = w_i \mid \forall i\right\} \Big|\frac{w,v}{2k} \left\{\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i \left\langle \mathbf{\Sigma}_*^{\dagger/2} x_i, v\right\rangle^2 \|v\|_2^{k-2} \geqslant \delta^2 \left(\frac{\sum_{i\in\mathcal{I}} w_i}{|\mathcal{I}|} - C\delta\right)\|v\|_2^k\right\} \tag{5.7}$$

Let $M = \mathbf{I} - \Pi$. Since $x_i = \Sigma_*^{1/2}\Sigma_*^{\dagger/2}x_i$, we have the following polynomial identity (in indeterminates Π, v) for any i:

$$\left\langle \Sigma_*^{-\dagger/2} x_i, \Sigma_*^{1/2} M v\right\rangle = \langle M x_i, v\rangle \ .$$

By using the (substitution) for manipulating SoS proofs and substituting v with the polynomial $\Sigma_*^{\dagger/2} M v$, we thus obtain:

$$\left\{\forall i \in [n]\ w_i^2 = w_i\right\} \Big|\frac{w,v}{2k} \left\{\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i \langle M x_i, v\rangle^2 \left\|\Sigma_*^{\dagger/2} M v\right\|_2^{k-2} \geqslant \delta^2 \left(\frac{\sum_{i\in\mathcal{I}} w_i}{|\mathcal{I}|} - C\delta\right)\left\|\Sigma_*^{\dagger/2} M v\right\|_2^k\right\} \tag{5.8}$$

Next, observe that $\mathcal{A}_{w,\Pi} \Big|\frac{w,\Pi}{2} \{w_i M x_i = 0\ \forall i\}$ and thus,

$$\mathcal{A}_{w,\Pi} \Big|\frac{w,v,\Pi}{4} \left\{\langle w_i M x_i, v\rangle^2 = w_i \langle M x_i, v\rangle^2 = 0\ \forall i\right\} \ .$$

Combining this with (5.8), we thus have:

$$\mathcal{A}_{w,\Pi} \Big|\frac{w,v}{2k} \left\{0 \geqslant \delta^2 \left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i - C\delta\right)\left\|\Sigma_*^{\dagger/2} M v\right\|_2^k\right\} \tag{5.9}$$

Using (3.5) to multiply throughout by the constant $1/\delta^2$ yields:

$$\mathcal{A}_{w,\Pi} \Big|\frac{w,v}{2k} \left\{0 \geqslant \left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i - C\delta\right)\left\|\Sigma_*^{\dagger/2} M v\right\|_2^k\right\} \tag{5.10}$$

Applying Lemma 5.4.6 with $a = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i$, $b = C\delta$ and $Z = \left\|\Sigma_*^{\dagger/2} Mv\right\|_2^k$, we obtain:

$$\mathcal{A}_{w,\Pi} \left|\frac{w,v}{2k+t}\right. \left\{ 0 \geqslant \left(\left(\frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} w_i \right)^t - (C\delta)^t \right) \left\|\Sigma_*^{\dagger/2} Mv\right\|_2^k \right\} \tag{5.11}$$

Let λ_{max} be the largest eigenvalue of Σ_*. By applying (3.5) and multiplying by $1/\lambda_{max}$ throughout, we can work with $1/\lambda_{max}\Sigma_*$ and thus assume that $\lambda_{max} = 1$. Let λ_{min} be the smallest non-zero eigenvalue of Σ_*. Then, $\lambda_{min} = \frac{1}{\kappa}$.

Recall, $\Sigma_* = U\Lambda U^\top$ and $\Pi_* = UU^\top$. Then, from the above, $\Sigma_* - \lambda_{min}\Pi_* \succeq 0$ and thus, we have:

$$\left|\frac{v,\Pi}{2}\right. \left\{ \lambda_{min} v^\top M\Pi_* Mv \leqslant v^\top M\Sigma_* Mv \right\}.$$

Using the (3.5) repeatedly we thus obtain:

$$\left|\frac{v}{k}\right. \left\{ \lambda_{min}^{k/2} \left(v^\top M\Pi_* Mv \right)^{k/2} \leqslant \left(v^\top M\Sigma_* Mv \right)^{k/2} \right\}. \tag{5.12}$$

Since $\lambda_{max} = 1$ and $M^2 = M$, we have:

$$\mathcal{A}_{w,\Pi} \left|\frac{v,\Pi}{4}\right. \left\{ v^\top M\Sigma_* Mv \leqslant \|Mv\|_2^2 = v^\top Mv = v^\top(\mathbf{I} - \Pi)v = \|v\|_2^2 - \|\Pi v\|_2^2 \leqslant \|v\|_2^2 \right\}$$

Using the (3.5) repeatedly again, we obtain:

$$\mathcal{A}_{w,\Pi} \left|\frac{v,\Pi}{4}\right. \left\{ \left(v^\top M\Sigma_* Mv \right)^{k/2} \leqslant \|v\|_2^k \right\} \tag{5.13}$$

Using (5.12) and (5.13) with (5.11), we thus have:

$$\begin{aligned} \mathcal{A}_{w,\Pi} \left|\frac{v,w}{2k}\right. \Bigg\{ & \left(\frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} w_i \right)^t \left(v^\top M\Pi_* Mv \right)^{k/2} \\ & \leqslant \frac{1}{\lambda_{min}^{k/2}} \left(\frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} w_i \right)^t \left(v^\top M\Sigma_* Mv \right)^{k/2} \\ & \leqslant \frac{1}{\lambda_{min}^{k/2}} \delta^t \|v\|_2^k \Bigg\}. \end{aligned} \tag{5.14}$$

Let $g \sim \mathcal{N}(0, I)$. Then, using the above with the substitution $v = g$, we have:

$$\mathcal{A}_{w,\Pi} \left|\frac{v,w}{2k}\right. \left\{ \left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i\right)^t \operatorname{tr}(M\Pi_* M)^{k/2} = \left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i\right)^t (\mathbf{E} g^\top M\Pi_* M g)^{k/2} \right.$$
$$\left. \leqslant \left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i\right)^t \mathbf{E}(g^\top M\Pi_* M g)^{k/2} \leqslant \frac{1}{\lambda_{min}^{k/2}}\delta^t \left\|Mg\right\|_2^k = r^{k/2}\frac{1}{\lambda_{min}^{k/2}}\delta^t \right\}, \quad (5.15)$$

where the inequality follows from the SoS Hölder's inequality.

Next,

$$\left\{\Pi^2 = \Pi\right\} \left|\frac{\Pi}{2}\right. \left\{\|\Pi\|_F^2 = \operatorname{tr}(\Pi^2) = \operatorname{tr}(\Pi) = r\right\}.$$

And also,

$$\left\{\Pi^2 = \Pi\right\} \left|\frac{\Pi}{2}\right. \left\{M^2 = (I-\Pi)^2 = I - 2\Pi + \Pi^2 = I - \Pi = M\right\}.$$

Thus,

$$\mathcal{A}_{w,\Pi} \left|\frac{\Pi}{2}\right. \left\{ \|\Pi - \Pi_*\|_F^2 = \|\Pi\|_F^2 + \|\Pi_*\|_F^2 - 2\operatorname{tr}(\Pi\Pi_*) = 2r - 2\operatorname{tr}(\Pi\Pi_*) \right.$$
$$\left. = 2\operatorname{tr}((I-\Pi)\Pi_*) = 2\operatorname{tr}(M\Pi_*) = 2\operatorname{tr}(M^2\Pi_*) = 2\operatorname{tr}(M\Pi_* M) \right\}.$$

$$\mathcal{A}_{w,\Pi} \left|\frac{\Pi}{2}\right. \left\{ \|\Pi - \Pi_*\|_F^2 = \|\Pi\|_F^2 + \|\Pi_*\|_F^2 - 2\operatorname{tr}(\Pi\Pi_*) = 2r - 2\operatorname{tr}(\Pi\Pi_*) \right.$$
$$\left. = 2\operatorname{tr}((I-\Pi)\Pi_*) = 2\operatorname{tr}(M\Pi_*) = 2\operatorname{tr}(M^2\Pi_*) = 2\operatorname{tr}(M\Pi_* M) \right\}.$$

Using (3.5) and combining with (5.15), we thus obtain:

$$\mathcal{A}_{w,\Pi} \left|\frac{\Pi,w}{2k+t}\right. \left\{ \left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i\right)^t \|\Pi - \Pi_*\|_F^k = \left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i\right)^t 2^{k/2}\operatorname{tr}(M\Pi_* M)^{k/2} \leqslant (2r/\lambda_{min})^{k/2}\delta^t \right\}.$$

Noting that $\lambda_{min} = 1/\kappa$ completes the proof.

□

5.4.3 High-Entropy Pseudo-distributions: Proof of Lemma 5.4.2

Fact 5.4.7 (Similar to the proof of Lemma 4.3 in [KKK19]). *Let $\tilde{\mu}$ be a pseudo-distribution of degree at least 2 on $w_1, w_2, \ldots, w_n$ that satisfies $\{w_i^2 = w_i \forall i\} \cup \{\sum_{i=1}^n w_i = \alpha n\}$ and minimizes $\left\|\sum_{i=1}^n \tilde{\mathbb{E}}[w_i]\right\|_2^2$. Then, $\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \tilde{\mathbb{E}}[w_i] \geqslant \alpha$.*

We defer the proof of this Fact to Appendix 5.6.1.

Proof of Lemma 5.4.2. From Fact 5.4.7, we have that $\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \tilde{\mathbb{E}}[w_i] \geqslant \alpha$. Applying Hölder's inequality for pseudo-distributions with $f = 1$ and $g = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} w_i$ gives:

$$\frac{1}{|\mathcal{I}|^t}\tilde{\mathbb{E}}\left(\sum_{i\in\mathcal{I}} w_i\right)^t \geqslant \frac{1}{|\mathcal{I}|^t}\left(\sum_{i\in\mathcal{I}} \tilde{\mathbb{E}}\, w_i\right)^t \geqslant \alpha^t\,.$$

□

5.4.4 Rounding Pseudo-distributions to a Large List: Proof of Lemma 5.4.3

In this subsection, we analyze Algorithm 118 and show that it returns a list $\mathcal{L}'$ that contains a projection matrix $\hat{\Pi}$ close to Π_*. The key step in our proof is the following lemma:

Lemma 5.4.8. *Given $t \in \mathbb{N}$, and an instance of Sub$_{\mathcal{D}}(\alpha, \Sigma_*)$ such that $\mathcal{I}$ is k-certifiably (C, δ)-anti-concentrated, let $\tilde{\mu}$ be a degree-$(2k+t)$ pseudo-distribution satisfying $\mathcal{A}_{w,\Pi}$ and minimizing $\|\tilde{\mathbb{E}}_{\tilde{\mu}}[w]\|_2$. Then,*

$$\frac{1}{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left(\sum_{i\in\mathcal{I}} w_i\right)^t\right]}\sum_{S\subseteq\mathcal{I},|S|\leqslant t}\binom{\mathcal{I}}{S}\tilde{\mathbb{E}}_{\tilde{\mu}}\left[w_S\,\|\Pi - \Pi_*\|_F^k\right] \leqslant \left(\frac{8\delta}{\alpha}\right)^t (2r\kappa)^{k/2}\,.$$

where $\binom{\mathcal{I}}{S}$ is the coefficient of the monomial indexed by S.

Proof. From Lemma 5.4.1, we have for every $t, \ell \in \mathbb{N}$,

$$\mathcal{A}_{w,\Pi}\left|\frac{w,\Pi}{t+k}\right.\left\{\frac{1}{|\mathcal{I}|^t}\left(\sum_{i\in\mathcal{I}} w_i\right)^t \|\Pi - \Pi_*\|_F^k \leqslant (2r\kappa)^{k/2}\delta^t\right\}.$$

Since $\tilde{\mu}$ satisfies $\mathcal{A}_{w,\Pi}$ and has degree at least $t+k$, taking pseudo-expectation yields:

$$\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\frac{1}{|\mathcal{I}|^t}\left(\sum_{i\in\mathcal{I}} w_i\right)^t \|\Pi-\Pi_*\|_F^k\right] \leqslant (2r\kappa)^{k/2}\delta^t\,.$$

Since $\tilde{\mu}$ satisfies $\mathcal{A}_{w,\Pi}$ and minimizes $\left\|\tilde{\mathbb{E}}_{\tilde{\mu}}\, w\right\|_2$, Lemma 5.4.2 yields: $\frac{1}{|\mathcal{I}|^t}\,\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left(\sum_{i\in\mathcal{I}} w_i\right)^t\right] \geqslant \alpha^t$. Multiplying both sides by $\frac{|\mathcal{I}|^t}{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left(\sum_{i\in\mathcal{I}} w_i\right)^t\right]} \leqslant \frac{1}{\alpha^t}$, we obtain:

$$\frac{1}{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left(\sum_{i\in\mathcal{I}} w_i\right)^t\right]}\,\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left(\sum_{i\in\mathcal{I}} w_i\right)^t \|\Pi-\Pi_*\|_F^k\right] \leqslant \left(\frac{8\delta}{\alpha}\right)^t (2r\kappa)^{k/2}\,. \tag{5.16}$$

For any monomial w_S, let $w_{S'}$ be its multilinearization. Then, observe that:

$$\left\{w_i^2 = w_i \mid \forall i\right\} \vdash_t^w \{w_S = w_{S'}\}\,.$$

Therefore, we have

$$\mathcal{A}_{w,\Pi} \vdash_t^w \left\{\left(\sum_{i\in\mathcal{I}} w_i\right)^t \|\Pi-\Pi_*\|_F^k = \sum_{S\subseteq\mathcal{I},|S|\leqslant t}\binom{\mathcal{I}}{S} w_S \|\Pi-\Pi_*\|_F^k\right\}\,. \tag{5.17}$$

Combining equations 5.16 and 5.17, we have

$$\frac{1}{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left(\sum_{i\in\mathcal{I}} w_i\right)^t\right]}\sum_{S\subseteq\mathcal{I},|S|\leqslant t}\binom{\mathcal{I}}{S}\tilde{\mathbb{E}}_{\tilde{\mu}}\left[w_S\|\Pi-\Pi_*\|_F^k\right] \leqslant \left(\frac{8\delta}{\alpha}\right)^t (2r\kappa)^{k/2}\,. \tag{5.18}$$

which concludes the proof.

□

Next, we show that sampling a subset of size t indicated by thee w's proportional to the marginal pseudo-distribution on this set results in an empirical estimator that is close to Π_* with constant probability.

Lemma 5.4.9. *Given $t\in\mathbb{N}$, let $\tilde{\mu}$ be a pseudo-distribution of degree at least $t+2k$ satisfying $\mathcal{A}_{w,\Pi}$ and minimizing $\|\tilde{\mathbb{E}}_{\tilde{\mu}}[w]\|_2$. Let $S\subseteq\mathcal{I}$, $|S|\leqslant t$ be chosen randomly with probability proportional to $\binom{\mathcal{I}}{S}\tilde{\mathbb{E}}_{\tilde{\mu}}[w_S]$. Let $\tilde{\mu}_S$ be the pseudo-distribution obtained by reweighting $\tilde{\mu}$ by the SoS polynomial w_S^2. Then, with probability at least $9/10$ over the draw of S, $\left\|\tilde{\mathbb{E}}_{\tilde{\mu}_S}[\Pi]-\Pi_*\right\|_F^k \leqslant 10(2r\kappa)^{k/2}(8\delta)^t\alpha^{-t}$.*

Proof. Rewriting the conclusion of Lemma 5.4.8, we have:

$$\frac{1}{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left(\sum_{i\in\mathcal{I}} w_i\right)^t\right]} \sum_{S\subseteq\mathcal{I},|S|\leqslant t} \binom{\mathcal{I}}{S} \tilde{\mathbb{E}}_{\tilde{\mu}}[w_S] \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[w_S \left\|\Pi-\Pi_*\right\|_F^k\right]}{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_S]} \leqslant \left(\frac{8\delta}{\alpha}\right)^t (2r\kappa)^{k/2}. \tag{5.19}$$

Further, $\sum_{S\subseteq\mathcal{I},|S|\leqslant t} \binom{\mathcal{I}}{S} \tilde{\mathbb{E}}_{\tilde{\mu}}[w_S] = \tilde{\mathbb{E}}_{\tilde{\mu}} \left(\sum_{i\in\mathcal{I}} w_i\right)^t$. Thus, $\frac{\binom{\mathcal{I}}{S}\tilde{\mathbb{E}}_{\tilde{\mu}}[w_S]}{\tilde{\mathbb{E}}_{\tilde{\mu}}\left(\sum_{i\in\mathcal{I}} w_i\right)^t}$ is a probability distribution, ζ, over $S\subseteq\mathcal{I}, |S|\leqslant t$. Thus, we can rewrite (5.19) as simply:

$$\mathbf{E}_{S\sim\zeta}\left[\frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_S \left\|\Pi-\Pi_*\right\|_F^k]}{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_S]}\right] \leqslant \left(\frac{8\delta}{\alpha}\right)^t (2r\kappa)^{k/2}.$$

By Markov's inequality, a $S\sim\zeta$ satisfies $\frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_S\|\Pi-\Pi_*\|_F^k]}{\tilde{\mathbb{E}}[w_S]} \leqslant 10(2r\kappa)^{k/2}(8\delta)^t\alpha^{-t}$ with probability at least $9/10$. Finally, observe that by Fact 5.3.1, $\tilde{\mathbb{E}}_{\tilde{\mu}_S}\left\|\Pi-\Pi_*\right\|_F^k = \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_S\|\Pi-\Pi_*\|_F^k]}{\tilde{\mathbb{E}}_{\tilde{\mu}}[w_S]}$. Thus, with probability at least $9/10$ over the choice of $S\sim\zeta$, $\tilde{\mathbb{E}}_{\tilde{\mu}_S}\left[\left\|\Pi-\Pi_*\right\|_F^k\right] \leqslant 10(2r\kappa)^{k/2}(8\delta)^t\alpha^{-t}$. By Cauchy-Schwarz inequality applied with $f=1$ and $g=\|(\Pi-\Pi_*)\|_F^k$, we have: $\left\|\tilde{\mathbb{E}}_{\tilde{\mu}}[(\Pi-\Pi_*)]\right\|_F^k \leqslant \tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left\|\Pi-\Pi_*\right\|_F^k\right]$. Thus, $\left\|\tilde{\mathbb{E}}_{\tilde{\mu}_S}[\Pi]-\Pi_*\right\|_F^k \leqslant 10(2r\kappa)^{k/2}(8\delta)^t\alpha^{-t}$. This completes the proof.

□

Proof of Lemma 5.4.3. We note that since $\mathcal{D}$ is k-certifiably (c,δ)-anti-concentrated, sampling $n_0 = (kd\log(d))^k$ suffices for the uniform distribution over $\mathcal{I}$ to be k-certifiably $(c,2\delta)$-anti-concentrated (this follows from Lemma 5.5.6). We then observe that by Lemma 5.4.7, $\frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left(\sum_{i\in\mathcal{I}} w_i\right)^t\right]}{|\mathcal{I}|^t} = \frac{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left(\sum_{i\in\mathcal{I}} w_i\right)^t\right]}{\tilde{\mathbb{E}}_{\tilde{\mu}}\left[\left(\sum_{i\in[n]} w_i\right)^t\right]} \geqslant \alpha^t$. Therefore, with probability at least $9\alpha^t/10$, $w_S\subset\mathcal{I}$ and the conclusion of Lemma 5.4.9 holds. However, the resulting matrix $\tilde{\Pi} = \tilde{\mathbb{E}}_{\tilde{\mu}_S}[\Pi]$ need not be a Projection matrix.

From Lemma 5.4.9, we can now conclude $\|\tilde{\Pi}-\Pi_*\|_F^2 \leqslant \left(10(2r\kappa)^{k/2}(8\delta/\alpha)^t\right)^{2/k} \leqslant cr\kappa(\delta/\alpha)^{2t/k}$. Setting $t = \Omega\left(\frac{k\log(r\kappa/\eta')}{\log(\delta/\alpha)}\right)$ in Lemma 5.4.9 suffices to bound $\|\tilde{\Pi}-\Pi_*\|_F^2 \leqslant \eta'$. It follows that with probability at least $9\alpha^t/10$, for all $i\in[d]$,

$$\lambda_i^2(\tilde{\Pi}) = \lambda_i^2(\Pi_*) \pm \left(10(2r\kappa)^{k/2}(8\delta)^t\alpha^{-t}\right) = \lambda_i^2(\Pi_*) \pm \eta'$$

Since Π_* is an actual rank-r Projection matrix, for $i\in[r]$, $\lambda_i^2(\tilde{\Pi})\in[1-\eta',1+\eta']$ and for $i\in[r+1,d]$, $\lambda_i^2(\tilde{\Pi})\in[-\eta',\eta']$. Recall, $\tilde{\Pi} = U\Lambda U^\top$ is the full Eigenvalue decomposition and therefore, $\|(\Lambda-\mathbf{I})\|_2^2 \leqslant \eta'$. Further, since $\sum_{i=1}^r \lambda_i^2(\tilde{\Pi}) \geqslant (1-\eta')r$ and trace is exactly r, $\|U_{\setminus r}\Lambda_{\setminus r}U_{\setminus r}^\top\|_F^2 = \sum_{i=r+1}^n \lambda_i^2(\tilde{\Pi}) \leqslant r\eta'$.

Now recall, $\hat{\Pi} = U_r U_r^\top$ is the corresponding Projection matrix we obtain in Algorithm 118, where U_r are the eigenvectors corresponding to the top-r eigenvalues of $\tilde{\Pi}$. Therefore,

$$\begin{aligned}\|\hat{\Pi} - \Pi_*\|_F^2 &= \|\hat{\Pi} - \tilde{\Pi} + \tilde{\Pi} - \Pi_*\|_F^2 \\ &\leqslant 2\left(\|\tilde{\Pi} - \Pi_*\|_F^2 + \|U_r U_r^\top - U\Lambda U^\top\|_F^2\right) \\ &\leqslant 4\left(\eta' + \|U_r(\Lambda_r - \mathbf{I}_r)U_r^\top\|_F^2 + \|U_{\setminus r}\Lambda_{\setminus r}U_{\setminus r}^\top\|_F^2\right) \\ &\leqslant 4\left(\eta' + \|(\Lambda_r - \mathbf{I}_r)\|_2^2\|\hat{\Pi}\|_F^2 + r\eta'\right) \leqslant 6r\eta'\end{aligned} \tag{5.20}$$

Setting $\eta' = \eta/6r$, we get $t = \Delta\left(\frac{k\log(r\kappa/\eta)}{\log(\delta/\alpha)}\right)$, for a sufficiently large constant Δ. Repeating $O(1/\alpha^t)$ times, with probability $99/100$, the resulting list contains a Projection matrix $\hat{\Pi}$ such that $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant \eta$. The claim follows by choosing $\delta = \alpha/2$. The running time is dominated by computing a $(t+2k)$-degree pseudo-distribution which requires $n^{O(k^2\log(r\kappa/\eta))}$ time.

□

5.4.5 Pruning the List: Proof of Lemma 5.4.4

Fact 5.4.10 (Concentration of Quadratic Forms of Subgaussians)**.** *Let x be a 1-subgaussian random variable on $\mathcal{R}^d$, i.e, $\mathbf{E}\exp(v^\top(x-\mu)) \leqslant \exp(\|v\|^2\sigma^2/2)$ for all $v \in \mathcal{R}^d$. Then, for any a matrix A and for any $t > 0$, we have*

$$\mathbf{Pr}\left[\left|\|Ax\|_2^2 - \mathbf{E}\|Ax\|_2^2\right| > t\right] \leqslant 2\exp\left(-\min\left(\frac{t^2}{\|A^\top A\|_F^2}, \frac{t}{\|A^\top A\|_2}\right)\right)$$

Fact 5.4.11 (Subspace Distance)**.** *Let Π_1, Π_2 be rank-r Projection matrices. Then, $\|(\mathbf{I}-\Pi_2)\Pi_1\|_F^2 = \frac{1}{2}\|\Pi_1 - \Pi_2\|_F^2$.*

Proof. Using $\|M\|_F^2 - \mathrm{Tr}\left[M^\top M\right]$, we have

$$\begin{aligned}\|(\mathbf{I} - \Pi_2)\Pi_1\|_F^2 = \mathrm{tr}\left[((\mathbf{I} - \Pi_2)\Pi_1)^\top(\mathbf{I} - \Pi_2)\Pi_1\right] &= \mathrm{tr}\left[\Pi_1(\mathbf{I} - \Pi_2)(\mathbf{I} - \Pi_2)\Pi_1\right] \\ &= \mathrm{tr}\left[\Pi_1\right] - \mathrm{tr}\left[\Pi_1\Pi_2\right] \\ &= \frac{1}{2}\left(\mathrm{tr}\left[\Pi_1^2\right] + \mathrm{tr}\left[\Pi_2^2\right] - 2\,\mathrm{tr}\left[\Pi_1\Pi_2\right]\right) \\ &= \frac{1}{2}\|\Pi_1 - \Pi_2\|_F^2\end{aligned} \tag{5.21}$$

where we repeatedly use $\Pi_1 = \Pi_1^2$, $\Pi_2 = \Pi_2^2$ and the cyclic property of the trace. $\square$

Lemma 5.4.12 (Testing Distinct Subspaces with One Sample)**.** *Let Σ_1 be any rank-r Covariance matrix and let $\mathcal{D}$ be a mean-zero subgaussian distribution with covariance Σ_1. Let Π_1 be the corresponding rank-r Projection matrix for Σ_1 and Π_2 be any fixed rank r Projection matrix. Then, for any $0 < \zeta < 1$,*

$$\begin{aligned}\mathbf{Pr}_{x\sim\mathcal{D}}\left[\|(\mathbf{I}-\Pi_2)x\|_2^2 \geqslant \frac{(1-\zeta)\lambda_{\min}}{2}\|\Pi_1-\Pi_2\|_F^2\right] \\ \geqslant 1-\exp\left(-c\min\left(\frac{\zeta^2}{\kappa^4},\frac{\zeta}{\kappa^2}\right)\left(\frac{\|\Pi_1-\Pi_2\|_F^2}{\|(\mathbf{I}-\Pi_2)\Pi_1\|_F^2}\right)\right),\end{aligned}$$

for a fixed constant c.

Proof. Let $\mathcal{D}'$ be a 1-subgaussian distribution. Observe,

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\|(\mathbf{I}-\Pi_2)x\|_2^2\right] = \mathop{\mathbb{E}}_{g\sim\mathcal{D}'}\left[\|(\mathbf{I}-\Pi_2)\Sigma_1^{\dagger/2}g\|_2^2\right] &\geqslant \mathop{\mathbb{E}}_{g\sim\mathcal{D}'}\left[\lambda_{\min}\|(\mathbf{I}-\Pi_2)\Pi_1 g\|_2^2\right] \\ &= \lambda_{\min}\|(\mathbf{I}-\Pi_2)\Pi_1\|_F^2 \\ &= \frac{\lambda_{\min}}{2}\|\Pi_1-\Pi_2\|_F^2\end{aligned} \tag{5.22}$$

where we use $\mathbb{E}_{g\sim\mathcal{D}'}[\|Mg\|_2^2] = \mathbb{E}_{g\sim\mathcal{D}'}\left[g^\top M^T M g\right] = Tr[M^T M]$ and Fact 5.4.11. Similarly, $\mathbb{E}_{x\sim\mathcal{D}}[\|(\mathbf{I}-\Pi_2)x\|_2^2] \leqslant \frac{\lambda_{\max}}{2}\|\Pi_1-\Pi_2\|_F^2$. Since $((\mathbf{I}-\Pi_2)$ is a projector, $\|(\mathbf{I}-\Pi_2)\Sigma_1\|_2 \leqslant \lambda_{\max}$. Applying Fact 5.4.10 with $A = ((\mathbf{I}-\Pi_2)\Sigma_1)^\top(\mathbf{I}-\Pi_2)\Sigma_1$, $\|A\|_2^2 = \|(\mathbf{I}-\Pi_2)\Sigma_1\|_2^2 \leqslant \lambda_{\max}^2\|(\mathbf{I}-\Pi_2)\Sigma_1\|_2^2$, $\|A^\top A\|_F = \|((\mathbf{I}-\Pi_2)\Sigma_1)^\top(\mathbf{I}-\Pi_2)\Sigma_1\|_F \leqslant \lambda_{\max}^2\cdot\|(\mathbf{I}-\Pi_2)\Pi_1\|_F$ and $t = \zeta\lambda_{\min}\|\Pi_1-\Pi_2\|_F^2/2$:

$$\begin{aligned}&\mathbf{Pr}_{x\sim\mathcal{D}}\left[\left|\|(\mathbf{I}-\Pi_2)x\|_2^2 - \frac{\lambda_{\min}}{2}\|\Pi_1-\Pi_2\|_F^2\right| > \frac{\zeta\lambda_{\min}}{2}\|\Pi_1-\Pi_2\|_F^2\right] \\ &\leqslant 2\exp\left(-c\min\left(\frac{\zeta^2}{\kappa^4},\frac{\zeta}{\kappa^2}\right)\left(\frac{\|\Pi_1-\Pi_2\|_F^2}{\|(\mathbf{I}-\Pi_2)\Pi_1\|_2^2}\right)\right)\end{aligned} \tag{5.23}$$

Rearranging the terms yields the claim. $\square$

We are now ready to prove Lemma 5.4.4:

Proof of Lemma 5.4.4. Let $100/\alpha^\gamma$ be the number of fresh samples we draw from $\mathrm{Sub}_{\mathcal{D}}(\alpha,\Sigma_*)$, for some $\gamma \geqslant 1$. Observe, by Markov, with probability $99/100$, there are at least γ contiguous

samples drawn from the inlier set $\mathcal{I}$. For the samples that are not inliers, we have no guarantees on the projector we add to our list $\mathcal{L}$. Let the i-th iteration of Algorithhm 119 correspond to $x_i, x_{i+1}, \dots x_{i+\gamma-1} \sim \mathcal{D}$. For a fixed projector $\hat{\Pi} \in \mathcal{L}'$ such that $\|\hat{\Pi} - \Pi_*\|_F^2 = \Omega(\kappa^4 \log(r\kappa) \log(1/\alpha)/\alpha^2\gamma)$, it follows from Lemma 5.4.12, that with probability at least $1 - \Omega(\alpha^{-(\kappa^4 \log(r\kappa) \log(1/\alpha)/\alpha^2\gamma)})$,

$$\|(\mathbf{I} - \hat{\Pi})x_i\|_2^2 \geqslant \frac{\lambda_{\min}}{4}\|\Pi_1 - \Pi_2\|_F^2 \geqslant \frac{\lambda_{\min}\kappa^4 \log(r\kappa) \log(1/\alpha)}{\alpha^2\gamma}$$

We then repeat the test γ times independently and thus with probabiltiy at least $1 - \Omega\left(\frac{1}{\alpha^{\log(r\kappa)\frac{\log(1/\alpha)}{\alpha^2}}}\right)$, there exists $\ell \in [\gamma]$, $\|(\mathbf{I} - \hat{\Pi})x_{i+\ell}\|_2^2 \geqslant \frac{\lambda_{\min}\kappa^4 \log(r\kappa) \log(1/\alpha)}{\alpha^2\gamma}$.

Since our list size is at most $O(1/\alpha^{\log(r\kappa)/\alpha^2})$, we can union bound over the failure probability for each projector in the list $\mathcal{L}'$. Therefore, with probability at least $99/100$, simultaneously for all projectors $\hat{\Pi} \in \mathcal{L}$, if $\|\hat{\Pi} - \Pi_*\|_F^2 = \Omega(\kappa^4 \log(r\kappa) \log(1/\alpha)/\alpha^2\gamma)$, there exists $\ell \in [\gamma]$,

$$\|(\mathbf{I} - \hat{\Pi})x_{i+\ell}\|_2^2 \geqslant \frac{\lambda_{\min}}{2}\|\hat{\Pi} - \Pi_*\|_F^2 > \lambda_{\min}\kappa^4 t \log(1/\alpha) \tag{5.24}$$

Recall, $\mathcal{D}$ has covariance Σ_* and if $x \sim \mathcal{D}$ $\mathbb{E}_{x\sim\mathcal{D}}[\|x\|_2^2] = \operatorname{tr}[\Sigma_*]$. By Markov, with probability at least $99/100$, $\|x\|_2^2 = O(\operatorname{tr}[\Sigma_*])$. Dividing out (5.24) by $\|x\|_2^2$, with probability at least $99/100$,

$$\frac{\|(\mathbf{I} - \hat{\Pi})x\|_2^2}{\|x\|_2^2} > \frac{\lambda_{\min}\kappa^4 t \log(1/\alpha)}{\operatorname{tr}[\Sigma_*]} = \Omega\left(\frac{\kappa^4 t \log(1/\alpha)}{\alpha^2\gamma}\right)$$

where the last inequality follows from $\lambda_{\min}/\operatorname{tr}[\Sigma_*] \geqslant 1$. Therefore, the set of projectors in the sub-list $\mathcal{L}_i'$ in Algorithm 119 only contains projectors $\hat{\Pi}$ such that $\|\hat{\Pi} - \Pi_*\|_F^2 \leqslant \left(\frac{\kappa^4 \log(r\kappa) \log(1/\alpha)}{\alpha^2\gamma}\right)$. By Lemma 5.4.9, $\mathcal{L}'$ is guanteed to have a projector $\bar{\Pi}$ such that $\|\bar{\Pi} - \Pi_*\|_F^2 \leqslant 0.1$. Observe,

$$\begin{aligned}\frac{\|(\mathbf{I} - \bar{\Pi})x\|_2^2}{\|x\|_2^2} = \|(\mathbf{I} - \bar{\Pi})x/\|x\|_2\|_2^2 = \|(\mathbf{I} - \bar{\Pi})\Pi_* g/\|g\|_2\|_2^2 &\leqslant \|(\mathbf{I} - \bar{\Pi})\Pi_*\|_F^2 \\ &= \frac{\|\bar{\Pi} - \Pi_*\|_F^2}{2} \\ &\ll \left(\frac{\kappa^4 \log(r\kappa) \log(1/\alpha)}{\alpha^2\gamma}\right)\end{aligned}$$

Therefore, $\mathcal{L}_i'$ is non-empty. Algorithm 119 selects one projector from $\mathcal{L}_i'$ arbitrarily, which completes the proof. □

5.5 Certifiable Anti-Concentration

In this section, prove basic facts about certifiable anti-concentration. We start by recalling the definition again.

Definition 5.5.1 (Certifiable Anti-Concentration)**.** *A zero-mean distribution D with covariance Σ is $2k$-certifiably $(\delta, C\delta)$-anti-concentrated if there exists a univariate polynomial p of degree d such that:*

1. $\left\lfloor \frac{v}{2k} \right\rfloor \left\{ \|v\|_2^{2k-2} \left\langle \Sigma^{\dagger/2} x, v \right\rangle^2 + \delta^2 p^2 \left(\left\langle \Sigma^{\dagger/2} x, v \right\rangle \right) \geqslant \frac{\delta^2 \|v\|_2^{2k}}{4} \right\}$.
2. $\left\lfloor \frac{v}{2k} \right\rfloor \left\{ \mathbf{E}_{x \sim D} \left[p^2 \left(\left\langle \Sigma^{\dagger/2} x, v \right\rangle \right) \right] \leqslant C\delta \|v\|_2^{2k} \right\}$.

A set $\mathcal{S}$ is $2k$-certifiably (C, δ)-anti-concentrated if the uniform distribution on $\mathcal{S}$ is $2k$-certifiably (C, δ)-anti-concentrated.

As discussed earlier, this definition is obtained by a important but technical modification of the definition used in [KKK19, RY20a]. We verify basic properties of this notion here and establish that natural distributions such as Gaussians do satisfy it. We first prove that natural distributions like the Gaussians and uniform distribution on the unit sphere are certifiably anti-concentrated.

Theorem 121. *(Certifiable Anti-Concentration of Gaussians.) Given $0 < \delta \leqslant 1/2$, there exists $k = O\left(\frac{\log^5(1/\delta)}{\delta^2}\right)$ such that $\mathcal{N}(0, \Sigma)$ is k-certifiably (C, δ)-anti-concentrated.*

Our proof of Theorem 121 will rely on the following construction of a low-degree polynomial with certain important properties:

Lemma 5.5.2 (Core Indicator for Strictly Sub-Exponential Tails)**.** *Given a univariate distribution $\mathcal{D}$ with mean 0 and variance $\sigma \leqslant 1$ such that*

1. ***Anti-Concentration:*** *for all $\eta > 0$, $\mathbf{Pr}_{x \sim \mathcal{D}}[|x| \leqslant \eta\sigma] \leqslant c_1 \eta$,*
2. ***Strictly Sub-Exponential Tail:*** *for all $k_1 < 2$, $\mathbf{Pr}_{x \sim \mathcal{D}}[|x| \geqslant t\sigma] \leqslant \exp(-t^{2/k_1}/c_2)$,*

for some fixed $c_1, c_2 > 1$. Then, for any $\delta > 0$, there exists a degree $d = O\left(\frac{\log^{(4+k_1)/(2-k_1)}(1/\delta)}{\delta^{2/(2-k_1)}}\right)$ even polynomial q satisfying:

1. *$|x| \leqslant \delta$, $q(x) = 1 \pm \delta$, and,*
2. *$\sigma^2 \mathbb{E}_{x \sim \mathcal{D}} [q^2(x)] \leqslant 10 c_1 c_2 \delta$.*

We will also use the following basic fact about even polynomials.

Lemma 5.5.3 (Structure of Even Polynomials)**.** *For any even univariate polynomial $q(t)$ of degree d, $\|v\|_2^{2d} q^2(\langle x, v\rangle / \|v\|_2)$ is a polynomial in vector-valued indeterminate v and further,*

$$\left|\frac{v}{2d}\right. \left\{\|v\|_2^{2d} q^2(\langle x, v\rangle / \|v\|_2) \geqslant 0\right\}.$$

Proof. The conclusion requires us to prove that $\|v\|_2^{2d} q^2(\langle x, v\rangle / \|v\|_2)$ is a sum-of-squares polynomial in vector-valued variable v. Let $q(t) = \sum_{i \in d} c_i t^i$. Since $q(t)$ is even,

$$q(t) = \frac{1}{2}(q(t) + q(-t)) = \frac{1}{2}\left(\sum_{i \in [d]} c_i t^i + c_i (-t)^i\right) = \sum_{1 \leqslant i \leqslant d/2} c_{2i} t^{2i}.$$

Thus, in particular, d is even and $q(t) = r(t^2)$ for some polynomial r of degree $d/2$. Substituting $t = \langle x, v\rangle / \|v\|_2$, we have; $\|v\|_2^{2d} q^2(\langle x, v\rangle / \|v\|_2) = \|v\|_2^{2d} \left(\sum_{i \leqslant d/2} c_{2i} \frac{\langle x, v\rangle^{2i}}{\|v\|_2^{2i}}\right)^2 = \left(\sum_{i \leqslant d/2} c_{2i} \|v\|_2^{d-2i} \langle x, v\rangle^{2i}\right)^2$ which is a sum-of-squares polynomial in v. □

Now, we are ready to prove that Gaussians are certifiably anti-concentrated under our new definition:

Proof of Theorem 121. Let $x \sim \mathcal{N}(0, \Sigma)$. We begin with the following polynomial :

$$p(v) = \|v\|_2^d q(\left\langle \Sigma^{\dagger/2} x, v \right\rangle / \|v\|_2)$$

where q is the degree $d = \Theta\left(\frac{\log^5(1/\delta)}{\delta^2}\right)$ polynomial from Lemma 5.5.2. By Fact 5.5.3, p is indeed a univariate polynomial in v. We will prove that $\mathcal{N}(0, \Sigma)$ is $2d$-certifiably (C, δ)-subgaussian for some some absolute constant $C > 0$ using the polynomial p.

Consider the polynomial $g(x) = x^2 + \delta^2 q^2(x) - \delta^2/4$. If $|x| > \delta$ then, $g(x) \geqslant 3\delta^2/4 \geqslant 0$. On the other hand, if $|x| \leqslant \delta$, using that $q^2(x) = (1 \pm \delta)^2 \geqslant \frac{1}{4}$ for every $\delta \leqslant 1/2$, $g(x) \geqslant 0$. Thus, g is a univariate, non-negative polynomial. Using Fact 2.2.10 we thus obtain:

$$\left|\frac{x}{2d}\right. \left\{x^2 + \delta^2 q^2(x) \geqslant \delta^2/4\right\},$$

or, equivalently, $x^2 + \delta^2 q^2(x) - \delta^2/4 = s(x)$ for a SoS polynomial s of degree at most $2d$. Since q is even, the LHS is invariant under the transformation $x \to -x$. Thus, s is an even polynomial.

Substituting $x = \frac{\left\langle \Sigma^{\dagger/2}x, v\right\rangle}{\|v\|_2}$, we thus have:

$$\frac{\left\langle \Sigma^{\dagger/2}x, v\right\rangle^2}{\|v\|_2^2} + \delta^2 q^2\left(\frac{\left\langle \Sigma^{\dagger/2}x, v\right\rangle}{\|v\|_2}\right) - \frac{\delta^2}{2} = s\left(\frac{\left\langle \Sigma^{\dagger/2}x, v\right\rangle}{\|v\|_2}\right)$$

Multiplying out by $\|v\|_2^{2d}$ and using the definition of p gives us the polynominal (in v) identity:

$$\|v\|_2^{2d-2}\left\langle \Sigma^{\dagger/2}x, v\right\rangle^2 + \delta^2 p^2\left(\left\langle \Sigma^{\dagger/2}x, v\right\rangle\right) - \frac{\delta^2\|v\|_2^{2d}}{4} = \|v\|_2^{2d} s\left(\frac{\left\langle \Sigma^{\dagger/2}x, v\right\rangle}{\|v\|_2}\right)$$

Since s is an even polynomial, it follows from Fact 5.5.3, $\|v\|_2^{2d} s\left(\frac{\left\langle \Sigma^{\dagger/2}x, v\right\rangle}{\|v\|_2}\right)$ is a sum-of-squares in v. Thus, we can conclude:

$$\left\{\frac{v}{2d}\right\}\left\{\|v\|_2^{2d-2}\left\langle \Sigma^{\dagger/2}x, v\right\rangle^2 + \delta^2 p^2\left(\left\langle \Sigma^{\dagger/2}x, v\right\rangle\right) \geqslant \frac{\delta^2\|v\|_2^{2d}}{4}\right\}$$

which completes the proof of the first inequality in Definition 2.2.12. By rotational invariance of Gaussians, $\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\langle x, v\rangle^\ell\right]$ is just a function of $\|v\|_2^{2\ell}$. Thus $\|v\|_2^{2t}\,\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma)}\left[q^2\left(\frac{\left\langle \Sigma^{\dagger/2}x, v\right\rangle}{\|v\|}\right)\right]$ is a polynomial in $\|v\|_2^2$. Since $\Sigma^{\dagger/2}x$ has variance 1, it follows from the definition of p and q that $\mathbf{E}_{x\sim D} p^2\left(\left\langle \Sigma^{\dagger/2}x, v\right\rangle\right) \leqslant C\delta\|v\|_2^d$, for $C = 10c_1c_2$. Therefore, applying Fact 2.2.10

$$\left\{\frac{\|v\|_2^2}{2d}\right\}\left\{\mathbf{E}_{x\sim D} p^2\left(\left\langle \Sigma^{\dagger/2}x, v\right\rangle\right) \leqslant C\delta\|v\|_2^{2d}\right\}$$

□

The proof above naturally extends to the uniform distribution on the unit sphere.

Theorem 122. *(Certifiable Anti-Concentration of Gaussians.) Given $0 < \delta \leqslant 1/2$, there exists $k = O\left(\frac{\log^5(1/\delta)}{\delta^2}\right)$ such that the uniform distribution on the unit sphere is k-certifiably (C, δ)-anti-concentrated.*

Next, we observe that our definition of certifiable anti-concentration is invariant under linear transformations:

Lemma 5.5.4. *(Affine Invariance.) Let $x \sim \mathcal{D}$ such that $\mathcal{D}$ is k-certifiably (C, δ)-anti-concentrated distribution. Then, for any $A \in \mathcal{R}^{m\times d}$, the random variable Ax has a k-certifiably (C, δ)-anti-concentrated distribution.*

In particular, this yields that certifiable-anti-concentration is preserved under taking linear projections of a distribution.

Corollary 5.5.5. *(Closure under taking projections) Let $x \sim \mathcal{D}$ such that $\mathcal{D}$ is k-certifiably (C,δ)-anti-concentrated distribution on $\mathcal{R}^d$. Let V be any subspace of $\mathcal{R}^d$ and let Π_V be the associated projection matrix. Then, the random variable $\Pi_V x$ has a k-certifiably (C,δ)-anti-concentrated distribution.*

Next, we show that anti-concentration is preserved under sampling, i.e. if $\mathcal{D}$ is anti-concentrated, then the uniform distribution over n samples from $\mathcal{D}$ is also anti-concentrated.

Lemma 5.5.6. *(Certifiable Anti-Concentration under Sampling.) Let $\mathcal{D}$ be k-certifiably (c,δ)-anti-concentrated Sub-Exponential distribution such that the certifying polynomial p has coefficients bounded by $d^{O(k)}$. Let $\mathcal{S}$ be a set of $n = \Omega((kd\log(d))^{O(k)}/C\delta)$ i.i.d. samples from $\mathcal{D}$. Then, with probability at least $1-1/d$, the uniform distribution on $\mathcal{S}$ is k-certifiably $(2c,\delta)$-anti-concentrated.*

Proof. Let p be a degree-k that witnesses anti-concentration of $\mathcal{D}$. We show that p also witnesses anti-concentration of the uniform distribution on $\mathcal{S}$, denoted by $\mathcal{D}'$. First, we observe that property 1 in definition 2.2.12 is point-wise and continues to hold for x sampled from $\mathcal{D}'$.

Further, we know that

$$\vdash_{\frac{v}{2k}} \left\{\mathbf{E}_{x\sim\mathcal{D}}\left[p^2\left(\left\langle \Sigma^{\dagger/2}x, v\right\rangle\right)\right] \leqslant C\delta\|v\|_2^{2k}\right\} \tag{5.25}$$

Since p^2 is a square polynomial, we can represent it in the monomial basis as $p^2\left(\left\langle \Sigma^{\dagger/2}x, v\right\rangle\right) = \left\langle c(\Sigma^{\dagger/2}x)c(\Sigma^{\dagger/2}x)^\top, (1,v)_{\otimes k}(1,v)_{\otimes k}^\top\right\rangle$, where $c(\Sigma^{\dagger/2}x)$ are the coefficients of $\mathbf{E}_{x\sim\mathcal{D}}p\left(\left\langle \Sigma^{\dagger/2}x, v\right\rangle\right)$ and $(1,v)_{\otimes k}$ are all monomials of degree at most d. For notational convenience let $c_x = c(\Sigma^{\dagger/2}x)$.

Since $\mathbf{E}_{x\sim\mathcal{D}}p^2\left(\left\langle \Sigma^{\dagger/2}x, v\right\rangle\right) = \left\langle \mathbf{E}_{x\sim\mathcal{D}}c_xc_x^\top, (1,v)_{\otimes k}(1,v)_{\otimes k}^\top\right\rangle$, and $\mathbf{E}_{x\sim\mathcal{D}}c_xc_x^\top = \mathbf{E}_{x\sim\mathcal{D}'}c_xc_x^\top + \left(\mathbf{E}_{x\sim\mathcal{D}}c_xc_x - \mathbf{E}_{x\sim\mathcal{D}'}c_xc_x^\top\right)$, using linearity of expectation and the (substitution) in (5.25),

$$\begin{aligned}
&\vdash_{\frac{v}{2k}} \left\{\left\langle \mathbf{E}_{x\sim\mathcal{D}'}c_xc_x^\top, (1,v)_{\otimes k}(1,v)_{\otimes k}^\top\right\rangle + \left\langle |\mathbf{E}_{x\sim\mathcal{D}}c_xc_x^\top - \mathbf{E}_{x\sim\mathcal{D}'}c_xc_x^\top|, (1,v)_{\otimes k}(1,v)_{\otimes k}^\top\right\rangle \leqslant C\delta\|v\|_2^{2k}\right\} \\
&\vdash_{\frac{v}{2k}} \left\{\left\langle \mathbf{E}_{x\sim\mathcal{D}'}c_xc_x^\top, (1,v)_{\otimes k}(1,v)_{\otimes k}^\top\right\rangle \leqslant \left\|\mathbf{E}_{x\sim\mathcal{D}}c_xc_x^\top - \mathbf{E}_{x\sim\mathcal{D}'}c_xc_x^\top\right\|_2\|v\|_2^{2k} + C\delta\|v\|_2^{2k}\right\}
\end{aligned} \tag{5.26}$$

where the second equation follows from Fact 3.2.19. Observe, it suffices to bound each entry,

i.e. for all $i, j \in [d^{2k}]$, $\left(\mathbb{E}_{x\sim\mathcal{D}'}\left[(c_x)_i(c_x)_j\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[(c_x)_i(c_x)_j\right]\right)^2 \leqslant (C^2\delta^2/d^{2k})$, with probability at least $1 - 1/d$. Then, using concentration of polynomials of Sub-exponential random variables, for all $i, j \in [d^k]$,

$$\mathbf{Pr}_{x\sim\mathcal{D}}\left[\left(\mathop{\mathbb{E}}_{x\sim\mathcal{D}'}\left[(c_x)_i(c_x)_j\right] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(c_x)_i(c_x)_j\right]\right)^2 > \epsilon^2\right] \leqslant \exp\left(-\left(\frac{\epsilon n}{\mathbb{E}_{x\sim\mathcal{D}}[(c_x)_i(c_x)_j]^2}\right)^{\frac{1}{2k}}\right)$$

Setting $\epsilon = C\delta/d^{2k}$ and union bounding over all i and j,

$$\mathbf{Pr}\left[\sum_{i,j\in[d^k]}\left(\mathbb{E}_{\mathcal{D}'}[(c_x)_i(c_x)_j] - \mathbb{E}_{\mathcal{D}}[(c_x)_i(c_x)_j]\right)^2 > C^2\delta^2\right] \leqslant d^{2k}\exp\left(-\left(\frac{n}{d^{O(k)}}\right)^{\frac{1}{2k}}\right)$$

where the bound on $\mathbb{E}_{x\sim\mathcal{D}}[(c_x)_i(c_x)_j]^2$ follows from our assumption on the coefficients of p. Setting $n = \Omega((kd\log(d))^{O(k)}/C\delta)$ suffices to bound the above probability by $1/d$. □

5.6 Appendix

We begin with showing that a d-dimension Gaussian vector that spans an $r \leqslant d$ subspace is δ-anti-concentrated in the subspace, for any $\delta > 0$.

Proposition 5.6.1 (Anti-Concentration)**.** *For all $\delta > 0$, $\mathbf{Pr}_{x\sim\mathcal{N}(0,\Sigma)}[|\langle x, v\rangle| \leqslant \delta\sqrt{v^\top\Sigma v}] \leqslant \delta$ whenever $v^\top\Sigma v > 0$.*

Proof. Let Σ be a rank-r covariance matrix and $\mathcal{N}(0,\Sigma)$ be the corresponding Gaussian distribution over vectors in $\mathbb{R}^d$. Let Π be the corresponding rank-r projection matrix. We first observe that only the subspace of $\mathbb{R}^d$ spanned by Σ has non-zero measure. Restricted to this subspace, we show that $x \sim \mathcal{N}(0,\Sigma)$ is δ-anti-concentrated for all $\delta > 0$. Note, this is equivalent to considering vectors of the form Πv for any $v \in \mathbb{R}^d$. Recall, the PDF of a multivariate Gaussian denoted by $\mathcal{N}(0,\Sigma)$ is given by

$$p(x,\Sigma) = \frac{1}{\sqrt{\det^\dagger(2\pi\Sigma^*)}}\exp\left(-\frac{1}{2}x^T\Sigma^\dagger x\right)$$

where $(\Sigma)^\dagger$ inverts the non-zero eigenvalues of Σ and $\det^\dagger$ is the pseudo-determinant. Now, we observe that for any non-zero $v \in \mathbb{R}^d$ and $x \sim \mathcal{N}(0, \Sigma)$, $\{\langle \Pi v, x\rangle = 0\}$ defines a rank-$(k-1)$ subspace. It is well known that the Gaussian measure on a lower dimensional subspace of span(Σ) is 0. Formally,

$$\int_{\langle \Pi v, x\rangle = 0} dp(x, \Sigma) = 0 \tag{5.27}$$

Therefore, for all $v \in \mathbb{R}^d$, $\mathbf{Pr}_{x\sim\mathcal{N}(0,\Sigma)}[\langle x, \Pi v\rangle = 0] = 0$. For all v in the kernel of Σ, $v^T\Sigma v = 0$.

For any v such that the quadratic form in non-zero, from stability of Gaussians, it follows that $\langle x, v\rangle \sim \mathcal{N}(0, v^T\Sigma v)$. Recall, the PDF of a univariate Gaussian denoted by $\mathcal{N}(0, v^\top\Sigma v)$ is given by

$$p(x) = \frac{1}{\sqrt{2\pi v^\top \Sigma v}} \exp\left(-\frac{x^2}{v^\top \Sigma v}\right)$$

Then,

$$\begin{aligned}\mathbf{Pr}\left[|x| \leqslant \delta\sqrt{v^\top\Sigma v}\right] &= \int_{-\delta\sqrt{v^\top\Sigma v}}^{\delta\sqrt{v^\top\Sigma v}} \frac{1}{\sqrt{2\pi v^\top \Sigma v}} \exp\left(-\frac{x^2}{v^\top \Sigma v}\right) dx \\ &\leqslant \int_{-\delta\sqrt{v^\top\Sigma v}}^{\delta\sqrt{v^\top\Sigma v}} \frac{1}{\sqrt{2\pi v^\top \Sigma v}} dx \leqslant \delta\end{aligned}$$

□

Using standard concentration arguments, we can derive a robust version of anti-concentration on a set of samples drawn from

Proposition 5.6.2 (Anti-Concentration of Gaussian Samples)**.** *Fix any $\delta > 0$ and let $\{x_1, x_2, \ldots, x_n\} \sim \mathcal{N}(0, \Sigma)$. Then, whenever $n \geqslant n_0$ for some $n_0 = \Omega(d/\delta^2)$, with probability at least $1 - 1/e^d$ over the draw of x_is, for every v such that $v^\top\Sigma v > 0$, $\frac{1}{n}\sum_{i=1}^n \mathbf{1}\left(|\langle x_i, v\rangle| < 2\delta\sqrt{v^\top\Sigma v}\right) \leqslant 2\delta$.*

Proof. By Proposition 5.6.1, for each $i \in [n]$, for all v, $\mathbf{Pr}[|\langle x_i, v\rangle| \leqslant \delta\sqrt{v^\top\Sigma v}] \leqslant \delta$. Therefore,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i\in[n]} \mathbf{1}\left(|\langle x_i, v\rangle| < \delta\sqrt{v^\top\Sigma v}\right)\right] = \frac{1}{n}\sum_{i\in[n]} \mathbf{Pr}\left[|\langle x_i, v\rangle| < \delta\sqrt{v^\top\Sigma v}\right] \leqslant \delta$$

By Chernoff, for any v,

$$\mathbf{Pr}\left[\frac{1}{n}\sum_{i\in[n]} \mathbf{1}\left(|\langle x_i, v\rangle| < \delta\sqrt{v^\top\Sigma v}\right) \geqslant 2\delta\right] \leqslant \exp\left(-\frac{4\delta n}{3}\right) \tag{5.28}$$

Next, we construct a $\delta/\sqrt{d}$ net in $\mathbb{R}^d$, denoted by $\mathcal{T}$, such that for any v, there exists $v' \in \mathcal{T}$ in the net and $\|v - v'\|_2 \leqslant \delta/\sqrt{d}$. By standard constructions, $|\mathcal{T}| \leqslant (\sqrt{d}/\delta)^d$. Then, by setting $n = \Omega(d \log(d/\delta))$, with probability at least $1 - 1/e^d$, for all $v' \in \mathcal{T}$,

$$\frac{1}{n} \sum_{i \in [n]} \mathbf{1}\left(|\langle x_i, v' \rangle| < \delta \sqrt{v'^\top \Sigma v'} \right) \leqslant 2\delta$$

By construction, for all $v \notin \mathcal{T}$, $|\langle x_i, v - v' \rangle| \leqslant \|x_i\|_2 \delta/\sqrt{d} \leqslant 2\delta$ and the claim follows. □

5.6.1 Proof of Fact 5.4.7

For completeness, we provide a proof of Fact 5.4.7. The proof strategy is similar to the proof of Lemma 4.3 in [KKK19].

Fact 5.5.2 *(High-Entropy Psuedo-Distribution Restated.) Let $\tilde{\mu}$ be a pseudo-distribution of degree at least 2 on $w_1, w_2, \ldots, w_n$ that satisfies $\{w_i^2 = w_i \forall i\} \cup \{\sum_{i=1}^n w_i = \alpha n\}$ and minimizes $\left\|\sum_{i=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}[w_i]\right\|_2^2$. Then, $\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \tilde{\mathbb{E}}_{\tilde{\mu}}[w_i] \geqslant \alpha$.*

Proof. Let $u = \frac{1}{\alpha n} \tilde{\mathbb{E}}[w]$ be a non-negative vector with entries summing to 1. Let $u_{\mathcal{I}} = \sum_{i \in \mathcal{I}} u_i$ denote the fraction of of mass on the inliers and $u_{\mathcal{O}} = 1 - u_{\mathcal{I}}$. Let $\tilde{\mu}$ be the minimal pseudo-distribution. For sake of contradiction, assume $u_{\mathcal{I}} < \alpha$. We can then exhibit a pseudo-distribution $\tilde{\mu}'$ that satisfies $\mathcal{A}$ and $\left\|\sum_{i=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}'}[w_i]\right\|_2^2 < \left\|\sum_{i=1}^n \tilde{\mathbb{E}}_{\tilde{\mu}}[w_i]\right\|_2^2$, contradicting minimality. Consider the real distribution $\tilde{\mu}^*$ that is supported on the inliers and $\Pi = \Pi_*$. This distribution clearly satisfies $\mathcal{A}_{w,\Pi}$ and thus any convex combination of $\tilde{\mu}'$ and $\tilde{\mu}$ also satisfies $\mathcal{A}_{w,\Pi}$. For some $\lambda > 0$, let $\tilde{\mu}_\lambda = \lambda \tilde{\mu}^* + (1 - \lambda)\tilde{\mu}$ be the corresponding mixed distribution.

We begin with lower bounding $\|u\|_2^2$ in terms of $u_{\mathcal{I}}$ and $u_{\mathcal{O}}$. It is easy to see that the minimum norm is obtained by spreading the mass $u_{\mathcal{I}}$ uniformly over the inliers and $u_{\mathcal{O}}$ uniformly over the outliers. Therefore,

$$\|u\|_2^2 \geqslant \left(\frac{u_{\mathcal{I}}}{\alpha n}\right)^2 \cdot \alpha n + \left(\frac{u_{\mathcal{O}}}{(1-\alpha)n}\right)^2 \cdot (1-\alpha)n = \frac{1}{\alpha n}\left(u_{\mathcal{I}}^2 + \left(u_{\mathcal{O}}^2 \cdot \frac{\alpha}{1-\alpha}\right)\right)$$

Now, consider $u_\lambda = \frac{1}{\alpha n} \tilde{\mathbb{E}}_{\tilde{\mu}_\lambda} w$. Then,

$$\|u_\lambda\|_2^2 = (1-\lambda)^2 \|u\|_2^2 + \frac{\lambda^2}{\alpha n} + 2\lambda(1-\lambda)\frac{u_{\mathcal{I}}}{\alpha n}$$

Therefore,

$$\begin{aligned}\|u_\lambda\|_2^2 - \|u_\lambda\|_2^2 &\geqslant \frac{\lambda}{\alpha n}\left((2-\lambda)\left(u_{\mathcal{I}}^2 + u_{\mathcal{O}}^2 \cdot \frac{\alpha}{1-\alpha}\right) - \lambda - 2(1-\lambda)\frac{u_{\mathcal{O}}}{\alpha n}\right) \\ &\geqslant \frac{\lambda(2-\lambda)}{\alpha n}\left(u_{\mathcal{I}}^2 + u_{\mathcal{O}}^2 \cdot \frac{\alpha}{1-\alpha} - u_{\mathcal{I}}\right)\end{aligned} \tag{5.29}$$

By assumption, $u_{\mathcal{I}} < \alpha$ and thus

$$\begin{aligned}u_{\mathcal{I}}^2 + (1-u_{\mathcal{I}})^2 \cdot \frac{\alpha}{1-\alpha} - u_{\mathcal{I}} &= \frac{(1-\alpha)u_{\mathcal{I}}(u_{\mathcal{I}}-1) + \alpha(1-u_{\mathcal{I}})^2}{1-\alpha} \\ &= \frac{(1-u_{\mathcal{I}})\left(\alpha(1-u_{\mathcal{I}}) - (1-\alpha)u_{\mathcal{I}}\right)}{1-\alpha} \\ &> 0\end{aligned}$$

Therefore, picking λ such that (5.29) is strictly greater than 0 suffices. □

5.6.2 Proof of Lemma 5.5.2

In this Subsection, we describe our construction of the core indicator polynomial. Our construction is derived from the polynomial approximation to the sign function in [DRST09] with a key difference. We do not require an upper envelope to the sign function, and thus obtain a simpler polynomial, which is even.

Lemma 5.5.2 *(Core Indicator Restated.) Given a univariate distribution $\mathcal{D}$ with mean 0 and variance $\sigma \leqslant 1$ such that*

1. ***Anti-Concentration:*** *for all $\eta > 0$, $\mathbf{Pr}_{x\sim\mathcal{D}}[|x| \leqslant \eta\sigma] \leqslant c_1\eta$,*
2. ***Sub-Exponential Tail:*** *for all $k < 2$, $\mathbf{Pr}_{x\sim\mathcal{D}}[|x| \geqslant t\sigma] \leqslant \exp(-t^{2/k}/c_2)$,*

for some fixed $c_1, c_2 > 1$. Then, for any $\delta > 0$, there exists a degree $d = O\left(\frac{\log^{(4+k)/(2-k)}(1/\delta)}{\delta^{2/(2-k)}}\right)$ even polynomial q such that for all $|x| \leqslant \delta$, $q(x) = 1 \pm \delta$ and $\sigma^2\,\mathbb{E}_{x\sim\mathcal{D}}\left[q^2(x)\right] \leqslant 10c_1c_2\delta$.

We start with recalling the following basic fact about growth of polynomials.

Fact 5.6.3. *(Growth of Polynomials [Riv74].) Let $a(x)$ be a polynomial of degree at most d such that $|a(x)| \leqslant b$ for all $x \in [-1,1]$. Then, $|a(x)| \leqslant b|2x|^d$ for all $|x| > 1$.*

We first show the existence of a low-degree indicator approximator polynomial that is even. We use an approximation to the sign function from [DRST09] :

Lemma 5.6.4. *(Sign Polynomial.) Let $a = \Theta(\epsilon^2/\log(1/\epsilon))$. There exists a degree-$O(1/a)$ polynomial $\ell(x)$ such that :*

1. *for all $|x| \in [a, 1]$, $\ell(x) \in [\mathit{sign}(x) - \epsilon^2, \mathit{sign}(x) + \epsilon^2]$*
2. *for all $x \in [-a, a]$, $\ell(x) \in [1 - \epsilon^2, 1 + \epsilon^2]$*
3. *ℓ is monotonically increasing in $(-\infty, -1] \cup [1, \infty)$*
4. *ℓ is an odd polynomial.*
5. *$|\ell(x)| \leqslant (1 + \epsilon^2)(|2x|)^d$ for all $|x| > 1$*

Proof. The first three properties follow from the construction in Theorem 4.5 [DRST09]. The fourth property follows from observing this polynomial has the form $\ell(x) = xr(x^2)$. From Fact 5.6.3, we can conclude that $|\ell(x)| \leqslant (1 + \epsilon^2)(|2x|)^d$ for all $|x| > 1$. □

Lemma 5.6.5. *(Indicator Polynomial.) Given $\delta > 0$ and $L \geqslant 1$, let $\epsilon^2 = \delta/L$. Then, there exists a polynomial q of degree $d = O(L \log(L/\delta)/\delta)$ such that $q(0) = 1$ and*

1. *q is an even polynomial.*
2. *$q(x) \in [-3\epsilon^2, 3\epsilon^2]$ for all $x \in [2\delta, L] \cup [-L, -2\delta]$.*
3. *$q(x) \in [-1 - \epsilon^2, 1 + \epsilon^2]$ for all $x \in [\delta, 2\delta] \cup [-2\delta, -\delta]$.*
4. *$q(x) \in [1 - 3\epsilon^2, 1 + 3\epsilon^2]$ for all $x \in [-\delta, \delta]$.*
5. *$q(x) < 4(|4x|)^d$ for all $|x| > L$.*

Proof. Let ℓ be the polynomial from Lemma 5.6.4. We then define

$$q(x) = \frac{\ell\left(\frac{x+\delta}{L} + a\right) - \ell\left(\frac{x-\delta}{L} - a\right)}{2\ell(\delta/L + a)}$$

It is easy to see $q(0) = 1$, since ℓ is an odd polynomial. Next, we observe that q is an even polynomial:

$$q(-x) = \frac{p\left(\frac{-x+\delta}{L} + a\right) - p\left(\frac{-x-\delta}{L} - a\right)}{2p(\delta/L + a)} = q(x)$$

Now, for all $x \in [\delta + 2aL, L]$, $\ell\left(\frac{x+\delta}{L} + a\right) = \text{sign}\left(\frac{x+\delta}{L} + a\right) \pm \epsilon^2 = 1 \pm \epsilon^2$ and $\ell\left(\frac{x-\delta}{L} - a\right) = 1 \pm \epsilon^2$ and thus assuming $\delta > \alpha$, $q(x) = \pm(4\epsilon^2)/2(1 \pm \epsilon^2) = \pm 3\epsilon^2$. A similar argument holds for $x \in [-L, -\delta - aL]$. Now, we show that $q(x)$ is close to 1 for $x \in [-\delta, \delta]$. Here,

$\ell\left(\frac{x+\delta}{L}+a\right)=1\pm\epsilon^2$ and $\ell\left(\frac{x-\delta}{L}-a\right)=-1\pm\epsilon^2$. Therefore, $q(x)=\frac{2\pm2\epsilon^2}{2\pm\epsilon^2}=1\pm3\epsilon^2$. Setting $aL=\delta$ suffices, therefore q has degree at most $O(L\log(L/\delta)/\delta)$. Further, for all $|x|\in[\delta,\delta+aL]$, $q(x)=\pm(1+\epsilon^2)$. Finally, for $|x|>L$, $q(x)\leqslant 4(|4x|)^d$. □

We can now blackbox the proof of Lemma A.1 from [KKK19] since the aforementioned Lemma constructs an appropriate polynomial to approximate the indicator function. Additionally, the polynomial we obtain is even and suffices for Lemma 5.5.2.

Chapter 6

Learning a Two-Layer Neural Network

6.1 Introduction

Neural networks have achieved remarkable success in solving many modern machine learning problems which were previously considered to be intractable. With the use of neural networks now being wide-spread in numerous communities, the optimization of neural networks is an object of intensive study.

Common usage of neural networks involves running stochastic gradient descent (SGD) with simple non-linear activation functions, such as the extremely popular ReLU function, to learn an incredibly large set of weights. This technique has enjoyed immense success in solving complicated classification tasks with record-breaking accuracy. However, theoretically the behavior and convergence properties of SGD are very poorly understood, and few techniques are known which achieve provable bounds for the training of large neural networks. This is partially due to the hardness of the problem – there are numerous formulations where the problem is known to be NP-hard [BR92, Jud88, BDL18, MR18]. Nevertheless, given the importance and success in solving this problem in practice, it is important to understand the source of this hardness.

Typically a neural network can be written in the following form: $\mathbf{A} = \boldsymbol{U}^i(\cdots \boldsymbol{U}^3 f(\boldsymbol{U}^2 f(\boldsymbol{U}^1 \mathcal{X}))$, where i is the depth of the network, $\mathcal{X} \in \mathbb{R}^{d \times n}$ is a matrix with columns corresponding to individual d-dimensional input samples, and $\mathbf{A}$ is the output labeling of $\mathcal{X}$. The functions f are applied entry-wise to a matrix, and are typically non-linear. Perhaps the most popular activation used in practice is the ReLU, given by $f(x) = \max\{0, x\}$. Here each $\boldsymbol{U}^i$ is an unknown linear map, representing the "weights", which maps inputs from one layer to the next layer. In the

reconstruction problem, when it is known that $\mathbf{A}$ and $\mathcal{X}$ are generated via the above model, the goal is to recover the matrices $\boldsymbol{U}^1, \ldots, \boldsymbol{U}^i$.

In this chapter, we consider the problem of learning the weights of two layer networks with a single non-linear layer. Such a network can be specified by two weight matrices $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$ and $\boldsymbol{V}^* \in \mathcal{R}^{k \times d}$, such that, on a d-dimensional input vector $x \in \mathcal{R}^d$, the classification of the network is given by $\boldsymbol{U}^* f(\boldsymbol{V}^* x) \in \mathcal{R}^m$. Given a training set $\mathcal{X} \in \mathcal{R}^{d \times n}$ of n examples, along with their labeling $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$, where $\mathbf{E}$ is a (possibly zero) noise matrix, the learning problem is to find $\boldsymbol{U}$ and $\boldsymbol{V}$ for which

$$\|\boldsymbol{U} - \boldsymbol{U}^*\|_F + \|\boldsymbol{V} - \boldsymbol{V}^*\|_F \leqslant \varepsilon$$

We consider two versions of this problem. First, in the noiseless (or realizable) case, we observe $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ precisely. In this setting, we demonstrate that exact recovery of the matrices $\boldsymbol{U}^*, \boldsymbol{V}^*$ is possible in polynomial time. Our algorithms, rather than exploiting smoothness of activation functions, exploit combinatorial properties of rectified activation functions. Additionally, we consider the more general noisy case, where we instead observe $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$, where $\mathbf{E}$ is a noise matrix which can satisfy various conditions. Perhaps the most common assumption in the literature [GKLW18, GLM17, JSA15] is that $\mathbf{E}$ has mean 0 and is sub-Gaussian. Observe that the first condition is equivalent to the statement that $\mathbb{E}[\mathbf{A} \mid \mathcal{X}] = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$. While we primarily focus on designing polynomial time algorithms for this model of noise, in Section 6.7 we demonstrate fixed-parameter tractable (in the number k of ReLUs) algorithms to learn the underlying neural network for a much wider class of noise matrices $\mathbf{E}$. We predominantly consider the *identifiable* case where $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$ has full column rank, however we also provide supplementary algorithms for the exact case when $m < k$. Our algorithms are robust to the behavior of $f(x)$ for positive x, and therefore generalize beyond the ReLU to a wider class of rectified functions f such that $f(x) = 0$ for $x \leqslant 0$ and $f(x) > 0$ otherwise.

It is known that stochastic gradient descent cannot converge to the ground truth parameters when f is ReLU and $\boldsymbol{V}^*$ is orthonormal, even if we have access to an infinite number of samples [LSSS14]. This is consistent with empirical observations and theory, which states that over-parameterization is crucial to train neural networks successfully [Har14, SC16]. In contrast, in this work we demonstrate that we can approximate the optimal parameters in the noisy case, and obtain the optimal parameters exactly in the realizable case, in polynomial time, without over-parameterization. In other words, we provide algorithms that do not succumb to spurious local minima, and can converge to the global optimum efficiently, without over-parametrization.

6.1.1 Our Contributions

We now state our results more formally. We consider 2-layer neural networks with ReLU-activation functions f. Such a neural network is specified by matrices $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$ and $\boldsymbol{V}^* \in \mathcal{R}^{k \times d}$. We are given d-dimensional input examples $x^i \in \mathcal{R}^d$, which form the columns of our input matrix $\mathcal{X}$, and also give the network's m-dimensional classification of $\mathcal{X}$, which is $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$, where f is applied entry-wise. We note that our formulation corresponds to having one non-linear layer.

Worst Case Upper Bounds. In the worst case setting, no properties are assumed on the inputs $\mathcal{X}, \mathbf{A}$. While this problem is generally assumed to be intractable, we show, perhaps surprisingly, that when $\text{rank}(\mathbf{A}) = k$ and $k = O(1)$, polynomial time exact algorithms do exist. One of our primary techniques throughout this work is the leveraging of combinatorial aspects of the ReLU function. For a row $f(\boldsymbol{V}^* \mathcal{X})_{i,*}$, we define a *sign pattern* of this row to simply be the subset of positive entries of the row. Thus, a sign pattern of a vector in $\mathcal{R}^n$ is simply given by the orthant of $\mathcal{R}^n$ in which it lies. We first prove an upper bound of $O(n^k)$ on the number of orthants which intersect with an arbitrary k-dimensional subspace of $\mathcal{R}^n$. Next, we show how to enumerate these sign patterns in time $n^{k+O(1)}$.

We use this result to give an $n^{O(k)}$ time algorithm for the neural network learning problem in the *realizable case*, where $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ for some fixed *rank-k* matrices $\boldsymbol{U}^*, \boldsymbol{V}^*$. After fixing a sign pattern of $f(\boldsymbol{V}^* \mathcal{X})$, we can effectively "remove" the non-linearity of f. Even so, the learning problem is still non-convex, and cannot be solved in polynomial time in the general case (even for fixed k). We show, however, that if the *rank* of $\mathbf{A}$ is k, then it is possible to use a sequence of linear programs to recover $\boldsymbol{U}^*, \boldsymbol{V}^*$ in polynomial time given the sign pattern, which allows for an $n^{O(k)}$ overall running time. Our theorem is stated below.

Theorem 123. *Given $\mathbf{A} \in \mathcal{R}^{m \times n}, \mathcal{X} \in \mathcal{R}^{d \times n}$, such that $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ and $\mathbf{A}$ is rank k, there is an algorithm that finds $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}, \boldsymbol{V}^* \in \mathcal{R}^{k \times d}$ such that $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ and runs in time poly$(n, m, d) \min\{n^{O(k)}, 2^n\}$.*

Worst Case Lower Bounds. Our upper bound relies crucially on the fact that $\mathbf{A}$ is rank k, which is full rank when $k \leqslant d, m$. We demonstrate that an $O(n^k)$ time algorithm is no longer possible without this assumption by proving the NP-hardness of the realizable learning problem when $\text{rank}(\mathbf{A}) < k$, which holds even for k as small as 2. Our hardness result is as follows.

Theorem 126. *For a fixed $\alpha \in \mathcal{R}^{m\times k}, \mathcal{X} \in \mathcal{R}^{d\times n}, \mathbf{A} \in \mathcal{R}^{m\times n}$, the problem of deciding whether there exists a solution $\boldsymbol{V} \in \mathcal{R}^{k\times d}$ to $\alpha f(\boldsymbol{V}\mathcal{X}) = \mathbf{A}$ is NP-hard even for $k = 2$. Furthermore, for the case for $k = 2$, the problem is still NP-hard when $\alpha \in \mathcal{R}^{m\times 2}$ is allowed to be a variable.*

Gaussian Inputs. Since non-convex optimization problems are known to be NP-hard in general, it is, perhaps, unsatisfying to settle for worst-case results. Typically, in the learning community, to make problems tractable it is assumed that the input data is drawn from some underlying distribution that may be unknown to the algorithm. So, in the spirit of learning problems, we make the common step of assuming that the samples in $\mathcal{X}$ have a standard Gaussian distribution. More generally, our algorithms work for arbitrary multi-variate Gaussian distributions over the columns of $\mathcal{X}$, as long as the covariance matrix is non-degenerate, i.e., full rank (see Remark 127). In this case, our running time and sample complexity will blow up by the condition number of the covariance matrix, which we can estimate first using standard techniques. For simplicity, we state our results here for $\mathbf{\Sigma} = \mathbb{I}$, though, for the above reasons, all of our results for Gaussian inputs $\mathcal{X}$ extend to all full rank $\mathbf{\Sigma}$

Furthermore, because many of our primary results utilize the combinatorial sparsity patterns of $f(\boldsymbol{V}\mathcal{X})$, where $\mathcal{X}$ is a Gaussian matrix, we do not rely on the fact that $f(x)$ is linear for $x > 0$. For this reason, our results generalize easily to other *non-linear* rectified functions f. In other words, any function f given by

$$f(x) = \begin{cases} 0 & \text{if } x \leqslant 0 \\ \phi(x) & \text{otherwise} \end{cases}$$

where $\phi(x) : [0, \infty] \to [0, \infty]$ is a continuous, injective function. In particular, our bounds do not change for polynomial valued $\phi(x) = x^c$ for $c \in \mathbb{N}$. Note, however, that our worst-case, non-distributional algorithms (stated earlier), where $\mathcal{X}$ is a fixed matrix, do not generalize to non-linear $\phi(x)$.

We first consider the noiseless setting, also referred to as the exact or realizable setting. Here $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ is given for rank k matrices $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$, where $\mathcal{X}$ has non-degenerate Gaussian marginals. The goal is then to recover the weights $(\boldsymbol{U}^*)^T, \boldsymbol{V}^*$ exactly up to a permutation of their rows (since one can always permute both sets of rows without effecting the output of the network). Note that for any positive diagonal matrix $\boldsymbol{D}$, $\boldsymbol{U}^* f(\boldsymbol{D}\boldsymbol{V}^*\mathcal{X}) = \boldsymbol{U}^*\boldsymbol{D} f(\boldsymbol{V}^*\mathcal{X})$ when f is the ReLU. Thus recovery of $(\boldsymbol{U}^*)^T, \boldsymbol{V}^*$ is always only possible up to a permutation and positive scaling. We now state our main theorem for the exact recovery of the weights in the realizable (noiseless) setting.

Theorem 134. *Suppose* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ *where* $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}, \boldsymbol{V}^* \in \mathcal{R}^{k \times d}$ *are both rank-*k*, and such that the columns of* $\mathcal{X} \in \mathcal{R}^{d \times n}$ *are mean* 0 *i.i.d. Gaussian. Then if* $n = \Omega(\text{poly}(d, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*)))$*, then there is a* $\text{poly}(n)$*-time algorithm which recovers* $(\boldsymbol{U}^*)^T, \boldsymbol{V}^*$ *exactly up to a permutation of the rows with high probability.*

To the best of our knowledge, this is the first algorithm which learns the weights matrices of a two-layer neural network with ReLU activation *exactly* in the noiseless case and with Gaussian inputs $\mathcal{X}$. Our algorithm first obtains good approximations to the weights $\boldsymbol{U}^*, \boldsymbol{V}^*$, and concludes by solving a system of judiciously chosen linear equations, which we solve using Gaussian elimination. Therefore, we obtain exact solutions in polynomial time, without needing to deal with convergence guarantees of continuous optimization primitives. Furthermore, to demonstrate the robustness of our techniques, we show that using results introduced in the concurrent and independent work of Ge et. al. [GKLW18], we can extend Theorem 134 to hold for inputs sampled from symmetric distributions (we refer the reader to Corollary 6.4.21). We note that [GKLW18] recovers the weight matrices up to additive error ε and runs in poly $\left(\frac{1}{\varepsilon}\right)$-time, whereas our algorithm has no ε dependency.

The runtime of our algorithm depends on the condition number $\kappa(\boldsymbol{V}^*)$ of $\boldsymbol{V}^*$, which is a fairly ubiquitous requirement in the literature for learning neural networks, and optimization in general [GKLW18, JSA15, LSW15, CMTV17, AGMR17, ZSJ+17, SJA16]. To address this dependency, in Lemma 6.4.22 we give a lower bound which shows at least a linear dependence on $\kappa(\boldsymbol{V}^*)$ is necessary in the sample and time complexity.

Next, we introduce an algorithm for approximate recovery of the weight matrices $\boldsymbol{U}^*, \boldsymbol{V}^*$ when $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$ for Gaussian marginals $\mathcal{X}$ and an i.i.d. sub-Gaussian mean-zero noise matrix $\mathbf{E}$ with variance σ^2.

Theorem 138. *Let* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$ *be given, where* $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}, \boldsymbol{V}^* \in \mathcal{R}^{k \times d}$ *are rank-*k*,* $\mathbf{E}$ *is a matrix of i.i.d. mean-zero sub-Gaussian random variables with variance* σ^2*, and such that the columns of* $\mathcal{X} \in \mathcal{R}^{d \times n}$ *are i.i.d. Gaussian. Then given* $n = \Omega\left(\text{poly}\left(d, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*), \sigma, \frac{1}{\varepsilon}\right)\right)$*, there is an algorithm that runs in* $\text{poly}(n)$ *time and w.h.p. outputs* $\boldsymbol{V}, \boldsymbol{U}$ *such that*

$$\|\boldsymbol{U} - \boldsymbol{U}^*\|_F \leqslant \varepsilon \qquad \|\boldsymbol{V} - \boldsymbol{V}^*\|_F \leqslant \varepsilon$$

Again, to the best of our knowledge, this work is the first which learns the weights of a 2-layer network in this noisy setting without additional constraints, such as the restriction that $\boldsymbol{U}$ be

positive. Recent independent and concurrent work, using different techniques, achieves similar approximate recovery results in the noisy setting [GKLW18]. We note that the algorithm of Goel et. al. [GK17] that [GKLW18] uses, crucially requires the linearity of the ReLU for $x > 0$, and thus the work of [GKLW18] does not generalize to the larger class of rectified functions which we handle. We also note that the algorithm of [GLM17] requires $\boldsymbol{U}^*$ to be non-negative. Finally, the algorithms presented in [JSA15] work for activation functions that are thrice differentiable and can only recover rows of $\boldsymbol{V}^*$ up to ± 1 scaling. Note, for the ReLU activation function, we need to resolve the signs of each row.

Fixed-Parameter Tractable Algorithms. For several harder cases of the above problems, we are able to provide Fixed-Parameter Tractable algorithms. First, in the setting where the "labels" are vector valued, i.e., $m > 1$, we note prior results, not restricted to ReLU activation, require the rank of $\boldsymbol{U}^*$ to be k [GKLW18, JSA15, GLM17]. This implies that $m \geqslant k$, namely, that the output dimension of the neural net is at least as large as the number k of hidden neurons. Perhaps surprisingly, however, we show that even when $\boldsymbol{U}^*$ does not have full column rank, we can still recover $\boldsymbol{U}^*$ exactly in the realizable case, as long as no two columns are non-negative scalar multiples of each other. Note that this allows for columns of the form $[u, -u]$ for $u \in \mathcal{R}^m$ as long as u is non-zero. Our algorithm for doing so is fixed paramater tractable in the condition number of $\boldsymbol{V}^*$ and the number of hidden neurons k. Our results rely on proving bounds on the sample complexity in order to obtain all 2^k possible sparsity patterns of the k-dimensional columns of $f(\boldsymbol{V}^*\mathcal{X})$.

Theorem 140. *Suppose $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ for $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}$ for any $m \geqslant 1$ such that no two columns of $\boldsymbol{U}^*$ are non-negative scalar multiples of each other, and $\boldsymbol{V}^* \in \mathcal{R}^{k\times n}$ has rank$(\boldsymbol{V}^*) = k$, and $n > \kappa^{O(k)}$poly(dkm). Then there is an algorithm which recovers $\boldsymbol{U}^*, \boldsymbol{V}^*$ exactly with high probability in time $\kappa^{O(k)}$poly(d, k, m).*

Furthermore, we generalize our results in the noisy setting to *arbitrary* error matrices $\|\mathbf{E}\|$, so long as they are independent of the Gaussians $\mathcal{X}$. In this setting, we consider a slightly different objective function, which is to find $\boldsymbol{U}, \boldsymbol{V}$ such that $\boldsymbol{U} f(\boldsymbol{V}\mathcal{X})$ approximates $\mathbf{A}$ well, where the measure is to compete against the optimal generative solution $\|\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X}) - \mathbf{A}\|_F = \|\mathbf{E}\|_F$. Our results are stated below.

Theorem 142. *Let $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X}) + \mathbf{E}$ be given, where $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ are rank-k, and $\mathbf{E} \in \mathcal{R}^{m\times n}$ is any matrix independent of $\mathcal{X}$. Then there is an algorithm which outputs*

$\boldsymbol{U} \in \mathcal{R}^{m\times k}, \boldsymbol{V} \in \mathcal{R}^{k\times d}$ in time $(\kappa/\varepsilon)^{O(k^2)}$poly$(n,d,m)$ such that with probability $1-\exp(-\sqrt{n})$ we have

$$\|\mathbf{A} - \boldsymbol{U}f(\boldsymbol{V}\mathcal{X})\|_F \leqslant \|\mathbf{E}\|_F + O\Big(\Big[\sigma_{\max}\varepsilon\sqrt{nm}\|\mathbf{E}\|_2\Big]^{1/2}\Big),$$

where $\|\mathbf{E}\|_2$ is the spectral norm of $\mathbf{E}$.

Note that the above error bounds depend on the flatness of the spectrum of $\mathbf{E}$. In particular, our bounds give a $(1+\varepsilon)$ approximation whenever the spectral norm of $\mathbf{E}$ is a $\sqrt{m}$ factor smaller than the Frobenius norm, as is in the case for a wide class of random matrices [Ver10b]. When this is not the case, we can scale ε by $1/\sqrt{m}$, to get an $(m\kappa/\varepsilon)^{O(k^2)}$-time algorithm which gives a $(1+\varepsilon)$ approximation for any error matrix $\mathbf{E}$ independent of $\mathcal{X}$ such that $\|\mathbf{E}\|_F = \Omega(\varepsilon\|\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})\|_F)$.

Sparse Noise. Finally, we show that for *sparse noise*, when the network is *low-rank* we can reduce the problem to the problem of exact recovery in the noiseless case. Here, by low-rank we mean that $m > k$. It has frequently been observed in practice that many pre-trained neural-networks exhibit correlation and a low-rank structure [DSD+13, DZB+14]. Thus, in practice it is likely that k need not be as large as m to well-approximate the data. For such networks, we give a polynomial time algorithm for Gaussian $\mathcal{X}$ for exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$. Our algorithm assumes that $\boldsymbol{U}^*$ has orthonormal columns, and satisfies an *incoherence* property, which is fairly standard in the numerical linear algebra community [CR07, CR09, KMO10, CLMW11, JNS13, Har14]. Formally, assume $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X}) + \mathbf{E}$ where $\mathcal{X}$ is i.i.d. Gaussian, and $\mathbf{E}$ is obtained from the following sparsity procedure. First, fix any matrix $\overline{\mathbf{E}}$, and randomly choose a subset of $nm - s$ entries for some $s < nm$, and set them equal to 0. The following result states that we can exactly recover $\boldsymbol{U}^*, \boldsymbol{V}^*$ in polynomial time even when $s = \Omega(mn)$.

Theorem 144 & Corollary 6.8.4. *Let $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ be rank k matrices, where $\boldsymbol{U}^*$ has orthonormal columns, $\max_i \|(\boldsymbol{U}^*)^T e_i\|_2^2 \leqslant \frac{\mu k}{m}$ for some μ, and $k \leqslant \frac{m}{\overline{\mu}\log^2(n)}$, where $\overline{\mu} = O\Big((\kappa(\boldsymbol{V}^*))^2\sqrt{k\log(n)\mu} + \mu + (\kappa(\boldsymbol{V}^*))^4\log(n)\Big)$. Here $\kappa(\boldsymbol{V}^*)$ is the condition number of $\boldsymbol{V}^*$. Let $\mathbf{E}$ be generated from the s-sparsity procedure with $s = \gamma nm$ for some constant $\gamma > 0$ and let $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}\mathcal{X}) + \mathbf{E}$. Suppose the sample complexity satisfies $n =$ poly$(d,m,k,\kappa(\boldsymbol{V}^*))$ Then on i.i.d. Gaussian input $\mathcal{X}$ there is a poly(n) time algorithm that recovers $\boldsymbol{U}^*, \boldsymbol{V}^*$ exactly up to a permutation and positive scaling with high probability.*

6.1.2 Related Work

Recently, there has been a flurry of work developing provable algorithms for learning the weights of a neural network under varying assumptions on the activation functions, input distributions, and noise models [SJA16, ABMM16, GKKT16, MR18, ZSJ+17, GKLW18, GLM17, ZSJ+17, Tia17a, LY17a, BG17, Sol17, GKM18, DG18]. In addition, there have been a number of works which consider lower bounds for these problems under a similar number of varying assumptions [GKKT16, LSSS14, ZLJ16, SJA16, ABMM16, BDL18, MR18]. We describe the main approaches here, and how they relate to our problem.

Learning ReLU Networks without noise. In the noiseless setting with Gaussian input, the results of Zhong et al. [ZSJ+17] utilize a similar strategy as ours. Namely, they first apply techniques from tensor decomposition to find a good initialization of the weights, whereafter they can be learned to a higher degree of accuracy using other methods. At this point our techniques diverge, as they utilize gradient descent on the initialized weights, and demonstrate good convergence properties for *smooth* activation functions. However, their results do not give convergence guarantees for non-smooth activation functions, including the ReLU and the more general class of rectified functions considered in this work. In this work, once we are given a good initialization, we utilize combinatorial aspects of the sparsity patterns of ReLU's, as well as solving carefully chosen linear systems, to obtain exact solutions.

Li and Yuan [LY17b] also analyize stochastic gradient descent, and demonstrate good convergence properties when the weight matrix $\boldsymbol{V}^*$ is known to be close to the identity, and $\boldsymbol{U}^* \in \mathcal{R}^{1\times k}$ is the all 1's vector. In [Tia17b], stochastic gradient descent convergence is also analyzed when $\boldsymbol{U}^* \in \mathcal{R}^{1\times k}$ is the all 1's vector, and when $\boldsymbol{V}^*$ is orthonormal. Moreover, [Tia17b] does not give bounds on sample complexity, and requires that a good initialization point is already given.

For uniformly random and sparse weights in $[-1, 1]$, Arora et al. [ABGM14] provide polynomial time learning algorithms. In [BG17], the learning of *convolutions neural networks* is considered, where they demonstrate global convergence of gradient descent, but do not provide sample complexity bounds.

Learning ReLU Networks with noise. Ge et al. [GLM17] considers learning a ReLU network with a single output dimension $\mathbf{A} = u^T f(\boldsymbol{V}\mathcal{X}) + \mathbf{E}$ where $u \in \mathcal{R}^k$ is restricted to be entry-wise positive and $\mathbf{E}$ is a zero-mean sub-Gaussian noise vector. In this setting, it is shown that the weights $u, \boldsymbol{V}$ can be approximately learned in polynomial time when the input $\mathcal{X}$ is i.i.d. Gaussian. However, in contrast to the algorithms in this work, the algorithm of [GLM17]

relies heavily on the non-negativity of u [Ge18], and thus cannot generalize to arbitrary u. Janzamin, Sedghi, and Anandkumar [JSA15] utilize tensor decompositions to approximately learn the weights in the presence of mean zero sub-Gaussian noise, when the activation functions are smooth and satisfy the property that $f(x) = 1 - f(-x)$. Using similar techniques, Sedghi and Anandkumar [SJA16] provide a polynomial time algorithm to approximate the weights, if the weights are sparse.

A more recent result of Ge et al. demonstrates polynomial time algorithms for learning weights of two-layer ReLU networks in the presence of mean zero sub-gaussian noise, when the input is drawn from a mixture of a symmetric and Gaussian distribution [GKLW18]. We remark that the results of [GKLW18] were independently and concurrently developed, and utilize substantially different techniques than ours that rely crucially on the linearity of the ReLU for $x > 0$ [Ge18]. For these reasons, their algorithms do not generalize to the larger class of rectified functions which are handled in this work. To the best of the our knowledge, for the case of Gaussian inputs, this work and [GKLW18] are the first to obtain polynomial time learning algorithms for this noisy setting.

Agnostic Learning. A variety of works study learning ReLU's in the more general *agnostic* learning setting, based off Valiant's original PAC learning model [Val84]. The agnostic PAC model allows for arbitrary noisy and distributions over observations, and the goal is to output a hypothesis function which approximates the output of the neural network. Note that this does not necessarily entail learning the weights of an underlying network. For instance, Arora et al. [ABMM16] gives an algorithm with $O(n^d)$ running time to minimize the empirical risk of a two-layer neural network. A closer analysis of the generalization bounds required in this algorithm for PAC learning is given in [MR18], which gives a $2^{\text{poly}(k/\varepsilon)}\text{poly}(n, m, d, k)$ time algorithm under the constraints that $\boldsymbol{U}^* \in \{1, -1\}^k$ is given a fixed input, and both the input examples $\mathcal{X}$ and the weights $\boldsymbol{V}^*$ are restricted to being in the unit ball. In contrast, our $(\kappa/\varepsilon)^{O(k^2)}$ time algorithm for general error matrices $\mathbf{E}$ improves on their complexity whenever $\kappa = O(2^{\text{poly}(k)})$, and moreover can handle arbitrarily large $\boldsymbol{V}^*$ and unknown $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$. We remark, however, that our loss function is different from that of the PAC model, and is in fact roughly equivalent to the empirical loss considered in [ABMM16].

Note that the above algorithms *properly* learn the networks. That is, they actually output weight matrices $\boldsymbol{U}, \boldsymbol{V}$ such that $\boldsymbol{U} f(\boldsymbol{V}\mathcal{X})$ approximates the data well under some measure. A relaxation if this setting is *improper* learning, where the output of the learning algorithm can be any efficiently computable function, and not necessarily the weights of neural network. Several works have been studied that achieve polynomial running times under varying assumptions

about the network parameters, such as [GKKT16, GK17]. The algorithm of [GK17], returns a "clipped" polynomial. In addition, [ZLJ16] gives polynomial time improper learning algorithms for multi-layer neural networks under several assumptions on the weights and activation functions.

Hardness. Hardness results for learning networks have an extensive history in the literature [Jud88, BR92]. Originally, hardness was considered for threshold activation functions $f(x) \in \{1, -1\}$, where it is known that even for two ReLU's the problem is NP-hard [BR92]. Very recently, there have been several concurrent and independent lower bounds developed for learning ReLU networks. The work of [BDL18] has demonstrated the hardness of a neural network with the same number of nodes as the hard network in this paper, albeit with two applications of ReLU's (i.e., two non-linear layers) instead of one. Note that the hardness results of this work hold for even a single non-linear layer. Also concurrently and independently, a recent result of [MR18] appears to demonstrate the same NP-hardness as that in this paper, albiet using a slightly different reduction. The results of [MR18] also demonstrate that *approximately* learning even a single ReLU is NP-hard. In addition, there are also NP-hardness results with respects to improper learning of ReLU networks [GKKT16, LSSS14, ZLJ16] under certain complexity theoretic assumptions.

Sparsity. One of the main techniques of our work involves analyzing the sparsity patterns of the vectors in the rowspan of $\mathbf{A}$. Somewhat related reasoning has been applied by Spielman, Wang, and Wright to the dictionary learning problem [SWW12]. Here, given a matrix $\mathbf{A}$, the problem is to recover matrices $\mathbf{B}, \mathcal{X}$ such that $\mathbf{A} = \mathbf{B}\mathcal{X}$, where $\mathcal{X}$ is sparse. They argue the uniqueness of such a factorization by proving that, under certain conditions, the sparsest vectors in the row span of $\mathbf{A}$ are the precisely rows of $\mathcal{X}$. This informs their later algorithm for the exact recovery of these sparse vectors using linear programming.

6.1.3 Our Techniques

One of the primary technical contributions of this work is the utilization of the combinatorial structure of sparsity patterns of the rows of $f(\boldsymbol{V}\mathcal{X})$, where f is a rectified function, to solve learning problems. Here, a sparsity pattern refers to the subset of coordinates of $f(\boldsymbol{V}\mathcal{X})$ which are non-zero, and a rectified function f is one which satisfies $f(x) = 0$ for $x \leqslant 0$, and $f(x) > 0$ otherwise.

Arbitrary Input. For instance, given $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ where $\boldsymbol{U}^*, \boldsymbol{V}^*$ are full rank and f is the ReLU, one approach to recovering the weights is to find k-linearly vectors v_i such that $f(v_i \mathcal{X})$ span precisely the rows of $\mathbf{A}$. Without the function $f(\cdot)$, one could accomplish this by solving a linear system. Of course, the non-linearity of the activation function complicates matters significantly. Observe, however, that if the sparsity pattern of $f(\boldsymbol{V}^* \mathcal{X})$ was known before hand, one could simple *remove* f on the coordinates where $f(\boldsymbol{V}^* \mathcal{X})$ is non-zero, and solve the linear system here. On all other coordinates, one knows that $f(\boldsymbol{V}^* \mathcal{X})$ is 0, and thus finding a linearly independent vector in the right row span can be solved with a linear system. Of course, naively one would need to iterate over 2^n possible sparsity patterns before finding the correct one. However, one can show that any k-dimensional subspace of $\mathcal{R}^n$ can intersect at most n^k orthants of $\mathcal{R}^n$, and moreover these orthants can be enumerated in $n^k \text{poly}(n)$ time given the subspace. Thus the rowspan of $\mathbf{A}$, being k-dimensional, can contain vectors with at most n^k patterns. This is the primary obervation behind our $n^k \text{poly}(n)$-time algorithm for exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$ in the noiseless case (for arbitrary $\mathcal{X}$).

As mentioned before, the prior result requires $\mathbf{A}$ to be rank-k, otherwise the row span of $f(\boldsymbol{V} \mathcal{X})$ cannot be recovered from the row span of $\mathbf{A}$. We show that this difficulty is not merely a product of our specific algorithm, by demonstrating that even for k as small as 2, if $\boldsymbol{U}^*$ is given as input then it is NP-hard to find $\boldsymbol{V}^*$ such that $\boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) = \mathbf{A}$, thus ruling out any general n^k time algorithm for the problem. For the case of $k = 2$, the problem is still NP-hard even when $\boldsymbol{U}^*$ is not given as input, and is a variable.

Gaussian Input. In response to the aformentioned hardness results, we relax to the case where the input $\mathcal{X}$ has Gaussian marginals. In the noiseless case, we *exactly* learn the weights $\boldsymbol{U}^*, \boldsymbol{V}^*$ given $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ (up to a positive scaling and permutation). As mentioned, our results utilize analysis of the sparsity patterns in the row-span of $\mathbf{A}$. One benefit of these techniques is that they are largely insensitive to the behavior of $f(x)$ for positive x, and instead rely on the rectified property $f(\cdot)$. Hence, this can include even exponential functions, and not solely the ReLU.

Our exact recovery algorithms proceed in two steps. First, we obtain an approximate version of the matrix $f(\boldsymbol{V}^* \mathcal{X})$. For a good enough approximation, we can exactly recover the sparsity pattern of $f(\boldsymbol{V}^* \mathcal{X})$. Our main insight is, roughly, that the only sparse vectors in the row span of $\mathbf{A}$ are precisely the rows of $f(\boldsymbol{V}^* \mathcal{X})$. Specifically, we show that the only vectors in the row span which have the same sparsity pattern as a row of $f(\boldsymbol{V}^* \mathcal{X})$ are scalar multiples of that row. Moreover, we show that no vector in the row span of $\mathbf{A}$ is supported on a strict subset of the support of a given row of $f(\boldsymbol{V}^* \mathcal{X})$. Using these facts, we can then set up a judiciously designed

linear system to find these vectors, which allows us to recover $f(\boldsymbol{V}^*\mathcal{X})$ and then $\boldsymbol{V}^*$ exactly. By solving linear systems, we avoid using iterative continuous optimization methods, which recover a solution up to additive error ε and would only provide rates of convergence in terms of ε. In contrast, Gaussian elimination yields exact solutions in a polynomial number of arithmetic operations.

The first step, finding a good approximation of $f(\boldsymbol{V}^*\mathcal{X})$, can be approached from multiple angles. In this work, we demonstrate two different techniques to obtain these approximations, the first being Independent Component Analysis (ICA), and the second being tensor decomposition. To illustrate the robustness of our exact recovery procedure once a good estimate of $f(\boldsymbol{V}^*\mathcal{X})$ is known, we show in Section 6.4.3 how we can bootstrap the estimators of recent, concurrent and independent work [GKLW18], to improve them from approximate recovery to exact recovery.

Independent Component Analysis. In the restricted case when $\boldsymbol{V}^*$ is orthonormal, we show that our problem can be modeled as a special case of *Independent Component Analysis* (ICA). The ICA problem approximately recovers a subspace $\mathbf{B}$, given that the algorithm observes samples of the form $y = \mathbf{B}x + \zeta$, where x is i.i.d. and drawn from a distribution that has moments bounded away from Gaussians, and ζ is a Gaussian noise vector. Intuitively, the goal of ICA is to find a linear transformation of the data such that each of the coordinates or features are as independent as possible. By rotational invariance of Gaussians, in this case $\boldsymbol{V}^*\mathcal{X}$ is also i.i.d. Gaussian, and we know that the columns of $f(\boldsymbol{V}^*\mathcal{X})$ have independent components and moments bounded away from a Gaussian. Thus, in the orthonormal case, our problem is well suited for the ICA framework.

Tensor Decomposition. A second, more general approach to approximating $f(\boldsymbol{V}^*\mathcal{X})$ is to utilize techniques from *tensor decomposition*. Our starting point is the generative model considered by Janzamin et. al. [JSA15], which matches our setting, i.e., $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$. The main idea behind this algorithm is to construct a tensor that is a function of both $\mathbf{A}, \mathcal{X}$ and captures non-linear correlations between them. A key step is to show that the resulting tensor has low CP-rank and the low-rank components actually capture the rows of the weight matrix $\boldsymbol{V}^*$. Intuitively, working with higher order tensors is necessary since matrix decompositions are only identifiable up to orthogonal components, whereas tensors have identifiable non-orthogonal components, and we are specifically interested in recovering approximations for non-orthonormal $\boldsymbol{V}^*$.

Next, we run a tensor decomposition algorithm to recover the low-rank components of the resulting tensor. While computing a tensor decomposition is NP-hard in general [HL13], there is a plethora of work on special cases, where computing such decompositions is tractable [BCMV14,

SWZ16, WA16, GVX14, GM15, BM16]. Tensor decomposition algorithms have recently become an invaluable algorithmic primitive and with applications in statistical and machine learning [JSA15, JSA14, GLM17, AGHK14a, BKS15].

However, there are several technical hurdles involved in utilizing tensor decompositions to obtain estimates of $\boldsymbol{V}^*$. The first is that standard analysis of these methods utilizes a generalized version of *Stein's Lemma* to compute the expected value of the tensor, which relies on the smoothness of the activation function. Thus, we first approximate $f(\cdot)$ closely using a Chebyshev polynomial $p(\cdot)$ on a sufficiently large domain. However, we cannot algorithmically manipulate the input to demand that $\mathbf{A}$ instead be generated as $\boldsymbol{U}^* p(\boldsymbol{V}^* \mathcal{X})$. Instead, we add a small mean-zero Gaussian perturbation to our samples and analyze the variation distance between $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \boldsymbol{G}$ and $\boldsymbol{U}^* p(\boldsymbol{V}^* \mathcal{X}) + \boldsymbol{G}$. For a good enough approximation p, this variation distance will be too small for any algorithm to distinguish between them, thus standard arguments imply the success of tensor decomposition algorithms when given the inputs $\mathbf{A} + \boldsymbol{G}$ and $\mathcal{X}$.

Next, a key step is to construct a non-linear transformation of the input by utilizing knowledge about the underlying density function for the distribution of $\mathcal{X}$, which we denote by $p(x)$. The non-linear function considered is the so-called Score Function, defined in [JSA14], which is the normalized m-th order derivative of the input probability distribution function $p(x)$. Computing the score function for an arbitrary distribution can be computationally challenging. However, as mentioned in [JSA14], we can use Hermite polynomials that help us compute a closed form for the score function, in the special case when $x \sim \mathcal{N}(0, \mathbf{I})$.

Sign Ambiguity. A further complication arises due to the fact that this form of tensor decomposition is agnostic to the signs of $\boldsymbol{V}$. Namely, we are guaranteed vectors v_i from tensor decomposition such that $\|v_i - \xi_i \boldsymbol{V}^*_{i,*}\|_F < \varepsilon$, where $\xi_i \in \{1, -1\}$ is some unknown sign. Prior works have dealt with this issue by considering restricted classes of smooth activation functions which satisfy $f(x) = 1 - f(-x)$ [JSA15]. For such functions, one can compensate for not knowing the signs by allowing for an additional affine transformation in the neural network. Since we consider non-affine networks and rectified functions $f(\cdot)$ which do not satisfy this restriction, we must develop new methods to recover the signs ξ_i to avoid the exponential blow-up needed to simply guess them.

For the noiseless case, if v_i is close enough to $\xi_i \boldsymbol{V}^*_{i,*}$, we can employ our previous results on the uniqueness of sparsity patterns in the row-span of $\mathbf{A}$. Namely, we can show that the sparsity pattern of $f(\xi v_i)$ will in fact be feasible in the row-span of $\mathbf{A}$, whereas the sparsity pattern of

$f(-\xi v_i)$ will not, from which we recover the signs ξ_i via a linear system.

In the presence of noise, however, the problem becomes substantially more complicated. Because we do not have the true row-span of $f(\boldsymbol{V}^*\mathcal{X})$, but instead a noisy row-span given by $\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X}) + \mathbf{E}$, we cannot recover the ξ_i's by feasibility arguments involving sparsity patterns. Our solution to the sign ambiguity in the noisy case is a projection-based scheme. Our scheme for determining ξ_i involves constructing a $2k-2$ dimensional subspace S, spanned by vectors of the form $f(\pm v_j\mathcal{X})$ for all $j \neq i$. We augment this subspace as $S^1 = S \cup \{f(v_i\mathcal{X})\}$ and $S^{-1} = S \cup \{f(-v_i\mathcal{X})\}$. We then claim that the length of the projections of the rows of $\mathbf{A}$ onto the S^ξ will be *smaller* for $\xi = \xi_i$ than for $\xi = -\xi_i$. Thus by averaging the projections of the rows of $\mathbf{A}$ onto these subspaces and finding the subspace which has the smaller projection length on average, we can recover the ξ_i's with high probability. Our analysis involves bounds on projections onto perturbed subspaces, and a spectral analysis of the matrices $f(\boldsymbol{W}\mathcal{X})$, where $\boldsymbol{W}$ is composed of up to $2k$ rows of the form $\boldsymbol{V}^*_{i,*}$ and $-\boldsymbol{V}^*_{i,*}$.

FPT Algorithms. In addition to our polynomial time algorithms, we also demonstrate how various seemingly intractable relaxations to our model, within the Gaussian input setting, can be solved in fixed-parameter tractable time in the number k of hidden units, and the condition numbers κ of $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$. Our first result demonstrates that, in the noiseless case, exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$ is still possible even when $\boldsymbol{U}^*$ is not rank k. Note that the assumption that $\boldsymbol{U}^*$ is rank k is required in many other works on learning neural networks [GLM17, GKLW18, JSA15, SJA16]

We demonstrate that taking $\text{poly}(d)\kappa^{O(k)}$ columns of $\mathcal{X}$, where κ is the condition number of $\boldsymbol{V}^*$, is sufficient to obtain 1-sparse vectors in the columns of $f(\boldsymbol{V}^*\mathcal{X})$. As a result, we can look for column of $\mathbf{A}$ which are positive scalar multiples of each other, and conclude that any such pair will indeed be a positive scaling of a column of $\boldsymbol{U}^*$ with probability 1. This allows for exact recovery of $\boldsymbol{U}^*$ for any $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}$ and $m \geqslant 1$, as long as no two columns of $\boldsymbol{U}^*$ are positive scalar multiples of each other. Thereafter, we can recover $\boldsymbol{V}^*$ by solving a linear system on the subset of 1-sparse columns of $f(\boldsymbol{V}\mathcal{X})$, and argue that the resulting constraint matrix is full rank. The result is a $\text{poly}(d,k,m)\kappa^O(k)$ time algorithm for exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$.

Our second FPT result involves a substantial generalization of the class of error matrices $\mathbf{E}$ which we can handle. In fact, we allow arbitrary $\mathbf{E}$, so long as they are independent of the input $\mathcal{X}$. Our primary technical observation is as follows. Suppose that we were given $f(v\mathcal{X}) + \mathbf{E}$, where $\mathbf{E}$ is an arbitrary, possibly very large, error vector, and $v \in \mathcal{R}^d$. Then one can look at the sign of each entry i, and consider it to be a noisy observation of which side of a halfspace the vector $\mathcal{X}_{*,i}$ lies within. In other words, we couch the problem as a noisy half-space learning

problem, where the half-space is given by the hyperplane normal to v, and the labeling of $\mathcal{X}_{*,i}$ is the sign of $(f(v\mathcal{X}) + \mathbf{E})_i$.

Now while the error on each entry will be large, resulting in nearly half of the labelings being flipped incorrectly, because $\mathbf{E}$ is *independent* of $\mathcal{X}$, we are able to adapt recent techniques in noisy-halfspace learning to recover v in polynomial time. In order to utilize these techniques without knowing anything about $\mathbf{E}$, we must first *smooth out* the error $\mathbf{E}$ by adding a large Gaussian matrix. The comparatively small value of $f(v\mathcal{X})$ is then able to shift the observed distribution of signs sufficiently to have non-trivial correlation with the true signs. Taking polynomially many samples, our algorithms detect this correlation, which will allow for accurate recovery of v.

To even obtain a matrix of the form $f(v\mathcal{X}) + \mathbf{E}$, where v is a row of $\boldsymbol{V}^*$, we can guess the pseudo-inverse of $\boldsymbol{U}^*$. To reduce the dependency on m, we first sketch $\boldsymbol{U}^*$ by a *subspace-embedding* $\mathcal{S} \in \mathcal{R}^{O(k)\times d}$, which will be a random Gaussian matrix and approximately preserve the column span of $\boldsymbol{U}^*$. In particular, this approximately preserves the spectrum of $\boldsymbol{U}^*$. The resulting matrix $\mathcal{S}\boldsymbol{U}^*$ has $O(k^2)$ entries, and, given the maximum singular value of the inverse (which can be guessed to a factor of 2), can be guessed accurately enough for our purposes in time $(\kappa/\varepsilon)^{O(k^2)}$, which dominates the overall runtime of the algorithm.

6.1.4 Roadmap

In Section 6.2 we introduce our $n^{O(k)}$ time exact algorithm when $\text{rank}(A) = k$ and arbitrary $\mathcal{X}$, for recovery of rank-k matrices $\boldsymbol{U}^*, \boldsymbol{V}^*$ such that $\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X}) = \mathbf{A}$. In this section, we also demonstrate that for a very wide class of distributions for *random* matices $\mathcal{X}$, the matrix $\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ is in fact full rank with high probability, and therefore can be solved with our exact algorithm. Then, in Section 6.3, we prove NP-hardness of the learning problem when $\text{rank}(A) < k$. Next, in Section 6.4, we give a polynomial time algorithm for exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$ in the case when $\mathcal{X}$ has Gaussian marginals in the realizable setting. Section 6.4.1 develops our Independenct Component Analysis Based algorithm, whereas Section 6.4.2 develops our more general exact recovery algorithm. In Section 6.4.3, we show how recent concurrent results can be bootstrapped via our technqiues to obtain exact recovery for a wider class of distributions.

In Section 6.5, we demonstrate how to extend our algorithm to the case where $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X}) + \mathbf{E}$ where $\mathbf{E}$ is mean 0 i.i.d. sub-Gaussian noise. Then in Section 6.6, we give a fixed-paramater tractable (FPT) (in k and $\kappa(\boldsymbol{V}^*)$) for the exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$ in the case where $\boldsymbol{U}^*$ does not have full column rank. We give our second FPT algorithm in Section 6.7, which finds weights

which approximate the optimal network for arbitrary error matrices $\mathbf{E}$ that are independent of $\mathcal{X}$. In Section 6.8, we demonstrate how the weights of certain *low-rank* networks, where $k < d, m$, can be recovered exactly in the presence of a class of arbitrary sparse noise in polynomial time.

6.1.5 Preliminaries

For a positive integer k, we write $[k]$ to denote the set $\{1, 2, \ldots, k\}$. We use the term *with high probability* (w.h.p.) in a parameter $r > 1$ to describe an event that occurs with probability $1 - \frac{1}{\text{poly}(r)}$. For a real r, we will often use the shorthand $\text{poly}(r)$ to denote a sufficiently large constant degree polynomial in r. Since for simplicity we do not seek to analyze or optimize the polynomial running time of our algorithms, we will state many of our error bounds within technical lemmas as $\frac{1}{\text{poly}(r)}$ where r constitutes some set of relevant parameters, with the understanding that this polynomial can be made arbitrarily large by increasing the sample complexity n of our algorithms by a polynomial factor.

In this work we use boldface font $\mathbf{A}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{W}$ to denote matrices, and non-boldface font x, y, u, v to denote vectors. For a vector x, we use $\|x\|_2$ to denote the ℓ_2 norm of x. For any matrix $\boldsymbol{W}$ with p rows and q columns, for all $i \in [p]$, let $\boldsymbol{W}_{i,*}$ denote the i-th row of $\boldsymbol{W}$, for all $j \in [q]$ let $\boldsymbol{W}_{*,j}$ denote the j-th column and let $\boldsymbol{W}_{i,j}$ denote the i, j-th entry of $\boldsymbol{W}$. Further, the singular value decomposition of $\boldsymbol{W}$, denoted by $\texttt{SVD}(\boldsymbol{W}) = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, is such that $\boldsymbol{U}$ is a $p \times r$ matrix with orthonormal columns, $\boldsymbol{V}^T$ is a $r \times q$ matrix with orthonormal rows and $\boldsymbol{\Sigma}$ is an $r \times r$ diagonal matrix, where r is the rank of $\boldsymbol{W}$. The entries along the diagonal are the singular values of $\boldsymbol{W}$, denoted by $\sigma_{\max} = \sigma_1(\boldsymbol{W}) \geqslant \sigma_2(\boldsymbol{W}) \geqslant \ldots \geqslant \sigma_r(\boldsymbol{W}) = \sigma_{\min}(\boldsymbol{W})$. We write $\|\boldsymbol{W}\|_F = \left(\sum_{p,q} \boldsymbol{W}_{p,q}^2\right)^{1/2}$ to denote the Frobenius norm of $\boldsymbol{W}$, and

$$\|\boldsymbol{W}\|_2 = \sup_x \frac{\|\mathbf{A}x\|_2}{\|x\|_2} = \sigma_{\max}(\boldsymbol{W})$$

to denote the spectral norm. We will write $\mathbb{I}_k$ to denote the $k \times k$ square identity matrix. We use the notation $\texttt{Proj}_{\boldsymbol{W}}(w)$ to denote the projection of the vector w onto the *row-span* of $\boldsymbol{W}$. In other words, if $x^* = \arg\min_x \|x\boldsymbol{W} - w\|_2$, then $\texttt{Proj}_{\boldsymbol{W}}(w) = x^*\boldsymbol{W}$. We now recall the condition number of a matrix $\boldsymbol{W}$.

Definition 6.1.1. *For a rank k matrix $\boldsymbol{W} \in \mathcal{R}^{p \times q}$, let $\sigma_{\max}(\boldsymbol{W}) = \sigma_1(\boldsymbol{W}) \geqslant \sigma_2(\boldsymbol{W}) \geqslant \ldots \geqslant \sigma_k(\boldsymbol{W}) = \sigma_{\min}(\boldsymbol{W})$ be the non-zero singular values of $\boldsymbol{W}$. Then the condition number $\kappa(\boldsymbol{W})$*

of $\boldsymbol{W}$ is given by

$$\kappa(\boldsymbol{W}) = \frac{\sigma_{\max}(\boldsymbol{W})}{\sigma_{\min}(\boldsymbol{W})}$$

Note that if $\boldsymbol{W}$ has full column rank (i.e., $k = q$), then if $\boldsymbol{W}^\dagger$ is the pseudo-inverse of $\boldsymbol{W}$ we have $\boldsymbol{W}^\dagger \boldsymbol{W} = \mathbb{I}_q$ and

$$\kappa(\boldsymbol{W}) = \|\boldsymbol{W}^\dagger\|_2 \|\boldsymbol{W}\|_2$$

where $\|\boldsymbol{W}\|_2 = \sigma_1(\boldsymbol{W})$ is the spectral norm of $\boldsymbol{W}$. Similarly if $\boldsymbol{W}$ has full row rank (i.e. $k = p$), then $\boldsymbol{W}\boldsymbol{W}^\dagger = \mathbb{I}_p$ and

$$\kappa(\boldsymbol{W}) = \|\boldsymbol{W}^\dagger\|_2 \|\boldsymbol{W}\|_2$$

A real m-th order tensor is $\mathcal{T} \in \otimes^m \mathcal{R}^d$ is the outer product of m d-dimensional Euclidean spaces. A third order tensor $\mathcal{T} \in \otimes \mathcal{R}^d$ is defined to be rank-1 if $\mathcal{T} = w \cdot a \otimes b \otimes c$ where $a, b, c \in \mathcal{R}^d$. Further, $\mathcal{T}$ has Candecomp/Parafac (CP) rank-k if it can be written as the sum of k rank-1 tensors, i.e.,

$$\mathcal{T} = \sum_{i=1}^{k} w_i a_i \otimes b_i \otimes c_i$$

is such that $w_i \in \mathcal{R}, a_i, b_i, c_i \in \mathcal{R}^d$. Next, given a function $f(x) : \mathcal{R}^d \to \mathcal{R}$, we use the notation $\nabla_x^m f(x) \in \otimes^m \mathcal{R}^d$ to denote the m-th order derivative operator w.r.t. the variable x, such that

$$[\nabla_x^m f(x)]_{i_1, i_2, \ldots i_m} = \frac{\partial f(x)}{\partial x_{i_1} \partial x_{i_2} \ldots \partial x_{i_m}}$$

.

In the context of the ReLU activation function, a useful notion to consider is that of a sign pattern, which will be used frequently in our analysis.

Definition 6.1.2. *For any matrix dimensions p, q, a sign pattern is simply a subset of $[p] \times [q]$. For a matrix $\boldsymbol{W} \in \mathcal{R}^{p \times q}$, we let $\texttt{sign}(\boldsymbol{W})$ be the sign pattern defined by*

$$\texttt{sign}(\boldsymbol{W}) = \{(i, j) \in [p] \times [q] \mid \boldsymbol{W}_{i,j} > 0\}$$

Intuitively, in the context of rectified activation functions, the sign pattern is an important notion since $\texttt{sign}(\boldsymbol{W})$ is invariant under application of f, in other words $\texttt{sign}(\boldsymbol{W}) = f(\texttt{sign}(\boldsymbol{W}))$. We similarly define a *sparsity-pattern* of a matrix $\boldsymbol{W} \in \mathcal{R}^{p \times q}$ as a subset of $[p] \times [q]$ where $\boldsymbol{W}$ is non-zero. Note that a sign and sparsity pattern of $\boldsymbol{W}$, taken together, specify precisely where the strictly positive, negative, and zero-valued entries are in $\boldsymbol{W}$.

We use the notation $\mathcal{N}(\mu, \sigma^2)$ to denote the Gaussian distribution with mean μ and variance σ^2. More generally, we write $\mathcal{N}(\mu, \mathbf{\Sigma})$ to denote a k-dimensional multi-variate Gaussian distribution with mean $\mu \in \mathcal{R}^k$ and variance $\mathbf{\Sigma} \in \mathcal{R}^{k \times k}$. We make use of the 2-stability of the Gaussian distribution several times in this work, so we now recall the following definition of stable random variables. We refer the reader to [Ind06] for a further discussion of such distributions.

Definition 6.1.3. *A distribution $\mathcal{D}_p$ is said to be p-stable if whenever $\mathcal{X}_1, \ldots, \mathcal{X}_n \sim \mathcal{D}_p$ are drawn independently, we have*

$$\sum_{i=1}^{n} a_i \mathcal{X}_i \sim \|a\|_p \mathcal{X}$$

for any fixed vector $a \in \mathcal{R}^n$, where $\mathcal{X} \sim \mathcal{D}_p$ is again distributed as a p-stable random variable. In particular, the Gaussian random variables $\mathcal{N}(0, \sigma^2)$ are p-stable for $p = 2$ (i.e., $\sum_i a_i g_i = \|a\|_2$, where $g, g_1, \ldots, g_n \sim \mathcal{N}(0, \sigma^2)$).

Finally, we remark that in this paper, we will work in the common real RAM model of computation, where arithmetic operations on real numbers can be performed in constant time.

6.2 Exact solution when rank$(\mathbf{A}) = k$

In this section, we consider the exact case of the neural network recovery problem. Given an input matrix $\mathcal{X} \in \mathcal{R}^{d \times n}$ of examples, and a matrix $\mathbf{A} \in \mathcal{R}^{m \times n}$ of classifications, the exact version of the recovery problem is to obtain rank-k matrices $\boldsymbol{U}^*, \boldsymbol{V}^*$ such that $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$, if such matrices exist. In this section we demonstrate the existence of an $n^{O(k)}$poly(md)-time algorithm for exact recovery when rank$(\mathbf{A}) = k$. We demonstrate that this assumption is likely necessary in Section 6.3, where we show that if rank$(\mathbf{A}) < k$ then the problem is NP-hard even for any $k \geqslant 2$ when the matrix $\boldsymbol{U}$ is given as input, and NP-hard for $k = 2$ when $\boldsymbol{U}^*$ is allowed to be a variable. This rules out the existence of a general $n^{O(k)}$ time algorithm for this problem.

The main theorem we prove in this section is that there is an algorithm with running time dominated by $\min\{n^{O(k)}, 2^n\}$ such that it recovers the underlying matrices $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ exactly. Intuitively, we begin by showing a structural result that there are at most $n^{O(k)}$ sign patterns that lie in the row space of $f(\boldsymbol{V}^* \mathcal{X})$ and we can efficiently enumerate over them using a linear program. For a fixed sign pattern in this set, we construct a sequence of k linear programs (LP) such that the i-th LP finds a vector y^i, $f(y^i)$ is in the row span of $f(\boldsymbol{V}^* \mathcal{X})$, subject to the fixed sign pattern, and the constraint that $f(y^i)$ is not a linear combination of $f(y^1), f(y^2), \ldots f(y^{i-1})$. We note that $f(y^i)$ being linearly independent is not a linear constraint, but we demonstrate how

it can be linearized in a straightforward manner.

Crucially, our algorithm relies on the fact that we have the row-span of $f(\boldsymbol{V}^*\mathcal{X})$. Note that this is implied by the assumption that $\mathbf{A}$ is rank k. Knowing the rowspan allows us to design the constraints in the prior paragraph, and thus solve the LP to recover the rows of $f(\boldsymbol{V}^*\mathcal{X})$. On the other hand, if the rank of $\mathbf{A}$ is less than k, then it no longer seems possible to efficiently determine the row span of $f(\boldsymbol{V}^*\mathcal{X})$. In fact, our NP-Hardness result of Section 6.3 demonstrates that, given $\boldsymbol{U}^*$ as input, if the rank of $\mathbf{A}$ is strictly less than k, the problem of determining the exact row-span of $f(\boldsymbol{V}^*\mathcal{X})$ is NP-Hard. The main result of this section is then as follows.

Theorem 123. *Given $\mathbf{A} \in \mathcal{R}^{m\times n}, \mathcal{X} \in \mathcal{R}^{d\times n}$, there is an algorithm that finds $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ such that $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ and runs in time poly$(nmd)\min\{n^{O(k)}, 2^n\}$, if rank$(\mathbf{A}) = k$.*

Let $\boldsymbol{V}' \in \mathcal{R}^{k\times n}$ be a basis for the row-span of $\mathbf{A}$. For two matrices $\mathbf{Y}, \mathbf{Z}$ of the same dimension, we will write $\mathbf{Y} \overset{\text{row}}{\simeq} \mathbf{Z}$ if the row spans of $\mathbf{Y}$ and $\mathbf{Z}$ are the same. The first step in our algorithm is to obtain a *feasible set* $\mathcal{S}$ of sign patterns, within which the true sign pattern of $f(\boldsymbol{V}^*\mathcal{X})$ must lie.

Lemma 6.2.1. *Given $\mathbf{A} \in \mathcal{R}^{m\times n}, \mathcal{X} \in \mathcal{R}^{d\times n}$, such that rank$(\mathbf{A}) = k$, there is an algorithm which runs in time $\min\{n^{O(k)}, 2^n\}$ and returns a set of sign patterns $\mathcal{S} \subset 2^{[m]\times[n]}$ with $|\mathcal{S}| = \min\{n^{O(k)}, 2^n\}$ such that for any rank-k matrices $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ such that $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ and any row $i \in [k]$, $\texttt{sign}((\boldsymbol{V}^*\mathcal{X})_i) = \texttt{sign}(S)$ for some $S \in \mathcal{S}$.*

Proof. Recall, $\mathbf{A}$ is rank k. Thus there is a subset $\boldsymbol{V}' \in \mathcal{R}^{k\times n}$ of k rows of $\mathbf{A}$ which span all the rows of $\mathbf{A}$. Critically, here we require that the rank of $\mathbf{A}$ is k and thus the row space of $\mathbf{A}$ is the same as that of $f(\boldsymbol{V}^*\mathcal{X})$. Since $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ and $\boldsymbol{V}', f(\boldsymbol{V}^*\mathcal{X})$ have the same dimensional row space, the row spaces of $\boldsymbol{V}'$ and $f(\boldsymbol{V}^*\mathcal{X})$ are precisely the same, and so there must be an invertible change of basis matrix $\boldsymbol{W}$ such that $\boldsymbol{W}\boldsymbol{V}' = f(\boldsymbol{V}^*\mathcal{X})$. Now note that $\texttt{sign}(\boldsymbol{V}^*\mathcal{X}) = \texttt{sign}(f(\boldsymbol{V}^*\mathcal{X})) = \texttt{sign}(\boldsymbol{W}\boldsymbol{V}')$, and thus it suffices to return a set of sign patterns $\mathcal{S}$ which contains $\texttt{sign}(\boldsymbol{W}\boldsymbol{V}')$. Therefore, consider any fixed sign pattern $S \subset [n]$, and fix a row $j \in [k]$, and consider the following feasibility linear program in the variables w_j

$$(w_j\boldsymbol{V}')_i \geqslant 1, \qquad \text{for all } i \in \texttt{sign}(S)$$

$$(w_j\boldsymbol{V}')_i \leqslant 0, \qquad \text{for all } i \notin \texttt{sign}(S)$$

Note that if the sign pattern S is feasible by some $w_j\boldsymbol{V}'$, then the above LP will be feasible with a suitably large positive scaling to w_j. Now the LP has k variables and n constraints, and thus a

solution is obtained by choosing the w_j that makes a subset of k linearly independent constraints tight. Observe in any such LP of the above form, there are at most $2n$ possible constraints that can ever occur. Thus if S is realizable as the sign pattern of some $w_j \boldsymbol{V}'$, then it is obtained by the unique solution to a system which chooses to make k of these constraints tight. Formally, if S, b are the constraints for which $w_j S \geqslant b$ in the LP, then a solution is given by $w_j S' = b'$ where S', b' are a subset of k of the constraints. Since there are at most $\binom{2n}{k} = O(n^k)$ such possible choices, it follows that there are at most $O(\min\{n^{O(k)}, 2^n\})$ realizable sign patterns, and these can be enumerated in $O(\min\{n^{O(k)}, 2^n\})$ time by simply checking the sign pattern which results from the solution (if one exists) to $w_j S' = b'$ taken over all subsets S', b' of constraints of size k.

□

Given access to the set of candidate sign patterns, $S \in \mathcal{S}$, and vectors $y^1, y^2, ..., y^{i-1} \in \mathcal{R}^n$, we can define the following iterative feasibility linear program, that at each iteration i finds a vector y^i which is equal to some vector in the row span of $\mathcal{X}$, and such that $f(y^1), f(y^2), \ldots, f(y^i)$ are all linearly independent and in the row span of $\mathbf{A}$.

Algorithm 1 : Iterative LP$(\mathcal{X}, S, y^1, y^2, \dots y^{i-1})$**.**

`Input`: Matrix $\mathcal{X}$, a sign pattern S, vectors $y^1, y^2, \dots y^{i-1}$ such that $f(y^1), f(y^2), \dots f(y^{i-1})$ are linearly independent.

1. Let y^i, z^i, w^i be variables in $\mathcal{R}^n$.
2. Let $\mathbf{Q} \in \mathcal{R}^{(i-1)\times n}$ be a matrix such that for all $j \in [i-1]$, $\mathbf{Q}_{j,*} = f(y^j)$. Construct the projection matrix $\mathbf{P}^{i-1}$ onto span $\{f(y^1), f(y^2), ..., f(y^{i-1})\}$. Note, the projection matrix is given by $\mathbf{P}^{i-1} = \mathbf{Q}^T(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}$.
3. Define $f_S(y^i)$ w.r.t. the sign pattern S such that

$$f_S(y^i_j) = \begin{cases} (y^i_j) & \text{if } j \in S \\ 0 & \text{otherwise} \end{cases}$$

`Output`: A feasible solution to the following LP:

$$\begin{array}{lll} \forall j \in [n] & y^i_j \geqslant 1, & \text{if } j \in S \\ \forall j \in [n] & y^i_j \leqslant 0, & \text{if } j \notin S \\ & y^i = w^i\mathcal{X} & \\ & f_S(y^i) = z^i\boldsymbol{V}' & \\ & f_S(y^i)(I - \mathbf{P}^{i-1}) \neq 0 & \end{array}$$

Remark 124. Observe, while the last constraint is not a linear constraint, it can be made linear by running $2n$ consecutive LP's, such that, for $t \in [n]$, in the $2t$-th LP we replace the constraint $f_S(y^i)(\mathbb{I} - \mathbf{P}^{i-1}) \neq 0$ above with

$$\left[f_S(y^i)\left(\mathbb{I} - \mathbf{P}^{i-1}\right)\right]_t \geqslant 1$$

and in the $(2t-1)$-th LP we replace constraint $f_S(y^i)\left(\mathbb{I} - \mathbf{P}^{i-1}\right) \neq 0$ with

$$\left[f_S(y^i)\left(\mathbb{I} - \mathbf{P}^{i-1}\right)\right]_t \leqslant -1$$

Note, the modified constraints are linear in the variables y^i. If there is a vector y^i which satisfies the above constraints such that $f_S(y^i)(\mathbb{I}-\mathbf{P}^{i-1}) \neq 0$, then by scaling y^i, w^i, z^i all by a sufficiently large positive constant, then y^i will also satisfy one of the $2n$ LPs described above, thus giving a solution to the original feasibility problem by returning the first feasible solution returned among

the $2n$ new LPs.

Using Algorithm 1 as a sub-routine, we iterate over all sign patterns $S \in \mathcal{S}$, such that we recover a linearly independent set of k vectors $f(y^1), f(y^2), \ldots f(y^k)$. Let $\mathbf{Y}$ be a matrix such that the j-th row corresponds to y^j. We then set up and solve two linear systems in $\boldsymbol{U}$ and $\boldsymbol{V}$, given by $\mathbf{A} = \boldsymbol{U} f(\mathbf{Y})$ and $\mathbf{Y} = \boldsymbol{V}\mathcal{X}$. We show that the solutions to the linear system correspond to $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$. Here, we note that since the optimal $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ are solutions to a linear system, we can recover them exactly.

Algorithm 2 : ExactNeuralNet$(\mathbf{A}, \mathcal{X}, \mathcal{S})$**.**

`Input`: Matrices $\mathbf{A}, \mathcal{X}$, a set of sign patterns $\mathcal{S}$.

1. For $i = 1, 2, \ldots, k$

   topsep=0pt,1temsep=-1ex,p1rtopsep=1ex,p1rsep=1ex $t = 1$.

   topsep=0pt,2temsep=-2ex,p2rtopsep=2ex,p2rsep=2ex While$(t \leqslant |\mathcal{S}|)$

   i. If `Iterative LP`$(\mathcal{X}, S_t, y^1, y^2, ..., y^{i-1})$ is feasible, let y^i be the output, and set $t = |\mathcal{S}| + 1$.

   ii. Else $t \leftarrow t + 1$.

2. Let $\mathbf{Y} \in \mathcal{R}^{k \times n}$ be the matrix with j-th row equal to y^j and let S be the corresponding sign pattern.
3. Let $\boldsymbol{U}^*$ be the solution to the linear system in $\boldsymbol{U}$ given by $\mathbf{A} = \boldsymbol{U} f_S(\mathbf{Y})$.
4. Let $\boldsymbol{V}^*$ be the solution to the linear system in $\boldsymbol{V}$ given by $\mathbf{Y} = \boldsymbol{V}\mathcal{X}$.

`Output`**:** $\boldsymbol{U}^*, \boldsymbol{V}^*$.

Lemma 6.2.2. *For any $i \in [k]$ vectors $y^1, y^2, ..., y^{i-1} \in \mathcal{R}^n$ and $S \in \mathcal{S}$, let y^i be a feasible solution to* `Iterative LP`$(\mathcal{X}, S, y^1, y^2, ..., y^{i-1})$*. Then all of the following hold:*

1. *y^i is in the row span of $\mathcal{X}$.*
2. *$f(y^i)$ is in the row span of $\mathbf{A}$.*
3. *$f(y^i)$ is independent of $f(y^1), f(y^2), ..., f(y^{i-1})$.*

Proof. The first condition follows due to the third constraint $y^i = w^i\mathcal{X}$. The first and second

constraint ensure that $f_S(y^i) = f(y^i)$, thus along with the fourth constraint and the fact that $\boldsymbol{V}'$ spans the rows of $\mathbf{A}$, the second condition follows. For the last condition, it suffices to show that if $\|f(y^i)(\mathbb{I} - \mathbf{P}^{i-1})\| \geqslant 1$ then $f(y^i)$ is not in the span of $\{f(y^1), \ldots, f(y^{i-1})\}$. Now if $f(y^i)(\mathbb{I} - \mathbf{P}^{i-1}) = z \neq 0$, then $f(y^i) = z + \texttt{Proj}_{i-1}(f(y^i))$, where $\texttt{Proj}_{i-1}(f(y^i))$ is the projection of $f(y^i)$ onto the subspace spanned by $\{f(y^1), \ldots, f(y^{i-1})\}$. If $f(y^i)$ was in this subspace, then we would have $\texttt{Proj}_{i-1}(f(y^i)) = y^i$, but this is impossible since $z \neq 0$, which completes the proof. □

Lemma 6.2.3. *Suppose that there exist matrices $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ with $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$. Then in the above algorithm, for each $i \in [k]$ $\texttt{Iterative LP}(\mathcal{X}, S_t, y^1, y^2, ..., y^{i-1})$ will be feasible for at least one $S_t \in \mathcal{S}$.*

Proof. The proof is by induction. For $i = 1$, since $f(\boldsymbol{V}^*\mathcal{X})$ has rank k and spans the rows of $\mathbf{A}$, it follows that there must be some $j \in [k]$ such that the j-th row $f(\boldsymbol{V}^*\mathcal{X})_j$ of $f(\boldsymbol{V}\mathcal{X})$ is in in the row span of $\boldsymbol{V}'$, and clearly $(\boldsymbol{V}^*\mathcal{X})_j$ is in the row span of $\mathcal{X}$. The last constraint is of the LP non-existent since $i = 1$. Furthermore, $(\boldsymbol{V}^*\mathcal{X})_j$ has some sign pattern S^*, and it must be that $S^* \in \mathcal{S}$ by construction of $\mathcal{S}$. Then there exists a positive constant $c > 0$ such that $(c\boldsymbol{V}^*\mathcal{X})_j$ satisfies the last constraints of $\texttt{Iterative LP}(\mathcal{X}, S^*, y^1, y^2, ..., y^{i-1})$ (made linear as described in Remark 124), and multiplying $(\boldsymbol{V}^*\mathcal{X})_j$ by a positive constant does not affect the fact that $(c\boldsymbol{V}^*\mathcal{X})_j$ is in the row space of $\mathcal{X}$ and $f(c\boldsymbol{V}^*\mathcal{X})_j$ is in the row space of $\mathbf{A}$ by closure of subspaces under scalar multiplication. Thus the $\texttt{Iterative LP}(\mathcal{X}, S^*, y^1, y^2, ..., y^{i-1})$ has a feasible point.

Now suppose we have feasible points $y^1, \ldots, y^{i-1}$, with $i \leqslant k$. Note that this guarantees that $f(y^1), \ldots, f(y^{i-1})$ are linearly independent. Since $f(\boldsymbol{V}^*\mathcal{X})$ spans the k-dimensional row-space of $\mathbf{A}$, there must be some j with $f(\boldsymbol{V}^*\mathcal{X})_j$ that is linearly independent of $f(y^1), \ldots, f(y^{i-1})$ such that $f(\boldsymbol{V}^*\mathcal{X})_j$ is in the row span of $\mathbf{A}$. Then $(\boldsymbol{V}^*\mathcal{X})_j$ is in the row span of $\mathcal{X}$, and similarly $(\boldsymbol{V}\mathcal{X})_j$ has some sign pattern S^*, and after multiplication by a suitably large constant it follows that the $\texttt{Iterative LP}(\mathcal{X}, S^*, y^1, y^2, ..., y^{i-1})$ will be feasible. The proposition follows by induction. □

Proof of Theorem 123. By Proposition 6.2.2, $f(y^1), \ldots, f(y^k)$ are independent, and give a solution to $f(\boldsymbol{V}\mathcal{X}) \overset{\text{row}}{\simeq} \mathbf{A}$. Thus we can find a $\boldsymbol{U} \in \mathcal{R}^{d\times k}$ in polynomial time via d independent linear regression problems that solves $\boldsymbol{U} f(\boldsymbol{V}\mathcal{X}) = \mathbf{A}$. By Proposition 6.2.1, there are at most $\min\{n^{O(k)}, 2^n\}$ sign patterns in the set $\mathcal{S}$, and solving for each iteration of $\texttt{Iterative LP}$ takes poly(nm)-time. Thus the total time is poly$(nmd)\min\{n^{O(k)}, 2^n\}$ as stated.

6.2.1 Rank$(\mathrm{A}) = k$ for random matrices $\mathcal{X}$.

We conclude this section with the observation that if the input $\mathcal{X}$ is drawn from a large class of independent distributions, then the resulting matrix $\boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ will in fact be rank k with high probability if $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ are rank k. Therefore, Algorithm 2 recovers $\boldsymbol{U}^*, \boldsymbol{V}^*$ in poly$(nmd) \min\{n^{O(k)}, 2^n\}$ for all such input matrices $\mathcal{X}$.

Lemma 6.2.4. *Suppose* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ *for rank* k *matrices* $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$ *and* $\boldsymbol{V}^* \in \mathcal{R}^{k \times d}$*, where* $\mathcal{X} \in \mathcal{R}^{d \times n}$ *is a matrix of random variables such that each column* $\mathcal{X}_{*,i}$ *is drawn i.i.d. from a distribution* $\mathcal{D}$ *with continuous p.d.f.* $p(x) : \mathcal{R}^d \to \mathcal{R}$ *such that* $p(x) > 0$ *almost everywhere in* $\mathcal{R}^d$*, and such that*

$$\inf_{v \in \mathcal{R}^d} \boldsymbol{Pr}_{x \sim \mathcal{D}}\Big[\langle v, x \rangle > 0\Big] > 10k \log(k/\delta)/n$$

Then rank$(A) = k$ *with probability* $1 - O(\delta)$.

Proof. By Sylverster's rank inequality, it suffices to show $f(\boldsymbol{V}^* \mathcal{X})$ is rank k. By symmetry and i.i.d. of the $\mathcal{X}_{ij}$'s in a fixed row i, each entry $f(\boldsymbol{V}^* \mathcal{X})_{ij}$ is non-zero with probability at least $10k \log(k/\delta)/n$ independently (within the row i). Then by Chernoff bounds, a fixed row $(\boldsymbol{V}^* \mathcal{X})_{i,*}$ will have at least k positive entries with probability at least $1 - 2^{-k \log(k/\delta)}$, and we can then union bound over all k rows to hold with probability at least $1 - O(\delta)$. Thus one can pick a $k \times k$ submatrix $\boldsymbol{W}$ of $f(\boldsymbol{V}^* \mathcal{X})$ such that, under some permutation $\boldsymbol{W}'$ of the columns of $\boldsymbol{W}$, the diagonal of $\boldsymbol{W}'$ is non-zero.

Since $\boldsymbol{V}^*$ is rank k, $\boldsymbol{V}^*$ is a surjective linear mapping of the columns of $\mathcal{X}$ from $\mathcal{R}^d$ to $\mathcal{R}^k$. Since $p(x) > 0$ almost everywhere, it follows that $p_{\boldsymbol{V}^*}(x) > 0$ almost everywhere, where $p_{\boldsymbol{V}^*}(x)$ is the continuous pdf of a column of $\boldsymbol{V}^* \mathcal{X}$. Then if $\mathcal{X}'$ is any matrix of k columns of $\mathcal{X}$, by independence of the columns, if $p_{k \times k} : \mathcal{R}^{k^2} \to \mathcal{R}$ is the joint pdf of all k^2 variables in $\boldsymbol{V}^* \mathcal{X}'$, it follows that $p_{k \times k}(x) > 0$ for all $x \in \mathcal{R}^{k^2}$. Thus, by conditioning on any sign pattern S of $\boldsymbol{V}^* \mathcal{X}'$, this results in a new pdf $p^S_{k \times k}$, which is simply $p_{k \times k}$ where the domain is restricted to an orthant Ω of R^{k^2}. Since $p_{k \times k}$ is continuous and non-zero almost everywhere, it follows that the support of the pdf $p^S_{k \times k} : \Omega \to \mathcal{R}$ is all of Ω. In particular, the Lesbegue measure of the support Ω inside of $\mathcal{R}^{k^2}$ is non-zero (note that this would not be true if $\boldsymbol{V}^*$ has rank $k' < k$, as the support on each column would then be confined to a subspace of $\mathcal{R}^k$, which would have Lesbegue measure zero in $\mathcal{R}^k$).

Now after conditioning on a sign pattern, $\det(\boldsymbol{W}')$ is a non-zero polynomial in s random variables, for $k \leqslant s \leqslant k^2$, and it is well known that such a function cannot vanish on any non-empty open set in $\mathcal{R}^s$ (see e.g. Theorem 2.6 of [Con], and note the subsequent remark on

replacing $\mathbb{C}^s$ with $\mathcal{R}^s$). It follows that the set of zeros of $\det(\boldsymbol{W}')$ contain no open set of $\mathcal{R}^s$, and thus has Lesbegue measure 0 in $\mathcal{R}^s$. By the remarks in the prior paragraph, we know that the Lesbegue measure (taken over $\mathcal{R}^s$) of the support of the joint distribution on the s variables is non-zero (after restricting to the orthant given by the sign pattern). In particular, the set of zeros of $\det(\boldsymbol{W}')$ has Lesbegue measure 0 inside of the support of the joint pdf of the non-zero variables in $\boldsymbol{W}'$. We conclude that the joint density of the variables of $\boldsymbol{W}'$, after conditioning on a sign pattern, integrated over the set of zeros of $\det(\boldsymbol{W}')$ will be zero, meaning that $\boldsymbol{W}'$ will have full rank almost surely, conditioned on the sign pattern event in the first paragraph when held with probability $1 - O(\delta)$. □

Remark 125. Note that nearly all non-degenerate distributions $\mathcal{D}$ on d-dimensional vectors will satisfy $\inf_{v \in \mathcal{R}^d} \Pr_{x \sim \mathcal{D}}\Big[\langle v, x\rangle > 0\Big] = c = \Omega(1)$. For instance any multi-variate Gaussian distribution with non-degenreate (full-rank) covariance matrix $\boldsymbol{\Sigma}$ will satisfy this bound with $c = 1/2$, and this will also hold for any symmetric i.i.d. distribution over the entries of $x \sim \mathcal{D}$. Thus it will suffice to take $n = \Omega(k \log(k/\delta))$ for the result to hold.

Corollary 6.2.5. *Let $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ for rank k matrices $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$ and $\boldsymbol{V}^* \in \mathcal{R}^{k \times d}$, where $\mathcal{X} \in \mathcal{R}^{d \times n}$ is a matrix of random variables such that each column $\mathcal{X}_{*,i}$ is drawn i.i.d. from a distribution $\mathcal{D}$ with continuous p.d.f. $p(x) : \mathcal{R}^d \to \mathcal{R}$ such that $p(x) > 0$ almost everywhere in $\mathcal{R}^d$, and such that*

$$\inf_{v \in \mathcal{R}^d} \boldsymbol{Pr}_{x \sim \mathcal{D}}\Big[\langle v, x\rangle > 0\Big] = \Omega(k \log(1/\delta)/n)$$

Then, there exists an algorithm such that, with probability $1 - O(\delta)$, recovers $\boldsymbol{U}^*$, $\boldsymbol{V}^*$ exactly and runs in time poly$(n, m, d, k) \min\{n^{O(k)}, 2^n\}$.

6.3 NP-Hardness

The goal of this section is to prove that the problem of deciding whether there exists $\boldsymbol{V} \in \mathcal{R}^{k \times d}$ that solves the equation $\alpha f(\boldsymbol{V}\mathcal{X}) = w$ for fixed input $\alpha \in \mathcal{R}^{m \times k}, X \in \mathcal{R}^{d \times n}, A \in \mathcal{R}^{m \times n}$, is NP-hard. We will first prove the NP-hardness of a geometric separability problem, which will then be used to prove NP-hardness for the problem of deciding the feasibility of $\alpha f(\boldsymbol{V}\mathcal{X}) = w$. Our hardness reduction is from a variant of Boolean SAT, used in [Meg88] to prove NP-hardness of a similar geometric seperability problem, called *reversible 6-SAT*, which we will now define. For a Boolean formula ψ on variables $\{u_1, \ldots, u_n, \overline{u_1}, \ldots, u_n\}$ (where $\overline{u_i}$ is the negation of u_i), let $\overline{\psi}$ be the formula where every variable u_i and $\overline{u_i}$ appearing in ψ is replaced with $\overline{u_i}$ and u_i

respectively. For instance, if $\psi = (u_1 \vee u_2 \vee \overline{u_3}) \wedge (\overline{u_2} \vee u_3)$ then $\overline{\psi} = (\overline{u_1} \vee \overline{u}_2 \vee u_3) \wedge (u_2 \vee \overline{u_3})$.

Definition 6.3.1. *A Boolean formula ψ is said to be reversible if ψ and $\overline{\psi}$ are both either satisfiable or not satisfiable.*

The reverse 6-SAT problem is then to, given a reversible Boolean formula ψ where each conjunct has exactly six literals per clause, determine whether or not ψ is satisfiable. Observe, if ξ is a satisfying assignment to the variables of a reversible formula ψ, then $\overline{\xi}$, obtained by negating each assignment of ξ, is a satisfying assignment to $\overline{\psi}$. The following can be found in [Meg88].

Proposition 6.3.2 (NP-Hardness of Reversible 6-SAT)**.** *[Meg88]] Given a reversible formula ψ in conjunctive normal form where each clause has exactly six literals, it is NP-hard to decide whether ψ is satisfiable.*

We now introduce the following *ReLU-seperability* problem, and demonstrate NP-hardness via a reduction from reversible 6-SAT.

Definition 6.3.3 (ReLU-separability.)**.** *Given two sets $P = \{p_1, ..., p_r\}, Q = \{q_1, \ldots, q_s\}$ of vectors in R^d, the ReLU-seperability is to find vectors $x, y \in \mathbb{R}^d$ such that*

- *For all $p_i \in P$, both $p_i^T x \leqslant 0$ and $p_i^T y \leqslant 0$.*
- *For all $q_i \in Q$, we have $f(q_i^T x) + f(q_i^T y) = 1$ where $f(\cdot) = \max(\cdot, 0)$ is the ReLU function.*

We say that an instance of ReLU-seperability is satisfiable if there exists such an $x, y \in \mathcal{R}^d$ that satisfy the above conditions.

Proposition 6.3.4. *It is NP-Hard to decide whether an instance of ReLU-seperability is satisfiable.*

Proof. Let $u_1, \ldots, u_n$ be the variables of the reversible 6-SAT instance ψ, and set $d = n + 2$, and let x, y be the solutions to the instance of ReLU separability which we will now describe. The vector x will be such that x_i represents the truth value of u_i, and y_i represents the truth value of $\overline{x_i} = \overline{u_i}$. For $j \in [n + 2]$, let $e_j \in \mathcal{R}^{n+2}$ be the standard basis vector with a 1 in the j-th coordinate and 0 elsewhere. For each $i \in [n]$, we insert e_i and $-e_i$ into Q. This ensures that $f(x_i) + f(y_i) = 1$ and $f(-x_i) + f(-y_i) = 1$. This occurs iff either $x_i = 1$ and $y_i = -1$ or $x_i = -1$ and $y_i = 1$, so y_i is the negation of x_i. In other words, the case $x_i = 1$ and $y_i = -1$

means u_i is true and $\overline{u_i}$ is false, and the case $x_i = -1$ and $y_i = 1$ means u_i is false and $\overline{u_i}$ is true. Now suppose we have a clause of the form $u_1 \vee \overline{u_2} \vee u_3 \vee u_4 \vee \overline{u_5} \vee u_6$ in ψ. Then this clause can be represented equivalently by the inequality $x_1 - x_2 + x_3 + x_4 - x_5 + x_6 \geqslant -5$.

To represent this affine constraint, we add additional constraints that force $x_{n+1} + x_{n+2} = 1/2$ and $y_{n+1} + y_{n+2} = 1/2$ (note that the $n+1$, and $n+2$ coordinates do not correspond to any of the n variables u_i). We force this as follows. Add e_{n+1} and e_{n+2} to Q, and add $-2e_{n+1}$, $-2e_{n+2}$ to Q. This forces $f(x_i) + f(y_i) = 1$ and $f(-2x_i) + f(-2y_i) = 1$ for each $i \in \{n+1, n+2\}$. For each $i \in \{n+1, n+2\}$ there are only two solutions, either $x_i = 1$ and $y_i = -1/2$ or $x_i = -1/2$ and $y_i = 1$. Finally, we add the vector $e_{n+1} + e_{n+2}$ to Q, which forces $f(x_{n+1}+x_{n+2})+f(y_{n+1}+y_{n+2}) = 1$. Now if $x_{n+1} = 1$, then x_{n+2} must be $-1/2$ since otherwise there is no solution to $2 + f(\cdot) = 1$, and we know $x_{n+2} \in \{1, -1/2\}$. This forces $y_{n+2} = 1$, which forces $x_{n+1} + x_{n+2} = 1/2 = y_{n+1} + y_{n+2}$, and a symmetric argument goes through when one assumes $y_{n+1} = 1$. This lets us write affine inequalities as follows. For the clause $u_1 \vee \overline{u_2} \vee u_3 \vee u_4 \vee \overline{u_5} \vee u_6$, we can write the corresponding equation $x_1 - x_2 + x_3 + x_4 - x_5 + x_6 \geqslant -5$ precisely as a point constraint, which for us is $(-1, 1, -1, -1, 1, -1, 0, 0, \ldots, 0, -10, -10) \in P$ (the two -10's are in coordinate positions $n+1$ and $n+2$). Now this also forces the constraint $y_1 - y_2 + y_3 + y_4 - y_5 + y_6 \geqslant -5$, but since the formula is reversible so we can assume WLOG that $\overline{u_1} \vee u_2 \vee \overline{u_3} \vee \overline{u_4} \vee u_5 \vee \overline{u_6}$ is also a conjunct and so the feasible set is not affected, and the first n coordinates of any solution x will indeed correspond to a satisfying assignment to ψ if one exists. Since reversible 6-SAT is NP-hard by Proposition 6.3.2, the stated result holds. □

Theorem 126. *For a fixed $\alpha \in \mathcal{R}^{m\times k}, \mathcal{X} \in \mathcal{R}^{d\times n}, \mathbf{A} \in \mathcal{R}^{m\times n}$, the problem of deciding whether there exists a solution $\boldsymbol{V} \in \mathcal{R}^{k\times d}$ to $\alpha f(\boldsymbol{V}\mathcal{X}) = \mathbf{A}$ is NP-hard even for $k = 2$. Furthermore, for the case for $k = 2$, the problem is still NP-hard when $\alpha \in \mathcal{R}^{m\times 2}$ is allowed to be a variable.*

Proof. Now we show the reduction from ReLU-separability to our problem. Given an instance (P, Q) of ReLU separability as in Definition 6.3.3, set $\alpha = [1, 1]$, and $w = [0, 0, \ldots, 0, 1, 1, \ldots, 1]$ so $w_i = 0$ for $i \leqslant r$ and $w_i = 1$ for $r < i \leqslant r + s$. Let $\mathcal{X} = [p_1,\ p_2,\ \ldots, p_r,\ q_1, \ldots,\ q_s] \in \mathbb{R}^{d\times(r+s)}$. Now suppose we have a solution $\boldsymbol{V} = [x, y]^T \in \mathbb{R}^{(r+s)\times 2}$ to $\alpha f(\boldsymbol{V}\mathcal{X}) = w$. This means $f(p_i^T x) + f(p_i^t y) = 0$ for all $p_i \in P$, so it must be that both $p_i^T x \leqslant 0$and $p_i^T y \leqslant 0$. Also, we have $f(q_i^T x) + f(q_i^t y) = 1$ for all $q_i \in Q$. These two facts together mean that x, y are a solution to ReLU-separability. Conversely, if solutions x, y to ReLU separability exist, then for all $p_i \in P$, both $p_i^T x \leqslant 0$ and $p_i^T y \leqslant 0$ implies $f(p_i^T x) + f(p_i^t y) = 0$, and for all $q_i \in Q$ we get $f(q_i^T x) + f(q_i^T y) = 1$, so $\boldsymbol{V} = [x, y]^T$ is a solution to our factoring problem. Using the NP-hardness of ReLU-separability by Proposition 6.3.4, the result follows. Note here that $k = 2$

is a constant, but for larger $\alpha \in \mathcal{R}^{m \times k}$ with m rows and k columns, we can pad the new entries with zeros to reduce the problem to the aforementioned one, which completes the proof for a fixed α.

Now for $k = 2$ and α a variable, we add the following constraints to reduce to the case of $\alpha = [1, 1]$, after which the result follows. First, we add 2 new columns and 1 new row to $\mathcal{X}$, giving $\mathcal{X}' \in \mathcal{R}^{(d+1)\times(r+s+2)}$. We set

$$\mathcal{X}' = \begin{bmatrix} \mathcal{X} & \vec{0} & \vec{0} \\ \vec{0}^T & 1 & -1 \end{bmatrix}$$

Where $\mathcal{X}$ is as in the last paragraph, where $\vec{0}$ is a column vector of the appropriate dimensions above. Also, we set $\mathbf{A}' = [\mathbf{A}, 1, 1] \in \mathcal{R}^{r+s+4}$. Let $\boldsymbol{V} = [x, y]^T$ as before. This ensures that $\alpha_1 f(x_{d+1}) + \alpha_2 f(y_{d+1}) = 1$ and $\alpha_1 f(-x_{d+1}) + \alpha_2 f(-y_{d+1}) = 1$. As before, we cannot have that both (x_{d+1}) and (y_{d+1}) are negative, or that both are positive, as then one of the two constraints would be impossible. WLOG, $(y_{d+1}) < 0$. Then we have $\alpha_1 f(x_{d+1}) = 1$, which ensures $\alpha_1 > 0$, and $\alpha_2 f(-y_{d+1}) = 1$, which ensures $\alpha_2 > 0$.

Now suppose we have a solution to $\boldsymbol{V} = [x, y]^T$ and $\alpha \in \mathcal{R}^2$ to this new problem with $\mathcal{X}', \mathbf{A}'$. Then we can set $x' = x/\alpha_1$ and $y' = y/\alpha_2$, and $\alpha' = [1, 1]$, and we argue that we have recovered a solution $[x', y']$ to ReLU separability. Note that $[1, 1]f([x', y']^T\mathcal{X}') = \mathbf{A}'$, since we can always pull a positive diagonal matrix in and out of f. Then restricting to the first $r + s$ columns of $\mathcal{X}', \mathbf{A}'$, we see that $[1, 1]f([x', y']^T\mathcal{X}) = \mathbf{A}$, thus $[x', y']$ are a solution to the neural-net learning problem as in the first paragraph, so as already seen we have that x', y' is a solution to ReLU-separability. Similarly, any solution x, y to ReLU separability can easily be extended to our learning problem by simply using $\boldsymbol{V} = \begin{bmatrix} x & 1 \\ y & -1 \end{bmatrix}$ and $\alpha = [1, 1]$, which completes the proof. □

6.4 A Polynomial Time Exact Algorithm for Gaussian Input

In this section, we study an exact algorithm for recovering the weights of a neural network in the realizable setting, i.e., the labels are generated by a neural network when the input is sampled from a Gaussian distribution. We also show that we can use independent and concurrent work of Ge et. al. [GKLW18] to extend our algorithms to the input being sampled from a symmetric distribution. Our model is similar to non-linear generative models such as those

for neural networks and generalized linear models already well-studied in the literature [SJA16, SA14, KKSK11, MM18], but with the addition of the ReLU activation function f and the second layer of weights $\boldsymbol{U}^*$. In other words, we receive as input i.i.d. Gaussian[1] input $\mathcal{X} \in \mathcal{R}^{d\times n}$ and the generated output is $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$, where $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}$ and $\boldsymbol{V}^* \in \mathcal{R}^{k\times d}$. For the remainder of the section, we assume that both $\boldsymbol{V}^*$ and $\boldsymbol{U}^*$ are rank k. Note that this implies that $d \geqslant k$ and $k \leqslant m$. In Section 6.6, however, we show that if we allow for a larger $((\kappa(\boldsymbol{V}^*))^{O(k)})$ sample complexity, we can recover $\boldsymbol{U}^*$ even when it is not full rank.

We note that the generative model considered in [SA14] matches our setting, however, it requires the function f to be differentiable and $\boldsymbol{V}^*$ to be sparse. In contrast, we focus on f being ReLU. The ReLU activation function has gained a lot of popularity recently and is ubiquitous in applications [Com94, Hyv99, FJK96, HO00, AGMS12, LAF+12, HK13]. As mentioned in Sedghi et. al. [SA14], if we make no assumptions on $\boldsymbol{V}^*$, the resulting optimal weight matrix is not identifiable. Here, we make no assumptions on $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ apart from them being full rank and show an algorithm that runs in polynomial time. The main technical contribution is then to recover the optimal $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ exactly, and not just up to ε-error. By solving linear systems at the final step of our algorithms, as opposed to iterative continuous optimization methods, our algorithms terminate after a polynomial number of arithmetic operations.

Formally, suppose there exist fixed rank-k matrices $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ such that $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$, and $\mathcal{X}$ is drawn from an i.i.d. Gaussian distribution. Note that we can assume that each row $\boldsymbol{V}_i^*$ of $\boldsymbol{V}^*$ satisfies $\|\boldsymbol{V}_i^*\|_2 = 1$ by pulling out a diagonal scaling matrix $\boldsymbol{D}$ with positive entries from f, and noting $\boldsymbol{U}^* f(\boldsymbol{D}\boldsymbol{V}^*\mathcal{X}) = (\boldsymbol{U}^*\boldsymbol{D}) f(\boldsymbol{V}^*\mathcal{X})$. Our algorithm is given as input both $\mathbf{A}$ and $\mathcal{X}$, and tasked with recovering the underlying generative neural network $\boldsymbol{U}^*, \boldsymbol{V}^*$. In the context of training neural networks, we consider $\mathcal{X}$ to be the feature vectors and $\mathbf{A}$ to be the corresponding labels. Note $\boldsymbol{U}^*, \boldsymbol{V}^*$ are oblivious to $\mathcal{X}$, and are fixed prior to the generation of the random matrix $\mathcal{X}$. In this section we present an algorithm that is polynomial in all parameters, i.e., in the rank k, the condition number of $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$, denoted by $\kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*)$ and n, m, d.

Given an approximate solution to $\boldsymbol{U}^*$, we show that there exists an algorithm that outputs $\boldsymbol{U}^*, \boldsymbol{V}^*$ exactly and runs in time polynomial in all parameters. We begin by giving an altenative algorithm for orthonormal $\boldsymbol{V}^*$ based on *Independent Component Analysis*. We believe that this perspective on learning neural networks may be useful beyond our results. Next, we will give a general algorithm for exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$ which does not require $\boldsymbol{V}^*$ to be orthonormal. This algorithm is based on the completely different approach of tensor decomposition, yet yields

[1] See Remark 127

the same polynomial running time for exact recovery in the noiseless case. We now pause for a brief aside on the generalization of our results to the non-identity covariance case.

Remark 127. While our results are stated for when the columns of $\mathcal{X}$ are Gaussian with identity covariance, they can naturally be extended to $\mathcal{X}$ with arbitrary non-degenerate (full-rank) covariance $\mathbf{\Sigma}$, by noting that $\mathcal{X} = \mathbf{\Sigma}^{1/2}\mathcal{X}'$ where $\mathcal{X}'$ is i.i.d. Gaussian, and then implicitly replacing $\boldsymbol{V}^*$ with $\boldsymbol{V}^*\mathbf{\Sigma}^{1/2}$ so that $f(\boldsymbol{V}^*\mathcal{X}) = f((\boldsymbol{V}^*\Sigma^{1/2})\mathcal{X}')$, and noting that $\kappa(\boldsymbol{V}^*\Sigma^{1/2})$ blows up by a $\sqrt{\kappa(\mathbf{\Sigma})}$ factor from $\kappa(\boldsymbol{V}^*)$. All our remaining results, which do not require $\boldsymbol{V}^*$ to be orthonormal, hold with the addition of polynomial dependency on $\sqrt{\kappa(\mathbf{\Sigma})}$, by just thinking of $\boldsymbol{V}^*$ as $\boldsymbol{V}^*\Sigma^{1/2}$ instead.

We use the sample covariance as our estimator for the true covariance $\mathbf{\Sigma}$ and have the following guarantee:

Lemma 6.4.1. *(Estimating Covariance of $\mathcal{X}$ [Ver18].) Let $\mathcal{X} \in \mathcal{R}^{d\times N}$ such that for all $i \in [N]$, $\mathcal{X}_{*,i} \sim \mathcal{N}(0, \mathbf{\Sigma})$. Let $\mathbf{\Sigma}_N = \frac{1}{N}\sum_{i\in[N]} \mathcal{X}_{*,i}\mathcal{X}_{*,i}^T$. With probability at least $1 - 2e^{-\delta}$,*

$$\|\mathbf{\Sigma} - \mathbf{\Sigma}_n\|_2 \leqslant c\frac{d+\delta}{N}\|\mathbf{\Sigma}\|_2$$

for a fixed constant c.

We can then estimate Σ using a holdout set of $N = \Omega(n^2\delta^2)$ samples, which suffices to get an accurate estimate of the covariance matrix. We point out that, other than the tensor decomposition algorithm of Section 6.4.2 and the noisy half-space learning routine in Section 6.7, our algorithms do not even need to estimate the covariance matrix $\mathbf{\Sigma}$ in the multivariate case in order to approximately (or exactly) recover $\boldsymbol{U}^*, \boldsymbol{V}^*$. With regards to our tensor decomposition algorithms, while our estimator for the covariance introduces small error in the computation of the Score Function and the resulting tensor decomposition, this can be handled easily in the perturbation analysis of Theorem 130 (refer to Remark 4 in [JSA15]). For our half-space learning algorithm in Section 6.7, the error caused by estimating Σ is negligible, and can be added to the "advesarial" error $\mathbf{B}$ of Theorem 141 which is already handled.

In the following warm-up Section 6.4.1, where it is assumed that $\boldsymbol{V}^*$ is orthonormal, we cannot allow $\mathcal{X}$ to have arbitrary covariance, since then $\boldsymbol{V}^*\mathbf{\Sigma}^{1/2}$ would not be orthonormal. However, for in the more general algorithm which follows in Section 6.4.2, arbitrary non-degenerate covariance $\mathbf{\Sigma}$ is allowed.

6.4.1 An Independent Component Analysis Algorithm for Orthonormal V^*

We begin with making the simplifying assumption that the optimal V^* has orthonormal rows, as a warm-up to our more general algorithm. Note, if V^* is orthonormal and $\mathcal{X}$ is standard normal, then by 2-stability of Gaussian random variables, $V^*\mathcal{X}$ is a matrix of i.i.d. Gaussian random variables. Since Gaussian random variables are symmetric around the origin, each column of $f(V^*\mathcal{X})$ is sparse, has i.i.d entries, and has moments bounded away from Gaussians. Using these facts, we form a connection to the Independent Component Analysis (ICA) problem, and use standard algorithms for ICA to recover an approximation to U^*.

The ICA problem approximately recovers a subspace $\mathbf{B}$, given that the algorithm observes samples of the form $y = \mathbf{B}x+\mathbf{E}$, where x is i.i.d. and drawn from a distribution that has moments bounded away from Gaussians and $\mathbf{E}$ is Gaussian noise. The ICA problem has a rich history of theoretical and applied work [Com94, FJK96, Hyv99, HO00, FKV04a, LAF+12, AGMS12, HK13]. Intuitively, the goal of ICA is to find a linear transformation of the data such that each of the coordinates or features are as independent as possible. For instance, if the dataset is generated as $y = \mathbf{B}x$, where $\mathbf{B}$ is an unknown affine transformation and x has i.i.d. components, with no noise added, then applying $\mathbf{B}^{-1}$ to y recovers the independent components exactly, as long as x is non-Gaussian. Note, if $x \sim \mathcal{N}(0, \mathbf{I}_m)$, then by rotational invariance of Gaussians, we can only hope to recover $\mathbf{B}$ up to a rotation and the identity matrix suffices as a solution.

Definition 6.4.2. *(Independent Component Analysis.) Given $\epsilon > 0$ and samples of the form $y_i = \mathbf{B}x_i + \mathbf{E}_i$, for all $i \in [n]$, such that $\mathbf{B} \in \mathcal{R}^{m\times m}$ is unknown and full rank, $x_i \in \mathcal{R}^m$ is a vector random variable with independent components and has fourth moments strictly less than that of a Gaussian, the ICA problem is to recover an additive error approximation to $\mathbf{B}$, i.e., recover a matrix $\widehat{\mathbf{B}}$ such that $\|\widehat{\mathbf{B}} - \mathbf{B}\|_F \leqslant \epsilon$.*

We use the algorithm provided in Arora et. al. [AGMS12] as a black box for ICA. We note that our input distribution is rectified Gaussian, which differs from the one presented in [AGMS12]. Observe, our distribution is invariant to permutations and *positive* scaling, is sub-Gaussian, and has moments that are bounded away from Gaussian. The argument in [AGMS12] extends to our setting, as conveyed to us via personal communication [Ge18]. We have the following formal guarantee :

Theorem 128. *(Provable ICA, [AGMS12] and [Ge18].) Suppose we are given samples of the form $y_i = \mathbf{B}x_i + \mathbf{E}_i$ for $i = 1, 2, \ldots, n$, where $\mathbf{B} \in \mathcal{R}^{m\times m}$, the vector $x_i \in \mathcal{R}^m$ has i.i.d.*

components and has fourth moments strictly bounded away from Gaussian, and $\mathbf{E}_i \in \mathcal{R}^m$ *is distributed as* $\mathcal{N}(0, \mathbb{I}_m)$*, there exists an algorithm that with high probability recovers* $\widehat{\mathbf{B}}$ *such that* $\|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{\Pi}\boldsymbol{D}\|_F \leqslant \epsilon$*, where* $\mathbf{\Pi}$ *is a permutation matrix and* $\boldsymbol{D}$ *is a diagonal matrix such that it is entry-wise positive. Further, the sample complexity is* $n = poly\left(\kappa(\mathbf{B}), \frac{1}{\epsilon}\right)$ *and the running time is poly*(n, m).

We remark that ICA analyses typically require $\mathbf{B}$ to be a square matrix, and recall that$\boldsymbol{U}^*$ is $m \times k$ for $m \geqslant k$. To handle this, we sketch our samples using a dense Gaussian matrix with exactly k columns, and show this sketch is rank preserving. We will denote the resulting matrix by $\mathbf{T}\boldsymbol{U}^*$.

Algorithm 3 : ExactNeuralNet$(\mathbf{A}, \mathcal{X})$

Input : Matrices $\mathbf{A} \in \mathcal{R}^{d \times n}$ and $\mathcal{X} \in \mathcal{R}^{r \times n}$ such that each entry in $\mathcal{X} \sim \mathcal{N}(0,1)$.

1. Let $\mathbf{T} \in \mathcal{R}^{k \times m}$ be a matrix such that for all $i \in [k]$, $j \in [m]$, $\mathbf{T}_{i,j} \sim \mathcal{N}(0,1)$. Let $\mathbf{TA}$ be the matrix obtained by applying the sketch to $\mathbf{A}$.
2. Consider the ICA problem where we receive samples of the form $\mathbf{TA} = \mathbf{T}\boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$.
3. Run the ICA algorithm, setting $\epsilon = \frac{1}{\text{poly}(m,d,k,\kappa(\boldsymbol{U}^*))}$, to recover $\widehat{\mathbf{TU}}$ such that $\|\widehat{\mathbf{TU}} - \mathbf{T}\boldsymbol{U}^*\boldsymbol{\Pi}\boldsymbol{D}\|_F \leqslant \frac{1}{\text{poly}(m,d,k,\kappa(\boldsymbol{U}^*))}$.
4. Let $\overline{\mathcal{X}}$ be the first $\ell = \text{poly}(d, m, k, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*))$ columns of $\mathcal{X}$, and let $\overline{\mathbf{A}} = \boldsymbol{U}^* f(\boldsymbol{V}^* \overline{\mathcal{X}})$. Let $\tau = \frac{1}{\text{poly}(\ell)}$ be a threshold. Then for all $i \in [k], j \in [\ell]$, set

$$\widehat{f(\boldsymbol{V}\overline{\mathcal{X}})}_{i,j} = \begin{cases} 0 & \text{if } \left((\widehat{\mathbf{TU}})^{-1}\mathbf{T}\overline{\mathbf{A}}\right)_{i,j} \leqslant \tau \\ \left((\widehat{\mathbf{TU}})^{-1}\mathbf{T}\overline{\mathbf{A}}\right)_{i,j} & \text{otherwise} \end{cases}$$

5. Let S_j be the sparsity pattern of the vector $\widehat{f(\boldsymbol{V}\overline{\mathcal{X}})}_{j,*}$. For all $j \in [k]$, and $r \in [k]$, solve the following linear system of equations in the unknowns $x_j^r \in \mathcal{R}^k$.

$$\begin{aligned} \forall i \in [\ell] \setminus S_j \quad & (x_j^r \overline{\mathbf{A}})_i = 0, \\ & (x_j^r)_r = 1 \end{aligned}$$

   Where $(x_j^r)_r$ is the r-th coordinate of x_j^r.
6. Set w_j to be the first vector x_j^r such that a solution exists to the above linear system.
7. Let $\boldsymbol{W} \in \mathcal{R}^{k \times \ell}$ be the matrix where the i-th row is given by $w_i \overline{\mathbf{A}}$. Flip the signs of the rows of $\boldsymbol{W}$ so that $\boldsymbol{W}$ has no strictly negative entries.
8. For each $i \in [k]$, solve the linear system $(\boldsymbol{W}_{i,*})_{S_i} = \boldsymbol{V}_{i,*} \overline{\mathcal{X}}_{S_i}$ for $\boldsymbol{V} \in \mathcal{R}^{k \times d}$, where the subscript S_i means restricting to the columns of S_i. Normalize $\boldsymbol{V}$ to have unit norm rows. Finally, solve the linear system $\mathbf{A} = \boldsymbol{U} f(\boldsymbol{V}\mathcal{X})$ for $\boldsymbol{U}$, using Gaussian Elimination.

Output : $\boldsymbol{U}, \boldsymbol{V}$.

Lemma 6.4.3. *(Rank Preserving Sketch.) Let $\mathbf{T} \in \mathcal{R}^{k \times m}$ be a matrix such that for all $i \in [k]$, $j \in [m]$, $\mathbf{T}_{i,j} \sim \mathcal{N}(0,1)$. Let $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$ such that rank$(\boldsymbol{U}^*) = k$ and $m > k$. Then,*

$\mathbf{T}\boldsymbol{U}^* \in \mathcal{R}^{k\times k}$ *has rank* k. *Further, with probability at least* $1-\delta$, $\kappa(\mathbf{T}\boldsymbol{U}^*) \leqslant (k^2 m/\delta)\kappa(\boldsymbol{U}^*)$.

Proof. Let $\mathbf{M}\boldsymbol{\Sigma}\mathbf{N}^T$ be the SVD of $\boldsymbol{U}^*$, such that $\mathbf{M} \in \mathcal{R}^{m\times k}$ and $\boldsymbol{\Sigma}\mathbf{N}^T \in \mathcal{R}^{k\times n}$. Since columns of $\mathbf{M}$ are orthonormal and Gaussians are rotationally invariant, $\mathbf{T}\mathbf{M} \in \mathcal{R}^{k\times k}$ is i.i.d. standard normal. Further, $\boldsymbol{\Sigma}\mathbf{N}^T$ has full row rank and thus has a right inverse, i.e., $\mathbf{N}\boldsymbol{\Sigma}^{-1}$. Then, $\text{rank}(\mathbf{T}\boldsymbol{U}) = \text{rank}(\mathbf{T}\mathbf{M}\boldsymbol{\Sigma}\mathbf{N}^T) \leqslant \text{rank}(\mathbf{T}\mathbf{M})$. Further $\mathbf{T}\mathbf{M} = \mathbf{T}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}$, and therefore $\text{rank}(\mathbf{T}\mathbf{M}) = \text{rank}(\mathbf{T}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}) \leqslant \text{rank}(\mathbf{T}\boldsymbol{U})$. Recall, $\mathbf{T}\mathbf{M}$ is a $m \times k$ matrix of standard Gaussian random variables and has a non-zero determinant with probability 1.

Next, $\kappa(\mathbf{T}\boldsymbol{U}^*) \leqslant \kappa(\mathbf{T})\kappa(\boldsymbol{U}^*)$. Note $\mathbf{T}$ is at least $k+1 \times k$ and by Theorem 3.1 in [RV10], with probability $1-\delta$, $\sigma_{\min}(\mathbf{T}) \geqslant k\delta$. Similarly, by Proposition 2.4 [RV10], with probability $1 - 1/e^{\Omega(1/\delta)}$, $\sigma_{\max}(\mathbf{T}) \leqslant km/\delta$. Union bounding over the two events, with probability at least $1 - 1/\text{poly}(k)$, $\kappa(\mathbf{T}) \leqslant \text{poly}(k)$ and thus $\kappa(\mathbf{T}\boldsymbol{U}^*) \leqslant \kappa(\boldsymbol{U})k^2 m/\delta$. □

Algorithmically, we sketch the samples $\mathbf{T}\mathbf{A}$ such that they are of the form $\mathbf{T}\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$. By Lemma 6.4.3, $\mathbf{T}\boldsymbol{U}^*$ is a square matrix and has rank k. Since $\boldsymbol{V}$ is orthonormal, each column of $f(\boldsymbol{V}^*\mathcal{X})$ has entries that are i.i.d. $\max\{\mathcal{N}(0,1), 0\}$. Note, the samples $\mathbf{T}\mathbf{A}$ now fit the ICA framework, the noise $\mathbf{E} = 0$, and thus we can approximately recover $\boldsymbol{U}^*$, without even looking at the matrix $\mathcal{X}$. Here, we set $\epsilon = \frac{1}{\text{poly}(m,d,k,\kappa(\boldsymbol{U}^*))}$ to get the desired running time. Recall, given the polynomial depedence on $1/\varepsilon$, we cannot recover $\boldsymbol{U}^*$ exactly.

Corollary 6.4.4. *(Approximate Recovery using ICA.) Given* $\mathbf{A} \in \mathcal{R}^{m\times n}, \mathcal{X} \in \mathcal{R}^{d\times n}$, *and a sketching matrix* $\mathbf{T} \in \mathcal{R}^{k\times m}$ *such that* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ *and for all* $i \in [k]$, $j \in [m]$, $\mathbf{T}_{i,j} \sim \mathcal{N}(0,1)$, *there exists an algorithm that outputs an estimator to* $\widehat{\mathbf{T}\boldsymbol{U}^*}$ *such that* $\|\widehat{\mathbf{T}\boldsymbol{U}} - \mathbf{T}\boldsymbol{U}^*\boldsymbol{\Pi}\boldsymbol{D}\|_F \leqslant \frac{1}{\text{poly}(m,d,k,\kappa(\boldsymbol{U}^*))}$, *where* $\boldsymbol{\Pi}$ *is a permutation matrix and* $\boldsymbol{D}$ *is strictly positive diagonal matrix. Further, the running time is poly* $(m, d, k, \kappa(\boldsymbol{U}^*))$.

Exact Recovery: By Corollary 6.4.4, running ICA on $\mathbf{T}\mathbf{A} = \mathbf{T}\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$, we recover $\mathbf{T}\boldsymbol{U}^*$ approximately up to a permutation and positive scaling of the column. Note that we can disregard the permutation by simply assuming $\boldsymbol{V}$ has been permuted to agree with the $\boldsymbol{\Pi}$. Let $\widehat{\mathbf{T}\boldsymbol{U}}$ be our estimate of $\mathbf{T}\boldsymbol{U}^*$. We then restrict our attention to the first $\ell = \text{poly}\left(d, m, k, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*)\right)$ columns of $\mathcal{X}$, and call this submatrix $\overline{\mathcal{X}}$, and $\overline{\mathbf{A}} = \boldsymbol{U}^* f(\boldsymbol{V}^*\overline{\mathcal{X}})$. We then multiply $\mathbf{T}\overline{\mathbf{A}}$ by the inverse $(\widehat{\mathbf{T}\boldsymbol{U}})^{-1}$, which we show allows us to recover $\boldsymbol{D}^{-1} f(\boldsymbol{V}^*\overline{X})$ up to additive ε error where ε is at most $O\left(\frac{1}{\text{poly}(d,m,k,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*)}\right)$. Since the sketch $\mathbf{T}$ will preserve rank, $\mathbf{T}\boldsymbol{U}$ will have an inverse, and thus $(\widehat{\mathbf{T}\boldsymbol{U}})$ will be invertible (we can always perturbe the entries of our estimate by $1/\text{poly}(n)$ to ensure this). The inverse can then be computed in a polynomial number of arithmetic operations via Gaussian elimination. By a simple thresholding argument, we show

that after rounding off the entries below $\tau = 1/\text{poly}(\ell)$ in $(\widehat{\mathbf{TU}})^{-1}\mathbf{T}\overline{\mathbf{A}}$, we in fact recover the *exact* sign pattern of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$.

Our main insight is now that the only sparse vectors in the row space of $\overline{\mathbf{A}}$ are precisely the rows (up to positive a scaling) of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$. Specifically, we show that the only vectors in the row span of $\boldsymbol{U}^* f(\boldsymbol{V}^*\overline{\mathcal{X}})$ which have the same sign and sparsity pattern as a row of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$ are positives scalings of the rows of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$. Here, by *sparsity pattern*, we mean the subset of entries of a row that are non-zero. Since each row of $f(\boldsymbol{V}^*\mathcal{X})$ is non-negative, the sign and sparsity patterns of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$ together specify where the non-zero entries are (which are therefore strictly positive).

Now after exact recovery of the sign pattern of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$, we can set up a linear system to find a vector in the row span of $\mathbf{A}$ with this sign pattern, thus recovering each row of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$ exactly. Critically, we exploit the combinatorial structure of ReLUs together with the fact that linear systems can be solved in a polynomial number of arithmetic operations. This allows for exact recovery of $\boldsymbol{U}^*$ thereafter. Recall that we assume the rows of $\boldsymbol{V}^*$ have unit length, which removes ambiguity in the positive scalings used for the rows of $\boldsymbol{V}*$ (and similarly the columns of $\boldsymbol{U}^*$).

We begin by showing that the condition number of $\boldsymbol{V}^*$ is inversely proportional to the minimum angle between the rows of $\boldsymbol{V}^*$, if they are interpreted as vectors in $\mathcal{R}^d$. This will allow us to put a lower bound on the number of disagreeing sign patterns between rows of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$ in Lemma 6.4.6. We will then use these results to prove the uniqueness of the sign and sparsity patterns of the rows of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$ in Lemma 6.4.8.

Lemma 6.4.5. *Let $\theta_{\min} \in [0, \pi]$ be the smallest angle between the lines spanned by two rows of the rank k matrix $\boldsymbol{V} \in \mathcal{R}^{k\times d}$ which unit norm rows, in other words $\theta_{\min} = \min_{i,j} \arccos(\langle \boldsymbol{V}_{i,*}, \boldsymbol{V}_{j,*}\rangle)$ where $\arccos$ takes values in the principle range $[0, \pi]$. Then $\kappa(\boldsymbol{V}) > \frac{c}{\theta_{\min}}$ for some constant c.*

Proof. Let i, j be such that $\arccos(|\langle \boldsymbol{V}_{i,*}, \boldsymbol{V}_{j,*}\rangle|) = \theta_{\min}$. Let $\boldsymbol{V}^-$ be the pseudo-inverse of $\boldsymbol{V}$. Since $\boldsymbol{V}$ has full row rank, it follows that $\boldsymbol{V}(\boldsymbol{V}^-)^T = I_k$, thus $\langle \boldsymbol{V}_{i,*}, \boldsymbol{V}^-_{j,*}\rangle = 0$ and $\langle \boldsymbol{V}_{j,*}, \boldsymbol{V}^-_{j,*}\rangle = 1$. The first fact implies that $\boldsymbol{V}^-_{j,*}$ is orthonormal to $\boldsymbol{V}_{i,*}$, and the second that $\cos(\theta(\boldsymbol{V}_{j,*}, \boldsymbol{V}^-_{j,*})) = (\|\boldsymbol{V}^-_{j,*}\|_2)^{-1}$ where $\theta(\boldsymbol{V}_{j,*}, \boldsymbol{V}^-_{j,*})$ is the angle between $\boldsymbol{V}_{j,*}$ and $\boldsymbol{V}^-_{j,*}$.

Now let $x = \boldsymbol{V}_{i,*}, y = \boldsymbol{V}^-_{j,*}/\|\boldsymbol{V}^-_{j,*}\|, z = \boldsymbol{V}_{j,*}$. Note that x, y, z are all points on the unit sphere in r dimensions, and since scaling does not effect the angle between two vectors, we have $\theta(x, y) = \theta(\boldsymbol{V}_{i,*}, \boldsymbol{V}^-_{j,*})$. We know $\theta(x, y) = \pi/2$, and $\theta_{\min} = \theta(x, z)$, so the law of cosines gives $\cos(\theta(y, z)) = \frac{2-\|y-z\|_2^2}{2}$. We have $\|y - z\|_2 = \|(y - x) - (z - x)\|_2 \geqslant |\sqrt{2} - \|z - x\|_2|$. Again

by the law of cosines, we have $\|z - x\|_2^2 = 2 - 2\cos(\theta_{\min})$. Since $\cos(x) \approx 1 - \Theta(x^2)$ for small x (consider the Taylor expansion), it follows that $\|z - x\|_2 \leqslant c'\theta_{\min}$ for some constant c'. So $\|y - z\|_2^2 \geqslant 2 - 2\sqrt{2}\|z - x\|_2 + \|z - x\|_2^2 \geqslant 2 - c''\theta_{\min}$ for another constant c''. It follows that

$$\cos(\theta(y, z)) \leqslant \frac{c''\theta_{\min}}{2}$$

From which we obtain $\|\boldsymbol{V}_{j,*}^-\|_2 \geqslant 2/(c''\theta_{\min})$. It follows that $\sigma_1(\boldsymbol{V}^-) \geqslant \|e_{j,*}^T \boldsymbol{V}^-\|_2 = \|\boldsymbol{V}_{j,*}^-\|_2 \geqslant \frac{2}{c''\theta_{\min}}$. Since the rows of $\boldsymbol{V}$ have unit norm, we have $\sigma_1(\boldsymbol{V}) \geqslant 1$, so $\kappa(\boldsymbol{V}) = \sigma_1(\boldsymbol{V})\sigma_1(\boldsymbol{V}^-) \geqslant \frac{2}{c''\theta_{\min}}$ which is the desired result setting $c = \frac{2}{c''}$. □

Lemma 6.4.6. *Fix any matrix $\boldsymbol{V} \in \mathcal{R}^{k\times d}$ with unit norm rows. Let $\mathcal{X} \in \mathcal{R}^{d\times \ell}$ be an i.i.d. Gaussian matrix for any $\ell \geqslant t\text{poly}(k, \kappa)$, where $\kappa = \kappa(\boldsymbol{V})$. For every pair $i, j \in [k]$ with $i \neq j$, with probability $1 - 1/\text{poly}(\ell)$ there are at least t coordinates $p \in [\ell]$ such that $(\boldsymbol{V}\mathcal{X})_{i,p} < 0$ and $(\boldsymbol{V}\mathcal{X})_{j,p} > 0$.*

Proof. We claim that $\mathbf{Pr}[(\boldsymbol{V}\mathcal{X})_{i,p} < 0, (\boldsymbol{V}\mathcal{X})_{j,p} > 0] = \Omega(1/\kappa)$. To see this, Consider the 2-dimensional subspace H spanned by $\boldsymbol{V}_{i,*}$ and $\boldsymbol{V}_{j,*}$. Let θ be the angle between $\boldsymbol{V}_{i,*}$ and $\boldsymbol{V}_{j,*}$ in the plane H. Then the event in question is the event that a random Gaussian vector, when projection onto this plane H, lies between two vectors with angle θ between each other. By the rotational invariance and spherical symmetric of Gaussians (see, e.g. [Bry12]), this probability is $\frac{\theta}{2\pi}$. Since $\kappa(\boldsymbol{V}) > \frac{c}{\theta_{\min}} = \Omega(\frac{1}{\theta})$ by Lemma 6.4.5, it follows that a random gaussian splits $\boldsymbol{V}_{i,*}$ and $\boldsymbol{V}_{j,*}$ with probability $\Omega(1/\kappa)$ as desired.

Thus on each column p of $f(\boldsymbol{V}\mathcal{X})$, $f(\boldsymbol{V}_{i,*}\mathcal{X}_{*,p}) < 0$ and $f(\boldsymbol{V}_{j,*}\mathcal{X}_{*,p}) > 0$ with probability at least $\Omega(1/\kappa)$. Using the fact that the entries in separate columns of $\boldsymbol{V}\mathcal{X}$ are independent, by Chernoff bounds, with probability greater than $1 - k^2\exp(\Omega(\ell/\kappa)) > 1 - 1/\text{poly}(\ell)$, after union bounding over all $O(k^2)$ ordered pairs i, j, we have that $f(\boldsymbol{V}_{i,*}\mathcal{X}) < 0$ and $f(\boldsymbol{V}_{j,*}\mathcal{X}) > 0$ on at least $\Omega(\ell/\kappa) > t$ coordinates.

□

Lemma 6.4.7. *Let $\mathbf{Z}_i$ be the i-th column of $(\boldsymbol{V}\mathcal{X})$, where $\boldsymbol{V}$ has rank k. Then the covariance of the coordinates of $\mathbf{Z}_i$ are given by the $k \times k$ posiitve definite covariance matrix $\boldsymbol{V}\boldsymbol{V}^T$, and the joint density function is given by:*

$$p(\mathbf{Z}_{i,1}, \ldots, \mathbf{Z}_{i,k}) = \frac{\exp\left(-\frac{1}{2}\mathbf{Z}_i^T(\boldsymbol{V}\boldsymbol{V}^T)^{-1}\mathbf{Z}_i\right)}{\sqrt{(2\pi)^k \det(\boldsymbol{V}\boldsymbol{V}^T)}}$$

In particular, the joint probability density of any subset of entries of $\boldsymbol{V}\mathcal{X}$ is smooth and everywhere non-zero.

Proof. Since $\mathbf{Z}_i = \boldsymbol{V}(\mathcal{X}_i^T)^T$, where $\mathcal{X}_i^T$ are i.i.d. normal random variables, t is well known that the covariance is given by $\boldsymbol{V}\boldsymbol{V}^T$ [Gut09], which is positive definite since $\boldsymbol{V}$ has full row rank. These are sufficient conditions ([Ash]) for the pdf to be given in the form as stated in the Proposition. Since distinct columns of $\boldsymbol{V}\mathcal{X}$ are statistically independent (as they are generated by separate columns of $\mathcal{X}$), the last statement of the proposition follows.

□

The following Lemma demonstrates that the the only vectors in the row span of $f(\boldsymbol{V}^*\mathcal{X})$ with the same sign and sparsity pattern as $f(\boldsymbol{V}^*\mathcal{X})_{i,*}$, for any given row i, are positive scalings of $f(\boldsymbol{V}^*\mathcal{X})_{i,*}$. Recall that a sparsity pattern $S \subseteq [n]$ of a vector $y \in \mathcal{R}^n$ is just set of coordinates $i \in S$ such that $y_i > 0$.

Lemma 6.4.8. *Let $\mathcal{X} \in \mathcal{R}^{d\times\ell}$ be an i.i.d. Gaussian matrix for any $\ell > t\text{poly}(k, \kappa(\boldsymbol{V}^*))$. Let S be the sparsity pattern of a fixed row $f(\boldsymbol{V}^*\mathcal{X})_{i,*}$, and let $\emptyset \subsetneq S' \subseteq S$. Then w.h.p. (in t), the only vectors in the row span of $f(\boldsymbol{V}^*\mathcal{X})$ with sparsity pattern S', if any exist, are non-zero scalar multiples of $f(\boldsymbol{V}^*\mathcal{X})_{i,*}$.*

Proof. Suppose $\mathbf{Z} = wf(\boldsymbol{V}^*\mathcal{X})$ had sparsity pattern S' and was not a scaling of $f(\boldsymbol{V}^*\mathcal{X})_{i,*}$. Then w is not 1-sparse, since otherwise it would be a scaling of a another row of $f(\boldsymbol{V}^*\mathcal{X})$, and by Proposition 6.4.6 no row's sparsity pattern is contained within any other row's sparsity pattern. Let $\boldsymbol{W}$ be $f(\boldsymbol{V}^*\mathcal{X})$ restricted to the rows corresponding to the non-zero coordinates in w, and write $\mathbf{Z} = w\boldsymbol{W}$ (where now w has also been restricted to the appropriate coordinates). Since $\boldsymbol{W}$ has at least 2 rows, and since the sparsity pattern of $w\boldsymbol{W}$ is contained within the sparsity pattern of $f(\boldsymbol{V}^*\mathcal{X})_{i,*}$, by Proposition 6.4.6, taking $t = 10k^2$, we know that there are at least $10k^2$ non-zero columns of $\boldsymbol{W}$ for which $w\boldsymbol{W}$ is 0, so let $\boldsymbol{W}'$ be the submatrix of all such columns.

Now for each row $\boldsymbol{W}_i'$ of $\boldsymbol{W}'$ with less than k non-zero entries, remove this row $\boldsymbol{W}_i'$ and also remove all columns of $\boldsymbol{W}'$ where $\boldsymbol{W}_i'$ was non-zero. Continue to do this removal iteratively until we obtain a new matrix $\boldsymbol{W}''$ where now every row has at least k non-zero entries. Observe that the resulting matrix $\boldsymbol{W}''$ has at least $9k^2$ columns. If there are no rows left, then since we only removed k columns for every row removed, this means there were at least $9k^2$ columns of $\boldsymbol{W}'$ which contained only zeros, which is a contradiction since by construction the columns of $\boldsymbol{W}'$ were non-zero to begin with. So, let $k' \leqslant k$ be the number of rows remaining in $\boldsymbol{W}''$. Note

that since the rows we removed were zero on the columns remaining in $\boldsymbol{W}''$, there must still be a vector w', which in particular is w restricted to the rows of $\boldsymbol{W}''$, which has no zero-valued entries and such that $w'\boldsymbol{W}'' = 0$.

Now observe once we obtain this matrix $\boldsymbol{W}''$, note that we have only conditioned on the sparsity pattern of the entries of $\boldsymbol{W}''$ (over the randomness of the Gaussians $\mathcal{X}$), but we have not conditioned on the values of the non-zero entries of $\boldsymbol{W}''$. Note that this conditioning does not change the continuity of the joint distributions of the columns of $\boldsymbol{W}''$, since this conditioning is simply restricting the columns to the non-zero intersection of half spaces which define this sign pattern. Since the joint density function of the columns of $\boldsymbol{V}\mathcal{X}$ is non-zero on all of $\mathcal{R}^k$ by Lemma 6.4.7, it follows that, after conditioning, any open set in this intersection of half spaces which defines the sparsity pattern of $\boldsymbol{W}''$ has non-zero probability measure with respects to the joint density function.

Given this, the argument now proceeds as in Lemma 6.2.4. Since each row of $\boldsymbol{W}''$ has at least k non-zero entries, we can find a square matrix $\boldsymbol{W}^\dagger \in \mathcal{R}^{k' \times k'}$ obtained by a taking a subset of $k' < 9k^2$ columns of $\boldsymbol{W}''$ and permuting them such that the diagonal of $\boldsymbol{W}^\dagger$ has a non-zero sign pattern. After conditioning on the sign pattern so that the diagonal is non-zero, the determinant $\det(\boldsymbol{W}^\dagger)$ of $\boldsymbol{W}^\dagger$ is a non-zero polynomial in s random variables with $k' \leqslant s \leqslant (k')^2$. By Lemma 6.4.7, the joint density function of these s variables is absolutely continuous and everywhere non-zero on the domain. Here the domain Ω is the intersection of half spaces given by the sign pattern conditioning.

Since Ω is non-empty, it has unbounded Lebesgue measure in $\mathcal{R}^s$. Since $\det(\boldsymbol{W}^\dagger)$ is a non-zero polynomial in s real variables, it is well known that $\det(\boldsymbol{W}^\dagger)$ cannot vanish on any non-empty open set in $\mathcal{R}^s$ (see e.g. Theorem 2.6 of [Con], and note the subsequent remark on replacing $\mathbb{C}^s$ with $\mathcal{R}^s$). It follows that the set of zeros of $\det(\boldsymbol{W}^\dagger)$ contain no open set of $\mathcal{R}^s$, and thus has Lesbegue measure 0 in Ω. Integrating the joint pdf of the s random variables over this subset of measure 0, we conclude that the probability that the realization of the random variables is in this set is 0. So the matrix $\boldsymbol{W}''$ has rank k', and so $w'\boldsymbol{W}'' = 0$ is impossible, a contradiction. It follows that $\mathbf{Z}$ is a scaling of a row of $f(\boldsymbol{V}^*\mathcal{X})$ as needed. □

We will now need the following perturbation bounds for the pseudo-inverse of matrices.

Proposition 6.4.9 (Theorem 1.1 [MZ10]). *Let $\mathbf{B}^\dagger$ denote the Moore–Penrose Pseudo-inverse of $\mathbf{B}$, and let $\|\mathbf{B}\|_2$ denote the operator norm of $\mathbf{B}$. Then for any $\mathbf{E}$ we have*

$$\|(\mathbf{B}+\mathbf{E})^\dagger - \mathbf{B}^\dagger\|_F \leqslant \sqrt{2}\max\left\{\|\mathbf{B}^\dagger\|_2^2, \|(\mathbf{B}+\mathbf{E})^\dagger\|_2^2\right\}\|\mathbf{E}\|_F$$

We prove the following corollary which will be useful to us.

Corollary 6.4.10. *For any* $\mathbf{B}, \mathbf{E}$ *and* $\frac{1}{4} \geqslant \varepsilon > 0$ *with* $\|\mathbf{B}\|_2 \geqslant 1$, $\|\mathbf{E}\|_F \leqslant \frac{\varepsilon}{\kappa^2}$ *and where* $\kappa = \kappa(\mathbf{B})$ *is the condition number of* $\mathbf{B}$. *Then we have*

$$\|(\mathbf{B}+\mathbf{E})^\dagger - \mathbf{B}^\dagger\|_F \leqslant O(\varepsilon)$$

and moreover, if $\mathbf{B}$ *has full column rank, then*

$$\|(\mathbf{B}+\mathbf{E})^\dagger\mathbf{B} - \mathbb{I}\|_F \leqslant O(\|\mathbf{B}\|_2\varepsilon)$$

Proof. We have $\|(\mathbf{B}+\mathbf{E})^\dagger - \mathbf{B}^\dagger\|_F \leqslant \max\left\{\|\mathbf{B}^\dagger\|_2^2, \|(\mathbf{B}+\mathbf{E})^\dagger\|_2^2\right\}\frac{O(\varepsilon)}{\kappa^2}$ by applying Proposition 6.4.9. In the first case, this is at most $\frac{1}{\sigma_{\min}^2(\mathbf{B})}\frac{O(\varepsilon)}{\kappa^2} = O(\varepsilon)$ as stated. Here we used the fact that $\|\mathbf{B}\|_2 = \sigma_{\max}(\mathbf{B}) \geqslant 1$, so $1/\sigma_{\min}(\mathbf{B}) \leqslant \kappa$. In the second case of the max, we have $\|(\mathbf{B}+\mathbf{E})^\dagger - \mathbf{B}^\dagger\|_F \leqslant \|(\mathbf{B}+\mathbf{E})^\dagger\|_2^2\frac{O(\varepsilon)}{\kappa^2} = \sigma_{\min}^{-2}(\mathbf{B}+\mathbf{E})\frac{O(\varepsilon)}{\kappa^2}$. By the Courant-Fisher theorem [2], using that $\|\mathbf{E}\|_2 \leqslant \|\mathbf{E}\|_F \leqslant 1/(4\kappa)$, we have

$$\sigma_{\min}(\mathbf{B}+\mathbf{E}) \geqslant \inf_{x:\|x\|_2=1}\|x(\mathbf{B}+\mathbf{E})\|_2 \geqslant \inf_{x:\|x\|_2=1}\Big|\,\|x\mathbf{B}\|_2 - \|x\mathbf{E}\|_2\,\Big|$$

$$\geqslant \sigma_{\min}(\mathbf{B}) - \frac{1}{4\kappa} \geqslant \sigma_{\min}(\mathbf{B})/2 \geqslant 1/(2\kappa)$$

where the minimum is taken over vectors x with the appropriate dimensions. Thus in both cases, we have $\|(\mathbf{B}+\mathbf{E})^\dagger - \mathbf{B}^\dagger\|_F \leqslant O(\varepsilon)$, so

$$\|(\mathbf{B}+\mathbf{E})^\dagger\mathbf{B} - \mathbb{I}\|_F = \|((\mathbf{B}+\mathbf{E})^\dagger - \mathbf{B}^\dagger)\mathbf{B}\|_F \leqslant \|\mathbf{B}\|_2 O(\varepsilon)$$

□

We now are ready to complete the proof of the correctness of Algorithm 3

Theorem 129. *(Exact Recovery for Orthonormal* $\boldsymbol{V}^*$*.) Given* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$*, for rank* k*-matrices* $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ *where* $\boldsymbol{V}^*$ *is orthonormal and* $\mathcal{X} \in \mathcal{R}^{d\times n}$ *which is i.i.d. Gaussian with* $n = poly(d, k, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*))$*, there is a* $poly(n)$*-time algorithm which recovers* $\boldsymbol{U}^*, \boldsymbol{V}^*$ *exactly with probability* $1 - \frac{1}{poly(d,m,k)}$.

Proof. By Corollary 6.4.4, after sketching $\mathbf{A}$ by a Gaussian matrix $\mathbf{T} \in \mathcal{R}^{k\times m}$ and running ICA on $\mathbf{TA}$ in $\text{poly}(d, m, k, \kappa(\boldsymbol{U}^*))$ time, we recover $\widehat{\mathbf{T}\boldsymbol{U}^*}$ such that $\|\widehat{\mathbf{T}\boldsymbol{U}} - \mathbf{T}\boldsymbol{U}^*\boldsymbol{\Pi}\boldsymbol{D}\|_F \leqslant$

[2] See https://en.wikipedia.org/wiki/Min-max_theorem

$\frac{1}{\text{poly}(d,k,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*))}$ for a sufficiently high constant-degree polynomial, such that $\boldsymbol{\Pi}$ is a permutation matrix and $\boldsymbol{D}$ is strictly positive diagonal matrix. We can disregard $\boldsymbol{\Pi}$ by assuming the rows of $\boldsymbol{V}^*$ have also been permuted by $\boldsymbol{\Pi}$, and we can disregard $\boldsymbol{D}$ by pulling this scaling into $\boldsymbol{V}^*$ (which can be done since it is a positive scaling). Thus $\|\widehat{\mathbf{T}\boldsymbol{U}} - \mathbf{T}\boldsymbol{U}^*\|_F \leqslant \frac{1}{\text{poly}(d,k,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*))}$

Observe now that we can assume that $1 \leqslant \|\mathbf{T}\boldsymbol{U}^*\|_2 \leqslant 2$ by guessing a scaling factor c to apply to $\mathbf{A}$ before running ICA. To guess this scaling factor, we can find the largest column (in L_2) y of $\mathbf{TA}$, and note that $y = (\mathbf{T}\boldsymbol{U}^*)f(\boldsymbol{V}^*\mathcal{X}_{*,j})$ for some j. Since $\|f(\boldsymbol{V}^*\mathcal{X}_{*,j})\|_2 \leqslant O(\sqrt{\log(n)})d$ with high probability for all $j \in [n]$ (using the Gaussian tails of $\mathcal{X}$), it follows that $\|y\|_2 \leqslant \sigma_{\max}(\mathbf{T}\boldsymbol{U}^*)O(\sqrt{\log(n)})d$. Since with w.h.p there is at least one column of $f(\boldsymbol{V}^*\mathcal{X})$ with norm at least $1/\text{poly}(n)$, it follows that $\|y\|_2 \geqslant \sigma_{\min}(\mathbf{T}\boldsymbol{U}^*)/\text{poly}(n) \geqslant \frac{\sigma_{\max}(\mathbf{T}\boldsymbol{U}^*)}{\text{poly}(n,\kappa)}$. Thus one can make $\log\left(\text{poly}(n,\kappa,d)\right) = O(\log(n))$ guesses in geometrically increasing powers of 2 between $\|y\|_2/O(\sqrt{\log(n)})d$ and $\|y\|_2\text{poly}(n,\kappa)$ to find a guess such that $\|c\mathbf{T}\boldsymbol{U}^*\|_2 \in (1,2)$ as desired. This will allow us to use Corollary 6.4.10 in the following paragraph.

Now let $\widehat{\mathbf{T}\boldsymbol{U}}^\dagger$ be the pseduo-inverse of $\widehat{\mathbf{T}\boldsymbol{U}}$, and let $\overline{\mathbf{A}} = \boldsymbol{U}^* f(\boldsymbol{V}^*\overline{\mathcal{X}})$ where $\overline{\mathcal{X}}$ is the first $\text{poly}(d,k,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*))$ columns of $\mathcal{X}$. We now claim that the sign pattern of $(\widehat{\mathbf{T}\boldsymbol{U}})^\dagger\mathbf{T}\overline{\mathbf{A}} = \widehat{\mathbf{T}\boldsymbol{U}}^\dagger\mathbf{T}\boldsymbol{U}^* f(\boldsymbol{V}^*\overline{\mathcal{X}})$ is exactly equal to that of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$ after rounding all entries of with value less than $1/\text{poly}(\ell)$ to 0. Note that since $\mathbf{T}\boldsymbol{U}^*$ is full rank, it has an inverse (which is given by the pseudoinverse $(\mathbf{T}\boldsymbol{U}^*)^\dagger$. Let $\mathbf{Z}$ be the resulting matrix after rounding performing this rounding to $\widehat{\mathbf{T}\boldsymbol{U}}^\dagger\mathbf{T}\mathbf{A}'$. We now apply Corollary 6.4.10, with $\mathbf{T}\boldsymbol{U}^* = \mathbf{B}$ and $\widehat{\mathbf{T}\boldsymbol{U}} = \mathbf{B} + \mathbf{E}$. Since we guesses $\sigma_{\max}(\mathbf{T}\boldsymbol{U}^*)$ up to a factor of 2 and normalized $\widehat{\mathbf{T}\boldsymbol{U}}$ by it, it follows that the entries of the diagonal matrix $\boldsymbol{D}$ are all at most 2 and at least $1/(2\kappa(\mathbf{T}\boldsymbol{U}^*))$, and then using the fact that $\|f(\boldsymbol{V}^*\overline{\mathcal{X}})\|_F < \|\boldsymbol{V}^*\overline{\mathcal{X}}\|_F \leqslant \sqrt{\ell}\|V^*\|_F \leqslant \sqrt{\ell k}$ w.h.p. in ℓ (using well-known upper bounds on the spectral norm of a rectangular Gaussian matrix, see e.g. Corollary 5.35 if [Ver10b]) we obtain

$$\begin{aligned}\|\mathbf{Z} - \boldsymbol{D}f(\boldsymbol{V}^*\overline{\mathcal{X}})\|_F &= \|\left(\widehat{\mathbf{T}\boldsymbol{U}}^\dagger(\mathbf{T}\boldsymbol{U}^*) - \mathbb{I}\right)\boldsymbol{D}f(\boldsymbol{V}^*\mathcal{X}')\|_F \\ &\leqslant \frac{1}{\text{poly}\left(d,k,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*)\right)}\end{aligned}$$

Note that algorithmically, instead of computing the inverse $\widehat{\mathbf{T}\boldsymbol{U}}^\dagger$, we can first randomly perturb $\widehat{\mathbf{T}\boldsymbol{U}}$ by an entry-wise additive $1/\text{poly}(n)$ to ensure it is full rank, and then compute the true inverse, which can be done via Gaussian elimination in polynomially many arithmetic operations. By the same perturbational bounds, our results do not change when using the $1/\text{poly}(n)$ perturbed inverse, as opposed to the original pseudo-inverse.

Now since the positive entries of $\boldsymbol{D}f(\boldsymbol{V}^*\overline{\mathcal{X}})$ have normal Gaussian marginals, and $\boldsymbol{D}$ is a diagonal matrix which is entry-wise at most 2 and at least $1/(2\kappa(\mathbf{T}\boldsymbol{U}^*))$, the probability that any non-zero entry of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$ is less than $1/\text{poly}(\ell)$ is at most $2\kappa(\mathbf{T}\boldsymbol{U}^*)/\text{poly}(\ell)$, and we can then union bound over $\text{poly}(d,k,m,\kappa)$ such entries in $\overline{\mathcal{X}}$. Note that by Lemma 6.4.3, $\kappa(\mathbf{T}\boldsymbol{U}^*) < \text{poly}(k,d,m)\kappa(\boldsymbol{U}^*)$ w.h.p. in k,d,m, so $\text{poly}(\ell) >> \kappa(\mathbf{T}\boldsymbol{U}^*)$. Conditioned on this, with probability $1 - 1/\text{poly}(d,m,k,\kappa)$ for sufficiently large $\ell = \text{poly}(d,k,m,\kappa)$, every strictly positive entry of $\boldsymbol{D}f(\boldsymbol{V}^*\overline{\mathcal{X}})$, and therefore of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$, is non-zero in $\mathbf{Z}$, and moreover, and every other entry will be 0 in $\mathbf{Z}$, which completes the claim that the sign and sparsity patterns of the two matrices are equal.

Given this, for each $i \in [k]$ we can then solve a linear system to find a vector w_j such that $(w_j\overline{\mathbf{A}})_p = 0$ for all p not in the sparsity pattern of $\mathbf{Z}_{i,*}$. In other words, the sparsity pattern of $(w_j\overline{\mathbf{A}})$ must be contained in the sparsity pattern of $\mathbf{Z}_{i,*}$, which is the sparsity pattern of $f(\boldsymbol{V}^*\overline{\mathcal{X}})_{i,*}$ be the prior argument. By Lemma 6.4.8, the only vector in the row span of $\overline{\mathbf{A}}$ (which is the same as the row span of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$ since $\boldsymbol{U}^*$ is full rank) which has a non-zero sparsity pattern contained in that of $f(\boldsymbol{V}^*\overline{\mathcal{X}})_{i,*}$ must be a non-zero scaling of $f(\boldsymbol{V}^*\overline{\mathcal{X}})_{i,*}$. It follows that there is a unique w_j, up to a scaling, such that $w_j\overline{\mathbf{A}}$ is zero outside of the sparsity pattern of $f(\boldsymbol{V}^*\overline{\mathcal{X}})_{i,*}$. Since at least one of the entries r of w_j is non-zero, there exists some scaling such that $w_j\mathbf{A}$ is zero outside of the sparsity pattern of $f(\boldsymbol{V}^*\overline{\mathcal{X}})_{i,*}$ *and* $(w_j)_r = 1$ (where $(w_j)_r$) is the r-th coordinate of w_j). Since the first constraint is satisfied uniquely up to a scaling, it follows that there will be a unique solution w_j^r to at least one of the $r \in [k]$ linear systems in Step 5 of Algorithm 3, which will therefore be optained by the linear system. This vector w_j we obtain from Steps 5 and 6 of Algorithm 3 will therefore be such that $w_j\overline{\mathbf{A}}$ is a non-zero scaling of $f(\boldsymbol{V}^*\overline{\mathcal{X}})_{i,*}$.

Then in Step 7 of Algorithm 3, we construct the matrix $\boldsymbol{W}$, and flip the signs appropriately so that each row of $\boldsymbol{W}$ is a strictly positive scaling of a row of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$. We then solve the linear system $(\boldsymbol{W}_{i,*})_{S_i} = \boldsymbol{V}_{i,*}\overline{\mathcal{X}}_{S_i}$ for the unknowns $\boldsymbol{V}$, which can be done with a polynomial number of arithmetic operations via Gaussian elimination. Recall here that S_i is the set of coordinates where $\boldsymbol{W}_{i,*}$, and therefore $f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}})$, is non-zero. Since at least $1/3$ of the signs in a given row i will be positive with probability $1 - 2^{-\Omega(\ell)}$ by Chernoff bounds, restricting to this subset S_i of columns gives the equation $\boldsymbol{W}_{i,*} = \boldsymbol{V}_{i,*}^*\overline{\mathcal{X}}_{S_i}$. Conditioned on S_i having at least d columns, we have that $\overline{\mathcal{X}}_{S_i}$ is full rank almost surely, since it is a matrix of Gaussians conditioned on the fact that every column lies in a fixed halfspace. To see this, apply induction on the columns of $\overline{\mathcal{X}}_{S_i'}$, and note at every step $i < d$, the Lesbegue measure of the span of the first i columns is 0 in this halfspace, and thus the $i+1$ column will not be contained in it almost surely. It follows that there is a unique solution $\boldsymbol{V}_{i,*}$ for each row i, which must therefore be the corresponding row of $\boldsymbol{V}^*$ (we

normalize the rows of $\boldsymbol{V}_{i,*}$ to have unit norm so that they are precisely the same). So we recover $\boldsymbol{V}^*$ exactly via these linear systems. Finally, we can solve the linear system $\mathbf{A} = \boldsymbol{U} f(\boldsymbol{V}^* \mathcal{X})$ for the variables $\boldsymbol{U}$ to recover $\boldsymbol{U}^*$ exactly in strongly polynomial time. Note that this linear system has a unique solution, since $f(\boldsymbol{V}^* \mathcal{X})$ is full rank w.h.p. by Lemma 6.2.4, which completes the proof. □

6.4.2 General Algorithm

We now show how to generalize the algorithm from the previous sub-section to handle non-orthonormal $\boldsymbol{V}^*$. Observe that when $\boldsymbol{V}^*$ is no longer orthonormal, the entries within a column of $\boldsymbol{V}^* \mathcal{X}$ are no longer independent. Moreover, due to the presence of the non-linear function $f(\cdot)$, no linear transformation will exist which can make the samples (i.e. columns of $f(\boldsymbol{V}^* \mathcal{X})$) independent entry-wise. While the entries do still have Gaussian marginals, they will have the non-trivial covariance matrix $\boldsymbol{V}^* (\boldsymbol{V}^*)^T \neq \mathbb{I}_k$. Thus it is no longer possible to utilize previously developed techniques from independent component analysis to recover good approximations to $\boldsymbol{U}^*$. This necessitates a new approach.

Our starting point is the generative model considered by Janzamin et. al. [JSA15], which matches our setting, i.e. $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$. The main idea behind this algorithm is to construct a tensor that is a function of both $\mathbf{A}, \mathcal{X}$ and then run a tensor decomposition algorithm to recover the low-rank components of the resulting tensor. While computing a tensor decomposition is NP-hard in general [HL13], there is a plethora of work on special cases, where computing such decompositions is tractable [BCMV14, SWZ16, WA16, GVX14, GM15, BM16]. Tensor decomposition algorithms have recently become an invaluable algorithmic primitive and found a tremendous number of applications in statistical and machine learning tasks [JSA15, JSA14, GLM17, AGHK14a, BKS15].

A key step is to construct a non-linear transform of the input by utilizing knowledge about the underlying pdf for the distribution of $\mathcal{X}$, which we denote by $p(x)$. The non-linear function considered is the so called Score Function, defined in [JSA14], which is the normalized m-th order derivative of the input probability distribution function $p(x)$.

Definition 6.4.11. *(Score Function.) Given a random vector $x \in \mathcal{R}^d$ such that $p(x)$ describes the corresponding probability density function, the m-th order score function $\mathcal{S}_m(x) \in \otimes^m \mathcal{R}^d$ is defined as*

$$\mathcal{S}_m(x) = (-1)^m \frac{\nabla_x^{(m)} p(x)}{p(x)}$$

The tensor that Janzamin et. al. [JSA14] considers is the cross moment tensor between $\mathbf{A}$ and $\mathcal{S}_3(\mathcal{X})$. This encodes the correlation between the output and the third order score function. Intuitively, working with higher order tensors is necessary since matrix decompositions are only identifiable up to orthogonal components, whereas tensor have identifiable non-orthogonal components, and we are specifically interested in recovering approximations for non-orthonormal $\boldsymbol{V}^*$. Computing the score function for an arbitrary distribution can be computationally challenging. However, as mentioned in Janzamin et. al. [JSA14], we can use orthogonal polynomials that help us compute the closed form for the score function $\mathcal{S}_{(m)}(x)$, in the special case when $x \sim \mathcal{N}(0, \mathbf{I})$.

Definition 6.4.12. *(Hermite Polynomials.) If the input is drawn from the multi-variate Gaussian distribution, i.e.* $x \sim \mathcal{N}(0, \mathbf{I})$*, then* $\mathcal{S}_{(m)}(x) = \mathcal{H}_m(x)$*, where* $H_m(x) = \frac{(-1)^m \nabla_x^{(m)} p(x)}{p(x)}$ *and* $p(x) = \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{\|x\|_2^2}{2}}$.

Since we know a closed form for the m-th order Hermite polynomial, the tensor $\mathcal{S}_{(m)}$ can be computed efficiently. The critical structural result in the algorithm of [JSA15] is to show that in expectation, the cross moment of the output and the score function actually forms a rank-k tensor, where the rank-1 components capture the rows of $\boldsymbol{V}^*$. Formally,

Lemma 6.4.13. *(Generalized Stein's Lemma [JSA15].) Let* $\mathbf{A}, \mathcal{X}$ *be input matrices such that* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$*, where* f *is a non-linear, thrice differentiable activation function. Let* $\mathcal{S}_3(x)$ *be the* 3*-rd order score function from Definition 6.4.11. Then,*

$$\widetilde{\mathcal{T}} = \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{A}_{*,i} \otimes \mathcal{S}_3(\mathcal{X}_{*,i})\right] = \sum_{j=1}^{k} \mathop{\mathbb{E}}_{x}\left[f'''(\boldsymbol{V}^* x)\right] \boldsymbol{U}^*_{*,j} \otimes \boldsymbol{V}^*_{j,*} \otimes \boldsymbol{V}^*_{j,*} \otimes \boldsymbol{V}^*_{j,*}$$

where f''' *is the third derivative of the activation function and* $x \sim p(x)$.

Note, $\widetilde{\mathcal{T}}$ is a 4-th order tensor and can be constructed from the input $\mathbf{A}$ and $\mathcal{X}$. The first mode of $\widetilde{\mathcal{T}}$ can be contracted by multiplying it with a random vector θ, therefore,

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbf{A}_{*,i} \otimes \mathcal{S}_3(\mathcal{X}_{*,i})\right] = \sum_{j=1}^{k} \lambda_j \boldsymbol{V}^*_{j,*} \otimes \boldsymbol{V}^*_{j,*} \otimes \boldsymbol{V}^*_{j,*}$$

where $\lambda_j = \mathbb{E}_x\left[f'''(\boldsymbol{V}^* x)\right] \langle \boldsymbol{U}^*_{*,j}, \theta \rangle$. Therefore, if we could recover the low-rank components of $\widetilde{\mathcal{T}}$ we would be obtain a approximate solution to $\boldsymbol{V}^*$. The main theorem in [JSA15] states that under a set of conditions listed below, there exists a polynomial time algorithm that recovers an

additive error approximation to $\boldsymbol{V}^*$. Formally,

Theorem 130. *(Approximate recovery [JSA15]) Let $\mathbf{A} \in \mathcal{R}^{m\times n}$, $\mathcal{X} \in \mathcal{R}^{d\times n}$ be inputs such that $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X}) + \eta$, where f is a non-linear thrice differentiable activation function, $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}$ has full column rank, $\boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ has full row rank, for all $i \in [n]$, $\mathcal{X}_{*,i} \sim \mathcal{N}(0, \mathbf{I})$ and η is mean zero sub-Gaussian noise with variance σ_{noise}. Then, there exists an algorithm that recovers $\widehat{\boldsymbol{V}}$ such that $\|\widehat{\boldsymbol{V}} - \boldsymbol{D}\boldsymbol{\Pi}\boldsymbol{V}^*\|_F \leqslant \epsilon$, where $\boldsymbol{D}$ is a diagonal ± 1 matrix and $\boldsymbol{\Pi}$ is a permutation matrix. Further, the algorithm runs in time*

$$\textit{poly}\left(m, d, k, \frac{1}{\epsilon}, \mathbb{E}\left[\|\mathbf{M}_3(x)\mathbf{M}_3(x)^T\|_2\right], \mathbb{E}\left[\|\mathcal{S}_2(x)\mathcal{S}_2(x)^T\|_2\right], \frac{1}{\lambda_{\min}}, \lambda_{\max}, \frac{\widetilde{\lambda}_{\max}}{\widetilde{\lambda}_{\min}}, \kappa(\boldsymbol{V}^*), \sigma_{noise}\right)$$

where $\mathcal{S}_3$ is the 3-rd order score function, $\mathbf{M}_3(x)\mathcal{R}^{d\times d^2}$ is the matricization of $\mathcal{S}_3$, $\lambda_j = \mathbb{E}_x\left[f'''(\boldsymbol{V}^*x)\right]\langle \boldsymbol{U}^*_{*,j}, \theta\rangle$, $\widetilde{\lambda}_j = \mathbb{E}_x\left[f''(\boldsymbol{V}^*x)\right]\langle \boldsymbol{U}^*_{*,j}, \theta\rangle$,$\kappa(\boldsymbol{V}^*)$ is the condition number, σ_{noise} is the variance of η and. Note, in the case where $\mathcal{X}_{*,i} \sim \mathcal{N}(0, \mathbf{I})$, $\mathbb{E}\left[\|\mathbf{M}_3(x)\mathbf{M}_3(x)^T\|_2\right] = O(d^3)$ and $\mathbb{E}\left[\|\mathcal{S}_2(x)\mathcal{S}_2(x)^T\|_2\right] = O(d^2)$.

Remark 131. We only use the Whitening, Tensor Decomposition and Unwhitening steps from Janzamin et. al. [JSA15], and therefore the sample complexity and running time only depends on Lemma 9 and Lemma 10 in [JSA15].

However, there are many technical challenges in extending the aforementioned result to our setting. We begin with using the estimator from Theorem 130 in the setting where the noise, η, is 0. The first technical challenge is the above theorem requires the activation function f to be thrice diffrentiable, however ReLU is not. To get around this, we use a result from approximation theory to show that ReLU can be well approximated every where with a low-degree polynomial.

Lemma 6.4.14. *(Approximating ReLU [GK17].) Let $f(x) = \max(0, x)$ be the ReLU function. Then, there exists a polynomial $p(x)$ such that*

$$\sup_{x\in[-1,1]} |f(x) - p(x)| \leqslant \eta$$

and $\deg(p) = O(\frac{1}{\eta})$ and $p([-1, 1]) \subseteq [0, 1]$.

This polynomial is at least thrice differentiable and can be easily extended to the domain we care about using simple transformations. We assume that the samples we observe are of the form $\boldsymbol{U}^*p(\boldsymbol{V}^*\mathcal{X})$ corrupted by small adversarial error. Formally, the label matrix $\mathbf{A}$ can be viewed as

being generated via $\mathbf{A} = \boldsymbol{U}^* p(\boldsymbol{V}^* \mathcal{X}) + \mathbf{Z}$, where $\mathbf{Z} = \boldsymbol{U}^* \Big(f(\boldsymbol{V}^* \mathcal{X}) - p(\boldsymbol{V}^* \mathcal{X}) \Big)$. We note that we only use the approximation as an analysis technique and show that we can get an approximate solution to $\boldsymbol{V}^*$. First, we make a brief remark regarding the normalization of the entries in $\mathbf{A}$.

Remark 132. Observe in both the noiseless and noisy cases, the latter being where $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$ where $\mathbf{E}$ is i.i.d. mean 0 with variance σ^2, that by scaling $\mathbf{A}$ by $1/\|\mathbf{A}_{*,\max}\|_2$, where $\|\mathbf{A}_{*,\max}\|_2$ is the largest column norm of $\mathbf{A}$, we can ensure that the resulting $\boldsymbol{U}^*$ has $\|\boldsymbol{U}^*\|_2 < m \max\{1, \sigma\} \kappa(\boldsymbol{U}^*)$, where σ^2 is the variance of the noise $\mathbf{E}$ (in the noisy case). To see why this is true, suppose this were not the case. Observe that w.h.p. at least half of the columns $\boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ which will have norm at least $\omega(1)\sigma_{\min}^{-1}(\boldsymbol{U}^*)$ (since w.h.p. half the columns of $f(\boldsymbol{V}^* \mathcal{X})$ have norm $\omega(1)$), thus if $\|\boldsymbol{U}^*\|_2 > m \max\{1, \sigma\} \kappa(\boldsymbol{U}^*)$ after normalization, then then at least half of the normalized columns of $\boldsymbol{U} f(\boldsymbol{V}^* \mathcal{X})$ will have norm $\omega(m \max\{1, \sigma\})$. By Markov inequality and a Chernoff bound, strictly less than $1/4$ of the columns of the original $\mathbf{E}$ can have norm $\omega(m\sigma)$ w.h.p., and since the normalized $\mathbf{E}$ is strictly smaller, by triangle inequality there will be a column of $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$ after normalization with larger than unit norm, a contradiction. Thus we can assume this normalization, giving $\eta << \frac{1}{\|\boldsymbol{U}^*\|_2}$ for sufficiently small $\eta = O(\frac{1}{\text{poly}(n,d,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*),\sigma)})$.

We now set η in Lemma 6.4.14 to be $\frac{1}{\text{poly}(n,d,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*),\sigma)}$. By the operator norm bound of Lemma 6.4.18, we know that $\|\boldsymbol{V}^* \mathcal{X}\|_F = O(\sqrt{nk})$, w.h.p., so $\|\mathbf{Z}\|_F = O(\|\boldsymbol{U}^*\|_2 \sqrt{nk}\eta) = O(\frac{1}{\text{poly}(n)})$ as needed. We again construct the same tensor, $\widetilde{\mathcal{T}} = \mathbb{E}\left[\sum_{i=1}^n \mathbf{A}_{*,i} \otimes \mathcal{S}_3(\mathcal{X}_{*,i})\right]$. Our analysis technique is now as follows. We add a light $\mathcal{N}(0,1)$ random matrix to our input $\mathbf{A}$, and argue that the variation distance between the distribution over inputs $\mathbf{A}$ (for a fixed $\mathcal{X}$), between the case of $\mathbf{A}$ using f and $\mathbf{A}$ using the polynomial p as a non-linear activation, is at most $1/\text{poly}(n)$. As a result, the input using ReLUs is statistically indistinguishable in variation distance from samples generated using the polynomial approximation to the ReLU function. Thus, any algorithm that succeeds on such a polynomial approximation must also succeed on the ReLU. Therefore, the algorithm from Theorem 130 still holds for approximate recovery using ReLUs. Formally,

Lemma 6.4.15. *The variational distance between n samples of the form $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \boldsymbol{G}$, where the columns of $\boldsymbol{G}$ are $\mathcal{N}(0, \mathbb{I}_d)$ and $\mathcal{X}$ is fixed, and $\mathbf{A}' = \boldsymbol{U}^* p(\boldsymbol{V}^* \mathcal{X}) + \boldsymbol{G} + \mathbf{Z}$ where $\|\mathbf{Z}\|_F = \frac{1}{poly(n)}$ is at most $\frac{1}{poly(n)}$.*

Proof. Given two independent Gaussian $\mathcal{N}(\mu_1, \mathbb{I}), \mathcal{N}(\mu_2, \mathbb{I})$, a standard result in probability theory is that their variations distance is $\Theta(\|\mu_1 - \mu_2\|_2)$ [Das08]. Thus the variation distance between

the i-th column of $\mathbf{A}$ and $\mathbf{A}'$ is $O(\|\mathbf{Z}_{*,i}\|_2)$. Since the columns of the input are independent, the overall distribution is a product distribution so the variation distance adds. Thus the total variation distance is at most $O(\|\mathbf{Z}_{i,*}\|_F^2) = \frac{1}{\text{poly}(n)}$ as needed. □

It follows from the above lemma that the algorithm corresponding to Theorem 130 cannot distinguish between receiving samples from the ReLU distribution with artificially added Gaussian noise or the samples from the polynomial approximation with small adversarial noise. Therefore, the algorithm recovers an approximation to the underlying weight matrix $\boldsymbol{V}^*$ in polynomial time. Formally, if we have an algorithm which can solve a class of problems coming from a distribution $\mathcal{D}$ with failure probability at most δ, then it can solve problems coming a distribution $\mathcal{D}'$ with failure probability at most $O(\delta + \delta')$, where δ' is the variational distance between $\mathcal{D}$ and $\mathcal{D}'$. Since δ' in our case is $\frac{1}{\text{poly}(n)}$, we can safely ignore this additional failure probability going forward. This is summarized in the following lemma, which follows directly from the definition of variation distance. Namely, that the probability of any event in one distribution can change by at most the variation distance in another distribution, in particular the event that an algorithm succeeds on that distribution.

Lemma 6.4.16. *Suppose we have an algorithm $\mathcal{A}$ that solves a problem $\mathcal{P}$ taken from a distribution $\mathcal{D}$ over $\mathcal{R}^n$ with probability $1 - \delta$. Let $\mathcal{D}'$ be a distribution over $\mathcal{R}^n$ with variation distance at most $\delta' \geqslant 0$ from $\mathcal{D}$. Then if $\mathcal{P}'$ is drawn from $\mathcal{D}'$, algorithm $\mathcal{A}$ will solve $\mathcal{P}'$ with probability $1 - O(\delta + \delta')$.*

Corollary 6.4.17. *(Approximate ReLU Recovery.) Let $\mathbf{A} \in \mathcal{R}^{m \times n}$, $\mathcal{X} \in \mathcal{R}^{d \times n}$ be inputs such that $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$, where f is the ReLU activation function, $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}$ has full column rank, $\boldsymbol{V}^* \in \mathcal{R}^{k \times d}$ has full row rank, for all $i \in [n]$, $\mathcal{X}_{*,i} \sim \mathcal{N}(0, \mathbf{I})$. Then, there exists an algorithm that recovers $\widehat{\boldsymbol{V}}$ such that $\|\widehat{\boldsymbol{V}} - \boldsymbol{D}\boldsymbol{\Pi}\boldsymbol{V}^*\|_F \leqslant \frac{1}{poly(n,m,d,\kappa(\boldsymbol{U}^*))}$, where $\boldsymbol{D}$ is a diagonal ± 1 matrix and $\boldsymbol{\Pi}$ is a permutation matrix. Further, the running time of this algorithm is poly$(n, m, d, \kappa(\boldsymbol{U}^*))$.*

First observe that we can assume WLOG that $\Pi = \mathbb{I}$, in other words that we recover an approximate $\boldsymbol{V}^*$ only up to its signs and not a permutation. We do this by simply (implicitly) permuting the rows of $\boldsymbol{V}^*$ to agree with our permutation, and permuting the columns of $\boldsymbol{U}^*$ by the same permutation. The resulting $\mathbf{A}$ is identical, and so we can assume that we know the permutation already.

Algorithm 4 : ExactNeuralNet$(\mathbf{A}, \mathcal{X})$

`Input :` Matrices $\mathbf{A} \in \mathcal{R}^{m \times n}$ and $\mathcal{X} \in \mathcal{R}^{d \times n}$ such that each entry in $\mathcal{X} \sim \mathcal{N}(0,1)$.

1. Let $\mathcal{S}_3(x) = H_3(x)$, where $H_3(x) = \frac{-\nabla_x^{(3)} p(x)}{p(x)}$ is the 3-rd order Hermite polynomial and and $p(x) = \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{\|x\|_2^2}{2}}$.
2. Let $\mathbf{A}' = \mathbf{A} + \boldsymbol{G}$ where $\boldsymbol{G} \in \mathcal{R}^{m \times n}$ and $\boldsymbol{G}_{i,j} \sim \mathcal{N}(0,1)$.
3. Compute the 4-th order tensor $\widetilde{\mathcal{T}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{A}'_{*,i} \otimes \mathcal{S}_3(\mathcal{X}_{*,i})$. Collapse the first mode using a random vector θ. By Lemma 6.4.13, $\widetilde{T}(\theta, \mathbf{I}, \mathbf{I}, \mathbf{I}) = \sum_{j=1}^{k} \lambda_j \boldsymbol{V}^*_{j,*} \otimes \boldsymbol{V}^*_{j,*} \otimes \boldsymbol{V}^*_{j,*}$, where $\lambda_j = \mathbb{E}_x \left[f'''(\boldsymbol{V}^* x) \right] \langle \boldsymbol{U}^*_{*,j}, \theta \rangle$.
4. Compute a CP-decomposition of $\widetilde{T}(\theta, \mathbf{I}, \mathbf{I}, \mathbf{I})$ using Tensor Power Method corresponding to Theorem 130, [JSA15], with accuracy parameter $\epsilon = \frac{1}{\text{poly}(d, m, \kappa(\boldsymbol{V}), \kappa(\boldsymbol{U}))}$ to obtain $\widehat{\boldsymbol{V}}$ such that $\|\widehat{\boldsymbol{V}} - \boldsymbol{D}\boldsymbol{\Pi}\boldsymbol{V}^*\|_F \leqslant \frac{1}{\text{poly}(d, m, \kappa(\boldsymbol{V}), \kappa(\boldsymbol{U}))}$, where $\boldsymbol{D}$ is a diagonal ± 1 matrix and $\boldsymbol{\Pi}$ is a permutation matrix.
5. Run the Recovering Signs Algorithm (5) on $\widehat{\boldsymbol{V}}$, $\mathbf{A}$ and $\mathcal{X}$ to obtain $\boldsymbol{V}^*$.
6. Using the matrix $\boldsymbol{V}^*$ obtained above, set up and solve the following linear system for the matrix $\boldsymbol{U}$:
$$\mathbf{A} = \boldsymbol{U} f(\boldsymbol{V}^* \mathcal{X}) \tag{6.1}$$
7. Let $\boldsymbol{U}^*$ be the solution to the above linear system.

`Output :` $\boldsymbol{U}^*, \boldsymbol{V}^*$.

Unfortunately, the ambiguity in signs resulting from the algorithm of Theorem 130 is a non-trivial difficulty, and must be resolved algorithmically. This is due to the fact that the ReLU is sensitive to negative scalings, as $f(\cdot)$ only commutes with positive scalings. Suppose the diagonal of $\boldsymbol{D}$ of Corollary 6.4.17 is given by the coefficents $\xi_i \in \{1, -1\}$. Then in order to recover the weights, we must recover the terms ξ_i, Naively trying each sign results in a running time of 2^k, which is no longer polynomial[3]. Thus, a considerably technical challenge will be to show how to determine the correct scaling for each row even in the presence of noise. We begin

[3]We remark that some prior results [JSA15] were able to handle this ambiguity by considering only a restricted class of smooth activation functions $f(\cdot)$ with the property that $f(x) = 1 - f(-x)$ for all $x \in \mathcal{R}$. Using affine transformations after application of the ReLU, this sign ambiguity for such activation functions can be accounted for. Since firstly the ReLU does not satisfy this condition and is non-trivially sensitive to the signs of its input,and secondly we are restricting to optimization over networks without affine terms, a more involved approach to dealing with sign ambiguity is required (especially for the noisy case).

with the case where there is no noise.

Recovering V^* from the Tensor Decomposition in the Noiseless Case. Recall that the tensor power method provides us with row vectors v_i such that $\|v_i - \xi_i \boldsymbol{V}^*_{i,*}\|_2 \leqslant \varepsilon$ where $\varepsilon = O\left(\frac{1}{\text{poly}(d,m,k,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*))}\right)$ for $\xi_i \in \{\boldsymbol{V}^*_{i,*}, -\boldsymbol{V}^*_{i,*}\}$. Thus, the tensor power method gives us a *noisy* version of *either* $\boldsymbol{V}^*_{i,*}$ or $-\boldsymbol{V}^*_{i,*}$, however we do not know which. A priori, it would require 2^k time to guess the correct signs of the k vectors v_i. In this section, we show that using the combinatorial sparsity patterns in the row span of $\mathbf{A}$, we can not only recover the signs, but recover the matrix $\boldsymbol{V}^*$ *exactly*. Our procedure is detailed in Algorithm 5 below, which takes the outputs v_i from the tensor power method and returns the true matrix $\boldsymbol{V}^*$ up to a permutation of the rows.

Algorithm 5: Exact Recovery of $\boldsymbol{V}^*$

`Input :` Matrices $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$, and $v_i^T \in \mathcal{R}^d$ s.t. $\|v_i - \xi_i \boldsymbol{V}^*_{i,*}\|_2 \leqslant \varepsilon$ for some $\varepsilon = O\left(\frac{1}{\text{poly}(d,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*))}\right)$ for some unknown $\xi_i \in \{1,-1\}$ and each $i = 1, 2, \ldots, k$.

1. Let $\overline{\mathcal{X}} \in \mathcal{R}^{d \times \ell}$ be the first $\ell = \text{poly}\left(k, d, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*)\right)$ columns of $\mathcal{X}$, and let $\overline{\mathbf{A}} = \boldsymbol{U}^* f(\boldsymbol{V}^* \overline{\mathcal{X}})$.

2. Let $\tau = \Theta(1/\text{poly}(\ell))$ be a thresholding value. Define the row vectors $v_i^+, v_i^- \in \mathcal{R}^\ell$ via

$$(v_i^+)_j = \begin{cases} f(v_i \overline{\mathcal{X}})_j & \text{if } f(v_i \overline{\mathcal{X}})_j > \tau \\ 0 & \text{otherwise} \end{cases} \qquad v_i^- = \begin{cases} f(-v_i \overline{\mathcal{X}})_j & \text{if } f(-v_i \overline{\mathcal{X}})_j > \tau \\ 0 & \text{otherwise} \end{cases}$$

   for $j = 1, 2, \ldots, \ell$.

3. Let S_i^+ be the sign pattern of v_i^+, and S_i^- be the sign pattern of v_i^-. For $q \in \{+,-\}$, solve define the r linear systems of equations in the variable $w_i^q \in \mathcal{R}^k$, where the r-th system is given by

$$(w_i^q \overline{\mathbf{A}})_j = 0 \quad \text{for } j \notin S_i^q$$

$$(w_i^q)_r = 1$$

   Where $(w_i^q)_r$ is the r-th coordinate of (w_i^q). Then let (w_i^q) be the vector returned from the first linear system which had a solution.

4. Let q' be such that the above linear system returns a solution $w_i^{q'}$ with the constraints given by $S_i^{q'}$ (and at least one of the constraints of the form $(w_i^{q'})_r = 1$). We output **FAIL** if this occurs for both $q \in \{+,-\}$.

5. Output $\boldsymbol{V}_{i,*} = z_i / \|z_i\|_2$ where z_i is the solution to the following linear system.

$$\text{for all } j \in S_i^{q'} \quad (z_i \overline{\mathcal{X}})_j = (w_i^q \overline{\mathbf{A}})_j$$

`Output :` $\boldsymbol{V}$ such that $\boldsymbol{V} = \boldsymbol{V}^*$.

Before we proceed, we recall a standard fact about the singular values of random Gaussian matrices.

Lemma 6.4.18 (Corollary 5.35 [Ver10b])**.** *Let $\mathcal{S} \in \mathcal{R}^{k \times n}$ be a matrix of i.i.d. normal $\mathcal{N}(0,1)$*

random variables, with $k < 10n$. Then with probability $1 - 2e^{-n/8}$, for all row vectors $w \in \mathcal{R}^\ell$ we have

$$\sqrt{n}/3\|w\|_2 \leqslant \|w\mathcal{S}\|_2 \leqslant 2\sqrt{n}\|w\|_2$$

In other words, we have $\sqrt{n}/3 \leqslant \sigma_{\min}(\mathcal{S}) \leqslant \sigma_{\max}(\mathcal{S}) \leqslant 2\sqrt{n}$.

Theorem 133. *With high probability in d, m, Algorithm 5 does not fail, and finds $\boldsymbol{V}$ such that $\boldsymbol{V} = \boldsymbol{V}^*$.*

Proof. Fix a $i \in [k]$, and WLOG suppose the input row v_i is such that $\|v_i - \boldsymbol{V}^*_{i,*}\|_2 \leqslant \varepsilon$ (i.e. WLOG suppose $\xi_i = 1$). Then $\|v_i\overline{\mathcal{X}} - \boldsymbol{V}^*_{i,*}\overline{\mathcal{X}}\|_2 \leqslant O(1)\sqrt{\ell}\varepsilon$ by the operator norm bound of Lemma 6.4.18, and since f can only decrease the distance between matrices, it follows that $\|f(v_i\overline{\mathcal{X}}) - f(\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})\|_2 \leqslant O(1)\sqrt{\ell}\varepsilon$. Similarly, we have $\|f(-v_i\overline{\mathcal{X}}) - f(-\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})\|_2 \leqslant O(1)\sqrt{\ell}\varepsilon$.

We now condition on the event that none of the non-zero entries of $f(\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})$, and $f(-\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})$ are less than $\tau = \Theta(1/\text{poly}(\ell))$ (where τ is as in Algorithm 5), which holds by a union bound with probability $1 - 1/\text{poly}(\ell)$ (high prob in d, m) and the fact that the non-zero entries of these matrices have folded Gaussian marginals (distributed as the absolute value of a Gaussian). Given this, it follows that the sign patterns S_i^+ and S_i^- of v_i^+ and v_i^- are precisely the sign patterns of $f(\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})$ and $f(-\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})$ respectively. Since $f(\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})$ is in the row space of $\mathbf{A}$, at least one of the the linear systems run on S_i^+ will have a unique solution given by taking $w_i^+ = ce_i^T\boldsymbol{U}^{-1}$ for an appropriate constant $c \neq 0$ such that one of the constraints of the form $(w_i^+)_r = 1$ is satisfied, and where $\boldsymbol{U}^{-1}$ is the left inverse of $\boldsymbol{U}$.

Now consider the matrix $\boldsymbol{W}$ such that $\boldsymbol{W}$ is $\boldsymbol{V}^*$ with the row $-\boldsymbol{V}^*_{i,*}$ appended at the end. Then Applying the same argument as in Lemma 6.4.6, we see that the sign patterns of every pair of rows of $f(\boldsymbol{W}\overline{\mathcal{X}})$ disagrees on at least $\text{poly}(k)$ signs w.h.p.. This is easily seen for all pairs which contain one of $\{\boldsymbol{V}^*_{i,*}, -\boldsymbol{V}^*_{i,*}\}$ by applying the exact argument of the lemma and noting that the condition number of the matrix $\boldsymbol{V}^*$ does not change after negating the i-th row. The pair $\{f(\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}}), f(-\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})\}$ itself disagrees on all sign patterns, which completes the proof of the claim. Note here that disagree means that, for any two rows y^i, y^j in question there are at least $\text{poly}(k)$ coordinates such that *both* $y^i_p > 0$ and $y^j_p < 0$ *and* vice-versa. Thus no sparsity pattern is contained within any other. Then by Lemma 6.4.8, it follows that with high probability the only vector in the row span of $f(\boldsymbol{W}\overline{\mathcal{X}})$ which has a sparsity pattern contained within S_i^- is a scalar multiple of $f(-\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})$. Since no vector with such a sparsity pattern exists in the row span of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$, the linear system with constraints given by S_i^- will be infeasible with high probability.

We conclude from the above $q' = +$ in the fourth step of Algorithm 5, and that $w_i^{q'} = w_i^+$ is such that the sign pattern of $w_i^+ \overline{\mathbf{A}}$ is S_i^+, which is also the sign pattern of $f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})$. Since $w_i^+ \overline{\mathbf{A}}$ is in the row span of $f(\boldsymbol{V}^* \overline{\mathcal{X}})$, again by Lemma 6.4.8, we conclude that $w_i^+ \overline{\mathbf{A}} = cf(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})$ for some constant $c > 0$ (we can enfoce $c > 0$ by flipping the sign of w_i^+ so that $w_i^+ \overline{\mathbf{A}}$ has no strictly negative entries). The linear system in step 5 solves the equation $z_i \overline{\mathcal{X}}_{S_i^+} = w_i^+ \overline{\mathbf{A}}_{S_i^+}$, where $\overline{\mathcal{X}}_{S_i^+}$ is $\overline{\mathcal{X}}$ restricted to the columns corresponding to indices in S_i^+, and similarly with $\overline{\mathbf{A}}_{S_i^+}$. This will have a unique solution if $\overline{\mathcal{X}}_{S_i^+}$ has full row rank. Since an index is included in S_i^+ with probability $1/2$ independently, it follows that $|S_i^+| > \ell/3 > \text{poly}(d)$ with probability $1 - 2^{-\Omega(\ell)}$. A column of $\overline{\mathcal{X}}_{S_i^+}$ is just an i.i.d. Gaussian vector conditioned on being in a fixed half-space. Then if the first $i < d$ columns of $\overline{\mathcal{X}}_{S_i^+}$ are independent, they span a $i-1$ dimensional subspace. The Lebesgue measure of this subspace intersected with the halfspace has has measure 0, since the half-space is d-dimensional and the subspace is i-dimensional. It follows that the probability that the $i+1$ column of $\overline{\mathcal{X}}_{S_i^+}$ is in this subspace is 0, from which we conclude by induction that $\overline{\mathcal{X}}_{S_i^+}$ has rank d. Thus the solution z_i is unique, and must therefore be equal to $\frac{1}{c}\boldsymbol{V}_{i,*}^*$ as we also have $\boldsymbol{V}_{i,*}^* \mathcal{X}_{S_i^+} = c w_i^+ \overline{\mathbf{A}}$ for $c > 0$. After normalizing z_i to have unit norm, we conclude $\boldsymbol{V}_{i,*} = z_i / \|z\|_2 = \boldsymbol{V}_{i,*}^*$ as needed.

□

Given the results developed thus far, the correctness of our algorithm for the exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$ in the realizable (noiseless) case follows immediately. Recall we can always assume WLOG that $\|\boldsymbol{V}_{i,*}^*\|_2 = 1$ for all rows $i \in [k]$.

Theorem 134. *(Exact Recovery for Gaussian Input.) Suppose* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ *where* $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}, \boldsymbol{V}^* \in \mathcal{R}^{k \times d}$ *are both rank-*k*, and such that* $\mathcal{X} \in \mathcal{R}^{d \times n}$ *is i.i.d. Gaussian. Assume WLOG that* $\|\boldsymbol{V}_{i,*}^*\|_2 = 1$ *for all rows* $i \in [k]$*. If* $n = \Omega(\text{poly}(d, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*)))$*, then Algorithm 4 runs in* $\text{poly}(n)$*-time and recovers* $(\boldsymbol{U}^*)^T, \boldsymbol{V}^*$ *exactly up to a permutation of the rows w.h.p. (in* d, m*).*

Proof. By Theorem 130 and Corollary 6.4.17, we can recover $\boldsymbol{D}\boldsymbol{V}^*$ up to $\varepsilon = \frac{1}{\text{poly}(d,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*))}$ error in polynomial time, and then by Theorem 133 we can not only recover the signs ξ_i that constitute the diagonal of $\boldsymbol{D}$, but also recover $\boldsymbol{V}^*$ *exactly* (all in a polynomial number of arithmetic operations). Given the fact that $f(\boldsymbol{V}^* \mathcal{X})$ is full rank by Lemma 6.2.4, the solution $\boldsymbol{U}$ to the linear system $\boldsymbol{U} f(\boldsymbol{V}^* \mathcal{X}) = \mathbf{A}$ is unique, and therefore equal to $\boldsymbol{U}^*$. This linear system can be solved in polynomial time by Gaussian elimination, and thus the runtime does not depend on the bit-complexity of $\boldsymbol{U}^*, \boldsymbol{V}^*$ in the real RAM model. So the entire procedure runs in time polyno-

mial in the sample complexity n, which is polynomial in all relevant parameters as stated in the Theorem. □

6.4.3 Extension to Symmetric Input Distributions.

The independent and concurrent work of Ge et al. [GKLW18] demonstrates the existence of an algorithm that approximately recovers $\boldsymbol{U}^*$, $\boldsymbol{V}^*$ in polynomial time, given that the input $\mathcal{X}$ is drawn from a mixture of a symmetric probability distribution and a Gaussian. In this section, we observe how our techniques can be combined with the those of [GKLW18] to achieve exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$ for this broader class of distributions. Namely, that we can replace running the tensor decomposition algorithm from [JSA15] with the algorithm of [GKLW18] instead to obtain good approximations to $\boldsymbol{U}^*, \boldsymbol{V}^*$, and then use our results on the uniqueness of sparsity patterns in the row-span of $f(\boldsymbol{V}^*\mathcal{X})$ to obtain exactly recovery. Only minor changes are needed in the proofs of our sparsity pattern uniqueness results (Lemmas 6.4.6 and 6.4.8) to extend them to mixtures of symmetric distributions and Gaussians.

.

Definition 6.4.19. *(Symmetric Distribution.) Let $x \in \mathcal{R}^d$ be a vector random variable and $\mathcal{D}$ be a probability distribution function such that $x \sim \mathcal{D}$. Then, $\mathcal{D}$ is a symmetric distribution if for all x, the probability of x and $-x$ is equal, i.e. $\mathcal{D}(x) = \mathcal{D}(-x)$.*

Ge et. al. [GKLW18] define an object called the distinguishing matrix, denoted by $\mathbf{M}$, and require that the minimum singular value of $\mathbf{M}$ is bounded away from 0.

Definition 6.4.20. *(Distinguishing Matrix [GKLW18].) Given an input distribution $\mathcal{D}$ the distinguishing matrix is defined as $\mathbf{N}^{\mathcal{D}} \in \mathcal{R}^{d^2 \times \binom{k}{2}}$, whose columns are indexed by i, j such that $1 \leqslant i < j \leqslant k$ and*

$$\mathbf{N}^{\mathcal{D}}_{i,j} = \frac{1}{n} \sum_{k \in [n]} (\boldsymbol{V}^*_{i,*}\mathcal{X}_{*,k})(\boldsymbol{V}^*_{j,*}\mathcal{X}_{*,k})(\mathcal{X}_{*,k} \otimes \mathcal{X}_{*,k}) \mathbf{1}\left\{(\boldsymbol{V}^*_{i,*}\mathcal{X}_{*,k})(\boldsymbol{V}^*_{j,*}\mathcal{X}_{*,k}) \leqslant 0\right\}$$

Similarly an augmented distinguishing matrix $\mathbf{M}^{\mathcal{D}} \in \mathcal{R}^{d^2 \times \left(\binom{k}{2}+1\right)}$ has all the same columns as $\mathbf{N}^{\mathcal{D}}$ with the last column being $\frac{1}{n}\sum_{k \in [n]} \mathcal{X}_{*,k} \otimes \mathcal{X}_{*,k}$.

In order to bound the singular values of the distinguishing matrix, Ge et. al. consider input distributions that are perturbations of symmetric distributions. In essence, given a desired target

distribution $\mathcal{D}$, the algorithm of Ge et. al. can handle a similar distribution $\mathcal{D}_\gamma$, which is obtained by mixing $\mathcal{D}$ with a Gaussian with random covariance.

More formally, the perturbation is paramaterized by $\gamma \in (0,1)$, which will define the mixing rate. It is required that $\gamma > \frac{1}{\text{poly}(N)}$ in order to achieve polynomial running time (where $\text{poly}(N)$ is the desired running time of the algorithm). First, let $\boldsymbol{G}$ be an i.i.d. entry-wise $\mathcal{N}(0,1)$ random Gaussian matrix, which will be used to give the random the covariance. To generate $\mathcal{D}_\gamma$, first define a new distribution $\mathcal{N}'_{\boldsymbol{G}}$ as follows. To sample a point from $\mathcal{N}'_{\boldsymbol{G}}$, first sample a Gaussian $g \sim \mathcal{N}(0, \mathbf{I}_d)$ and then output $\boldsymbol{G}g$. Then the perturbation $\mathcal{D}_\gamma$ of the input distribution $\mathcal{D}$ is a mixture between $\mathcal{D}$ and $\mathcal{N}'_{\boldsymbol{G}}$. To sample $\mathcal{X}_{*,i}$ from $\mathcal{D}_\gamma$, pick z as a Bernoulli random variable where $\mathbf{Pr}[z=1] = \gamma$ and $\mathbf{Pr}[z=0] = 1-\gamma$, then for $i \in [n]$

$$\mathcal{X}_{*,i} \sim \begin{cases} \mathcal{D} & \text{if } z = 0 \\ \boldsymbol{G}g & \text{otherwise} \end{cases}$$

If the input is drawn from a mixture distribution $\mathcal{D}_\gamma$, $\sigma_{\min}(\mathbf{M})$ is bounded away from 0. We refer the reader to Section 2.3 in [GKLW18] for further details. We observe that we can extend the main algorithmic result therein with our results on exact recover to recover $\boldsymbol{U}^*$, $\boldsymbol{V}^*$ with zero-error in polynomial time.

Theorem 135. *(Informal Theorem 7 in [GKLW18].) Let $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ be full rank k such that $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$, f is ReLU, and for all $i \in [n]$ $\mathcal{X}_{*,i} \sim \mathcal{D}_\gamma$ as defined above. Let $\mathbf{M}$ be the distinguishing matrix as defined in [GKLW18]. For all $i \in [n]$, let Γ be such that $\|\mathcal{X}_{*,i}\|_2 \leqslant \Gamma$. Then, there exists an algorithm than runs in time poly $\left(\Gamma, 1/\varepsilon, 1/\delta, \|\boldsymbol{U}^*\|_2, \frac{1}{\sigma_{\min}(\mathbb{E}[\mathcal{X}_{*,i}\mathcal{X}_{*,i}^T])}, \frac{1}{\sigma_{\min}(\boldsymbol{U}^*)}, \frac{1}{\sigma_{\min}(\mathbf{M})}\right)$ and with probability $1-\delta$ outputs a matrix $\widehat{\boldsymbol{U}}$ such that $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}^*\boldsymbol{\Pi}\boldsymbol{D}\|_F \leqslant \varepsilon$.*

We use the algorithm corresponding to the aforementioned theorem to obtain an approximation $\widehat{\boldsymbol{U}}$ to $\boldsymbol{U}^*$, and then obtain an approximation to $f(\boldsymbol{V}^*\mathcal{X})$ by multiplying $\mathbf{A}$ on the left by $\widehat{\boldsymbol{U}}^{-1}$. The error in our approximation of $\boldsymbol{V}^*$ obtained via $\widehat{\boldsymbol{U}}^{-1}\mathbf{A}$ is analyzed in Section 6.4.1. Given this approximation of $\boldsymbol{V}^*$, we observe that running steps 4-8 of our Algorithm 3 recovers $\boldsymbol{V}^*, \boldsymbol{U}^*$ exactly (see Remark 136 below). Note that the only part of Algorithm 3 that required $\boldsymbol{V}^*$ to be orthonormal is step 3 which runs ICA, which we are replacing here with the algorithm of Theorem 135.

Here we remove the random matrix $\mathcal{T}$ from Algorithm 3, as it is not needed if we are already given an approximation $\widehat{\boldsymbol{U}}$ of $\boldsymbol{U}^*$. Thus we proceed exactly as in Algorithm 3 by restricting $\mathcal{X}, \mathbf{A}$ to $\ell = \text{poly}(d, m, k, \frac{1}{\gamma}, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*))$ columns $\overline{\mathcal{X}}, \overline{\mathbf{A}}$, and then rounding the entries of

$\widehat{f(\boldsymbol{V}\mathcal{X})} = \widehat{\boldsymbol{U}}^{-1}\overline{\mathbf{A}}$ below τ to 0. Finally, we solve the same linear system as in Algorithm 3 to recover the rows of $f(\boldsymbol{V}^*\mathcal{X})$ exactly, from which $\boldsymbol{V}^*$ and then $\boldsymbol{U}^*$ can be exactly recovered via solving the final two linear systems in Algorithm 3. We summarized this formally as follows.

Corollary 6.4.21. *(Exact Recovery for Symmetric Input.) Suppose* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ *where* $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ *are both rank-*k*, for all* $i \in [n]$*,* $\mathcal{X}_{*,i} \sim \mathcal{D}_\gamma$*, and* $\|\mathcal{X}_{*,i}\|_2 \leqslant \Gamma$*. Assume WLOG that* $\|\boldsymbol{V}^*_{i,*}\|_2 = 1$ *for all rows* $i \in [k]$*. If*

$$n \geqslant \textit{poly}\left(d, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*), \Gamma, \frac{1}{\gamma}, \|\boldsymbol{U}^*\|_2, \frac{1}{\sigma_{\min}(\mathbb{E}[\mathcal{X}_{*,i}\mathcal{X}_{*,i}^T])}, \frac{1}{\sigma_{\min}(\boldsymbol{U}^*)}, \frac{1}{\sigma_{\min}(\mathbf{M})}\right)$$

then there exists an algorithm that runs in poly(n)*-time and recovers* $(\boldsymbol{U}^*)^T, \boldsymbol{V}^*$ *exactly up to a permutation of the rows w.h.p. (in* d, m*).*

Remark 136. To prove the correctness of Algorithm 5 on $\mathcal{D}_\gamma$, we need only to generalize Lemmas 6.4.6 and 6.4.8 which together give the uniqueness of sparsity patterns of the rows of $f(\boldsymbol{V}^*\mathcal{X})$ in the rowspan of $\mathbf{A}$. We note that Lemma 6.4.6 can be easily generalized by first conditioning on the input being Gaussian, which in $\mathcal{D}_\gamma$ occurs with γ probability, and then going applying the same argument, replacing $\frac{1}{\kappa}$ with $\frac{\gamma}{\kappa}$ everywhere. The only change in the statement of Lemma 6.4.6 is that we now require $\ell = t\text{poly}(k, \kappa, \frac{1}{\gamma})$ to handle $\mathcal{X} \sim \mathcal{D}_\gamma$.

Next, the proof of Lemma 6.4.8 immediately goes through as the argument in the proof which demonstrates the determinant in question is non-zero only requires that the distribution $\mathcal{D}_\gamma$ is non-zero everywhere in the domain. Namely, the proof requires that the support of $\mathcal{D}_\gamma$ is all of $\mathcal{R}^d$. Note, this condition is always the case for the mixture $\mathcal{D}_\gamma$ since Gaussians are non-zero everywhere in the domain.

6.4.4 Necessity of poly$(\kappa(\boldsymbol{V}^*))$ Sample Complexity

So far, our algorithms for the exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$ have had polynomial dependency on the condition numbers of $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$. In this section, we make a step towards justifying the necessity of these dependencies. As always, we work without loss of generality under the assumption that $\|\boldsymbol{V}_{i,*}\|_2 = 1$ for all rows $i \in [k]$. Specifically, demonstrate the following.

Lemma 6.4.22. *Any algorithm which, when run on* $\mathbf{A}, \mathcal{X}$*, where that* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$*, and* $\mathcal{X}$ *has i.i.d. Gaussian* $\mathcal{N}(0,1)$ *entries, recovers* $(\boldsymbol{U}^*)^T, \boldsymbol{V}^*$ *exactly (up to a permutation of the rows) with probability at least* $1 - c$ *for some sufficiently small constant* $c > 0$*, requires* $n = \Omega(\kappa(\boldsymbol{V}^*))$

samples.

Proof. We construct two instances of $\mathbf{A}^1 = \boldsymbol{U}^1 f(\boldsymbol{V}^1 \mathcal{X})$ and $\mathbf{A}^2 = \boldsymbol{U}^1 f(\boldsymbol{V}^2 \mathcal{X})$. Let

$$\boldsymbol{U}^1 = \begin{bmatrix} \sqrt{1+a^2}/2 & \sqrt{1+a^2}/2 \end{bmatrix} \qquad \boldsymbol{V}^1 = \begin{bmatrix} \frac{1}{\sqrt{1+a^2}} & \frac{a}{\sqrt{1+a^2}} \\ \frac{1}{\sqrt{1+a^2}} & -\frac{a}{\sqrt{1+a^2}} \end{bmatrix}$$

$$\boldsymbol{U}^2 = \begin{bmatrix} \sqrt{1+(2a)^2}/2 & \sqrt{1+(2a)^2}/2 \end{bmatrix} \qquad \boldsymbol{V}^2 = \begin{bmatrix} \frac{1}{\sqrt{1+(2a)^2}} & \frac{a}{\sqrt{1+(2a)^2}} \\ \frac{1}{\sqrt{1+(2a)^2}} & -\frac{(2a)}{\sqrt{1+(2a)^2}} \end{bmatrix}$$

Now note that for $a \in [0,1]$, the rows of $\boldsymbol{V}^1$ have unit norm, and $\kappa(\boldsymbol{V}^1) = \frac{1}{a}$. Now let $a^i = ia$, and note, however, that for the j-th sample $\mathcal{X}_{*,j} = [x_1^j, x_2^j]$ i.i.d. Gaussian, we have for $i \in \{1,2\}$

$$\boldsymbol{U}^i f(\boldsymbol{V}^i \mathcal{X}_{*,j}) = \frac{f(x_1^j + a^i x_2^j) + f(x_1^j - a^i x_2^j)}{2}$$

Now note that when $|x_1^j| > (2a)|x_2^j|$, for *both* $i \in \{1,2\}$ we have either

$$\mathbf{A}^i_{*,j} = \boldsymbol{U}^i f(\boldsymbol{V}^i \mathcal{X}_{*,j}) = 0$$

or

$$\mathbf{A}^i_{*,j} = \boldsymbol{U}^i f(\boldsymbol{V}^i \mathcal{X}_{*,j}) = x_1^j$$

And in either case we do not get any information about a. In such a case, the j-th column of $\mathbf{A}^1$ and $\mathbf{A}^2$ are the same. In particular, conditioned on a given $\mathcal{X}$ such that $|x_1^j| > (2a)|x_2^j|$ for all columns j, we have $\mathbf{A}^1 = \mathbf{A}^2$. Now note that the probability that one Gaussian is $\frac{1}{2a}$ times larger than another is $\Theta(\frac{1}{a})$, thus any algorithm that takes less than $c\frac{1}{a}$ samples, for some absolute constant $c > 0$, cannot distinguish between $\mathbf{A}^1$ and $\mathbf{A}^2$, since we will have $\mathbf{A}^1 = \mathbf{A}^2$ with $\Omega(1)$ probability in this case, which completes the proof. □

6.5 A Polynomial Time Algorithm for Gaussian input and Sub-Gaussian Noise

In the last section, we gave two algorithms for exact recovery of $\boldsymbol{U}^*, \boldsymbol{V}^*$ in the noiseless (exact) case. Namely, where the algorithm is given as input $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ and $\mathcal{X}$. Our general algorithm for this problem first utilized a tensor decomposition algorithm which allowed for

approximate recovery of $\boldsymbol{V}^*$, up to the signs of its rows. Observe that this procedure, given by Theorem 130, can handle mean zero subgaussian noise $\mathbf{E}$, such that $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$. In this section, we will show how to utilize this fact as a sub-procedure to recover approximately recover $\boldsymbol{U}^*, \boldsymbol{V}^*$ in this noisy case.

We begin with using the algorithm corresponding to Theorem 130 to get an approximate solution to $\boldsymbol{V}^*$, up to permutations and ± 1 scaling. We note that the guarantees of Theorem 130 still hold when the noise $\mathbf{E}$ is sub-Gaussian. Therefore, we obtain a matrix $\widetilde{\boldsymbol{V}}$ such that $\|\widehat{\boldsymbol{V}} - \boldsymbol{D}\boldsymbol{\Pi}\boldsymbol{V}^*\|_F \leqslant \epsilon$, where $\boldsymbol{D}$ is a diagonal ± 1 matrix and $\boldsymbol{\Pi}$ is a permutation matrix.

Algorithm 6: Recovering Signs$(v_i, \mathbf{A}, \mathcal{X})$

`Input :` Matrices $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$, and $v_i^T \in \mathcal{R}^d$ s.t. $\|v_i - \xi_i \boldsymbol{V}^*_{i,*}\|_2 \leqslant \varepsilon$ for some unknown $\xi_i \in \{1, -1\}$ and $i = 1, 2, \ldots, k$, where $\varepsilon = O(\frac{1}{\text{poly}(d,m,k,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*)}$.

1. Let $\overline{\mathcal{X}} \in \mathcal{R}^{d \times \ell}$ be the first $\ell = \text{poly}(k, d, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*), \sigma)$ columns of $\mathcal{X}$, and similarly define $\overline{\mathbf{E}}$, and let $\overline{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \overline{\mathcal{X}}) + \overline{\mathbf{E}}$.
2. For $i \in [k]$, let
$$S_{i,+} = \{f(v_j \overline{\mathcal{X}}), f(-v_j \overline{\mathcal{X}})\}_{j \neq i} \cup \{f(v_i \overline{\mathcal{X}})\}$$
and
$$S_{i,-} = \{f(v_j \overline{\mathcal{X}}, f(-v_j \overline{\mathcal{X}})\}_{j \neq i} \cup \{f(-v_i \overline{\mathcal{X}})\}$$
3. Let $\mathbf{P}_{S_{i,+}}$ be the orthogonal projection matrix onto the row span of vectors in $S_{i,+}$. Compute
$$a^+_{i,j} = \|\overline{\mathbf{A}}_{j,*}(\mathbb{I} - \mathbf{P}_{S_{i,+}})\|_2^2$$
$$a^-_{i,j} = \|\overline{\mathbf{A}}_{j,*}(\mathbb{I} - \mathbf{P}_{S_{i,-}})\|_2^2$$
For each $j \in [m]$.
4. Let $a_i^+ = \sum_j a^+_{i,j}$, and $a_i^- = \sum_j a^-_{i,j}$. If $a_i^+ < a_i^-$, set $\boldsymbol{V}_{i,*} = v_i$, otherwise set $\boldsymbol{V}_{i,*} = -v_i$.

`Output :` $\boldsymbol{V}$ such that $\|\boldsymbol{V} - \boldsymbol{V}^*\|_2 \leqslant \varepsilon$, thus recovering ξ_i for $i \in [k]$.

Recall that in the noiseless case, we needed to show that given a approximate version of $\boldsymbol{V}^*$ up to the signs of the rows, we can recover both the signs and $\boldsymbol{V}^*$ exactly in polynomial time. Formally, we were given rows v_i such that $\|v_i - \xi_i \boldsymbol{V}^*_{i,*}\|_2$ was small for some $\xi_i \in \{1, -1\}$, however we did not know ξ_i. This issue is a non-trivial one, as we cannot simply guess the ξ_i's

(there are 2^k possibilities), and moreover we cannot assume WLOG that the ξ_i's are 1 by pulling the scaling through the ReLU, which is only commutes with *positive* scalings. Our algorithm for recovery of the true signs ξ_i in the exact case relied on combinatorial results about the sparsity patterns of $f(\boldsymbol{V}^*\mathcal{X})$. Unfortunately, these combinatorial results can no longer be used as a black-box in the noisy case, as the sparsity patterns can be arbitrarily corrupted by the noise. Thus, we must develop a refined, more general algorithm for the recovery of the signs ξ_i in the noise case. Thus we begin by doing precisely this.

6.5.1 Recovering the Signs ξ_i with Subgaussian Noise

Lemma 6.5.1. *Let $g \in \mathcal{R}^n$ be a row vector of i.i.d. mean zero variables with variance σ, and let $\mathcal{S}$ be any fixed k dimensional subspace of $\mathcal{R}^n$. Let $\mathbf{P}_{\mathcal{S}} \in \mathcal{R}^{n\times n}$ be the projection matrix onto $\mathcal{S}$. Then for any $\delta > 0$ with probability $1-\delta$, we have*

$$\|g\mathbf{P}_{\mathcal{S}}\|_2 = \sigma\sqrt{k/\delta}$$

Proof. We can write $P_{\mathcal{S}} = \boldsymbol{W}^T\boldsymbol{W}$ for matrices $\boldsymbol{W} \in \mathcal{R}^{k\times n}$ with orthonormal rows. Then $\mathbb{E}\left[\|g\boldsymbol{W}^T\|_2^2\right] = \sigma^2 k$, and by Markov bounds with probability $1-\delta$ we have $\|g\boldsymbol{W}^T\|_2^2 = \|g\boldsymbol{W}^T\boldsymbol{W}\|_2^2 < \sigma^2 k/\delta$ as needed.

□

Lemma 6.5.2. *Let $\mathbf{Q} \in \mathcal{R}^{k\times\ell}$ be a matrix of row vectors for $\ell > poly(k)$ (for some sufficiently large polynomial) with $1 \leqslant \|\mathbf{Q}\|_2$ and let $\mathbf{P}_{\mathbf{Q}} = \mathbf{Q}^T(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}$ be the projection onto them. Let $\mathbf{E}$ be such that $\|\mathbf{E}\|_F \leqslant \frac{\varepsilon}{\left(\kappa(\mathbf{Q})\|\mathbf{Q}\|_2\right)^4}$, and let $\mathbf{P}_{\mathbf{Q}+\mathbf{E}}$ be the projection onto the rows of $\mathbf{Q}+\mathbf{E}$.*
Then for any vector $x^T \in \mathcal{R}^\ell$, we have

$$\|x\mathbf{P}_{\mathbf{Q}+\mathbf{E}}\|_2 = \|x\mathbf{P}_{\mathbf{Q}}\|_2 \pm O(\varepsilon\|x\|_2)$$

Proof. We have $\mathbf{P}_{\mathbf{Q}+\mathbf{E}} = (\mathbf{Q}+\mathbf{E})^T((\mathbf{Q}+\mathbf{E})^T(\mathbf{Q}+\mathbf{E}))^{-1}(\mathbf{Q}+\mathbf{E})$. Now

$$(\mathbf{Q}+\mathbf{E})^T(\mathbf{Q}+\mathbf{E}) = \mathbf{Q}^T\mathbf{Q} + \mathbf{E}^T\mathbf{Q} + \mathbf{Q}^T\mathbf{E} + \mathbf{E}^T\mathbf{E}$$

Further, $\|\mathbf{E}^T\mathbf{Q} + \mathbf{Q}^T\mathbf{E} + \mathbf{E}^T\mathbf{E}\|_F \leqslant \|\mathbf{E}\|_F\|\mathbf{Q}\|_2 + \|\mathbf{E}\|_F^2 \leqslant 2\frac{\varepsilon}{\kappa^4(\mathbf{Q})\|\mathbf{Q}\|_2^2}$. Thus we can write $(\mathbf{Q}+\mathbf{E})^T(\mathbf{Q}+\mathbf{E}) = \mathbf{Q}^T\mathbf{Q} + \mathbf{Z}$ where $\|\mathbf{Z}\|_F \leqslant 2\frac{\varepsilon}{\kappa^4(\mathbf{Q})\|\mathbf{Q}\|_2^2}$. Applying Corollary 6.4.10 with $\mathbf{B} = \mathbf{Q}^T\mathbf{Q}$, and $\mathbf{E} = \mathbf{Z}$, we can write $(\mathbf{Q}+\mathbf{E})^T(\mathbf{Q}+\mathbf{E}))^{-1} = (\mathbf{Q}^T\mathbf{Q})^{-1} + \mathbf{Z}'$, where $\|\mathbf{Z}'\|_F \leqslant$

$O(\frac{\varepsilon}{\kappa^2(\mathbf{Q})\|\mathbf{Q}\|_2^2})$. Thus

$$\begin{aligned}\mathbf{P}_{\mathbf{Q}+\mathbf{E}} &= (\mathbf{Q}+\mathbf{E})^T((\mathbf{Q}^T\mathbf{Q})^{-1}+\mathbf{Z}')(\mathbf{Q}+\mathbf{E}) \\ &= \mathbf{P}_{\mathbf{Q}} + \mathbf{Q}^T\mathbf{Z}'(\mathbf{Q}+\mathbf{E}) + \mathbf{Q}^T(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{E} + \mathbf{E}^T((\mathbf{Q}^T\mathbf{Q})^{-1}+\mathbf{Z}')(\mathbf{Q}+\mathbf{E})\end{aligned}$$

Therefore,

$$\begin{aligned}\|\mathbf{Q}^T\mathbf{Z}'(\mathbf{Q}+\mathbf{E})\|_F &\leqslant \|\mathbf{Q}^T\|_2\|\mathbf{Z}'(\mathbf{Q}+\mathbf{E})\|_F \\ &\leqslant \|\mathbf{Q}^T\|_2\|\mathbf{Z}'\|_F\|(\mathbf{Q}+\mathbf{E})\|_2 \\ &\leqslant \|\mathbf{Q}^T\|_2\|\mathbf{Z}'\|_F(\|\mathbf{Q}\|_2+\|\mathbf{E}\|_F) \\ &= O\left(\frac{\varepsilon}{\kappa^2(\mathbf{Q})}\right) = O(\varepsilon)\end{aligned}$$

Next, we have

$$\begin{aligned}\|\mathbf{Q}^T(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{E}\|_F &\leqslant \|\mathbf{Q}^T\|_2\|(\mathbf{Q}^T\mathbf{Q})^{-1}\|_2\|\mathbf{E}\|_F \\ &\leqslant \frac{\varepsilon}{\|\mathbf{Q}\|_2^3} \\ &< \varepsilon\end{aligned}$$

where in the second to last inequality we used the fact that $\|\mathbf{Q}\|_2 > 1$ so $\sigma_{\min}^{-2}(\mathbf{Q}) = \|(\mathbf{Q}^T\mathbf{Q})^{-1}\|_2 < 1/\kappa^2(\mathbf{Q})$. Applying the above bounds similarly, we have $\|\mathbf{E}^T(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}\|_F \leqslant O(\varepsilon)$, $\|\mathbf{E}^T(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{E}\|_F \leqslant O(\varepsilon)$, and $\|\mathbf{E}^T\mathbf{Z}'(\mathbf{Q}+\mathbf{E})\|_F \leqslant O(\varepsilon)$. We conclude $\mathbf{P}_{\mathbf{Q}+\mathbf{E}} = \mathbf{P}_{\mathbf{Q}} + \mathbf{Z}''$, where $\|\mathbf{Z}''\|_F \leqslant O(\varepsilon)$. It follows that for any $x \in \mathcal{R}^\ell$, we have

$$\begin{aligned}\|x\mathbf{P}_{\mathbf{Q}+\mathbf{E}}\|_2 &= \|x\mathbf{P}_{\mathbf{Q}}\|_2 \pm \|x\mathbf{Z}''\|_2 \\ &= \|x\mathbf{P}_{\mathbf{Q}}\|_2 \pm O(\varepsilon\|x\|_2)\end{aligned}$$

□

Lemma 6.5.3. *Let $\mathbf{Q} \in \mathcal{R}^{r\times\ell}$ for $1 \leqslant r \leqslant 2k$ be any matrix whose rows are formed by taking r distinct rows from the set $\{f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}}), f(-\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}})\}_{i\in[k]}$, where $\overline{\mathcal{X}}$ is $\mathcal{X}$ restricted to the first $\ell = \text{poly}(k, d, m, \kappa(\boldsymbol{V}^*))$ columns. Then w.h.p. (in ℓ), both $\|\mathbf{Q}\|_F^2 \leqslant 10r\ell$ and $\sigma_{\min}(\mathbf{Q}) = \Omega(\frac{\sqrt{\ell}}{(\kappa(\boldsymbol{V}^*))^2})$.*

Proof. The first bound $\|\mathbf{Q}\|_F^2 \leqslant 10r\ell$ follows from the fact that the $\|\cdot\|_2^2$ norm of each row is distributed as a χ^2 random variable, so the claim follows from standard tail bounds for such variables [LM00]. For the second claim, write $\mathbf{Q} = f(\boldsymbol{W}\overline{\mathcal{X}})$, where the rows of $\boldsymbol{W}$ are the

r distinct rows from the set $\{\boldsymbol{V}^*_{i,*}, -\boldsymbol{V}^*_{i,*}\}_{i\in[k]}$ corresponding to the rows of $\mathbf{Q}$. Let $\boldsymbol{W}^+$ be the subset of rows of the form $\boldsymbol{V}^*_{i,*}$, and $\boldsymbol{W}^-$ its complement. There now there is a rotation matrix $\mathbf{R}$ that rotates $\boldsymbol{W}^+$ to be lower triangular, so that the j-th row of $\boldsymbol{W}^+\mathcal{R}$ is supported on the first j columns. Let $\boldsymbol{W}$ be such that the pairs of rows $\{\boldsymbol{V}^*_{i,*}, -\boldsymbol{V}^*_{i,*}\}$ with the same index i are placed together. Then $\boldsymbol{W}\mathcal{R}$ is block-upper triangular, where the j-th pair of rows of the form $\{\boldsymbol{V}^*_{i,*}, -\boldsymbol{V}^*_{i,*}\}$ are supported on the first j columns. Since Gaussians are rotationally invariant, $\boldsymbol{W}\mathbf{R}\overline{\mathcal{X}}$ has the same distribution as $\boldsymbol{W}\overline{\mathcal{X}}$, thus we can assume that $\boldsymbol{W}$ is in this block lower triangular form, and $\boldsymbol{V}^*$ is in lower triangular form.

WLOG assume the rank of $\boldsymbol{W}$ is k (the following arguements will hold when the rank is $k' < k$). We now claim that we can write $\boldsymbol{W}_{r,*} = \alpha + \varphi e_k$, where α is in the span of $e_1, \ldots, e_{k-1}$ and $\varphi = \Omega(\frac{1}{\kappa})$ where $\kappa = \kappa(\boldsymbol{V}^*)$. To see this, note that if this were not the case, the projection of $\boldsymbol{W}_{r,*}$ onto the all prior rows with the same sign (i.e. all either of the form $\boldsymbol{V}^*_{i,*}$ or $-\boldsymbol{V}^*_{i,*}$) would be less than $\frac{1}{\kappa}$, since the prior span all of $\mathcal{R}^{k-1}$ on the first $k-1$ columns, and the only part of $\boldsymbol{W}_{r,*}$ outside of this span has weight φ. Let w be such that $w\boldsymbol{W}'$ is this projection, where $\boldsymbol{W}'$ excludes the last row of $\boldsymbol{W}$ WLOG this row is of the form $\boldsymbol{V}^*_{i,*}$, and WLOG $i = k$. Then can write $w\boldsymbol{W}' = v(\boldsymbol{V}^*)'$ where $(\boldsymbol{V}^*)'$ excludes the last row of $\boldsymbol{V}^*$. Then $\|[v,-1]\boldsymbol{V}^*\|_2 < \frac{1}{\kappa}$, and since $\|\boldsymbol{V}^*\|_2 \geqslant 1$ it follows that $\kappa(\boldsymbol{V}^*) > \kappa$, a contradiction since κ is defined as the condition number of $\boldsymbol{V}^*$.

Now let x^i be the i-th column of $\overline{\mathcal{X}}$, and let $\mathcal{E}_i$ be the event that $\varphi x^i_k > \lambda|\langle\alpha, (x^i_1, \ldots, x^i_{k-1})\rangle|$, where λ will later be set to be a sufficiently large constant. Since φx^i_k and $\langle\alpha, (x^i_1, \ldots, x^i_{k-1})\rangle$ are each distributed as a Gaussian with variance at most 1, by anti-concentration of Gaussians $\mathbf{Pr}\mathcal{E}_i = \Omega(\frac{1}{\lambda\kappa})$. So let $\mathcal{S} \subset [\ell]$ be the subset of i such that $\mathcal{E}_i$ holds, and let $\mathcal{X}_\mathcal{S}$ be $\overline{\mathcal{X}}$ restricted to this subset. Let $\boldsymbol{V}^*_{i,*}$ be the last column of $\boldsymbol{W}$ (WLOG we assume it is $\boldsymbol{V}^*_{i,*}$ and not $-\boldsymbol{V}^*_{i,*}$). We now upper bound the norm of the projection of $f(\boldsymbol{V}^*_{i,*}\mathcal{X}_\mathcal{S})$ onto the row span of $f(\boldsymbol{W}\mathcal{X}_\mathcal{S})$. Now $f(-\boldsymbol{V}^*_{i,*}\mathcal{X})$, if it exists as a row of $f(\boldsymbol{W}\mathcal{X})$, will be identically 0 on the coordinates in $\mathcal{S}$ (because $\boldsymbol{V}^*_{i,*}\mathcal{X}$ is positive on these coordinates by construction). So we can disregard it in the following analysis. By construction of $\mathcal{S}$ we can write

$$f(\boldsymbol{V}^*_{i,*}\mathcal{X}_\mathcal{S}) = f(\varphi(\mathcal{X}_\mathcal{S})_{k,*}) + b$$

where $\|b\|_2 \leqslant \|\frac{1}{\lambda}f(\boldsymbol{V}^*_{i,*}\mathcal{X}_\mathcal{S})\|_2$, where $(\mathcal{X}_\mathcal{S})_{k,*}$ is the k-th row of $\mathcal{X}_\mathcal{S}$. By the triangle inequality, the projection of $f(\boldsymbol{V}^*_{i,*}\mathcal{X}_\mathcal{S})$ onto the rowspan of $f(\boldsymbol{W}'\mathcal{X}_\mathcal{S})$ (where $\boldsymbol{W}'$ is $\boldsymbol{W}$ excluding $\boldsymbol{V}^*_{i,*}$), is

$$\|f(\boldsymbol{V}^*_{i,*}\mathcal{X}_\mathcal{S})\mathbf{P}_{\boldsymbol{W}'}\|_2 \leqslant \frac{1}{\lambda}\|f(\boldsymbol{V}^*_{i,*}\mathcal{X}_\mathcal{S})\|_2 + \|f(\varphi(\mathcal{X}_\mathcal{S})_{k,*})\mathbf{P}_{\boldsymbol{W}'}\|_2$$

where $\mathbf{P}_{\boldsymbol{W}'}$ is the projection onto the rowspan of $f(\boldsymbol{W}'\mathcal{X}_{\mathcal{S}})$. Crucially, observe that $f(\boldsymbol{W}'\mathcal{X}_{\mathcal{S}})$, and thus $\mathbf{P}_{\boldsymbol{W}'}$, does not depend on the k-th row $(\mathcal{X}_{\mathcal{S}})_{k,*}$ of $\mathcal{X}_{\mathcal{S}}$. Now

$$\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\mathbf{P}_{\boldsymbol{W}'}\|_2 \leqslant \|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\mathbf{P}_{\boldsymbol{W}'+1}\|_2$$

where $\mathbf{P}_{\boldsymbol{W}'+1}$ is the projection onto the row span of $\{f(\boldsymbol{W}'\mathcal{X}_{\mathcal{S}})_{j,*}\}_{\text{rows } j \text{ of } \boldsymbol{W}'} \bigcup \{\mathbf{1}\}$ where $\mathbf{1}$ is the all $1's$ vector. This holds since adding a vector to the span of the subspace being projected onto can only increase the length of the projection. Let $\mathbf{P}_1$ be the projection just onto the row $\mathbf{1}$.

Now observe that $\varphi f((\mathcal{X}_{\mathcal{S}})_{k,*})(\mathbb{I} - \mathbf{P}_1)$ is a mean 0 i.i.d. shifted rectified-Gaussian vector with variance strictly less than φ^2 (here rectified means 0 with prob $1/2$ and positive Gaussian otherwise). Moreover, the mean of the entries of $f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})$ is $\Theta(\varphi)$. The L_2 of these vectors are thus sums of sub-exponential random variables, so by standard sub-exponential concentration (see e.g. [Wai19]) we have $\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})(\mathbb{I} - \mathbf{P}_1)\|_2 = \Theta(\varphi)\sqrt{|S|}$, and moreover

$$\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\|_2 - \|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})(\mathbb{I} - \mathbf{P}_1)\|_2 = \Omega(\varphi)\sqrt{|\mathcal{S}|} \tag{6.2}$$

w.h.p in $\log(|\mathcal{S}|)$ where $|S| \geqslant \Theta(1)\frac{\ell}{\kappa\lambda} = \text{poly}(d, m, k, \kappa)$. Now by Lemma 6.5.1, we have

$$\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})(\mathbb{I} - \mathbf{P}_1)\mathbf{P}_{\boldsymbol{W}'+1}\|_2 \leqslant \varphi\sqrt{(2k+1)/\delta}$$

with probability $1 - \delta$, for some $\delta = 1/\text{poly}(k, d, m)$. So

$$\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})(\mathbb{I} - \mathbf{P}_1)\mathbf{P}_{\boldsymbol{W}'+1}\|_2 \leqslant O(\frac{\text{poly}(k, d, m)}{\sqrt{|\mathcal{S}|}})\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})(\mathbb{I} - \mathbf{P}_1)\|_2$$

Write $f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*}) = f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\mathbf{P}_{+1} + f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})(\mathbb{I} - \mathbf{P}_1)$. Then by triangle inequality, we can upper bound $\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\mathbf{P}_{\boldsymbol{W}'+1}\|_2$ by

$$\begin{aligned}
&\leqslant \|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\mathbf{P}_1\mathbf{P}_{\boldsymbol{W}'+1}\|_2 + \|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})(\mathbb{I} - \mathbf{P}_1)\mathbf{P}_{\boldsymbol{W}'+1}\|_2 \\
&\leqslant \|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\mathbf{P}_1\|_2 + O(\frac{\text{poly}(k, d, m)}{\sqrt{|S|}})\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})(\mathbb{I} - \mathbf{P}_1)\|_2 \\
&= \left(\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\|_2^2 - \|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})(\mathbb{I} - \mathbf{P}_1)\|_2^2\right)^{1/2} + O(\frac{\text{poly}(k, d, m)}{\sqrt{|S|}})\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\|_2
\end{aligned}$$

Using the bound from Equation 6.2, for some constants $c, c' < 1$ bounded away from 1, we have

$$\begin{aligned}
&= \|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\|_2(1-c) + O(\frac{\text{poly}(k,d,m)}{\sqrt{|S|}})\|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\|_2 \\
&\leqslant \|f(\varphi(\mathcal{X}_{\mathcal{S}})_{k,*})\|_2(1-c')
\end{aligned}$$

Thus $\|f(\boldsymbol{V}_{i,*}^*\mathcal{X}_{\mathcal{S}})\mathbf{P}_{\boldsymbol{W}'}\|_2 \leqslant \|f(\boldsymbol{V}_{i,*}^*\mathcal{X}_{\mathcal{S}})\mathbf{P}_{\boldsymbol{W}'+1}\|_2 \leqslant (1-\Theta(1))\|f(\boldsymbol{V}_{i,*}^*\mathcal{X}_{\mathcal{S}})\|_2$. Now by setting $\lambda > 2c'$ a sufficently large constant, the b term becomes negligible, and the above bound holds replacing c' with $c'/2$. Since we have $\|f(\boldsymbol{V}_{i,*}^*\mathcal{X}_{\mathcal{S}})\|_2 = \Theta(\varphi\sqrt{|\mathcal{S}|})$, and $\|f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}})\|_2 = \Theta(\sqrt{\ell})$, if $\overline{\mathbf{P}_{\boldsymbol{W}'}}$ is projection of onto the rows of $f(\boldsymbol{W}'\overline{X})$, we have

$$\begin{aligned}
\|f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}})\overline{\mathbf{P}_{\boldsymbol{W}'}}\|_2 &\leqslant \|f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}})\|_2\left(1-\Theta\left(\frac{\varphi}{\kappa\lambda}\right)\right) \\
&< \|f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}})\|_2(1-\Theta(\frac{1}{\kappa^2}))
\end{aligned}$$

Where here we recall $\lambda = \Theta(1)$. Since this argument used no facts about the row i we were choosing, it follows that the norm of the projection of any row onto the subspace spanned by the others others in $f(\boldsymbol{W}\overline{\mathcal{X}})$ is at most a $(1-\Theta(\frac{1}{\kappa^2}))$ factor less than the norm was before the projection. In particular, this implies that $f(\boldsymbol{W}\overline{\mathcal{X}})$ is full rank. Note by sub-exponential concentration, each row norm is $\Theta(\sqrt{\ell})$ w.h.p. We are now ready to complete the argument. Write $f(\boldsymbol{W}\overline{\mathcal{X}}) = \mathbf{B}\Sigma\mathbf{Q}^T$ in its singular value decomposition. Since the projection of one row onto another does not change by a row rotation,, we can rotate $\mathbf{Q}^T$ to be the identity, and consider $\mathbf{B}\Sigma$. Let u_i be a unit vector in the direction of the i-th row projected onto the orthogonal space to the prior rows. Now for any unit vector u, write it as $u = \sum_i u_i a_i$ (which we can do because $f(\boldsymbol{W}\overline{\mathcal{X}})$ is full rank). Noting that $\frac{\|f(\boldsymbol{W}\overline{\mathcal{X}})_{i,*}\|_2}{\|f(\boldsymbol{W}\mathcal{X}_{\mathcal{S}})_{i,*}\|_2} = O(\text{poly}(k)\kappa^2)$ for any row i, we have

$$\begin{aligned}
\|f(\boldsymbol{W}\overline{\mathcal{X}})u\|_2^2 &\geqslant \sum_i \langle u_i, u\rangle^2\Omega(\frac{\ell}{\kappa^4}) \\
&\geqslant \sum_i a_i^2\Omega(\frac{\ell}{\kappa^4}) \\
&= \Omega(\frac{\ell}{\kappa^4})
\end{aligned}$$

Thus $\sigma_{\min}(\mathbf{Q}) = \sigma_{\min}(f(\boldsymbol{W}\overline{\mathcal{X}})) = \Omega(\frac{\sqrt{\ell}}{\kappa^2})$ as needed, where recall we have been writing $\kappa = \kappa(\boldsymbol{V}^*)$. □

Using the bounds developed within the proof of the prior lemma gives the following corollary.

Corollary 6.5.4. *Let $\mathbf{P}_{S_{i,+}}, \mathbf{P}_{S_{i,-}}$ be as in Algorithm 6. Then*

$$\|f(\boldsymbol{V}_{i,*}^*\mathcal{X})\mathbf{P}_{S_{i,-}}\|_2 = \|f(\boldsymbol{V}_{i,*}^*\mathcal{X})\|_2\Big(1-\Omega(\frac{1}{\kappa(\boldsymbol{V}^*)^2 poly(k)})\Big)$$

and

$$\|f(-\boldsymbol{V}_{i,*}^*\mathcal{X})\mathbf{P}_{S_{i,+}}\|_2 = \|f(-\boldsymbol{V}_{i,*}^*\mathcal{X})\|_2\Big(1-\Omega(\frac{1}{\kappa(\boldsymbol{V}^*)^2 poly(k)})\Big)$$

Theorem 137. *Let $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X}) + \mathbf{E}$, where $\mathbf{E}$ is i.i.d. mean zero with variance σ^2. Then given $v_i^T \in \mathcal{R}^d$ such that. $\|v_i - \xi_i \boldsymbol{V}_{i,*}^*\|_2 \leqslant \varepsilon$ for some unknown $\xi_i \in \{1,-1\}$ and $i = 1,2,\ldots,k$, where $\varepsilon = O(\frac{1}{poly(d,m,k,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*))})$ is sufficiently small, with high probability, Algorithm 6 returns $\boldsymbol{V}$ such that $\|\boldsymbol{V} - \boldsymbol{V}^*\|_2 \leqslant \varepsilon$ in poly$(d,m,k,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*),\sigma)$ time.*

Proof. Consider a fixed $i \in [k]$, and WLOG assume $\xi_i = 1$, so we have $\|v_i - \boldsymbol{V}_{i,*}^*\|_2 \leqslant \varepsilon = O(\frac{1}{\text{poly}(d,m,k,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*))})$. We show $a_i^+ < a_i^-$ with high probability. Now fix a row j of $\overline{\mathbf{A}} = \boldsymbol{U}^* f(\boldsymbol{V}^*\overline{\mathcal{X}}) + \overline{\mathbf{E}}$ as in Algorithm 6, where $\overline{\mathbf{Q}}$ refers to restricting to the first $\ell = \text{poly}(k,d,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*),\sigma)$ columns of a matrix $\mathbf{Q}$. Note that we choose ε so that $\varepsilon < 1/\text{poly}(\ell)$ (which is achieved by taking n sufficiently large). This row is given by $\overline{\mathbf{A}}_{j,*} = \boldsymbol{U}_{j,*}^* f(\boldsymbol{V}^*\overline{\mathcal{X}}) + \overline{\mathbf{E}}_{j,*}$. As in the proof of Theorem 133, using Lemma 6.4.18 to bound $\|\overline{\mathcal{X}}\|_2$, we have $\|f(v_i\overline{\mathcal{X}}) - f(\xi\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}})\|_2 \leqslant O(\varepsilon\sqrt{\ell})$. We now using Lemma 6.5.2 to bound the projection difference between using approximate projection matrix $\mathbf{P}_{S_{i,+}}$ formed by our approximate vectors $f(v_i\overline{\mathcal{X}})$, and the true projection matrix $\mathbf{P}_{S_{i,+}}^*$ formed by the vectors $f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}})$. By Lemma 6.5.3, the condition number and the spectral norm of the matrix formed by the rows that span $\mathbf{P}_{S_{i,+}}^*$ are at most $O(r\text{poly}(k)\kappa^2)$ and $O(r\ell) = \text{poly}(k,d,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*),\sigma)$ respectively, w.h.p. (in k,d,m). Setting $\varepsilon = \varepsilon'/\text{poly}(k,d,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*),\sigma)$ sufficiently small, Lemma 6.5.2 gives $\|x\mathbf{P}_{S_{i,+}}\|_2 = \|x\mathbf{P}_{S_{i,+}}^*\|_2 \pm O(\varepsilon'\|x\|_2)$ for $\varepsilon' = \frac{1}{\text{poly}(k,d,m,\kappa(\boldsymbol{V}^*),\kappa(\boldsymbol{U}^*),\sigma}$ and any vector x.

Now we have $a_{i,j}^+ = (\|\overline{\mathbf{E}_{j,*}}(\mathbb{I}-\mathbf{P}_{S_{i,+}})\|_2 \pm O(\varepsilon'\sigma\sqrt{\ell}))^2 \leqslant \|\overline{\mathbf{E}_{j,*}}\|_2^2 + \pm O(\sigma^2\varepsilon'\ell\text{poly}(k,d,m))$. Here we used that $\|\mathbf{E}\|_2^2 \leqslant \sigma^2\ell\text{poly}(k,d,m)$, w.h.p. in k,d,m (by Chebyshev's inequality), and the fact that w.h.p. we have $\|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}}) + \overline{\mathbf{E}_{j,*}})\|_2 = O(\sigma\sqrt{\ell})$. Then, setting $\varepsilon' = \varepsilon''/\text{poly}(\ell)$, we have

$$\begin{aligned}
(a_{i,j}^-)^2 &= \|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}}) + \overline{\mathbf{E}_{j,*}})(\mathbb{I}-\mathbf{P}_{S_{i,-}})\|_2^2 \pm O(\varepsilon'') \\
&= \|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}}) + \overline{\mathbf{E}_{j,*}})\|_2^2 - \|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}}) + \overline{\mathbf{E}_{j,*}})\mathbf{P}_{S_{i,-}}\|_2^2 \pm O(\varepsilon'') \\
&\geqslant \|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}}) + \overline{\mathbf{E}_{j,*}})\|_2^2 - \Big(\|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^*\overline{\mathcal{X}})\mathbf{P}_{S_{i,-}}\|_2 + \|\|\overline{\mathbf{E}_{j,*}}\mathbf{P}_{S_{i,-}}\|_2\Big)^2 \pm O(\varepsilon'')
\end{aligned}$$

where the second equality follows by the Pythagorean Theorem. Applying Lemma 6.5.1 and

Corollary 6.5.4, writing $\kappa = \kappa(\boldsymbol{V}^*)$, with probability $1-\delta$ we have

$$(a_{i,j}^-)^2 \geqslant \|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}}) + \overline{\mathbf{E}_{j,*}})\|_2^2 \\ - \left(\|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2 (1 - \Omega(\frac{1}{\kappa^2 \text{poly}(k)})) + 2\sigma\sqrt{k/\delta}\right)^2 \pm O(\varepsilon'')$$

Setting $\delta < 1/\text{poly}(k,d,m)$ to get high probability gives

$$\begin{aligned}(a_{i,j}^-)^2 &\geqslant \|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}}) + \overline{\mathbf{E}_{j,*}})\|_2^2 - \|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2^2 (1 - \Omega(\frac{1}{\kappa^2 \text{poly}(k)})) \\ &- 4\sigma\sqrt{\frac{k}{\delta}}\|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2 \pm O(\varepsilon'\sqrt{\ell}\sigma + \frac{\sigma^2 k}{\delta}) \\ &= \Omega(\frac{1}{\kappa^2 \text{poly}(k)})\|\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2^2 + \|\overline{\mathbf{E}_{j,*}}\|_2^2 + 2\langle \boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}}), \overline{\mathbf{E}_{j,*}}\rangle \\ &- 4\sigma\sqrt{k/\delta}\|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2 \pm O(\varepsilon'\sqrt{\ell}\sigma + \frac{\sigma^2 k}{\delta})\end{aligned}$$

Thus,

$$\begin{aligned}a_i^- - a_i^+ > \sum_{j\in[m]} &\Omega(\frac{1}{\kappa^2 \text{poly}(k)})\|\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2^2 + 2\langle \boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}}), \overline{\mathbf{E}_{j,*}}\rangle \\ &- 4\sigma\sqrt{k/\delta}\|(\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2 \pm O(\varepsilon'\sqrt{\ell}\sigma + \frac{\sigma^2 k}{\delta})\end{aligned}$$

By Chebyshev's inequality, we have, $|2\langle \boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}}), \overline{\mathbf{E}_{j,*}}\rangle| < \text{poly}(dkm)\sigma\|\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2$ w.h.p. in d,k,m. Thus $\sum_j |2\langle \boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}}), \overline{\mathbf{E}_{j,*}}\rangle| < \text{poly}(dkm)\sigma\|\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2$. Now $\|\boldsymbol{U}_{*,i}^*\|_2 > 1/\kappa(\boldsymbol{U}^*)$, otherwise we would have $\|\boldsymbol{U}^* e_i\|_2 < 1/\kappa(\boldsymbol{U}^*)$, which is impossible by definition. Thus there is at least one entry of $\boldsymbol{U}_{*,i}^*$ with magnitude at least $1/(m\kappa(\boldsymbol{U}^*))$. So

$$\begin{aligned}\sum_j \|\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2^2 &\geqslant \frac{1}{m\kappa(\boldsymbol{U}^*)}\|f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2^2 \\ &= \Omega(\ell\frac{1}{m\kappa(\boldsymbol{U}^*)})\end{aligned}$$

where the last bound follows via bounds on χ^2 variables [LM00].

The above paragraph also demonstrates that $\frac{|2\langle \boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}}), \overline{\mathbf{E}_{j,*}}\rangle|}{\|\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2^2} \leqslant \frac{\text{poly}(dkm)\sigma}{\sqrt{\ell}}$, so taking ℓ sufficiently large this is less than $1/2$. Thus

$$\sum_{j\in[m]} \Omega(\frac{1}{\kappa^2 \text{poly}(k)})\|\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2^2 + 2\langle \boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}}), \overline{\mathbf{E}_{j,*}}\rangle > \frac{1}{2}\sum_{j\in[m]} \Omega(\frac{1}{\kappa^2 \text{poly}(k)})\|\boldsymbol{U}_{j,i}^* f(\boldsymbol{V}_{i,*}^* \overline{\mathcal{X}})\|_2^2$$

and we are left with

$$a_i^- - a_i^+ > \sum_{j\in[m]} \Omega(\frac{1}{\kappa^2 \text{poly}(k)})\|\boldsymbol{U}^*_{j,i} f(\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})\|_2^2 - 4\sigma\sqrt{k/\delta}\|(\boldsymbol{U}^*_{j,i} f(\boldsymbol{V}^*_{i,*}\overline{\mathcal{X}})\|_2 - O(\varepsilon'\sqrt{\ell}\sigma + \frac{\sigma^2 k}{\delta})$$
$$\geqslant \Omega(\ell \frac{1}{m\kappa^2\kappa(\boldsymbol{U}^*)\text{poly}(k)}) - O(\sigma\sqrt{k\ell/\delta} + \varepsilon'\sqrt{\ell}\sigma + \frac{\sigma^2 k}{\delta})$$

Taking $\delta = 1/\text{poly}(d,k,m)$ as before and ℓ sufficiently larger than $1/\delta^2$, the above becomes $a_i^- - a_i^+ = \Omega(\ell \frac{1}{m\kappa^2\kappa(\boldsymbol{U}^*)\text{poly}(k)}) = \omega(1)$ w.h.p. in d,k,m. Thus the algorithm correctly determines $\xi_i = 1$ after seeing $a_i^- > a_i^+$, and the analysis is symmetric in the case that $xi_i = -1$.

□

6.5.2 Recovering the Weights U^*, V^*

We have now shown in the prior section that given approximate $\boldsymbol{V}^*_{i,*}$'s where the signs ξ_i are unknown, we can recover the signs exactly in polynomial time in the noisy setting. Thus we recover $\boldsymbol{V}$ such that $\|\boldsymbol{V} - \boldsymbol{V}^*_{i,*}\|_2 \leqslant \varepsilon$ for some polynomially small ε. To complete our algorithm, we simply find $\boldsymbol{U}^*$ by solving the linear regression problem

$$\min_{\boldsymbol{U}} \|\boldsymbol{U} f(\boldsymbol{V}\mathcal{X}) - \mathbf{A}\|_F$$

It is well know that the solution to the above regression problem can be solved by computing the pseduoinverse of the matrix $f(\boldsymbol{V}\mathcal{X})$, thus the entire procedure can be carried out in polynomial time. The following lemma states that, even in the presence of noise, the solution $\boldsymbol{U}$ from the regression problem must in fact be very close to the true solution $\boldsymbol{U}^*$.

Lemma 6.5.5. *Let $\overline{\mathcal{X}}$ be the first $\ell = \text{poly}(k,d,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*),\sigma)$ columns of $\mathcal{X}$, and similarly define $\overline{\mathbf{E}}$. Set $\overline{\mathbf{A}} = \boldsymbol{U}^* f(\boldsymbol{V}^*\overline{\mathcal{X}}) + \overline{\mathbf{E}}$. Then given $\boldsymbol{V}$ such that $\|\boldsymbol{V} - \boldsymbol{V}^*\|_2 \leqslant \varepsilon$ where $\varepsilon = \frac{\varepsilon'}{\text{poly}(\ell)}$ for some $\varepsilon' > 0$, if $\boldsymbol{U}$ is the regression solution to $\min_U \|\boldsymbol{U} f(\boldsymbol{V}\overline{\mathcal{X}}) - \overline{\mathbf{A}}\|_2$, then $\|\boldsymbol{U} - \boldsymbol{U}^*\|_2 < O(\varepsilon')$ with high probability in m,d.*

Proof. Note $\|f(\boldsymbol{V}\overline{\mathcal{X}}) - f(\boldsymbol{V}^*\overline{\mathcal{X}})\|_2 \leqslant O(1)\sqrt{\ell}\varepsilon$ by standard operator norm bounds on Gaussian matrices (see Lemma 6.4.18), and the fact that $|a-b| > |f(a) - f(b)|$ for any $a, b \in \mathcal{R}$. The regression problem is solved row by row, so fix a row $i \in [m]$ and consider $\min_u \|uf(\boldsymbol{V}\overline{\mathcal{X}}) - \overline{\mathbf{A}}_{i,*}\|_2 = \min_u \|uf(\boldsymbol{V}\overline{\mathcal{X}}) - (\boldsymbol{U}^*_{i,*} f(\boldsymbol{V}^*\overline{\mathcal{X}}) + \overline{\mathbf{E}}_{i,*})\|_2 = \|uf(\boldsymbol{V}\overline{\mathcal{X}}) - (\boldsymbol{U}^*_{i,*} f(\boldsymbol{V}\overline{\mathcal{X}}) + \overline{\mathbf{E}_{i,*}} +$

$\boldsymbol{U}^*_{i,*}\mathbf{Z})\|_2$, where $\mathbf{Z}$ is a matrix such that $\|\mathbf{Z}\|_F \leqslant O(1)\sqrt{\ell}\varepsilon$. Now by the normal equations[4], if u^* is the above optimizer, we have

$$\begin{aligned} u^* &= \Big(\boldsymbol{U}^*_{i,*}f(\boldsymbol{V}\overline{\mathcal{X}}) + \overline{\mathbf{E}_{i,*}} + \boldsymbol{U}^*_{i,*}\mathbf{Z}\Big)f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big[f(\boldsymbol{V}\overline{\mathcal{X}})f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big]^{-1} \\ &= \boldsymbol{U}^*_{i,*} + \Big(\overline{\mathbf{E}_{i,*}} + \boldsymbol{U}^*_{i,*}\mathbf{Z}\Big)f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big[f(\boldsymbol{V}\overline{\mathcal{X}})f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big]^{-1} \\ &= \boldsymbol{U}^*_{i,*} + \overline{\mathbf{E}_{i,*}}f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big[f(\boldsymbol{V}\overline{\mathcal{X}})f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big]^{-1} + \boldsymbol{U}^*_{i,*}\mathbf{Z}' \end{aligned}$$

Where $\mathbf{Z}'$ is a matrix such that $\|\mathbf{Z}'\|_F = O\Big(\frac{\sqrt{\ell}\varepsilon\kappa(f(\boldsymbol{V}\overline{\mathcal{X}})}{\sigma_{\min}(f(\boldsymbol{V}\overline{\mathcal{X}}))}\Big)$. Note that we can scale $\overline{\mathbf{A}}$ at the beginning so that no entry is larger than ℓ^2, which implies w.h.p. that each row of $\boldsymbol{U}^*$ has norm at most ℓ^2. Thus

$$\|\boldsymbol{U}_{i,*}\mathbf{Z}'\|_F = O\Big(\frac{\ell^{5/2}\varepsilon\kappa(f(\boldsymbol{V}\overline{\mathcal{X}})}{\sigma_{\min}(f(\boldsymbol{V}\overline{\mathcal{X}}))}\Big)$$

Now note $\mathbb{E}\left[\|\overline{\mathbf{E}_{i,*}}f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big[f(\boldsymbol{V}\overline{\mathcal{X}})f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big]^{-1}\|_2^2\right] = O(\sqrt{\ell k}\frac{\sigma^2}{\sigma^2_{\min}(f(\boldsymbol{V}\overline{\mathcal{X}}))})$ using $\|f(\boldsymbol{V}\overline{\mathcal{X}}\|_2 = O(\sqrt{\ell k})$ by the same operator norm bounds as before. By Markov bounds, w.h.p. in m, d, we have

$$\|\overline{\mathbf{E}_{i,*}}f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big[f(\boldsymbol{V}\overline{\mathcal{X}})f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big]^{-1}\|_2^2 = O(\sqrt{\ell}\text{poly}(d,m)\frac{\sigma^2}{\sigma^2_{\min}(f(\boldsymbol{V}\overline{\mathcal{X}}))})$$

Now by Courant-Fischer theorem and application of the triangle inequality, we have $\sigma_{\min}(f(\boldsymbol{V}\overline{\mathcal{X}})) > \sigma_{\min}(f(\boldsymbol{V}^*\overline{\mathcal{X}})) - O(\sqrt{\ell}\varepsilon)$, and by Lemma 6.5.3 (see Section 6.5.1), we have $\sigma_{\min}(f(\boldsymbol{V}^*\overline{\mathcal{X}})) = \Omega(\frac{\sqrt{\ell}}{\kappa(\boldsymbol{V}^*)\text{poly}(k)})$, thus for ℓ sufficiently large we obtain $\sigma_{\min}(f(\boldsymbol{V}\overline{\mathcal{X}})) = \Omega(\frac{\sqrt{\ell}}{\kappa(\boldsymbol{V}^*)\text{poly}(k)})$, in which case we have

$$\|\overline{\mathbf{E}_{i,*}}f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big[f(\boldsymbol{V}\overline{\mathcal{X}})f(\boldsymbol{V}\overline{\mathcal{X}})^T\Big]^{-1}\|_2^2 = O\Big(\text{poly}(d,m)\frac{\sigma^2}{\sqrt{\ell}}\Big)$$

Setting $\ell, 1/\varepsilon$ to be sufficiently large polynomials in $(d, m, k, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*), \sigma)$, we obtain

$$\|u^* - \boldsymbol{U}^*_{i,*}\|_2 \leqslant O(\varepsilon'/\sqrt{m})$$

from which the Lemma follows. □

We now state our main theorem for recovery of the weight matrices in the noisey case.

Theorem 138. *Let $\mathbf{A} = \boldsymbol{U}^*f(\boldsymbol{V}^*\mathcal{X}) + \mathbf{E}$ be given, where $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V}^* \in \mathcal{R}^{k\times d}$ are rank-k and $\mathbf{E}$ is a matrix of i.i.d. mean zero subgaussian random variables with variance σ^2. Then*

[4]See https://en.wikipedia.org/wiki/Linear_least_squares

given $n = \Omega\Big(\text{poly}\Big(d, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*), \sigma, \frac{1}{\varepsilon}\Big)\Big)$*, there is an algorithm that runs in* $\text{poly}(n)$ *time and w.h.p. outputs* $\boldsymbol{V}, \boldsymbol{U}$ *such that*

$$\|\boldsymbol{U} - \boldsymbol{U}^*\|_F \leqslant \varepsilon \qquad \|\boldsymbol{V} - \boldsymbol{V}^*\|_F \leqslant \varepsilon$$

Proof. The proof of correctness of the Tensor Decomposition based approximate recovery of $\boldsymbol{V}^*$ up to the signs is the same as in the exact case, via Theorem 130. By Theorem 137, we can recover the signs ξ_i, and thus recover $\boldsymbol{V}$ so that $\|\boldsymbol{V} - \boldsymbol{V}^*\|_F \leqslant \varepsilon$. Observe that while the results in Section 6.5.1 were stated for $\varepsilon = \Theta\left(\frac{1}{\text{poly}(d,m,\kappa(\boldsymbol{U}^*),\kappa(\boldsymbol{V}^*)\sigma)}\right)$, they can naturally be generalized to any ε which is at least this small by increasing n by a $\text{poly}(1/\varepsilon)$ factor before running the tensor decomposition algorithm. Then by Lemma 6.5.5, we can recover $\boldsymbol{U}$ in polynomial time such that $\|\boldsymbol{U} - \boldsymbol{U}^*\|_F \leqslant \varepsilon$ as desired, which completes the proof. □

Remark 139. As in Remark 132, we have implicitly normalized the entire matrix $\mathbf{A}$ so that the columns of $\mathbf{A}$ have at most unit norm. If one seeks bounds for the recovery of the *unnormalized* $\boldsymbol{U}^*$, the error becomes $\|\boldsymbol{U} - \boldsymbol{U}^*\|_F \leqslant \varepsilon\|\boldsymbol{U}^*\|_2$. To see why this holds, note that the normalization factor of Remark 132 is at least $\Omega(\frac{1}{\|\boldsymbol{U}^*\|_2 + \sqrt{m\log(\ell)}})$, where $\ell = \text{poly}(d, m, \kappa(\boldsymbol{U}^*), \kappa(\boldsymbol{V}^*), \sigma)$ is as in Section 6.5.1, and $O(\sqrt{m\log(\ell)})$ is a bound on the max column norm of $\mathbf{E}$ by subgaussian concentration. Thus multiplying by the inverse of this normalization factor blows up the error to $\|\boldsymbol{U}^*\|_2\varepsilon$ after scaling ε down by a polynomial factor.

6.6 A Fixed-Parameter Tractable Exact Algorithm for Arbitrary Weight Matrixs

In the prior sections, we required that $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}$ have rank k in order to recover it properly. Of course, this is a natural assumption, making $\boldsymbol{U}^*$ identifiable. In this section, however, we show that even when $m < k$ and $\boldsymbol{U}^*$ does not have full column rank, we can still recover $\boldsymbol{U}^*\boldsymbol{V}^*$ exactly in the noiseless case where we are given $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ and $\mathcal{X}$, as long as the no two columns of $\boldsymbol{U}^*$ are non-negative scalar multiples of each other. Observe that this excludes columns from being entirely zero, but allows for columns of the form $[u, -u]$ for for $u \in \mathcal{R}^m$, as long as u is non-zero. Our algorithm requires $n = \text{poly}(d, k)\kappa^{\Omega(k)}$, samples, and runs in time $O(n\text{poly}(d, m, k))$. Here $\kappa = \kappa(\boldsymbol{V}^*)$ is the condition number of $\boldsymbol{V}^*$. Our algorithm does not have any dependency on the condition number of $\boldsymbol{U}^*$.

Algorithm 7 : FPTExactNeuralNet$(\boldsymbol{V}', \mathcal{X}, \mathcal{S})$**.**

`Input:` Matrices $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) \in \mathcal{R}^{d \times n}$ and $\mathcal{X} \in \mathcal{R}^{r \times n}$ such that each entry in $\mathcal{X} \sim \mathcal{N}(0, 1)$.

1. Find a subset S of columns of non-zero columns of $\mathbf{A}$ such that each for each $i \in S$ there is a $j \in S$, $j \neq i$, with $\mathbf{A}_{*,i} = c\mathbf{A}_{*,j}$ for some $c > 0$.
2. Partition S into S_r for $r \in [k]$ such that for each pair $i, j \in S_r$, $i \neq j$, we have $\mathbf{A}_{*,i} = c\mathbf{A}_{*,j}$ for some $c \in \mathcal{R}^{\neq 0}$.
3. For each $i \in [k]$, choose a representative $j_i \in S_i$, and let $\boldsymbol{U}_{*,i} = \mathbf{A}_{*,j_i}$. For each $j \in S_i$, let $c_{i,j}$ be such that $c_{i,j}\boldsymbol{U}_{*,i} = \mathbf{A}_{*,j}$.
4. let $\boldsymbol{W}$ be the matrix where the i-th row is given by the solution w_i to the following linear system:

$$\forall i \in [k] : \quad w_i \mathcal{X}_{*,j} = c_{i,j} \quad \text{if } j \in S_i$$

5. Set $\boldsymbol{V}_{i,*} = \boldsymbol{W}_{i,*} / \|\boldsymbol{W}_{i,*}\|_2$, and let $\boldsymbol{U}$ be the solution to the following linear system:

$$\boldsymbol{U} f(\boldsymbol{V}\mathcal{X}) = \mathbf{A}$$

`Output:` $(\boldsymbol{U}, \boldsymbol{V})$.

The runtime of our algorithm is polynomial in the sample complexity n and the size of the networks d, k, but simply requires poly$(d, k)\kappa^{\Omega(k)}$ samples in order to obtain columns of $f(\boldsymbol{V}^*\mathcal{X})$ which are 1-sparse, in which case the corresponding column of $\mathbf{A}$ will be precisely a positive scaling of a column of $\boldsymbol{U}^*$. In this way, we are able to progressively recover each column of $\boldsymbol{U}^*$ simply by finding columns of $\mathbf{A}$ which are scalar multiples of each other. The full algorithm, Algorithm 7, is given formally below.

Lemma 6.6.1. *For each $i \in [k]$, with probability $1-\delta$, at least d columns of $f(\boldsymbol{V}^*\overline{\mathcal{X}})$ are positive scalings of e_i^T, where $\overline{\mathcal{X}}$ is the first ℓ columns of $\mathcal{X}$ for $n = \Omega(d \log(k/\delta)\kappa^{O(k)})$. In other words, $|S_i| \geqslant d$.*

Proof. Let $\kappa = \kappa(\boldsymbol{V}^*)$. As in the proof of Lemma 6.5.3, we can assume that $\boldsymbol{V}^*$ is lower triangular by rotating the rows by a matrix $\mathbf{R}$, and noting that $\mathbf{R}\mathcal{X}$ has the same distribution as

$\mathcal{X}$ by the rotational invariance of Gaussians. We now claim that $\Pr[\|\boldsymbol{V}^* g\|_2 < \frac{1}{k\kappa}] = \Omega((\frac{1}{k\kappa})^k)$, where $g \sim \mathcal{N}(0, \mathbb{I}_d)$ is a Gaussian vector. To see this, since $\boldsymbol{V}^*$ is rank k and in lower triangular form, $\boldsymbol{V}^*$ is supported on its the first k columns. Thus it suffices to compute the value $\|\boldsymbol{V}^* g\|_2$ were $g \in \mathcal{R}^k$ is a k-dimensional Gaussian. By the anti-concentration of Gaussian, each $g_i < \frac{1}{k\kappa}$ with probability at least $\Omega(1/(k\kappa))$. Since the entries are independent, it follows that $\Pr[\|g\|_2 \leqslant \frac{1}{\sqrt{k}\kappa}] = \Omega(1/(k\kappa)^k)$. Let $\mathcal{E}_1$ be the event that this occurs. Since $\boldsymbol{V}^*$ has unit norm rows, it follows by Cauchy-Schwartz that conditioned on $\mathcal{E}_1$, we have $\tilde{g} = \boldsymbol{V}^* g$ satisfies $\|\tilde{g}\|_2 = O(\frac{1}{\kappa})$

Now consider the pdf of the k-dimensional multivariate Gaussian $\tilde{g}$ that has covariance $\Sigma = \boldsymbol{V}^*(\boldsymbol{V}^*)^T$, which is given by

$$p(x) = \frac{\exp\Big(-\frac{1}{2}x\Sigma^{-1}x\Big)}{\sqrt{(2\pi)^k \det(\Sigma)}}$$

for $x \in \mathcal{R}^k$. Now condition on the event $\mathcal{E}_2$ that $\tilde{g}$ is contained within the ball $\mathcal{B}$ of radius $O(\frac{1}{\kappa})$ centered at 0. Since $\mathcal{E}_1$ implies $\mathcal{E}_2$, we have $\Pr[\mathcal{E}_2] = \Omega(1/(k\kappa)^k)$ Now the eigenvalues of Σ are the squares of the singular values of $\boldsymbol{V}^*$, which are all between $1/\kappa$ and $\sqrt{k}$. So all eigenvalues of Σ^{-1} are between $1/k$ and κ^2. Thus for all $x \in \mathcal{B}$, we have

$$\frac{1}{2} \leqslant \frac{1}{e^{1/2}} \leqslant \exp\Big(-\frac{1}{2}x\Sigma^{-1}x\Big) \leqslant 1$$

It follows that

$$\sup_{x,y\in\mathcal{B}} \frac{p(x)}{p(y)} \leqslant 2$$

Now let $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_{2^k}$ be the intersection of all 2^k orthants in $\mathcal{R}^k$ with $\mathcal{B}$. The above bound implies that

$$\max_{i,j\in[2^k]} \frac{\int_{\mathcal{O}_i} p(x)dx}{\int_{\mathcal{O}_j} p(y)dy} \leqslant 2$$

Thus conditioned on $\mathcal{E}_2$ for the i.i.d. gaussian vector $\tilde{g} \sim \mathcal{N}(0, \Sigma) \in \mathcal{R}^k$, the probability that $\tilde{g}$ is in a given $\mathcal{O}_i$ is at most twice the probability that g is in $\mathcal{O}_j$ for any other j. Thus $\min_{i\in[2^k]} \Pr[\tilde{g} \in \mathcal{O}_i] > \frac{1}{2^{k+1}}$. Thus for any sign pattern $\mathcal{S}$ on k-dimensional vectors, and in particular for the sign partner $\mathcal{S}_i$ of e_i, the probability that $\tilde{g}$ has this sign pattern conditioned on $\mathcal{E}_2$ is at least $\frac{1}{2^{k+1}}$. Since $\Pr[\mathcal{E}_2] = \Omega(1/(k\kappa)^k)$, it follows that in $n = \Omega(d\log(k/\delta)(k\kappa)^{2k})$ repetitions, a scaling of e_i will appear at least d times in the columns of $f(\boldsymbol{V}^*\mathcal{X})$ with probability $1 - \delta/k$, and the Lemma follows by a union bound over e_i for $i \in [k]$.

□

Theorem 140. *Suppose* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ *for* $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}$ *for any* $m \geqslant 1$ *such that no two*

columns of $\boldsymbol{U}^*$ are non-negative scalar multiples of each other, and $\boldsymbol{V}^* \in \mathcal{R}^{k\times n}$ has rank$(\boldsymbol{V}^*) = k$, and $n > \kappa^{O(k)}$poly(dkm). Then Algorithm 7 recovers $\boldsymbol{U}^*, \boldsymbol{V}^*$ exactly with high probability in time $\kappa^{O(k)}$poly(d, k, m).

Proof. By Lemma 6.6.1, at least d columns of $f(\boldsymbol{V}^*\mathcal{X})$ will be scalar multiples of e_i for each i. Thus the set S of indices, as defined in Step 2 of Algorithm 7, will contain each column of $\boldsymbol{U}^*$ as a column. It suffices to show that no two columns of $\overline{\mathbf{A}}$ can be scalar multiples of each other if they are not a scalar multiple of $\boldsymbol{U}^*$. To see this, if two columns of $f(\boldsymbol{V}^*\mathcal{X})$ were not 1-sparse, then the distribution of $\boldsymbol{U}^*f(\boldsymbol{V}^*\mathcal{X})$ on these columns is supported on a t-dimensional manifold living inside $\mathcal{R}^m$, for some $t \geqslant 2$. In particular, this manifold is the conic hull of at least $t' \geqslant t \geqslant 2$ columns of $\boldsymbol{U}^*$ (where t' is the sparsity of the columns of $f(\boldsymbol{V}^*\mathcal{X})$. This follows from the fact that the conic hull of any subset of 2 columns of $\boldsymbol{U}^*$ is 2-dimensional, since no columns two of $\boldsymbol{U}^*$ are non-negative scalings of each other. Thus the probability that two draws from such a distribution lie within the same 1-dimensional subspace, which has measure 0 inside of any $t \geqslant 2$-dimensional conic hull, is therefore 0, which completes the claim.

To complete the proof of the theorem, by pulling a diagonal matix $\boldsymbol{D}$ through f, we can assume $\boldsymbol{U} = \boldsymbol{U}^*$. By construction then, $c_{i,j}$ is such that $(\boldsymbol{D}\boldsymbol{V}_{i,*}\mathcal{X}_{*,j}) = c_{i,j}$, as it is the scaling which takes $\boldsymbol{U}_{*,i}$ to $\mathbf{A}_{*,j}$. Thus w_i, as defined in step 4 of Algorithm 7, is the solution to a linear equation $w_i\mathcal{X}_{S_i} = c$ for some fixed vector c, where $\mathcal{X}_{S_i}$ is $\mathcal{X}$ restricted to the columns in S_i. Since $|S_i| \geqslant d$ by Lemma 6.6.1, to show that w_i is unique it suffices for $\mathcal{X}_{S_i}$ to be full rank. But as argued in the proof of Theorem 129, any subset of d columns of $\mathcal{X}$ will be rank d and invertible with probability 1. Thus w_i is unique, and must therefore be a scaling of $\boldsymbol{V}^*_{i,*}$, which we find by normalizing w_i to have unit norm. After this normalization, we can renormalize $\boldsymbol{U}$, or simply solve a linear system for $\boldsymbol{U}$ as in Step 5 of Algorithm 7. By Lemma 6.2.4, $f(\boldsymbol{V}^*\mathcal{X})$ will have full rank w.h.p., so the resulting $\boldsymbol{U}$ will be unique and therefore equal to $\boldsymbol{U}^*$ as desired.

□

6.7 A Fixed-Parameter Tractable Algorithm for Arbitrary Non-Adversarial Noise

In the noisy model, the observed matrix $\mathbf{A}$ is generated by a perturbation $\mathbf{E}$ of some neural network $\boldsymbol{U}^*f(\boldsymbol{V}^*\mathcal{X})$ with rank k matrices $\boldsymbol{U}^* \in \mathcal{R}^{m\times k}, \boldsymbol{V} \in \mathcal{R}^{k\times d}$, and i.i.d. Gaussian $\mathcal{N}(0, 1)$ input $\mathcal{X} \in \mathcal{R}^{d\times n}$. Formally, we are given as input $\mathcal{X}$ and $\mathbf{A} = \boldsymbol{U}^*f(\boldsymbol{V}\mathcal{X}) + \mathbf{E}$, which is a noisy

observation of the underlying network $\boldsymbol{U}^* f(\boldsymbol{V}\mathcal{X})$, and tasked with recovering approximations to this network. In Section 6.5, we showed that approximate recovery of the weight matrices $\boldsymbol{U}^*, \boldsymbol{V}^*$ is possible in polynomial time when the matrix $\mathbf{E}$ was i.i.d. mean 0 and sub-Gaussian. In this section, we generalize our noise model substantially to include all error matrices $\mathbf{E}$ which *do not depend* on the input matrix $\mathcal{X}$. Our goal is then to obtain $\boldsymbol{U}, \boldsymbol{V}$ such that

$$\|\boldsymbol{U} f(\boldsymbol{V}\mathcal{X}) - \mathbf{A}\|_F \leqslant (1+\varepsilon)\|\mathbf{E}\|_F$$

Thus we would like to be able to recover a good approximation to the observed input, where we are competing against the cost $\mathrm{OPT} = \|\mathbf{A} - \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})\|_2 = \|\mathbf{E}\|_2$. Observe that this is a slightly different objective than before, where our goal was to recover the actual weights $\boldsymbol{U}^*, \boldsymbol{V}^*$ approximately. This is a product of the more general noise model we consider in this Section. The loss function here can be thought of as recovering $\boldsymbol{U}, \boldsymbol{V}$ which approximate the observed classification nearly as well as the optimal generative $\boldsymbol{U}^*, \boldsymbol{V}^*$ do. This is more similar to the empirical loss considered in other words [ABMM16]. The main result of this section is the development of a fixed parameter tractable algorithm which returns $\boldsymbol{U}, \boldsymbol{V}$ such that

$$\|\mathbf{A} - \boldsymbol{U} f(\boldsymbol{V}\mathcal{X})\|_F \leqslant \|\mathbf{E}\|_F + O\left(\left[\sigma_{\min}\varepsilon\sqrt{nm}\|\mathbf{E}\|_2\right]^{1/2}\right) \tag{6.3}$$

Where $\sigma_{\max} = \sigma_{\max}(\boldsymbol{U}^*)$, and $\|\mathbf{E}\|_2$ is the spectral norm of $\mathbf{E}$. In this section, to avoid clustering, we will write $\sigma_{\max}, \sigma_{\min}$, and κ to denote the singular values and condition number of $\boldsymbol{U}^*$. Our algorithm has no dependency on the condition number of $\boldsymbol{V}^*$. The runtime of our algorithm is $(\frac{\kappa}{\varepsilon})^{O(k^2)}\mathrm{poly}(n, r, d)$, which is fixed-parameter tractable in $k, \kappa, \frac{1}{\varepsilon}$. Here the sample complexity n satisfies $n = \Omega(\mathrm{poly}(r, d, \kappa, \frac{1}{\varepsilon}))$.

We remark that the above bound in Equation 6.3 may at first seem difficult to parse. Intuitively, this bound will be a $(1+\varepsilon)$ multiplicative approximation whenever the Frobenius norm of $\mathbf{E}$ is roughly an $\sqrt{m}$ factor larger than the spectral norm–in other words, when the error $\mathbf{E}$ is relatively flat. Note that these bounds will hold when $\mathbf{E}$ is drawn from a very wide class of random matrices, including matrices with heavier tails (see [Ver10b] and discussion below). When this is not the case, and $\|\mathbf{E}\|_2 \approx \|\mathbf{E}\|_F$, then we lose an additive $\sqrt{m}$ factor in the error guarantee. Note that this can be compensated by scaling ε by a $\frac{1}{\sqrt{m}}$ factor, in which case we will get a $(1+\varepsilon)$ multiplicative approximation for any $\mathbf{E}$ which is not too much smaller than $\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ (meaning $\|\mathbf{E}\|_F = \Omega(\varepsilon\|\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})\|_F)$). The runtime in this case will be $(m\kappa/\varepsilon)^{O(k^2)}$, which is still $(\kappa/\varepsilon)^{O(k^3)}$ whenever $m = O(2^k)$. Note that if the noise $\mathbf{E}$ becomes arbitrarily smaller than the signal $\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$, then the multiplicative approximation of Equation 6.3 degrades, and

instead becomes an additive guarantee.

To see why this is a reasonable bound, we must first examine the normalizations implicit in our problem. As always we assume that $\boldsymbol{V}^*$ has unit norm rows. Using the 2-stability of Gaussians, we know that $\mathbb{E}\left[(\boldsymbol{V}^*\mathcal{X})_{i,j}^2\right] = 1$ for any $i, j \in [k]\times[n]$, and by symmetry of Gaussians we have that $\mathbb{E}\left[f(\boldsymbol{V}^*\mathcal{X})_{i,j}^2\right] = 1/2$. By linearity of expectation we have $\mathbb{E}\left[\|f(\boldsymbol{V}^*\mathcal{X})\|_F^2\right] = kn/2$. Since $\sigma_{\min}^2\|f(\boldsymbol{V}^*\mathcal{X})\|_F^2 \leqslant \|\boldsymbol{U}f(\boldsymbol{V}^*\mathcal{X})\|_F^2 \leqslant \sigma_{\max}^2\|f(\boldsymbol{V}^*\mathcal{X})\|_F^2$, it follows that

$$\frac{\sigma_{\min}^2(\boldsymbol{U})kn}{2} \leqslant \mathbb{E}\left[\|\boldsymbol{U}f(\boldsymbol{V}^*\mathcal{X})\|_F^2\right] \leqslant \frac{\sigma_{\max}^2(\boldsymbol{U})kn}{2}$$

Thus for the scale of the noise $\mathbf{E}$ to be within a $\Omega(1)$ factor of the average squared entry of $\boldsymbol{U}f(\boldsymbol{V}^*\mathcal{X})$ on average, we expect $\|\mathbf{E}\|_F = O(\sigma_{\max}\sqrt{nk})$ and $\|\mathbf{E}\|_F = \Omega(\sigma_{\min}\sqrt{nk})$

Now consider the case where $\mathbf{E}$ is a random matrix, coming from a very broad class of distributions. Since $\mathbf{E} \in \mathcal{R}^{m\times n}$ with $n >> m$, one of the main results of random matrix theory is that many such matrices are *approximately* isometries [Ver10b]. Thus, for a such a random matrix $\mathbf{E}$ normalized to be within a constant of the signal, we will have $\|\mathbf{E}\|_2 = O(\sigma_{\max}\sqrt{\frac{nk}{m}})$. This gives

$$\|\mathbf{A} - \boldsymbol{U}f(\boldsymbol{V}^*\mathcal{X})\|_F \leqslant \|\mathbf{E}\|_F(1 + O(\varepsilon))$$

after scaling ε by a quadratic factor. In general, we get multiplicative approximations whenever either the spectrum of $\mathbf{E}$ is relatively flat, or when we allow $(m\kappa/\varepsilon)^{O(k^2)}$ runtime. Note that in both cases, for the above bound to be a $\|\mathbf{A} - \boldsymbol{U}f(\boldsymbol{V}^*\mathcal{X})\|_F \leqslant (1 + \varepsilon)\|\mathbf{E}\|_F$ approximation, we must have $\|\mathbf{E}\|_F = \Omega(\varepsilon\|\boldsymbol{U}^*f(\boldsymbol{V}^*\mathcal{X})\|_F)$ as noted above. Otherwise, the error we are trying to compete against is too small when compared to the matrices in question to obtain a multiplicative approximation.

6.7.1 Main Algorithm

Our algorithm is then formally given in Figure 8. Before presenting it, we first recall some fundamental tools of numerical linear algebra. First, we recall the notion of a subspace-embedding.

Definition 6.7.1 (Subspace Embedding)**.** *Let $\boldsymbol{U} \in \mathcal{R}^{m\times k}$ be a rank-k matrix and, let $\mathcal{F}$ be family of random matrices with m columns, and let $\mathcal{S}$ be a random matrix sampled from $\mathcal{F}$. Then we*

say that $\mathcal{S}$ is a $(1 \pm \delta)$-ℓ_2-subspace embedding for the column space of $\boldsymbol{U}$ if for all $x \in \mathcal{R}^k$,

$$\|\mathcal{S}\boldsymbol{U}x\|_2 = (1 \pm \delta)\|\boldsymbol{U}x\|_2$$

Note in the above definition, $\mathcal{S}$ is a subspace embedding for the column span of $\boldsymbol{U}$, meaning for any other basis $\boldsymbol{U}'$ spanning the same columns as $\boldsymbol{U}$, we have that $\mathcal{S}$ is also a $(1 \pm \delta)$-ℓ_2-subspace embedding for $\boldsymbol{U}'$. For brevity, we will generally say that $\mathcal{S}$ is a subspace embedding for a matrix $\boldsymbol{U}$, with the understanding that it is in fact a subspace embedding for *all* matrices with the same column span as $\boldsymbol{U}$. Note that if $\mathcal{S}$ is a subspace embedding for a rank-k matrix $\boldsymbol{U}$ with largest and smallest singular values $\sigma_{\max}$ and $\sigma_{\min}$ respectively, then $\mathcal{S}\boldsymbol{U}$ is rank-k with largest and smallest singular values each in the range $(1 \pm \delta)\sigma_{\max}$ and $(1 \pm \delta)\sigma_{\min}$ respectively. The former fact can be seen by the necessity that $\|\mathcal{S}\boldsymbol{U}x\|_2$ be non-zero for all non-zero $x \in \mathcal{R}^k$, and the latter by the fact that $\max_{x \in \mathcal{R}^k,\, \|x\|_2 = 1} \|\mathcal{S}\boldsymbol{U}x\|_2 = (1 \pm \delta) \max_{x \in \mathcal{R}^k,\, \|x\|_2 = 1} \|\boldsymbol{U}x\|_2 = \sigma_{\max}$, and the same bound holds replacing $\max$ with $\min$. Our algorithm will utilize the following common family of random matrices.

Algorithm 8 : Learning Neural Nets with Gaussian Inputs and Arbitrary Non-Adversarial Noise$(\mathbf{A}, \mathcal{X})$**.**

1. Generate a random matrix $\mathcal{S} \in \mathcal{R}^{c_1 k/\delta^2 \times m}$ of i.i.d. Gaussian $\mathcal{N}(0, 1/k)$ variables for some sufficiently large constant c_1 and $\delta = 1/10$.
2. Enumerate all $k \times c_1 k/\delta^2$ matrices $\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^\nu$ with entries of the form $\frac{1}{\sigma_{\min}(\boldsymbol{U})}(1 + \frac{\varepsilon^4}{ck^4})^{-i}$ for integers $0 \leqslant i \leqslant c' k^8 (1/\varepsilon^8)\kappa^8$ for sufficiently large constants c, c' and any $\frac{1}{\varepsilon} > k$. Note that $\nu = 2^{O(k^2 \log(\frac{1}{\varepsilon} k\kappa))}$.
3. For $i = 1, 2, \ldots, \nu$
   (a) Generate a matrix $\boldsymbol{G} \in \mathcal{R}^{k \times n}$ s.t. $\boldsymbol{G}$ consists of i.i.d. $\mathcal{N}\left(0, \Theta\left(\left(\frac{\varepsilon^{-2}\kappa^2 k \|\mathbf{M}\mathcal{S}\mathbf{E}\|_F}{\sqrt{n}}\right)^2\right)\right)$ random variables. Note, we can guess the value $\|\mathbf{M}\mathcal{S}\mathbf{E}\|_F$ in $O(\log(n))$ powers of 2 around $\|\mathbf{M}\mathcal{S}\mathbf{A}\|_F$.
   (b) For each row $p \in [k]$ and $q \in [n]$, let $y_q = \mathtt{sign}(\mathbf{M}^i \mathcal{S}\mathbf{A} + \boldsymbol{G})_{p,q}$.
   (c) For each $p = 1, 2, \ldots, [k]$, let w_i^p be the solution to the following convex program:
   $$\max_{w,} \sum_{i=1}^{n} y_i \langle w, X_{*,i} \rangle$$
   $$\text{subject to } \|w\|_2^2 \leqslant 1$$
   (d) Let $\boldsymbol{V}^i \in \mathcal{R}^{k \times d}$ be the matrix with p-th row equal to w_i^p.
4. Let $\boldsymbol{U}$ and $\boldsymbol{V}^{i^*}$ be the matrices that achieve the minimum value of the linear regression problem
   $$\arg\min_{\boldsymbol{U}, \boldsymbol{V}^i} \|\mathbf{A} - \boldsymbol{U} f(\boldsymbol{V}^i \mathcal{X})\|_F^2$$

Output: $(\boldsymbol{U}, \boldsymbol{W}^{i^*})$.

Proposition 6.7.2 (Gaussian Subspace Embedding [Sar06])**.** *Fix any rank k-matrix matrix $\boldsymbol{U} \in \mathcal{R}^{m \times k}$, and let $\mathcal{S} \in \mathcal{R}^{c_1 k/\varepsilon^2 \times m}$ be a random matrix where every entry is generated i.i.d. Gaussian $\mathcal{N}(0, 1/k)$, for some sufficiently large constant c_1. Then with probability $99/100$, $\mathcal{S}$ is a subspace embedding for $\boldsymbol{U}$.*

Our algorithm is then as follows. We first sketch the input matrix $\mathbf{A}$ by a $O(k) \times m$ Gaussian matrix $\mathcal{S}$, and condition on it being a subspace embedding for $\boldsymbol{U}$. We then left multiply by $\mathcal{S}$ to obtain $\mathcal{S}\mathbf{A} = \mathcal{S}\boldsymbol{U} f(\boldsymbol{V}\mathcal{X}) + \mathcal{S}\mathbf{E}$. Now we would like to ideally recover $f(\boldsymbol{V}\mathcal{X})$, and since $\mathcal{S}$ is a subspace embedding for $\boldsymbol{U}$, we know that $\mathcal{S}\boldsymbol{U}$ has full column rank and thus has an left

inverse. Since we do not know $\boldsymbol{U}$, we must guess the left inverse $(\mathcal{S}\boldsymbol{U})^{-1} \in \mathcal{R}^{k \times O(k)}$ of $\mathcal{S}\boldsymbol{U}$. We generate guesses $\mathbf{M}^i$ of $(\mathcal{S}\boldsymbol{U})^{-1}$, and try each of them. For the right guess, we know that after left multiplying by $\mathbf{M}^i$ we will have $\mathbf{M}^i\mathcal{S}\mathbf{A} = f(\boldsymbol{V}\mathcal{X}) + \mathbf{M}^i\mathcal{S}\mathbf{E} + \mathbf{Z}$, where $\mathbf{Z}$ is some error matrix which arises from our error in guessing $(\mathcal{S}\boldsymbol{U})^{-1}$.

We then observe that the signs of each row of this matrix can be thought of as labels to a noisy halfspace classification problem, where the sign of $(\mathbf{M}^i\mathcal{S}\mathbf{A})_{p,q}$ is a noisy observation of the sign of $\langle \boldsymbol{V}_{p,*}, \mathcal{X}_{*,q} \rangle$. Using this fact, we then run a convex program to recover each row $\boldsymbol{V}_{p,*}$. In order for recovery to be possible, there must be some non-trivial correlation between the labeling of these signs, meaning the sign of $(\mathbf{M}^i\mathcal{S}\mathbf{A})_{p,q}$, and the true sign of $\langle \boldsymbol{V}_{p,*}, \mathcal{X}_{*,q} \rangle$. In order to accomplish this, we must *spread out* the error $\mathbf{E}$ to allow the value of $\langle \boldsymbol{V}_{p,*}, \mathcal{X}_{*,q} \rangle$ to have an effect on the observed sign a non-trivial fraction of the time. We do this by adding a matrix $\boldsymbol{G}$ such that the i-th row $\boldsymbol{G}_{i,*}$ consists of i.i.d. $\mathcal{N}\Big(0, \Theta\Big(\Big(\frac{\varepsilon^{-2}\kappa^2 k \|\mathbf{M}\mathcal{S}\mathbf{E}\|_F}{\sqrt{n}}\Big)^2\Big)\Big)$ random variables to $\mathbf{M}^i\mathcal{S}\mathbf{A}$. We will simply guess the value $\|\mathbf{M}\mathcal{S}\mathbf{E}\|_F$ here in $O(\log(n))$ powers of 2 around $\|\mathbf{M}\mathcal{S}\mathbf{A}\|_F$. We prove a general theorem (Theorem 141) about the recovery of hyperplanes $v \in \mathcal{R}^d$ when given noisy labels from a combination of ReLU observations, adversarial, and non-adversarial noise components. Finally, we solve for $\boldsymbol{U}$ by regression. The full procedure is described formally given in Algorithm 8.

6.7.2 Analysis

First note that by our earlier bounds on the singular values of $\mathcal{X}$ (Proposition 6.4.18), we have $\|f(\boldsymbol{V}^*\mathcal{X})\|_F \leqslant O(\sqrt{nk})$, thus if $\|\mathbf{E}\|_F > \sigma_{\max}(\boldsymbol{U}^*)\frac{\sqrt{nk}}{\varepsilon}$, we can simply return $\boldsymbol{U}^* = 0, \boldsymbol{V}^* = 0$, and obtain our desired competitive approximation with the cost $OPT = \|\mathbf{E}\|_F$. Thus, where can now assume that $\|\mathbf{E}\|_F < \sigma_{\max}(\boldsymbol{U}^*)\frac{\sqrt{nk}}{\varepsilon}$.

By Proposition 6.7.2, with probability $99/100$ we have both that $\mathcal{S}\boldsymbol{U}^*$ is rank-k and that the largest and smallest singular values of $\mathcal{S}\boldsymbol{U}^*$ are perturbed by at most a $(1 \pm \delta)$ factor, meaning $\sigma_{\max}(\boldsymbol{U}^*) = (1 \pm \delta)\sigma_{\max}(\mathcal{S}\boldsymbol{U}^*)$ and $\sigma_{\min}(\boldsymbol{U}^*) = (1 \pm \delta)\sigma_{\in}(\mathcal{S}\boldsymbol{U}^*)$, from which it follows that $\kappa(\boldsymbol{U}^*) = (1 \pm O(\delta))\kappa(\mathcal{S}\boldsymbol{U}^*)$. Note that we can repeat the algorithm $O(n)$ times to obtain this result with probability $1 - \exp(-n)$ at least once by Hoeffding bounds. So we can now condition on this and assume the prior bounds on the singular values and rank of $\mathcal{S}\boldsymbol{U}^*$. Thus we will now write $\sigma_{\max} = \sigma_{\max}(\mathcal{S}\boldsymbol{U}^*)$, $\sigma_{\min} = \sigma_{\min}(\mathcal{S}\boldsymbol{U}^*)$, and $\kappa = \kappa(\mathcal{S}\boldsymbol{U}^*)$, with the understanding that these values have been perturbed by a $(1 \pm 3\delta) < (1 \pm 1/2)$ factor.

We can assume that we know κ and $\sigma_{\min}(\boldsymbol{U}^*)$ up to a factor of 2 by guessing them in geometrically increasing intervals. Note that we can assume $\sigma_{\max}$ is within a poly(n) factor of the

largest column norm of $\mathbf{A}$, since otherwise $\|\mathbf{E}\|_F$ would necessarily be larger than $\sigma_{\max}(\boldsymbol{U}^*)\frac{\sqrt{nk}}{\varepsilon}$. Given this column norm, we obtain an interval $[a,b] \subset \mathcal{R}$ $\frac{a}{b} = \text{poly}(n,\kappa)$, such that both κ and $\sigma_{\min}(\boldsymbol{U}^*)$ must live inside $[a,b]$. Then we can make $O(\log^2(\frac{a}{b})) = O(\log^2(n\kappa))$ guesses to find κ and $\sigma_{\min}(\boldsymbol{U}^*)$ up to a factor of 2. Thus guessing the correct approximations to $\kappa, \sigma_{\min}(\boldsymbol{U}^*)$ will not effect our run time bounds, since our overall complexity is already polynomial in n and κ. Similarly, we can also guess the value of $\|\mathbf{M}\mathcal{S}\mathbf{E}\|_F$ up to a factor of 2 using $O(\log(n\kappa))$ guesses, as is needed in step $3a$ of Algorithm 8.

The following Proposition gives the error bound needed for the right guess of the inverse $(\mathcal{S}\boldsymbol{U})^{-1}$

Proposition 6.7.3. *On the correct guess of $\sigma_{\max}(\mathcal{S}\boldsymbol{U}^*)$ (up to a constant factor of 2 error), there is an $i \in [\nu]$ such that $\mathbf{M}^i = (\mathcal{S}\boldsymbol{U}^*)^{-1} + \boldsymbol{\Lambda}$ where $\|\boldsymbol{\Lambda}\|_\infty \leqslant \frac{\varepsilon^4}{\sigma_{\min}\kappa^4 k^4}$.*

Proof. First note that no entry in $(\mathcal{S}\boldsymbol{U}^*)^{-1}$ can be greater than $\frac{1}{\sigma_{\min}}$ (since $\sigma_{\min}$ is the smallest singular value of $\mathcal{S}\boldsymbol{U}^*$, and therefore $\frac{1}{\sigma_{\min}}$ is the largest singular value of $(\mathcal{S}\boldsymbol{U}^*)^{-1}$. Thus there is a guess of $\mathbf{M}^i$ such that for each entry (p,q) of $(\mathcal{S}\boldsymbol{U}^*)^{-1}$ in the range $(\frac{1}{\sigma_{\min}(1/\varepsilon)^4\kappa^4k^4}, \frac{1}{\sigma_{\min}})$, we have $\mathbf{M}^i_{p,q} = (\mathcal{S}\boldsymbol{U}^*)^{-1}_{p,q}(1 \pm \frac{1}{(1/\varepsilon)^4\kappa^4k^4}) = (\mathcal{S}\boldsymbol{U}^*)^{-1} \pm \frac{1}{\sigma_{\min}(1/\varepsilon)^4\kappa^4k^4}$. For all other entries less than $\frac{1}{\sigma_{\min}(1/\varepsilon)^4\kappa^4k^4}$, we get $\mathbf{M}^i_{p,q} = (\mathcal{S}\boldsymbol{U}^*)^{-1}_{p,q} \pm \frac{1}{\sigma_{\min}(1/\varepsilon)^4\kappa^4k^4}$ by setting $\mathbf{M}^i_{p,q} = \frac{1}{\sigma_{\min}(1/\varepsilon)^4\kappa^4k^4}$ (which is the lowest guess of value which we make for the coordinates of $\mathbf{M}^i$), from which the proposition follows. □

6.7.3 Learning Noisy Halfspaces:

By Proposition 6.7.3, we know that for the correct guess of $\mathbf{M}^i$ we can write $\mathbf{M}^i\mathcal{S}\mathbf{A} = f(\boldsymbol{V}^*\mathcal{X}) + (\mathbf{M}^i\mathcal{S}\mathbf{E}) + \mathbf{Z}$ where $\mathbf{Z} = \boldsymbol{\Lambda}\mathcal{S}\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$. Thus $\mathbf{M}^i\mathcal{S}\mathbf{A}$ can be thought of as a noisy version of $f(\boldsymbol{V}^*\mathcal{X})$. We observe now that our problem can be viewed as the problem of learning a halfspace in $\mathcal{R}^d$ with noisy labels. Specifically, we are obtain examples of the form $\mathcal{X}_{*,q}$ with the label $y_q = \text{Sign}\Big(f(\boldsymbol{V}^*_{p,*}\mathcal{X}_{*:q}) + (\mathbf{M}^i\mathcal{S}\mathbf{E})_{p,q} + \mathbf{Z}_{p,q}\Big) \in \{1,-1\}$, and our goal is to recover $\boldsymbol{V}^*_{p,*}$ from these labeled examples $\{y_q\}$. Note that if the labeled examples were of the form $\mathcal{X}_{*,q}$ and $\text{Sign}(\langle \boldsymbol{V}^*_{p,*}, \mathcal{X}_{*,q}\rangle)$, then this would correspond to the noiseless learning problem for half-spaces. Unfortunately, our problem is not noiseless, as it will often be the case that $\text{Sign}\Big(f(\boldsymbol{V}_{p,*}\mathcal{X}_{*:q}) + (\mathbf{M}^i\mathcal{S}\mathbf{E})_{p,q} + \mathbf{Z}_{p,q}\Big) \neq \text{Sign}(\langle \boldsymbol{V}_{p,*}, \mathcal{X}_{*,q}\rangle)$ (in fact, this will happen very close to half of the time). We will demonstrate, however, that recovery of $\boldsymbol{V}_{p,*}$ is still possible by showing that there is a non-trivial correlation between the labels y_q and the true sign. To do this, we show the following more general result.

Theorem 141. *Given n i.i.d. Gaussian examples $\mathcal{X} \in \mathcal{R}^{d\times n}$ with labels $y_q = Sign\Big((f(\boldsymbol{V}\mathcal{X}) + \boldsymbol{G} + \mathbf{B})_{p,q}\Big) \in \{1,-1\}$ where $\boldsymbol{G}$ is an arbitrary fixed matrix independent of $\mathcal{X}$, and $\mathbf{B}$ is any matrix such that $\|\mathbf{B}\|_F \leqslant \frac{\sqrt{n}}{\omega}$ for any $\omega = o(\sqrt{n})$. Then if $v_{p,*}$ is the solution to the convex program in step 3c of Figure 8 run on the inputs examples $\mathcal{X}$ and $\{y_q\}$, then with probability $1 - e^{-n^{1/2}/10}$ we have*

$$\|v_{p,*} - \boldsymbol{V}_{p,*}\|_2^2 = O\left(\sqrt{\omega}\frac{\|\boldsymbol{G}\|_F}{\sqrt{n}}\left(\frac{\sqrt{d}}{\sqrt{n}} + \frac{1}{n^{1/4}} + \frac{\log(\omega)}{\omega}\right)\right)$$

Before we prove the theorem, we first show that our setting fits into this model. Observe that in our setting, $\boldsymbol{G} = \mathbf{M}^i\mathcal{S}\mathbf{E}$, and $\mathbf{B} = \mathbf{Z}$. Note that the Gaussian matrix added in Step 3a of Algorithm 8 is a component of proof of Theorem 141, and different than the $\boldsymbol{G}$ here. Namely, for Theorem 141 to work, one must first add Gaussian matrix to *smear out* the fixed noise matrix $\mathbf{M}^i\mathcal{S}\mathbf{E}$. See the proof of Theorem 141 for further details. The following Proposition formally relates our setting to that of Theorem 141.

Proposition 6.7.4. *We have $\|(\mathbf{M}^i\mathcal{S}\mathbf{E})\|_F = O(\frac{1}{\sigma_{\min}}\sqrt{m}\|\mathbf{E}\|_2)$, and $\|\mathbf{Z}\|_F = \|\boldsymbol{\Lambda}\mathcal{S}\boldsymbol{U}^* f(\boldsymbol{V}\mathcal{X})\|_2 \leqslant \sqrt{n}\frac{2}{(1/\varepsilon)^4\kappa^4k^2}$*

Proof. Since $\mathcal{S}\boldsymbol{U}^*$ is $\kappa = \sigma_{\max}/\sigma_{\min}$ conditioned (as conditioned on by the success of $\mathcal{S}$ as a subspace embedding for $\boldsymbol{U}^*$), it follows that for any row p, we have $\|((\mathcal{S}\boldsymbol{U}^*)^{-1})_{p,*}\|_2 \leqslant \frac{1}{\sigma_{\min}}$. Thus by Proposition 6.7.3 we have $\|\mathbf{M}^i_{p,*}\|_2 \leqslant \frac{1}{\sigma_{\min}} + \frac{1}{\sigma_{\min}(1/\varepsilon)^4\kappa^4k^3} \leqslant \frac{2}{\sigma_{\min}}$, and by Proposition 6.4.18, noting that $\mathcal{S}$ can be written as a i.i.d. matrix of $\mathcal{N}(0,1)$ variables scaled by $\frac{1}{\sqrt{k}}$, we have $\|\mathbf{M}^i_{p,*}\mathcal{S}\|_2 \leqslant \frac{2}{\sigma_{\min}}\sqrt{\frac{2m}{k}}$. Applying this over all $O(k)$ rows, it follows that $\|\mathbf{M}^i\mathcal{S}\mathbf{E}\|_F = O(\frac{1}{\sigma_{\min}}\sqrt{m}\|\mathbf{E}\|_2)$, where $\|\mathbf{E}\|_2$ is the spectral norm of $\mathbf{E}$.

For the second, note the bound $\|\boldsymbol{\Lambda}\|_\infty \leqslant 1/(\sigma_{\min}(1/\varepsilon)^4\kappa^4k^4)$ from Proposition 6.7.3 implies that $\|\boldsymbol{\Lambda}_{p,*}\|_2 \leqslant 1/(\sigma_{\min}(1/\varepsilon)^4\kappa^4k^3)$ (using that $k > c_1/\delta$ where δ is as in Figure 8), so $\|\boldsymbol{\Lambda}_{p,*}\mathcal{S}\boldsymbol{U}^*\|_2 \leqslant \frac{\sigma_{\max}}{\sigma_{\min}(1/\varepsilon)^4\kappa^4k^3} \leqslant \frac{1}{(1/\varepsilon)^4\kappa^4k^3}$. Now by Proposition 6.4.18, we have that the largest singular value of $\mathcal{X}$ is at most $2\sqrt{n}$ with probability at least $1 - 2e^{-n/8}$, which we now condition on. Thus $\|\boldsymbol{V}\mathcal{X}\|_F \leqslant 2\sqrt{nk}$, from which it follows $\|f(\boldsymbol{V}\mathcal{X})\|_F \leqslant 2\sqrt{nk}$, giving $\|\mathbf{Z}_{p,*}\|_2 \leqslant 2\sqrt{n}\frac{1}{(1/\varepsilon)^4\kappa^4k^{5/2}}$ for every $p \in [k]$, so $\|\mathbf{Z}\|_F \leqslant \sqrt{n}\frac{2}{(1/\varepsilon)^4\kappa^4k^2}$ as needed.

□

By Theorem 141 and Proposition 6.7.4, we obtain the following result.

Corollary 6.7.5. *Let i be such that $\mathbf{M}^i = (\mathcal{S}\boldsymbol{U}^*)^{-1} + \boldsymbol{\Lambda}$, where $\|\boldsymbol{\Lambda}\|_\infty \leqslant 1/(\sigma_{\min}(1/\varepsilon)^4\kappa^4k^4)$*

as in Proposition 6.7.3, and let $\boldsymbol{W}^i$ be the solution to the convex program as defined in Step 3d of the algorithm in Figure 8. Then with probability $1 - \exp(-\sqrt{n}/20)$, for every row $p \in [k]$ we have

$$\|\boldsymbol{V}_{p,*} - \boldsymbol{W}^i_{p,*}\|_2^2 \leqslant \frac{\varepsilon\sqrt{m}\|\mathbf{E}\|_2}{\sigma_{\min}\sqrt{n}}$$

Proof. By Proposition 6.7.4 we can apply Theorem 141 with $\omega = \varepsilon^{-4}\kappa^4 k^2$ and $\|\boldsymbol{G}\|_F = O(\frac{1}{\sigma_{\min}}\sqrt{m}\|\mathbf{E}\|_2)$, we obtain the stated result for a single row p with probability at least $1 - e^{-\sqrt{n}/10}$ after taking $n = \text{poly}(\kappa, d)$ sufficiently large. Union bounding over all k rows gives the desired result.

□

Proof of Theorem 141 To prove the theorem, we will use techniques from [PV13]. Let $v \in \mathcal{R}^d$ be fixed with $\|v\|_2 = 1$, and let $\mathcal{X} \in \mathcal{R}^{d\times n}$ be a matrix of i.i.d. Gaussian $\mathcal{N}(0,1)$ variables. Let y_q be a noisy observation of the value $\text{sgn}(\langle v, \mathcal{X}_{*,q}\rangle)$, such that the y_q's are independent for different q. We say that the y_q's are *symmetric* if $\mathbb{E}\left[y_q \mid \mathcal{X}_{*,q}\right] = \theta_q(\langle v, \mathcal{X}_{*,q}\rangle)$ for each $q \in [n]$. In other words, the expectation of the noisy label y_q given the value of the sample $\mathcal{X}_{*,q}$ depends only on the value of the inner product $\langle v, \mathcal{X}_{*,q}\rangle$. We consider now the following requirement relating to the correlation between y_q and $\text{sgn}(\langle v, \mathcal{X}_{*,q}\rangle)$.

$$\mathbb{E}_{g\sim\mathcal{N}(0,1)}\Big[\theta_q(g)g\Big] = \lambda_q \geqslant 0 \tag{6.4}$$

Note that the Gaussian g in Equation 6.4 can be replaced with the identically distributed variable $\langle v, X_{*,q}\rangle$. In this case, Equation 6.4 simply asserts that there is indeed some correlation between the observed labels y_q and the ground truth $\text{sgn}(\langle v, \mathcal{X}_{*,q}\rangle)$. When this is the case, the following convex program is proposed in [PV13] for recovery of v

$$\max_{w,\,\|w\|_2\leqslant 1} \sum_{q=1}^{n} y_q\langle w, X_{*,q}\rangle \tag{6.5}$$

We remark that we must generalize the results of [PV13] here in order to account for θ_q depending on q. Namely, since $\mathbf{E}$ is not identically distribution, we must demonstrate bounds on the solution to the above convex program for the range of parameters $\{\lambda_q\}_{q\in[n]}$.

Now fix a row $p \in [k]$ and let $v = \boldsymbol{V}^*_{p,*}$. We will write $\boldsymbol{G}' = \boldsymbol{G} + \boldsymbol{G}''$, where $\boldsymbol{G}''$ is an i.i.d. Gaussian matrix distributed $(\boldsymbol{G}'')_{i,j} \sim \mathcal{N}(0, \eta^2)$ for all $i, j \in [k] \times [n]$, where $\eta = 100\sqrt{\omega}\|\boldsymbol{G}\|_F/\sqrt{n}$. For technical reasons, we replace the matrix $\boldsymbol{G}$ with $\boldsymbol{G}'$ be generating and

adding $\boldsymbol{G}''$ to our matrix $(f(\boldsymbol{V}\mathcal{X}) + \boldsymbol{G} + \mathbf{B})$. Then the setting of Theorem 141, we have $y_q = \text{Sign}((f(\boldsymbol{V}^*\mathcal{X}) + \boldsymbol{G}' + \mathbf{B})_{p,q})$. Note that by the definition of η, at most $\frac{n}{100\omega}$ entries in $\boldsymbol{G}$ can be larger than $\eta/10 = 10\sqrt{\omega}\|\boldsymbol{G}\|_F/\sqrt{n}$. Let $\mathbf{B}'$ be the matrix of entries of $\boldsymbol{G}$ which do not satisfy this, so we instead write $y_q = \text{Sign}((f(\boldsymbol{V}^*\mathcal{X}) + \boldsymbol{G}' + \mathbf{B}' + \mathbf{B})_{p,q})$, where $\boldsymbol{G}' = \boldsymbol{G} + \boldsymbol{G}'' - \mathbf{B}'$. Thus $\boldsymbol{G}'_{p,q} \sim \mathcal{N}(\mu_{p,q}, \eta^2)$ where $\mu_{p,q} < 10\sqrt{\omega}\|\boldsymbol{G}\|_F/\sqrt{n} = \eta/10$. Note that $\mathbf{B}'$ is $\frac{n}{100\omega}$ sparse, as just argued.

Note that the above model does not fully align with the aforementioned model, because $\mathbf{B}$ is an arbitrary matrix that can depend potentially on $f(\boldsymbol{V}^*\mathcal{X})$, and not just $\langle \boldsymbol{V}^*_{p,*}, \mathcal{X}_{*,q}\rangle$. So instead, suppose hypothetically that in the place of y_q we were given the labels $y'_q = \text{Sign}((f(\boldsymbol{V}^*\mathcal{X}) + \boldsymbol{G}')_{p,q})$, which indeed satisfies the above model. Note that we have also removed $\mathbf{B}'$ from the definition of y'_q, since we will handle it at the same time as we handle $\mathbf{B}$. In this case we can write $\mathbb{E}\left[y'_q \mid \mathcal{X}_{*,q}\right] = \mathbb{E}_g[\text{sign}(f(\langle \mathcal{X}_{*,q}, \boldsymbol{V}^*_{p,*}\rangle) + g_{p,q}) \;\middle|\; \langle \mathcal{X}_{*,q}, \boldsymbol{V}_{p,*}\rangle]$ where $g_{p,q} \sim \mathcal{N}(\boldsymbol{G}_{p,q} - \mathbf{B}'_{p,q}, \eta^2)$ is a Gaussian independent of X.

Proposition 6.7.6 gives the corresponding value of λ for this model.

Proposition 6.7.6. *The function θ_q as defined by the hypothetical labels y'_q satisfies Equation 6.4 with $\lambda_q \geqslant \frac{c}{\eta}$ for some constant $c > 0$, where $\eta = 100\sqrt{\omega}\|\boldsymbol{G}\|_F/\sqrt{n}$.*

Proof. We can write $\mathbb{E}\left[y'_q \mid \mathcal{X}_{*,q}\right] = \mathbb{E}[\text{sign}(f(\langle \mathcal{X}_{*,q}, \boldsymbol{V}_{p,*}\rangle) + g_{p,q}) \;\middle|\; \langle \mathcal{X}_{*,q}, \boldsymbol{V}_{p,*}\rangle]$ where $g_{p,q} \sim \mathcal{N}(\boldsymbol{G}_{p,q} - \mathbf{B}_{p,q}, \eta^2)$. Let $\mu_q = \boldsymbol{G}_{p,q} - \mathbf{B}_{p,q}$ (for a fixed row p. Then $\theta(z) = 1 - 2\Pr[g \leqslant -f(z)]$, and Equation 6.4 can be evaluated by integration by parts. Let $p_q(z) = \frac{1}{\sqrt{2\pi\eta^2}} e^{-\frac{(z-\mu_q)^2}{2\eta^2}}$ is the p.d.f. of $g_{p,q}$. Note by the prior paragraphs we have $\eta^2 > 10\mu_q^2$ for all q. Then we have

$$\begin{aligned}
\lambda &= \mathbb{E}\left[\theta'(g)\right] = \mathbb{E}\left[2p(-f(z))\right] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sqrt{\frac{2}{\pi(\eta^2)}} e^{-(f(z)+\mu_q)^2/(2(\eta^2))}\right] \\
&= \sqrt{\frac{2}{\pi(\eta^2)}}\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[e^{-\frac{f(z)^2+2\mu_q f(z)+\mu_q^2}{2\eta^2}}\right] \\
&= \Omega(\frac{1}{\eta})
\end{aligned}$$

□

Now for any $z \in \mathcal{R}^d$ with $\|z\|_2 \leqslant 1$, let $h(z) = \frac{1}{n}\sum_{q=1}^n y_q\langle z, X_{*,q}\rangle$, and let $h'(z) = \frac{1}{n}\sum_{q=1}^n y'_q\langle z, X_{*,q}\rangle$. Observe that the hypothetical function h' corresponds to the objective function of Equation 6.5

with values of y'_q which satisfy the model of 6.4, whereas h, corresponding to the labels y_q which we actually observe, does not. Let $B_2^d = \{x \in \mathcal{R}^d \mid \|x\|_2 \leqslant 1\}$ and let $\mathcal{B}_2^d = B - B = \{x - z \mid x, y \in B\}$ be the Minkowski difference. The following follows immediately from [PV13].

Lemma 6.7.7 (Lemma 4.1 [PV13])**.** *For any $z \in B_2^d$, we have $\mathbb{E}[h'(z)] = \frac{1}{n}\sum_{q=1}^n \lambda_q \langle z, \boldsymbol{V}_{p,*}\rangle$ and thus because h' is a linear function, we have*

$$\mathbb{E}[h'(\boldsymbol{V}_{p,*}) - h'(z)] = \mathbb{E}[h'(\boldsymbol{V}_{p,*} - z)] = \frac{1}{n}\sum_{q=1}^n \lambda_q(1 - \langle \boldsymbol{V}_{p,*}, z\rangle) \geqslant \frac{1}{n}\sum_{q=1}^n \frac{\lambda_q}{2}\|\boldsymbol{V}_{p,*} - z\|_2^2$$

We now cite Proposition 4.2 of [PV13]. We remark that while the proposition is stated for the concentration of the value of $h'(z)$ around its expectation when the λ_q are are all uniformly the same $\lambda_q = \lambda$, we observe that this fact has no bearing on the proof of Proposition 6.7.8 below. This is because only the $y_q \in \{1, -1\}$ depend on the λ_q's, and the concentration result of Proposition 6.7.8, in fact, holds for *any* possible values of the y_q's. Thus one could replace $h'(z)$ below with any function of the form $\hat{h}(z) = \frac{1}{n}\sum_{q=1}^n y_q\langle z, g_q\rangle$ for any values of $y_q \in \{1, -1\}$, and the following concentration result would hold as long as $\{g_q\}_{q\in[n]}$'s is a collection of independent $\mathcal{N}(0, \mathbb{I}_d)$ variables.

Proposition 6.7.8 (Proposition 4.2 [PV13])**.** *For each $t > 0$, we have*

$$\Pr[\sup_{z\in\mathcal{B}_2^d} \left|h'(z) - \mathbb{E}[h'(z)]\right| \geqslant \frac{4\sqrt{d}}{\sqrt{n}} + t] \leqslant 4\exp(-\frac{-nt^2}{8})$$

We now demonstrate how to utilize these technical results in our setting. First, however, we must bound $\sup_{z\in B}|h'(z) - h(z)|$, since in actuality we will need bounds on the value of $h(z)$. We first introduce a bound on the expected number of flips between the signs $y_{p,*}$ and $y'_{p,*}$.

Proposition 6.7.9. *Let $T = \{q \in [n] \mid y_q \neq y'_q\}$. Then with probability $1 - e^{-10\sqrt{n}}$, we have $|T| \leqslant 11\frac{n}{\omega}$.*

Proof. We have $\|\mathbf{B}_{p,*}\|_1 \leqslant \sqrt{n}\|\mathbf{B}_{p,*}\|_2 \leqslant n/\omega$ by the original assumption on $\mathbf{B}$ in Theorem 141. Then $\Pr[q \in T]$ is at most the probability $\boldsymbol{G}'_{p,q}$ is in some interval of size $2|\mathbf{B}_{p,q}|$, which is at most $2|\mathbf{B}_{p,q}|$ by the anti-concentration of Gaussians. Thus $\mathbb{E}[|T|] \leqslant 2\|\mathbf{B}_{p,*}\|_1 \leqslant 2n/\omega$, and by Chernoff bounds $\Pr[|T| > 10n/\omega] < e^{-10\sqrt{n}}$ as needed. To handle $\mathbf{B}'$, we simple recall that $\mathbf{B}'$ was $\frac{n}{100\omega}$ sparse, and thus can flip at most $\frac{n}{100\omega} < n/\omega$ signs. □

Proposition 6.7.10. *Let h, h' be defined as above. Let $\hat{w} \in B_2^r$ be the solution to the optimization problem*

$$\max_{w,\, \|w\|_2} n\, h(w) = \max_{w,\, \|w\|_2} \sum_{q=1}^{n} y_q \langle w, \mathcal{X}_{*,q} \rangle \tag{6.6}$$

Then if with probability $1 - \exp(-\sqrt{n})$ we have

$$\Pr[\sup_{z \in B} \left| h'(z) - h(z) \right| \leqslant \frac{3\log(\omega)}{\omega}] \geqslant 1 - e^{-\sqrt{n}}$$

Proof. Let $S \subset \{x \in \mathcal{R}^d \mid \|x\|_\infty \leqslant 1\}$ be an ε-net for $\varepsilon = 1/n^3$. Standard results demonstrate the existence of S with $|S| < 2^{12d\log(n)}$ (see e.g. [Ver10b, Woo14b]). Fix $z \in S$ and observe $|h'(z) - h(z)| = \frac{2}{n}\sum_{q \in T} |\langle z, \mathcal{X}_{*,q} \rangle|$. Note that we can assume $\|z\|_2 = 1$, since increasing the norm to be on the unit sphere can only make $|h'(z) - h(z)|$ larger. By Proposition 6.7.9, we have $|T| \leqslant n/\tau$, where $\tau = \frac{\omega}{11}$ with probability $1 - e^{-10\sqrt{n}}$, so we can let $\mathcal{F} = \{T' \subset [n] \mid |T'| \leqslant n/\tau\}$. Note $|\mathcal{F}| \leqslant n(e\tau)^{n/\tau}$. Fix $T' \in \mathcal{F}$. The sum $\sum_{q \in T'} |\langle z, \mathcal{X}_{*,q} \rangle|$ is distributed as the L_1 of a Gaussian $\mathcal{N}(0,1)$ vector in $|T'|$, dimensions, and is $\sqrt{|T'|}$-Lipschitz with respect to L_2, i.e. $|\|x\|_1 - \|y\|_1| \leqslant \|x - y\|_1 \leqslant \sqrt{|T'|}\|x - y\|_2$. So by Lipschitz concentration (see [Ver10b] (Proposition 5.34)), we have $\Pr[\frac{1}{n}\sum_{q \in T'} |\langle z, \mathcal{X}_{*,q} \rangle| > \frac{\log(e\tau)}{\tau}] \leqslant \exp(-\log^2(e\tau)n/\tau)$. We can then union bound over all $T' \in \mathcal{F}$ and $z \in S$ to obtain the result with probability

$$1 - \exp\Big(- \frac{n\log^2(e\tau)}{\tau} + \frac{n\log(e\tau)}{\tau} + \log(n) + 12r\log(n) \Big) > 1 - \exp\Big(- \log^2(\tau)n/(2\tau) \Big)$$

So let $\mathcal{E}_1$ be the event that $\sum_{q \in T'} |\langle z, \mathcal{X}_{*,q} \rangle| < \sqrt{\log(\tau)}|T'|$ for all $T' \in \mathcal{F}$ and $z \in S$. Now fix $w \in \mathcal{R}^d$ with $\|w\|_2 \leqslant 1$, and let $y \in S$ be such that $\|y - z\|_2 \leqslant 1/n^3$. Observing that h and h' are linear functions, we have $|h(z) - h'(z)| \leqslant |h(y) - h'(y)| + |h(z - y) - h'(z - y)| \leqslant \frac{\log(e\tau)}{\tau} + |h(z-y) - h'(z-y)|$. Now condition on the event $\mathcal{E}_2$ that $\|X\|_F^2 \leqslant 10nd$, where $\Pr[\mathcal{E}_2] > 1 - \exp(nd)$ by standard concentration results for χ^2 distributions [LM00]. Conditioned on $\mathcal{E}_2$ we have $|h(z-y)| + |h'(z-y)| \leqslant 4\sqrt{10nd}/n^3 \leqslant 1/\tau$, giving $|h(z) - h'(z)| \leqslant \frac{3\log(e\tau)}{\tau} < \frac{3\log(\omega)}{\omega}$, from which the proposition follows after union bounding over the events $\mathcal{E}_1, \mathcal{E}_2$ and Proposition 6.7.9, which hold with probability $1 - (\exp(-\log^2(\tau)n/(2\tau)) + \exp(nd) + \exp(-10\sqrt{n})) > 1 - \exp(-\sqrt{n})$.

□

Lemma 6.7.11. *Let $\hat{w}$ be the solution to the optimization Problem in Equation 6.5 for our input labels $y_q = sign((f(\boldsymbol{V}\mathcal{X}) + \boldsymbol{G}' + \mathbf{B} + \mathbf{B}')_{p,q})$. Then w with probability $1 - e^{-n^{1/2}/10}$, we have $\|\hat{w} - \boldsymbol{V}_{p,*}\|_2^2 = O(\sqrt{\omega}\|\boldsymbol{G}\|_F/\sqrt{n})\Big(\frac{4\sqrt{d}}{\sqrt{n}} + \frac{1}{n^{1/4}} + \frac{6\log(\omega)}{\omega}\Big)$ for some constant c.*

Proof. Applying Lemma 6.7.7, and a union bound over the probabilities of failure in Proposition 6.7.8 with $t = n^{1/4}$ and Proposition 6.7.10, we have

$$\begin{aligned}
0 &\leqslant h(\hat{w}) - h(\boldsymbol{V}_{p,*}) \\
&\leqslant h'(\hat{w}) - h'(\boldsymbol{V}_{p,*}) + \frac{6\log(\omega)}{\omega} \\
&= h'\left(\hat{w} - \boldsymbol{V}_{p,*}\right) + \frac{6\log(\omega)}{\omega} \\
&\leqslant \mathbb{E}\left[h'\left(\hat{w} - \boldsymbol{V}_{p,*}\right)\right] + \frac{4\sqrt{d}}{\sqrt{n}} + \frac{1}{n^{1/4}} + \frac{6\log(\omega)}{\omega} \\
&\leqslant -\frac{\lambda}{2}\|\hat{w} - \boldsymbol{V}_{p,*}\|_2^2 + \frac{4\sqrt{d}}{\sqrt{n}} + \frac{1}{n^{1/4}} + \frac{6\log(\omega)}{\omega}
\end{aligned}$$

Applying Proposition 6.7.6, which yields $\frac{1}{\lambda} = O(\eta) = O(\sqrt{\omega}\|\boldsymbol{G}\|_F/\sqrt{n})$ completes the proof. □

Proof of Theorem 141. The proof of the theorem follows directly from Lemma 6.7.11. □

6.7.4 Completing the Analysis

We will now need the following straightforward lemma to complete the proof.

Theorem 142. *Let* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}\mathcal{X}) + \mathbf{E}$ *be the input, where each entry of* $\mathcal{X} \in \mathcal{R}^{d\times n}$ *is i.i.d.* $\mathcal{N}(0,1)$ *and* $\mathbf{E}$ *independent of* $\mathcal{X}$*. Then the algorithm in Figure 8 outputs* $\boldsymbol{U} \in \mathcal{R}^{m\times k}, \boldsymbol{V} \in \mathcal{R}^{k\times d}$ *in time* $2^{O(k^2\log((1/\varepsilon)\kappa))}$*poly*$(n,d)$ *such that with probability* $1 - \exp(-\sqrt{n})$ *we have*

$$\|\mathbf{A} - \boldsymbol{U}f(\boldsymbol{V}\mathcal{X})\|_F \leqslant \|\mathbf{E}\|_F + O\left(\left[\sigma_{\min}\varepsilon\sqrt{nm}\|\mathbf{E}\|_2\right]^{1/2}\right)$$

Where $\|\mathbf{E}\|_2$ *is the spectral norm of* $\mathbf{E}$*.*

Proof. Let $\boldsymbol{W}^i \in \mathcal{R}^{k\times r}$ be as in Corollary 6.7.5. Then, taking $n = \text{poly}(d, \kappa, \frac{1}{\varepsilon})$ large enough, we have $\boldsymbol{W}^i = \boldsymbol{V} + \boldsymbol{\Gamma}$ where $\|\boldsymbol{\Gamma}_{p,*}\|_F^2 \leqslant \frac{\varepsilon k\sqrt{m}\|\mathbf{E}\|_2}{\sigma_{\min}\sqrt{n}}$ for each row p with probability $1 - \exp(-r^4)$ by Corollary 6.7.5. Then applying the spectral norm bound on Gaussian matrices from Proposition 6.4.18, we obtain that $\|\boldsymbol{V}_{p,*}\mathcal{X} - \boldsymbol{W}^i_{p,*}\mathcal{X}\|_F^2 = O(\sqrt{n}\frac{\varepsilon k\sqrt{m}\|\mathbf{E}\|_2}{\sigma_{\min}})$ with probability at least $1 - e^{-9n}$. Since f just takes the maximum of 0 and the input, it follows that $\|f(\boldsymbol{W}^i\mathcal{X}) - f(\boldsymbol{V}\mathcal{X})\|_F^2 = O(\sqrt{n}\frac{\varepsilon k\sqrt{m}\|\mathbf{E}\|_2}{\sigma_{\min}})$, and therefore $\|\boldsymbol{U}^* f(\boldsymbol{W}^i\mathcal{X}) - \boldsymbol{U}^* f(\boldsymbol{V}\mathcal{X})\|_F^2 = O(\sigma_{\max}^2\sqrt{n}\frac{\varepsilon k\sqrt{m}\|\mathbf{E}\|_2}{\sigma_{\min}})$, which is at most $O(\sigma_{\min}\varepsilon\sqrt{nm}\|\mathbf{E}\|_2)$ after rescaling ε by a $\frac{1}{\kappa^2 k}$ factor. Now if $\boldsymbol{U}$ is the minimizer to the

regression problem $\min_{\boldsymbol{U}} \|\mathbf{A} - \boldsymbol{U} f(\boldsymbol{W}^i \mathcal{X})\|_F^2$ in step 5 of Figure 8, then note

$$\|\mathbf{A} - \boldsymbol{U} f(\boldsymbol{W}^i \mathcal{X})\|_F \leqslant \|\mathbf{A} - \boldsymbol{U}^* f(\boldsymbol{W}^i \mathcal{X})\|_F \leqslant \|\mathbf{E}\|_F + O\Big(\Big[\sigma_{\min} \varepsilon \sqrt{nm} \|\mathbf{E}\|_2\Big]^{1/2}\Big)$$

as needed.

For the probability of failure, note that Corollary 6.7.5 holds with probability $1 - \exp(-\Omega(\sqrt{n}))$. To apply this, we needed only to condition on the fact that $\mathcal{S}$ was a subspace embedding for $\boldsymbol{U}$, which occurs with probability $99/100$ for a single attempt. Running the algorithm $O(n)$ times, by Hoeffding bounds at least one trial will be successful with probability $1 - \exp(-\Omega(\sqrt{n}))$ as desired. To analyze runtime, note that we try at most $\text{poly}(nd)$ guesses of $\mathcal{S}$ and guesses of $\sigma_{\min}$ and κ. Moreover, there are at most $\nu = (\frac{\kappa}{\varepsilon})^{O(k^2)}$ guesses $\mathbf{M}^i$ carried out in Step 2 of Figure 8). For every such guess, we run the optimization program in step $3c$. Since the program has a linear function and a convex unit ball constraint, it is will known that such programs can be solved in polynomial time [BV04]. Finally, the regression problem in step 4 is linear, and thus can be solved in $\text{poly}(n)$ time, which completes the proof.

□

6.8 A Polynomial Time Algorithm for Exact Weight Recovery with Sparse Noise

In this section, we examine recovery procedures for the weight matrices of a *low-rank* neural network in the presence of arbitrarily large sparse noise. Here, by low rank, we mean that $m > k$. It has frequently been observed in practice that many pre-trained neural-networks exhibit correlation and a low-rank structure [DSD+13, DZB+14]. Thus, in practice it is likely that k need not be as large as m to well-approximate the data.

More formally, we are given $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) + \mathbf{E}$ where $\mathbf{E}$ is some sparse noise matrix with possibly very large entries. We show that under the assumption that $\boldsymbol{U}^*$ has orthonormal columns and satisfies an incoherence assumptions (which is fairly standard in the numerical linear algebra community) [CR07, CR09, KMO10, CLMW11, JNS13, Har14], we can recover the weights $\boldsymbol{U}^*, \boldsymbol{V}^*$ exactly, even when the sparsity of the matrices is a constant fraction of the number of entries. Our algorithm utilizes results on the recovery of low-rank matrices in the presence of a sparse noise. The error matrix $\mathbf{E} \in \mathcal{R}^{m \times n}$ is a sparse matrix whose non-zero entries are uniformly chosen from the set of all coordinates of an arbitrary matrix $\overline{\mathbf{E}}$. Formally,

we define the following noise procedure:

Definition 6.8.1. *(Sparse Noise.) A matrix* $\mathbf{E}$ *is said to be generated from a* s*-sparse-noise procedure if there is an arbitrary matrix* $\overline{\mathbf{E}}$*, such that* $\mathbf{E}$ *is generated by setting all but* $s \leqslant mn$ *entries of* $\overline{\mathbf{E}}$ *to be* 0 *uniformly at random.*

Definition 6.8.2. *(Incoherence.) A rank* k *matrix* $\mathbf{M} \in \mathcal{R}^{m \times n}$ *is said to be* μ*-incoherent if* $svd(\mathbf{M}) = \mathbf{P}\Sigma\mathbf{Q}$ *is the singular value decomposition of* $\mathbf{M}$ *and*

$$\begin{aligned} \max_i \|\mathbf{P}^T e_i\|_2^2 &\leqslant \frac{\mu k}{m} \\ \max_i \|\mathbf{Q} e_i\|_2^2 &\leqslant \frac{\mu k}{n} \end{aligned} \tag{6.7}$$

and

$$\max_i \|\mathbf{P}\mathbf{Q}\|_\infty \leqslant \sqrt{\frac{\mu k}{nm}} \tag{6.8}$$

Remark 143. The values $\|\mathbf{P}e_i\|_2^2$ and $\|\mathbf{Q}e_i\|_2^2$ are known as the (left and right, respectively) *leverage-scores* of $\mathbf{M}$. For an excellent survey on leverage scores, we refer the reader to [M+11]. We note that the set of leverage scores of $\mathbf{M}$ does not depend on the choice of orthonormal basis $\mathbf{P}$ or $\mathbf{Q}$ [Woo14b]. Thus, to obtain the bounds given in Equation 6.7, it suffices let $\mathbf{P}$ be any matrix with orthonormal columns which spans the columns of $\mathbf{M}$, and similarly it suffices to let $\mathbf{Q}$ be any matrix with orthonormal rows which spans the rows of $\mathbf{M}$.

Lemma 6.8.3. *The entire matrix* $\boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$*, where* $\mathcal{X}$ *is i.i.d. Gaussian,* $(\boldsymbol{U}^*)^T, \boldsymbol{V}^*$ *have orthonormal rows, and* $\boldsymbol{U}^*$ *is* μ*-incoherent, meaning* $\max_i \|(\boldsymbol{U}^*)^T e_i\|_2^2 \leqslant \frac{\mu k}{m}$*, is* $\overline{\mu}$*-incoherent for*

$$\overline{\mu} = O\Big((\kappa(\boldsymbol{V}^*))^2 \sqrt{k \log(n) \mu} + \mu + (\kappa(\boldsymbol{V}^*))^4 \log(n)\Big)$$

Proof. For $t \in \{\max, \min\}$, let $\sigma_t = \sigma_t(\boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}))$. Let $\mathbf{P}\Sigma\mathbf{Q}$ be the SVD of $\boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$, and let For any i, since $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ are orthonormal we have

$$\begin{aligned} \|\mathbf{Q}^T e_i\|_2^2 \leqslant \frac{\|\Sigma\mathbf{Q}^T e_i\|_2^2}{\sigma_{min}^2} &= \frac{\|\boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X}) e_i\|_2^2}{\sigma_{min}^2} \\ &= \frac{\|f(\boldsymbol{V}^* \mathcal{X}) e_i\|_2^2}{\sigma_{min}^2} \\ &\leqslant \frac{\|\boldsymbol{V}^* \mathcal{X} e_i\|_2^2}{\sigma_{min}^2} \end{aligned}$$

Now each entry in of $\boldsymbol{V}^*\mathcal{X}$ is an i.i.d. Gaussian, and so is at most $10\sqrt{\log(n)}$ with probability $1 - e^{-10n}$, so $\|\boldsymbol{V}^*\mathcal{X}e_i\|_2^2 \leqslant 100k\log(n)$ with probability $1 - e^{-9n}$ by a union bound. Since the columns of $\boldsymbol{U}^*$ are orthonormal, $\sigma_{\min}^2 = \sigma_{\min}^2(f(\boldsymbol{V}^*\mathcal{X}))$, which is at least $\frac{n}{(\kappa(\boldsymbol{V}^*))^4}$ by Lemma 6.5.3. Thus we have that $\|\mathbf{Q}^T e_i\|_2^2 = O(k(\kappa(\boldsymbol{V}^*))^4\log(n)/n)$. This shows the $O((\kappa(\boldsymbol{V}^*))^4\log(n))$-incoherehnce for the second part of Equation 6.7, and the first part follows from the μ-incoherence assumption on U. The incoherence bound of $(\kappa(\boldsymbol{V}^*))^2\sqrt{k\log(n)\mu}$ for Equation 6.8 follows by applying Cauchy Schwartz to the LHS and using the bounds just obtained for Equation 6.7. □

Theorem 144. *(Extending Theorem 1.1 in [CLMW11].) If* $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X}) + \mathbf{E}$ *where* $\mathbf{E}$ *is produced by the sparsity procedure outlined above with* $s \leqslant \gamma nm$ *for a fixed constant* $\gamma > 0$*. Then if* $\boldsymbol{U}^*$ *has orthonormal columns, is* μ*-incoherent,* $\mathcal{X}$ *is Gaussian, and the sample complexity satisfies* $n = poly(d, m, k, \kappa(\boldsymbol{V}^*))$*, then there is a polynomial time algorithm which, given only* $\mathbf{A}$*, outputs both matrices* $\mathbf{M} = \boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ *and* $\mathbf{E}$*, given that* $k \leqslant \frac{m}{\overline{\mu}\log^2(n)}$*, where* $\overline{\mu} = O\Big((\kappa(\boldsymbol{V}^*))^2\sqrt{k\log(n)\mu} + \mu + (\kappa(\boldsymbol{V}^*))^4\log(n)\Big)$.

Proof. The results of [CLMW11] demonstrate that solving

$$\begin{aligned} \min_{\mathbf{Y}} \quad & \|\mathbf{A} - \mathbf{Y}\|_F^2 \\ \text{s.t.} \quad & \text{rank}(\mathbf{Y}) \leqslant k \end{aligned}$$

recovers the optimal low-rank matrix given that conditions of the previous lemma are satisfied. That is, if we do not care about running time, the above optimization problem recovers $\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$ exactly. However, the above problem is highly non-convex and instead we optimize over the nuclear norm.

$$\begin{aligned} \min_{\mathbf{Y},\mathbf{E}} \quad & \|\mathbf{Y}\|_* + \|\mathbf{E}\|_1 \\ \text{s.t.} \quad & \mathbf{Y} + \mathbf{E} = \mathbf{A} \end{aligned}$$

By Theorem 1.1 in [CLMW11], we know that the solution to the above problem is unique and equal to $\boldsymbol{U}^* f(\boldsymbol{V}^*\mathcal{X})$. It remains to show that the above optimization problem can be solved in polynomial time. Note, the objective function is convex. As mentioned in [LSW15], we can then run an interior point algorithm and it is well known that in order to achieve additive error ϵ, we need to iterate poly$(\log(1/\epsilon))$ times. Observe, for exact recovery we require a dual certificate that can verify optimality. Section 2.3 in [CLMW11] uses a modified analysis of the golfing scheme introduced by [Gro11] to create a dual certificate for the aforementioned convex

program. We observe that this construction of the dual is independent of the kind of factorization we desire and only requires $\mathbf{Y}$ to be rank k. Given that $\boldsymbol{U}^*, \boldsymbol{V}^*, \mathcal{X}, \mathbf{E}$ have polynomially bounded bit complexity, this immediately implies a polynomial time algorithm to recover $\boldsymbol{U}^* f(\boldsymbol{V}^* \mathcal{X})$ in unfactored form. □

As an immediate corollary of the above theorem, our exact algorithms of Section 6.4 can be applied to the matrix $\mathbf{M}$ of Theorem 144 to recover $\boldsymbol{U}^*, \boldsymbol{V}^*$. Formally,

Corollary 6.8.4. *Let $\boldsymbol{U}^* \in \mathcal{R}^{m \times k}, \boldsymbol{V}^* \in \mathcal{R}^{k \times d}$ be rank k matrices, where $\boldsymbol{U}^*$ has orthonormal columns, $\max_i \|(\boldsymbol{U}^*)^T e_i\|_2^2 \leqslant \frac{\mu k}{m}$ for some μ, and $k \leqslant \frac{m}{\overline{\mu} \log^2(n)}$, where $\overline{\mu} = O\Big((\kappa(\boldsymbol{V}^*))^2 \sqrt{k \log(n)} \mu + \mu + (\kappa(\boldsymbol{V}^*))^4 \log(n)\Big)$. Here $\kappa(\boldsymbol{V}^*)$ is the condition number of $\boldsymbol{V}^*$. Let $\mathbf{E}$ be generated from the s-sparsity procedure with $s = \gamma nm$ for some constant $\gamma > 0$ and let $\mathbf{A} = \boldsymbol{U}^* f(\boldsymbol{V} \mathcal{X}) + \mathbf{E}$. Suppose the sample complexity satisfies $n = poly(d, m, k, \kappa(\boldsymbol{V}^*))$ Then on i.i.d. Gaussian input $\mathcal{X}$ there is a poly(n) time algorithm that recovers $\boldsymbol{U}^*, \boldsymbol{V}^*$ exactly up to a permutation and positive scaling with high probability.*

Acknowledgements

The authors would like to thank Anima Anandkumar, Mark Bun, Rong Ge, Sam Hopkins and Rina Panigrahy for useful discussions.

Part II

Nearly Optimal Algorithms for Learning Latent Models

Chapter 7

Low-Rank Approximation with $1/\epsilon^{1/3}$ Matrix-Vector Products

7.1 Introduction

Iterative methods, and in particular Krylov subspace methods, are ubiquitous in scientific computing. Algorithms such as power iteration, Golub-Kahan Bidiagonalization, Arnoldi iteration, and the Lanczos iteration, are used in basic subroutines for matrix inversion, solving linear systems, linear programming, low-rank approximation, and numerous other fundamental linear algebra primitives [Saa81, LS13]. A common technique in the analysis of Krylov methods is the use of Chebyshev polynomials, which can be applied to the singular values of a matrix to implement an approximate interval or step function [MH02, Riv20]. Further, Chebyshev polynomials reduce the degree required to accurately approximate such functions, leading to significantly fewer iterations and faster running time. In this paper we investigate the power of Krylov methods for low-rank approximation in the matrix-vector product model.

The Matrix-Vector Product Model. In this model, there is an underlying matrix $\mathbf{A}$, which is often implicit, and for which the only access to $\mathbf{A}$ is via matrix-vector products. Namely, the algorithm chooses a query vector v^1, obtains the product $\mathbf{A} \cdot v^1$, chooses the next query vector v^2, which is any randomized function of v^1 and $\mathbf{A} \cdot v^1$, then receives $\mathbf{A} \cdot v^2$, and so on. If $\mathbf{A}$ is a non-symmetric matrix, we assume access to products of the form $\mathbf{A}^\top v$ as well. We refer to the minimal number q of queries needed by the algorithm to solve a problem with constant probability as the *query complexity*. We note that upper bounds on the query complexity

immediately translate to running time bounds for the RAM model, when $\mathbf{A}$ is explicit, since a matrix-vector product can be implemented in $\text{nnz}(\mathbf{A})$ time, i.e., the number of non-zero entries in the matrix. Since this model captures a large family of iterative methods, it is natural to ask whether Krylov subspace based methods yield optimal algorithms, where the complexity measure of interest is the number of matrix-vector products.

This model and related vector-matrix-vector query models were formalized for a number of problems in [SWYZ19, RWZ20], though the model is standard for measuring efficiency in scientific computing and numerical linear algebra, see, e.g., [BFG96]; in that literature, methods that use only matrix-vector products are called *matrix-free*. Subsequently, for the problem of estimating the top eigenvector, nearly tight bounds were obtained in [SAR18, BHSW20]. Also, for the problem of estimating the trace of a positive semidefinite matrix, tight bounds were obtained in [MMMW21] (see, also [WWZ14], where tight bounds were shown in the related vector-matrix-vector query model). For recovering a planted clique from a random graph, upper and lower bounds were obtained in [RWYZ21]. In the non-adaptive setting, where $v^1, \ldots, v^q$, are chosen before making any queries to $\mathbf{A}$, this is equivalent to the *sketching model*, which is thoroughly studied on its own (see, e.g., [Nel11, Woo14b]), and in the context of data streams [Mut05, LNW14b].

Why is the matrix $\mathbf{A}$ implicit? A small query complexity q leads to an algorithm running in time $\mathcal{O}(T(\mathbf{A}) \cdot q + P(n, d, q))$, where $T(\mathbf{A})$ is the time to multiply the $n \times d$ matrix $\mathbf{A}$ by an arbitrary vector, and $P(n, d, q)$ is the time needed to form the queries and process the query responses, which is typically small. When the matrix $\mathbf{A}$ is given as a list of $\text{nnz}(\mathbf{A})$ non-zero entries, then $T(\mathbf{A}) \leqslant \text{nnz}(\mathbf{A})$. However, in many problems $\mathbf{A}$ is not given explicitly, and it is too expensive to write $\mathbf{A}$ down. Indeed, one may be given $\mathbf{A}$ but want to compute a low-rank approximation to the "covariance" (Gram) matrix $\mathbf{A}^\top\mathbf{A}$, and computing $\mathbf{A}^\top\mathbf{A}$ is too slow [MW17a]. More generally, one may be given $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and a function $f : \mathbb{R} \to \mathbb{R}$, and want to compute matrix-vector products with the generalized matrix function $f(\mathbf{A}) = \mathbf{U}f(\mathbf{\Sigma})\mathbf{V}^\top$, where $\mathbf{U}$ has orthonormal columns, $\mathbf{V}^\top$ has orthonormal rows, $\mathbf{\Sigma}$ is a diagonal matrix, and f is applied entry-wise to each entry on the diagonal.

The covariance matrix corresponds to $f(x) = x^2$, and other common functions f include the matrix exponential $f(x) = e^x$ and low-degree polynomials. For instance, when $\mathbf{A}$ is the adjacency matrix of an undirected graph, $f(x) = x^3/6$ is used to count the number of triangles [Tso08, Avr10]. Yet another example is when $\mathbf{A}$ is the Hessian $\mathbf{H}$ of a neural network with a huge number of parameters, for which it is often impossible to compute or store the entire Hessian [GKX19]. Typically $\mathbf{H} \cdot v$, for any chosen vector v, is computed using Pearlmutter's

trick [Pea94]. However, even with Pearlmutter's trick and distributed computation on modern GPUs, it takes 20 hours to compute the eigendensity of a single Hessian $\mathbf{H}$ with respect to the cross-entropy loss on the CIFAR-10 dataset from a set of fixed weights for ResNet-18 [KH+09], which has approximately 11 million parameters [HZRS16, GKX19]. This time is directly proportional to the number of matrix-vector products, and therefore minimizing this quantity is crucial.

Algorithms and Lower Bounds for Low-Rank Approximation. The low-rank approximation problem is well studied in numerical linear algebra, with countless applications to clustering, data mining, principal component analysis, recommendation systems, and many more. (For surveys on low-rank approximation, see the monographs [KV09, Mah11, Woo14b] and references therein.) In this problem, given an implicit $n \times d$ matrix $\mathbf{A}$, the goal is to output a matrix $\mathbf{Z} \in \mathbb{R}^{d \times k}$ with orthonormal columns such that

$$\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_X \leqslant (1+\epsilon) \min_{\mathbf{U}:\mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \|\mathbf{A}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\|_X, \tag{7.1}$$

where $\|\cdot\|_X$ denotes some norm. Note that given $\mathbf{Z}$, one can compute $\mathbf{A}\mathbf{Z}$ with an additional k queries, which will be negligible, and then $(\mathbf{A}\mathbf{Z}) \cdot \mathbf{Z}^\top$ is a rank-k matrix written in factored form, i.e., as the product of an $n \times k$ matrix and a $k \times d$ matrix. Among other things, low-rank approximation provides (1) a compression of $\mathbf{A}$ from nd parameters to $(n+d)k$ parameters, (2) faster matrix-vector products, since $\mathbf{A}\mathbf{Z} \cdot \mathbf{Z}^\top \cdot y$ can be computed in $O((n+d)k)$ time for an arbitrary vector y, as opposed to the $O(nd)$ time needed to compute $\mathbf{A} \cdot y$, and (3) de-noising, as often matrices $\mathbf{A}$ are close to low-rank (e.g., they are the product of latent factors) but only high rank due to noise.

Despite its tremendous importance, the optimal matrix-vector product complexity of low-rank approximation is unknown for any commonly used norm. The best known upper bound is due to Musco and Musco [MM15], who achieve $\tilde{\mathcal{O}}(k/\epsilon^{1/2})$ queries[1] for both the case when $\|\cdot\|_X$ is the commonly studied Frobenius norm $\|\mathbf{B}\|_F = \left(\sum_{i,j} \mathbf{B}_{i,j}^2\right)^{1/2}$ as well as when $\|\cdot\|_X$ is the Spectral (operator) norm $\|\mathbf{B}\|_2 = \sup_{\|y\|_2=1} \|\mathbf{B}y\|_2$.

On the lower bound front, there is a trivial lower bound of k, since $\mathbf{A}$ may be full rank and achieving (7.1) requires k matrix-vector products since one must reconstruct the column span of $\mathbf{A}$ exactly. However, *no lower bounds in terms of the approximation factor ϵ were known*. We note that Simchowitz, Alaoui and Recht [SAR18] prove lower bounds for approximating the top r eigenvalues of a symmetric matrix; however these guarantees are incomparable to those that

[1] We let $\tilde{\mathcal{O}}(f) = f \cdot \text{poly}(\log(dk/\epsilon))$.

follow from a low-rank approximation, even when the norm $\|\cdot\|_X$ is the operator norm (see Appendix 7.6 for a brief discussion).

Relationship to the Sketching Literature. Low-rank approximation has been extensively studied in the sketching literature which, when $\mathbf{A}$ is given explicitly, can achieve $\mathcal{O}(\text{nnz}(\mathbf{A}))$ time both for the Frobenius norm [CW13, MM13a, NN13a], as well as for Schatten-p norms [LW20]. However, these works require reading all of the entries in $\mathbf{A}$, and thus do not apply to any of the settings mentioned above. Further, the matrix-vector query model is especially important for problems such as trace estimation, where a low-rank approximation is used to first reduce the variance [MMMW21]. As trace estimation is often applied to implicit matrices, e.g., in computing Stochastic Lanczos Quadrature (SLQ) for Hessian eigendensity estimation [GKX19], in studying the effects of batch normalization and residual connections in neural networks [YGKM20], and in computing a disentanglement regularizer for deep generative models [PPZ+20], sketching algorithms for low-rank approximation often do not apply.

Another important application is low-rank approximation of covariance matrices [MW17a], for which the covariance matrix is not given explicitly. Here, we have a data matrix $\mathbf{A}$ and we want a low-rank approximation for $\mathbf{A}\mathbf{A}^\top$. Even when $\mathbf{S}$ is a sparse sketching matrix, the matrix $\mathbf{S}\mathbf{A}$ is no longer sparse, and one needs to multiply $\mathbf{S}\mathbf{A}$ by $\mathbf{A}^\top$ to obtain a sketch of $\mathbf{S}\mathbf{A}\mathbf{A}^\top$, which is a dense matrix-matrix multiplication. Moreover, when viewed in the matrix-vector product model, sketching algorithms obtain provably worse query complexity than existing iterative algorithms (see Table 1.2 for a comparison). Further, as modern GPUs often do not exploit sparsity, *even when the matrix $\mathbf{A}$ is given, a GPU may not be able to take advantage of sparse queries*, which means the total time taken is proportional to the number of matrix-vector products.

Motivating Schatten-p Norms. The Schatten norms for $1 \leqslant p < 2$ are more robust than the Frobenius norm, as they dampen the effect of large singular values. In particular, the Schatten-1 norm, also known as the nuclear norm, has been widely used for robust PCA [XCS10, CLMW11, YPCC16] as well as a convex relaxation of matrix rank in matrix completion [CR09, CP10], low-dimensional Euclidean embeddings [RFP10, TDSL00, RS00], image denoising [GZZF14, GXM+17] and tensor completion [YZ16]. In contrast, for $p > 2$, Schatten norms are more sensitive to large singular values and provide an approximation to the operator norm. In particular, for a rank r matrix, it is easy to see that setting $p = \log(r)/\eta$ yields a $(1+\eta)$-approximation to the operator norm (i.e., $p = \infty$). While the Block Krylov algorithm of Musco and Musco [MM15] implies a matrix-vector query upper bound of $\tilde{\mathcal{O}}\left(k/\epsilon^{1/2}\right)$ for Schatten-∞ low-rank approximation,

the exact complexity of this problem remains an outstanding open problem. When $p > 2$, we can interpolate between Frobenius and operator norm, and setting p to be a large fixed constant can be a proxy for Schatten-∞ low-rank approximation, with significantly fewer matrix-vector products (see Theorem 150).

Our Central Question. The main question of our work is:

What is the matrix-vector product complexity of low-rank approximation for the Frobenius norm, and more generally, for other matrix norms?

7.1.1 Our Results

We begin by stating our results for Frobenius and more generally, Schatten-p norm low-rank approximation for any $p \geqslant 1$; see Table 1.2 for a summary.

Theorem 145 (Query Upper Bound, informal Theorem 150)**.** *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, a target rank $k \in [d]$, an accuracy parameter $\epsilon \in (0,1)$ and any (not necessarily constant) $p \in [1, \mathcal{O}(\log(d)/\epsilon)]$, there exists an algorithm that uses $\tilde{\mathcal{O}}\left(kp^{1/6}/\epsilon^{1/3}\right)$ matrix-vector products and outputs a $d \times k$ matrix $\mathbf{Z}$ with orthonormal columns such that with probability at least $99/100$,*

$$\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} \leqslant (1+\epsilon) \min_{\mathbf{U}:\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \|\mathbf{A}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\|_{\mathcal{S}_p}.$$

When $p \geqslant \log(d)/\epsilon$, we get $\tilde{\mathcal{O}}\left(k/\sqrt{\epsilon}\right)$ matrix-vector products.

We note that for Frobenius norm low-rank approximation (Schatten p for $p = 2$), we improve the prior matrix-vector product bound of $\tilde{\mathcal{O}}(k/\epsilon^{1/2})$ by Musco and Musco [MM15] to $\tilde{\mathcal{O}}(k/\epsilon^{1/3})$. For Schatten-$p$ low-rank approximation for $p \in [1,2)$, we improve work of Li and Woodruff [LW20] who require query complexity at least $\Omega(k^{2/p}/\epsilon^{4/p+1})$, which is a polynomial factor worse in both k and $1/\epsilon$ than our $\tilde{\mathcal{O}}(k/\epsilon^{1/3})$ bound.

For $p > 2$, [LW20] obtain a query complexity of $\Omega(\min(n,d)^{1-2/p})$. We drastically improve this to $\tilde{\mathcal{O}}(k/\epsilon^{1/3})$, which does not depend on d or n at all. Setting $p = \log(d)/\epsilon$ suffices to obtain a $(1+\epsilon)$-approximation to the spectral norm ($p = \infty$), and we obtain an $\tilde{\mathcal{O}}\left(k/\sqrt{\epsilon}\right)$ query algorithm, matching the best known bounds for spectral low-rank approximation [MM15]. When $p > \log(d)/\epsilon$, we can simply run Block Krylov for $p = \infty$.

Remark 146 (Comments on the RAM Model)**.** Although our focus is on minimizing the num-

ber of matrix-vector products, which is the key resource in the applications described above, our bounds also improve the running time of low-rank approximation algorithms when the matrix $\mathbf{A}$ has a small number of non-zero entries and is explicitly given. For simplicity, we state our bounds and those of previous work without using algorithms for fast matrix multiplication; similar improvements hold when using such algorithms. When $\text{nnz}(\mathbf{A}) = O(n)$, for Frobenius norm low-rank approximation, work in the sketching literature, and in particular [ACW17] (building off of [CW13, NN13a, Coh16]), achieves $O(nk^2/\epsilon)$ time. In contrast, in this setting our runtime is $\tilde{\mathcal{O}}(nk^2/\epsilon^{2/3})$. Similarly, for Schatten-p low-rank approximation for $p \in [1, 2)$, the previous best [LW20] requires $\tilde{\Omega}(nk^{4/p}/\epsilon^{(8/p-2)})$ time, while for $p > 2$ [LW20] requires $\tilde{\Omega}(nd^{2(1-2/p)}(k/\epsilon)^{4/p})$ time. In both cases our runtime is only $\tilde{\mathcal{O}}(nk^2p^{1/3}/\epsilon^{2/3})$. We obtain analogous improvements when the sparsity $\text{nnz}(\mathbf{A})$ is allowed to be $n(k/\epsilon)^C$ for a small constant $C > 0$.

Next, we state our lower bounds on the matrix-vector query complexity of Schatten-p low-rank approximation.

Theorem 147 (Query Lower Bound for constant p, informal Theorem 153 and Theorem 155)**.** *Given $\varepsilon > 0$, and a fixed constant $p \geqslant 1$, there exists a distribution $\mathcal{D}$ over $n \times n$ matrices such that for $\mathbf{A} \sim \mathcal{D}$, any algorithm that with at least constant probability outputs a unit vector v such that $\|\mathbf{A}\left(\mathbf{I} - vv^\top\right)\|_{\mathcal{S}_p}^p \leqslant (1+\varepsilon)\min_{\|u\|_2=1}\|\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p}^p$ must perform $\Omega(1/\varepsilon^{1/3})$ matrix-vector queries to $\mathbf{A}$.*

Remark 148. We note that this is the first lower bound as a function of ϵ for this problem, even for the well-studied case of $p = 2$, achieving an $\Omega(1/\epsilon^{1/3})$ bound, which is tight for any constant k, simultaneously for all constant $p \geqslant 1$.

Remark 149. Braverman, Hazan, Simchowitz and Woodworth [BHSW20] and Simchowitz, Alaoui and Recht [SAR18] establish eigenvalue estimation lower bounds that we use in our arguments, but their results do not directly imply low-rank approximation lower bounds for any matrix norm that we are aware of, including spectral low-rank approximation, i.e., $p = \infty$ (see Appendix 7.6).

Matrix Polynomials and Streaming Algorithms. Since our algorithms are based on iterative methods, they generalize naturally to low-rank approximations of matrices of the form $\mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^\ell$ and $\left(\mathbf{A}^\top\mathbf{A}\right)^\ell$ for any integer ℓ, given $\mathbf{A}$ as input. We defer the details to Appendix 7.7.

Since we work in the matrix-vector model, our algorithms naturally extend to the multi-pass turnstile streaming setting. Notably, for $p > 2$, with $\mathcal{O}\Big(\log(d/\epsilon)p^{1/6}/\epsilon^{1/3}\Big)$ passes we are able to improve the $\tilde{\mathcal{O}}\left(n\left(\frac{kn^{1-2/p}}{\epsilon^2} + \frac{k^{2/p}+n^{1-2/p}}{\epsilon^{2+2/p}}\right)\right)$ memory bound of [LW20] to $\tilde{\mathcal{O}}\left(nk/\epsilon^{1/3}\right)$. We defer the details to Appendix 7.8.

7.2 Additional Related Work

Existing approaches to solve low-rank approximation problems under several norms fall into two broad categories: iterative methods and linear sketching. Iterative methods, such as Krylov subspace based methods, are captured by the matrix-vector product framework, whereas linear sketching allows for the choice of a matrix $\mathbf{S} \in \mathbb{R}^{t\times n}$, where t is the number of "queries", and then observes the product $\mathbf{S}\cdot\mathbf{A}$ and so on (see [Woo14b] and references therein). The model has important applications to streaming and distributed algorithms and several recent works have focused on estimating spectral norms and the top singular values [AN13, LNW14a, LW16b, BBK+21b], estimating Schatten and Ky-Fan norms [LW16b, LW17, LW16a, BKKS19] and low-rank approximation [CW13, MM13b, NN13a, BDN15, Coh16].

In addition to studying unitarily invariant norms, such as the Schatten norm, there also has been significant amount of work on studying low-rank approximation under matrix ℓ_p norms [SWZ17, BBB+19, SWZ20, MW21] and weighted low-rank approximation [SJ03, RSW16, BWZ19], settings in which the problem is known to be NP-Hard. Finally, there has been a recent flurry of work on sublinear time algorithms for low-rank approximation under various structural assumptions on the input [MW17c, BW18, IVWW19, SW19, BCW20b] and in quantum-inspired models [KP16, CLW18, Tan19, RSML18, GLT18, GSLW19, CCH+20].

7.3 Preliminaries

Given an $n \times d$ matrix $\mathbf{A}$ with rank r, and $n \geqslant d$, we can compute its singular value decomposition, denoted by $SVD(\mathbf{A}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, such that $\mathbf{U}$ is an $n \times r$ matrix with orthonormal columns, $\mathbf{V}^\top$ is an $r \times d$ matrix with orthonormal rows and $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix. The entries along the diagonal are the singular values of $\mathbf{A}$, denoted by $\sigma_1, \sigma_2 \ldots \sigma_r$. Given an integer $k \leqslant r$, we define the truncated singular value decomposition of $\mathbf{A}$ that zeros out all but the top k singular values of $\mathbf{A}$, i.e., $\mathbf{A}_k = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^\top$, where $\mathbf{\Sigma}_k$ has only k non-zero entries along the diagonal. It is well-known that the truncated SVD computes the best rank-k approximation to $\mathbf{A}$

under any unitarily invariant norm, but in particular for any Schatten-p norm (defined below), we have $\mathbf{A}_k = \min_{\text{rank}(\mathbf{X})=k} \|\mathbf{A} - \mathbf{X}\|_{\mathcal{S}_p}$. More generally, for any matrix $\mathbf{M}$, we use the notation $\mathbf{M}_k$ and $\mathbf{M}_{\setminus k}$ to denote the first k components and all but the first k components respectively. We use $\mathbf{M}_{i,*}$ and $\mathbf{M}_{*,j}$ to refer to the i^{th} row and j^{th} column of $\mathbf{M}$ respectively.

We use the notation $\mathbf{I}_k$ to denote a *truncated identity matrix*, that is, a square matrix with its top k diagonal entries equal to one, and all other entries zero. The dimension of $\mathbf{I}_k$ will be determined by context.

Definition 7.3.1 (Orthogonal Projection Matrices)**.** *Given a $d \times d$ symmetric matrix $\mathbf{P}$ and $k \in [d]$, $\mathbf{P}$ is a rank-k orthogonal projection matrix if rank$(\mathbf{P}) = k$ and $\mathbf{P}^2 = \mathbf{P}$.*

It follows from the above definition that $\mathbf{P}$ has eigenvalues that are either 0 or 1 and admits a singular value decomposition of the form $\mathbf{U}\mathbf{U}^\top$ where $\mathbf{U}$ has k orthonormal columns.

Definition 7.3.2 (Unitary Matrices)**.** *Given a symmetric matrix $\mathbf{U} \in \mathbb{R}^{d \times d}$ we say $\mathbf{U}$ is a unitary matrix if $\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$.*

Definition 7.3.3 (Rotation Matrices)**.** *Given a symmetric matrix $\mathbf{R} \in \mathbb{R}^{d \times d}$ we say $\mathbf{R}$ is a rotation matrix if $\mathbf{R}$ is unitary and det $(\mathbf{R}) = 1$.*

Fact 7.3.4 (Courant-Fischer for Singular Values)**.** *Given an $n \times d$ matrix $\mathbf{A}$ with singular values $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_d$, the following holds: for all $i \in [d]$,*

$$\sigma_i = \max_{S:\ dim(S)=i} \ \min_{x \in S:\ \|x\|_2=1} \|x^\top \mathbf{A}\|_2.$$

Fact 7.3.5 (Weyl's Inequality for Singular Values (see Exercise 22 [Tao20]))**.** *Given $n \times d$ matrices $\mathbf{X}, \mathbf{Y}$, for any $i, (j-1) \in [d]$ such that $i + j \leqslant d$,*

$$\sigma_{i+j}(\mathbf{X} + \mathbf{Y}) \leqslant \sigma_i(\mathbf{X}) + \sigma_{j+1}(\mathbf{Y}).$$

Fact 7.3.6 (Bernoulli's Inequality)**.** *For any $x, p \in \mathbb{R}$ such that $x \geqslant -1$ and $p \geqslant 1$, $(1+x)^p \geqslant 1 + px$.*

Schatten Norms and Trace Inequalities. We recall some basic facts for Schatten-p norms. We also require the following trace and operator inequalities.

Definition 7.3.7 (Schatten-p Norm)**.** *Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, let $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_d$ be the*

singular values of $\mathbf{A}$*. Then, for any* $p \in [0, \infty)$*, the Schatten-*p *norm of* $\mathbf{A}$ *is defined as*

$$\|\mathbf{A}\|_{\mathcal{S}_p} = Tr\left[\left(\mathbf{A}^\top \mathbf{A}\right)^{p/2}\right]^{1/p} = \left(\sum_{i\in[d]} \sigma_i^p(\mathbf{A})\right)^{1/p}.$$

Fact 7.3.8 (Schatten-p norms are Unitarily Invariant)**.** *Given an* $n \times d$ *matrix* $\mathbf{M}$*, for any* $m \times n$ *matrix* $\mathbf{U}$ *with orthonormal columns, a norm* $\|\cdot\|_X$ *is defined to be unitarily invariant if* $\|\mathbf{UM}\|_X = \|\mathbf{M}\|_X$*. The Schatten-*$p$ *norm is unitarily invariant for all* $p \geqslant 1$*.*

There exists a closed-form expression for the low-rank approximation problem under Schatten-p norms:

Fact 7.3.9 (Schatten-p Low-Rank Approximation)**.** *Given a matrix* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and an integer* $k \in \mathbb{N}$*,*

$$\mathbf{A}_k = \arg\min_{rank(\mathbf{X})\leqslant k} \|\mathbf{A} - \mathbf{X}\|_{\mathcal{S}_p},$$

where $\mathbf{A}_k$ *is the truncated SVD of* $\mathbf{A}$*.*

Fact 7.3.10 (Araki–Lieb–Thirring Inequality [Ara90])**.** *Given PSD matrices* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$*, for any* $r \geqslant 1$*, the following inequality holds:*

$$Tr\left[(\mathbf{BAB})^r\right] \leqslant Tr\left[\mathbf{B}^r\mathbf{A}^r\mathbf{B}^r\right].$$

Further, for $0 < r < 1$*, the reverse holds*

$$Tr\left[(\mathbf{BAB})^r\right] \geqslant Tr\left[\mathbf{B}^r\mathbf{A}^r\mathbf{B}^r\right].$$

Fact 7.3.11 (Mahler's Orthogonal Operator Inequality, Theorem 1.7 in [Mah90])**.** *Given* $p \geqslant 2$*, and matrices* $\mathbf{P}$ *and* $\mathbf{Q}$ *such that the row (column) span of* $\mathbf{P}$ *is orthogonal to the row (column) span of* $\mathbf{Q}$*, the following inequality holds:*

$$\|\mathbf{P}\|_{\mathcal{S}_p}^p + \|\mathbf{Q}\|_{\mathcal{S}_p}^p \leqslant \|\mathbf{P} + \mathbf{Q}\|_{\mathcal{S}_p}^p.$$

Fact 7.3.12 (Hölder's Inequality for Schatten-p Norms, Corollary 4.2.6 [Bha13])**.** *Given matrices* $\mathbf{A}, \mathbf{B}^\top \in \mathbb{R}^{n\times d}$ *and* $p \in [1, \infty)$*, the following holds*

$$\|\mathbf{AB}\|_{\mathcal{S}_p} \leqslant \|\mathbf{A}\|_{\mathcal{S}_q} \cdot \|\mathbf{B}\|_{\mathcal{S}_r},$$

for any q, r such that $\frac{1}{p} = \frac{1}{q} + \frac{1}{r}$.

We also require *pinching inequalities* that were originally introduced to relate norms for partitioned operators over direct sums of Hilbert spaces. In our context, these inequalities simplify to norm inequalities for block matrices:

Fact 7.3.13 (Pinching Inequalities for Schatten-p Norms, [BKL02])**.** *Let $\mathbf{M} \in \mathbb{R}^{td \times td}$ be the following block matrix*

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{(1,1)} & \mathbf{M}_{(1,2)} & \cdots & \mathbf{M}_{(1,t)} \\ \mathbf{M}_{(2,1)} & \mathbf{M}_{(2,2)} & \cdots & \mathbf{M}_{(1,t)} \\ \vdots & & \ddots & \vdots \\ \mathbf{M}_{(t,1)} & \mathbf{M}_{(t,2)} & \cdots & \mathbf{M}_{(t,t)} \end{bmatrix},$$

where for all $i, j \in [t]$, $\mathbf{M}_{(i,j)} \in \mathbb{R}^{d \times d}$. For all $p \geqslant 1$, the following inequality holds:

$$\left(\sum_{i \in [t]} \|\mathbf{M}_{(i,i)}\|_{\mathcal{S}_p}^p \right)^{1/p} \leqslant \|\mathbf{M}\|_{\mathcal{S}_p}.$$

We also require a norm compression inequality that is a special case of Conjecture 7.3.15 (and known to be true), when each block is aligned in the following sense:

Fact 7.3.14 (Aligned Norm Compression Inequality, Section 4.3 in [Aud08])**.** *Let $\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 & \mathbf{M}_2 \\ \mathbf{M}_3 & \mathbf{M}_4 \end{pmatrix}$ such that there exist scalars $\alpha_1, \alpha_2, \beta_1, \beta_2$ such that $\mathbf{M}_1 = \alpha_1 \mathbf{X}$, $\mathbf{M}_2 = \alpha_2 \mathbf{X}$, $\mathbf{M}_3 = \beta_1 \mathbf{Y}$ and $\mathbf{M}_4 = \beta_2 \mathbf{Y}$. Then, for any $p \geqslant 2$,*

$$\|\mathbf{M}\|_{\mathcal{S}_p} \leqslant \left\| \begin{pmatrix} \|\mathbf{M}_1\|_{\mathcal{S}_p} & \|\mathbf{M}_2\|_{\mathcal{S}_p} \\ \|\mathbf{M}_3\|_{\mathcal{S}_p} & \|\mathbf{M}_4\|_{\mathcal{S}_p} \end{pmatrix} \right\|_{\mathcal{S}_p}.$$

Finally, we note a related conjecture, Audenaert's Norm Compression Conjecture [Aud08], a question in functional analysis concerning operator inequalities (see also [AK12]):

Conjecture 7.3.15 (Schatten-p Norm Compression)**.** *Let $\mathbf{M}$ be a partitioned operator (block matrix) such that $\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 & \mathbf{M}_2 \\ \mathbf{M}_3 & \mathbf{M}_4 \end{pmatrix}$. Let $\mathbf{C}_{\mathbf{M},p} = \begin{pmatrix} \|\mathbf{M}_1\|_{\mathcal{S}_p} & \|\mathbf{M}_2\|_{\mathcal{S}_p} \\ \|\mathbf{M}_3\|_{\mathcal{S}_p} & \|\mathbf{M}_4\|_{\mathcal{S}_p} \end{pmatrix}$ be a 2×2 matrix that denotes the Schatten-p compression of $\mathbf{M}$ for any $p \geqslant 1$. Then, $\|\mathbf{M}\|_{\mathcal{S}_p} \geqslant \|\mathbf{C}_{\mathbf{M},p}\|_{\mathcal{S}_p}$ if $1 \leqslant p \leqslant 2$, and $\|\mathbf{M}\|_{\mathcal{S}_p} \leqslant \|\mathbf{C}_{\mathbf{M},p}\|_{\mathcal{S}_p}$ if $2 \leqslant p < \infty$.*

We only require special cases of this that are known to be true.

Random Matrix Theory. Next, we recall some basic facts for Wishart ensembles from random matrix theory (we refer the reader to [Tao12] for a comprehensive overview).

Definition 7.3.16 (Wishart Ensemble)**.** *An $n \times n$ matrix $\mathbf{W}$ is sampled from a Wishart Ensemble, $\mathit{Wishart}(n)$, if $\mathbf{W} = \mathbf{X}\mathbf{X}^\top$ such that for all $i, j \in [n]$ $\mathbf{X}_{i,j} \sim \mathcal{N}\left(0, \frac{1}{n}\mathbf{I}\right)$.*

Fact 7.3.17 (Norms of a Wishart Ensemble)**.** *Let $\mathbf{W} \sim \mathit{Wishart}(n)$ such that $n = \Omega(1/\varepsilon^3)$. Then, with probability $99/100$, $\|\mathbf{W}\|_{op} \leqslant 5$ and for any fixed constant p, $\|\mathbf{I} - \frac{1}{5}\mathbf{W}\|_{\mathcal{S}_p}^p = \Theta\left(\frac{1}{\varepsilon^{1/3}}\right)$.*

7.4 Algorithms for Schatten-p LRA

In this section, we focus on obtaining algorithms for low-rank approximation in Schatten-p norm, simultaneously for all real, not necessarily constant, $p \in [1, \mathcal{O}(\log(d)/\epsilon)]$. For the special case of $p \in \{2, \infty\}$, Musco and Musco [MM15] showed an algorithm with matrix-vector query complexity $\tilde{O}(k/\epsilon^{1/2})$, given below as Algorithm 152. We show that the number of matrix-vector products we require scales proportional to $\tilde{O}\left(kp^{1/6}/\epsilon^{1/3}\right)$ instead. Finally, recall when $p > \log(d)/\epsilon$, it suffices to run Block Krylov for $p = \infty$, which requires $\mathcal{O}(\log(d/\epsilon)k/\sqrt{\epsilon})$ matrix-vector products.

Theorem 150 (Optimal Schatten-p Low-Rank Approximation)**.** *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, a target rank $k \in [d]$, an accuracy parameter $\epsilon \in (0, 1)$ and any $p \in [1, \mathcal{O}(\log(d)/\epsilon)]$, Algorithm 151 performs $\mathcal{O}\left(\frac{kp^{1/6}\log(d/\epsilon)}{\epsilon^{1/3}} + \log(d/\epsilon)k\sqrt{p}\right)$ matrix-vector products and outputs a $d \times k$ matrix $\mathbf{Z}$ with orthonormal columns such that with probability at least $9/10$,*

$$\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} \leqslant (1+\epsilon) \min_{\mathbf{U}:\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \|\mathbf{A}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\|_{\mathcal{S}_p}.$$

Further, in the RAM model, the algorithm runs in time $\mathcal{O}\left(\frac{nnz(\mathbf{A})p^{1/6}k\log^2(d/\epsilon)}{\epsilon^{1/3}} + \frac{np^{(\omega-1)/6}k^{\omega-1}}{\epsilon^{(\omega-1)/3}}\right)$.

We first introduce the following lemmas from Musco and Musco [MM15] that provide convergence bounds for the performance of Block Krylov Iteration (Algorithm 152) :

Lemma 7.4.1 (Gap Independent Block Krylov with Arbitrary Accuracy)**.** *Let $\mathbf{A}$ be an $n \times d$ matrix, k be the target rank and $\gamma > 0$ be an accuracy parameter. Then, initializing Algorithm 152 with block size k and running for $q = \Omega\left(\log(d/\gamma)/\sqrt{\gamma}\right)$ iterations outputs a $d \times k$ matrix $\mathbf{Z}$*

such that with probability $99/100$*, for all* $i \in [k]$*,*

$$\|\mathbf{A}\mathbf{Z}_{*,i}\|_2^2 = \sigma_i^2 \pm \gamma\sigma_{k+1}^2.$$

Further, the total number of matrix-vector products is $\mathcal{O}(kq)$ *and the running time in the RAM model is* $\mathcal{O}\Big(nnz(\mathbf{A})kq + n\left(kq\right)^2 + \left(kq\right)^\omega\Big)$*.*

The aforementioned lemma follows directly from Theorem 1 in [MM15], using the per-vector error guarantee (3).

Lemma 7.4.2 (Gap Dependent Block Krylov, Theorem 13 [MM15])**.** *Let* $\mathbf{A}$ *be an* $n \times d$ *matrix and* $\gamma > 0$*, be an accuracy parameter and* $p, k \in \mathcal{N}$ *be such that* $b \geqslant k$*. Let* $\sigma_1, \sigma_2 \ldots \sigma_d$ *be the singular values of* $\mathbf{A}$*. Then, initializing Algorithm 152 with block size* b *and running for* $q = \Omega\left(\log(n/\gamma)\sqrt{\sigma_k}/\sqrt{\sigma_k - \sigma_b}\right)$ *iterations outputs a* $d \times k$ *matrix* $\mathbf{Z}$ *such that with probability* $99/100$*, for all* $i \in [k]$

$$\|\mathbf{A}\mathbf{Z}_{*,i}\|_2^2 = \sigma_i^2 \pm \gamma\sigma_{k+1}^2.$$

Further, the total number of matrix-vector products is $\mathcal{O}(pq)$ *and the running time in the RAM model is* $\mathcal{O}\Big(nnz(\mathbf{A})bq + n\left(bq\right)^2 + \left(bq\right)^\omega\Big)$*.*

Algorithm 151 (Optimal Schatten-p Low-rank Approximation)**.**

Input: *An* $n \times d$ *matrix* $\mathbf{A}$*, target rank* $k \leqslant d$*, accuracy parameter* $0 < \varepsilon < 1$*, and* $p \geqslant 1$*.*

1. *Let* $\gamma_1 = \varepsilon^{2/3}/p^{1/3}$*. Run Block Krylov Iteration (Algorithm 152) on* $\mathbf{A}^\top$ *with block size* k*, and number of iterations* $q = \mathcal{O}\Big(\log(d/\gamma_1)/\sqrt{\gamma_1} + \log(d/\epsilon)\sqrt{p}\Big)$*. Let* $\mathbf{W}_1 \in \mathbb{R}^{n\times k}$ *be the corresponding output with orthonormal columns.*

2. *Let* $\gamma_2 = \varepsilon$ *and let* $s = \mathcal{O}\Big(p^{-1/3}k/\varepsilon^{1/3}\Big)$*. Run Block Krylov Iteration (Algorithm 152) on* $\mathbf{A}^\top$ *with block size* s*, and number of iterations* $q = \mathcal{O}\Big(\log(d/\gamma_2)\sqrt{p}\Big)$*. Let* $\mathbf{W}_2 \in \mathbb{R}^{n\times k}$ *be the corresponding output with orthonormal columns.*

3. *Run Block Krylov on* $\mathbf{A}$ *with target rank* $k+1$ *and number of iterations* $q = \mathcal{O}\Big((\log(dp) + \log(d/\epsilon))\sqrt{p}\Big)$*, and let* $\hat{\mathbf{Z}}_1$ *be the resulting* $d\times(k+1)$ *output matrix. Compute* $\hat{\sigma}_1^2 = \|\mathbf{A}(\hat{\mathbf{Z}}_1)_{*,1}\|_2^2$ *and* $\hat{\sigma}_{k+1}^2 = \|\mathbf{A}(\hat{\mathbf{Z}}_1)_{*,k+1}\|_2^2$*, rough estimates of the* 1*-st and* $(k+1)$*-st singular values of* $\mathbf{A}$*. Run Block Krylov on* $\mathbf{A}$ *with target rank* s*, where* $s = \mathcal{O}\Big(p^{-1/3}k/\varepsilon^{1/3}\Big)$ *and iterations* $q = \mathcal{O}\Big(\log(d/\epsilon)\sqrt{p}\Big)$*, and let* $\hat{\mathbf{Z}}_2$ *be the resulting* $d \times s$ *output matrix. Compute* $\hat{\sigma}_s^2 = \|\mathbf{A}(\hat{\mathbf{Z}}_2)_{*,s}\|_2^2$*, an estimate to the*

s-th singular value of $\mathbf{A}$.

4. *If* $\hat{\sigma}_1^2 \geqslant (1+0.5/p)\hat{\sigma}_{k+1}^2$, *set* $\mathbf{Z} = \mathbf{Z}_1$. *Else, if* $\hat{\sigma}_s^2 \leqslant \hat{\sigma}_{k+1}^2/\left(1+0.5/p\right)$, *set* $\mathbf{Z}$ *to be an orthonormal basis for* $\mathbf{A}^\top \mathbf{W}_2 \mathbf{W}_2^\top$ *and otherwise set* $\mathbf{Z}$ *to be an orthonormal basis for* $\mathbf{A}^\top \mathbf{W}_1 \mathbf{W}_1^\top$.

Output: *A matrix* $\mathbf{Z} \in \mathbb{R}^{d\times k}$ *with orthonormal columns such that*

$$\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p \leqslant (1+\epsilon) \min_{\mathbf{U}:\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \|\mathbf{A}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\|_{\mathcal{S}_p}^p.$$

Next, we prove the following key lemma relating the Schatten-p norm of row and column projections applied to a matrix $\mathbf{A}$ to the Schatten-p norm of the matrix itself. We can interpret this lemma as an extension of the Pythagorean Theorem to Schatten-p spaces and believe this lemma is of independent interest. We note that we appeal to *pinching inequality* for partitioned operators to obtain this lemma.

Lemma 7.4.3 (Schatten-p Norms for Orthogonal Projections)**.** *Let* $\mathbf{A}$ *be an* $n \times d$ *matrix, let* $\mathbf{P}$ *be an* $n \times n$ *matrix, and let* $\mathbf{Q}$ *be a* $d \times d$ *matrix such that both* $\mathbf{P}$ *and* $\mathbf{Q}$ *are orthogonal projection matrices of rank* k *(see Definition 7.3.1). Then, the following inequality holds for all* $p \geqslant 1$*:*

$$\|\mathbf{A}\|_{\mathcal{S}_p}^p \geqslant \|\mathbf{P}\mathbf{A}\mathbf{Q}\|_{\mathcal{S}_p}^p + \|\left(\mathbf{I} - \mathbf{P}\right)\mathbf{A}\left(\mathbf{I} - \mathbf{Q}\right)\|_{\mathcal{S}_p}^p.$$

Proof. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the SVD of $\mathbf{A}$, where $\mathbf{U} \in \mathbb{R}^{n\times d}$ and $\mathbf{V}^\top \in \mathbb{R}^{d\times d}$ have orthonormal columns and rows respectively. We construct unitary matrices $\mathbf{R}$ and $\mathbf{S}$, such that $\mathbf{R} \in \mathbb{R}^{n\times n}$ and $\mathbf{S} \in \mathbb{R}^{d\times d}$ that satisfy the following constraints:

1. $\mathbf{R}^\top \mathbf{I}_k \mathbf{R}\mathbf{A}\mathbf{S}^\top \mathbf{I}_k \mathbf{S} = \mathbf{P}\mathbf{A}\mathbf{Q}$, and
2. $\mathbf{R}^\top \left(\mathbf{I} - \mathbf{I}_k\right)\mathbf{R}\mathbf{A}\mathbf{S}^\top \left(\mathbf{I} - \mathbf{I}_k\right)\mathbf{S} = \left(\mathbf{I} - \mathbf{P}\right)\mathbf{A}\left(\mathbf{I} - \mathbf{Q}\right)$,

where the trunctated Identity matrix, $\mathbf{I}_k$, left multiplying $\mathbf{A}$ is $n \times n$ and right multiplying $\mathbf{A}$ is $d \times d$.

Recall, since $\mathbf{P}$ is a rank-k projection matrix, it admits a decomposition $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$ such that $\mathbf{X}$ has k orthonormal columns and similarly $\mathbf{I} - \mathbf{P} = \mathbf{Y}\mathbf{Y}^\top$, where $\mathbf{Y}$ has $n - k$ orthonormal columns. Further, since $\mathbf{X}$ and $\mathbf{Y}$ span disjoint subspaces, and the union of their span is $\mathbb{R}^n$, the matrix $(\mathbf{X} \mid \mathbf{Y})$, obtained by concatenating their columns, is unitary. Then, it suffices to set

$\mathbf{R} = (\mathbf{X} \mid \mathbf{Y})^\top$. To see this, observe,

$$\mathbf{R}^\top \mathbf{I}_k \mathbf{R} = (\mathbf{X} \mid 0) \cdot \begin{pmatrix} \mathbf{X}^\top \\ 0 \end{pmatrix} = \mathbf{X}\mathbf{X}^\top = \mathbf{P},$$

and similarly,

$$\mathbf{R}^\top (\mathbf{I} - \mathbf{I}_k) \mathbf{R} = \mathbf{Y}\mathbf{Y}^\top = \mathbf{I} - \mathbf{P}.$$

We repeat the above argument for the projection matrix $\mathbf{Q}$. Let $\mathbf{Q} = \mathbf{W}\mathbf{W}^\top$, where $\mathbf{W}$ is $d \times k$ and has orthonormal columns, and $\mathbf{I} - \mathbf{Q} = \mathbf{Z}\mathbf{Z}^\top$, where $\mathbf{Z}$ is $d \times (d-k)$ and has orthonormal columns. Observe, it suffices to set $\mathbf{S} = (\mathbf{W} \mid \mathbf{Z})^\top$, since $\mathbf{S}$ is unitary and $\mathbf{S}^\top \mathbf{I}_k \mathbf{S} = \mathbf{Q}$ and $\mathbf{S}^\top (\mathbf{I} - \mathbf{I}_k) \mathbf{S} = \mathbf{I} - \mathbf{Q}$. Note, by construction, we satisfy the two aforementioned constraints.

Let $\hat{\mathbf{A}} = \mathbf{R}\mathbf{A}\mathbf{S}^\top$. Since $\mathbf{R}$ and $\mathbf{S}$ are unitary, it follows from unitary invariance of the Schatten-p norm that

$$\|\hat{\mathbf{A}}\|_{\mathcal{S}_p} = \|\mathbf{R}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \mathbf{S}^\top\|_{\mathcal{S}_p} = \|\mathbf{A}\|_{\mathcal{S}_p} \tag{7.2}$$

Further, observe for any $n \times d$ matrix $\mathbf{M}$, we have have the following block decomposition

$$\begin{aligned} \mathbf{M} &= \mathbf{I}_k \mathbf{M} \mathbf{I}_k + \mathbf{I}_k \mathbf{M} (\mathbf{I} - \mathbf{I}_k) + (\mathbf{I} - \mathbf{I}_k) \mathbf{M} \mathbf{I}_k + (\mathbf{I} - \mathbf{I}_k) \mathbf{M} (\mathbf{I} - \mathbf{I}_k) \\ &= \begin{pmatrix} \mathbf{M}_{1:k,1:k} & \mathbf{M}_{1:k,k+1:d} \\ \mathbf{M}_{k+1:n,1:k} & \mathbf{M}_{k+1:n,k+1:d} \end{pmatrix}, \end{aligned}$$

where the notation $\mathbf{M}_{i:i',j:j'}$ picks the $(i'-i+1) \times (j'-j+1)$ sized sub-matrix corresponding to the rows indices $[i, i']$ and column indices $[j, j']$. Since appending rows and columns of 0's does not change the singular values, we have $\|\mathbf{I}_k \mathbf{M} \mathbf{I}_k\|_{\mathcal{S}_p} = \|\mathbf{M}_{1:k,1:k}\|_{\mathcal{S}_p}$ and $\|(\mathbf{I} - \mathbf{I}_k) \mathbf{M} (\mathbf{I} - \mathbf{I}_k)\|_{\mathcal{S}_p} = \|\mathbf{M}_{k+1:n,k+1:d}\|_{\mathcal{S}_p}$. Setting $\mathbf{M} = \hat{\mathbf{A}}$, we have

$$\begin{aligned} \|\hat{\mathbf{A}}\|_{\mathcal{S}_p}^p &= \left\| \begin{pmatrix} \hat{\mathbf{A}}_{1:k,1:k} & \hat{\mathbf{A}}_{1:k,k+1:d} \\ \hat{\mathbf{A}}_{k+1:n,1:k} & \hat{\mathbf{A}}_{k+1:n,k+1:d} \end{pmatrix} \right\|_{\mathcal{S}_p}^p \\ &\geqslant \|\hat{\mathbf{A}}_{1:k,1:k}\|_{\mathcal{S}_p}^p + \|\hat{\mathbf{A}}_{k+1:n,k+1:d}\|_{\mathcal{S}_p}^p \\ &= \|\mathbf{I}_k \hat{\mathbf{A}} \mathbf{I}_k\|_{\mathcal{S}_p}^p + \|(\mathbf{I} - \mathbf{I}_k) \hat{\mathbf{A}} (\mathbf{I} - \mathbf{I}_k)\|_{\mathcal{S}_p}^p, \end{aligned} \tag{7.3}$$

where the inequality follows from using the *pinching inequality* on the block matrix (see Fact 7.3.13). By the unitary invariance of the Schatten-p norm, we have

$$\|\mathbf{I}_k \hat{\mathbf{A}} \mathbf{I}_k\|_{\mathcal{S}_p}^p = \|\mathbf{R}^\top \mathbf{I}_k \hat{\mathbf{A}} \mathbf{I}_k \mathbf{S}\|_{\mathcal{S}_p}^p = \|\mathbf{P}\mathbf{A}\mathbf{Q}\|_{\mathcal{S}_p}^p,$$

and similarly,

$$\left\|\left(\mathbf{I}-\mathbf{I}_k\right)\hat{\mathbf{A}}\left(\mathbf{I}-\mathbf{I}_k\right)\right\|_{\mathcal{S}_p}^p=\left\|\mathbf{R}^\top\left(\mathbf{I}-\mathbf{I}_k\right)\hat{\mathbf{A}}\left(\mathbf{I}-\mathbf{I}_k\right)\mathbf{S}\right\|_{\mathcal{S}_p}^p=\left\|\left(\mathbf{I}-\mathbf{P}\right)\mathbf{A}\left(\mathbf{I}-\mathbf{Q}\right)\right\|_{\mathcal{S}_p}^p.$$

Plugging these two bounds back into Equation (7.3), along with Equation (7.2), we can conclude,

$$\|\mathbf{A}\|_{\mathcal{S}_p}^p\geqslant\|\mathbf{PAQ}\|_{\mathcal{S}_p}^p+\left\|\left(\mathbf{I}-\mathbf{P}\right)\mathbf{A}\left(\mathbf{I}-\mathbf{Q}\right)\right\|_{\mathcal{S}_p}^p.$$

□

Algorithm 152 (Block Krylov Iteration, [MM15])**.**

Input: *An $n\times d$ matrix $\mathbf{A}$, target rank k, iteration count q and a block size parameter s such that $k\leqslant s\leqslant d$.*

1. *Let $\mathbf{U}$ be a $n\times s$ matrix such that each entry is drawn i.i.d. from $\mathcal{N}(0,1)$. Let $\mathbb{K}=\left[\mathbf{A}^\top\mathbf{U};(\mathbf{A}^\top\mathbf{A})\mathbf{A}^\top\mathbf{U};(\mathbf{A}^\top\mathbf{A})^2\mathbf{A}^\top\mathbf{U};\ldots;(\mathbf{A}^\top\mathbf{A})^q\mathbf{A}^\top\mathbf{U}\right]$ be the $d\times s(q+1)$ Krylov matrix obtained by concatenating the matrices $\mathbf{A}^\top\mathbf{U},\ldots,\left(\mathbf{A}^\top\mathbf{A}\right)^q\mathbf{A}^\top\mathbf{U}$.*
2. *Compute an orthonomal basis $\mathbf{Q}$ for the column span of $\mathbb{K}$. Let $\mathbf{M}=\mathbf{Q}^\top\mathbf{A}^\top\mathbf{AQ}$.*
3. *Compute the top k left singular vectors of $\mathbf{M}$, and denote them by $\mathbf{Y}_k$.*

Output: $\mathbf{Z}=\mathbf{QY}_k$

Note, despite establishing Lemma 7.4.3, it is not immediately apparent how to lower bound $\|\mathbf{AZZ}^\top\|_{\mathcal{S}_p}^p$, where $\mathbf{Z}$ is a candidate solution. Next, we show how to translate a guarantee on the Euclidean norm of $\mathbf{A}$ times a column of $\mathbf{Z}$ to a lower bound on $\|\mathbf{AZZ}^\top\|_{\mathcal{S}_p}^p$.

Lemma 7.4.4 (Per-Vector Guarantees to Schatten Norms)**.** *Let $\mathbf{A}$ be an $n\times d$ matrix with singular values denoted by $\{\sigma_i(\mathbf{A})\}_{i\in[d]}$. Let $\mathbf{Z}$ be a $d\times k$ matrix with orthonormal columns that is output by Algorithm 152, such that for all $i\in[k]$, with probability at least 99/100, $\|\mathbf{AZ}_{*,i}\|_2^2\geqslant\sigma_i^2(\mathbf{A})-\gamma_i\sigma_{k+1}^2(\mathbf{A})$, for some $\gamma\in(0,1)$. Then, for any $p\geqslant1$, we have*

$$\|\mathbf{AZZ}^\top\|_{\mathcal{S}_p}^p\geqslant\|\mathbf{A}_k\|_{\mathcal{S}_p}^p-\sum_{i\in[k]}\mathcal{O}(\gamma_i p)\,\sigma_{k+1}^2(\mathbf{A})\,\sigma_i^{p-2}(\mathbf{A}).$$

Proof. First, we observe that it suffices to show that $\sigma_i(\mathbf{AZ})^2\geqslant\|\mathbf{A}z_i\|_2^2$, where z_i is shorthand for $\mathbf{Z}_{*,i}$, the i-th columm of $\mathbf{Z}$. Assuming this inequality holds, we can complete the proof as

follows: we know that for all $i \in [k]$,

$$\begin{aligned} \sigma_i^2(\mathbf{AZ}) \geqslant \|\mathbf{A}z_i\|_2^2 &\geqslant \sigma_i^2(\mathbf{A}) - \gamma\sigma_{k+1}^2(\mathbf{A}) \\ &= \sigma_i^2(\mathbf{A})\left(1 - \gamma\frac{\sigma_{k+1}^2(\mathbf{A})}{\sigma_i^2(\mathbf{A})}\right) \end{aligned} \tag{7.4}$$

Then, taking $p/2$-th powers in (7.4),

$$\begin{aligned} \sigma_i^p(\mathbf{AZ}) &\geqslant \sigma_i^p(\mathbf{A})\left(1 - \gamma\frac{\sigma_{k+1}^2(\mathbf{A})}{\sigma_i^2(\mathbf{A})}\right)^{p/2} \\ &\geqslant \sigma_i^p(\mathbf{A})\left(1 - \mathcal{O}\left(\frac{\gamma p\sigma_{k+1}^2(\mathbf{A})}{\sigma_i^2(\mathbf{A})}\right)\right) \\ &= \sigma_i^p(\mathbf{A}) - \mathcal{O}(\gamma p)\,\sigma_{k+1}^2(\mathbf{A})\,\sigma_i^{p-2}(\mathbf{A}) \end{aligned} \tag{7.5}$$

where the second inequality follows from the generalized Bernoulli inequality (see Fact 7.3.6). Summing over all $i \in [k]$, we can conclude

$$\|\mathbf{AZ}\|_{\mathcal{S}_p}^p \geqslant \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - \sum_{i\in[k]} \mathcal{O}(\gamma p)\,\sigma_{k+1}^2(\mathbf{A})\,\sigma_i^{p-2}(\mathbf{A}).$$

Therefore, it remains to show that $\sigma_i(\mathbf{AZ})^2 \geqslant \|\mathbf{A}z_i\|_2^2$. First, we recall that Algorithm 152 outputs $\{z_i\}_{i\in[k]}$ such that $z_i = \mathbf{Q}\tilde{z}_i$, where $\mathbf{Q}$ is an orthonormal basis for the Krylov space $\mathbb{K}$ (an $d \times s(q+1)$ matrix) and $\tilde{z}_i$ is the i-th singular vector of $\mathbf{Q}^\top\mathbf{A}^\top\mathbf{AQ}$. Note that the $\tilde{z}_i$'s are $s(q+1)$-dimensional vectors. Let $\mathbf{W\Omega W}^\top$ be the SVD of $\mathbf{Q}^\top\mathbf{A}^\top\mathbf{AQ}$. Then, $\mathbf{QW\Omega W}^\top\mathbf{Q}^\top$ is the SVD of $\mathbf{QQ}^\top\mathbf{A}^\top\mathbf{AQQ}^\top$. To see this, let the i-th column of $\mathbf{QW}$ be denoted by $\mathbf{QW}_{*,i}$. Then,

$$\langle \mathbf{QW}_{*,i}, \mathbf{QW}_{*,i}\rangle = \mathbf{W}_{*,i}^\top\mathbf{Q}^\top\mathbf{QW}_{*,i} = 1$$

and similarly for any $j \neq i$,

$$\langle \mathbf{QW}_{*,i}, \mathbf{QW}_{*,j}\rangle = \mathbf{W}_{*,i}^\top\mathbf{Q}^\top\mathbf{QW}_{*,j} = 0$$

where we use that $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$ and the columns of $\mathbf{W}$ are orthonormal, which holds by definition. Therefore, $z_i = \mathbf{Q}\tilde{z}_i$ is the i-th singular vector of $\mathbf{QQ}^\top\mathbf{A}^\top\mathbf{AQQ}^\top$. Let $\tilde{\mathbf{Z}}$ be the matrix obtained

by stacking the vectors $\tilde{z}_i$ together. Then, we have

$$\begin{aligned}\sigma_i(\mathbf{AZ})^2 = \sigma_i^2(\mathbf{AQ}\tilde{\mathbf{Z}}) &= \sigma_i^2(\mathbf{AQ}) \\ &= \sigma_i^2(\mathbf{AQQ}^\top) \\ &= z_i^T\mathbf{QQ}^\top\mathbf{A}^\top\mathbf{AQQ}^\top z_i \\ &= z_i^\top\mathbf{A}^\top\mathbf{A}z_i\end{aligned} \tag{7.6}$$

where the first equality follows from the definition of $\tilde{\mathbf{Z}}$, the second follows from observing that $\tilde{\mathbf{Z}}$ are the singular vectors of $\mathbf{AQ}$ as shown above, the third follows from $\mathbf{Q}^\top$ having orthonormal rows, the fourth from z_i being the i-th singular vector of $\mathbf{AQQ}^\top$ and the last from observing that z_i is in the column span of $\mathbf{Q}$ and thus $\mathbf{QQ}^\top z_i = z_i$. This concludes the proof. □

Next, we show a lemma relating a high-accuracy per vector guarantee to cost on the residual subspace.

Lemma 7.4.5 (High-Accuracy Per-Vector Guarantee to Residual Cost)**.** *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, integer $p \geqslant 1$, $k \in [d]$, $\ell \in [k]$ and orthonormal vectors $\{w_i\}_{i\in[\ell]}$ such that $\|\mathbf{A}^\top w_i\|_2^2 \geqslant \sigma_i^2 - poly\,(\epsilon/d)\,\sigma_{k+1}^2$ and $(\sigma_\ell - \sigma_{\ell+1})/\sigma_\ell \geqslant \epsilon/d$. Let $\mathbf{W}$ be the matrix formed by stacking together the w_i's as columns. Then,*

$$\|\mathbf{A}_\ell^\top\left(\mathbf{I} - \mathbf{WW}^\top\right)\|_F^2 \leqslant poly\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2,$$

where $\mathbf{A}_\ell$ is the matrix obtained by truncating all but the top ℓ singular values of $\mathbf{A}$.

Proof. By Pythagorean Theorem,

$$\begin{aligned}\|\mathbf{A}^\top\mathbf{WW}^\top\|_F^2 &= \|\mathbf{A}_\ell^\top\mathbf{WW}^\top\|_F^2 + \|(\mathbf{A}-\mathbf{A}_\ell)^\top\mathbf{WW}^\top\|_F^2 \\ &\leqslant \|\mathbf{A}_\ell^\top\mathbf{WW}^\top\|_F^2 + \sigma_{\ell+1}^2\left(\|\mathbf{W}\|_F^2 - \|\mathbf{U}_\ell^\top\mathbf{W}\|_F^2\right),\end{aligned} \tag{7.7}$$

where $\mathbf{A} = \mathbf{U\Sigma V}^\top$. Further,

$$\|\mathbf{A}_\ell^\top\mathbf{WW}^\top\|_F^2 = \|\mathbf{\Sigma}_\ell\mathbf{U}_\ell^\top\mathbf{W}\|_F^2 \leqslant \sum_{i\in[\ell]}\sigma_i^2 - \sigma_\ell^2\left(\ell - \|\mathbf{U}_\ell^\top\mathbf{W}\|_F^2\right), \tag{7.8}$$

where the last inequality is obtained by making the Euclidean norm of all of the $\mathbf{U}_\ell\mathbf{W}$'s in $[\ell-1]$to be 1, and the ℓ-th row to be $\|(\mathbf{U}_\ell^\top\mathbf{W})_{\ell,*}\|_2^2 = \|\mathbf{U}_\ell^\top\mathbf{W}\|_F^2 - (\ell-1)$. Rearranging

Equation (7.8), we have

$$\ell - \|\mathbf{U}_\ell^\top \mathbf{W}\|_F^2 \leqslant \frac{\|\mathbf{A}_\ell^\top\|_F^2 - \|\mathbf{A}_\ell^\top \mathbf{W}\|_F^2}{\sigma_\ell^2}. \tag{7.9}$$

Now, observe $\|\mathbf{W}\|_F^2 = \ell$, and substituting (7.9) back into (7.8),

$$\|\mathbf{A}^\top \mathbf{W}\|_F^2 \leqslant \|\mathbf{A}_\ell^\top \mathbf{W}\|_F^2 + \frac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\left(\|\mathbf{A}_\ell^\top\|_F^2 - \|\mathbf{A}_\ell^\top \mathbf{W}\|_F^2\right). \tag{7.10}$$

Next, we can use the guarantee's on the w_i to obtain a lower bound on $\|\mathbf{A}^\top \mathbf{W}\|_F^2$ as follows:

$$\|\mathbf{A}^\top \mathbf{W}\|_F^2 = \sum_{i\in[\ell]} \|\mathbf{A}^\top w_i\|_2^2 \geqslant \|\mathbf{A}_\ell^\top\|_F^2 - \ell \cdot \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2, \tag{7.11}$$

Combining equations (7.10) and (7.11), we have

$$\begin{aligned} \|\mathbf{A}_\ell^\top - \mathbf{A}_\ell^\top \mathbf{W}\mathbf{W}^\top\|_F^2 &= \left(\|\mathbf{A}_\ell^\top\|_F^2 - \|\mathbf{A}_\ell^\top \mathbf{W}\|_F^2\right) \\ &\leqslant \frac{\ell \text{poly}(\epsilon/d)\sigma_{k+1}^2}{1-(\sigma_{\ell+1}/\sigma_\ell)^2} \\ &\leqslant \text{poly}(\epsilon/d)\sigma_{k+1}^2, \end{aligned} \tag{7.12}$$

which concludes the proof. □

Next, we need a lemma relating the Schatten-p norm of $\mathbf{AZ}$ to that of $\mathbf{W}^\top\mathbf{A}$, where $\mathbf{Z}$ is an arbitrary orthonormal basis and $\mathbf{W}$ is an orthonormal basis for $\mathbf{AZ}$.

Lemma 7.4.6. *Given a full-rank $n \times d$ matrix $\mathbf{A}$, let $\mathbf{W}$ be a $n \times k$ matrix with orthonormal columns. Further, let $\mathbf{Z}$ be an $d \times k$ matrix with orthonormal columns such that $\mathbf{Z}$ is a basis for $\mathbf{A}^\top\mathbf{W}$. Then, for all $i \in [k]$,*

$$\sigma_i\left(\mathbf{AZ}\right)^p \geqslant \sigma_i\left(\mathbf{A}^\top\mathbf{W}\right)^p$$

Proof. We use the following fact that for two matrices $\mathbf{A}$ and $\mathbf{B}$, we have that for all i, $\sigma_i(\mathbf{A}\cdot\mathbf{B}) \leqslant \sigma_i(\mathbf{A}) \cdot \sigma_1(\mathbf{B})$; see, e.g., (2) in [LC15] and references [33-36] therein.

Using this fact, we have

$$\begin{aligned}\sigma_i(\mathbf{A}^\top\mathbf{W}) = \sigma_i(\mathbf{A}^\top\mathbf{W}\mathbf{W}^T) &= \sigma_i(\mathbf{Z}\mathbf{Z}^T\mathbf{A}^\top\mathbf{W}\mathbf{W}^T)\\ &\leqslant \sigma_i(\mathbf{Z}\mathbf{Z}^T\mathbf{A}^\top)\cdot\sigma_1(\mathbf{W}\mathbf{W}^T)\\ &= \sigma_i(\mathbf{Z}\mathbf{Z}^T\mathbf{A}^\top)\\ &= \sigma_i(\mathbf{Z}^T\mathbf{A}),\end{aligned}$$

where we have used that $\sigma_1(\mathbf{W}\mathbf{W}^T) = 1$ since $\mathbf{W}\mathbf{W}^T$ is a projection matrix, and the fact that $\mathbf{Z}\mathbf{Z}^T$ is a basis for the column span of $\mathbf{A}^\top\mathbf{W}$. Raising both sides to the p-th power establishes the lemma.

□

Finally, we can combine the two aforementioned lemmas to obtain the following corollary:

Corollary 7.4.7 (Changing Basis for high-accuracy vectors)**.** *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, integer $p \geqslant 1$, $k \in [d]$, $\ell \in [k]$ and orthonormal vectors $\{w_i\}_{i\in[\ell]}$ such that $\|\mathbf{A}^\top w_i\|_2^2 \geqslant \sigma_i^2 - \mathit{poly}\,(\epsilon/d)\,\sigma_{k+1}^2$ and $(\sigma_\ell - \sigma_{\ell+1})/\sigma_\ell \geqslant \epsilon/d$. Let $\mathbf{W}$ be the matrix formed by stacking together the w_i's as columns and let $\mathbf{Z}$ be an orthonormal basis for $\mathbf{A}^\top\mathbf{W}\mathbf{W}^\top$. Then,*

$$\|\mathbf{A}_\ell\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_F^2 \leqslant \mathit{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2.$$

Proof. We closely follow the proof in Lemma 7.4.5. By Pythagorean Theorem,

$$\begin{aligned}\|\mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_F^2 &= \|\mathbf{A}_\ell\mathbf{Z}\mathbf{Z}^\top\|_F^2 + \|(\mathbf{A} - \mathbf{A}_\ell)\,\mathbf{Z}\mathbf{Z}^\top\|_F^2\\ &\leqslant \|\mathbf{A}_\ell\mathbf{Z}\mathbf{Z}^\top\|_F^2 + \sigma_{\ell+1}^2\left(\|\mathbf{Z}\|_F^2 - \|\mathbf{V}_\ell\mathbf{Z}\|_F^2\right),\end{aligned} \tag{7.13}$$

where $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Further,

$$\|\mathbf{A}_\ell\mathbf{Z}\mathbf{Z}^\top\|_F^2 = \|\mathbf{\Sigma}_\ell\mathbf{V}_\ell^\top\mathbf{Z}\|_F^2 \leqslant \sum_{i\in[\ell]}\sigma_i^2 - \sigma_\ell^2\left(\ell - \|\mathbf{V}_\ell^\top\mathbf{Z}\|_F^2\right), \tag{7.14}$$

where the last inequality is obtained by making the Euclidean norm of all of the $\mathbf{V}_\ell^\top\mathbf{Z}$'s in $[\ell-1]$to be 1, and the ℓ-th row to be $\|(\mathbf{V}_\ell^\top\mathbf{Z})_{\ell,*}\|_2^2 = \|\mathbf{V}_\ell^\top\mathbf{Z}\|_F^2 - (\ell-1)$. Rearranging Equation (7.14), we have

$$\ell - \|\mathbf{V}_\ell^\top\mathbf{Z}\|_F^2 \leqslant \frac{\|\mathbf{A}_\ell\|_F^2 - \|\mathbf{A}_\ell\mathbf{Z}\|_F^2}{\sigma_\ell^2}. \tag{7.15}$$

Now, observe $\|\mathbf{Z}\|_F^2 = \ell$, and substituting (7.15) back into (7.14),

$$\|\mathbf{AZ}\|_F^2 \leqslant \|\mathbf{A}_\ell \mathbf{Z}\|_F^2 + \frac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\left(\|\mathbf{A}_\ell\|_F^2 - \|\mathbf{A}_\ell \mathbf{Z}\|_F^2\right). \tag{7.16}$$

Next, we can use the guarantee's on the w_i to obtain a lower bound on $\|\mathbf{A}^\top \mathbf{W}\|_F^2$ as follows:

$$\|\mathbf{A}^\top \mathbf{W}\|_F^2 = \sum_{i\in[\ell]} \|\mathbf{A}^\top w_i\|_2^2 \geqslant \|\mathbf{A}_\ell^\top\|_F^2 - \ell \cdot \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2, \tag{7.17}$$

Observe $\|\mathbf{A}^\top \mathbf{W}\|_F^2 = \sum_{i\in[\ell]} \sigma_i^2(\mathbf{A}^\top \mathbf{W})$. By Lemma 7.4.6, we know that for all i, $\sigma_i^2(\mathbf{A}^\top \mathbf{W}) \leqslant \sigma_i^2(\mathbf{AZ})$. Therefore, we can restate the above equation as follows:

$$\|\mathbf{AZ}\|_F^2 \geqslant \|\mathbf{A}^\top \mathbf{W}\|_F^2 \geqslant \|\mathbf{A}_\ell\|_F^2 - \ell \cdot \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2, \tag{7.18}$$

Combining equations (7.16) and (7.18), we have

$$\begin{aligned} \|\mathbf{A}_\ell - \mathbf{A}_\ell \mathbf{Z}\mathbf{Z}^\top\|_F^2 &= \left(\|\mathbf{A}_\ell\|_F^2 - \|\mathbf{A}_\ell \mathbf{Z}\|_F^2\right) \\ &\leqslant \frac{\ell \text{poly}(\epsilon/d)\sigma_{k+1}^2}{1-(\sigma_{\ell+1}/\sigma_\ell)^2} \\ &\leqslant \text{poly}(\epsilon/d)\sigma_{k+1}^2, \end{aligned} \tag{7.19}$$

which concludes the proof. □

We now have all the ingredients we need to complete the proof of Theorem 150.

Proof of Theorem 150. Observe, using Lemma 7.4.1 with probability at least $97/100$, Step 1 of Algorithm 151 outputs $\hat{\sigma}_i$'s such that for all $i \in [k+1]$, $\hat{\sigma}_i^2 = (1 \pm 0.1/p)\,\sigma_i^2$ and $\hat{\sigma}_s^2 = (1 \pm 0.1/p)\,\sigma_s^2$, for $s = \mathcal{O}\left(kp^{-1/3}/\epsilon^{1/3}\right)$. Condition on this event.

At a high level, we proceed via a case analysis: either the Schatten-p norm of the tail is large compared to the $(k+1)$-st singular value, and we don't require a highly accurate solution, or the Schatten-p norm of the tail is small, and increasing the block size induces a gap. We formalize this intuition into a proof.

No large singular values. Let us first consider the case where $\sigma_1 < (1+1/p)\,\sigma_{k+1}$. We yet again split into cases, and consider the case where the Schatten-p norm of the tail is small, i.e.

$\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p \leqslant \frac{k}{p^{1/3}\epsilon^{1/3}} \cdot \sigma_{k+1}^p$. Observe, for any $t \in [1, d-k-1]$,

$$\frac{k}{p^{1/3}\epsilon^{1/3}} \cdot \sigma_{k+1}^p \geqslant \|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p \geqslant \sum_{i=k+1}^{k+1+t} \sigma_i^p \geqslant t\sigma_{k+1+t}^p. \tag{7.20}$$

Then, setting $t = \frac{(1+1/p)^p k}{\epsilon^{1/3}p^{1/3}} = \Theta\left(\frac{k}{\epsilon^{1/3}p^{1/3}}\right)$, we have $\sigma_{k+1+t} \leqslant \sigma_{k+1}/\left(1+1/p\right)$. It suffices to show that we can detect this gap for some $s \geqslant k+1+t$. Recall, we know that $\hat{\sigma}_{k+1} = (1 \pm 0.1/p)\sigma_{k+1}$ and $\hat{\sigma}_s = (1 \pm 0.1/p)\sigma_s$. Then, we have

$$\hat{\sigma}_s \leqslant \left(1 + \frac{0.1}{p}\right)\sigma_s \leqslant \left(1 + \frac{0.1}{p}\right)\sigma_{k+1+t} \leqslant \left(1 + \frac{0.1}{p}\right) \cdot \left(\frac{1}{1+1/p}\right)\sigma_{k+1} \leqslant \frac{1}{\left(1 + \frac{0.5}{p}\right)}\hat{\sigma}_{k+1}. \tag{7.21}$$

Therefore, Algorithm 151 outputs $\mathbf{Z}$, an orthonormal basis for $\mathbf{A}^\top\mathbf{W}_2$, where $\mathbf{W}_2$ is obtained by running Algorithm 152 on $\mathbf{A}^\top$, initialized with a block size of $\Theta\left(\frac{k}{\epsilon^{1/3}p^{1/3}}\right)$ and run for $\mathcal{O}\Big(\log(d/\epsilon)\sqrt{p}\Big)$ iterations. Observe, since $\sigma_{k+1+t} \leqslant \sigma_{k+1}/\left(1+1/p\right)$, this suffices to demonstrate a gap that depends on p as follows: $\frac{\sigma_k}{\sigma_k - \sigma_{k+t+1}} \leqslant p$. Recall, we account for this gap by running $\mathcal{O}\Big(\log(d)\sqrt{p}\Big)$ iterations. Using the gap dependent analysis (Lemma 7.4.2), we can conclude that with probability at least $99/100$, for all $i \in [k]$,

$$\|\mathbf{A}^\top(\mathbf{W}_2)_{*,i}\|_2^2 \geqslant \sigma_i^2 - \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2. \tag{7.22}$$

Then, applying Lemma 7.4.4 with $\mathbf{W}_2\mathbf{W}_2^\top$ satisfying the guarantee in (7.22), we have

$$\begin{aligned}\|\mathbf{A}^\top\mathbf{W}_2\mathbf{W}_2^\top\|_{\mathcal{S}_p}^p &\geqslant \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - \text{poly}\left(\frac{\epsilon}{d}\right)\sum_{i\in[k]}\sigma_{k+1}^2\sigma_i^{p-2} \\ &\geqslant \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^p.\end{aligned} \tag{7.23}$$

where the last inequality uses that $\sigma_1 < (1+1/p)\sigma_{k+1}$ and $(1+1/p)^{p-2} = \mathcal{O}(1)$. Next, we use Lemma 7.4.3 to relate $\|\mathbf{A}^\top\mathbf{W}_2\mathbf{W}_2^\top\|_{\mathcal{S}_p}^p$ to $\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p$, where $\mathbf{Z}$ is an orthonormal basis for $\mathbf{A}^\top\mathbf{W}_2\mathbf{W}_2^\top$ as output by the algorithm. Setting $\mathbf{Q} = \mathbf{Z}\mathbf{Z}^\top$ and $\mathbf{P} = \mathbf{W}_2\mathbf{W}_2^\top$, we observe that $\|\mathbf{P}\mathbf{A}\mathbf{Q}\|_{\mathcal{S}_p}^p = \|\mathbf{A}^\top\mathbf{W}_2\mathbf{W}_2^\top\|_{\mathcal{S}_p}^p = \|\mathbf{W}_2\mathbf{W}_2^\top\mathbf{A}\|_{\mathcal{S}_p}^p$ and $\|\left(\mathbf{I}-\mathbf{P}\right)\mathbf{A}\left(\mathbf{I}-\mathbf{Q}\right)\|_{\mathcal{S}_p}^p =$

$\|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p$. Then, invoking Lemma 7.4.3 and plugging in Equation (7.23), we have

$$\begin{aligned}\|(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q})\|_{\mathcal{S}_p}^p = \|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p &\leqslant \|\mathbf{A}\|_{\mathcal{S}_p}^p - \|\mathbf{A}^\top\mathbf{W}_2\mathbf{W}_2^\top\|_{\mathcal{S}_p}^p \\ &\leqslant \|\mathbf{A}\|_{\mathcal{S}_p}^p - \|\mathbf{A}_k\|_{\mathcal{S}_p}^p + \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^p \\ &\leqslant \left(1+\text{poly}\left(\frac{\epsilon}{d}\right)\right)\|\mathbf{A}-\mathbf{A}_k\|_{\mathcal{S}_p}^p,\end{aligned} \tag{7.24}$$

which concludes the analysis in this case.

As shown in Equation 7.21, we can detect a gap between σ_{k+1+t} and σ_{k+1} by comparing $\hat{\sigma}_s$ and $\hat{\sigma}_{k+1}$. When 7.21 does not hold, we know that $\hat{\sigma}_s \geqslant (1+0.5/p)\,\hat{\sigma}_{k+1}$ and Algorithm 151 outputs $\mathbf{Z}$, an orthonormal basis for $\mathbf{A}^\top\mathbf{W}_1\mathbf{W}_1^\top$. Since we have $(1\pm 0.1/p)$-approximate estimates to these quantities, we can conclude that $\sigma_s \geqslant (1+0.1/p)\,\sigma_{k+1}$. Then, we have

$$\|\mathbf{A}-\mathbf{A}_k\|_{\mathcal{S}_p}^p \geqslant s\cdot\sigma_s^p = \Omega\left(\frac{k}{\epsilon^{1/3}p^{1/3}}\right)\sigma_{k+1}^p. \tag{7.25}$$

It therefore remains to consider the case where $\|\mathbf{A}-\mathbf{A}_k\|_{\mathcal{S}_p}^p > \frac{ck}{p^{1/3}\epsilon^{1/3}}\cdot\sigma_{k+1}^p$, for a fixed universal constant c. Here, we note that the tail is large enough that an additive error of $\mathcal{O}\left(\epsilon^{2/3}p^{1/3}\right)\sigma_{k+1}^2$ on each of the top-k singular values suffices. Formally, it follows from Lemma 7.4.1 (setting $\gamma = \epsilon^{2/3}p^{-1/3}$, and invoking it for $\mathbf{A}^\top$) that initializing Algorithm 152 with block size k and running for $\mathcal{O}\left(\log(d/\epsilon)p^{1/6}/\epsilon^{1/3}\right)$ iterations suffices to output a $n\times k$ matrix $\mathbf{W}_1$ such that with probability at least $99/100$, for all $i\in[k]$,

$$\|\mathbf{A}^\top(\mathbf{W}_1)_{*,i}\|_2^2 \geqslant \sigma_i^2 - \epsilon^{2/3}p^{-1/3}\sigma_{k+1}^2. \tag{7.26}$$

Then, invoking Lemma 7.4.4 with $\mathbf{A}^\top$ and $\mathbf{W}_1$ as defined above, we have

$$\begin{aligned}\|\mathbf{A}^\top\mathbf{W}_1\mathbf{W}_1^\top\|_{\mathcal{S}_p}^p &= \|\mathbf{W}_1\mathbf{W}_1^\top\mathbf{A}\|_{\mathcal{S}_p}^p \\ &\geqslant \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - \sum_{i\in[k]}\mathcal{O}\left(\epsilon^{2/3}p^{-1/3}p\right)\sigma_{k+1}^2\sigma_i^{p-2} \\ &\geqslant \|\mathbf{A}_k\|_{\mathcal{S}_p}^p - \mathcal{O}\left(k\epsilon^{2/3}p^{2/3}\right)\sigma_{k+1}^p\end{aligned} \tag{7.27}$$

where the last inequality uses that $\sigma_1 < (1+1/p)\sigma_{k+1}$ and $(1+1/p)^p = \mathcal{O}(1)$. Recall, in this case, Algorithm 151 outputs $\mathbf{Z}\mathbf{Z}^\top$ where $\mathbf{Z}$ is an orthonormal basis for $\mathbf{A}^\top\mathbf{W}_1\mathbf{W}_1^\top$. Next, we invoke Lemma 7.4.3 to relate $\|\mathbf{A}^\top\mathbf{W}_1\mathbf{W}_1^\top\|_{\mathcal{S}_p}^p$ to $\|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p$. Setting $\mathbf{Q} = \mathbf{Z}\mathbf{Z}^\top$ and $\mathbf{P} = \mathbf{W}_1\mathbf{W}_1^\top$, we observe that $\|\mathbf{P}\mathbf{A}\mathbf{Q}\|_{\mathcal{S}_p}^p = \|\mathbf{W}_1\mathbf{W}_1^\top\mathbf{A}\|_{\mathcal{S}_p}^p$ and $\|(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q})\|_{\mathcal{S}_p}^p =$

$\|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p$. Then, invoking Lemma 7.4.3 and plugging in Equation (7.27), we have

$$\begin{aligned}\|(\mathbf{I}-\mathbf{P})\,\mathbf{A}\,(\mathbf{I}-\mathbf{Q})\|_{\mathcal{S}_p}^p = \|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p &\leqslant \|\mathbf{A}\|_{\mathcal{S}_p}^p - \|\mathbf{W}_1\mathbf{W}_1^\top\mathbf{A}\|_{\mathcal{S}_p}^p \\ &\leqslant \|\mathbf{A}\|_{\mathcal{S}_p}^p - \|\mathbf{A}_k\|_{\mathcal{S}_p}^p + \mathcal{O}\Big(k\epsilon^{2/3}p^{2/3}\Big)\,\sigma_{k+1}^p \\ &\leqslant (1+\mathcal{O}(p\epsilon))\,\|\mathbf{A}-\mathbf{A}_k\|_{\mathcal{S}_p}^p,\end{aligned} \tag{7.28}$$

where the last inequality follows from our assumption on the Schatten-p norm of the tail, given the case we are in. Taking the $(1/p)$-th root, and recalling that $\epsilon < 1/2$, we obtain

$$\|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} \leqslant (1+\mathcal{O}(\epsilon))\,\|\mathbf{A}-\mathbf{A}_k\|_p, \tag{7.29}$$

which concludes the case where $\ell = 0$.

Large Singular Values. Next, we consider the case where $\sigma_1 > (1+1/p)\,\sigma_{k+1}$. Then, let $\ell \in [k]$ be the largest integer such that $\sigma_\ell > (1+0.5/p)\,\sigma_{k+1}$ and $\sigma_{\ell+1} < (1-\epsilon/d)\sigma_\ell$. Observe, such an ℓ is guaranteed to exist.

We then note that in all settings Algorithm 151 runs $\Omega(\log(d/\epsilon)\sqrt{p}$ iterations on $\mathbf{A}^\top$ and since exists a gap of size p between σ_ℓ and σ_{k+1} it follows from Lemma 7.4.2 that running Block Krylov Iteration that Algorithm 151 always outputs an orthonormal matrix $\mathbf{W}$ s.t. for all $i \in [\ell]$,

$$\|\mathbf{A}^\top\mathbf{W}_{*,i}\|^2 \geqslant \sigma_i^2 - \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2. \tag{7.30}$$

Further, for all $i \in [\ell+1, k]$, we have

$$\|\mathbf{A}^\top\mathbf{W}_{*,i}\|^2 \geqslant \sigma_i^2 - \gamma_i\sigma_{k+1}^2, \tag{7.31}$$

where γ_i is determined by whether $\mathbf{W} = \mathbf{W}_1$ or $\mathbf{W} = \mathbf{W}_2$, as we discuss later.

We note that we cannot simply take $p/2$-th powers here (for large p) as this would introduce cross terms that scale proportional to $\sigma_i(\mathbf{A})$, which can be significantly larger than $\sigma_{k+1}(\mathbf{A})$. Instead, we require a finer analysis by splitting $\mathbf{A}$ into a head and tail term. Further, we let $\mathbf{Z}$ be an orthonormal basis for $\mathbf{A}^\top\mathbf{W}\mathbf{W}^\top$.

We are now ready to bound $\|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}$. By the triangle inequality,

$$\|\mathbf{A}\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} \leqslant \|\mathbf{A}_\ell\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} + \|(\mathbf{A}-\mathbf{A}_\ell)\left(\mathbf{I}-\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} \tag{7.32}$$

By Corollary 7.4.7, we know that $\|\mathbf{A}_\ell \left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_F^2 \leqslant \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}^2$, and since all Schatten norms are within a $\sqrt{d}$ factor of the Frobenius norm, we have

$$\|\mathbf{A}_\ell \left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} \leqslant \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}.$$

Substituting this back into Equation (7.32), we have

$$\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} \leqslant \mathcal{O}\left(\frac{\epsilon}{d}\right)\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p} + \underbrace{\|(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}}_{7.33.1}. \tag{7.33}$$

It remains to bound term 7.33.1 above. By triangle inequality, we have

$$\begin{aligned}&\|(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p \\ &\leqslant \left(\|(\mathbf{I} - \mathbf{W}\mathbf{W}^\top)(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} + \|\mathbf{W}\mathbf{W}^\top(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}\right)^p\end{aligned} \tag{7.34}$$

We bound the two terms on the RHS independently. To upper bound $\|\mathbf{W}\mathbf{W}^\top(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p$ we use the relation between Frobenius and Schatten norms, and recall that by definition, $\mathbf{W}\mathbf{W}^\top\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right) = \mathbf{W}\mathbf{W}^\top\mathbf{A} - \mathbf{W}\mathbf{W}^\top\mathbf{A}\mathbf{Z}\mathbf{Z}^\top = 0$, and thus

$$\begin{aligned}\|\mathbf{W}\mathbf{W}^\top(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} &\leqslant \sqrt{k}\|\mathbf{W}\mathbf{W}^\top(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_F \\ &= \sqrt{k}\|\mathbf{W}\mathbf{W}^\top\mathbf{A}_\ell(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top)\|_F \\ &\leqslant \sqrt{k}\|\mathbf{A}_\ell(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top)\|_F \\ &\leqslant \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1},\end{aligned} \tag{7.35}$$

where the last inequality follows from Corollary 7.4.7. Therefore, combining the above, we have

$$\|\mathbf{W}\mathbf{W}^\top(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} \leqslant \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1}, \tag{7.36}$$

It remains to upper bound $\|(\mathbf{I} - \mathbf{W}\mathbf{W}^\top)(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p$. Recall, $\mathbf{W}$ is the orthonormal basis output by Block Krylov run on $\mathbf{A}^\top$, and in Algorithm 151 $\mathbf{W}$ is either $\mathbf{W}_1$ or $\mathbf{W}_2$. Let

$\mathbf{Z}$ is a basis for $\mathbf{A}^\top \mathbf{W}\mathbf{W}^\top$. Then, applying Lemma 7.4.3 with $\mathbf{Q} = \mathbf{Z}\mathbf{Z}^\top$ and $\mathbf{P} = \mathbf{W}\mathbf{W}^\top$, we have

$$\begin{aligned}\|\left(\mathbf{I} - \mathbf{W}\mathbf{W}^\top\right)(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p &\leqslant \|(\mathbf{A} - \mathbf{A}_\ell)\|_{\mathcal{S}_p}^p - \|\mathbf{W}\mathbf{W}^\top\left(\mathbf{A} - \mathbf{A}_\ell\mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p \\ &= \sum_{j\in[\ell+1,d]} \sigma_j^p - \sum_{j\in[k]} \sigma_j^p\left(\mathbf{W}^\top(\mathbf{A} - \mathbf{A}_\ell)\mathbf{Z}\right)\end{aligned} \tag{7.37}$$

Next, we show that for all $j \in [k]$, $\sigma_j\left(\mathbf{W}^\top(\mathbf{A} - \mathbf{A}_\ell)\mathbf{Z}\right) \geqslant \sigma_{j+\ell}\left(\mathbf{W}^\top\mathbf{A}\right)$. Here, we invoke Fact 7.3.5 for $\mathbf{X} = \mathbf{W}^\top(\mathbf{A} - \mathbf{A}_\ell)\mathbf{Z}$ and $\mathbf{Y} = \mathbf{W}^\top\mathbf{A}_\ell\mathbf{Z}$, with $i = j$ and $j = \ell$. Note, the precondition on the indices i, j in Fact 7.3.5 is satisfied since $\mathbf{X}, \mathbf{Y}$ are $n \times k$ matrices, and $j \in [k]$ and $\ell < k$. Then, we have

$$\begin{aligned}\sigma_{j+\ell}\left(\mathbf{W}^\top\mathbf{A}\mathbf{Z}\right) &= \sigma_{j+\ell}\left(\mathbf{W}^\top(\mathbf{A} - \mathbf{A}_\ell\mathbf{Z}) + \mathbf{W}^\top\mathbf{A}_\ell\mathbf{Z}\right) \\ &\leqslant \sigma_j\left(\mathbf{W}^\top(\mathbf{A} - \mathbf{A}_\ell)\mathbf{Z}\right) + \sigma_{\ell+1}\left(\mathbf{W}^\top\mathbf{A}_\ell\mathbf{Z}\right),\end{aligned}$$

but $\mathbf{W}^\top\mathbf{A}_\ell\mathbf{Z}$ is a rank $\leqslant \ell$ matrix, and thus $\sigma_{\ell+1}\left(\mathbf{W}^\top\mathbf{A}_\ell\mathbf{Z}\right) = 0$. Therefore, we can conclude,

$$\|\left(\mathbf{I} - \mathbf{W}\mathbf{W}^\top\right)(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p \leqslant \sum_{j\in[\ell+1,d]} \sigma_j^p - \sum_{j\in[\ell+1,k+\ell]} \sigma_j^p\left(\mathbf{W}^\top\mathbf{A}\mathbf{Z}\right). \tag{7.38}$$

However, now we observe that $\mathbf{W}\mathbf{W}^\top\mathbf{A}\mathbf{Z}\mathbf{Z}^\top = \mathbf{W}\mathbf{W}^\top\mathbf{A}$, and thus $\sigma_j\left(\mathbf{W}^\top\mathbf{A}\mathbf{Z}\right) = \sigma_j(\mathbf{W}\mathbf{A})$. Recall, for all $j \in [k]$, it follows from Equation (7.5) in the proof of Lemma 7.4.4 that $\sigma_j^p(\mathbf{W}^\top\mathbf{A}) = \sigma_j^p(\mathbf{A}^\top\mathbf{W}) \geqslant \sigma_j^p(\mathbf{A}) - \mathcal{O}(\gamma_j p)\,\sigma_{k+1}^2\sigma_j^{p-2}$. Further, by definition, for $j \in [\ell+1, k]$, $\sigma_j \leqslant (1 + 1/p)\,\sigma_{k+1}$ and thus, for all $j \in [\ell+1, k]$,

$$\begin{aligned}\sigma_j^p(\mathbf{W}^\top\mathbf{A}) &\geqslant \sigma_j^p - \mathcal{O}\left(\gamma_j p\,(1 + 1/p)^{p-2}\right)\sigma_{k+1}^p \\ &\geqslant \sigma_j^p - \mathcal{O}(\gamma_j p)\,\sigma_{k+1}^p.\end{aligned} \tag{7.39}$$

Recall, we can correctly determine whether $\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p \leqslant \frac{k}{p^{1/3}\epsilon^{1/3}}\sigma_{k+1}^p$ or not, up to error in estimating the singular values as shown earlier, in the analysis for the case where there are no large singular values. In particular, let us first consider the case where this is true. Repeating the

argument in equations (7.20), (7.21), we can conclude that Algorithm 151 outputs $\mathbf{W} = \mathbf{W}_2$ and thus for all $j \in [\ell+1, k]$, $\gamma_j = \text{poly}(\epsilon/d)$. Therefore, substituting this back into Equation (7.38), we have

$$\begin{aligned}\left\|\left(\mathbf{I} - \mathbf{W}\mathbf{W}^\top\right)(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p &\leqslant \sum_{j\in[\ell+1,d]} \sigma_j^p - \left(\sum_{j\in[\ell+1,k]} \sigma_j^p - \mathcal{O}(\gamma_j p)\,\sigma_{k+1}^p\right) \\ &\leqslant \|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p + \sum_{j\in[\ell+1,k]} \mathcal{O}(\gamma_j p)\,\sigma_{k+1}^p \\ &\leqslant \left(1 + \text{poly}\left(\frac{\epsilon}{d}\right)\right)\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p,\end{aligned} \tag{7.40}$$

concluding the analysis in this case.

Next, consider the case where $\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p > \frac{k}{p^{1/3}\epsilon^{1/3}}\sigma_{k+1}^p$. Then, repeating the analysis in Equation (7.25), we know that Algorithm 151 outputs $\mathbf{W} = \mathbf{W}_2$, and thus for all $j \in [\ell+1, k]$, $\gamma_j = \epsilon^{2/3}p^{-1/3}$ as shown in Equation (7.26). Again, substituting this back into Equation (7.38), we have

$$\begin{aligned}\left\|\left(\mathbf{I} - \mathbf{W}\mathbf{W}^\top\right)(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\right\|_{\mathcal{S}_p}^p &\leqslant \|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p + \sum_{j\in[\ell+1,k]} \mathcal{O}(\gamma_j p)\,\sigma_{k+1}^p \\ &\leqslant \|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p + \mathcal{O}\left(kp \cdot \epsilon^{2/3}p^{-1/3}\right)\sigma_{k+1}^p \\ &\leqslant (1 + \mathcal{O}(\epsilon p))\,\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p}^p,\end{aligned} \tag{7.41}$$

where the last inequality follows from our assumption. Taking the $(1/p)$-th root and substituting equations (7.36) and (7.41) back into (7.34), we can conclude

$$\|(\mathbf{A} - \mathbf{A}_\ell)\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p \leqslant (1 + \mathcal{O}(\epsilon))\,\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p} + \text{poly}\left(\frac{\epsilon}{d}\right)\sigma_{k+1} \leqslant (1 + \mathcal{O}(\epsilon))\,\|\mathbf{A} - \mathbf{A}_k\|_{\mathcal{S}_p},$$

which when substituted into Equation (7.32) concludes the analysis.

Next, we analyze the running time and matrix-vector products. Running Algorithm 152 with block size k for $q = \mathcal{O}\left(\log(d)p^{1/6}/\epsilon^{1/3}\right)$ iterations requires $\mathcal{O}\left(\frac{\text{nnz}(\mathbf{A})kp^{1/6}\log(d)}{\epsilon^{1/3}}\right)$ time and $\mathcal{O}\left(\frac{kp^{1/6}\log(d)}{\epsilon^{1/3}}\right)$ matrix-vector products. Similarly, running with block size $\mathcal{O}\left(k/\left(\epsilon p\right)^{1/3}\right)$ for $q = \mathcal{O}\left(\log(d/\epsilon)\sqrt{p}\right)$ iterations requires $\mathcal{O}\left(\frac{\text{nnz}(\mathbf{A})kp^{1/6}\log(d/\epsilon)}{\epsilon^{1/3}}\right)$ time and $\mathcal{O}\left(\frac{kp^{1/6}\log(d)}{\epsilon^{1/3}}\right)$ matrix-vector products. Finally, we observe that to obtain a $(1+1/p)$-approximation to σ_1 and σ_{k+1}, we need $\mathcal{O}\left(\log(d)\sqrt{p}\right)$ iterations with blocksize $k+1$ and this requires $\mathcal{O}\left(\log(d)\sqrt{p}k\right)$ matrix-vector products. Note, our setting of the exponent of p and ϵ was chosen to balance the two cases, and

this concludes the proof.

□

7.5 Query Lower Bounds

Next, we show that the ϵ-dependence obtained by our algorithms for Schatten-p low-rank approximation is optimal in the restricted computation model of matrix-vector products. The matrix-vector product model is defined as follows: given a matrix $\mathbf{A}$, our algorithm is allowed to make adaptive matrix-vector queries to $\mathbf{A}$, where one matrix-vector query is of the form $\mathbf{A}v$, for any $v \in \mathbb{R}^d$. Our lower bounds are information-theoretic and rely on the hardness of estimating the smallest eigenvalue of a Wishart ensemble, as established in recent work of Braverman, Hazan, Simchowitz and Woodworth [BHSW20].

We split the lower bounds into the case of $p \in [1, 2]$ and $p > 2$. For $p \in [1, 2]$, we have a simple argument based on the Araki-Lieb-Thirring inequality (Fact 7.3.10), whereas for $p > 2$, our lower bounds require an involved argument using a norm compression inequality for partitioned operators (Fact 7.3.14).

7.5.1 Lower Bounds for $p \in [1, 2]$

The main lower bound we prove in this sub-section is as follows:

Theorem 153 (Query Lower Bound for $p \in [1, 2]$)**.** *Given $\varepsilon > 0$, and $p \in [1, 2]$, there exists a distribution $\mathcal{D}$ over $n \times n$ matrices such that for $\mathbf{A} \sim \mathcal{D}$, any randomized algorithm that with probability at least $9/10$ outputs a rank-1 matrix $\mathbf{B}$ such that $\|\mathbf{A} - \mathbf{B}\|_{\mathcal{S}_p}^p \leqslant (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p$ must make $\Omega(1/\varepsilon^{1/3})$ matrix-vector queries to $\mathbf{A}$.*

We require the following theorem on the hardness of computing the minimum eigenvalue of a Wishart Matrix, introduced recently by Braverman, Hazan, Simchowitz and Woodworth [BHSW20]:

Theorem 154 (Computing Min Eigenvalue of Wishart, Theorem 3.1 [BHSW20])**.** *Given $\epsilon \in (0, 1)$, there exists a function $\mathbf{d} : (0, 1) \to \mathbb{N}$ such that for all $d \geqslant \mathbf{d}(\epsilon)$, the following holds. Let $\mathbf{W} \sim$ Wishart(d) be a Wishart matrix and $\{\lambda_i\}_{i\in[d]}$ be the eigenvalues of $\mathbf{W}$, in descending order. Then, there exists a universal constant c^* such that:*

1. *Let ζ_1 be the event that $\lambda_d(\mathbf{W}) \leqslant c_1/d^2$, ζ_2 be the event that $\lambda_{d-1}(\mathbf{W}) - \lambda_d(\mathbf{W}) \geqslant c_2/d^2$ and ζ_3 be the event that $\|\mathbf{W}\|_{op} \leqslant 5$, where c_1 and c_2 are constants that depend only on ϵ. Then, $\mathbf{Pr}_{\mathbf{W}}[\zeta_1 \cap \zeta_2 \cap \zeta_3] \geqslant 1 - \frac{c^*\sqrt{\epsilon}}{2}$.*

2. *Any randomized algorithm that makes at most $(1-\epsilon)d$ adaptive matrix-vector queries and outputs an estimate $\hat{\lambda}_d$ must satisfy*

$$\mathbf{Pr}_{\mathbf{W}}\left[|\hat{\lambda}_d - \lambda_d| \geqslant \frac{1}{4d^2}\right] \geqslant c^*\sqrt{\epsilon}.$$

We also use the following lemma from [BHSW20] bounding the minimum eigenvalue of a Wishart ensemble:

Lemma 7.5.1 (Non-Asymptotic Spectra of Wishart Ensembles, Corollary 3.3 [BHSW20])**.** *Let $\mathbf{W} \sim \mathit{Wishart}(n)$ be such that $n = \Omega(1/\varepsilon^3)$. Then, there exists a universal constant $c_2 > 0$ such that*

$$\mathbf{Pr}\left[\lambda_n(\mathbf{W}) \geqslant \frac{1}{n^2}\right] \geqslant c_2, \quad \textit{and} \quad \mathbf{Pr}\left[\lambda_n(\mathbf{W}) < \frac{1}{2n^2}\right] \geqslant \frac{c_2}{2}.$$

We are now ready to prove Theorem 153. Our high level approach is to show that we can take any solution that is a $(1+\varepsilon)$-relative-error Schatten-p low-rank approximation to the hard instance $\mathbf{I} - \frac{1}{5}\mathbf{W}$, where $\mathbf{W}$ is a Wishart ensemble, and extract from it an accurate estimate of the minimum eigenvalue of $\mathbf{W}$, thus appealing to the hardness stated in (2) of Theorem 154 above.

Proof of Theorem 153. Let $n = \Theta\left(1/\epsilon^{1/3}\right)$ and let $\mathbf{A} = \mathbf{I} - \frac{1}{5}\mathbf{W}$ be an $n \times n$ instance where $\mathbf{W} \sim \mathsf{Wishart}(n)$. Let ζ_1 be the event that $\|\mathbf{W}\|_{\text{op}} \leqslant 5$. It follows from Fact 7.3.17 that ζ_1 holds with probability at least $99/100$, and we condition on this event. Let ζ_2 be the event that $\lambda_n(\mathbf{W}) \geqslant \frac{1}{n^2} = \frac{\varepsilon^{2/3}}{c^*}$ and ζ_3 be the event that $\lambda_n(\mathbf{W}) < \frac{1}{2n^2} = \frac{\varepsilon^{2/3}}{2c^*}$.

Then, conditioning on ζ_2, we have that

$$1 - \frac{1}{5}\lambda_n(\mathbf{W}) \leqslant 1 - \frac{\varepsilon^{2/3}}{5c^*}. \tag{7.42}$$

Similarly, conditioning on ζ_3, we have that

$$1 - \frac{1}{5}\lambda_n(\mathbf{W}) \geqslant 1 - \frac{\varepsilon^{2/3}}{10c^*}. \tag{7.43}$$

We observe that for $p \in [1,2]$, using Bernoulli's inequality (Fact 7.3.6) we have

$$\left(1 - \frac{1}{5}\lambda_n(\mathbf{W})\right)^p \geqslant 1 - \frac{p}{5}\lambda_n(\mathbf{W})$$

and since $(1-x)^p \leqslant (1-x)$ for any $x \in (0,1)$, we also have that,

$$\left(1 - \frac{1}{5}\lambda_n(\mathbf{W})\right)^p \leqslant 1 - \frac{1}{5}\lambda_n(\mathbf{W})$$

Therefore, we can conclude, $\left(1 - \frac{1}{5}\lambda_n(\mathbf{W})\right)^p = 1 - \Theta\left(\lambda_n(\mathbf{W})\right)$.

$$\|\mathbf{A}\|_{\mathcal{S}_p}^p = \sum_{i\in[n]} \lambda_i^p\left(\mathbf{I} - \frac{1}{5}\mathbf{W}\right) \leqslant \sum_{i\in[n]} \lambda_i\left(\mathbf{I} - \frac{1}{5}\mathbf{W}\right) \leqslant \mathcal{O}\Big(\frac{1}{\epsilon^{1/3}}\Big) \tag{7.44}$$

where the last inequality follows from the fact that $n = \sqrt{c^*}/\epsilon^{1/3}$. Let $\mathbf{A}_1$ denote the best rank-1 approximation to $\mathbf{A}$. Then, it follows from Equation (7.44) that

$$\epsilon\|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p \leqslant \epsilon\|\mathbf{A}\|_{\mathcal{S}_p}^p \leqslant \mathcal{O}\Big(\epsilon^{2/3}\Big) \tag{7.45}$$

Observe, any $(1+\epsilon)$-approximate relative-error Schatten-p low-rank approximation algorithm for $k=1$ outputs a matrix $vv^\top$ such that

$$\begin{aligned}\|\mathbf{A}\left(\mathbf{I} - vv^\top\right)\|_{\mathcal{S}_p}^p &\leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p \\ &\leqslant \|\mathbf{A}\|_{\mathcal{S}_p}^p - \|\mathbf{A}\|_{\text{op}}^p + \Theta(\epsilon^{2/3})\end{aligned} \tag{7.46}$$

By definition of the Schatten-p norm we have:

$$\begin{aligned}\|\mathbf{A}\left(\mathbf{I} - vv^\top\right)\|_{\mathcal{S}_p}^p &= \text{Tr}\left[\left(\left(\mathbf{I} - vv^\top\right)^2\mathbf{A}^2\left(\mathbf{I} - vv^\top\right)^2\right)^{p/2}\right] \\ &\geqslant \text{Tr}\left[\left(\mathbf{I} - vv^\top\right)^p\mathbf{A}^p\left(\mathbf{I} - vv^\top\right)^p\right] \\ &= \text{Tr}\left[\mathbf{A}^p - \mathbf{A}^p vv^\top\right] \\ &= \|\mathbf{A}\|_{\mathcal{S}_p}^p - \text{Tr}\left[\left(vv^\top\right)^{p/2}\left(\mathbf{A}^2\right)^{p/2}\left(vv^\top\right)^{p/2}\right] \\ &\geqslant \|\mathbf{A}\|_{\mathcal{S}_p}^p - \text{Tr}\left[\left(vv^\top\mathbf{A}^2vv^\top\right)^{p/2}\right] \\ &= \|\mathbf{A}\|_{\mathcal{S}_p}^p - \|\mathbf{A}vv^\top\|_{\mathcal{S}_p}^p \\ &= \|\mathbf{A}\|_{\mathcal{S}_p}^p - \|\mathbf{A}v\|_2^p\end{aligned} \tag{7.47}$$

where the first and last inequality follows from the reverse Araki-Lieb-Thirring inequality (Fact

7.3.10). Combining equations (7.46) and (7.47), we have that

$$\|\mathbf{A}\|_{\text{op}}^p \geqslant \|\mathbf{A}v\|_2^p \geqslant \|\mathbf{A}\|_{\text{op}}^p - \Theta(\epsilon^{2/3}) \tag{7.48}$$

Next, we observe that $\mathbf{A}v = (\mathbf{I} - 1/5\mathbf{W})\, v$ can be computed with one additional matrix-vector product and

$$\|\mathbf{A}\|_{\text{op}}^p = \left(1 - \frac{1}{5}\lambda_n(\mathbf{W})\right)^p = 1 - \frac{p}{5}\lambda_n(\mathbf{W}) + \mathcal{O}\Big(\lambda_n^2(\mathbf{W})\Big) \tag{7.49}$$

Consider the estimator $\hat{\lambda}(\mathbf{W}) = \frac{5}{p}\left(1 - \|\left(\mathbf{I} - \frac{1}{5}\mathbf{W}\right)v\|_2^p\right)$. Combining equations (7.48) and (7.49), we can conclude

$$\hat{\lambda}(\mathbf{W}) = \lambda_{\min}(\mathbf{W}) \pm \Theta(\epsilon^{2/3}).$$

obtaining an additive error estimate to the minimum eigenvalue of $\mathbf{W}$ by computing an additional matrix-vector product. It follows that we satisfy conditions (1) and (2) in Theorem 154 and thus any algorithm for computing a rank-1 approximation to the matrix $\mathbf{A} = \mathbf{I} - \frac{1}{5}\mathbf{W}$ in Schatten p norm must make at least $\frac{1}{\epsilon^{1/3}}$ queries to the aforementioned matrix, completing the proof. The claim follows from Theorem 154. □

7.5.2 Lower Bound for $p > 2$

We now consider the case when $p > 2$. We note that the previous approach no longer works since we cannot lower bound the cost of $\|\,(\mathbf{I} - \mathbf{W}/5)\left(\mathbf{I} - vv^\top\right)\|_{\mathcal{S}_p}$, as the Araki-Lieb-Thirring inequality reverses (see application in Equation 7.47). Therefore, we require a new approach, and appeal to a special case of Conjecture 7.3.15 that is known to be true, i.e. the Aligned Norm Compression inequality (see Fact 7.3.14). The main theorem we prove in this sub-section is as follows:

Theorem 155 (Query Lower Bound for $p > 2$)**.** *Given $\varepsilon > 0$, and $p \geqslant 2$ such that $p = \mathcal{O}(1)$, there exists a distribution $\mathcal{D}$ over $n \times n$ matrices such that for $\mathbf{A} \sim \mathcal{D}$, any randomized algorithm that with probability at least $99/100$ outputs a unit vector u such that $\|\mathbf{A} - \mathbf{A}uu^\top\|_{\mathcal{S}_p}^p \leqslant (1 + \varepsilon)\|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p$ must make $\Omega\left(1/\varepsilon^{1/3}\right)$ matrix-vector queries to $\mathbf{A}$.*

We first introduce a sequence of key lemmas required for our proof.

Corollary 7.5.2 (Special Case of Lemma 7.4.1)**.** *Given $\gamma \in [0, 1]$, a vector $v \in \mathbb{R}^d$ and an*

$n \times d$ matrix $\mathbf{A}$, let $t = \log(n/\gamma)/(c\sqrt{\gamma})$, for a fixed universal constant c. Then, there exists an algorithm that computes t matrix-vector products with $\mathbf{A}$ and outputs a unit vector u such that with probability at least $99/100$,

$$\|\mathbf{A}\|_{op}^2 - \|\mathbf{A}u\|_2^2 \leqslant O\left(\gamma\sigma_2^2\right).$$

where σ_2 is the second largest singular value of $\mathbf{A}$.

Next, we prove a key lemma relating the norm of a matrix to norms of orthogonal projections applied to the matrix. We note that this lemma is straight forward and holds for arbitrary vectors unit u, v if Conjecture 7.3.15 holds. However, we show that we can transform our matrix to have structure such that we can apply Fact 7.3.14 instead.

Lemma 7.5.3 (Orthogonal Projectors to Block Matrices)**.** *Given an $n \times d$ matrix $\mathbf{A}$, $p > 2$ and unit vectors $u \in \mathbb{R}^d, v \in \mathbb{R}^n$, such that $\left(\mathbf{I} - vv^\top\right)\mathbf{A}uu^\top = 0$. Then, we have*

$$\|\mathbf{A}\|_{\mathcal{S}_p} \leqslant \|\begin{pmatrix} \|vv^\top\mathbf{A}uu^\top\|_{\mathcal{S}_p} & \|vv^\top\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p} \\ 0 & \|\left(\mathbf{I} - vv^\top\right)\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p} \end{pmatrix}\|_{\mathcal{S}_p}.$$

Proof. Let $\mathbf{I} - vv^\top = \mathbf{Y}\mathbf{Y}^\top$, where $\mathbf{Y}$ has $n-1$ orthonormal columns. Further, since v and $\mathbf{Y}$ span disjoint subspaces, and the union of their span is $\mathbb{R}^n$, the matrix $(v \mid \mathbf{Y})$, obtained by concatenating their columns is unitary. Then, let $\mathbf{R} = (v \mid \mathbf{Y})^\top$ and observe, $\mathbf{R}$ has orthonormal rows and columns (since $\mathbf{R}$ is unitary). Next, let $\mathbf{I} - uu^\top = \mathbf{Z}\mathbf{Z}^\top$, where $\mathbf{Z}$ is $d \times (d-1)$ and has orthonormal columns. Let $\mathbf{S} = (u \mid \mathbf{Z})^\top$, and observe $\mathbf{S}$ has orthonormal rows and columns.

Let $\hat{\mathbf{A}} = \mathbf{R}\mathbf{A}\mathbf{S}^\top$, which admits the following block-matrix form:

$$\hat{\mathbf{A}} = \begin{pmatrix} v^\top \\ \mathbf{Y}^\top \end{pmatrix} \cdot \mathbf{A} \cdot (u \mid \mathbf{Z}) = \begin{pmatrix} v^\top \\ \mathbf{Y}^\top \end{pmatrix} (\mathbf{A}u \mid \mathbf{A}\mathbf{Z}) = \begin{pmatrix} v^\top\mathbf{A}u & v^\top\mathbf{A}\mathbf{Z} \\ \mathbf{Y}^\top\mathbf{A}u & \mathbf{Y}^\top\mathbf{A}\mathbf{Z} \end{pmatrix}$$

Since $\mathbf{R}$ and $\mathbf{S}$ are unitary, it follows from unitary invariance of the Schatten-p norm that

$$\|\mathbf{A}\|_{\mathcal{S}_p} = \|\hat{\mathbf{A}}\|_{\mathcal{S}_p} = \|\begin{pmatrix} v^\top\mathbf{A}u & v^\top\mathbf{A}\mathbf{Z} \\ \mathbf{Y}^\top\mathbf{A}u & \mathbf{Y}^\top\mathbf{A}\mathbf{Z} \end{pmatrix}\|_{\mathcal{S}_p} = \|\begin{pmatrix} v^\top\mathbf{A}u & v^\top\mathbf{A}\mathbf{Z} \\ 0 & \mathbf{Y}^\top\mathbf{A}\mathbf{Z} \end{pmatrix}\|_{\mathcal{S}_p}, \tag{7.50}$$

where the last equality follows from observing that

$$\|\mathbf{Y}^\top \mathbf{A}u\|_F = \|\mathbf{Y}\mathbf{Y}^\top \mathbf{A}uu^\top\|_F = \|\left(\mathbf{I} - vv^\top\right)\mathbf{A}uu^\top\|_F = 0$$

and therefore $\mathbf{Y}^\top \mathbf{A}u$ is a matrix of all 0s. Next, we append a set of $d-2$ columns of 0's to make the top left and top right block the same size. Since this does not change the singular values, we have

$$\|\mathbf{A}\|_{\mathcal{S}_p} = \|\begin{pmatrix} v^\top \mathbf{A}u & 0 & v^\top \mathbf{A}\mathbf{Z} \\ 0 & 0 & \mathbf{Y}^\top \mathbf{A}\mathbf{Z} \end{pmatrix}\|_{\mathcal{S}_p} \tag{7.51}$$

Next, we construct a rotation matrix $\mathbf{R}$ such that on right multiplying a row vector by $\mathbf{R}$, the first $d-1$ coordinates remain the same and on the remaining coordinates, the vector $v^\top \mathbf{A}\mathbf{Z}$ gets mapped to $ce_1^\top$ for some scalar c. Let $\mathbf{S}$ be the $d-1 \times d-1$ rotation matrix such that $v^\top \mathbf{A}\mathbf{Z}\mathbf{S} = ce_1^\top$. Then, $\mathbf{R} = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{S} \end{pmatrix}$ and it is easy to verify that $\mathbf{R}$ is unitary. Therefore,

$$\begin{pmatrix} v^\top \mathbf{A}u & 0 & v^\top \mathbf{A}\mathbf{Z} \\ 0 & 0 & \mathbf{Y}^\top \mathbf{A}\mathbf{Z} \end{pmatrix} \cdot \mathbf{R} = \begin{pmatrix} v^\top \mathbf{A}u & 0 & ce_1^\top \\ 0 & 0 & \mathbf{Y}^\top \mathbf{A}\mathbf{Z}\mathbf{S} \end{pmatrix}$$

Now, we observe the final matrix above has a block matrix form we can apply the Aligned Norm Compression inequality from Fact 7.3.14, with $\alpha_1 = v^\top \mathbf{A}u$, $\alpha_2 = c$, $\beta_1 = 0$ and $\beta_2 = 0$, and therefore

$$\begin{aligned} \|\mathbf{A}\|_{\mathcal{S}_p} = \|\begin{pmatrix} v^\top \mathbf{A}u & 0 & ce_1^\top \\ 0 & 0 & \mathbf{Y}^\top \mathbf{A}\mathbf{Z}\mathbf{S} \end{pmatrix}\|_{\mathcal{S}_p} &\leqslant \|\begin{pmatrix} \|v^\top \mathbf{A}u\|_{\mathcal{S}_p} & 0 & \|ce_1^\top\|_{\mathcal{S}_p} \\ 0 & 0 & \|\mathbf{Y}^\top \mathbf{A}\mathbf{Z}\mathbf{S}\|_{\mathcal{S}_p} \end{pmatrix}\|_{\mathcal{S}_p} \\ &= \|\begin{pmatrix} \|vv^\top \mathbf{A}uu^\top\|_{\mathcal{S}_p} & \|vv^\top \mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_{\mathcal{S}_p} \\ 0 & \|\mathbf{Y}\mathbf{Y}^\top \mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_{\mathcal{S}_p} \end{pmatrix}\|_{\mathcal{S}_p} \end{aligned} \tag{7.52}$$

where the last equality follows from unitary invariance and substituting the definition of $\mathbf{Y}\mathbf{Y}^\top$ and $\mathbf{Z}\mathbf{Z}^\top$ completes the proof.

□

Fact 7.5.4 (SVD of a 2×2 Matrix)**.** *Given a* 2×2 *matrix* $\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *let* $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ *be the SVD*

of M. *Then,*

$$\mathbf{\Sigma}_{1,1} = \sqrt{\frac{a^2+b^2+c^2+d^2+\sqrt{\left(a^2+b^2-c^2-d^2\right)^2+4\left(ac+bd\right)^2}}{2}},$$

and

$$\mathbf{\Sigma}_{2,2} = \sqrt{\frac{a^2+b^2+c^2+d^2-\sqrt{\left(a^2+b^2-c^2-d^2\right)^2+4\left(ac+bd\right)^2}}{2}}.$$

Now, we are ready to prove Theorem 155.

Proof of Theorem 155. Let $\mathbf{A} = \mathbf{I} - \frac{1}{5}\mathbf{W}$ where $\mathbf{W}$ is an $n \times n$ Wishart matrix as in the proof of Theorem 153 and we have by hypothesis that there is an algorithm that with probability at least $99/100$, outputs a unit vector u such that $\|\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p}^p \leqslant (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_1\|_{\mathcal{S}_p}^p$. Let $v = \mathbf{A}u/\|\mathbf{A}u\|_2$ and observe, $\left(\mathbf{I} - vv^\top\right)\mathbf{A}uu^\top = 0$. Further, by the unitary invariance of the Schatten-p norm,

$$\|vv^\top \mathbf{A}uu^\top\|_{\mathcal{S}_p} = |v^\top \mathbf{A}u| = \frac{|u^\top \mathbf{A}^\top \mathbf{A}u|}{\|\mathbf{A}u\|_2} = \|\mathbf{A}u\|_2. \tag{7.53}$$

Similarly,

$$\begin{aligned}
\|vv^\top \mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p} &= \sqrt{\|v^\top \mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_2^2} = \sqrt{\|v^\top \mathbf{A}\|_2^2 - \|v^\top \mathbf{A}uu^\top\|_2^2} \\
&= \sqrt{\frac{\|u^\top \mathbf{A}^\top \mathbf{A}\|_2^2}{\|\mathbf{A}u\|_2^2} - \|\mathbf{A}u\|_2^2} \\
&\leqslant \sqrt{\frac{\|u^\top \mathbf{A}^\top\|_2^2 \cdot \|\mathbf{A}\|_{\mathrm{op}}^2}{\|\mathbf{A}u\|_2^2} - \|\mathbf{A}u\|_2^2} \\
&\leqslant \epsilon^{1/3}\sigma_2,
\end{aligned} \tag{7.54}$$

where we use sub-multiplicativity of the ℓ_2 norm and Corollary 7.5.2 with $\gamma = \epsilon^{2/3}$. Note that we can assume w.l.o.g. that Corollary 7.5.2 holds since we can just iterate Block Krylov $q = (1/c\epsilon^{1/3})$ times, for a sufficiently large constant c, starting the iterations with the vector u output by the algorithm hypothesized for the theorem, and pay only $(1/c\epsilon^{1/3})$ extra matrix-vector

products. Since $vv^\top \mathbf{A} + \mathbf{A}uu^\top - vv^\top \mathbf{A}uu^\top$ has rank at most 3,

$$\begin{aligned} \|\left(\mathbf{I} - vv^\top\right)\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p}^p &= \|\mathbf{A} - vv^\top\mathbf{A} - \mathbf{A}uu^\top + vv^\top\mathbf{A}uu^\top\|_{\mathcal{S}_p}^p \\ &\geqslant \|\mathbf{A} - \mathbf{A}_3\|_{\mathcal{S}_p}^p \\ &= \Omega\left(\frac{1}{\epsilon^{1/3}}\right), \end{aligned} \tag{7.55}$$

where the last inequality follows from Fact 7.3.17.

Let $\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \|vv^\top\mathbf{A}uu^\top\|_{\mathcal{S}_p} & \|vv^\top\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p} \\ \|\left(\mathbf{I} - vv^\top\right)\mathbf{A}uu^\top\|_{\mathcal{S}_p} & \|\left(\mathbf{I} - vv^\top\right)\mathbf{A}\left(\mathbf{I} - uu^\top\right)\|_{\mathcal{S}_p} \end{pmatrix}^\top$. Then, it follows from Fact 7.5.4 that

$$\begin{aligned} \mathbf{\Sigma}_{1,1}(\mathbf{M}) &= \frac{1}{\sqrt{2}} \cdot \sqrt{a^2 + c^2 + d^2 + \sqrt{\left(a^2 - c^2 - d^2\right)^2 + 4\left(ac\right)^2}} \\ &= \frac{1}{\sqrt{2}} \cdot \sqrt{a^2 + c^2 + d^2 + \left(c^2 + d^2 - a^2\right) + \Theta\left(\frac{4a^2c^2}{c^2 + d^2 - a^2}\right)} \\ &= \sqrt{c^2 + d^2 + \Theta\left(\frac{a^2c^2}{c^2 + d^2 - a^2}\right)}, \end{aligned} \tag{7.56}$$

where we use that $b = 0$, $c, a \leqslant 1$ and $1 \ll d$ and the Taylor expansion of $\sqrt{x + y}$ for $x, y \geqslant 0$. Similarly,

$$\mathbf{\Sigma}_{2,2}(\mathbf{M}) = \sqrt{a^2 - \Theta\left(\frac{a^2c^2}{c^2 + d^2 - a^2}\right)}. \tag{7.57}$$

Then, using equations (7.56) and (7.57) we can bound the Schatten-p norm of $\mathbf{M}$ as follows:

$$\|\mathbf{M}\|_{\mathcal{S}_p}^p \leqslant \underbrace{\left(c^2 + d^2 + \Theta\left(\frac{a^2c^2}{c^2 + d^2 - a^2}\right)\right)^{p/2}}_{7.58.1} + \underbrace{\left(a^2 - \Theta\left(\frac{a^2c^2}{c^2 + d^2 - a^2}\right)\right)^{p/2}}_{7.58.2}. \tag{7.58}$$

We now bound each of the terms above. Consider the first term:

$$\begin{aligned}\left(c^2+d^2+\Theta\left(\frac{a^2c^2}{c^2+d^2-a^2}\right)\right)^{p/2} &= \Bigg(\|vv^\top\mathbf{A}\left(\mathbf{I}-uu^\top\right)\|_{\mathcal{S}_p}^2 \\ &\qquad + \|\left(\mathbf{I}-vv^\top\right)\mathbf{A}\left(\mathbf{I}-uu^\top\right)\|_{\mathcal{S}_p}^2 + \Theta\left(\varepsilon^{2/3}\|\mathbf{A}u\|_2^2\right)\Bigg)^{p/2} \\ &\leqslant \left(\Theta\left(\varepsilon^{2/3}\right)+\|\mathbf{A}\left(\mathbf{I}-uu^\top\right)\|_{\mathcal{S}_p}^2\right)^{p/2} \\ &\leqslant \left(1+\mathcal{O}\left(\epsilon^{2p/3}\right)\right)\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p,\end{aligned} \tag{7.59}$$

where we use equation (7.53), (7.54), and (7.55), and $\|\mathbf{A}\left(\mathbf{I}-uu^\top\right)\|_{\mathcal{S}_p}^2 \leqslant (1+\epsilon)^{2/p}\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^2$. The last inequality follows from observing that

$$\varepsilon^{2/3} \leqslant \mathcal{O}\left(\varepsilon^{4/3}\cdot\frac{1}{\epsilon^{2/3p}}\right) \leqslant \mathcal{O}\left(\varepsilon^{4/3}\cdot\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^2\right).$$

We can now bound the second term in Equation 7.58 as follows:

$$\left(a^2-\Theta\left(\frac{a^2c^2}{c^2+d^2-a^2}\right)\right)^{p/2} = \left(\|\mathbf{A}u\|_2^2-\Theta\left(\epsilon^{2/3}\|\mathbf{A}u\|_2^2\right)\right)^{p/2} \leqslant \|\mathbf{A}u\|_2^p. \tag{7.60}$$

Then, we have

$$\|\mathbf{M}\|_{\mathcal{S}_p}^p \leqslant \left(1+\mathcal{O}\left(\epsilon^{2p/3}\right)\right)\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p + \|\mathbf{A}u\|_2^p.$$

It follows from Lemma 7.5.3, that $\|\mathbf{M}\|_{\mathcal{S}_p}^p \geqslant \|\mathbf{A}\|_{\mathcal{S}_p}^p$ and thus

$$\begin{aligned}\|\mathbf{A}u\|_2^p &\geqslant \|\mathbf{A}\|_{\mathcal{S}_p}^p - \left(1+\mathcal{O}\left(\epsilon^{2p/3}\right)\right)\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p \\ &= \|\mathbf{A}\|_{\text{op}}^p - \mathcal{O}\left(\epsilon^{2p/3}\right)\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p \\ &\geqslant \|\mathbf{A}\|_{\text{op}}^p - \mathcal{O}\left(\epsilon\|\mathbf{A}-\mathbf{A}_1\|_{\mathcal{S}_p}^p\right) \\ &\geqslant \|\mathbf{A}\|_{\text{op}}^p - \mathcal{O}\left(\varepsilon^{2/3}\right)\end{aligned} \tag{7.61}$$

where the second to last inequality follows from recalling $p \geqslant 2$. The remainder of the proof is as in that following (7.48) in the proof of Theorem 153. □

7.6 Extending Prior Work on Lower Bounds

In this section, we briefly discuss prior work on estimating top singular/eigenvalues in the matrix-vector product model and why existing approaches do not immediately imply a lower bound for low-rank approximation, under any unitarily invariant norm, including Frobenius and spectral norm.

In a sequence of works, Braverman, Hazan, Simchowitz and Woodworth [BHSW20] and Simchowitz, Alaoui and Recht [SAR18] establish eigenvalue estimation lower bounds in the matrix-vector query model. We draw on their techniques and use the hard instance at the heart of their lower bound, but require additional techniques to obtain a lower bound for low-rank approximation.

The main theorem (Theorem 2.2 of [SAR18]), for k =1, states that any randomized algorithm which outputs a vector v such that with constant probability

$$v^\top |\mathbf{A}| v >= (1 - \mathcal{O}(\text{gap}))\|\mathbf{A}\|_{\text{op}},$$

requires $\Omega\left(1/\sqrt{\text{gap}}\right)$ matrix-vector products, where $|\mathbf{A}| = (\mathbf{A}^2)^{1/2}$ has the same singular values as $\mathbf{A}$ and gap $\in (0, 1)$. However, this guarantee is too weak to imply a lower bound for spectral low-rank approximation.

Indeed, for this theorem to be meaningful in our setting, we require setting gap $= \Theta(\epsilon)$. However, there exist input matrices $\mathbf{A}$, e.g., $\mathbf{A} = \text{diag}\,(1 + \epsilon, 1, \ldots, 1, 0)$, and vector $v = \Theta\left(\sqrt{\epsilon}\right) e_1 + ((1 - \Theta(\epsilon))\, e_n$ such that

$$\|\mathbf{A}(\mathbf{I} - vv^\top)\|_{\text{op}} \leqslant (1 + \epsilon)\, \sigma_2(\mathbf{A}),$$

i.e. v yields a valid low-rank approximation but $v^\top \mathbf{A} v$ is only $\Theta(\epsilon)$. Note, here the gap is $\Theta(1)$ instead of the required $1 - \epsilon$ and thus we obtain no lower bound for spectral low-rank approximation.

Moreover, it can be shown that when $\mathbf{A}$ is the hard instance considered in [SAR18], i.e. $\mathbf{A} = \mathbf{G} + \lambda uu^T$, where $\mathbf{G}$ is a Gaussian Orthogonal Ensemble (GOE) and u is a random unit vector on the sphere, there exists a vector v that does not satisfy the guarantee of Theorem 2.2, yet yields a spectral low-rank approximation. In particular, consider $v = \Theta(\sqrt{\epsilon})r_1 + (1 - \Theta(\epsilon))\, r_d$ where r_1 is the largest singular vector of $|\mathbf{A}|$ and r_d is the smallest singular vector. Since the smallest $O(1)$ singular values of a $d \times d$ GOE can be shown to be $O(1/d)$, and $\mathbf{A}$ is a rank-1

perturbation of a GOE, similar to the diagonal case above, we can show

$$\|\mathbf{A}\left(\mathbf{I} - vv^T\right)\|_{\text{op}} \leqslant (1+\epsilon)\,\sigma_2(\mathbf{A}),$$

yet $v^\top|\mathbf{A}|v$ is only $\Theta(\epsilon)$. Therefore, it is not possible to obtain a lower bound for low-rank approximation from Theorem 2.2 in a black-box manner.

7.7 Low Rank Approximation of Matrix Polynomials

We note that polynomials of matrices are implicitly defined, even in the RAM model, and computing them explicitly would be prohibitively expensive and may destroy any sparsity structure. The proof just follows from running our algorithm on $\mathbf{M} = \left(\mathbf{A}^\top\mathbf{A}\right)^\ell$. It is straightforward to simulate a matrix-vector product of the form $\mathbf{M}v$ using access to matrix-vector products for $\mathbf{A}$ and $\mathbf{A}^\top$ with an $\mathcal{O}(\ell)$ overhead.

Theorem 156 (Low Rank Approximation of Matrix Polynomials)**.** *Given an $n \times d$ matrix $\mathbf{A}$, $\ell \in \mathbb{N}$, target rank k and an accuracy parameter $\varepsilon > 0$, let $\mathbf{M} = \left(\mathbf{A}^\top\mathbf{A}\right)^\ell$ or $\mathbf{M} = \mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^\ell$. Then, for any $p \geqslant 1$, there exists an algorithm that uses at most $\mathcal{O}\Big(k\ell\log(nk)p^{1/6}/\varepsilon^{1/3}\Big)$ matrix-vector products and with probability at least $9/10$ outputs a matrix $\mathbf{Z} \in \mathbb{R}^{d\times k}$ with orthonormal columns such that,*

$$\|\mathbf{M}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p} \leqslant (1+\varepsilon)\min_{\mathbf{U}:\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k}\|\mathbf{M}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\|_{\mathcal{S}_p}.$$

The only prior work we are aware of is the algorithm of [MM15], which would achieve a worse $\mathcal{O}\Big(k\ell\log(nk)/\varepsilon^{1/2}\Big)$ number of matrix-vector products for the Frobenius norm and match our guarantee for the spectral norm.

7.8 Improved Streaming Bounds

In the streaming model, the input matrix is initialized to all zeros, and at each time step, the (i,j)-th entry is updated. The updates can be positive or negative, and the goal is to output a low-rank approximation, without storing the whole matrix. The number of passes required by our algorithm is proportional to the number of *adaptive* matrix-vector queries we require. As an immediate corollary of this observation, we obtain the following formal guarantee:

Corollary 7.8.1 (Schatten LRA in a Stream)**.** *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, a target rank $k \in [d]$, an accuracy parameter $\epsilon \in (0,1)$ and any $p \geqslant 1$, there exists a streaming algorithm that makes $\mathcal{O}\Big(\log(d/\epsilon)p^{1/6}/\epsilon^{1/3}\Big)$ passes over the input, requires $\mathcal{O}\Big(nk/\epsilon^{1/3}\Big)$ space, and outputs a $d \times k$ matrix $\mathbf{Z}$ with orthonormal columns such that with probability at least $9/10$,*

$$\|\mathbf{A}\left(\mathbf{I} - \mathbf{Z}\mathbf{Z}^\top\right)\|_{\mathcal{S}_p}^p \leqslant (1+\epsilon) \min_{\mathbf{U}:\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \|\mathbf{A}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\|_{\mathcal{S}_p}^p.$$

The only prior work on low-rank approximation in a stream is by Boutsidis, Woodruff and Zhong, who consider the special case of $p = 2$ [BWZ16]. They obtain a single pass algorithm that requires $\mathcal{O}(nk/\epsilon + \text{poly}(k/\epsilon))$ space and a two pass algorithm that requires $\mathcal{O}(nk + \text{poly}(k/\epsilon))$ space. For general p, we note that recent work by Li and Woodruff [LW20] can be used to derive a streaming algorithm that obtains a worse space dependence but only requires a single pass: for $1 \leqslant p < 2$, the space required is $\tilde{\mathcal{O}}\left(n\left(\frac{k+k^{2/p}}{\epsilon^2} + \frac{k^{2/p}}{\epsilon^{1+2/p}}\right)\right)$ and for $p > 2$, the space required is $\tilde{\mathcal{O}}\left(n\left(\frac{kn^{1-2/p}}{\epsilon^2} + \frac{k^{2/p}+n^{1-2/p}}{\epsilon^{2+2/p}}\right)\right)$.

We note that for $p < 2$, we obtain a polynomially better dependence on ϵ and for $p > 2$, the space complexity of our algorithm is linear in n, as compared to $n^{2-2/p}$ above. The optimal space complexity of Schatten-p low-rank approximation (for $p \neq 2$) in a single pass remains open.

Chapter 8

PSD Low-Rank Approximation

8.1 Introduction

Low-rank approximation is one of the most common dimensionality reduction techniques, whereby one replaces a large matrix $\mathbf{A}$ with a low-rank factorization $\mathbf{U} \cdot \mathbf{V} \approx \mathbf{A}$. Such a factorization provides a compact way of storing $\mathbf{A}$ and allows one to multiply $\mathbf{A}$ quickly by a vector. It is used as an algorithmic primitive in clustering [DFK+04, McS01], recommendation systems [DKR02], web search [AFKM01, Kle99], and learning mixtures of distributions [AM05, KSV05], and has numerous other applications.

A large body of recent work has looked at *relative-error* low-rank approximation, whereby given an $n \times n$ matrix $\mathbf{A}$, an accuracy parameter $\epsilon > 0$, and a rank parameter k, one seeks to output a rank-k matrix $\mathbf{B}$ for which

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2, \tag{8.1}$$

where for a matrix $\mathbf{C}$, $\|\mathbf{C}\|_F^2 = \sum_{i,j} \mathbf{C}_{i,j}^2$, and $\mathbf{A}_k$ denotes the best rank-k approximation to $\mathbf{A}$ in Frobenius norm. $\mathbf{A}_k$ can be computed exactly using the singular value decomposition, but takes time $O(n^\omega)$, where ω is the matrix multiplication constant. We refer the reader to the survey [Woo14a] and references therein.

For worst-case matrices, it is not hard to see that any algorithm achieving (8.1) must spend at least $\Omega(\text{nnz}(\mathbf{A}))$ time, where $\text{nnz}(\mathbf{A})$ denotes the number of non-zero entries (sparsity) of $\mathbf{A}$. Indeed, without reading most of the non-zero entries of $\mathbf{A}$, one could fail to read a single large entry, thus making one's output matrix $\mathbf{B}$ an arbitrarily bad approximation.

A flurry of recent work [KP16, MW17c, BW18, CLW18, Tan19, RSML18, GLT18, IVWW19, SW19, GSLW19] has looked at the possibility of achieving *sublinear* time algorithms (classical and quantum) for low-rank approximation. In particular, Musco and Woodruff [MW17c] consider the important case of positive-semidefinite (PSD) matrices. PSD matrices include as special cases covariance matrices, correlation matrices, graph Laplacians, kernel matrices and random dot product models. Further, the special case where the input itself is low-rank (PSD Matrix Completion) has applications in quantum state tomography [GLF+10]. Subsequently, Bakshi and Woodruff [BW18] considered low-rank approximation of the closely related family of Negative-type (Euclidean Squared) distance matrices. Negative-type metrics include as special cases ℓ_1 and ℓ_2 metrics, spherical metrics and hypermetrics, as well as effective resistances in graphs [DL09, TD87, CRR+96, CKM+11]. Negative-type metrics have found various applications in algorithm design and optimization [ALN08, SS11, KMP14, MST15].

Musco and Woodruff show that it is possible to output a low-rank matrix $\mathbf{B}$ in factored form achieving (8.1) in $\widetilde{O}(nk/\epsilon^{2.5} + nk^{\omega-1}/\epsilon^{2(\omega-1)})$ time, while reading only $\widetilde{O}(nk/\epsilon^{2.5})$ entries of $\mathbf{A}$. They also showed a lower bound that any algorithm achieving (8.1) must read $\Omega(nk/\epsilon)$ entries, and closing the gap between these bounds has remained an open question. Similarly, Bakshi and Woodruff exploit the structure of Negative-type metrics to reduce to the PSD case and obtain a bi-criteria algorithm that requires $\widetilde{O}(nk/\epsilon^{2.5})$ queries. The gap in the sample complexity and the requirement of a bi-criteria guarantee remained open.

Next we consider PSD matrices that have been corrupted by a small amount of noise. A drawback of algorithms achieving (8.1) is that they cannot tolerate any amount of unstructured noise. For instance, if one slightly corrupts a few off-diagonal entries, making the input matrix $\mathbf{A}$ no longer PSD, then it is impossible to detect such corruptions in sublinear time, making the relative-error guarantee (8.1) information-theoretically impossible. Motivated by this, we also introduce a new framework where an adversary corrupts the input by adding a noise matrix $\mathbf{N}$ to a psd matrix $\mathbf{A}$. We assume that the Frobenius norm of the corruption is bounded relative to the Frobenius norm of $\mathbf{A}$, i.e., $\|\mathbf{N}\|_F^2 \leqslant \eta\|\mathbf{A}\|_F^2$. We also assume the corruption is well-spread, i.e., each row of $\mathbf{N}$ has ℓ_2^2-norm at most a fixed constant factor larger than ℓ_2^2-norm of the corresponding row of $\mathbf{A}$.

This model captures small perturbations to PSD matrices that we may observe in real-world datasets, as a consequence of round-off or numerical errors in tasks such as computing Laplacian pseudoinverses, and systematic measurement errors when computing a covariance matrix. One important application captured by our model is low-rank approximation of corrupted *correlation matrices*. Finding a low-rank approximation of such matrices occurs when measured correlations

are asynchronous or incomplete, or when models are stress-tested by adjusting individual correlations. Low-rank approximation of correlation matrices also has many applications in finance [Hig02].

Given that it is information-theoretically impossible to obtain the relative-error guarantee (8.1) in the *robust model*, we relax our notion of approximation to the following well-studied additive-error guarantee:

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant \|\mathbf{A}-\mathbf{A}_k\|_F^2 + (\epsilon+\eta)\|\mathbf{A}\|_F^2. \tag{8.2}$$

This additive-error guarantee was introduced by the seminal work of Frieze et. al. [FKV04b], and triggered a long line of work on low-rank approximation from a computational perspective. Frieze et al. showed that it is possible to achieve (8.2) in $O(\text{nnz}(\mathbf{A}))$ time. Further, given access to an oracle for computing row norms of $\mathbf{A}$, 8.2 is achievable in sublinear time. More recently, the same notion of approximation was used to obtain sublinear sample complexity and running time algorithms for *distance matrices* [BW18, IVWW19], and a quantum algorithm for recommendation systems [KP16], which was subsequently dequantized [Tan19].

This raises the question of how robust are our sublinear low-rank approximation algorithms for structured matrices, if we relax to additive-error guarantees and allow for corruption. In particular, can we obtain additive-error low-rank approximation algorithms for PSD matrices that achieve sublinear time and sample complexity in the presence of noise? We characterize when such robust algorithms are achievable in sublinear time.

8.1.1 Our Results

We begin with stating our results for low-rank approximation for structured matrices. Our main result is an optimal algorithm for low-rank approximation of PSD matrices:

Theorem 166 *(Informal Sample-Optimal PSD LRA.) Given a PSD matrix $\mathbf{A}$, there exists an algorithm that queries $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ and outputs a rank k matrix $\mathbf{B}$ such that with probability $99/100$, $\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$, and the algorithm runs in time $\widetilde{O}(n \cdot (k/\epsilon)^{\omega-1})$.*

Remark 157. Our algorithm matches the sample complexity lower bound of Musco and Woodruff, up to logarithmic factors, which shows that any randomized algorithm that outputs a $(1+\epsilon)$-

Problem	Prior Work		Our Results		Query Lower Bound
	Query	Run Time	Query	Run Time	
PSD LRA	$O\left(\frac{nk}{\epsilon^{2.5}}\right)$	$O\left(\frac{nk^{\omega-1}}{\epsilon^{2\omega-2}}+\frac{nk}{\epsilon^{2.5}}\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	[MW17c]		Thm. 166		[MW17c]
PSD LRA PSD Output	$O\left(nk\left(\frac{k}{\epsilon^2}+\frac{1}{\epsilon^3}\right)\right)$	$O\left(nk^{\omega-1}\left(\frac{k}{\epsilon^\omega}+\frac{1}{\epsilon^{3\omega-3}}\right)\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	[MW17c]		Thm. 166		[MW17c]
Negative-Type LRA	$O\left(\frac{nk}{\epsilon^{2.5}}\right)$	$O\left(\frac{nk^{\omega-1}}{\epsilon^{2\omega-2}}+\frac{nk}{\epsilon^{2.5}}\right)$	$O^*\left(\frac{nk}{\epsilon}\right)$	$O^\dagger\left(\frac{nk^{\omega-1}}{\epsilon^{\omega-1}}\right)$	$\Omega\left(\frac{nk}{\epsilon}\right)$
	Bi-criteria, [BW18]		No Bi-criteria, Thm. 173		[BW18]
Coreset Ridge Regression	$O\left(\frac{ns_\lambda^2}{\epsilon^4}\right)$	$O\left(\frac{ns_\lambda^\omega}{\epsilon^\omega}\right)$	$O^*\left(\frac{ns_\lambda}{\epsilon^2}\right)$	$O^\dagger\left(\frac{ns_\lambda^{\omega-1}}{\epsilon^{2\omega-2}}\right)$	$\Omega\left(\frac{ns_\lambda}{\epsilon^2}\right)$
	[MW17c]		Thm. 174		Thm 176

Table 8.1: Comparison with prior work. The notation O^* and $O^\dagger$ represent existence of matching lower bounds for query complexity and running time (assuming the fast matrix multiplication exponent ω is 2) respectively. The notation s_λ is used to denote the statistical dimension of ridge regression. All bounds are stated ignoring polylogarithmic factors in n, k and ϵ.

relative-error low-rank approximation for a PSD matrix $\mathbf{A}$ must read $\Omega(nk/\epsilon)$ entries. Our running time also improves that of Musco and Woodruff and is optimal if the matrix multiplication exponent ω is 2.

Remark 158. We can extend our algorithm such that the low-rank matrix $\mathbf{B}$ we output is also PSD with the same query complexity and running time. In comparison, the algorithm of Musco and Woodruff accesses $\widetilde{O}(nk/\epsilon^3 + nk^2/\epsilon^2)$ entries in $\mathbf{A}$ and runs in time $\widetilde{O}(n(k/\epsilon)^\omega + nk^{\omega-1}/\epsilon^{3(\omega-1)})$.

At the core of our analysis is a sample optimal algorithm for Spectral Regression: $\min_{\mathbf{X}} \|\mathbf{DX}-\mathbf{E}\|_2^2$. We show that when $\mathbf{D}$ has orthonormal columns and $\mathbf{E}$ is arbitrary, we can sketch the problem by sampling rows proportional to the leverage scores of $\mathbf{D}$ and approximately preserve the minimum cost. This is particularly surprising since our sketch only computes sampling probabilities by reading entries in $\mathbf{D}$, while being completely agnostic to the entries in $\mathbf{E}$. Here, we also prove a spectral approximate matrix product guarantee for our one-sided leverage score sketch, which may be of independent interest. We note that such a guarantee for leverage score sampling does not appear in prior work, and we discuss the technical challenges we need to overcome in the subsequent section.

The techniques we develop for PSD low-rank approximation also extend to computing a low-rank approximation for distance matrices that arise from negative-type (Euclidean-squared) metrics. Here, our input is a pair-wise distance matrix $\mathbf{A}$ corresponding to a point set $\mathcal{P} =$

$\{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^d$ such that $\mathbf{A}_{i,j} = \|x_i - x_j\|_2^2$. We obtain an optimal algorithm for computing a low-rank approximation of such matrices:

Theorem 173 *(Informal Sample-Optimal LRA for Negative-Type Metrics.) Given a negative-type distance matrix $\mathbf{A}$, there exists an algorithm that queries $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ and outputs a rank k matrix $\mathbf{B}$ such that with probability $99/100$, $\|\mathbf{A} - \mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$, and the algorithm runs in time $\widetilde{O}(n \cdot (k/\epsilon)^{\omega-1})$.*

Remark 159. Prior work of Bakshi and Woodruff [BW18] obtains a $\widetilde{O}(nk/\epsilon^{2.5})$ query algorithm that outputs a rank-$(k+4)$ matrix $\mathbf{B}$ such that $\|\mathbf{A} - \mathbf{B}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$. We show that the bi-criteria guarantee is not necessary, thereby resolving an open question in their paper.

Structured Regression. The sample-optimal algorithm for PSD Low-Rank Approximation also leads to a faster algorithm for Ridge Regression, when the design matrix is PSD. Given a PSD matrix $\mathbf{A}$, a vector y and a regularization parameter λ, we consider the following optimization problem: $\min_{x \in \mathbb{R}^n} \|\mathbf{A}x - y\|_2^2 + \lambda\|x\|_2^2$. This problem is often referred to as Ridge Regression and has been the focus of numerous theoretical and practical works (see [Gru17] and references therein).

Theorem 174 *(Informal Ridge Regression.) Given a PSD matrix $\mathbf{A}$, a regularization parameter λ and statistical dimension $s_\lambda = Tr\left[(\mathbf{A}^2 + \lambda\mathbf{I})^{-1}\mathbf{A}^2\right]$, there exists an algorithm that queries $\widetilde{O}(ns_\lambda/\epsilon^2)$ entries of $\mathbf{A}$ and with probability $99/100$ outputs a $(1+\epsilon)$ approximate solution to the Ridge Regression objective and runs in $\widetilde{O}(n(s_\lambda/\epsilon^2)^{\omega-1})$ time.*

Remark 160. Our result improves on prior work by Musco and Woodruff [MW17c], who obtain an algorithm that queries $\widetilde{O}(ns_\lambda^2/\epsilon^4)$ entries in $\mathbf{A}$ and runs in $\widetilde{O}(n(s_\lambda/\epsilon^2)^\omega)$ time.

Remark 161. Since our algorithm works for all y simultaneously, we obtain a low-rank *coreset* of the design matrix (in factored form) that preserves the Ridge Regression cost up to a $(1+\epsilon)$ factor. Further, in Theorem 176, we prove a matching lower bound on the query complexity for any coreset construction.

Robust Low-Rank Approximation. Next, we consider a robust form of low-rank approximation problem, where the input is a PSD matrix corrupted by noise. In this setting, we have query access to the corrupted matrix $\mathbf{A} + \mathbf{N}$, where $\mathbf{A}$ is PSD and $\mathbf{N}$ is such that $\|\mathbf{N}\|_F^2 \leqslant \eta\|\mathbf{A}\|_F^2$. Further, for all $i \in [n]$ $\|\mathbf{N}_{i,*}\|_2^2 \leqslant c\|\mathbf{A}_{i,*}\|_2^2$, for a fixed constant c. The diagonal of a PSD matrix carries crucial information since the largest diagonal entry upper bounds all off-diagonal entries.

Therefore, a reasonable adversarial strategy is to corrupt the largest diagonal entries and make them close to the small diagonal entries, which enables the resulting matrix to have large off-diagonal entries that are hard to find. Capturing this intuition we parameterize our algorithms and lower bounds by the largest ratio between a diagonal entry of $\mathbf{A}$ and $\mathbf{A}+\mathbf{N}$, denoted by $\phi_{\max} = \max_{j\in[n]} \mathbf{A}_{j,j}/|(\mathbf{A}+\mathbf{N})_{j,j}|$.

Theorem 178. *(Informal lower bound.) Let $\epsilon > \eta > 0$. Given $\mathbf{A}+\mathbf{N}$ such that $\mathbf{A}$ is PSD and $\mathbf{N}$ is a corruption matrix as defined above, any randomized algorithm that with probability at least $2/3$ outputs a rank-k approximation up to additive error $(\epsilon+\eta)\|\mathbf{A}\|_F^2$ must read $\Omega\left(\phi_{\max}^2 nk/\epsilon\right)$ entries of $\mathbf{A}+\mathbf{N}$.*

Remark 162. Any algorithm must incur additive error $\eta\|\mathbf{A}\|_F^2$, since $\mathbf{A}$ is not even identifiable below additive-error $\eta\|\mathbf{A}\|_F^2$.

Remark 163. In our hard instance, $\phi_{\max}^2$ can be as large as $\epsilon n/k$, which implies a sample-complexity lower bound of $\Omega(n^2)$. While this lower bound precludes sublinear algorithms for arbitrary PSD matrices, we observe that in many applications $\phi_{\max}$ can be significantly smaller. For instance, if $\mathbf{A}$ is a correlation matrix, we know that the true diagonal entries of $\mathbf{A}+\mathbf{N}$ are 1 and can ignore any corruption on them to bound $\phi_{\max}$ by 1.

Motivated by the aforementioned observation, we introduce algorithms for robust low-rank approximation, parameterized by the corruption on the diagonal entries. We obtain the following theorem:

Theorem 183 *(Informal Robust LRA.) Given $\mathbf{A}+\mathbf{N}$, which satisfies our noise model, there exists an algorithm that queries $\widetilde{O}(\phi_{\max}^2 nk/\epsilon)$ entries in $\mathbf{A}+\mathbf{N}$ and computes a rank k matrix $\mathbf{B}$ such that with probability at least $99/100$, $\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant \|\mathbf{A}-\mathbf{A}_k\|_F^2 + (\epsilon+\sqrt{\eta})\|\mathbf{A}\|_F^2$.*

Remark 164. While the sample complexity of this algorithm matches the sample complexity in the lower bound, it incurs additive-error $\sqrt{\eta}\|\mathbf{A}\|_F^2$ as opposed to $\eta\|\mathbf{A}\|_F^2$. An interesting open question here is whether we can achieve additive-error $o(\sqrt{\eta}\|\mathbf{A}\|_F^2)$, though we note that when $\eta^2 \leqslant \epsilon$, this just changes the additive error guarantee of our low-rank approximation by a constant factor.

Remark 165. Our techniques extend to low-rank approximation of correlation matrices, and we obtain a sample complexity of $\widetilde{O}(nk/\epsilon)$, which is optimal. In fact, the hard instance in [MW17c] implies an $\Omega(nk/\epsilon)$ lower bound on the sample complexity, even in the presence of no noise.

Surprisingly, corrupting a correlation matrix does not increase the sample complexity and only incurs an additive error of $\sqrt{\eta}\|\mathbf{A}\|_F^2$ (see Corollary 8.4.11 for a formal statement).

8.2 Preliminaries and Notation

Given an $m \times n$ matrix $\mathbf{A}$ with rank r, we can compute its singular value decomposition, denoted by $\texttt{SVD}(\mathbf{A}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, such that $\mathbf{U}$ is an $m \times r$ matrix with orthonormal columns, $\mathbf{V}^\top$ is an $r \times n$ matrix with orthonormal rows and $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix. The entries along the diagonal are the singular values of $\mathbf{A}$, denoted by $\sigma_1, \sigma_2 \ldots \sigma_r$. Given an integer $k \leqslant r$, we define the truncated singular value decomposition of $\mathbf{A}$ that zeros out all but the top k singular values of $\mathbf{A}$, i.e., $\mathbf{A}_k = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^\top$, where $\mathbf{\Sigma}_k$ has only k non-zero entries along the diagonal. It is well known that the truncated SVD computes the best rank-k approximation to $\mathbf{A}$ under the Frobenius norm, i.e., $\mathbf{A}_k = \min_{\text{rank}(\mathbf{X}) \leqslant k} \|\mathbf{A} - \mathbf{X}\|_F$. More generally, for any matrix $\mathbf{M}$, we use the notation $\mathbf{M}_k$ and $\mathbf{M}_{\setminus k}$ to denote the first k components and all but the first k components respectively. We use $\mathbf{M}_{i,*}$ and $\mathbf{M}_{*,j}$ to refer to the i^{th} row and j^{th} column of $\mathbf{M}$ respectively. For an $n \times n$ PSD matrix $\mathbf{A}$, we denote the singular (eigenvalue) decomposition by $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$. Further, since $\mathbf{\Sigma}_{i,i} \geqslant 0$, let $\mathbf{A}^{1/2} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{U}^\top$ be the square root of $\mathbf{A}$. Note that $\mathbf{A}_{i,j} = \langle \mathbf{A}_{i,*}^{1/2}, \mathbf{A}_{j,*}^{1/2} \rangle$. By Cauchy-Schwarz, for all $i, j \in [n]$, $\mathbf{A}_{i,j}^2 = \langle \mathbf{A}_{i,*}^{1/2}, \mathbf{A}_{j,*}^{1/2} \rangle^2 \leqslant \|\mathbf{A}_{i,*}^{1/2}\|_2^2 \cdot \|\mathbf{A}_{j,*}^{1/2}\|_2^2 = \mathbf{A}_{i,i} \cdot \mathbf{A}_{j,j}$. We use $\mathsf{nnz}(\mathbf{A})$ to denote the number of non-zero entries (sparsity) of $\mathbf{A}$. We use operator and spectral norm interchangeably to denote $\|\mathbf{M}\|_2 = \max_{\|y\|_2=1} \|\mathbf{M}y\|_2$. We also use the notation $\mathbf{M}^\dagger$ to denote the Moore-Penrose pseudoinverse.

8.3 Relative Error PSD Low-Rank Approximation

In this section, we describe our main algorithm for *relative-error* PSD Low-Rank Approximation, where we query only $\widetilde{O}(nk/\epsilon)$ of the input matrix $\mathbf{A}$. This improves the best known algorithm by Musco and Woodruff that queries $\widetilde{O}(nk/\epsilon^{2.5})$ and matches their query lower bound of $\Omega(nk/\epsilon)$ up to polylogarithmic factors [MW17c]. Formally, we prove the following:

Theorem 166. *(Sample-Optimal PSD Low-Rank Approximation.) Given an $n \times n$ PSD matrix $\mathbf{A}$, an integer k, and $1 > \epsilon > 0$, Algorithm 11 samples $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ and outputs matrices $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n \times k}$ such that with probability at least $9/10$,*

$$\|\mathbf{A} - \mathbf{M}\mathbf{N}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Further, the algorithm runs in $\widetilde{O}(n(k/\epsilon)^{\omega-1} + (k/\epsilon^3)^{\omega})$ *time.*

We begin by defining various statistical quantities associated with a given matrix, such as the leverage and ridge-leverage scores. The leverage score of a given row measures the importance of this row in composing the row span. Leverage scores have found numerous applications in regression, preconditioning, linear programming and graph sparsification [Sar06, SS11, LS15, CLM+15]. In the special case of graphs, they are referred to as *effective resistances*.

Definition 8.3.1. *(Leverage Scores.) Given a matrix* $\mathbf{M} \in \mathbb{R}^{n \times m}$*, let* $\boldsymbol{m}_i = \mathbf{M}_{i,*}$ *be the* i*-th row of* $\mathbf{M}$*. Then, for all* $i \in [n]$ *the* i*-th row leverage score of* $\mathbf{M}$ *is given by*

$$\tau_i(\mathbf{M}) = \boldsymbol{m}_i(\mathbf{M}^\top \mathbf{M})^\dagger \boldsymbol{m}_i^\top$$

The column leverage scores can be defined analogously. Note, in the special case where $\mathbf{M}$ has orthonormal columns, the row leverage scores of $\mathbf{M}$ are simply the ℓ_2^2 norms of the rows i.e., $\tau_i(\mathbf{M}) = \|\mathbf{m}_i\|_2^2$. It is well-known that sampling rows of a matrix proportional to the leverage scores satisfies the subspace embedding property (Spectral Sparsification for Graphs) and leads to faster algorithms for ℓ_2-norm Regression. Recall, for an $n \times m$ matrix $\mathbf{A}$, a leverage score sampling matrix $\mathbf{S} = \mathbf{D}\mathbf{\Omega}^\top$, where $\mathbf{D}$ is a $t \times t$ diagonal matrix and $\mathbf{\Omega}$ is an $n \times t$ sampling matrix. For all $j \in [t]$, select row index $i \in [n]$ with probability $p_i = \tau_i(\mathbf{A}) / \sum_i \tau_i(\mathbf{A})$ and set $\mathbf{\Omega}_{i,j} = 1$ and $\mathbf{D}_{j,j} = 1/\sqrt{tp_i}$.

Lemma 8.3.2. *(Subspace Embedding.) Given a matrix* $\mathbf{A} \in \mathbb{R}^{n \times m}$*,* $\epsilon > 0$*, and a leverage score sampling matrix* $\mathbf{S}$ *with* $t = O(mlog(m)/\epsilon^2)$ *rows, with probability at least* $99/100$*, for all* $x \in \mathbb{R}^m$

$$\|\mathbf{S}\mathbf{A}x\|_2^2 = (1 \pm \epsilon)\|\mathbf{A}x\|_2^2$$

This simply follows from an application of the Matrix Chernoff bound. Observe that the sketch preserves all the singular values of $\mathbf{A}$ up to a factor of $1 \pm \epsilon$. We refer the reader to a recent survey for more details [Woo14a]. Next, we recall that leverage score sampling results in a fast algorithm for regression.

Lemma 8.3.3. *(Fast Regression, Theorem 38 [CW13].) Given matrices* $\mathbf{A} \in \mathbb{R}^{n \times m}, \mathbf{B} \in \mathbb{R}^{n \times d}$ *such that rank*$(\mathbf{A}) \leqslant r$ *and* $\epsilon > 0$*, sample* $O(r \log(r) + r/\epsilon)$ *rows of* $\mathbf{A}, \mathbf{B}$ *proportional to the leverage scores of* $\mathbf{A}$ *to obtain a sketch* $\mathbf{S}$ *such that* $\mathbf{Y}^* = \arg\min_{\mathbf{Y}} \|\mathbf{S}\mathbf{A}\mathbf{Y} - \mathbf{S}\mathbf{B}\|_F^2$*. Then, with*

probability at least $1 - c$,

$$\|\mathbf{A}\mathbf{Y}^* - \mathbf{B}\|_F^2 \leqslant (1+\epsilon)\min_{\mathbf{Y}} \|\mathbf{A}\mathbf{Y} - \mathbf{B}\|_F^2$$

for a fixed small constant c. *Further, the time to compute* $\mathbf{Y}^*$ *is* $O(\mathsf{nnz}(\mathbf{A})\log(r/\epsilon) + (n+d)(r/\epsilon)^{\omega-1} + \mathit{poly}(r/\epsilon))$.

Note, the terms in the running time follow from using Cohen's construction for OSNAP [Coh16]. Leverage score sampling matrices also approximately preserve norms in affine spaces, which leads to faster algorithms for multi-response regression, i.e., $\min_{\mathbf{X}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2$, where $\mathbf{B}$ now has a large number of columns.

Lemma 8.3.4. *(Affine Embeddings, Theorem 39 [CW13].) Given matrices* $\mathbf{A} \in \mathbb{R}^{n\times m}$ *, such that rank*$(\mathbf{A}) = r$*, and* $\mathbf{B} \in \mathbb{R}^{n\times d}$*, let* $\mathbf{S}$ *be a leverage score sampling matrix with* $t = O(r/\epsilon^2)$ *rows. Further, let* $\mathbf{X}^*$ *be the optimizer for* $\min_{\mathbf{X}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2$ *and let* $\mathbf{B}^* = \mathbf{A}\mathbf{X}^* - \mathbf{B}$. *Then, with probability at least* $1 - c$*, for all* $\mathbf{X} \in \mathbb{R}^{m\times d}$

$$\|\mathbf{S}\mathbf{A}\mathbf{X} - \mathbf{S}\mathbf{B}\|_F^2 - \|\mathbf{S}\mathbf{B}^*\|_F^2 = (1\pm\epsilon)\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2 - \|\mathbf{B}^*\|_F^2$$

for a fixed small constant c.

An important application of the above lemma (which we use extensively) is to sketch constrained regression problems, for example, when the matrix $\mathbf{X}$ has a fixed small rank. Since affine embeddings approximately preserve the cost of all affine spaces up to a fixed shift, this guarantee in particular holds for $\mathbf{X}$ with small rank. Recall, an important caveat here is that the cost of the sketched problem is not a relative-error approximation to the cost of the original problem since we cannot estimate $\|\mathbf{B}^*\|_F^2$ in general. However, the upshot here is that the aforementioned guarantee still suffices for optimization since the fixed shift does not change the optimizer.

The next tool we use is input-sparsity time low-rank approximation. This was achieved by Clarkson and Woodruff [CW13] and the exact dependence on k, ϵ was improved in subsequent works [MM13b, NN13a, BDN15, Coh16]. While the standard low-rank approximation guarantee achieves relative-error under Frobenius norm, here we will require a spectral norm bound, which follows from results of [CEM+15, CMM17].

Lemma 8.3.5. *(Input-Sparsity Spectral LRA [CEM+15, CMM17].) Given a matrix* $\mathbf{A} \in \mathbb{R}^{n\times d}$*,* $\epsilon, \delta > 0$ *and* $k \in \mathbb{N}$*, let* $k' = O(k/\epsilon)$. *Then, there exists an algorithm that outputs a matrix*

$\mathbf{Z}^\top \in \mathbb{R}^{k' \times n}$ *such that with probability at least* $1 - \delta$*,*

$$\|\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\top\|_2^2 \leqslant O\left(\frac{\epsilon}{k}\right)\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2$$

in time and query complexity $\widetilde{O}\left(\mathsf{nnz}(\mathbf{A}) + (n+d)\mathit{poly}(k/\epsilon\delta)\right)$*.*

Proof. By Lemma 18 from [CEM+15] it suffices to use any obvious subspace embedding matrix with $\epsilon = O(1)$ and $k = k/\epsilon$. Here, we use OSNAP in the regime that requires $\widetilde{O}(k/\epsilon^2)$ rows and sparsity $\text{polylog}(k)/\epsilon$ [NN13a]. Instantiating this OSNAP construction with $\epsilon = O(1)$ and $k = k/\epsilon$ results in $\mathbf{Z}^\top$ with k/ϵ rows in the desired running time. □

Next we define the ridge leverage scores of a matrix. The ridge leverage scores were used as sampling probabilities in the context of linear regression and spectral approximation [LMP13, KLM+17, AM15], and low-rank approximation [CMM17, MW17c]. Intuitively, the ridge leverage scores can be thought of as adding a regularization term that attenuates the smaller singular directions such that they are sampled with proportionately lower probability.

Definition 8.3.6. *(Ridge Leverage Scores.) Given a matrix* $\mathbf{M} \in \mathbb{R}^{n \times m}$ *and an integer* k*, let* $\boldsymbol{m}_i = \mathbf{M}_{i,*}$ *be the* i*-th row of* $\mathbf{M}$*. Then, for all* $i \in [n]$*, the* i*-th rank-*k *ridge leverage score of* $\mathbf{M}$ *is*

$$\rho_i^k(\mathbf{M}) = \boldsymbol{m}_i \left(\mathbf{M}^\top\mathbf{M} + \frac{\|\mathbf{M} - \mathbf{M}_k\|_F^2}{k}\mathbf{I}\right)^\dagger \boldsymbol{m}_i^\top$$

Since we typically use the row ridge leverage scores to define a probability distribution over the rows and sample according to this distribution, it is crucial that their sum is small as this controls the number of rows we would need to sample. This follows from a straightforward calculation:

Lemma 8.3.7. *(Lemma 4 from [CMM17].) Let* $\rho_i^k(\mathbf{M})$ *be the* i*-th ridge leverage score of* $\mathbf{M}$*. Then,*

$$\sum_{i \in [n]} \rho_i^k(\mathbf{M}) \leqslant 2k$$

Cohen et. al. [CMM17] show that the ridge leverage scores of a matrix can be approximated up to a small constant in $O(\mathsf{nnz}(\mathbf{A}))$ time, however this involves reading the entire matrix $\mathbf{A}$. For the special case of $\mathbf{A}$ being PSD, Musco and Musco [MM17] show that the ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$ can be approximated up to a small constant using a so-called Nyström approximation.

Lemma 8.3.8. *(Lemma 4 of [MW17c].) Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and integer k, there exists an algorithm that accesses $O(nk\log(k/\delta))$ entries in $\mathbf{A}$ and computes $\hat{\rho}_i^k(\mathbf{A}^{\frac{1}{2}})$ for all $i \in [n]$, such that with probability $1-\delta$,*

$$\rho_i^k(\mathbf{A}^{1/2}) \leqslant \hat{\rho}_i^k(\mathbf{A}^{1/2}) \leqslant 3\rho_i^k(\mathbf{A}^{1/2})$$

and runs in time $O(n(k\log(k/\delta))^{\omega-1})$, where ω is the matrix multiplication exponent.

Note, while it is not known how to compute ridge leverage scores of a PSD matrix in sublinear time, Musco and Woodruff [MW17c] show that the ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$ are a coarse approximation to the ridge leverage scores of $\mathbf{A}$.

Lemma 8.3.9. *(Lemma 5 in [MW17c].) Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, for all $i \in [n]$,*

$$\rho_i^k(\mathbf{A}) \leqslant 2\sqrt{\frac{n}{k}}\rho_i^k(\mathbf{A}^{\frac{1}{2}})$$

Musco and Woodruff then show that sampling columns of $\mathbf{A}$, according to the corresponding ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$, suffices to obtain a column projection-cost preserving sketch (PCP), if we oversample by a $\sqrt{n/k}$ factor. Projection-cost preserving sketches were introduced by Feldmen et. al. [FSS13] and Cohen et. al. [CEM+15] and studied in the context of low-rank approximation in [CMM17, MW17c, BW18].

Lemma 8.3.10. *(Column PCP from [MW17c].) Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, integer k and $\epsilon > 0$, for all $j \in [n]$ let $\bar{\rho}_j^k(\mathbf{A}^{\frac{1}{2}})$ be a constant approximation to the column-ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$. Let $q_j = \bar{\rho}_j^k(\mathbf{A}^{\frac{1}{2}})/\sum_j \bar{\rho}_j^k(\mathbf{A}^{\frac{1}{2}})$ and let $t = O\left(\sqrt{\frac{n}{k}}\sum_j \bar{\rho}_j^k(\mathbf{A}^{\frac{1}{2}})\log(k/\delta)/\epsilon^2\right) = O\left(\sqrt{nk}\log(k/\delta)/\epsilon^2\right)$. Construct $\mathbf{C} \in \mathbb{R}^{n\times t}$ by sampling t columns of $\mathbf{A}$ and setting each one to be $\frac{1}{\sqrt{tq_j}}\mathbf{A}_{*,j}$ with probability q_j. Then, with probability $1-\delta$, for any rank-k projection matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$,*

$$(1-\epsilon)\|\mathbf{A}-\mathbf{X}\mathbf{A}\|_F^2 \leqslant \|\mathbf{C}-\mathbf{X}\mathbf{C}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{X}\mathbf{A}\|_F^2$$

Further, such a $\mathbf{C}$ can be computed by accessing $\widetilde{O}(nk)$ entries in $\mathbf{A}$ and in time $O(nk^{\omega-1})$.

This result also implies that the resulting matrix $\mathbf{C}$ is a *Spectral-Frobenius PCP* for $\mathbf{A}$ (Lemma 24 in [MW17c]), i.e., for any rank-k projection matrix $\mathbf{X}$,

$$(1-\epsilon)\|\mathbf{A}-\mathbf{X}\mathbf{A}\|_2^2 - \frac{\epsilon}{k}\|\mathbf{A}-\mathbf{A}_k\|_F^2 \leqslant \|\mathbf{C}-\mathbf{X}\mathbf{C}\|_2^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{X}\mathbf{A}\|_2^2 + \frac{\epsilon}{k}\|\mathbf{A}-\mathbf{A}_k\|_F^2 \quad (8.3)$$

As noted by Musco and Woodruff, the resulting matrix $\mathbf{C}$ is not even square and thus is it unclear how to sample rows of $\mathbf{C}$ to obtain a row-PCP in sublinear time and queries. In particular, the ridge-leverage scores of rows of $\mathbf{C}$ can be an n/k-factor larger than the corresponding ridge-leverage scores of $\mathbf{A}^{\frac{1}{2}}$. Instead, Musco and Woodruff sample rows of $\mathbf{C}$ proportional to the rank-k/ϵ^2 ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$. In addition, they show the stronger guarantee that a *Spectral-Frobenius PCP* holds (by Lemma 8 of [MW17c]) for PSD Matrices.

Lemma 8.3.11. *(Spectral-Frobenius PCP.) Given a PSD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, an integer k and $\epsilon > 0$, let $\mathbf{C} \in \mathbb{R}^{n\times t}$ be a column PCP for $\mathbf{A}$, following Lemma 8.3.10. Let $k' = k/\epsilon^2$. For all $i \in [n]$, let $\bar{\rho}_i^{k'}(\mathbf{A}^{\frac{1}{2}})$ be a constant approximation to the rank-k' row ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$. Let $p_i = \bar{\rho}_i^{k'}(\mathbf{A}^{\frac{1}{2}})/\sum_i \bar{\rho}_i^{k'}(\mathbf{A}^{\frac{1}{2}})$ and let $t = O\left(\sqrt{\frac{n}{k}}\sum_i \bar{\rho}_i^{k'}(\mathbf{A}^{\frac{1}{2}})\log(n)/\epsilon\right) = O\left(\sqrt{nk}\log(n)/\epsilon^3\right)$. Then, with probability $1 - c$, for all rank-k' projection matrices $\mathbf{X}$,*

$$(1-\epsilon)\|\mathbf{C}-\mathbf{C}\mathbf{X}\|_2^2 - \frac{\epsilon}{k}\|\mathbf{A}-\mathbf{A}_k\|_F^2 \leqslant \|\mathbf{R}-\mathbf{R}\mathbf{X}\|_2^2 \leqslant (1+\epsilon)\|\mathbf{C}-\mathbf{C}\mathbf{X}\|_2^2 + \frac{\epsilon}{k}\|\mathbf{A}-\mathbf{A}_k\|_F^2$$

We observe that we could compute a low-rank approximation to $\mathbf{R}$ in input sparsity time, which already requires querying $\Omega(\mathsf{nnz}(\mathbf{R})) = \Omega(nk/\epsilon^4)$ entries in $\mathbf{A}$ and is far from optimal in terms of the dependence on ϵ. It is here that we digress from the approach of Musco and Woodruff. We observe that the dependence on n and k is optimal and thus we instantiate the aforementioned column and row PCPs with $\epsilon = O(1)$ and $k = k/\epsilon$. While this results in weaker PCP guaranteees, the resulting matrix $\mathbf{R}$ is a $\sqrt{nk/\epsilon} \times \sqrt{nk/\epsilon}$ matrix and we can now afford to read all of it and thus we can compute a rank-k low-rank approximation to $\mathbf{R}$ using the input sparsity time algorithm of Clarkson and Woodruff [CW13].

However, the main technical challenge here is that we can no longer use the approach of [CMM17, MW17c, BW18] to use the low-rank approximation for $\mathbf{R}$ and solve regression problems to recover an ϵ-approximate low-rank matrix for $\mathbf{A}$. In particular, we can now only hope for an $O(1)$ approximation if we use the standard technique of iteratively solving regression problems. Our first insight is that computing a Spectral Low-Rank Approximation to $\mathbf{R}$ results in a *structured* projection matrix for $\mathbf{C}$, from which we can compute a *structured* projection matrix for $\mathbf{A}$. Further, this structured projection can be computed with only $\widetilde{O}(nk/\epsilon)$ queries. We first describe how this structured projection matrix for $\mathbf{A}$ results in an efficient low-rank approximation algorithm.

8.3.1 Structured Projections to Low-Rank Approximation

Our starting point is a structural result based on the Spectral-Frobenius projection (SF) property introduced by Clarkson and Woodruff in the context of approximating arbitrary matrices with low-rank PSD matrices [CW17]. In this subsection, we show that if we are given a projection matrix that satisfies the SF property, we can obtain a query-optimal algorithm for Low-Rank Approximation. We begin by defining this property:

Definition 8.3.12. *((ϵ, k)-SF Projection.) Given any matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, integer k, and $\epsilon > 0$, a projection matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ is (ϵ, k)-SF w.r.t. $\mathbf{A}$ if*

$$\|\mathbf{A} - \mathbf{AP}\|_2^2 \leqslant \frac{\epsilon}{k}\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

or

$$\|\mathbf{A} - \mathbf{PA}\|_2^2 \leqslant \frac{\epsilon}{k}\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Intuitively, the following structural result of Clarkson and Woodruff relates an (ϵ, k)-SF projection to a relative-error low-rank approximation. We leverage this connection heavily in subsequent sections.

Lemma 8.3.13. *(Structured Projections and Low-Rank Approximation [CW17].) Let $\mathbf{P} \in \mathbb{R}^{n \times n}$ be an (ϵ, k)-SF projection w.r.t $\mathbf{A}$, then*

$$\|\mathbf{A} - \mathbf{PA}_k\mathbf{P}\|_F^2 \leqslant (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Ignoring computational and query complexity constraints, suppose we were given a matrix $\mathbf{Q} \in \mathbb{R}^{n \times k'}$ with orthonormal columns such that $\mathbf{P} = \mathbf{QQ}^\top$ is an (ϵ, k)-SF Projection, where k' is the dimension of the space $\mathbf{P}$ projects onto. Note, for now it suffices to set $k' = \text{poly}(k/\epsilon)$. As a consequence of Lemma 8.3.13, we observe that solving the following constrained regression problem suffices to obtain a $(1 + \epsilon)$-relative error solution to the Low-Rank Approximation problem:

$$\min_{\text{rank}(\mathbf{X}) \leqslant k} \|\mathbf{A} - \mathbf{QXQ}^\top\|_F^2 \tag{8.4}$$

However, there are several challenges pertaining to this approach. As noted above, it is not immediately clear how to obtain such a $\mathbf{Q}$ with nk/ϵ queries to $\mathbf{A}$. Further, it is not immediately clear how to solve Equation 8.4 efficiently. While we have reduced to optimizing over $k' \times k'$ sized matrices $\mathbf{X}$ with rank at most k, the problem still seems intractable in sublinear time and

queries.

Algorithm 9 : Structured Projection to Low-Rank Approximation

Input: A PSD Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, integer k, $\epsilon > 0$, an orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{n \times k'}$ such that the projection matrix $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ satisfies $\|\mathbf{A} - \mathbf{P}\mathbf{A}\|_2^2 \leqslant \frac{\epsilon}{k}\|\mathbf{A} - \mathbf{A}_k\|_F^2$

1. Consider the optimization problem:

$$\min_{\text{rank}(\mathbf{X}) \leqslant k} \|\mathbf{A} - \mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2$$

2. For all $i \in [n]$, compute the leverage scores, $\tau_i(\mathbf{Q})$. Since $\mathbf{Q}$ has orthonormal columns, $\tau_i(\mathbf{Q}) = \|\mathbf{Q}_{i,*}\|_2^2$ and can be computed exactly. Let $p = \{p_1, p_2 \dots p_n\}$ denote a distribution over rows of $\mathbf{A}$ for which $p_i = \tau_i(\mathbf{Q}) / \sum_{i'} \tau_{i'}(\mathbf{Q})$.

3. Let $t = k'/\epsilon^2$. Construct a *leverage score sampling* matrix $\mathbf{S}$ by sampling t rows of $\mathbf{A}$, such that $\mathbf{S} = \mathbf{D}\mathbf{\Omega}^\top$, where $\mathbf{D}$ is a $t \times t$ diagonal matrix and $\mathbf{\Omega}$ is an $n \times t$ sampling matrix. For all $j \in [t]$, select row index $i \in [n]$ with probability p_i and set $\mathbf{\Omega}_{i,j} = 1$ and $\mathbf{D}_{j,j} = 1/\sqrt{tp_i}$. Repeat this sampling process to construct another *leverage score sampling* matrix $\mathbf{T}$.

4. Consider the sketched optimization problem :

$$\min_{\text{rank}(\mathbf{X}) \leqslant k} \|\mathbf{S}\mathbf{A}\mathbf{T} - \mathbf{S}\mathbf{Q}\mathbf{X}\mathbf{Q}^\top\mathbf{T}\|_F^2$$

   Compute $\mathbf{SAT}$, $\mathbf{P}_{\mathbf{SQ}}$, $\mathbf{P}_{\mathbf{Q}^\top\mathbf{T}}$,$(\mathbf{SQ})^\dagger$ and $(\mathbf{Q}^\top\mathbf{T})^\dagger$, where $\mathbf{P}_{\mathbf{SQ}}$ and $\mathbf{P}_{\mathbf{Q}^\top\mathbf{T}}$ are the projections onto $\mathbf{SQ}$ and $\mathbf{Q}^\top\mathbf{T}$ respectively. Compute $\mathsf{SVD}(\mathbf{P}_{\mathbf{SQ}}\mathbf{SAT}\mathbf{P}_{\mathbf{Q}^\top\mathbf{T}})$. By Theorem 167 the sketched problem is minimized by $\mathbf{X}^* = (\mathbf{SQ})^\dagger[\mathbf{P}_{\mathbf{SQ}}\mathbf{SAT}\mathbf{P}_{\mathbf{Q}^\top\mathbf{T}}]_k(\mathbf{Q}^\top\mathbf{T})^\dagger$.

5. Let $\mathbf{U}^* \in \mathbb{R}^{k' \times k}$ be an orthonormal basis for the columns of $\mathbf{X}^*$. Compute an orthonormal basis $\mathbf{M}$ for $\mathbf{QU}^*$. Consider the following regression problem: $\min_{\mathbf{Y} \in \mathbb{R}^{k \times n}} \|\mathbf{A} - \mathbf{M}\mathbf{Y}\|_F^2$. For all $i \in [n]$, compute $\tau_i(\mathbf{M}) = \|\mathbf{M}_{i,*}\|_2^2$. Let $q = \{q_1, q_2, \dots, q_n\}$ be a distribution over the rows of $\mathbf{A}$ such that $q_i = \tau_i(\mathbf{M}) / \sum_{i' \in [n]} \tau_{i'}(\mathbf{M})$. Let $\mathbf{W}$ be a *leverage score sampling* matrix with k/ϵ rows sampled proportional to q.

6. Consider the sketched regression problem: $\min_{\mathbf{Y} \in \mathbb{R}^{k \times n}} \|\mathbf{W}\mathbf{A} - \mathbf{W}\mathbf{M}\mathbf{Y}\|_F^2$. Let $\mathbf{N}$ be the minimizer to this regression problem computed using the algorithm from Lemma 8.3.3.

Output: $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n \times k}$ such that $\|\mathbf{A} - \mathbf{M}\mathbf{N}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$

We begin by describing how to solve the optimization problem in Equation 8.4 with $\widetilde{O}(nk/\epsilon)$ queries given that we have access to $\mathbf{Q}$ and $\mathbf{Q}\mathbf{Q}^\top$ is an (ϵ, k)-SF Projection. At a high level, our approach is to sketch the problem by sampling rows and columns proportional to the row leverage scores of $\mathbf{Q}$. We observe that since $\mathbf{Q}$ has orthonormal columns, the row leverage scores of $\mathbf{Q}$ are simply the ℓ_2^2 norms of corresponding rows. Therefore, we create sampling matrices $\mathbf{S}$ and $\mathbf{T}$ that sample $\text{poly}(k')$ rows proportional to the leverage scores of $\mathbf{Q}$ and consider the resulting optimization problem:

$$\min_{\text{rank}(X)\leqslant k} \|\mathbf{SAT} - \mathbf{SQXQ}^\top\mathbf{T}\|_F^2 \tag{8.5}$$

We then show that the minimizer for Equation 8.5 is an approximate minimizer for Equation 8.4. Further, the optimization problem in Equation 8.5 is referred to as Generalized Low-Rank Approximation and admits a closed form solution:

Theorem 167. *(Generalized Low-Rank Approximation [FT07].) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times k'}$ and $\mathbf{C} \in \mathbb{R}^{k'\times n}$ and $k \in \mathbb{N}$. Then, the Generalized Low-Rank Approximation problem*

$$\min_{rank(\mathbf{X})\leqslant k} \|\mathbf{A} - \mathbf{BXC}\|_F^2$$

is minimized by $\mathbf{X} = \mathbf{B}^\dagger[\mathbf{P}_B\mathbf{A}\mathbf{P}_C]_k\mathbf{C}^\dagger$, where $\mathbf{P}_B, \mathbf{P}_C$ are the projection matrices onto $\mathbf{B}$ and $\mathbf{C}$ respectively.

We apply the above theorem to Equation 8.5. Both the query complexity and running time here contribute a lower-order term and we can afford to compute the SVD for each term. Let $\mathbf{X}^*$ be the solution to the sketched optimization problem in Equation 8.5. Then, we can compute $\mathbf{U}^*$, an orthonormal column basis for $\mathbf{X}^*$ and consider $\mathbf{M}$, an orthonormal basis for $\mathbf{QU}^* \in \mathbb{R}^{n\times k}$ to be one of the low-rank factors for $\mathbf{A}$. To find the second factor, we set up the following regression problem:

$$\min_{\mathbf{N}\in\mathbb{R}^{k\times n}} \|\mathbf{A} - \mathbf{MN}\|_F^2 \tag{8.6}$$

Again, $\mathbf{M}$ has orthonormal columns and thus we can efficiently compute the corresponding row leverage scores and sample k/ϵ rows. By Lemma 8.3.4 this achieves a $(1+\epsilon)$-approximation to the optimal cost in Equation 8.6 and obtains an $\mathbf{N}^*$ with $\widetilde{O}(nk/\epsilon)$ queries to $\mathbf{A}$. At this stage, we have obtained a $(1+\epsilon)$-approximate rank-k solution to Equation 8.4 and Lemma 8.3.13 implies that we are done. We now formalize this argument:

Theorem 168. *(Structured Projection to Low-Rank Approximation.) Given a rank-k' projection*

matrix $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$*, such that* $\mathbf{P}$ *is an* (ϵ, k)*-SF projection, Algorithm 9 queries* $\widetilde{O}(nk/\epsilon + k'^2/\epsilon^4)$ *entries in* $\mathbf{A}$ *and with probability* $99/100$ *outputs* $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n\times k}$ *such that*

$$\|\mathbf{A} - \mathbf{M}\mathbf{N}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Further, Algorithm 9 runs in time $\widetilde{O}(n(k/\epsilon)^{\omega-1} + nk'^{\omega-1} + (k'/\epsilon^2)^\omega)$.

Proof. Since $\mathbf{P}$ is an (ϵ, k)-SF projection and $\mathbf{A}_k$ is a feasible solution to $\min_{\text{rank}(\mathbf{Y})\leqslant k} \|\mathbf{A} - \mathbf{P}\mathbf{Y}\mathbf{P}\|_F^2$, from Lemma 8.3.13 we have

$$\min_{\text{rank}(\mathbf{Y})\leqslant k} \|\mathbf{A} - \mathbf{P}\mathbf{Y}\mathbf{P}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2 \tag{8.7}$$

Since $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$, we can substitute it in Equation 8.7 to get $\min_{\text{rank}(\mathbf{Y})\leqslant k} \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^\top\mathbf{Y}\mathbf{Q}\mathbf{Q}^\top\|_F^2$. We further relax this by optimizing over all rank-k matrices $\mathbf{X} \in \mathbb{R}^{k'\times k'}$ instead of matrices of the form $\mathbf{Q}\mathbf{Y}\mathbf{Q}^\top$. Therefore,

$$\min_{\text{rank}(\mathbf{X})\leqslant k} \|\mathbf{A} - \mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2 \tag{8.8}$$

where we are now optimizing over a $k' \times k'$ matrix $\mathbf{X}$, which is considerably smaller than $\mathbf{Y}$. Let $\mathbf{S}, \mathbf{T}^\top \in \mathbb{R}^{k'/\epsilon^2\times n}$ be the *leverage score sampling* matrices as defined in Algorithm 9. Observe, from Lemma 8.3.4 we know $\mathbf{S}$ has a sufficient number of rows to be an affine embedding for Equation 8.8. However, we cannot directly apply the affine embedding guarantee since $\mathbf{A} - \mathbf{Q}\mathbf{X}\mathbf{Q}^\top$ is not an affine space. Let $\mathbf{H}$ be $k' \times n$ matrix, let $\mathbf{H}^* = \arg\min_{\mathbf{H}} \|\mathbf{A} - \mathbf{Q}\mathbf{H}\|_F^2$ and let $\mathbf{A}^*$ be $\mathbf{A} - \mathbf{Q}\mathbf{H}^*$. Then, with probability at least $1 - c_1$, for all $\mathbf{H}$,

$$\|\mathbf{S}\mathbf{A} - \mathbf{S}\mathbf{Q}\mathbf{H}\|_F^2 - \|\mathbf{S}\mathbf{A}^*\|_F^2 = (1\pm\epsilon)\|\mathbf{A} - \mathbf{Q}\mathbf{H}\|_F^2 + \|\mathbf{A}^*\|_F^2 \tag{8.9}$$

Since Equation 8.9 holds for all $\mathbf{H}$, in particular it holds for all rank-k matrices $\mathbf{X}$ such that $\mathbf{H} = \mathbf{X}\mathbf{Q}^\top$. Therefore, with probability at least $1 - c_1$, for all rank k matrices $\mathbf{X}$,

$$\|\mathbf{S}\mathbf{A} - \mathbf{S}\mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2 - \|\mathbf{S}\mathbf{A}^*\|_F^2 = (1\pm\epsilon)\|\mathbf{A} - \mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2 + \|\mathbf{A}^*\|_F^2 \tag{8.10}$$

Here, we observe that while we cannot estimate $\|\mathbf{A}^*\|_F^2$ accurately, it is a fixed matrix independent of $\mathbf{X}$ and thus we can still approximately optimize. Let ζ_1 be the event that Equation 8.10 holds. We now use the sampling matrix $\mathbf{T}$ to sketch $\|\mathbf{S}\mathbf{A} - \mathbf{Z}\mathbf{Q}^\top\|_F^2$. Let $\mathbf{Z}' = \arg\min_{\mathbf{Z}} \|\mathbf{S}\mathbf{A} - \mathbf{Z}\mathbf{Q}^\top\|_F^2$ and let $\mathbf{S}\mathbf{A}' = \mathbf{S}\mathbf{A} - \mathbf{Z}'\mathbf{Q}^\top$. Then, with probability at least $1 - c_2$, for

all $\mathbf{Z}$,

$$\|\mathbf{SAT} - \mathbf{ZQ}^\top\mathbf{T}\|_F^2 - \|\mathbf{SA}'\mathbf{T}\|_F^2 = (1 \pm \epsilon)\|\mathbf{SA} - \mathbf{ZQ}^\top\|_F^2 + \|\mathbf{SA}'\|_F^2 \tag{8.11}$$

In particular, the above equation holds for all rank-k matrices $\mathbf{X}$ such that $\mathbf{Z} = \mathbf{SQ}^\top\mathbf{X}$. Let ζ_2 be the event that the aforementioned equation holds. Combining equations 8.10 and 8.11 and conditioning on ζ_1 and ζ_2, for all rank-k matrices $\mathbf{X}$,

$$\|\mathbf{SAT}-\mathbf{SQXQ}^\top\mathbf{T}\|_F^2-\|\mathbf{SA}'\mathbf{T}\|_F^2 = (1\pm\epsilon)^2\left(\|\mathbf{A} - \mathbf{QXQ}^\top\|_F^2 + \|\mathbf{SA}^*\|_F^2 + \|\mathbf{A}^*\|_F^2\right)+\|\mathbf{SA}'\|_F^2 \tag{8.12}$$

Here, we observe that while the sketch does not preserve the cost of all $\mathbf{X}$ up to relative error $(1+\epsilon)$, the additive error $\Delta \leqslant (1+\epsilon)\left(\|\mathbf{SA}^*\|_F^2 + \|\mathbf{A}^*\|_F^2 + \|\mathbf{SA}'\|_F^2 + \|\mathbf{SA}'\mathbf{T}\|_F^2\right)$ is fixed and is independent of $\mathbf{X}$. Let $\mathbf{X}^* = \arg\min_{\text{rank}(\mathbf{X})\leqslant k}\|\mathbf{SAT} - \mathbf{SQXQ}^\top\mathbf{T}\|_F^2$. Then, union bounding over ζ_1 and ζ_2, with probability $1 - c_1 - c_2$,

$$\|\mathbf{A} - \mathbf{QX}^*\mathbf{Q}^\top\|_F^2 \leqslant (1+\epsilon)\min_{\text{rank}(\mathbf{X})\leqslant k}\|\mathbf{A} - \mathbf{SQXQ}^\top\|_F^2 \tag{8.13}$$

Therefore, it suffices to efficiently compute $\mathbf{X}^*$. By Theorem 167, we know that the sketched optimization problem above is minimized by $\mathbf{X}^* = (\mathbf{SQ})^\dagger[\mathbf{P}_{\mathbf{SQ}}\mathbf{SATP}_{\mathbf{Q}^\top\mathbf{T}}]_k(\mathbf{Q}^\top\mathbf{T})^\dagger$, which can be computed exactly as shown in Step 4 of Algorithm 9. We note that we can now explicitly compute $\mathbf{SAT}$ by querying the relevant entries in $\mathbf{A}$. Further, we can compute $\mathbf{SQ}$ and $\mathbf{Q}^\top\mathbf{T}$ without querying $\mathbf{A}$ at all. Recalling equation 8.13 we can approximate the optimal low rank approximation cost:

$$\|\mathbf{A} - \mathbf{QX}^*\mathbf{Q}^\top\|_F^2 \leqslant (1+O(\epsilon))\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

While we have now approximately minimized the optimization problem from Equation 8.4, recall our goal was to obtain a rank-k approximation to $\mathbf{A}$ in factored form i.e., outputting $n \times k$ matrices $\mathbf{M}, \mathbf{N}^\top$ such that the low rank approximation is given by $\mathbf{MN}$. Towards this end, we compute $\mathbf{U}^*$, an orthonormal column basis for $\mathbf{X}^*$ such that $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^*$. Substituting this in the above equation we have

$$\|\mathbf{A} - \mathbf{QU}^*\mathbf{V}^*\mathbf{Q}^\top\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2 \tag{8.14}$$

Let $\mathbf{M} = \mathbf{QU}^* \in \mathbb{R}^{n\times k}$ be one of the low-rank factors for $\mathbf{A}$. To find the second factor, we observe :

$$\min_{\mathbf{Y}\in\mathbb{R}^{k\times n}}\|\mathbf{A} - \mathbf{MY}\|_F^2 \leqslant \|\mathbf{A} - \mathbf{MV}^*\mathbf{Q}^\top\|_F^2 \tag{8.15}$$

and therefore, approximately optimizing $\min_{\mathbf{Y}\in\mathbb{R}^{k\times n}}\|\mathbf{A} - \mathbf{MY}\|_F^2$ suffices. Again, $\mathbf{M}$ has or-

thonormal columns and thus we can efficiently compute the corresponding leverage scores to create a sketch $\mathbf{W}$ with $O(k/\epsilon)$ rows. From Lemma 8.3.3, with probability at least $1 - c_3$ for all $\mathbf{Y}$,

$$\|\mathbf{WA} - \mathbf{WMY}\|_F^2 = (1 \pm \epsilon)\|\mathbf{A} - \mathbf{MY}\|_F^2$$

Let $\mathbf{N}$ be the optimal solution for the sketched problem as defined in Algorithm 9. Then, with probability at least $1 - c_3$,

$$\|\mathbf{A} - \mathbf{MN}\|_F^2 \leqslant \left(\frac{1+\epsilon}{1-\epsilon}\right) \min_{\mathbf{Y} \in \mathbb{R}^{k \times n}} \|\mathbf{A} - \mathbf{MY}\|_F^2 \tag{8.16}$$

We conclude correctness by union bounding over the failure probabilities of all the sketches and observing that with probability at least $99/100$,

$$\|\mathbf{A} - \mathbf{MN}\|_F^2 \leqslant (1 + O(\epsilon))\|\mathbf{A} - \mathbf{MV}^*\mathbf{Q}^\top\|_F^2 \leqslant (1 + O(\epsilon))\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

where the inequalities follow from Equations 8.6, 8.15 and 8.14.

Finally, we analyze the query complexity and running time of our algorithm. Since Algorithm 9 is given $\mathbf{Q}$ as input, computing the leverage scores in Step 2 requires no queries to $\mathbf{A}$ and requires $O(nk')$ time. Next, observe we do not have to explicitly compute $\mathbf{SA}$ or $\mathbf{AT}$, since $\mathbf{SAT}$ is simply a submatrix of $\mathbf{A}$ with $(k'/\epsilon^2)^2$ entries appropriately scaled, it suffices to query them. $\mathbf{SAT}$ can be computed in $O(k'^2/\epsilon^4)$ time. Next, we compute $\mathsf{SVD}(\mathbf{SQ})$ and $\mathsf{SVD}(\mathbf{Q}^\top\mathbf{T})$, which requires no queries to $\mathbf{A}$ and time $O(k'^\omega/\epsilon^2)$. We can then compute $(\mathbf{SQ})^\dagger, (\mathbf{Q}^\top\mathbf{T})^\dagger, \mathbf{P}_{\mathbf{SQ}}$ and $\mathbf{P}_{\mathbf{Q}^\top\mathbf{T}}$ from the aforementioned SVDs. Next, we compute the matrix $\mathbf{P}_{\mathbf{SQ}}\mathbf{SATP}_{\mathbf{Q}^\top\mathbf{T}}$, which requires no extra queries to $\mathbf{A}$ and time $O((k'/\epsilon^2)^\omega)$, which is also the time required to compute its SVD. We can then compute $\mathbf{X}^*$ in Step 4 with a total of $O(k'^2/\epsilon^4)$ queries to $\mathbf{A}$ in time $O(nk' + (k'/\epsilon^2)^\omega)$.

In Step 5, we can compute $\mathbf{U}^*$ by computing the SVD of $\mathbf{X}^*$ and compute $\mathbf{M}$ in time $O(nk'^{\omega-1} + k'^\omega)$ and do not require any queries to $\mathbf{A}$. In Step 6, computing $\mathbf{WA}$ requires $O(nk/\epsilon)$ queries to $\mathbf{A}$, since $\mathbf{W}$ has $\widetilde{O}(k/\epsilon)$ rows. Note, this step contributes the leading term to the query complexity and it is crucial $\mathbf{W}$ does not have more rows. By Lemma 8.3.3, $\mathbf{N}$ can be computed in time $\widetilde{O}(nk/\epsilon + n(k/\epsilon)^{\omega-1} + k^3/\epsilon)$. Overall, Algorithm 9 requires $\widetilde{O}(nk/\epsilon + k'^2/\epsilon^4)$ queries to $\mathbf{A}$ and runs in time $\widetilde{O}(n(k/\epsilon)^{\omega-1} + nk'^{\omega-1} + (k'/\epsilon^2)^\omega)$. □

In light of Theorem 168, to obtain a low rank approximation for $\mathbf{A}$, it suffices to obtain an SF Projection. In particular, it suffices to obtain a matrix $\mathbf{Q} \in \mathbb{R}^{n \times k'}$, for $k' = \text{poly}(k, 1/\epsilon)$ such that $\mathbf{P} = \mathbf{QQ}^\top$ is an (ϵ, k)-SF projection, by querying $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$. One possible approach

to computing such a $\mathbf{Q}$ is to use the following result by Musco and Woodruff:

Theorem 169. *(Theorem 25, [MW17c].) Given a PSD matrix $\mathbf{A}$, integer k, $\epsilon > 0$, there exists an algorithm that reads $\widetilde{O}(nk/\epsilon^6 + nk^2/\epsilon^2)$ entries of $\mathbf{A}$ and with probability at least $99/100$, outputs $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n\times k}$ such that*

$$\|\mathbf{A} - \mathbf{MN}\|_2^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_2^2 + \frac{\epsilon}{k}\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Instantiating this theorem with $\epsilon = O(1)$ and $k = k/\epsilon$, we obtain a matrix $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n\times k/\epsilon}$ such that

$$\begin{aligned}\|\mathbf{A} - \mathbf{MN}^\top\|_2^2 &\leqslant O(1)\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_2^2 + O\left(\frac{\epsilon}{k}\right)\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2 \\ &\leqslant O\left(\frac{\epsilon}{k}\right)\|\mathbf{A} - \mathbf{A}_k\|_F^2\end{aligned}$$

where the last inequality follows from observing $\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_F^2 \leqslant \|\mathbf{A} - \mathbf{A}_k\|_F^2$ and

$$\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{j=k+1}^{n} \sigma_j^2(\mathbf{A}) \geqslant \left(\frac{k}{\epsilon} - k\right)\sigma_{k/\epsilon}^2 \geqslant \left(\frac{k}{\epsilon} - k\right)\|\mathbf{A} - \mathbf{A}_{k/\epsilon}\|_2^2$$

We can then compute an orthonormal basis for $\mathbf{M}$ and denote it by $\mathbf{Q}$. Here, we observe $\mathbf{P} = \mathbf{QQ}^\top$ is an (ϵ, k)-SF projection matrix. Further, the algorithm of Musco and Woodruff instantiated with the above parameters queries $\widetilde{O}(nk^2/\epsilon^2)$ entries in $\mathbf{A}$. As a corollary of Theorem 168, providing the rank-k/ϵ projection matrix $\mathbf{Q}$ as input to Algorithm 9, implies an algorithm for low rank approximation which queries $\widetilde{O}(nk^2/\epsilon^2)$ entries in $\mathbf{A}$. This already improves the ϵ-dependence in the query complexity of best known algorithm for PSD low-rank approximation, since the algorithm of Musco and Woodruff requires $\widetilde{O}(nk/\epsilon^{2.5})$ queries [MW17c]. Note, this algorithm has worse dependence on k. However, our goal is to obtain linear dependence on both k and $1/\epsilon$. Towards this end, we focus on obtaining an SF projection with fewer queries to $\mathbf{A}$.

8.3.2 Spectral Regression

In this subsection, we consider the *Spectral Regression* problem. This problem is a natural generalization of least-squares regression, when the response variable is a matrix. *Spectral Regression* arises in the context of Regularized Least Squares Classification, for instance [CLL+10]. Given matrices $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{X} \in \mathbb{R}^{d\times m}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$, the *Spectral Regression* problem considers the

following optimization problem:

$$\min_{\mathbf{X}} \|\mathbf{AX} - \mathbf{B}\|_2$$

We note that this is natural variant of multi-response regression where we minimize the difference between $\mathbf{AX}$ and $\mathbf{B}$ in spectral norm as opposed to the extensively studied and well-understood Frobenius norm. To the best of our knowledge the only relevant related work on *Spectral Regression* is by Clarkson and Woodruff [CW09] and Cohen et. al. [CNW15]. Both these works provide oblivious sketches to reduce the dimension of the problem, which unfortunately do not suffice for our application. Instead of *Spectral Regression* in its full generality we focus on the following special case:

Given an $n \times n$ PSD matrix $\mathbf{A}$, a rank parameter k, and an accuracy parameter ϵ, let $\mathbf{C}$ be a $n \times \sqrt{nk/\epsilon}$ matrix such that it is a *column PCP* for $\mathbf{A}$, satisfying the guarantees of Lemma 8.3.10, instantiated with $k = k/\epsilon$, and $\epsilon = O(1)$. Let $\mathbf{Z}^\top$ be a $k/\epsilon \times \sqrt{nk/\epsilon}$ matrix with orthonormal rows such that the corresponding projection matrix $\mathbf{ZZ}^\top$ is an $(O(1), k/\epsilon)$-SF Projection for $\mathbf{C}$. Then, we consider the following *Spectral Regression* problem:

$$\min_{\mathbf{W} \in \mathbb{R}^{n \times k/\epsilon}} \|\mathbf{C} - \mathbf{WZ}^\top\|_2 \tag{8.17}$$

Our main technical contribution here is to obtain a new algorithm to solve this optimization problem. We subsequently show how understanding this special case is crucial to obtaining *optimal* algorithms for low rank approximation of PSD matrices. The techniques we develop here may be of independent interest and find applications to other problems. Formally, we prove the following:

Theorem 170. *(Approximate Spectral Regression.) Let $\mathbf{C} \in \mathbb{R}^{n \times \sqrt{nk/\epsilon}}$ be a column PCP for $\mathbf{A}$ satisfying the guarantees of Lemma 8.3.10 instantiated with $k = k/\epsilon$ and $\epsilon = O(1)$. Let $\mathbf{Z} \in \mathbb{R}^{\sqrt{nk/\epsilon} \times k/\epsilon}$ be an orthonormal matrix such that $\mathbf{ZZ}^\top$ is an $(O(1), k/\epsilon)$-SF projection for $\mathbf{C}$. Then, Algorithm 10 queries $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ and with probability $99/100$ computes $\widehat{\mathbf{W}}$ such that*

$$\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 \leqslant \widetilde{O}(1)\left(\min_{\mathbf{W} \in \mathbb{R}^{n \times k/\epsilon}} \|\mathbf{C} - \mathbf{WZ}^\top\|_2^2 + \frac{\epsilon}{k}\|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2\right)$$

Further, the algorithm runs in time $\widetilde{O}(nk/\epsilon + (k/\epsilon)^\omega)$.

Algorithm 10 : Approximate Spectral Regression

Input: A PSD Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, integer k, and $\epsilon > 0$. $\mathbf{C} \in \mathbb{R}^{n \times \sqrt{nk/\epsilon}}$, a *column PCP* for $\mathbf{A}$ satisfying the guarantees of Lemma 8.3.10 instantiated with $k = k/\epsilon$ and $\epsilon = O(1)$. $\mathbf{Z} \in \mathbb{R}^{\sqrt{nk/\epsilon} \times k/\epsilon}$ be an orthonormal matrix such that $\mathbf{Z}\mathbf{Z}^\top$ is an $(O(1), k/\epsilon)$-SF projection for $\mathbf{C}$.

1. Consider the *Spectral Regression* problem:

$$\min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2$$

   Let $t = \sqrt{nk/\epsilon}$. For all $j \in [t]$, compute $\tau_j(\mathbf{Z}^\top) = \|\mathbf{Z}_{j,*}\|_2^2$. Let $q = \{q_1, q_2, \ldots, q_t\}$ be a distribution of columns of $\mathbf{C}$ such that for all $j \in [t]$, $q_j = \min(\tau_j(\mathbf{Z}^\top), 1)$.

2. Construct a sampling matrix $\mathbf{S}$ such that $\mathbf{C}\mathbf{S}$ selects each column $\mathbf{C}_{*,j}$ independently with probability q_j and scales it by $1/\sqrt{q_j}$. Similarly, construct $\mathbf{Z}^\top\mathbf{S}$. Consider the sketched optimization problem :

$$\min_{\mathbf{W}} \|\mathbf{C}\mathbf{S} - \mathbf{W}\mathbf{Z}^\top\mathbf{S}\|_2^2$$

3. Compute $(\mathbf{Z}^\top\mathbf{S})^\dagger = \mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}$. Let $\widehat{\mathbf{W}} = \mathbf{C}\mathbf{S}(\mathbf{Z}^\top\mathbf{S})^\dagger$ be the solution to the sketched optimization problem.

Output: $\widehat{\mathbf{W}} \in \mathbb{R}^{n \times k/\epsilon}$ such that $\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 \leqslant \widetilde{O}(1) \min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_F^2 + \widetilde{O}(\epsilon/k)\|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2$

We begin by characterizing the optimal solution and optimal cost for the *Spectral Regression* problem. We prove a structural result that shows the optimal solution for *Spectral Regression* is given by projecting $\mathbf{C}$ away from the span of $\mathbf{Z}^\top$. This matches the characterization of the optimal solution to regression under the Frobenius norm, given by the well-known normal equations. Recall, by definition of the Moore-Penrose pseudoinverse, this projection matrix is $(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top$.

Then, the optimal cost for Equation 8.17 is $\|\mathbf{C} - \mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top\|_2^2$ and is achieved by $\mathbf{W}^* = \mathbf{C}(\mathbf{Z}^\top)^\dagger$. Intuitively, we show that any feasible $\mathbf{W}$ must incur the above cost by analyzing $\|y^\top(\mathbf{C} - \mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top)\|_2^2$ for a fixed vector y. This enables us to exploit the geometry of Euclidean space and instantiate y as needed to relate it back to the spectral norm.

Lemma 8.3.14. *(Characterizing Opt for Spectral Regression.) Let* $\mathbf{C}$ *and* $\mathbf{Z}$ *be matrices as*

defined in Theorem 170. Let $\mathbf{W}^* = \mathbf{C}(\mathbf{Z}^\top)^\dagger = \mathbf{C}\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}$*, such that* $\mathbf{W}^*\mathbf{Z}^\top$ *is the projection of* $\mathbf{C}$ *on the colspan*$(\mathbf{Z})$*. Let* $\mathbf{C}^* = \mathbf{C} - \mathbf{W}^*\mathbf{Z}^\top$ *be the projection of* $\mathbf{C}$ *orthogonal to colspan*$(\mathbf{Z})$*. Then,*

$$\|\mathbf{C}^*\|_2^2 = \min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2$$

and the corresponding minimizer is $\mathbf{W}^*$*.*

Proof. Note, by definition $\|\mathbf{C} - \mathbf{W}^*\mathbf{Z}^\top\|_2^2 = \|\mathbf{C}^*\|_2^2$ and since $\mathbf{W}^*$ is feasible,

$$\min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2 \leqslant \|\mathbf{C}^*\|_2^2.$$

Therefore, it suffices to show any $\mathbf{W}$ must incur cost at least $\|\mathbf{C}^*\|_2^2$. By definition, we have

$$\mathbf{C} = \mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top + \mathbf{C}(\mathbf{I} - (\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top) = \mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top + \mathbf{C}^*$$

By definition of spectral norm, $\|\mathbf{C}-\mathbf{W}\mathbf{Z}^\top\|_2^2 \geqslant \|y^\top\mathbf{C}-y^\top\mathbf{W}\mathbf{Z}^\top\|_2^2$, for all y such that $\|y\|_2 = 1$. Next, for any unit vector $y \in \mathbb{R}^n$,

$$\begin{aligned}
\|y^\top\mathbf{C} - y^\top\mathbf{W}\mathbf{Z}^\top\|_2^2 &= \|y^\top(\mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top + \mathbf{C}^*) - y^\top\mathbf{W}\mathbf{Z}^\top\|_2^2 \\
&= \|y^\top\mathbf{C}^* - y^\top(\mathbf{W} - \mathbf{W}^*)\mathbf{Z}^\top\|_2^2 \\
&= \|y^\top\mathbf{C}^*\|_2^2 + \|y^\top(\mathbf{W} - \mathbf{W}^*)\mathbf{Z}^\top\|_2^2 + 2\langle y^\top\mathbf{C}^*, y^\top(\mathbf{W} - \mathbf{W}^*)\mathbf{Z}^\top\rangle
\end{aligned} \tag{8.18}$$

We observe that $\mathbf{W}\mathbf{Z}^\top = \mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top$ is the projection of $\mathbf{C}$ on the rowspan of $\mathbf{Z}^\top$ and $\mathbf{C}^*$ is the projection of $\mathbf{C}$ on the orthogonal complement of rowspan of $\mathbf{Z}^\top$. Therefore, $\langle\mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top, \mathbf{C}^*\rangle = 0$. Further, for any y, $y^\top(\mathbf{W} - \mathbf{W}^*)\mathbf{Z}^\top$ is in the row span of $\mathbf{Z}^\top$ and is thus perpendicular to $y^\top\mathbf{C}^*$. Plugging this back in to Equation 8.18, we have

$$\|y^\top\mathbf{C} - y^\top\mathbf{W}\mathbf{Z}^\top\|_2^2 = \|y^\top\mathbf{C}^*\|_2^2 + \|y^\top(\mathbf{W} - \mathbf{W}^*)\mathbf{Z}^\top\|_2^2 \geqslant \|y^\top\mathbf{C}^*\|_2^2 \tag{8.19}$$

where the inequality follows from non-negativity of norms. Since Equation 8.19 holds for all y, we can pick y such that $\|y^\top\mathbf{C}^*\|_2^2 = \|\mathbf{C}^*\|_2^2$. Therefore, $\|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2 \geqslant \|y^\top\mathbf{C} - y^\top\mathbf{W}\mathbf{Z}^\top\|_2^2 \geqslant \|\mathbf{C}^*\|_2^2$. This completes the proof. □

Next, we sketch the *Spectral Regression* problem from Equation 8.17 such that we approximately preserve the spectral norm cost of all $\mathbf{W} \in \mathbb{R}^{n\times k/\epsilon}$. A natural approach here would be to follow the Affine Embedding idea for Frobenius norm and hope a similar guarantee holds for

spectral norm as well. However, since $\mathbf{Z}$ could have rank as large as k/ϵ and we can no longer obtain a relative-error $(1+\epsilon)$-approximate Affine Embedding even for Frobenius norm without incurring a larger dependence on ϵ. Instead, we relax the notion of approximation for our sketch. We note that it suffices to construct a sketch $\mathbf{S}$ such that if

$$\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}} \|\mathbf{CS} - \mathbf{WZ}^\top\mathbf{S}\|_2^2$$

then

$$\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 \leqslant \widetilde{O}(1)\left(\min_{\mathbf{W}\in\mathbb{R}^{n\times k/\epsilon}} \|\mathbf{C} - \mathbf{WZ}^\top\|_2^2 + \frac{\epsilon}{k}\|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2\right)$$

as stated in Theorem 170. Note, here we only need to weakly preserve the cost of the optimal $\mathbf{W}$ for the sketched problem as opposed to preserving the cost of all matrices $\mathbf{W}$. At a high level, this comes down to analyzing the spectrum of $\|\mathbf{C}^*\mathbf{SS}^\top\|_2$ We begin with the definition of the Approximate Matrix Multiplication (AMM) guarantee and discuss its application in approximately minimizing *Spectral Regression*.

Definition 8.3.15. *((ϵ,k)-Spectral AMM.) Given matrices $\mathbf{A}\in\mathbb{R}^{n\times m}$ and $\mathbf{B}\in\mathbb{R}^{m\times d}$, a sketch $\mathbf{\Pi}\in\mathbb{R}^{m\times t}$ satisfies (ϵ,k)-Spectral AMM if with probability at least $1-\delta$,*

$$\|\mathbf{A\Pi\Pi}^\top\mathbf{B} - \mathbf{AB}\|_2 \leqslant \epsilon\sqrt{\left(\|\mathbf{A}\|_2^2 + \frac{\|\mathbf{A}\|_F^2}{k}\right)\cdot\left(\|\mathbf{B}\|_2^2 + \frac{\|\mathbf{B}\|_F^2}{k}\right)}$$

Approximate Matrix Multiplication was introduced by Drineas et al. [DKM06] with respect to the Frobenius norm, as opposed to the spectral norm above. Subsequent work by Cohen et al. [CNW15] studied the spectral norm bound and showed that any sketch $\mathbf{\Pi}$ that is an oblivious subspace embedding (i.e., satisfies Lemma 8.3.2 with $\mathbf{\Pi}$ being an oblivious sketch) implies an AMM guarantee, as long as $\mathbf{\Pi}$ has $\Theta(k+\log(1/\delta)/\epsilon^2)$ columns. The *Spectral AMM* property combined with an $O(1)$-Subspace Embedding suffice to approximately minimize the Spectral Regression problem :

Theorem 171. *(Theorem 3, [CNW15]) Let $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{\Pi}$ be as defined above. If $\mathbf{\Pi}$ is an $(\sqrt{\epsilon}, rank(\mathbf{A}))$-Spectral AMM for $\mathbf{U}_A$ and $(\mathbf{I}-\mathbf{P_A})\mathbf{B}$, and an $O(1)$-Subspace Embedding for $\mathbf{A}$, and $\widehat{\mathbf{X}} = \arg\min_{\mathbf{X}}\|\mathbf{\Pi AX} - \mathbf{\Pi B}\|_2^2$, then with probability $99/100$,*

$$\|\mathbf{A}\widehat{\mathbf{X}} - \mathbf{B}\|_2^2 \leqslant (1+\epsilon)\|\mathbf{P_AB} - \mathbf{B}\|_2^2 + \frac{\epsilon}{k}\|\mathbf{P_AB} - \mathbf{B}\|_F^2$$

where $\mathbf{U_A}$ is an orthonormal basis for $\mathbf{A}$ and $\mathbf{P_A}$ is the projection onto the span of $\mathbf{A}$.

However, all the constructions presented for the sketch in [CEM+15] are either oblivious sketches or require sampling proportional to both $\mathbf{A}$ and $\mathbf{B}$. Applying an oblivious sketch $\mathbf{S}$ in our problem requires computing $\mathbf{CS}$ which would query $\Omega(\mathsf{nnz}(\mathbf{C})) = \Omega(n^{1.5}\sqrt{k/\epsilon})$ entries in $\mathbf{A}$. Therefore, the main challenge here is to construct a sampling matrix $\mathbf{S}$ while reading $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ such that $\mathbf{S}$ is an $(\widetilde{O}(1), k/\epsilon)$-*Spectral AMM* and an $O(1)$-Subspace Embedding. We construct $\mathbf{S}$ by sampling $\widetilde{O}(k/\epsilon)$ columns of $\mathbf{Z}^\top$ proportional to the leverage scores of $\mathbf{Z}^\top$. While it is easy to show $\mathbf{S}$ is a Subspace Embedding, observe that our sampling probabilities are computed without reading $\mathbf{C}$.

Proof of Theorem 170. As a starting point, we observe that yet again, since $\mathbf{Z}^\top$ has orthonormal rows, the leverage scores are simply the ℓ_2^2 norms of the columns of $\mathbf{Z}^\top$. Therefore, one possible approach is to construct a *leverage score sampling* sketch $\mathbf{S}$ for $\mathbf{C}$, by sampling columns proportional to the leverage scores of $\mathbf{Z}^\top$. We note we can afford to sample at most $\widetilde{O}(k/\epsilon)$ columns, since our algorithm queries all entries in the resulting sketched matrix $\mathbf{CS}$.

Further, for reasons to be discussed later, it is crucial that we sample columns of $\mathbf{C}$ independently, as opposed to the standard way of sampling with replacement we have used thus far. The independent sampling process can be described as follows: for all $j \in [\sqrt{nk/\epsilon}]$, we sample $\mathbf{C}^*_{*,j}$ with probability $\min(\|\mathbf{Z}^\top_{*,j}\|_2^2, 1)$. We use the following lemma from [CMM17] to show that independently sampling columns satisfies some desirable properties.

Lemma 8.3.16. *(Lemma 21, [CMM17].) Given a matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, for all $j \in [m]$ let $\bar{\rho}_j^k(\mathbf{M}) = \Theta(\rho_j^k(\mathbf{M})$ be estimates of the rank-k column ridge-leverage scores of $\mathbf{M}$ and let $q_j = \min(\bar{\rho}_j^k(\mathbf{M})\log(k/\delta)/\epsilon^2, 1)$. Then, construct $\mathbf{MS}$ by selecting each column $\mathbf{M}_{*,j}$ with probability q_j and scale it by $1/\sqrt{q_j}$. Then, with probability at least $1-\delta$, $\mathbf{MS}$ has $\sum_{j\in[m]} \bar{\rho}_j^k \cdot \log(k/\delta)/\epsilon^2$ columns and*

$$(1-\epsilon)\mathbf{MSS}^\top\mathbf{M}^\top - \frac{\epsilon}{k}\|\mathbf{M}-\mathbf{M}_k\|_F^2\mathbf{I} \preceq \mathbf{MM}^\top \preceq (1+\epsilon)\mathbf{MSS}^\top\mathbf{M}^\top + \frac{\epsilon}{k}\|\mathbf{M}-\mathbf{M}_k\|_F^2\mathbf{I}$$

The above lemma independently samples columns proportional to the *ridge leverage scores*. In our setting, we can set the ridge parameter $\lambda = 0$, and sample according to the exact leverage scores of $\mathbf{Z}^\top$. Formally, let $q = \{q_1, q_2, \ldots, q_m\}$ be the corresponding distribution over columns of $\mathbf{Z}^\top$ such that $q_j = \min(\|\mathbf{Z}^\top\|_2^2 \log(k), 1)$. Since $\mathbf{Z}^\top$ has k/ϵ orthonormal rows, the leverage scores sum up to $\text{rank}(\mathbf{Z}^\top) \leqslant k/\epsilon$. We then use Lemma 8.3.16 by setting $\epsilon = 1/10, \delta = 0.01$ and

thus with probability at least $99/100$, $\mathbf{Z}^\top\mathbf{S}$ has $\sum_{j\in[\sqrt{nk/\epsilon}]}\tau_j(\mathbf{Z}^\top)\log(n) = \widetilde{O}(k/\epsilon)$ rows and

$$\frac{9}{10}\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z} \preceq \mathbf{Z}^\top\mathbf{Z} \preceq \frac{11}{10}\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z} \tag{8.20}$$

If the guarantee in Equation 8.20 holds for a sketch $\mathbf{S}$, we refer to $\mathbf{S}$ as satisfying an $O(1)$-Subspace Embedding property. Observe, this is equivalent to $\mathbf{S}$ preserving all singular values of $\mathbf{Z}^\top$ up to a constant.

We can now obtain a closed form solution for the *Spectral Regression* problem in the sketched space. By Lemma 8.3.14, the optimal solution to the optimization problem in Step 2 of Algorithm 10 is given by $\widehat{\mathbf{W}} = \mathbf{C}\mathbf{S}(\mathbf{Z}^\top\mathbf{S})^\dagger$. Since $\mathbf{S}$ satisfies the $O(1)$-Subspace Embedding property in Equation 8.20, it preserves the rank of $\mathbf{Z}^\top$. Therefore, $\mathbf{Z}^\top\mathbf{S}$ has full row rank and $(\mathbf{Z}^\top\mathbf{S})^\dagger = \mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}$ and thus $\widehat{\mathbf{W}} = \mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}$ is the optimal solution. Next, we bound the cost of $\widehat{\mathbf{W}}$ in the original problem. Let $\mathbf{P}_{\mathbf{Z}^\top} = \mathbf{Z}^\dagger\mathbf{Z}^\top$ be the orthogonal projection matrix onto $\mathbf{Z}^\top$. Using the fact that $\|\mathbf{M}\|_2^2 = \max_{\|y\|_2=1}\|y^\top\mathbf{M}\|_2^2$ and the Pythagorean Theorem for Euclidean space we have

$$\begin{aligned}\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 &= \|\mathbf{C} - \mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\|_2^2 \\ &= \max_{\|y\|_2^2=1}\|y^\top\mathbf{C}\mathbf{P}_{\mathbf{Z}^\top} - y^\top\mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{P}_{\mathbf{Z}^\top}\|_2^2 + \\ &\qquad \|y^\top\mathbf{C}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top}) - y^\top\mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top})\|_2^2\end{aligned} \tag{8.21}$$

Here, we observe $y^\top\mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$ is a vector in the row space of $\mathbf{Z}^\top$ and$(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top})$ is the projection on the orthogonal complement of rowspan$(\mathbf{Z}^\top)$, thus this evaluates to 0. Since $\mathbf{C}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top}) = \mathbf{C}^*$, we can upper bound $\|y^\top\mathbf{C}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^\top})\|_2$ by $\|\mathbf{C}^*\|_2^2$. Similarly, we can upper bound the first term by its spectral norm. Therefore, plugging this back into Equation 8.21,

$$\begin{aligned}\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 &\leqslant \|\mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top - \mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\|_2^2 + \|\mathbf{C}^*\|_2^2 \\ &= \|\left(\mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z} - \mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}\right)(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\|_2^2 + \|\mathbf{C}^*\|_2^2 \\ &\leqslant \|\mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z} - \mathbf{C}\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2\|(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\|_2^2 + \|\mathbf{C}^*\|_2^2\end{aligned} \tag{8.22}$$

where we use that $\mathbf{Z}^\top$ has orthonormal columns and the sub-multiplicativity of the spectral norm. From Equation 8.20, it follows that for all $i \in [k/\epsilon]$, $\sigma_i^2(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z}) = (1 \pm 0.1)^2\sigma_i^2(\mathbf{Z}^\top\mathbf{Z}) = (1 \pm 0.1)^2$. Therefore, $\|(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}\|_2^2 = 1/\sigma_{\min}^2(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z}) \leqslant 100/81$. Substituting this back

into Equation 8.22, we have

$$\begin{aligned}\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 &\leqslant O(1)\|(\mathbf{C}(\mathbf{Z}^\top)^\dagger\mathbf{Z}^\top - \mathbf{C})\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2 + \|\mathbf{C}^*\|_2^2 \\ &\leqslant O(1)\|\mathbf{C}^*\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2 + \|\mathbf{C}^*\|_2^2\end{aligned} \tag{8.23}$$

where the last inequality follows from the definition of $\mathbf{C}^*$. In order to bound the cost above, we focus on analyzing $\|\mathbf{C}^*\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2$. Since we want to compare $\|\mathbf{C}^*\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2$ to $\|\mathbf{C}^*\mathbf{Z}\|_2^2$, a natural way to proceed would be to interpret this term as an instance of *Approximate Matrix Product*. Therefore, we next show that the leverage score sampling matrix $\mathbf{S}$ satisfies the *Spectral AMM* property for $\mathbf{C}^*$ and $\mathbf{Z}^\top$. Here, we want to analyze how sampling columns of $\mathbf{C}^*$ proportional to the *leverage scores* of $\mathbf{Z}^\top$ affects the spectrum of $\mathbf{C}^*$. An important tool in this analysis is the following result by Rudelson and Vershynin on how the spectral norm of a matrix degrades when we sample a uniformly random subset of rows of a matrix:

Theorem 172. *(Theorem 1.8 in [RV07]) Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, let $\mathcal{Q}$ be a uniformly random subset of $[n]$ s.t. $\mathbb{E}[\mathcal{Q}] = q$. Let $\mathbf{A}_{|\mathcal{Q}}$ denote the submatrix restricted to the rows indexed by $\mathcal{Q}$. Then,*

$$\mathbb{E}\left[\left\|\mathbf{A}_{|\mathcal{Q}}\right\|_2\right] = O\left(\sqrt{\frac{q}{n}}\|\mathbf{A}\|_2 + \sqrt{\log(q)}\|\mathbf{A}\|_{(n/q)}\right)$$

where $\|\mathbf{A}\|_{(n/q)}$ is the average of the largest n/q ℓ_2-norms of columns of $\mathbf{A}$.

We extend the above statement to rectangular matrices:

Corollary 8.3.17 (Spectral Decay for Rectangular Matrices)**.** *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, s.t. for all $j, j' \in [m]$, $\|\mathbf{A}_{*,j}\|_2^2 = \Theta(\|\mathbf{A}_{*,j'}\|_2^2)$. Let $\mathcal{Q}$ be a uniformly random subset of $[n]$ s.t. $\mathbb{E}[\mathcal{Q}] = q$. Let $b = \lceil n/m \rceil$ and $\mathbf{A}_{|\mathcal{Q}}$ denote the submatrix restricted to the rows indexed by $\mathcal{Q}$. Then,*

$$\mathbb{E}\left[\left\|\mathbf{A}_{|\mathcal{Q}}\right\|_2\right] = O\left(\sqrt{\frac{q}{n}}\|\mathbf{A}\|_2 + \sqrt{\log(q)/b}\|\mathbf{A}\|_{(n/q)}\right)$$

Proof. First, consider the case when $m \geqslant n$. To see this, let $\mathsf{SVD}(\mathbf{A}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{U}\mathbf{\Sigma}$ is an $n \times n$ matrix. Now, $\|\mathbf{A}\|_2 = \|\mathbf{U}\mathbf{\Sigma}\|_2$ and applying Theorem 172 to $\mathbf{U}\mathbf{\Sigma}$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{A}_{|\mathcal{Q}}\right\|_2\right] = \mathbb{E}\left[\left\|(\mathbf{U}\mathbf{\Sigma})_{|\mathcal{Q}}\right\|_2\right] &= O\left(\sqrt{\frac{q}{n}}\|\mathbf{U}\mathbf{\Sigma}\|_2 + \sqrt{\log(q)}\|\mathbf{U}\mathbf{\Sigma}\|_{(n/q)}\right) \\ &= O\left(\sqrt{\frac{q}{n}}\|\mathbf{A}\|_2 + \sqrt{\log(q)}\|\mathbf{A}\|_{(n/q)}\right)\end{aligned} \tag{8.24}$$

where we repeatedly use that $\mathbf{V}^T$ has orthonormal rows. Here, we note that since the columns of $\mathbf{A}$ have the same squared norm up to a constant, $\|\mathbf{A}\|_{(n/q)} = \Theta(\|\mathbf{A}\|_{1\to 2})$, i.e. the max column

norm of $\mathbf{A}$.

Next, consider the case where $m < n$. Let $b = \lceil n/m \rceil$. In order to analyze the spectral norm of $\mathbf{A}_{\mathcal{Q}}$, we create b copies of $\mathbf{A}$ and concatenate them such that the resulting matrix $\mathbf{A}^*$ has more columns than rows. Applying Equation (8.24) to $\mathbf{A}^*$ and substituting the average with max, we have

$$\mathbb{E}\left[\left\|\mathbf{A}^*_{|\mathcal{Q}}\right\|_2\right] = O\left(\sqrt{\frac{q}{n}}\|\mathbf{A}^*\|_2 + \sqrt{\log(q)}\|\mathbf{A}^*\|_{1\to 2}\right) \tag{8.25}$$

Observe, $\mathbf{A}^*_{|\mathcal{Q}}$ selects uniformly random rows of $\mathbf{A}^*$ and $\left\|\mathbf{A}^*_{|\mathcal{Q}}\right\|_2 = \max_{\|x\|_2=1} \|x^\top \mathbf{A}^*_{|\mathcal{Q}}\|_2$ and for any vector x, $\|x^\top \mathbf{A}^*_{|\mathcal{Q}}\|_2 = \sqrt{b}\|x^\top \mathbf{A}_{|\mathcal{Q}}\|_2$. Therefore, $\mathbb{E}\left[\left\|\mathbf{A}^*_{|\mathcal{Q}}\right\|_2\right] = \sqrt{b} \cdot \mathbb{E}\left[\left\|\mathbf{A}_{|\mathcal{Q}}\right\|_2\right]$ and $\|\mathbf{A}^*\|_2 = \sqrt{b} \cdot \|\mathbf{A}^*\|_2$. Finally, it is easy to see that since the columns of $\mathbf{A}^*$ are copies of columns of $\mathbf{A}$, the max column norm does not change. Therefore, (8.24) to $\mathbf{A}^*$, we have

$$\mathbb{E}\left[\left\|\mathbf{A}_{|\mathcal{Q}}\right\|_2\right] = O\left(\sqrt{\frac{q}{n}}\|\mathbf{A}\|_2 + \sqrt{\log(q)/b}\|\mathbf{A}\|_{1\to 2}\right) = O\left(\sqrt{\frac{q}{n}}\|\mathbf{A}^*\|_2 + \sqrt{\log(q)}\|\mathbf{A}^*\|_{(n/q)}\right) \tag{8.26}$$

and the claim follows. □

Intuitively, there are two technical challenges in applying Corollary 8.3.17. First, a *leverage score sampling* matrix need not sample columns uniformly at random, since we have no control over the column norms of $\mathbf{Z}^\top$. Second, the $\|\cdot\|_{(n/q)}$ norm only shrinks when all columns of $\mathbf{A}$ have roughly the same squared norm. We overcome these challenges by partitioning the matrix, first according to row norms, such that each partition does indeed have the same row norm, up to a factor of 2. Next, we further partition each such matrix according to the sampling probabilities, such that within each partition, the sampling process is *close* to uniform sampling. Formally,

Lemma 8.3.18. *(Weak Spectral Approximate Matrix Product.) Let $\mathbf{Z}, \mathbf{C}^*$ and $\mathbf{S}$ be as defined in Lemma 8.3.14. Then, with probability at least $99/100$, $\mathbf{S}$ satisfies $(\widetilde{O}(1), k/\epsilon)$-Spectral AMM, i.e.,*

$$\|\mathbf{C}^*\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2 \leqslant \widetilde{O}(1)\left(\frac{\epsilon}{k}\|\mathbf{C}^*\|_F^2 + \|\mathbf{C}^*\|_2^2\right)$$

Proof. By sub-multiplicativity of the spectral norm and $\mathbf{S}$ being an $O(1)$-subspace embedding for $\mathbf{Z}^\top$, we have

$$\begin{aligned}\|\mathbf{C}^*\mathbf{S}\mathbf{S}^\top\mathbf{Z}\|_2^2 &\leqslant \|\mathbf{C}^*\mathbf{S}\|_2^2 \cdot \|\mathbf{S}^\top\mathbf{Z}\|_2^2 \\ &\leqslant O(1)\|\mathbf{C}^*\mathbf{S}\|_2^2\end{aligned} \tag{8.27}$$

where the second inequality follows from $\mathbf{Z}^\top$ having orthonormal rows.

We begin by observing that Corollary 8.3.17 requires the squared row norms of $\mathbf{C}^*$ to be roughly the same, which need not be the case in general. Note, here the sampling matrix subsamples columns of $\mathbf{C}^*$, as opposed to rows in Corollary 8.3.17. Thus, we partition the rows of $\mathbf{C}^*$ into $O(\log(n))$ blocks such that either the squared column norms are the same up to a factor of 2 or they are at most $\|\mathbf{C}^*\|_F^2/\text{poly}(n)$. Formally, for all $\ell \in [c\log(n)]$, let

$$\mathcal{B}_\ell = \left\{ i \in [n] : \frac{\|\mathbf{C}^*\|_F^2}{2^{\ell+1}} \leqslant \|\mathbf{C}^*_{i,*}\|_2^2 \leqslant \frac{\|\mathbf{C}^*\|_F^2}{2^\ell} \right\}$$

represent the blocks for rows with large squared norm. Let $\mathcal{B}_r = [n] \setminus \cup_{\ell \in [\log(n)]} \mathcal{B}_\ell$ be the remaining rows, which have norm at most $\|\mathbf{C}^*\|_F^2/\text{poly}(n)$. Since the set of indices in the blocks form a partition of the rows of $\mathbf{C}^*$, we can write $\|\mathbf{C}^*\|_F^2 = \sum_{\ell \in [\log(n)]} \|\mathbf{C}^*_{\mathcal{B}_\ell}\|_F^2 + \|\mathbf{C}^*_{\mathcal{B}_r}\|_F^2$. Similarly, we can bound the spectral norm as follows:

$$\begin{aligned} \|\mathbf{C}^*\mathbf{S}\|_2^2 = \max_{\|y\|_2^2=1} \|\mathbf{C}^*\mathbf{S}y\|_2^2 &\leqslant O\left(\sum_{\ell \in [\log(n)]} \|\mathbf{C}^*_{\mathcal{B}_\ell}\mathbf{S}y\|_2^2 + \|\mathbf{C}^*_{\mathcal{B}_r}\mathbf{S}y\|_2^2 \right) \\ &\leqslant O\left(\sum_{\ell \in [\log(n)]} \|\mathbf{C}^*_{\mathcal{B}_\ell}\mathbf{S}\|_2^2 + \|\mathbf{C}^*_{\mathcal{B}_r}\mathbf{S}\|_2^2 \right) \end{aligned} \tag{8.28}$$

We now handle the two separately. Since $\mathbf{S}$ is an unbiased estimator of the squared Frobenius norm of $\mathbf{Z}^\top$, it is an unbiased estimator of the squared Frobenius norm of $\mathbf{C}^*$. Therefore, with probability at least $99/100$,

$$\|\mathbf{C}^*_{|\mathcal{B}_r}\mathbf{S}\|_2^2 \leqslant \|\mathbf{C}^*_{|\mathcal{B}_r}\mathbf{S}\|_F^2 = O(\|\mathbf{C}^*_{|\mathcal{B}_r}\|_F^2) \leqslant \frac{\|\mathbf{C}^*\|_F^2}{\text{poly}(n)} << \frac{\epsilon}{k}\|\mathbf{C}^*\|_F^2 \tag{8.29}$$

For the remaining terms, we cannot use this näive analysis as this would only leave us with an upper bound of $\|\mathbf{C}^*\|_F^2$, which is too large.

If instead of a leverage score sampling matrix, $\mathbf{S}$ were a uniform sampling sketch that samples k/ϵ columns of $\mathbf{C}^*$ in expectation, we could apply Corollary 8.3.17 for each ℓ, with $q = k/\epsilon$ and $n = \sqrt{nk/\epsilon}$ and $b = \lceil \sqrt{nk}/(\sqrt{\epsilon}|\mathcal{B}_\ell|) \rceil$, to obtain

$$\begin{aligned} \mathbb{E}\left[\left\| \mathbf{C}^*_{|\mathcal{B}_\ell}\mathbf{S} \right\|_2^2 \right] = \sqrt{\frac{n\epsilon}{k}}\, \mathbb{E}\left[\left\| (\mathbf{C}^*_{|\mathcal{B}_\ell})_Q \right\|_2^2 \right] &\leqslant O\left(\left\| \mathbf{C}^*_{|\mathcal{B}_\ell} \right\|_2^2 + \frac{\epsilon \log(k/\epsilon)|\mathcal{B}_\ell|}{k} \left\| (\mathbf{C}^*)^\top_{|\mathcal{B}_\ell} \right\|^2_{\sqrt{\epsilon n/k}} \right) \\ &\leqslant O\left(\left\| \mathbf{C}^*_{|\mathcal{B}_\ell} \right\|_2^2 + \frac{\epsilon \log(k/\epsilon)}{k} \left\| \mathbf{C}^*_{|\mathcal{B}_\ell} \right\|_F^2 \right) \end{aligned} \tag{8.30}$$

where Q is the subset of columns selected by $\mathbf{S}$ and the second inequality follows from observing that the all the row norms of $(\mathbf{C}^*)_{|\mathcal{B}_\ell}$ are within a factor of 2 of each other and thus the max squared row norm times the size of the set is the squared Frobenius norm.

Using Equations 8.29 and 8.30 to upper bound the two terms in Equation 8.28 suffices to finish the proof. Unfortunately, a similar analysis does not immediately go through when we replace a uniform sampling matrix with a leverage score sketch. Instead, we partition the sketch $\mathbf{S}$ into buckets such that each bucket corresponds to rows in $\mathbf{S}$ that scale columns of $\mathbf{C}^*$ within a factor of 2. For notational convenience, let $m = \sqrt{nk/\epsilon}$ and $t = k/\epsilon$. Recall, we construct $\mathbf{S}$ by sampling the j-th column of $\mathbf{Z}^\top$ independently with probability $q_j = \min(\|\mathbf{Z}_{j,*}\|_2^2 \log(k), 1)$ and scale this column by $1/\sqrt{q_j}$. We group the scaling factors into buckets. Note, if for some j, $q_j < 1/n^3$, we can ignore the corresponding column.

Let ζ_j be the indicator for a column of $\mathbf{Z}^\top$ to be sampled by $\mathbf{S}$. Then, $\mathbf{Pr}[\zeta_j = 1] = q_j = \min(\|\mathbf{Z}_{*,j}^\top\|_2^2 \log(k), 1)$. Since $q_j \leqslant 1/n^3$, we can union bound over at most m such events and conclude with probability at least $1 - 1/n^2$, for all $j \in m$, no column $\mathbf{Z}_{*,j}^\top$ is sampled such that $q_j \leqslant 1/n^3$. Further, since $q_j \leqslant 1$, $1/\sqrt{q_j} \in [1, n^{1.5}]$. Therefore, it suffices to bucket values in the range $[1, n^{1.5}]$. For all $h \in [c\log(n)]$, let $\mathcal{S}$ denote the set of column indices from $\mathbf{Z}^\top$ that were sampled by the sketch $\mathbf{S}$. Then,

$$\mathcal{T}_h = \left\{ j \in \mathcal{S} : 2^h \leqslant \frac{1}{\sqrt{q_j}} \leqslant 2^{h+1} \right\}$$

Let $\mathbf{S}_{\mathcal{T}_h}$ be the subset of rows of $\mathbf{S}$ which are indexed by the set $\mathcal{T}_h$. Since t is fixed and the scaling factors in $\mathcal{T}_h$ differ by at most a factor of 2, the corresponding sampling probabilities in $\mathbf{D}$ differ by at most $\sqrt{2}$, which is still not uniform. To fix this, we change the sampling process and independently sample each column indexed by $j \in \mathcal{T}_h$ with probability 2^{h+1}, while still scaling it by $1/\sqrt{tq_j}$. Let this new distribution be denoted by q'. Under the new sampling process, we now sample rows independently and therefore, we are at least as likely to see all the rows sampled by $\mathbf{S}_{\mathcal{T}_h}$ in our new sampling process. Therefore, it now holds that

$$\mathop{\mathbb{E}}_{q}\left[\|\mathbf{C}^*_{|\mathcal{B}_\ell}\mathbf{S}_{\mathcal{T}_h}\|_2^2\right] \leqslant \mathop{\mathbb{E}}_{q'}\left[\|\mathbf{C}^*_{|\mathcal{B}_\ell}\mathbf{S}_{\mathcal{T}_h}\|_2^2\right]$$

Further, in the new sampling process, each row restricted to the set $\mathcal{T}_h$ is uniformly sampled with

probability $1/2^{(h+1)/2}$ and thus we can apply Corollary 8.3.17 to $\mathbf{C}^*_{|\mathcal{B}_\ell}\mathbf{S}_{\mathcal{T}_h}$.

$$\begin{aligned}\mathop{\mathbb{E}}_{q'}\left[\|\mathbf{C}^*_{|\mathcal{B}_\ell}\mathbf{S}_{\mathcal{T}_h}\|_2^2\right] &\leqslant O\left(\|\mathbf{C}^*_{|\mathcal{B}_\ell}\|_2^2 + \frac{\epsilon\log(k/\epsilon)|\mathcal{B}_\ell|}{k}\left\|(\mathbf{C}^*_{|\mathcal{B}_\ell})^\top\right\|_{(\sqrt{\epsilon n/k})}\right)\\ &\leqslant O\left(\|\mathbf{C}^*_{|\mathcal{B}_\ell}\|_2^2 + \frac{\epsilon\log(k/\epsilon)}{k}\left\|\mathbf{C}^*_{|\mathcal{B}_\ell}\right\|_F^2\right)\end{aligned} \tag{8.31}$$

where the second inequality follows from squared row norms in $\mathbf{C}^*_{|\mathcal{B}_\ell}$ being equal up to a factor of 2. Therefore, with probability at least $1 - 1/c'\log(n)$,

$$\left\|\mathbf{C}^*_{|\mathcal{B}_\ell}\mathbf{S}_{\mathcal{T}_h}\right\|_2^2 \leqslant \widetilde{O}\left(\|\mathbf{C}^*_{|\mathcal{B}_\ell}\|_2^2 + \frac{\epsilon\log(k/\epsilon)}{k}\|\mathbf{C}^*_{|\mathcal{B}_\ell}\|_F^2\right) \tag{8.32}$$

Let η_h be the event that the above bound holds. Then, union bounding over all $c\log(n)$ such events, with probability at least $99/100$, simultaneously for all h,

$$\begin{aligned}\|\mathbf{C}^*_{|\mathcal{B}_\ell}\mathbf{S}\|_2^2 &\leqslant O\left(\sum_{h\in[c\log(n)]}\|\mathbf{C}^*_{|\mathcal{B}_\ell}\mathbf{S}_{\mathcal{T}_h}\|_2^2\right)\\ &\leqslant \widetilde{O}\left(\|\mathbf{C}^*_{|\mathcal{B}_\ell}\|_2^2 + \frac{\epsilon}{k}\|\mathbf{C}^*_{|\mathcal{B}_\ell}\|_F^2\right)\end{aligned} \tag{8.33}$$

which follows from Equation 8.32. Substituting this back into Equation 8.28,

$$\begin{aligned}\|\mathbf{C}^*\mathbf{S}\|_2^2 &\leqslant O\left(\sum_{\ell\in[\log(n)]}\|\mathbf{C}^*_{|\mathcal{B}_\ell}\mathbf{S}\|_2^2 + \|\mathbf{C}^*_{|\mathcal{B}_r}\mathbf{S}\|_2^2\right)\\ &\leqslant \widetilde{O}\left(\|\mathbf{C}^*\|_2^2 + \frac{\epsilon}{k}\|\mathbf{C}^*\|_F^2\right)\end{aligned} \tag{8.34}$$

where the second inequality follows from Equation 8.33 and observing that $\|\mathbf{C}^*_{\mathcal{B}_\ell}\|_2^2 \leqslant \|\mathbf{C}^*\|_2^2$ and $\sum_\ell \|\mathbf{C}^*_{\mathcal{B}_\ell}\|_F^2 = \|\mathbf{C}^*\|_F^2$, which completes the proof. □

Combining the above lemma with 8.21, and observing that $\|\mathbf{C}^*\|_F^2 \leqslant \|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2$, we can bound the cost of $\widehat{\mathbf{W}}$

$$\|\mathbf{C} - \widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 \leqslant \widetilde{O}\left(\min_{\mathbf{W}}\|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2 + \frac{\epsilon}{k}\|\mathbf{C} - \mathbf{C}_{k/\epsilon}\|_F^2\right)$$

which completes the correctness proof of Theorem 170. Next, we analyze the running time. In Step 1 of Algorithm 10, we compute a distribution over the columns of $\mathbf{Z}^\top$, which does not require reading any entries in $\mathbf{A}$ and takes time $\sqrt{nk/\epsilon}\cdot k/\epsilon = \sqrt{n}(k/\epsilon)^{1.5}$. Step 2 requires computing $\mathbf{CS}$ and $\mathbf{Z}^\top\mathbf{S}$. Note since $\mathbf{S}$ samples $\widetilde{O}(k/\epsilon)$ columns in $\mathbf{C}$, we have to query $n\cdot$

$\widetilde{O}(k/\epsilon)$ entries in $\mathbf{A}$ to explicitly compute $\mathbf{CS}$ and can be computed in as much time. Since $\mathbf{Z}^\top$ has fewer rows the running time is dominated by computing $\mathbf{CS}$. For Step 3, we compute $(\mathbf{Z}^\top\mathbf{S}\mathbf{S}^\top\mathbf{Z})^{-1}$, which requires no queries to $\mathbf{A}$ and runs in time $\widetilde{O}((k/\epsilon)^\omega)$ and thus $(\mathbf{Z}^\top\mathbf{S})^\dagger$ can be computed in the same time. Therefore, the total query complexity of Algorithm 10 is $\widetilde{O}(nk/\epsilon)$ and the running time is $\widetilde{O}(nk/\epsilon + (k/\epsilon)^\omega)$, which concludes the proof.

8.3.3 Sample-Optimal Algorithm

In this subsection, we describe our main algorithm for PSD Low-Rank Approximation. Given a PSD matrix $\mathbf{A}$, our algorithm queries $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ and runs in time $\widetilde{O}(n(k/\epsilon)^{\omega-1} + (k/\epsilon^3)^\omega)$. This resolves an open question on the ϵ-dependence of the query complexity and matches the lower bound of $\Omega(nk/\epsilon)$ up to polylog factors from [MW17c]. At a high level, our algorithm consists of two stages: first, we use the existing machinery developed by Musco and Woodruff to obtain weak PCPs by setting ϵ to be a constant. By observing that their algorithms have linear dependence on the rank, we can afford to rank-(k/ϵ) PCPs instead. This enables us to find a structured subspace that contains a spectral low-rank approximation for the PCP.

Since our PCPs are accurate only up to $O(1)$-error, we cannot directly extract a $(1+\epsilon)$ relative error approximation for $\mathbf{A}$. However, we show that the PCPs have enough structure to obtain a structured subspace that spans a $(1+\epsilon)$-approximate solution for $\mathbf{A}$. A key ingredient to recover this structured subspace is an efficient algorithm for *Spectral Regression*.

Following the approach of Musco and Woodruff we use the ridge leverage scores of $\mathbf{A}^{1/2}$ to compute $\mathbf{C}$, a column PCP for $\mathbf{A}$ and $\mathbf{R}$ a row PCP for $\mathbf{C}$, with a minor tweak: we instantiate their theorems (Lemmas 8.3.10 and 8.3.11) with $k = k/\epsilon$ and $\epsilon = O(1)$. While the precise guarantees satisfied by our PCPs are weaker than the PCPs used by Musco and Woodruff, the dimensions of our PCPs are smaller.

Algorithm 11 : Sample Optimal PSD Low-Rank Approximation

Input: A PSD Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, integer k, and $\epsilon > 0$.

1. Let $t = c\sqrt{\frac{nk}{\epsilon}}\log(n)$, for some constant c and let $k' = \widetilde{O}(k/\epsilon)$. For all $j \in [n]$, let $\bar{\rho}_j^{k'}(\mathbf{A}^{1/2})$ be the approximate column ridge-leverage scores that satisfy Lemma 8.3.8. Let $q = \{q_1, q_2 \dots q_n\}$ denote a distribution over columns of $\mathbf{A}$ such that $q_j = \rho_j^{k'}(\mathbf{A}^{1/2}) / \sum_j \rho_j^{k'}(\mathbf{A}^{1/2})$.
2. Construct a *column PCP* for $\mathbf{A}$ by sampling t columns of $\mathbf{A}$ such that each column is set to $\frac{\mathbf{A}_{*,j}}{\sqrt{tq_j}}$ with probability q_j, for all $j \in [n]$. Let $\mathbf{C}$ be the resulting $n \times t$ matrix that satisfies the guarantee of Lemma 8.3.10 instantiated with $k = k'$ and $\epsilon = O(1)$.
3. Construct a *row PCP* for $\mathbf{C}$ by sampling t rows of $\mathbf{C}$ such that each row is set to $\frac{\mathbf{C}_{i,*}}{\sqrt{tq_i}}$ with probability q_i, for all $i \in [n]$. Let $\mathbf{R}$ be the resulting $t \times t$ matrix that satisfies the guarantee of Lemma 8.3.11 instantiated with $k = k/\epsilon$ and $\epsilon = O(1)$.
4. Run the *input-sparsity* algorithm from Lemma 9.4.1 to compute a rank-k/ϵ matrix $\mathbf{Z}$ with orthonormal columns such that $\|\mathbf{R} - \mathbf{R}\mathbf{Z}\mathbf{Z}^\top\|_2^2 \leqslant O\left(\frac{\epsilon}{k}\right)\|\mathbf{R} - \mathbf{R}_{k/\epsilon}\|_F^2$.
5. Run Algorithm 10 with parameters k, ϵ on the *Spectral Regression* problem

$$\min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2$$

   Let $\widehat{\mathbf{W}}$ be the output of Algorithm 10. Compute an orthonormal basis $\mathbf{Q}$ for $\mathbf{W}$. Note, $\mathbf{Q}\mathbf{Q}^\top$ is an $(O(1), k/\epsilon)$-SF projection for $\mathbf{A}$.
6. Run Algorithm 9 with input $\mathbf{A}$, $\mathbf{Q}$, k and ϵ to approximately minimize $\|\mathbf{A} - \mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2$ over rank k matrices $\mathbf{X}$. Let $\mathbf{M}, \mathbf{N}$ be the output of Algorithm 9.

Output: $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n \times k}$ such that $\|\mathbf{A} - \mathbf{M}\mathbf{N}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$

In particular, we obtain a row PCP $\mathbf{R}$, which is a $\sqrt{nk/\epsilon} \times \sqrt{nk/\epsilon}$ matrix (ignoring poly-logarithmic factors) and we can afford to read all of it. The input-sparsity time algorithm from Lemma 9.4.1 queries $\mathsf{nnz}(\mathbf{R}) = \widetilde{O}(nk/\epsilon)$ entries to obtain a rank-(k/ϵ) matrix $\mathbf{Z}$ with orthonormal columns such that

$$\|\mathbf{R} - \mathbf{R}\mathbf{Z}\mathbf{Z}^\top\|_2^2 \leqslant \frac{\epsilon}{k}\|\mathbf{R} - \mathbf{R}_k\|_F^2 \tag{8.35}$$

Since $\mathbf{R}$ is a *Spectral-Frobenius PCP* for $\mathbf{C}$, $\mathbf{Z}\mathbf{Z}^\top$ satisfies $\|\mathbf{C} - \mathbf{C}\mathbf{Z}\mathbf{Z}^\top\|_2^2 \leqslant \frac{\epsilon}{k}\|\mathbf{C} - \mathbf{C}_k\|_F^2$. Since $\mathbf{C}$ is a *Spectral-Frobenius PCP* for $\mathbf{A}$, it suffices to obtain a projection for the column space of $\mathbf{C}$ that also satisfies the above guarantee. Therefore, we solve the following *Spectral Regression*

problem: $\min_{\mathbf{W}} \|\mathbf{C} - \mathbf{W}\mathbf{Z}^\top\|_2^2$. Recall, we can approximately optimize this using Algorithm 10. Let $\widehat{\mathbf{W}}$ be the resulting solution. We can then compute an orthonormal basis for $\widehat{\mathbf{W}}$ (denoted by $\mathbf{Q}$) and show that $\mathbf{Q}\mathbf{Q}^\top$ is an $(O(1), k/\epsilon)$-SF projection for $\mathbf{A}$. Then, we can obtain a low rank approximation for $\mathbf{A}$ by simply running Algorithm 9.

Proof of Theorem 166. Let $k' = \widetilde{O}(k/\epsilon)$. It follows from Lemma 8.3.8 that we can compute the rank-k' ridge leverage scores of $\mathbf{A}^{1/2}$, up to a constant factor using the algorithm of Musco and Musco [MM17]. By Lemma 8.3.9, the ridge leverage scores of $\mathbf{A}^{1/2}$ are a $\sqrt{\epsilon n/k'}$-approximation to the ridge leverage scores of $\mathbf{A}$. Let q be a distribution over rows and columns of $\mathbf{A}$ as defined in Algorithm 11. Since we sample $t = O(\sqrt{nk/\epsilon}\log(n))$ columns of $\mathbf{A}$ proportional to q, instantiating Lemma 8.3.11 with $k = k'$ and $\epsilon = 0.1$, we obtain a mixed Spectral-Frobenius column PCP $\mathbf{C}$ such that with probability at least $1 - c_1$, for all rank-k' projections $\mathbf{X}$,

$$\frac{9}{10}\|\mathbf{A}-\mathbf{X}\mathbf{A}\|_2^2 - \frac{1}{10k'}\|\mathbf{A}-\mathbf{A}_{k'}\|_F^2 \leqslant \|\mathbf{C}-\mathbf{X}\mathbf{C}\|_2^2 \leqslant \frac{11}{10}\|\mathbf{A}-\mathbf{X}\mathbf{A}\|_2^2 + \frac{1}{10k'}\|\mathbf{A}-\mathbf{A}_{k'}\|_F^2 \quad (8.36)$$

Let ζ_1 be the indicator for $\mathbf{C}$ satisfying the above guarantee. Similarly, sampling t rows of $\mathbf{C}$ proportional to q, results in a mixed Spectral-Frobenius row PCP for $\mathbf{R}$ such that with probability at least $1 - c_2$, for all rank-k' projection matrices $\mathbf{X}$,

$$\frac{9}{10}\|\mathbf{C}-\mathbf{C}\mathbf{X}\|_2^2 - \frac{1}{10k'}\|\mathbf{C}-\mathbf{C}_{k'}\|_F^2 \leqslant \|\mathbf{R}-\mathbf{R}\mathbf{X}\|_2^2 \leqslant \frac{11}{10}\|\mathbf{C}-\mathbf{C}\mathbf{X}\|_2^2 + \frac{1}{10k'}\|\mathbf{C}-\mathbf{C}_{k'}\|_F^2 \quad (8.37)$$

Further, it is well-known that with the same probability $\|\mathbf{R} - \mathbf{R}_{k'}\|_F^2 = \|\mathbf{C} - \mathbf{C}_{k'}\|_F^2$. Let ζ_2 be the event that $\mathbf{R}$ satisfies the above guarantee. Next, we compute a Spectral Low-Rank Approximation for $\mathbf{R}$, using the algorithm from Lemma 9.4.1, with $k = k'$ and $\epsilon = 0.1$. As a result, with probability at least $1 - c_3$, we obtain a rank-k' matrix $\mathbf{Z} \in \mathbb{R}^{t \times k}$, such that $\mathbf{Z}\mathbf{Z}^\top$ is a $(0.1, k')$-SF projection for $\mathbf{R}$, i.e.,

$$\|\mathbf{R} - \mathbf{R}\mathbf{Z}\mathbf{Z}^\top\|_2^2 \leqslant \frac{1}{10k'}\|\mathbf{R} - \mathbf{R}_{k'}\|_F^2 \quad (8.38)$$

Let ζ_3 be the event that $\mathbf{Z}$ satisfies the above guarantee. Union bounding over $\zeta_1, \zeta_2, \zeta_3$, we know that all of them hold with probability at least $1 - (c_1 + c_2 + c_3)$. Since $\mathbf{R}$ is a Spectral-Frobenius

row PCP for $\mathbf{C}$ and $\mathbf{Z}\mathbf{Z}^\top$ is a rank-k' projection matrix, it follows from Equation 8.37

$$\begin{aligned}\|\mathbf{C}-\mathbf{C}\mathbf{Z}\mathbf{Z}^\top\|_2^2 &\leqslant \frac{10}{9}\|\mathbf{R}-\mathbf{R}\mathbf{Z}\mathbf{Z}^\top\|_2^2+\frac{1}{9k'}\|\mathbf{R}-\mathbf{R}_{k'}\|_F^2\\ &\leqslant \frac{1}{10k'}\|\mathbf{R}-\mathbf{R}_k\|_F^2+\frac{1}{9k'}\|\mathbf{R}-\mathbf{R}_{k'}\|_F^2\\ &\leqslant \tilde{O}\left(\frac{\epsilon}{k}\right)\|\mathbf{C}-\mathbf{C}_{k'}\|_F^2\end{aligned} \tag{8.39}$$

where the second inequality follows from Equation 8.35 and the third follows from the fact that PCPs preserve Frobenius norm up to a constant factor. While conditioning on ζ_3, it follows from Equation 8.39 that $\mathbf{Z}\mathbf{Z}^\top$ is an $(\tilde{O}(1), k/\epsilon)$-SF projection for $\mathbf{C}$, our goal is to compute an SF projection for $\mathbf{A}$. Since $\mathbf{Z}\mathbf{Z}^\top$ is a $t\times t$ matrix, it does not even match the dimensions of $\mathbf{A}$. Therefore, we set up the following *Spectral Regression* problem:

$$\min_{\mathbf{W}\in\mathbb{R}^{n\times k'}}\|\mathbf{C}-\mathbf{W}\mathbf{Z}^\top\|_2^2 \tag{8.40}$$

Let $\widehat{\mathbf{W}}$ be the approximate minimizer of the above problem obtained by running Algorithm 10. Then, it follows from Theorem 170 that with probability at least $99/100$,

$$\begin{aligned}\|\mathbf{C}-\widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2 &\leqslant \tilde{O}(1)\left(\min_{\mathbf{W}}\|\mathbf{C}-\mathbf{W}\mathbf{Z}^\top\|_2^2+\frac{\epsilon}{k}\|\mathbf{C}-\mathbf{C}_{k'}\|_F^2\right)\\ &\leqslant \tilde{O}(1)\left(\|\mathbf{C}-\mathbf{C}_{k'}\|_2^2+\frac{\epsilon}{k}\|\mathbf{C}-\mathbf{C}_{k'}\|_F^2\right)\\ &\leqslant \tilde{O}(1)\left(\frac{\epsilon}{k}\|\mathbf{C}-\mathbf{C}_{k'}\|_F^2\right)\end{aligned} \tag{8.41}$$

where the second inequality follows from $\|\mathbf{C}-\mathbf{C}\mathbf{Z}\mathbf{Z}^\top\|_2^2\leqslant\tilde{O}\left(\frac{\epsilon}{k}\right)\|\mathbf{C}-\mathbf{C}_{k'}\|_F^2$ (by definition of an SF projection) and observing that $\mathbf{W}=\mathbf{C}\mathbf{Z}^\top$ is a feasible solution to Equation 8.40. Let ζ_4 be the event that Equation 8.41 holds. Next, let $\mathbf{Q}$ be an orthonormal basis for $\mathbf{W}$. We observe that $\mathbf{Q}\mathbf{Q}^\top\mathbf{C}$ is the orthogonal projection of $\mathbf{C}$ onto the subspace spanned by $\mathbf{Q}$ and the matrix $\widehat{\mathbf{W}}\mathbf{Z}^\top$ also lies in the subspace. Therefore, by the Pythagorean Theorem, for any fixed unit vector y,

$$\|\mathbf{C}y-\mathbf{Q}\mathbf{Q}^\top\mathbf{C}y\|_2^2\leqslant\|\mathbf{C}y-\widehat{\mathbf{W}}\mathbf{Z}^\top y\|_2^2\leqslant\|\mathbf{C}-\widehat{\mathbf{W}}\mathbf{Z}^\top\|_2^2$$

Picking y such that $\|\mathbf{C}y-\mathbf{Q}\mathbf{Q}^\top\mathbf{C}y\|_2^2=\|\mathbf{C}-\mathbf{Q}\mathbf{Q}^\top\mathbf{C}\|_2^2$, and combining it with Equation 8.41 we have

$$\|\mathbf{C}-\mathbf{Q}\mathbf{Q}^\top\mathbf{C}\|_2^2\leqslant\tilde{O}(1)\left(\frac{\epsilon}{k}\|\mathbf{C}-\mathbf{C}_{k'}\|_F^2\right) \tag{8.42}$$

Conditioning on event ζ_1, we know that $\|\mathbf{C}-\mathbf{C}_{k'}\|_F^2=\|\mathbf{A}-\mathbf{A}_{k'}\|_F^2$. Since $\mathbf{Q}\mathbf{Q}^\top$ is a rank-

k' projection matrix and $\mathbf{C}$ is a mixed Spectral-Frobenius column PCP for $\mathbf{A}$, it follows from Equation 8.36,

$$\begin{aligned}\|\mathbf{A}-\mathbf{Q}\mathbf{Q}^\top\mathbf{A}\|_2^2 &\leqslant \frac{10}{9}\|\mathbf{C}-\mathbf{Q}\mathbf{Q}^\top\mathbf{C}\|_2^2+\frac{1}{9k'}\|\mathbf{A}-\mathbf{A}_{k'}\|_F^2\\ &\leqslant \widetilde{O}(1)\left(\frac{\epsilon}{k}\|\mathbf{A}-\mathbf{A}_{k'}\|_F^2\right)\end{aligned} \tag{8.43}$$

where the last inequality follows from Equation 8.42. Therefore, $\mathbf{Q}\mathbf{Q}^\top$ is a $(0.1,k')$-SF projection for $\mathbf{A}$. Finally, we run Algorithm 9 on $\min_{\text{rank}(\mathbf{X})=k}\|\mathbf{A}-\mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2$. Here, we note that for Algorithm 9, a $(0.1,k')$-SF projection is equivalent to an (ϵ,k)-SF projection, up to poly-logarithmic factors. Therefore, Theorem 168 holds as is. Then, by Theorem 168, we know that with probability at least $99/100$ Algorithm 9 outputs matrices $\mathbf{M},\mathbf{N}$ such that $\|\mathbf{A}-\mathbf{M}\mathbf{N}^\top\|_F^2\leqslant(1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$. Let ζ_5 be the event that the aforementioned algorithm succeeds. Then, union bounding over $\zeta_1,\zeta_2,\zeta_3,\zeta_4$ and ζ_5, with probability at least $9/10$, $\mathbf{M},\mathbf{N}$ is a relative-error Low-Rank Approximation for $\mathbf{A}$, which concludes correctness.

Next, we analyze the query complexity and running time of Algorithm 11. Step 1 computes the rank-k' *ridge leverage scores* of $\mathbf{A}^{1/2}$ and by Lemma 8.3.8, requires reading $O(nk'\log(k'))=\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ and runs in time $\widetilde{O}(n(k/\epsilon)^{\omega-1})$. Steps 2 and 3 require no queries to $\mathbf{A}$ and the sampling can be performed in $O(n)$ time. In step 4, the *input sparsity* algorithm from Lemma 9.4.1 queries $\mathsf{nnz}(\mathbf{R})=t^2=\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ and runs in $\widetilde{O}(nk/\epsilon+\sqrt{n}\text{poly}(k/\epsilon))=\widetilde{O}(nk/\epsilon)$ time. We know from Theorem 168 that Step 5 requires $\widetilde{O}(nk/\epsilon)$ queries to $\mathbf{A}$ and runs in $\widetilde{O}(nk/\epsilon+(k/\epsilon)^\omega)$ time. Finally, in Step 6, we run Algorithm 9 such that $\mathbf{Q}$ is a $n\times k'$ matrix. Therefore, it follows from Theorem 168, that the total number of queries to $\mathbf{A}$ is $\widetilde{O}(nk/\epsilon+k^2/\epsilon^6)$ and the running time is $\widetilde{O}(n(k/\epsilon)^{\omega-1}+(k/\epsilon^3)^\omega)$. The final query complexity and running time follows, and this concludes the proof.

Outputting a PSD Low-Rank Approximation. Here, we extend our algorithm to show that we can obtain a relative-error low-rank approximation matrix $\mathbf{B}$ such that $\mathbf{B}$ itself is a PSD matrix, using the same sample complexity and running time as in Theorem 166. Outputting a PSD low-rank approximation was first considered by Clarkson and Woodruff [CW17], who obtain an input-sparsity algorithm for arbitrary $\mathbf{A}$. When $\mathbf{A}$ is PSD, Musco and Woodruff show that this problem can be solved with $\widetilde{O}(nk/\epsilon^3+nk^2/\epsilon^2)$ queries, in time $\widetilde{O}(n(k/\epsilon)^\omega+nk^{\omega-1}/\epsilon^{3(\omega-1)})$.

We run Algorithm 11 till Step 5, i.e., we recover $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{Q}^\top$ is a SF projection for $\mathbf{A}$.

We then modify Algorithm 9 by considering the following optimization problem instead:

$$\min_{\substack{\text{rank}(\mathbf{X})\leqslant k\\ \mathbf{X}\succeq 0}} \|\mathbf{A}-\mathbf{Q}\mathbf{X}\mathbf{Q}^\top\|_F^2 \tag{8.44}$$

As before, we sketch on both sides by sampling proportional to the leverage scores of $\mathbf{Q}$. Let the resulting sampling matrices be denoted by $\mathbf{S},\mathbf{T}$. Then, we have the following sketched optimization problem:

$$\min_{\substack{\text{rank}(\mathbf{X})\leqslant k\\ \mathbf{X}\succeq 0}} \|\mathbf{S}\mathbf{A}\mathbf{T}-\mathbf{S}\mathbf{Q}\mathbf{X}\mathbf{Q}^\top\mathbf{T}\|_F^2 \tag{8.45}$$

Following Step 4 in Algorithm 9, we can compute $\mathbf{SAT}$, $\mathbf{P}_{\mathbf{SQ}}$, $\mathbf{P}_{\mathbf{Q}^\top\mathbf{T}}$. We then compute $\widehat{\mathbf{X}} = (\mathbf{SQ})^\dagger\mathbf{P}_{\mathbf{SQ}}\mathbf{SATP}_{\mathbf{Q}^\top\mathbf{T}}(\mathbf{Q}^\top\mathbf{T})^\dagger$ and $\mathbf{X}^* = [(\widehat{\mathbf{X}}+\widehat{\mathbf{X}}^\top)/2]_{k+}$, where for any matrix $\mathbf{M}$, $[\mathbf{M}]_{k+}$ is defined by setting all but the top-k positive eigenvalues to 0. Finally, we output $\mathbf{N}\mathbf{N}^\top$ where $\mathbf{N}=\mathbf{Q}(\mathbf{X}^*)^{1/2}$.

Corollary 8.3.19. *(Outputting a PSD Low-Rank Approximation.) Given an $n\times n$ PSD matrix $\mathbf{A}$, an integer k, and $1>\epsilon>0$, there exists an algorithm that samples $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ and outputs a rank-k $\mathbf{M}\mathbf{M}^\top$ such that with probability at least $9/10$,*

$$\|\mathbf{A}-\mathbf{M}\mathbf{M}^\top\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$$

Further, the algorithm runs in $\widetilde{O}(n(k/\epsilon)^{\omega-1}+(k/\epsilon^3)^\omega)$ time.

Proof. We first note that an extension of Lemma 8.3.13 holds for outputting a PSD matrix as well. As a consequence of the following lemma, obtaining an approximate solution to the optimization problem in Equation 8.44 suffices.

Lemma 8.3.20. *(Structured Projections and PSD LRA [CW17].) Let $\mathbf{P}\in\mathbb{R}^{n\times n}$ be an (ϵ,k)-SF projection w.r.t $\mathbf{A}$, then*

$$\|\mathbf{A}-\mathbf{P}\mathbf{A}_{k+}\mathbf{P}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A}-\mathbf{A}_{k+}\|_F^2$$

We then use the analysis of Lemma 15 from [CW17] to conclude that $\mathbf{X}^*$ is the minimizer for Equation 8.45. Finally, we note that the running time and query complexity is dominated by computing $\mathbf{Q}$ and thus is the same as Theorem 166. Computing $\mathbf{X}^*$ requires no additional queries to $\mathbf{A}$ and only contributes a lower order term to the running time. □

8.3.4 Negative-Type Distances

In this subsection, we consider the problem of computing low-rank approximation for distance matrices. Here, the input matrix $\mathbf{A}$ is formed by the pairwise distances between a set of points $\mathbf{P} = \{p_1, \ldots, p_n\}$ in an underlying metric space d, i.e., $A_{i,j} = d(p_i, p_j)$. Low-rank approximation for distance matrices was introduced by Bakshi and Woodruff [BW18] who obtained sublinear time *additive-error* algorithms for arbitrary metrics. Subsequently, Indyk et. al. [IVWW19] provided sample-optimal algorithms for additive-error low-rank approximation. For arbitrary distance matrices, it is known that relative-error algorithms require $\Omega(\text{nnz}(\mathbf{A}))$ queries [BW18].

Here, we focus on the special case of negative-type (Euclidean Squared) metrics [Sch38]. Negative-type metrics have numerous applications in algorithm design since it is possible to optimize over them using a semidefinite program (SDP). One significant algorithmic application of negative-type metrics appears in the Arora-Rao-Vazirani algorithm for the Sparsest Cut problem [ARV09]. We refer the reader to extensive subsequent work on embeddability of such metrics and the references therein [ALN08, ALN07, CGR05]. It is well-known that negative-type metrics include ℓ_1 and ℓ_2 metrics, spherical metrics and hyper metrics [DL09, TD87]. Therefore, our algorithms extend to distance matrices that arise from all such metrics.

For negative-type metrics, Bakshi and Woodruff obtain a bi-criteria relative-error low-rank approximation algorithm that queries $\widetilde{O}(nk/\epsilon^{2.5})$ entries in $\mathbf{A}$ and output a rank $k+4$ matrix. In contrast, we obtain a sample-optimal algorithm that does not require a bi-criteria guarantee. As noted above, our algorithm works for any distance matrix where the distance can be realized as a negative-type metric.

Theorem 173 (Sample-Optimal Negative-Type LRA)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a negative-type distance matrix. Given $\epsilon > 0$ and $k \in [n]$, there exists an algorithm that queries $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ and outputs matrices $\mathbf{M}, \mathbf{N}^\top \in \mathbb{R}^{n\times k}$ such that with probability $99/100$,*

$$\|\mathbf{A} - \mathbf{MN}\|_F^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Further, the algorithm runs in time $\widetilde{O}(n(k/\epsilon)^{\omega-1})$.

To demonstrate the connection between negative-type metrics and PSD matrices, we observe that a negative-type distance matrix $\mathbf{A}$ can be realized as the distances corresponding to a point set $\mathcal{P} = \{x_1, x_2, \ldots x_n\}$ such that $\mathbf{A}_{i,j} = \|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2\langle x_i, x_j\rangle$. Therefore, we can rewrite $\mathbf{A}$ as $\mathbf{R}_1 + \mathbf{R}_2 - 2\mathbf{B}$, where for all $j \in [n]$, $(\mathbf{R}_1)_{i,j} = \|x_i\|_2^2$, $\mathbf{R}_2 = \mathbf{R}_1^\top$ and $\mathbf{B}$ is

PSD. Further, we can obtain query access to $\mathbf{B}$ by simply assuming w.l.o.g. that x_1 is centered at the origin and the i-th entry in the first row corresponds to $\|x_i\|_2^2$. Therefore, we can simulate our PSD low-rank approximation algorithms on the matrix $\mathbf{B}$ by only having query access to $\mathbf{A}$.

Our main contribution here is to show that if $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ is an $(O(1), k/\epsilon)$-SF projection matrix for $\mathbf{B}$, then adjoining $\mathbf{Q}^\top$ with the row span of $\mathbf{R}_1$ and $\mathbf{R}_2$ results in an SF-projection matrix for $\mathbf{A}$. Here, the row span of $\mathbf{R}_1$ is $\mathbf{1}^\top/\sqrt{n}$ and $\mathbf{R}_2$ is v such that for all $i \in [n]$ $v_i = \|x_i\|_2^2/\sum_i \|x_i\|_2^2$. We note that once we obtain an SF projection for $\mathbf{A}$, we can run Algorithm 9 to output a $(1+\epsilon)$ relative-error low-rank approximation.

Lemma 8.3.21 (Structured Projections for Distance Matrices)**.** *Let $\mathbf{A}$ be a negative-type matrix such that $\mathbf{A} = \mathbf{R}_1 + \mathbf{R}_2 - 2\mathbf{B}$, as defined above and let $\epsilon > 0$. Given an $(O(1), k/\epsilon)$-SF projection $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ for $\mathbf{B}$, let $\mathbf{\Omega}^\top$ be a basis for $\mathbf{Q}^\top$ appended with the basis vectors for rowspan$(\mathbf{R}_1)$ and rowspan$(\mathbf{R}_2)$. Then, with probability at least $99/100$,*

$$\min_{\text{rank}(\mathbf{X}) \leqslant k} \|\mathbf{A} - \mathbf{\Omega}\mathbf{X}\mathbf{\Omega}^\top\|_2^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Proof. By Lemma 7 in [CW17], for any symmetric matrices $\mathbf{Y}, \mathbf{Z}$ such that $(\mathbf{Y} - \mathbf{Z})\mathbf{Z} = 0$ and projection matrix $\mathbf{P}$, the following holds:

$$\|\mathbf{Y} - \mathbf{P}\mathbf{Z}\mathbf{P}\|_F^2 = \|\mathbf{Y} - \mathbf{Z}\|_F^2 + \|\mathbf{Z} - \mathbf{P}\mathbf{Z}\mathbf{P}\|_F^2 + 2\text{Tr}\left[(\mathbf{Y} - \mathbf{Z})(\mathbf{I} - \mathbf{P})\mathbf{Z}\mathbf{P}\right] \tag{8.46}$$

Applying Equation 8.46 with $\mathbf{Y} = \mathbf{A}$ and $\mathbf{Z} = \mathbf{A}_k$, for any projection matrix $\mathbf{P}$, we have

$$\|\mathbf{A} - \mathbf{P}\mathbf{A}_k\mathbf{P}\|_F^2 = \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \|\mathbf{A}_k - \mathbf{P}\mathbf{A}_k\mathbf{P}\|_F^2 + 2\text{Tr}\left[(\mathbf{A} - \mathbf{A}_k)(\mathbf{I} - \mathbf{P})\mathbf{A}_k\mathbf{P}\right] \tag{8.47}$$

Next, we bound the $\|\mathbf{A}_k - \mathbf{P}\mathbf{A}_k\mathbf{P}\|_F^2$ as follows:

$$\begin{aligned} \|\mathbf{A}_k - \mathbf{P}\mathbf{A}_k\mathbf{P}\|_F^2 &\leqslant 2\|\mathbf{A}_k(\mathbf{I} - \mathbf{P})\|_F^2 \\ &\leqslant 2k\|\mathbf{A}_k(\mathbf{I} - \mathbf{P})\|_2^2 \\ &\leqslant 2k\|\mathbf{A}(\mathbf{I} - \mathbf{P})\|_2^2 \end{aligned} \tag{8.48}$$

To bound the trace, we use the Von Neuman trace inequality,

$$\begin{aligned} 2\text{Tr}\left[(\mathbf{A}-\mathbf{A}_k)(\mathbf{I}-\mathbf{P})\mathbf{A}_k\mathbf{P}\right] &= 2\text{Tr}\left[(\mathbf{A}-\mathbf{A}_k)(\mathbf{I}-\mathbf{P})^2\mathbf{A}_k\mathbf{P}\right] \\ &\leqslant 2\sum_{i\in[n]}\sigma_i((\mathbf{A}-\mathbf{A}_k)(\mathbf{I}-\mathbf{P}))\sigma_i((\mathbf{I}-\mathbf{P})\mathbf{A}_k\mathbf{P}) \\ &\leqslant 2k\|(\mathbf{A}-\mathbf{A}_k)(\mathbf{I}-\mathbf{P})\|_2\|(\mathbf{I}-\mathbf{P})\mathbf{A}_k\mathbf{P}\|_2 \\ &\leqslant 2k\|\mathbf{A}(\mathbf{I}-\mathbf{P})\|_2^2 \end{aligned} \tag{8.49}$$

It suffices to bound $\|\mathbf{A}(\mathbf{I}-\mathbf{P})\|_2^2$ for $\mathbf{P}=\mathbf{\Omega}\mathbf{\Omega}^\top$. Since $\mathbf{A}=\mathbf{R}_1+\mathbf{R}_2-2\mathbf{B}$ and $(\mathbf{R}_1+\mathbf{R}_2)(\mathbf{I}-\mathbf{P})=0$, we have

$$\begin{aligned} \|\mathbf{A}(\mathbf{I}-\mathbf{P})\|_2^2 &\leqslant 2\|\mathbf{B}(\mathbf{I}-\mathbf{P})\|_2^2 \\ &\leqslant 2\|\mathbf{B}(\mathbf{I}-\mathbf{Q}\mathbf{Q}^\top)\|_2^2 \\ &\leqslant O\left(\frac{\epsilon}{k}\right)\|\mathbf{B}-\mathbf{B}_{k+2}\|_F^2 \end{aligned}$$

To relate $\|\mathbf{B}-\mathbf{B}_k\|_F^2$ back to $\mathbf{A}$, observe

$$\begin{aligned} \|\mathbf{A}-\mathbf{A}_k\|_F^2 = \|\mathbf{R}_1+\mathbf{R}_2-2\mathbf{B}-\mathbf{A}_k\|_F^2 &= 4\|\mathbf{B}-(\mathbf{R}_1+\mathbf{R}_2-\mathbf{A}_k)/2\|_F^2 \\ &\geqslant 4\|\mathbf{B}-\mathbf{B}_{k+2}\|_F^2 \end{aligned}$$

Therefore, $\|\mathbf{A}(\mathbf{I}-\mathbf{P})\|_2^2 \leqslant O(\epsilon/k)\|\mathbf{A}-\mathbf{A}_k\|_F^2$. We can thus bound Equations 8.49 and 8.48 with $O(\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_F^2$. Substituting this into Equation 8.47, we conclude that $\|\mathbf{A}-\mathbf{P}\mathbf{A}_k\mathbf{P}\|_F^2 \leqslant (1+O(\epsilon))\|\mathbf{A}-\mathbf{A}_k\|_F^2$, for $\mathbf{P}=\mathbf{\Omega}\mathbf{\Omega}^\top$ and the claim follows. □

Recall, we can compute an SF projection for the PSD matrix $\mathbf{B}$ efficiently using Algorithm 11 and then solve the optimization problem in Lemma 8.3.21 using Algorithm 9. We can therefore reduce low-rank approximation of negative-type matrices to PSD low-rank approximation with only $O(n)$ additional queries and Theorem 173 follows.

8.3.5 Ridge Regression

We consider the following regression problem: given a PSD matrix $\mathbf{A}$, a vector y and a ridge parameter λ,

$$\min_x \|\mathbf{A}x-y\|_2^2+\lambda\|x\|_2^2.$$

As a corollary of Theorem 166, we obtain a faster algorithm for the aforementioned problem. We begin with the following simple lemma from [MW17c]:

Lemma 8.3.22 (Lemma 26 in [MW17c]). *Given a PSD matrix $\mathbf{A}$, vector y, and $\lambda > 0$, let $\mathbf{B}$ be a matrix such that $\|\mathbf{A} - \mathbf{B}\|_2^2 \leqslant \epsilon^2\lambda$. Then, for any vector $\tilde{x}$ such that*

$$\|\mathbf{B}\tilde{x} - y\|_2^2 + \lambda\|\tilde{x}\|_2^2 \leqslant (1+\epsilon')\left(\min_x \|\mathbf{B}x - y\|_2^2 + \lambda\|x\|_2^2\right)$$

we have

$$\|\mathbf{A}\tilde{x} - y\|_2^2 + \lambda\|\tilde{x}\|_2^2 \leqslant (1+\epsilon')(1+5\epsilon)\left(\min_x \|\mathbf{A}x - y\|_2^2 + \lambda\|x\|_2^2\right)$$

Therefore, it suffices to find a rank-k matrix $\mathbf{B}$ such that $\|\mathbf{A}-\mathbf{B}\|_2^2 \leqslant \epsilon^2\lambda$. Let $\widetilde{s}_\lambda$ be an upper bound on the statistical dimension $s_\lambda = \mathrm{Tr}\left[(\mathbf{A}^2 + \lambda\mathbf{I})^{-1}\mathbf{A}^2\right]$ Setting $k = \widetilde{s}_\lambda/\epsilon^2$, we can bound $\|\mathbf{A} - \mathbf{A}_k\|_F^2$ as follows:

$$\begin{aligned}
\frac{\epsilon^2}{\widetilde{s}_\lambda}\|\mathbf{A} - \mathbf{A}_k\|_F^2 &\leqslant \epsilon^2 \frac{\sum_{i=k+1}^n \lambda_i^2(\mathbf{A})}{\sum_{i=1}^n \lambda_i^2(\mathbf{A})/(\lambda_i^2(\mathbf{A}) + \lambda)} \\
&\leqslant \epsilon^2 \frac{\sum_{i=k+1}^n \lambda_i^2(\mathbf{A})}{\sum_{i=k+1}^n \lambda_i^2(\mathbf{A})/(\lambda_i^2(\mathbf{A}) + \lambda)} \\
&\leqslant c\epsilon^2\lambda
\end{aligned}$$

We can then solve the regression problem $\min_x \|\mathbf{B}x - y\|_2^2 + \lambda\|x\|_2^2$ exactly in time $O(nk^{\omega-1})$ and obtain the following result:

Theorem 174 (Ridge Regression). *Given a PSD matrix $\mathbf{A}$, a regularization parameter λ and an upper bound $\widetilde{s}_\lambda$ on the statistical dimension $s_\lambda = \mathrm{Tr}\left[(\mathbf{A}^2 + \lambda\mathbf{I})^{-1}\mathbf{A}^2\right]$, there exists an algorithm that queries $\widetilde{O}(n\widetilde{s}_\lambda/\epsilon^2)$ entries of $\mathbf{A}$ and with probability $99/100$ outputs $\widehat{x}$ such that for all $y \in \mathbb{R}^d$,*

$$\|\mathbf{A}\widehat{x} - y\|_2^2 + \lambda\|\widehat{x}\|_2^2 \leqslant (1+\epsilon)\left(\min_x \|\mathbf{A}x - y\|_2^2 + \lambda\|x\|_2^2\right)$$

Further, the algorithm runs in $\widetilde{O}(n(\tilde{s}_\lambda/\epsilon^2)^{\omega-1})$ time.

Remark 175. Observe that we can derive a data structure from our algorithm that preserves the objective cost (up to $1+\epsilon$) for all x and y simultaneously and thus we obtain a coreset for Ridge Regression.

To complement the above algorithmic result, we present a new lower bound for coreset constructions for ridge regression, which matches our upper bound in all parameters. At a high level,

our hard instance for constant s_λ consists of $1/\epsilon^2$ blocks of all 1s, each of size $\epsilon\sqrt{n} \times \epsilon\sqrt{n}$, placed randomly across the matrix. Since any coreset construction must preserve the cost of all x, y, we pick pairs (x, y) to be the eigenvectors of $\mathbf{A}$ (scaled appropriately) and show that in order to preserve the cost of all pairs, the coreset algorithm must find all the blocks, which requires $\Omega(n/\epsilon^2)$ queries to $\mathbf{A}$. Repeating the above construction s_λ-times suffices to obtain a linear lower bound in terms of s_λ. Formally,

Theorem 176 (Coreset Lower Bound for Ridge Regression)**.** *Given a PSD matrix $\mathbf{A}$ and $\epsilon, \lambda > 0$ let $s_\lambda = Tr\left[(\mathbf{A}^2 + \lambda\mathbf{I})^{-1}\mathbf{A}^2\right]$ denote the statistical dimension of $\mathbf{A}$. Then, any coreset construction $\mathcal{C}$ that with constant probability, preserves the ridge regression cost up to $(1+\epsilon)$ simultaneously for all x, y, must read $\Omega(ns_\lambda/\epsilon^2)$ entries in $\mathbf{A}$.*

We recall the lower bound instance for low rank approximation of PSD matrices shown by Musco and Woodruff :

Theorem 177 (Lower Bound for PSD LRA ([MW17c]))**.** *Given an $n \times n$ PSD matrix $\mathbf{A}$, $\epsilon_0 > 0$ and $k_0 \in [n]$, any randomized algorithm that outputs a rank k_0 matrix $\mathbf{B}$ such that with probability at least $9/10$,*

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leqslant (1+\epsilon_0)\|\mathbf{A} - \mathbf{A}_{k_0}\|_F^2$$

must query $\Omega(nk_0/\epsilon_0)$ entries in $\mathbf{A}$.

We consider the hard distribution defined by Musco and Woodruff, and show that we can obtain a low rank approximation to this instance with strengthened parameters by using a coreset for ridge regression.

Definition 8.3.23 (Hard Input Distribution for LRA ([MW17c]))**.** *Let $\mathbf{M}$ be an $n \times n$ matrix and let $\epsilon_0 > 0$, $k_0 \in [n]$. Let $\gamma(n, \epsilon_0, k_0)$ be a distribution over $\mathbf{M}$ such that $\mathcal{S} \subset [n]$ is a uniformly random subset of size $n/2$, which is further partitioned into subsets $\mathcal{S}_1, \mathcal{S}_2, \ldots \mathcal{S}_{k_0}$ such that for all $\ell \in [k]$, $\mathcal{S}_\ell$ is picked uniformly at random and $|\mathcal{S}_\ell| = n/(2k_0)$. For each subset $\mathcal{S}_\ell$, let $\mathbf{A}_{\mathcal{S}_\ell}$ denote the principle submatrix of A indexed by the set $\mathcal{S}_\ell$. Then, with probability $1/2$, $\mathbf{A}_{\mathcal{S}_\ell}$ is such that all the diagonal entries are set to 1 and a uniformly random principle submatrix of $\mathbf{A}_{\mathcal{S}_\ell}$, indexed by the set $\mathcal{T}_\ell$, such that $|\mathcal{T}_\ell| = c\sqrt{\epsilon_0|\mathcal{S}_\ell|}$ is set to all 1s. With the remaining probability $\mathbf{A}_{\mathcal{S}_\ell}$ is set to the $\mathbf{I}$.*

We show that we can derive a low-rank matrix $\mathbf{B}$ that satisfies the relative-error guarantee above from a coreset for ridge regression.

Proof of Theorem 176. We show a proof by contradiction, where the high level idea is that a coreset for ridge regression can be used to derive a low-rank approximation to $\mathbf{A}$, when $\mathbf{A}$ is picked from $\gamma(n, O(1), s_\lambda/\epsilon^2)$ (the hard distribution defined in 8.3.23). First, we observe that with probability at least $99/100$, the input distribution has $\Omega(s_\lambda/\epsilon^2)$ blocks that contain a principle submatrix with all 1s. To see this let $X_1, \ldots X_{s_\lambda/\epsilon^2}$ be indicators for the corresponding blocks $\mathbf{A}_{\mathcal{S}_\ell}$ having a principle submatrix of all 1s. Then,

$$\mathbb{E}\left[\sum_{\ell\in[s_\lambda/\epsilon^2]} X_\ell\right] = \frac{\epsilon^2 n}{2s_\lambda} \tag{8.50}$$

Since the X_ℓ's are independent, by a Chernoff bound we have

$$\mathbf{Pr}\left[\right] \sum_{\ell\in[s_\lambda/\epsilon^2]} X_\ell \leqslant (1-\delta)\frac{\epsilon^2 n}{2s_\lambda} \leqslant \exp\left(-\frac{c\delta\epsilon^2 n}{s_\lambda}\right) \tag{8.51}$$

For $n \geqslant \Omega(s_\lambda/\epsilon^2)$, we can bound the above probability by $1/100$. We begin by showing that for our input instance, $s_\lambda = \Theta(n/\lambda)$ and thus the aforementioned equations differ by $O(\epsilon n/s_\lambda)$. To see this observe

$$s_\lambda = \sum_{i\in[n]} \frac{\sigma_i^2(\mathbf{A})}{\sigma_i^2(\mathbf{A})+\lambda} \tag{8.52}$$

Then, there are $s_\lambda/2\epsilon^2$ large eigenvalues, each of magnitude $\epsilon\sqrt{n/s_\lambda}$ and thus the total contribution is

$$\frac{s_\lambda}{2\epsilon^2}\cdot\frac{(\epsilon^2 n/s_\lambda)}{(\epsilon^2 n/s_\lambda)+\lambda} = \frac{n}{\epsilon^2 n/s_\lambda+\lambda}$$

The remaining eigenvalues simply contribution $1/(1+\lambda)$ to the sum and since there are at most n of them, the total contribution is $n/(1+\lambda)$. Therefore, we can conclude $s_\lambda = \Theta(n/\lambda)$.

For a block in $\mathbf{A}$ indexed by ℓ, let $\hat{x}_\ell$ be the eigenvector supported on indices in $\mathcal{S}_\ell$ and let $\hat{y}_\ell = \sqrt{n/s_\lambda}\hat{x}_\ell$. For non-identity blocks, $\mathbf{A}\hat{x}_\ell = |\mathcal{T}_\ell|\hat{x}_\ell = \sqrt{\epsilon^2 n/s_\lambda}\hat{x}_\ell$ and the regression cost is

$$\|(1-\epsilon)\sqrt{n/s_\lambda}\hat{x}_\ell\|_2^2 + \lambda = (1-2\epsilon)n/s_\lambda + cn/s_\lambda \tag{8.53}$$

When the block indexed by ℓ is the identity block, we get $\mathbf{A}\hat{x}_\ell = \hat{x}$ and the regression cost is

$$\|(\sqrt{n/s_\lambda}-1)\hat{x}_\ell\|_2^2 + \lambda = (n/s_\lambda + 1 - 2\sqrt{n/s_\lambda}) + cn/s_\lambda \tag{8.54}$$

Instead consider a vector that intersects an eigenvector $\tilde{x}_\ell$ on a $(1-\gamma)$-fraction of the support and the rest is arbitrary. Then, when an an all 1s block exists, $\mathbf{A}\tilde{x}_\ell \geqslant (1-\gamma)^2|\mathcal{T}_\ell|\hat{x}_\ell = (1-$

$\gamma)^2\sqrt{\epsilon^2 n/s_\lambda}\hat{x}_\ell$ and thus the regression cost is at most

$$(1-\epsilon(1-2\gamma))^2 n/s_\lambda + cn/s_\lambda$$

Further, when the block is simply the identity, a similar calculation shows that the regression cost is at least $(1-2\epsilon\gamma)^2 n/s_\lambda + cn/s_\lambda$. Therefore, the ridge regression cost determines the existence of a $(1-2\gamma)$-fraction of an all 1s principle submatrix even when $\hat{x}_\ell$ intersects with an eigenvector on a $(1-\gamma)$-fraction of coordinates.

Consider a coreset $\mathcal{C}$ for the above instance. Since this coreset preserves the ridge regression objective upto a $(1+\epsilon/1000)$ factor for all x, y, as per our above discussion we can query the coreset on the tuples $(\hat{x}_\ell, \hat{y}_\ell)$, which represent the eigenvectors of each block, to determine if a block contains a principle submatrix with all 1s. However, a priori we do not know the support of the eigenvector within each block $\mathbf{A}_{\mathcal{S}_\ell}$.

Instead we query the coreset on all possible supports and show we can determine the right one as follows: let $\tilde{x}$ be supported on a set that intersects with a principle submatrix of all 1s on at most a γ fraction. Observe that $\mathbf{A}\tilde{x} \leqslant \frac{1}{\gamma}(\gamma^2\epsilon\tilde{x})$ and thus the ridge regression cost can be lower bounded as follows:

$$(1-\epsilon\gamma)^2 n/s_\lambda + cn/s_\lambda \tag{8.55}$$

We therefore take the set of all vectors on which the coreset cost is less than the above cost and let the resulting list be $\mathcal{L}$. Note, this list must include the eigenvectors and further, only includes vectors which intersect an all 1's submatrix on a $1-\gamma$-fraction. Therefore, picking a set of $\epsilon^2 n/s_\lambda$ vectors that have maximum support suffices.

Since we detect a $(1-\gamma)$-fraction of all principle submatrices in $\mathbf{A}$, it follows that we can output a $1+c'$-approximate low-rank approximation for $\mathbf{A}$, for a fixed small constant c'. To see this, observe that the optimal low-rank approximation to $\mathbf{A}$ is given by the matrix that selects all the principle submatrices with all 1s and thus $\|\mathbf{A}-\mathbf{A}_{k_0}\|_F^2 = n - k_0 = n - s_\lambda/\epsilon^2$. Further, our approximation to $\mathbf{A}_k$, denoted by $\mathbf{B}$, matches $\mathbf{A}_k$ on a $(1-\gamma)$-fraction of each principle submatrix of all 1s and thus we match $\mathbf{A}_k$ on these entries. Subsequently we bound the additional cost that $\mathbf{B}$ incurs which $\mathbf{A}_k$ does not. This includes entries that are 1 in $\mathbf{A}_k$ and 0 in $\mathbf{B}$ and vice versa.

To bound the cost of the entries that exist in $\mathbf{A}_k$ but do not exist in $\mathbf{B}$, observe on each principle submatrix, $\mathbf{B}$ and $\mathbf{A}$ intersect in at least $(1-\gamma)^2$-fraction of the entries and thus the remaining entries are at most $(1-(1-\gamma)^2)\cdot\frac{\epsilon^2 n}{cs_\lambda}\cdot\frac{s_\lambda}{\epsilon^2} = 4\gamma n/c$, since the size of each block is $\frac{\epsilon^2 n}{cs_\lambda}$ and the number of blocks are at most $\frac{s_\lambda}{\epsilon^2}$. Finally, observe that since we do not pick exact

eigenvectors we can have non-zero off diagonal entries in $\mathbf{B}$ that do not exist in $\mathbf{A}_k$. However, we have at most γ-fraction of the support on each indicator vector remaining and contributing to two rectangular blocks of 1s, each of size $\gamma \cdot \frac{\epsilon^2 n}{c s_\lambda} \cdot \frac{s_\lambda}{\epsilon^2} = \gamma n/c$. Therefore, the additional non-zero entries in $\mathbf{B}$ that do not appear in $\mathbf{A}$ are $2\gamma n/c$ in number.

Therefore, the overall cost $\|\mathbf{A} - \mathbf{B}\|_F^2 \leqslant (1 + 6\gamma/c)n$. By setting the constants γ and c and observing that $s_\lambda/\epsilon^2 \ll n$, we obtain a $(1 + c')$-low-rank approximation to $\mathbf{A}$, for any arbitrary small constant c'. Therefore, our reduction suffices to solve the hard instance above and a lower bound of $\Omega(nk_0/\epsilon_0) = \Omega(ns_\lambda/\epsilon^2)$ queries follows. □

8.4 Robust Low-Rank Approximation

One drawback of relative-error guarantees is that the corresponding algorithms cannot tolerate any amount of noise. Therefore, we introduce a robust model for low-rank approximation by relaxing the requirement from relative-error guarantees to additive-error guarantees. In the robustness model we consider, we begin with an $n \times n$ PSD matrix $\mathbf{A}$. An adversary is then allowed to arbitrarily corrupt $\mathbf{A}$ by adding a corruption matrix $\mathbf{N}$ such that the corruption in each row is an fixed constant times the ℓ_2^2 row norm of the row and the total corruption is an η-fraction of squared Frobenius norm of $\mathbf{A}$. While the adversary may corrupt any number of entries of $\mathbf{A}$, the norm of the corruption matrix is bounded and the algorithm has query access to $\mathbf{A} + \mathbf{N}$. We parameterize our lower bound and algorithms by the largest ratio between a diagonal entry of $\mathbf{A}$ and $\mathbf{A} + \mathbf{N}$, denoted by $\phi_{\max} = \max_{j \in [n]} \mathbf{A}_{j,j}/|(\mathbf{A} + \mathbf{N})_{j,j}|$. This captures the intuition that the diagonal is crucial for sublinear time low rank approximation and the sample complexity degrades as we corrupt larger diagonals entries.

8.4.1 Lower Bound for Robust PSD Low-Rank Approximation

In this subsection, we show a query lower bound of $\Omega(\eta^2 n^2 k/\epsilon^2) = \Omega(\phi_{\max}^2 nk/\epsilon)$ for any algorithm that outputs a low rank approximation up to additive-error $(\epsilon + \eta)\|\mathbf{A}\|_F^2$. Note, obtaining error smaller than $\eta\|\mathbf{A}\|_F^2$ is information-theoretically impossible and reflected in the query lower bound.

Our lower bound holds for randomized algorithms, and uses Yao's minimax principle [Yao77]. The overall strategy is to demonstrate a lower bound for deterministic algorithms on a carefully chosen input distribution. We construct our input distributions as follows: let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a

block diagonal matrix with such that $\mathbf{B}_1$ is $5\epsilon/\eta \times 5\epsilon/\eta$ randomly positioned, non-contiguous block with all entries $\sqrt{n\eta^2/5\epsilon}$ and $\mathbf{B}_2$ is the identity matrix on the remaining indices. $\mathbf{A}$ is clearly a PSD matrix since each principle submatrix is PSD. Observe $\|\mathbf{A}\|_F^2 = (25\epsilon/\eta^2) \cdot (n\eta^2/5\epsilon) + (n - 5\epsilon/\eta) = (1 + 5\epsilon)n - o(n)$. Further, the dense block $\mathbf{B}_1$ contributes a total squared Frobenius norm of at least $4\epsilon n$ and the diagonal entry contributes an η/ϵ fraction of each row. Since $\epsilon \geqslant \eta$, the diagonal contributes at most the entire ℓ_2^2 norm. The corresponding diagonal of $\mathbf{B}_1$ also has ℓ_2^2 squared norm $5\epsilon/\eta \cdot n\eta^2/5\epsilon = \eta n$.

At a high level, the adversary can then corrupt the diagonal and set each diagonal entry to be 1, making it hard for the algorithm to find rows corresponding to $\mathbf{B}_1$. We show that any ϵ-additive-error low-rank approximation must detect at least one entry in $\mathbf{B}_1$ to adaptively sample the corresponding row and column, but the diagonals no longer provide any useful information. Thus any algorithm must query most entries in $\mathbf{A}$. Further, in our construction, note $\phi_{\max} = \sqrt{n\eta^2/5\epsilon}$.

We first describe intuitively why a low rank approximation needs to recover many rows from the block $\mathbf{B}_1$. Since $\mathbf{A}$ has this block structure, the best rank-1 approximation satisfies $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = n - |\mathbf{B}_1|$. Therefore, assuming the cardinality of $\mathbf{B}_1$ is negligible, in order to obtain an overall error bound of $\epsilon\|\mathbf{A}\|_F^2 \geqslant \epsilon n$, the algorithm must find a constant fraction of off-diagonal entries in $\mathbf{B}_1$. This is because $\mathbf{B}_1$ contributes at least $5\epsilon n$ norm. However, since the diagonals no longer convey any information about the off-diagonals, and the block $\mathbf{B}_1$ is placed on a random subset of indices, any deterministic algorithm must read arbitrary off-diagonal entries until it finds a non-zero entry. Since there are only $25\epsilon^2/\eta^2$ non-zeros in $\mathbf{B}_1$, to find one in expectation (over the input distribution) requires sampling $\epsilon^2 n^2/\eta^2$ entries. While the above serves well as intuition, a rigorous proof requires many additional steps. We begin by defining a distribution over the input matrices :

Definition 8.4.1. *Given $n \in \mathbb{N}, \epsilon > \eta > 0$, let $\mathcal{S} \subset [n]$ be a uniformly random subset of size $\lceil 5\epsilon/\eta \rceil$. Let $\mu(n, \epsilon, \eta)$ be a distribution over matrices $\mathbf{M} \in \mathbb{R}^{n \times n}$ such that $\forall i \in [n]$, $\mathbf{M}_{i,i} = 1$ and $\forall i, j \in \mathcal{S}$, $\mathbf{M}_{i,j} = \sqrt{\eta^2 n/5\epsilon}$. All remaining entries in $\mathbf{M}$ are 0.*

Next, we show that any $\mathbf{M}$ sampled from $\mu(n, \epsilon, \eta)$ can be decomposed into $\mathbf{A} + \mathbf{N}$ such that $\mathbf{A}$ is PSD and $\|\mathbf{N}\|_F^2 \leqslant \eta\|\mathbf{A}\|_F^2$. To see this, let $\mathbf{N}$ be a diagonal matrix such that for all $i \in \mathcal{S}$, $\mathbf{N}_{i,i} = -\sqrt{\eta^2 n/5\epsilon}$ and let $\mathbf{A} = \mathbf{M} - \mathbf{N}$. We give an algebraic proof that $\mathbf{A}$ is PSD, but $\mathbf{A}$ can also be decomposed into a rank-1 block of all $\sqrt{\eta^2 n/5\epsilon}$-s corresponding to all $i, j \in \mathcal{S}$ and

identity on the remaining indices. For all $x \in \mathbb{R}^n$,

$$\begin{aligned} x^T \mathbf{A} x &= \sum_{i,j \in \mathcal{S}} \mathbf{A}_{i,j} x_i x_j + \sum_{i,j \notin \mathcal{S}} \mathbf{A}_{i,j} x_i x_j \\ &= \sqrt{\eta^2 n / 5\epsilon} \sum_{i,j \in \mathcal{S}} x_i x_j + \sum_{i \notin \mathcal{S}} x_i^2 \\ &= \sqrt{\eta^2 n / 5\epsilon} \left(\sum_{i \in \mathcal{S}} x_i \right)^2 + \sum_{i \notin \mathcal{S}} x_i^2 \\ &\geqslant 0 \end{aligned} \tag{8.56}$$

and thus $\mathbf{A}$ is PSD. Further, $\|\mathbf{A}\|_F^2 = (25\epsilon^2/\eta^2) \cdot (n\eta^2/5\epsilon) + (n - 5\epsilon/\eta) = (1 + 5\epsilon)n - 5\epsilon/\eta$. Then, $\|\mathbf{N}\|_F^2 = 5\epsilon/\eta \cdot \eta^2 n/\epsilon = \eta \|\mathbf{A}\|_F^2$, as desired. Intuitively, we show that if $\mathbf{B}$ is a rank-k matrix that is a good low-rank approximation for $\mathbf{M}$ sampled from μ, then it cannot be a good low-rank approximation for $\mathbf{I}$. To this end, we consider a distribution where $\mathbf{M}$ is drawn from $\mu(n, \epsilon, \eta)$ with probability $1/2$ and is $\mathbf{I}_{n \times n}$ with probability $1/2$.

Definition 8.4.2. *(Hard Distribution) Given $n \in \mathbb{N}, \epsilon > \eta > 0$, let $\nu(n, \epsilon, \eta)$ be a distribution over $\mathbf{M} \in \mathbb{R}^{n \times n}$ such that with probability $1/2$, $\mathbf{M}$ is sampled from $\mu(n, \epsilon, \eta)$ and with probability $1/2$, $\mathbf{M} = \mathbf{I}_{n \times n}$.*

We now show that a low-rank approximation to $\mathbf{M}$ can be used as a certificate to separate the mixture $\nu(n, \epsilon, \eta)$ since it can distinguish between the input being identity or far from it. Thus if the distributions are close in a statistical sense, any algorithm to distinguish between the two would require querying many entries in $\mathbf{M}$. Formally,

Lemma 8.4.3. *(LRA as a Distinguisher.) Let $\mathbf{M}$ be a matrix drawn from $\mu(n, \epsilon, \eta)$ and let $\mathbf{B}$ be a rank-k matrix that is the candidate low-rank approximation to $\mathbf{M}$ such that $\|\mathbf{M} - \mathbf{B}\|_F^2 \leqslant \epsilon n$. Then, $\|\mathbf{M} - \mathbf{I}\|_F^2 > 1.1\epsilon n$.*

Proof. Since $\|\mathbf{M} - \mathbf{B}\|_F^2 \leqslant \epsilon n$, $\mathbf{B}$ must have at least $4\epsilon n$ mass on the off-diagonal entries of $\mathbf{M}$. So, $\mathbf{B}$ must have at least $10\epsilon^2/\eta^2$ non-zero off-diagonal entries. Therefore, it must have at least $5\epsilon^2/\eta^2$ entries with squared mass $\epsilon n/2$. To see why, assume there is a subset of at least $12\epsilon^2/\eta^2$ entries, each being at most $\sqrt{n\eta^2/10\epsilon}$. Restricted to only these entries, the squared Frobenius norm difference between $\mathbf{M}$ and $\mathbf{B}$ is already at least $1.2\epsilon n$, contradicting our assumption. Given that there exists a subset of $5\epsilon^2/\eta^2$ off-diagonal entries having squared mass $1.2\epsilon n$, $\|\mathbf{B} - \mathbf{I}\|_F^2 > 1.2\epsilon n$, and thus $\mathbf{B}$ is not an additive error low-rank approximation for $\mathbf{I}$. □

Theorem 178. *(Lower bound for PSD Matrices.) Let $\mathbf{A}$ be a PSD matrix, $k \in \mathbb{Z}$ and $\epsilon > 0$*

be any constant. Let $\mathbf{N}$ *be an arbitrary matrix such that* $\|\mathbf{N}\|_F^2 \leqslant \eta\|\mathbf{A}\|_F^2$*. Any randomized algorithm* $\mathcal{A}$ *that only has query access to* $\mathbf{A}+\mathbf{N}$*, with probability at least* $2/3$*, computes a rank-*k *matrix* $\mathbf{B}$ *such that*

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant \|\mathbf{A}-\mathbf{A}_k\|_F^2 + \epsilon\|\mathbf{A}\|_F^2$$

must read $\Omega\left(\phi_{\max}^2 nk/\epsilon\right)$ *entries of* $\mathbf{A}+\mathbf{N}$ *on some input, possibly adaptively, in expectation.*

Proof. Let Algorithm $\mathcal{A}$ be a deterministic algorithm that outputs a rank-k matrix $\mathbf{B}$ such that it is an additive-error low-rank approximation $\mathbf{M}$. Let $T \subset [n^2]$ be the subset of entries read by $\mathcal{A}$. Let $L(\mu)$ denote the distribution of T conditioned on $\mathbf{M} \sim \mu(n,\epsilon,\eta)$ and $L(i)$ be the distribution of T conditioned on $\mathbf{M}=\mathbf{I}$. By Lemma 8.4.3, since the output of $\mathcal{A}$ can be used to distinguish between the two distributions, it is well-known that the success probability over the randomness in T is at most $1/2 + D_{TV}(L(\mu),L(i))/2$ [BY02]. Since we assume $\mathcal{A}$ succeeds with probability at least $2/3$,

$$D_{TV}(L(\mu),L(i)) \geqslant 1/3 \tag{8.57}$$

It remains to upper bound D_{TV} in terms of $|T|$. Recall, $\mathcal{S}$ is the random set of indices where $\mu(n,\epsilon,\eta)$ is non-zero. Let $\widehat{\mathcal{S}}$ be the subset of $\mathcal{S}$ restricted to the off-diagonal entries of $\mathbf{M}$. When $\mathbf{M} \sim \mu(n,\epsilon,\eta)$, $\forall i,j \in \widetilde{\mathcal{S}}$, $\mathbf{M}_{i,j}$ is non-zero and when $\mathbf{M}=\mathbf{I}$, the same entries are 0. Observe, for all $i,j \notin \widetilde{\mathcal{S}}$, $\mathbf{M}_{i,j}$ are fixed. Further, $\mathcal{S}$ is a uniform subset of $[n]$. Therefore,

$$\mathbf{Pr}\left[(i,j) \in T \mid (i,j) \in \mathcal{S}\right] = \frac{|T|\epsilon^2}{\eta^2 n^2} \tag{8.58}$$

Then, with probability at least $1-|T|\epsilon^2/\eta^2n^2$, $\mathcal{A}$ queries the same entries for both $L(\mu)$ and $L(i)$. Therefore

$$D_{TV}(L(\mu),L(i)) \leqslant |T|\epsilon^2/\eta^2n^2.$$

Combined with Equation 8.57, if $\mathcal{A}$ succeeds with probability at least $2/3$, $|T|\epsilon^2/\eta^2n^2 \geqslant 1/3$ and thus $|T| = \Omega(\eta^2n^2/\epsilon^2)$. Given that any deterministic algorithm must query $\Omega(\eta^2n^2/\epsilon^2) = \Omega(\phi_{\max}^2 n/\epsilon)$ entries for $\nu(n,\epsilon,\eta)$, to now obtain a linear dependence on the rank k, we can use the standard approach of creating k disjoint copies of the block $\mathbf{B}_1$ in the hard distribution, as shown in [MW17c]. The theorem follows from Yao's minimax principle. □

8.4.2 Robust Sublinear Low-Rank Approximation Algorithms

In this subsection, we provide a robust algorithm for the model discussed above. We parameterize our algorithms and lower bound by the largest ratio between a diagonal entry of $\mathbf{A}$ and $\mathbf{A}+\mathbf{N}$, denoted by $\phi_{\max} = \max_{j\in[n]} \mathbf{A}_{j,j}/|(\mathbf{A}+\mathbf{N})_{j,j}|$. In addition, we provide robust *PCP* constructions, by introducing a new sampling procedure to construct projection-cost preserving sketches. Our sampling procedure is straightforward: we sample each column proportional to the diagonal entry in that column. This sampling requires n queries to the matrix $\mathbf{A}$ to obtain an additive-error projection cost preservation guarantee. Further, for the special case of correlation matrices, we can uniformly sample columns of $\mathbf{A}$ to obtain a smaller matrix such that all rank k projections in the column and row space are preserved.

For our algorithms, we assume we know $\phi_{\max}$. In practice, this assumption may not hold, but we can query as many entries in $\mathbf{A}+\mathbf{N}$ as our budget allows, given that correctness holds only when the queries are at least $\widetilde{O}(\phi_{\max}^2 nk/\epsilon)$. Since we read the diagonals of $\mathbf{A}+\mathbf{N}$ and we know $\phi_{\max}$, we can obtain an upper bound on $\mathbf{A}_{i,i}$ and $\mathbf{A}_{j,j}$. Therefore, whenever we query an off-diagonal entry in $\mathbf{A}+\mathbf{N}$, we can truncate it to $\phi_{\max}\sqrt{|(\mathbf{A}+\mathbf{N})_{i,i}|\cdot|(\mathbf{A}+\mathbf{N})_{j,j}|}$ without increasing the corruption in our input.

Robust Projection-Cost Preserving Sketches. Here, we show that diagonal sampling is a robust sampling procedure to create *projection-cost preserving sketches*. We begin by relating the ℓ_2^2 row (or column) norms of a PSD matrix to it's spectral norm. Let $\mathbf{A}$ be a PSD matrix and let $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ be the SVD for $\mathbf{A}$.

Lemma 8.4.4. *Given an $n\times n$ PSD matrix $\mathbf{A}$, for all $i\in[n]$, $\|\mathbf{A}_{i,*}\|_2^2 \leqslant \|\mathbf{A}\|_2\cdot\mathbf{A}_{i,i}$.*

Proof. Observe, $\mathbf{A}_{i,*} = \mathbf{U}_i\mathbf{\Sigma}\mathbf{U}^T$ and $\mathbf{A}_{i,i} = (\mathbf{U}_{i,*}\mathbf{\Sigma}\mathbf{U}^T)_i = \sum_{j=1}^n \sigma_j(\mathbf{A})\mathbf{U}_{i,j}^2$. Then,

$$\begin{aligned}\|\mathbf{A}_{i,*}\|_2^2 = \mathbf{A}_{i,*}\mathbf{A}_{i,*}^T = \mathbf{U}_{i,*}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{U}_{i,*}^T &= \sum_{j=1}^n \sigma_j^2(\mathbf{A})\mathbf{U}_{i,j}^2\\ &\leqslant \|\mathbf{A}\|_2\sum_{j=1}^n \sigma_j\mathbf{U}_{i,j}^2\\ &= \|\mathbf{A}\|_2\cdot\mathbf{A}_{i,i}\end{aligned}$$

□

An immediate consequence of Lemma 8.4.4 is that the ℓ_2^2 norm of a row or column of a PSD

matrix is at most $\frac{\|\mathbf{A}\|_F}{\mathbf{A}_{i,i}}$. Note, this precludes matrices where most of the mass in concentrated on a small number of rows or columns. Recall, we observe as input the matrix $\mathbf{A}+\mathbf{N}$ and our goal is to obtain a PCP for this in sublinear time and queries.

Musco and Musco [MM17] describe how to approximately compute the ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$ (if $\mathbf{A}$ is PSD) using a Nystrom approximation. [MW17c] use this method to compute the ridge leverage scores of $\mathbf{A}^{\frac{1}{2}}$ with $O(nk)$ queries, where $\mathbf{A} = \mathbf{A}^{\frac{1}{2}} \cdot \mathbf{A}^{\frac{1}{2}}$. However, these approaches do not apply when we perturb the input and it may no longer be PSD. Therefore, the best known construction by Cohen et. al. [CMM17] would require $\Omega(nnz(\mathbf{A}))$ time to compute approximate ridge-leverage scores of $\mathbf{A}$. Note, this does not use the structure that $\mathbf{A}$ has.

In contrast, we show that sampling columns proporitional to the diagonal entries suffices to obtain a PCP. Note, we only need to query the diagonal of $\mathbf{A}$ to compute the distribution over columns exactly. The main technical challenge here is to obtain the correct dependence on n and k and account for the perturbation to the input, given that our sampling probabilities are straightforward to compute and do not rely on spectral properties of $\mathbf{A}+\mathbf{N}$. Note, the following is a structural result and while we do not know $\mathbf{A}$, we can still show the following :

Theorem 179. *(Robust Spectral Bound.) Let $\mathbf{A}$ be an $n \times n$ PSD matrix and $\mathbf{N}$ be an arbitrary matrix such that $\|\mathbf{N}\|_F^2 \leqslant \eta\|\mathbf{A}\|_F^2$ and for all $j \in [n]$, $\|\mathbf{N}_{*,j}\|_2^2 \leqslant c\|\mathbf{A}_{*,j}\|_2^2$, for any fixed constant c. Let $\phi_{\max} = \max_j \mathbf{A}_{j,j}/(\mathbf{A}+\mathbf{N})_{j,j}$ and let $q = \{q_1, q_2 \dots q_n\}$ be a distribution over the columns of $\mathbf{A}+\mathbf{N}$ such that for all j, $q_j = (\mathbf{A}+\mathbf{N})_{j,j}/Tr[\mathbf{A}+\mathbf{N}]$ and let $t = O\left(\phi_{\max}\sqrt{n}k^2\log(n/\delta)/\epsilon^2\right)$. Then, construct a sampling matrix $\mathbf{T}$ that samples t columns of $\mathbf{A}+\mathbf{N}$ such that it samples column$(\mathbf{A}+\mathbf{N})_{*,j}$ with probability q_j and scales it by $1/\sqrt{tq_j}$. With probability at least $1-\delta$, for any rank-k orthogonal projection $\mathbf{X}$,*

$$\mathbf{A}\mathbf{A}^T - \left(\frac{\epsilon}{k}\right)\|\mathbf{A}\|_F^2\mathbf{I} \preceq \mathbf{A}\mathbf{T}(\mathbf{A}\mathbf{T})^T \preceq \mathbf{A}\mathbf{A}^T + \left(\frac{\epsilon}{k}\right)\|\mathbf{A}\|_F^2\mathbf{I}$$

Proof. First, we note that we cannot explicitly compute $\mathbf{A}\mathbf{T}$, but we can show that the sampling probabilities we have access to result in a PCP for $\mathbf{A}$. Let $\mathbf{Y} = \mathbf{A}\mathbf{T}(\mathbf{A}\mathbf{T})^T - \mathbf{A}\mathbf{A}^T$. For notational convenience let $\mathbf{A}_j = \mathbf{A}_{*,j}$. We can then write $\mathbf{Y} = \sum_{j\in[t]}\left(\mathbf{C}_{*,j}\mathbf{C}_{*,j}^T - \frac{1}{t}\mathbf{A}\mathbf{A}^T\right) = \sum_{j\in[t]}\mathbf{X}_j$, where $\mathbf{X}_j = \frac{1}{t}(\frac{1}{q_j}\mathbf{A}_j\mathbf{A}_j^T - \mathbf{A}\mathbf{A}^T)$ with probability q_j. We observe that $\mathbb{E}[\mathbf{X}_j] = \mathbb{E}[\mathbf{C}_{*,j}\mathbf{C}_{*,j}^T - \frac{1}{t}\mathbf{A}\mathbf{A}^T] = 0$, and therefore, $\mathbb{E}[\mathbf{Y}] = 0$. Next, we bound the operator norm of $\mathbf{Y}$. To this end, we use the Matrix Bernstein inequality, which in turn requires a bound on the operator norm of $\mathbf{X}_j$

and variance of $\mathbf{Y}$. Recall,

$$\begin{aligned}
\|\mathbf{X}_j\|_2 &= \left\| \frac{1}{tq_j}\mathbf{A}_j\mathbf{A}_j - \frac{1}{t}\mathbf{A}\mathbf{A}^T \right\|_2 \\
&\leqslant \frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]}{t(\mathbf{A}+\mathbf{N})_{j,j}}\|\mathbf{A}_j\|_2^2 + \frac{1}{t}\|\mathbf{A}\|_2^2 \\
&\leqslant \frac{2\mathrm{Tr}\left[\mathbf{A}\right] + |\mathrm{Tr}\left[\mathbf{N}\right]|}{t(\mathbf{A}+\mathbf{N})_{j,j}}\left((1+\eta)\|\mathbf{A}\|_2\mathbf{A}_{j,j}\right) \\
&\leqslant \frac{c(\mathrm{Tr}\left[\mathbf{A}\right] + |\mathrm{Tr}\left[\mathbf{N}\right]|)\|\mathbf{A}\|_2\phi_{\max}}{t} \\
&\leqslant \frac{c\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\|\mathbf{A}\|_2}{t}
\end{aligned} \tag{8.59}$$

where we use triangle inequality for operator norm to obtain the first inequality, triangle inequality up to a factor of 2 for ℓ_2^2 norms for the second inequality, $\|\mathbf{N}_j\|_2^2 \leqslant \eta\|\mathbf{A}_{*,j}\|_2^2$ and $\|\mathbf{A}_j\|_2^2 \leqslant \|\mathbf{A}\|_2 \cdot \mathbf{A}_{j,j}$ (from Lemma 8.4.4) for the third inequality and definition of $\phi_{\max}$ and $\eta = O(1)$ for the fourth. Finally, we relate the trace of $\mathbf{A}$ and $\mathbf{N}$ to their respective Frobenius norm using Cauchy-Schwarz:

$$\mathrm{Tr}\left[\mathbf{A}\right] = \sum_{i=1}^{n}\sigma_i(\mathbf{A}) \leqslant \sqrt{\sum_{i=1}^{n}\sigma_i^2(\mathbf{A})\cdot n} = \sqrt{n\|\mathbf{A}\|_F^2}$$

and

$$|\mathrm{Tr}\left[\mathbf{N}\right]| = \left|\sum_{i=1}^{n}\sigma_i(\mathbf{A})\right| \leqslant \sqrt{\sum_{i=1}^{n}\sigma_i^2(\mathbf{N})\cdot n} = \sqrt{n\|\mathbf{N}\|_F^2} \leqslant \sqrt{n\eta}\|\mathbf{A}\|_F$$

where the last inequality follows from $\|\mathbf{N}\|_F \leqslant \sqrt{\eta}\|\mathbf{A}\|_F$. Next, we bound $\mathbf{Var}\mathbf{Y} \leqslant \mathbb{E}\left[\mathbf{Y}^2\right]$.

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{Y}^2\right] &= t\,\mathbb{E}\left[\left((\mathbf{AT})_{*,j}(\mathbf{AT})_{*,j}^T - \frac{1}{t}\mathbf{A}\mathbf{A}^T\right)^2\right] \\
&= t\,\mathbb{E}\left[\left((\mathbf{AT})_{*,j}(\mathbf{AT})_{*,j}^T\right)^2 + \frac{1}{t^2}\left(\mathbf{A}\mathbf{A}^T\right)^2 - \frac{2}{t}(\mathbf{AT})_{*,j}(\mathbf{AT})_{*,j}^T\mathbf{A}\mathbf{A}^T\right] \\
&= \frac{1}{t}\left(\sum_{j\in[n]}\frac{(\mathbf{A}_j\mathbf{A}_j^T)^2}{q_j} + (\mathbf{A}\mathbf{A}^T)^2 - \sum_{j\in[n]} 2\mathbf{A}_j\mathbf{A}_j^T\mathbf{A}\mathbf{A}^T\right) \\
&\preceq \frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]}{t\mathbf{A}_{j,j}}\left(\sum_{j\in[n]}(\mathbf{A}_j\mathbf{A}_j^T)^2\right) \\
&\preceq \frac{c\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\|\mathbf{A}\|_2}{t}\mathbf{A}\mathbf{A}^T
\end{aligned}
\tag{8.60}
$$

where we use linearity of expectation, $(\mathbf{A}\mathbf{A}^T)^2 \succeq 0$ and $\|\mathbf{A}_{j,*}\|_2^2 \leqslant \|\mathbf{A}\|_2 \cdot \mathbf{A}_{j,j}$. Applying the Matrix Bernstein inequality,

$$
\begin{aligned}
\mathbf{Pr}\left[\right] \|\mathbf{Y}\|_2 \geqslant \epsilon\|\mathbf{A}\|_F^2 &\leqslant 2n\exp\left(-\frac{\epsilon^2\|\mathbf{A}\|_F^4}{\frac{c\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\|\mathbf{A}\|_2^3}{t} + \frac{2\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\|\mathbf{A}\|_2\left(\epsilon\|\mathbf{A}\|_F^2\right)}{3t}}\right) \\
&\leqslant 2n\exp\left(-\frac{\epsilon^2 t}{c\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\|\mathbf{A}\|_2}\left(\frac{\|\mathbf{A}\|_F^4}{\|\mathbf{A}\|_2^2 + \epsilon\|\mathbf{A}\|_F^2}\right)\right) \\
&\leqslant 2n\exp\left(-\frac{\epsilon^2 t}{c'\phi_{\max}\sqrt{n}}\right) \\
&\leqslant \delta/2
\end{aligned}
$$

where the last inequality follows from setting $t = O(\phi_{\max}\sqrt{n}\log(n/\delta)/\epsilon^2)$. To yield the claim, we set $\epsilon = \epsilon/k$. □

We use the above spectral bound to show that sampling proportional to diagonal entries preserves the projection cost of the columns of $\mathbf{A}$ on to any k-dimensional subspace up to an additive $(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$.

Theorem 180. *(Column Projection-Cost Preservation.) Given $\mathbf{A}+\mathbf{N}$, where $\mathbf{A}$ is an $n \times n$ PSD matrix and $\mathbf{N}$ is an arbitrary noise matrix as defined above, $k \in \mathbb{Z}$ and $\epsilon > \eta > 0$, let $q = \{q_1, q_2 \ldots q_n\}$ be a probability distribution over the columns of $\mathbf{A}+\mathbf{N}$ such that $q_j = \frac{(\mathbf{A}+\mathbf{N})_{j,j}}{Tr[\mathbf{A}+\mathbf{N}]}$. Let $t = O\left(\phi_{\max}\sqrt{n}k^2\log(\frac{n}{\delta})/\epsilon^2\right)$. Then, construct $\mathbf{C}$ using t columns of $\mathbf{A}+\mathbf{N}$ and set each one to $\frac{(\mathbf{A}+\mathbf{N})_{*,j}}{\sqrt{tq_j}}$ with probability q_j. With probability at least $1-c$, for any rank-k orthogonal*

projection $\mathbf{X}$*,*

$$\|\mathbf{C} - \mathbf{X}\mathbf{C}\|_F^2 = \|\mathbf{A} - \mathbf{X}\mathbf{A}\|_F^2 \pm (\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$$

for a fixed constant c*.*

Proof. Here, the matrix $\mathbf{C}$ is actually a matrix we can compute. Observe that we can relate $\mathbf{C}$ to the sampling matrix $\mathbf{T}$ as defined in Theorem 179 as $\mathbf{C} = (\mathbf{A} + \mathbf{N})\mathbf{T}$. We follow the proof strategy of the relative error guarantees in [CMM17] and additive error guarantees in [BW18] but note, our spectral bounds from Theorem 179 apply to matrices that we do not actually compute. Observe, $\|\mathbf{A} - \mathbf{X}\mathbf{A}\|_F^2 = \text{Tr}\left[(\mathbb{I} - \mathbf{X})\mathbf{A}\mathbf{A}^T(\mathbb{I} - \mathbf{X})\right]$. Then,

$$\begin{aligned}
\text{Tr}\left[(\mathbb{I} - \mathbf{X})\mathbf{A}\mathbf{A}^T(\mathbb{I} - \mathbf{X})\right] &= \text{Tr}\left[\mathbf{A}\mathbf{A}^T\right] + \text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}\right] - \text{Tr}\left[\mathbf{A}\mathbf{A}^T\mathbf{X}\right] - \text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\right] \\
&= \text{Tr}\left[\mathbf{A}\mathbf{A}^T\right] + \text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}\right] - \text{Tr}\left[\mathbf{A}\mathbf{A}^T\mathbf{X}\mathbf{X}\right] - \text{Tr}\left[\mathbf{X}\mathbf{X}\mathbf{A}\mathbf{A}^T\right] \\
&= \text{Tr}\left[\mathbf{A}\mathbf{A}^T\right] + \text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}\right] - \text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}\right] - \text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}\right] \\
&= \text{Tr}\left[\mathbf{A}\mathbf{A}^T\right] - \text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}\right]
\end{aligned} \tag{8.61}$$

where we used the fact that for any projection matrix $\mathbf{X} = \mathbf{X}^2$ in addition to the cyclic property of the trace. Similarly,

$$\|\mathbf{C} - \mathbf{X}\mathbf{C}\|_F^2 = \text{Tr}\left[(\mathbb{I} - \mathbf{X})\mathbf{C}\mathbf{C}^T(\mathbb{I} - \mathbf{X})\right] = \text{Tr}\left[\mathbf{C}\mathbf{C}^T\right] - \text{Tr}\left[\mathbf{X}\mathbf{C}\mathbf{C}^T\mathbf{X}\right] \tag{8.62}$$

We first relate $\text{Tr}\left[\mathbf{A}\mathbf{A}^T\right]$ and $\text{Tr}\left[\mathbf{C}\mathbf{C}^T\right]$. Recall,

$$\mathbb{E}\left[\text{Tr}\left[\mathbf{C}\mathbf{C}^T\right]\right] = \mathbb{E}\left[\|\mathbf{C}\|_F^2\right] = \|\mathbf{A} + \mathbf{N}\|_F^2 \leqslant \text{Tr}\left[\mathbf{A}\mathbf{A}^T\right] + 2\sqrt{\eta}\|\mathbf{A}\|_F^2$$

Using a scalar Chernoff bound, we show that with probability at least $1 - 1/\text{poly}(n)$, $\|\mathbf{C}\|_F^2 = (1 \pm \epsilon)\|\mathbf{A} + \mathbf{N}\|_F^2$. This is equivalent to $|\|\mathbf{C}\|_F^2 - \|\mathbf{A} + \mathbf{N}\|_F^2| \leqslant \epsilon\|\mathbf{A} + \mathbf{N}\|_F^2$. Observe, for all $j \in [t]$, $\mathbf{C}_{*,j} = \frac{1}{\sqrt{q_{j'}t}}(\mathbf{A} + \mathbf{N})_{*,j'}$ for some $j' \in [n]$. Then,

$$\begin{aligned}
\|\mathbf{C}_{*,j}\|_2^2 = \frac{1}{q_{j'}t}\|(\mathbf{A} + \mathbf{N})_{*,j'}\|_2^2 &= \frac{\text{Tr}\left[\mathbf{A} + \mathbf{N}\right]\epsilon^2}{\phi_{\max}\sqrt{n}k\log(n/\delta)(\mathbf{A} + \mathbf{N})_{j',j'}}\|(\mathbf{A} + \mathbf{N})_{*,j'}\|_2^2 \\
&\leqslant \frac{c\sqrt{n}\|\mathbf{A}\|_F\epsilon^2}{\sqrt{n}\log(n/\delta)}\|\mathbf{A}\|_2 \\
&\leqslant \frac{c\epsilon^2}{k\log(n/\delta)}\|\mathbf{A} + \mathbf{N}\|_F^2
\end{aligned} \tag{8.63}$$

where we use $\text{Tr}\left[\mathbf{A}\right] \leqslant \sqrt{n}\|\mathbf{A}\|_F$, $\text{Tr}\left[\mathbf{N}\right] \leqslant \sqrt{\eta n}\|\mathbf{A}\|_F$ and $t = O(\phi_{\max}\sqrt{n}k\log(n/\delta)/\epsilon^2)$. Therefore, $\frac{k\log(n/\delta)}{\epsilon^2\|\mathbf{A}\|_F^2}\|\mathbf{C}_{*,j}\|_2^2 \in [0,1]$. By a Chernoff bound,

$$\begin{aligned}\mathbf{Pr}\left[\right] \|\mathbf{C}\|_F^2 \geqslant (1+2\epsilon)\|\mathbf{A}+\mathbf{N}\|_F^2 = \mathbf{Pr}\left[\right] \frac{k\log(n/\delta)}{\epsilon^2\|\mathbf{A}+\mathbf{N}\|_F^2}\|\mathbf{C}\|_F^2 \geqslant \frac{k\log(n/\delta)}{\epsilon^2}(1+\epsilon) \\ \leqslant \exp\left(-\frac{k\epsilon^2\log(n/\delta)}{\epsilon^2}\right) \\ \leqslant \frac{\delta}{2}\end{aligned} \tag{8.64}$$

We can repeat the above argument to lower bound $\|\mathbf{C}\|_F^2$. Therefore, with probability $1-\delta$, we have

$$|\|\mathbf{C}\|_F^2 - \|\mathbf{A}+\mathbf{N}\|_F^2| \leqslant \epsilon\|\mathbf{A}+\mathbf{N}\|_F^2$$

Here, we can upper bound this by observing $\|\mathbf{A}+\mathbf{N}\|_F^2 \leqslant \|\|_F^2 + \|\mathbf{N}\|_F^2 + 2\langle\mathbf{A},\mathbf{N}\rangle \leqslant \|\mathbf{A}\|_F^2 + 3\sqrt{\eta}\|\mathbf{A}\|_F^2$. Therefore,

$$|\|\mathbf{C}\|_F^2 - \|\mathbf{A}\|_F^2| \leqslant \epsilon\|\mathbf{A}\|_F^2 + (1+\epsilon)\sqrt{\eta}\|\mathbf{A}\|_F^2 \leqslant (\epsilon+2\sqrt{\eta})\|\mathbf{A}\|_F^2 \tag{8.65}$$

Next, we relate $\text{Tr}\left[\mathbf{X}\mathbf{C}\mathbf{C}^T\mathbf{X}\right]$ and $\text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}\right]$. First, we observe

$$\mathbf{C}\mathbf{C}^T = (\mathbf{A}\mathbf{T}+\mathbf{N}\mathbf{T})(\mathbf{A}\mathbf{T}+\mathbf{N}\mathbf{T})^T = (\mathbf{A}\mathbf{T})(\mathbf{A}\mathbf{T})^T+(\mathbf{A}\mathbf{T})(\mathbf{N}\mathbf{T})^T+(\mathbf{N}\mathbf{T})(\mathbf{A}\mathbf{T})^T+(\mathbf{N}\mathbf{T})(\mathbf{N}\mathbf{T})^T \tag{8.66}$$

We begin by first bounding $\text{Tr}\left[\mathbf{X}(\mathbf{A}\mathbf{T})(\mathbf{A}\mathbf{T})^T\mathbf{X}\right]$. Observe, $\mathbf{X}$ is a rank k projection matrix and we can represent it as $\mathbf{Z}\mathbf{Z}^T$, where $\mathbf{Z} \in \mathbb{R}^{n\times k}$ and has orthonormal columns. By the cyclic property of the trace, we have

$$\text{Tr}\left[\mathbf{Z}\mathbf{Z}^T(\mathbf{A}\mathbf{T})(\mathbf{A}\mathbf{T})^T\mathbf{Z}\mathbf{Z}^T\right] = \text{Tr}\left[\mathbf{Z}^T(\mathbf{A}\mathbf{T})(\mathbf{A}\mathbf{T})^T\mathbf{Z}\right] = \sum_{j\in[k]}\mathbf{Z}_{*,j}^T(\mathbf{A}\mathbf{T})(\mathbf{A}\mathbf{T})^T\mathbf{Z}_{*,j}$$

Similarly, $\text{Tr}\left[\mathbf{Z}\mathbf{Z}^T\mathbf{A}\mathbf{A}^T\mathbf{Z}\mathbf{Z}^T\right] = \sum_{j\in[k]}\mathbf{Z}_{*,j}^T\mathbf{A}\mathbf{A}^T\mathbf{Z}_{*,j}$. By Theorem 179 , we have

$$\sum_{j\in[k]}\left(\mathbf{Z}_{*,j}^T\mathbf{A}\mathbf{A}^T\mathbf{Z}_{*,j}-\left(\frac{\epsilon}{k}\right)\|\mathbf{A}\|_F^2\mathbf{Z}_{*,j}^T\mathbf{I}\mathbf{Z}_{*,j}\right)\leqslant\sum_{j\in[k]}\left(\mathbf{Z}_{*,j}^T(\mathbf{A}\mathbf{T})(\mathbf{A}\mathbf{T})^T\mathbf{Z}_{*,j}\right)$$
$$\leqslant\sum_{j\in[k]}\left(\mathbf{Z}_{*,j}^T\mathbf{A}\mathbf{A}^T\mathbf{Z}_{*,j}+\left(\frac{\epsilon}{k}\right)\|\mathbf{A}\|_F^2\mathbf{Z}_{*,j}^T\mathbf{I}\mathbf{Z}_{*,j}\right) \tag{8.67}$$

Since $\mathbf{Z}_{*,j}^T\mathbf{Z}_{*,j}=1$ and $\text{Tr}\left[\mathbf{Z}^T\mathbf{A}\mathbf{A}^T\mathbf{Z}\right]=\text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}\right]$, we have

$$\text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}\right]-\epsilon\|\mathbf{A}\|_F^2\leqslant\text{Tr}\left[\mathbf{X}(\mathbf{A}\mathbf{T})(\mathbf{A}\mathbf{T})^T\mathbf{X}\right]\leqslant\text{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{A}^T\mathbf{X}\right]+\epsilon\|\mathbf{A}\|_F^2 \tag{8.68}$$

Next, we focus on $\text{Tr}\left[\mathbf{X}(\mathbf{N}\mathbf{T})(\mathbf{N}\mathbf{T})^T\mathbf{X}\right]=\|\mathbf{X}\mathbf{N}\mathbf{T}\|_F^2$. Observe, since $\mathbf{T}$ is an unbiased estimator of Frobenius norm, by Markov's inequality we can show with probability at least $1-c$, $\|\mathbf{X}\mathbf{N}\mathbf{T}\|_F=c\|\mathbf{N}\|_F=O(\sqrt{\eta})\|\mathbf{A}\|_F$. Therefore, we can upper bound $\text{Tr}\left[\mathbf{X}(\mathbf{N}\mathbf{T})(\mathbf{N}\mathbf{T})^T\mathbf{X}\right]$ by $O(\eta)\|\mathbf{A}\|_F^2$. Now, we focus on the cross terms. By Cauchy-Schwartz, and a Markov bound, with probability at least $1-c$,

$$\text{Tr}\left[\mathbf{X}(\mathbf{A}\mathbf{T})(\mathbf{N}\mathbf{T})^T\mathbf{X}\right]\leqslant\|\mathbf{A}\mathbf{T}\|_F\cdot\|\mathbf{N}\mathbf{T}\|_F\leqslant O(\sqrt{\eta})\|\mathbf{A}\|_F^2 \tag{8.69}$$

Combining equations 8.65, 8.68, 8.69 and union bounding over the success of the random events, with probability $1-c$,

$$\|\mathbf{A}-\mathbf{X}\mathbf{A}\|_F^2-O(\epsilon+\sqrt{\eta})\|\mathbf{A}\|_F^2\leqslant\|\mathbf{C}-\mathbf{X}\mathbf{C}\|_F^2\leqslant\|\mathbf{A}-\mathbf{X}\mathbf{A}\|_F^2+O(\epsilon+\sqrt{\eta})\|\mathbf{A}\|_F^2$$

□

Robust Row Projection Cost Preserving Sketches. We now extend the diagonal sampling algorithm to construct a row projection cost preserving sketch for the matrix $\mathbf{C}$. We note that following the construction for $\mathbf{A}$ does not immediately give a row PCP for $\mathbf{C}$ since $\mathbf{C}$ is no longer PSD or even square matrix. Here, all previous approaches to construct a PCP with sublinear queries hit a roadblock, since the matrix $\mathbf{C}$ need not have any well defined structure apart from being a scaled subset of the columns of a PSD matrix. However, we show that sampling rows of $\mathbf{C}$ proportional to the diagonal entries of $\mathbf{A}$ results in a row PCP.

We begin by relating the row norms of $\mathbf{C}$ to the row norms of $\mathbf{A}$. Note, we do not expect to obtain concentration here, since such a sampling procedure would then help us estimate row

norms of $\mathbf{A}$ up to a constant and we would be done by using [FKV04b]. Therefore, we obtain the following one-sided guarantee:

Lemma 8.4.5. *Let $\mathbf{AT} \in \mathbb{R}^{n\times t}$ be a column projection-cost preserving sketch for $\mathbf{A}$ as described in Theorem 179. For all $i \in [n]$, with probability at least $1 - 1/n^c$,*

$$\|(\mathbf{AT})_{i,*}\|_2^2 \leqslant O\left(\log(n)\max\left\{\|\mathbf{A}_{i,*}\|_2^2, \frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\right\}\right)$$

where c is a fixed constant.

Proof. Observe that $\|(\mathbf{AT})_{i,*}\|_2^2 = \sum_{j\in[t]}(\mathbf{AT})_{i,j}^2$, where $(\mathbf{AT})_{i,j}^2 = \frac{\mathrm{Tr}[\mathbf{A}+\mathbf{N}]}{t(\mathbf{A}+\mathbf{N})_{j,j}}\mathbf{A}_{i,j}^2$ with probability $\frac{(\mathbf{A}+\mathbf{N})_{j,j}}{\mathrm{Tr}[\mathbf{A}+\mathbf{N}]}$. Then, $\mathbb{E}[\|(\mathbf{AT})_{i,*}\|_2^2] = \sum_{i=1}^n \mathbf{A}_{i,j}^2 = \|\mathbf{A}_{i,*}\|_2^2$. Next, we compute the variance of $\|(\mathbf{AT})_{i,*}\|_2^2$. $\mathbf{Var}\left[\|(\mathbf{AT})_{i,*}\|_2^2\right] = t\mathbf{Var}\left[(\mathbf{AT})_{i,j}^2\right] \leqslant \mathbb{E}\left[(\mathbf{AT})_{i,j}^4\right]$. Then,

$$\begin{aligned}
t\,\mathbb{E}\left[(\mathbf{AT})_{i,j}^4\right] = \sum_{j\in[n]}\frac{1}{tq_j}\mathbf{A}_{i,j}^4 &\leqslant \sum_{j\in[n]}\frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]}{t(\mathbf{A}+\mathbf{N})_{j,j}}\mathbf{A}_{i,j}^2\mathbf{A}_{i,i}\mathbf{A}_{j,j} \\
&\leqslant \frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]\phi_{\max}\mathbf{A}_{i,i}}{t}\|\mathbf{A}_{i,*}\|_2^2 \\
&\leqslant \left(\frac{2\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\right)^2 + \|\mathbf{A}_{i,*}\|_2^4 \qquad \text{[AM-GM]}
\end{aligned}$$

where we use $\mathbf{A}_{i,j}^2 \leqslant \mathbf{A}_{i,i}\mathbf{A}_{j,j}$, which follows from applying Cauchy-Schwarz to $\langle\mathbf{A}_{i,*}^{1/2}, \mathbf{A}_{j,*}^{1/2}\rangle$, i.e.,

$$\mathbf{A}_{i,j}^2 = \langle\mathbf{A}_{i,*}^{1/2}, \mathbf{A}_{j,*}^{1/2}\rangle^2 \leqslant \|\mathbf{A}_{i,*}^{1/2}\|_2^2\|\mathbf{A}_{j,*}^{1/2}\|_2^2 = \mathbf{A}_{i,i}\mathbf{A}_{j,j}$$

Similarly, we bound

$$(\mathbf{AT})_{i,j}^2 = \frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]}{t(\mathbf{A}+\mathbf{N})_{j,j}}\mathbf{A}_{i,j}^2 \leqslant \frac{2\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F}{t}\mathbf{A}_{i,i}$$

Applying Bernstein's inequality,

$$\begin{aligned}
&\mathbf{Pr}\left[\right] |\|(\mathbf{AT})_{i,*}\|_2^2 - \|\mathbf{A}_{i,*}\|_2^2 \geqslant \gamma \max\left\{\|\mathbf{A}_{i,*}\|_2^2, \frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\right\} \\
&\leqslant 2\exp\left(-\frac{\gamma^2 \max\left\{\|\mathbf{A}_{i,*}\|_2^4, \left(\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\right)^2\right\}}{\left(\frac{2\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\right)^2 + \|\mathbf{A}_{i,*}\|_2^4 + \gamma\max\left\{\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}\|\mathbf{A}_{i,*}\|_2^2}{t}, \left(\frac{\mathrm{Tr}[\mathbf{A}]\mathbf{A}_{i,i}}{t}\right)^2\right\}}\right) \\
&\leqslant 2\exp\left(-\frac{\gamma\max\left\{\|\mathbf{A}_{i,*}\|_2^4, \left(\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\right)^2\right\}}{c'\left(\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\right)^2 + c'\|\mathbf{A}_{i,*}\|_2^4}\right)
\end{aligned}$$

where $\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\|\mathbf{A}_{i,*}\|_2^2 \leqslant \left(\frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i,i}}{t}\right)^2 + \|\mathbf{A}_{i,*}\|_2^4$ follows from the AM-GM inequality. Setting $\gamma = \Omega(\log(n))$ completes the proof.

□

To construct a row projection cost preserving sketch of $\mathbf{C}$, we sample t rows of $\mathbf{C}$ proportional to the corresponding diagonal entries of $\mathbf{A}$. Formally, we consider a probability distribution, $p = \{p_1, p_2, \ldots p_n\}$, over the rows of $\mathbf{C}$ such that $p_i = \frac{\mathbf{A}_{i,i}}{\mathrm{Tr}[\mathbf{A}]}$. Let $\mathbf{R}$ be a $t \times t$ matrix where each row of $\mathbf{R}$ is set to $\frac{1}{\sqrt{tp_i}}\mathbf{C}_{i,*}$ with probability p_i. As before, $\mathbf{R}$ can be represented as $\mathbf{SC} = \mathbf{S}(\mathbf{AT}+\mathbf{NT})$. We first obtain a spectral guarantee for $\mathbf{SAT}$, while we cannot actually compute this.

Theorem 181. *(Spectral Bounds.) Let* $\mathbf{AT}$ *be an* $n \times t$ *matrix constructed as shown in Theorem 179. Let* $p = \{p_1, p_2 \ldots p_n\}$ *be a probability distribution over the rows of* $\mathbf{AT}$ *such that* $p_i = \frac{(\mathbf{A}+\mathbf{N})_{i,i}}{Tr[\mathbf{A}+\mathbf{N}]}$. *Let* $t = O\left(\frac{\sqrt{n}k^2}{\epsilon^2}\log(\frac{n}{\delta})\right)$. *Construct a sampling matrix* $\mathbf{S}$ *that samples* t *rows of* $\mathbf{AT}$ *such that row* $(\mathbf{AT})_{i,*}$ *is picked with with probability* p_i *and scaled by* $\frac{1}{\sqrt{tp_i}}$. *Then, with probability at least* $1-\delta$,

$$(\mathbf{AT})^T(\mathbf{AT}) - \frac{\epsilon}{k}\|\mathbf{A}\|_F^2\mathbf{I} \preceq (\mathbf{SAT})^T(\mathbf{SAT}) \preceq (\mathbf{AT})^T(\mathbf{AT}) + \frac{\epsilon}{k}\|\mathbf{A}\|_F^2\mathbf{I}$$

Proof. Let $\mathbf{Y} = (\mathbf{SAT})^T(\mathbf{SAT}) - (\mathbf{AT})^T(\mathbf{AT})$. For notational convenience let $(\mathbf{AT})_i = (\mathbf{AT})_{i,*}$ and $(\mathbf{SAT})_i = (\mathbf{SAT})_{i,*}$. We can then write $\mathbf{Y} = \sum_{i\in[t]}\left((\mathbf{SAT})_i^T(\mathbf{SAT})_i - \frac{1}{t}(\mathbf{AT})^T(\mathbf{AT})\right) = \sum_{i\in[t]}\mathbf{X}_i$, where $\mathbf{X}_i = \frac{1}{t}\left(\frac{1}{p_i}(\mathbf{AT})_i^T(\mathbf{AT})_i - (\mathbf{AT})^T(\mathbf{AT})\right)$ with probability p_i. We observe that $\mathbb{E}[\mathbf{X}_i] = \mathbb{E}[(\mathbf{SAT})_i^T(\mathbf{SAT})_i - \frac{1}{t}(\mathbf{AT})^T(\mathbf{AT})] = \sum_i \frac{p_i}{p_i}(\mathbf{AT})_i^T(\mathbf{AT})_i - (\mathbf{AT})^T(\mathbf{AT}) = 0$, and therefore, $\mathbb{E}[\mathbf{Y}] = 0$. Next, we bound the operator norm of $\mathbf{Y}$. To this end, we use the Matrix Bernstein inequality, which in turn requires a bound on the operator norm of $\mathbf{X}_i$ and variance of

$\mathbf{Y}$. Recall, for some $i' \in [n]$

$$\begin{aligned}
\|\mathbf{X}_i\|_2 &= \left\| \frac{1}{tp_{i'}} (\mathbf{AT})_{i'}^T (\mathbf{AT})_{i'} - \frac{1}{t} (\mathbf{AT})^T (\mathbf{AT}) \right\|_2 \\
&\leqslant \frac{1}{tp_{i'}} \|(\mathbf{AT})_{i'}^T (\mathbf{AT})_{i'}\|_2 + \frac{1}{t} \|(\mathbf{AT})^T (\mathbf{AT})\|_2 \\
&= \frac{\|(\mathbf{AT})_{i'}\|_2^2}{tp_{i'}} + \frac{\|(\mathbf{AT})\|_2^2}{t} \\
&\leqslant \frac{\log(n)}{tp_{i'}} \max \left\{ \|\mathbf{A}_{i',*}\|_2^2, \frac{\phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \mathbf{A}_{i,i}}{t} \right\} + \frac{\|(\mathbf{AT})\|_2^2}{t} && \text{[by Lemma 8.4.5]} \\
&\leqslant \frac{\log(n)}{t} \max \left\{ \phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \|\mathbf{A}\|_2, \frac{(\phi_{\max} \sqrt{n} \|\mathbf{A}\|_F)^2}{t}, \|(\mathbf{AT})\|_2^2 \right\} && \text{[by Lemma 8.4.4]} \\
&\leqslant \frac{\phi_{\max} \sqrt{n} \log(n) \|\mathbf{A}\|_F^2}{t} \left(1 + \frac{\sqrt{n}}{t} \right) \leqslant \frac{2\phi_{\max} \sqrt{n} \log(n) \|\mathbf{A}\|_F^2}{t}
\end{aligned} \tag{8.70}$$

where the last inequality uses that $t = \Omega(\sqrt{n})$. Next, we bound $\mathbf{Var}\mathbf{Y} \leqslant \mathbb{E}\left[\mathbf{Y}^2\right]$ as follows

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Y}^2\right] &= t \left(\sum_{i \in [n]} \frac{p_i}{t^2 p_i^2} ((\mathbf{AT})_i^T (\mathbf{AT})_i)^2 + \frac{1}{t^2} ((\mathbf{AT})^T (\mathbf{AT}))^2 - \sum_{i \in [n]} \frac{2p_i}{p_i t^2} (\mathbf{AT})_i^T (\mathbf{AT})_i (\mathbf{AT})^T (\mathbf{AT}) \right) \\
&= \frac{1}{t} \left(\sum_{i \in [n]} \frac{((\mathbf{AT})_i^T (\mathbf{AT})_i)^2}{p_i} + ((\mathbf{AT})^T (\mathbf{AT}))^2 - \sum_{i \in [n]} 2 (\mathbf{AT})_i^T (\mathbf{AT})_i (\mathbf{AT})^T (\mathbf{AT}) \right) \\
&\preceq \frac{1}{t} \left(\sum_{i \in [n]} \frac{((\mathbf{AT})_i^T (\mathbf{AT})_i)^2}{p_i} \right) \\
&\preceq \frac{\log(n)}{t} \max \left\{ \phi_{\max} \sqrt{n} \|\mathbf{A}\|_F \|\mathbf{A}\|_2, \frac{(\phi_{\max} \sqrt{n} \|\mathbf{A}\|_F)^2}{t} \right\} \left(\sum_{i \in [n]} (\mathbf{AT})_i^T (\mathbf{AT})_i \right) \\
&\preceq \frac{c \log(n) \sqrt{n} \|\mathbf{A}\|_F^2 \|(\mathbf{AT})\|_2^2}{t} \mathbf{I}_{n \times n}
\end{aligned} \tag{8.71}$$

where we again use Lemma 8.4.4 and Theorem 180. Observe,

Applying the Matrix Bernstein inequality with equations 8.70 and 8.71

$$\begin{aligned}\mathbf{Pr}\left[\right]\|\mathbf{Y}\|_2 \geqslant \epsilon\|\mathbf{A}\|_F^2 &\leqslant 2n\exp\left(-\frac{\epsilon^2\|\mathbf{A}\|_F^4}{\frac{c\log(n)\sqrt{n}\phi_{\max}\|\mathbf{A}\|_F^2\|(\mathbf{AT})\|_2^2}{t}+\frac{\epsilon\sqrt{n}\log(n)\phi_{\max}}{3t}\|\mathbf{A}\|_F^4}\right)\\ &\leqslant 2n\exp\left(-\frac{\epsilon^2 t}{c'\phi_{\max}\sqrt{n}\log(n)}\right)\end{aligned} \tag{8.72}$$

where the second inequality uses Theorem 179, to conclude that with probability at least $1-\delta/2$ $\|\mathbf{AT}\|_2^2 \leqslant \|\mathbf{A}\|_2^2+\epsilon/k\|\mathbf{A}\|_F^2 \leqslant O(\|\mathbf{A}\|_F^2)$. Therefore, it suffices to set $t = \frac{c'\phi_{\max}\sqrt{n}\log(n))}{\epsilon^2}\log(n/\delta)$, to bound the above probability by $\delta/2$. Union bounding over the error for both PCPs, and setting $\epsilon = \epsilon/k$, we can conclude that with probability at least $1-\delta$,

$$(\mathbf{AT})^T(\mathbf{AT}) - \frac{\epsilon}{k}\|\mathbf{A}\|_F^2\mathbf{I} \preceq (\mathbf{SAT})^T(\mathbf{SAT}) \preceq (\mathbf{AT})^T(\mathbf{AT}) + \frac{\epsilon}{k}\|\mathbf{A}\|_F^2\mathbf{I}$$

when $t = \Omega\left(\frac{\phi_{\max}\sqrt{n}k^2}{\epsilon^2}\right)$.

□

We use the spectral bound from Theorem 181 to obtain a row projection-cost preservation guarantee. We follow the same proof strategy as Theorem 180, while requiring modified version of the scalar Chernoff bound. We do away with the head-tail split from [CMM17],[MW17c] and [BW18] and analyze the projection-cost guarantee directly. This enables us to obtain a better ϵ dependence than [MW17c] and [BW18]. Note, our ϵ dependence matches that of [CMM17] but our row projection cost preserving sketch can be computed in sub-linear time, albeit for PSD matrices.

Theorem 182. *(Row Projection-Cost Preservation.) Given as input $\mathbf{A}+\mathbf{N}$ let $\mathbf{C}$ be an $n \times t$ matrix as defined in Theorem 180 such that $\mathbf{C} = \mathbf{AT}+\mathbf{NT}$. Let $p = \{p_1, p_2 \ldots p_n\}$ be a probability distribution over the rows of $\mathbf{C}$ such that $p_j = \frac{(\mathbf{A}+\mathbf{N})_{j,j}}{Tr[\mathbf{A}+\mathbf{N}]}$. Let $t = O\left(\frac{\phi_{max}\sqrt{n}k^2\log^2(n)}{\epsilon^2}\right)$. Then, construct $\mathbf{R}$ using t rows of $\mathbf{C}$ and set each one to $\frac{\mathbf{C}_{i,*}}{\sqrt{tp_i}}$ with probability p_i. With probability at least $1-c$, for any rank-k orthogonal projection $\mathbf{X}$,*

$$\|\mathbf{R}-\mathbf{RX}\|_F^2 = \|\mathbf{C}-\mathbf{CX}\|_F^2 \pm O(\epsilon+\sqrt{\eta})\|\mathbf{A}\|_F^2$$

for a fixed constant c.

Proof. Note, $\mathbf{R} = \mathbf{SC} = \mathbf{SAT}+\mathbf{SNT}$, where $\mathbf{S}$ and $\mathbf{T}$ are the corresponding sampling matrices.

Observe, $\|\mathbf{C} - \mathbf{CX}\|_F^2 = \text{Tr}\left[(\mathbb{I} - \mathbf{X})\mathbf{C}^T\mathbf{C}(\mathbb{I} - \mathbf{X})\right]$. Then,

$$\begin{aligned}
\text{Tr}\left[(\mathbb{I} - \mathbf{X})\mathbf{C}^T\mathbf{C}(\mathbb{I} - \mathbf{X})\right] &= \text{Tr}\left[\mathbf{C}^T\mathbf{C}\right] + \text{Tr}\left[\mathbf{X}\mathbf{C}^T\mathbf{C}\mathbf{X}\right] - \text{Tr}\left[\mathbf{C}^T\mathbf{C}\mathbf{X}\right] - \text{Tr}\left[\mathbf{X}\mathbf{C}^T\mathbf{C}\right] \\
&= \text{Tr}\left[\mathbf{C}^T\mathbf{C}^T\right] + \text{Tr}\left[\mathbf{X}^T\mathbf{C}^T\mathbf{C}\mathbf{X}\right] - \text{Tr}\left[\mathbf{C}^T\mathbf{C}\mathbf{X}\mathbf{X}\right] - \text{Tr}\left[\mathbf{X}\mathbf{X}\mathbf{C}^T\mathbf{C}\right] \\
&= \text{Tr}\left[\mathbf{C}^T\mathbf{C}\right] + \text{Tr}\left[\mathbf{X}\mathbf{C}^T\mathbf{C}\mathbf{X}\right] - \text{Tr}\left[\mathbf{X}\mathbf{C}^T\mathbf{C}\mathbf{X}\right] - \text{Tr}\left[\mathbf{X}\mathbf{C}^T\mathbf{C}\mathbf{X}\right] \\
&= \text{Tr}\left[\mathbf{C}^T\mathbf{C}\right] - \text{Tr}\left[\mathbf{X}\mathbf{C}^T\mathbf{C}\mathbf{X}\right] \\
&= \text{Tr}\left[\mathbf{C}^T\mathbf{C}\right] - \text{Tr}\left[\mathbf{X}(\mathbf{AT} + \mathbf{NT})^T(\mathbf{AT} + \mathbf{NT})\mathbf{X}\right]
\end{aligned} \tag{8.73}$$

where we used the fact that for any projection matrix $X = X^2$ in addition to the cyclic property of the trace. Here, for analyzing the cross and tail terms, we observe that with probability $1 - c$, $\|\mathbf{X}(\mathbf{AT})^T\|_F \leqslant O(1)\|\mathbf{A}\|_F$ and $\|\mathbf{X}(\mathbf{NT})^T\|_F^2 \leqslant O(\eta)\|\mathbf{A}\|_F^2$. Therefore,

$$\text{Tr}\left[\mathbf{X}(\mathbf{AT} + \mathbf{NT})^T(\mathbf{AT} + \mathbf{NT})\mathbf{X}\right] = \text{Tr}\left[\mathbf{X}(\mathbf{AT})^T\mathbf{ATX}\right] \pm O(\sqrt{\eta})\|\mathbf{A}\|_F^2 \tag{8.74}$$

Similarly,

$$\begin{aligned}
\|\mathbf{R} - \mathbf{RX}\|_F^2 = \text{Tr}\left[(\mathbb{I} - \mathbf{X})\mathbf{R}^T\mathbf{R}(\mathbb{I} - \mathbf{X})\right] &= \text{Tr}\left[\mathbf{R}^T\mathbf{R}\right] - \text{Tr}\left[\mathbf{X}\mathbf{R}^T\mathbf{R}\mathbf{X}\right] \\
&= \text{Tr}\left[\mathbf{R}^T\mathbf{R}\right] - \text{Tr}\left[\mathbf{X}(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{X}\right] \pm O(\sqrt{\eta})\|\mathbf{A}\|_F^2
\end{aligned} \tag{8.75}$$

Here, we observe $\|\mathbf{SAT}\|_F^2$ is an unbiased estimator for $\|\mathbf{A}\|_F^2$ and $\|\mathbf{SNT}\|_F^2$ is an unbiased estimator for $\|\mathbf{N}\|_F^2$. Using the same idea as above, we can bound the cross and tail terms by $O(\sqrt{\eta})\|\mathbf{A}\|_F^2$. Our goal is show that Equations 8.73 and 8.75 are related up to additive error $O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$. We first relate $\text{Tr}\left[\mathbf{C}^T\mathbf{C}\right]$ and $\text{Tr}\left[\mathbf{R}^T\mathbf{R}\right]$. Recall, $\mathbb{E}\left[\text{Tr}\left[\mathbf{R}^T\mathbf{R}\right]\right] = \mathbb{E}\left[\|\mathbf{R}\|_F^2\right] = \|\mathbf{C}\|_F^2 = \text{Tr}\left[\mathbf{C}^T\mathbf{C}\right]$. Using a scalar Chernoff bound, we show that with probability at least $1 - 1/\text{poly}(n)$, $|\|\mathbf{R}\|_F^2 - \|\mathbf{C}\|_F^2| \leqslant \epsilon\|\mathbf{A}\|_F^2$. Observe, for all $i \in [t]$, $\mathbf{R}_{*,i} = \frac{1}{\sqrt{p_{i'}t}}\mathbf{C}_{i',*}$ for some

$i' \in [n]$. Then,

$$\begin{aligned}\|\mathbf{R}_{i,*}\|_2^2 = \frac{1}{p_{i'}t}\|\mathbf{C}_{i',*}\|_2^2 &= \frac{\phi_{max}\sqrt{n}\|\mathbf{A}\|_F\epsilon^2}{\phi_{max}\sqrt{nk}\log(n)\log(n/\delta)\mathbf{A}_{i',i'}}\|\mathbf{C}_{i',*}\|_2^2 \\ &\leqslant \frac{c(1+\eta)\|\mathbf{A}\|_F\epsilon^2}{\sqrt{k}\log(n/\delta)\mathbf{A}_{i',i'}}\max\left\{\|\mathbf{A}_{i',*}\|_2^2, \frac{\phi_{\max}\sqrt{n}\|\mathbf{A}\|_F\mathbf{A}_{i',i'}}{t}\right\} \\ &\leqslant \frac{c\epsilon^2}{\sqrt{k}\log(n/\delta)}\max\left\{\|\mathbf{A}\|_2\|\mathbf{A}\|_F, \frac{\|\mathbf{A}\|_F^2\epsilon^2}{\sqrt{k}\log(n/\delta)}\right\} \\ &\leqslant \frac{c\epsilon^2}{\sqrt{k}\log(n/\delta)}\|\mathbf{A}+\mathbf{N}\|_F^2\end{aligned} \tag{8.76}$$

where we use $\mathbf{C}_{i',*} = (\mathbf{AT})_{i',*} + (\mathbf{NT})_{i',*}$, $\|(\mathbf{NT})_{i,*}\|_2^2 \leqslant (\eta)\|(\mathbf{AT})_{i,*}\|_2^2$ for all i and Lemma 8.4.5 to bound $\|(\mathbf{AT})_{i,*}\|_2^2$. Therefore, $\frac{\sqrt{k}\log(n/\delta)}{c\epsilon^2\|\mathbf{A}\|_F^2}\|\mathbf{R}_{i,*}\|_2^2 \in [0,1]$. Note, $\|\mathbf{R}\|_F^2$ is an unbiased estimator for $\|\mathbf{A}+\mathbf{N}\|_F^2$. Using a Chernoff bound,

$$\begin{aligned}\mathbf{Pr}\left[\right]\|\mathbf{R}\|_F^2 \geqslant (1+\epsilon)\|\mathbf{A}+\mathbf{N}\|_F^2 = \mathbf{Pr}\left[\right]\frac{\sqrt{k}\log(n/\delta)}{\epsilon^2\|\mathbf{A}\|_F^2}\|\mathbf{R}\|_F^2 \geqslant \frac{\sqrt{k}\log(n/\delta)}{\epsilon^2}(1+\epsilon) \\ \leqslant \exp\left(-\frac{\sqrt{k}\epsilon^2\log(n/\delta)}{\epsilon^2}\right) \leqslant \frac{\delta}{10}\end{aligned} \tag{8.77}$$

Therefore, with probability at least $1-\delta/10$, $|\|\mathbf{R}\|_F^2 - \|\mathbf{A}+\mathbf{N}\|_F^2| \leqslant \epsilon\|\mathbf{A}+\mathbf{N}\|_F^2$. Note, we can then bound $\|\mathbf{A}+\mathbf{N}\|_F^2 \leqslant \|\mathbf{A}\|_F^2 + 2\sqrt{\eta}\|\mathbf{A}\|_F^2$. Therefore,

$$|\|\mathbf{R}\|_F^2 - \|\mathbf{A}\|_F^2| \leqslant O(\epsilon+\sqrt{\eta})\|\mathbf{A}\|_F$$

Recall, by equation 8.65, with probability $\delta/10$, $\|\mathbf{A}\|_F^2 = (1 \pm (\epsilon + 2\sqrt{\eta}))\|\mathbf{C}\|_F^2$ and thus we have that $\|\mathbf{R}\|_F^2 - \|\mathbf{C}\|_F^2 \leqslant 3\epsilon\|\mathbf{A}\|_F^2$. We can repeat the above argument to lower bound $\|\mathbf{R}\|_F^2$. Therefore, with probability $1-\delta$, we have

$$|\|\mathbf{R}\|_F^2 - \|\mathbf{C}\|_F^2| \leqslant O(\epsilon+\sqrt{\eta})\|\mathbf{A}\|_F^2 \tag{8.78}$$

Next, we relate Tr $\left[\mathbf{X}(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{X}\right]$ and Tr $\left[\mathbf{X}(\mathbf{AT})^T\mathbf{ATX}\right]$. Observe, $\mathbf{X}$ is a rank k projection matrix and we can represent it as $\mathbf{ZZ}^T$, where $\mathbf{Z} \in \mathbb{R}^{n\times k}$ and has orthonormal columns. By the cyclic property of the trace, we have

$$\text{Tr}\left[\mathbf{ZZ}^T(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{ZZ}^T\right] = \text{Tr}\left[\mathbf{Z}^T(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{Z}\right] = \sum_{j\in[k]}\mathbf{Z}_{*,j}^T(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{Z}_{*,j}$$

Similarly, Tr $\left[\mathbf{Z}\mathbf{Z}^T(\mathbf{AT})^T\mathbf{ATZ}\mathbf{Z}^T\right] = \sum_{j\in[k]} \mathbf{Z}_{*,j}^T(\mathbf{AT})^T\mathbf{ATZ}_{*,j}$. By Theorem 181 , we have

$$\sum_{j\in[k]} \left(\mathbf{Z}_{*,j}^T(\mathbf{AT})^T\mathbf{ATZ}_{*,j}\right) = \sum_{j\in[k]} \left(\mathbf{Z}_{*,j}^T(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{Z}_{*,j} \pm \frac{\epsilon}{k}\|\mathbf{A}\|_F^2\mathbf{Z}_{*,j}^T\mathbf{IZ}_{*,j}\right) \tag{8.79}$$

Since $\mathbf{Z}_{*,j}^T\mathbf{Z}_{*,j} = 1$ and Tr $\left[\mathbf{Z}^T(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{Z}\right]$ = Tr $\left[\mathbf{X}(\mathbf{AT})^T\mathbf{ATX}\right]$, we obtain

$$\text{Tr}\left[\mathbf{X}(\mathbf{AT})^T\mathbf{ATX}\right] - \epsilon\|\mathbf{A}\|_F^2 \leqslant \text{Tr}\left[\mathbf{X}(\mathbf{SAT})^T(\mathbf{SAT})\mathbf{X}\right] \leqslant \text{Tr}\left[\mathbf{X}(\mathbf{AT})^T\mathbf{ATX}\right] + \epsilon\|\mathbf{A}\|_F^2 \tag{8.80}$$

Combining equations 8.80,8.78, 8.74 and 8.75 with probability $1 - c$,

$$\|\mathbf{C} - \mathbf{CX}\|_F^2 - O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2 \leqslant \|\mathbf{R} - \mathbf{RX}\|_F^2 \leqslant \|\mathbf{C} - \mathbf{CX}\|_F^2 + O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$$

□

Algorithm 12 : Robust PSD Low-Rank Approximation

Input: A Matrix $\mathbf{A}+\mathbf{N}$, integer k, $\epsilon > 0$ and $\phi_{\max} = \max_j \mathbf{A}_{j,j}/(\mathbf{A}+\mathbf{N})_{j,j}$

1. Let $t = \frac{c\phi_{\max}^2\sqrt{n}k\log^2(n)}{\epsilon^2}$, for some constant c. Let $q = \{q_1, q_2 \dots q_n\}$ denote a distribution over columns of $\mathbf{A}+\mathbf{N}$ such that $q_j = \frac{(\mathbf{A}+\mathbf{N})_{j,j}}{\mathrm{Tr}[\mathbf{A}+\mathbf{N}]}$. Construct a column PCP for $\mathbf{A}+\mathbf{N}$ by sampling t columns of $\mathbf{A}+\mathbf{N}$ such that each column is set to $\frac{(\mathbf{A}+\mathbf{N})_{*,j}}{\sqrt{tq_j}}$ with probability q_j. Let $\mathbf{C}$ be the resulting $n \times t$ matrix that satisfies the guarantee of Theorem 180.

2. Let $p = \{p_1, p_2 \dots p_n\}$ denote a distribution over rows of $\mathbf{C}$ such that $p_i = \frac{(\mathbf{A}+\mathbf{N})_{i,i}}{\mathrm{Tr}[\mathbf{A}+\mathbf{N}]}$. Construct a row PCP for $\mathbf{C}$ by sampling t rows of $\mathbf{C}$ such that each row is set to $\frac{\mathbf{C}_{i,*}}{\sqrt{tp_i}}$ with probability p_i. Let $\mathbf{R}$ be the resulting $t \times t$ matrix that satisfies the guarantee of Theorem 182. Sample $\Theta(n)$ entries uniformly at random from $\mathbf{A}$ and rescale such that $\widetilde{v}^2 = \Theta(\|\mathbf{A}\|_F^2)$.

3. Let $\mu = \phi_{\max}\sqrt{|(\mathbf{A}+\mathbf{N})_{i,i}| \cdot |(\mathbf{A}+\mathbf{N})_{i',i'}|}$ For all $i \in [t]$, let $\mathcal{X}_i = \sum_{j\in[\epsilon^3 t/k^3]} \mathcal{X}_{i,j}$ such that $\mathcal{X}_{i,j} = (k^3/\epsilon^3)\mathbf{R}_{i,i'}^2$, with probability $1/t$, for all $i' \in [t]$. Here, we query the entry corresponding to $\mathbf{R}_{i,i'}$ in $\mathbf{A}+\mathbf{N}$ and truncate it to μ. Let $\tau = \phi_{\max}^2 n\widetilde{v}^2/t^2$. If $\mathcal{X}_i > \tau$, sample row $\mathbf{R}_{i,*}$ with probability 1. For the remaining rows, sample $nk/(\epsilon t)$ rows uniformly at random.

4. Run the sampling algorithm from Frieze-Kannan-Vempala [FKV04b] to compute a $t \times k$ matrix $\mathbf{S}$ such that $\|\mathbf{R} - \mathbf{R}\mathbf{S}\mathbf{S}^T\|_F^2 \leqslant \|\mathbf{R} - \mathbf{R}_k\|_F^2 + \epsilon\|\mathbf{R}\|_F^2$. Consider the regression problem
$$\min_{\mathbf{X}\in\mathcal{R}^{n\times k}} \|\mathbf{C} - \mathbf{X}\mathbf{S}^T\|_F^2.$$
Sketch the problem using the leverage scores of $\mathbf{S}^T$, as shown in Lemma 8.3.3, to obtain a sampling matrix $\mathbf{E}$ with $O(\frac{k}{\epsilon})$ columns. Compute
$$\mathbf{X}_\mathbf{C} = \arg\min_{\mathbf{X}\in\mathcal{R}^{n\times k}} \|\mathbf{C}\mathbf{E} - \mathbf{X}\mathbf{S}^T\mathbf{E}\|_F^2.$$
Let $\mathbf{X}_\mathbf{C}\mathbf{S}^T = \mathbf{U}\mathbf{V}^T$ be such that $\mathbf{U} \in \mathbb{R}^{n\times k}$ has orthonormal columns.

5. Consider the regression problem
$$\min_{\mathbf{X}\in\mathbb{R}^{k\times n}} \|\mathbf{A} - \mathbf{U}\mathbf{X}\|_F^2.$$
Sketch the problem as above, following Lemma 8.3.3 to obtain a sampling matrix $\mathbf{E}'$ with $O(\frac{k}{\epsilon})$ rows. Compute
$$\mathbf{X}_\mathbf{A} = \arg\min_\mathbf{X} \|\mathbf{E}'\mathbf{A} - \mathbf{E}'\mathbf{U}\mathbf{X}\|_F^2$$

Output: $\mathbf{M} = \mathbf{U}$, $\mathbf{N}^T = \mathbf{X}_\mathbf{A}$

Full Algorithm. Next, we describe a sublinear time and query robust algorithm for low-rank approximation of PSD matrices. We show that querying $\widetilde{O}(\phi_{\max}^2 nk/\epsilon)$ entries of $\mathbf{A}$ suffices. While we assume we know $\phi_{\max}$, in practice this need not be the case. Therefore, given a budget for the total number of queries, denoted by β, we can run the algorithm by querying a $\sqrt{\beta} \times \sqrt{\beta}$ submatrix (as described in Algorithm 12), but correctness only holds when $\beta \geqslant \widetilde{\Theta}(\phi_{\max}^2 nk/\epsilon)$. Recall, whenever we read an entry in $(\mathbf{A}+\mathbf{N})_{i,j}$, we can truncate it to $\phi_{\max}\sqrt{|(\mathbf{A}+\mathbf{N})_{i,i}| \cdot |(\mathbf{A}+\mathbf{N})_{j,j}|}$. We can compute these thresholds by simply reading the diagonal of $\mathbf{A}+\mathbf{N}$.

We proceed by constructing column and row projection-cost preserving sketches of $\mathbf{A}+\mathbf{N}$, to obtain a $t \times t$ matrix $\mathbf{R}$, where $t = \widetilde{O}(\phi_{\max}\sqrt{n}k^2/\epsilon^2)$. Instead of reading the entire matrix, we sample $\epsilon^3 t/k^3$ entries in each row of $\mathbf{R}$, and read these entries. Ideally we would want to estimate ℓ_2^2 norms of rows of $\mathbf{R}$ to then use a result of Frieze-Kannan-Vempala [FKV04b] to show that there exists an $s \times t$ matrix $\mathbf{S}$ such that the row space of $\mathbf{S}$ contains a good rank-k approximation, where $s = c\phi_{\max}^2 nk/\epsilon t$, for some constant c. However, we show that is it not possible to obtain accurate estimates of the row norms of each row of $\mathbf{R}$ with high probability.

Instead, we describe a new sampling procedure that ends up sampling rows of $\mathbf{R}$ with the same probability as Frieze-Kannan-Vempala. Once we compute a good low-rank approximation for $\mathbf{R}$ we can follow the approach of [CMM17],[MW17c] and [BW18], where we set up two regression problems, and use fast approximate regression to compute a low rank approximation for $\mathbf{A}$. The main theorem we prove in this section is as follows:

Theorem 183. *(Robust PSD LRA.) Let k be an integer and $\epsilon > \eta > 0$. Given a matrix $\mathbf{A}+\mathbf{N}$, where $\mathbf{A}$ is PSD and $\mathbf{N}$ is a corruption term such that $\|\mathbf{N}\|_F^2 \leqslant \sqrt{\eta}\|\mathbf{A}\|_F^2$ and for all $i \in [n]$ $\|\mathbf{N}_{i,*}\|_2^2 \leqslant c\|\mathbf{A}_{i,*}\|_2^2$, for a fixed constant c, Algorithm 12 samples $\widetilde{O}\left(\phi_{\max}^2 nk/\epsilon\right)$ entries in $\mathbf{A}+\mathbf{N}$ and computes matrices $\mathbf{M}, \mathbf{N}^T \in \mathbb{R}^{n\times k}$ such that with probability at least $99/100$,*

$$\|\mathbf{A} - \mathbf{MN}\|_F^2 \leqslant \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$$

We begin with the following simple lemma for approximating the Frobenius norm :

Lemma 8.4.6. *(Approximating Frobenius Norm.) Given as input an $n \times n$ matrix $\mathbf{A}+\mathbf{N}$, there exists an algorithm that reads $O(\phi_{\max}^2 n)$ entries in $\mathbf{A}$ and outputs an estimator $\tilde{v}$ such that with probability at least $1 - \frac{1}{n^c}$, $\tilde{v} = \Theta(\|\mathbf{A}\|_F^2)$.*

Proof. There are multiple ways to see this. Observe, in Theorem 182, we show that sampling $\frac{\phi_{\max}^2 n \log(n)}{\epsilon^2}$ entries results in row projection-cost preserving sketch $\mathbf{R}$ such that $\|\mathbf{R}\|_F^2 = (1 \pm$

$\epsilon)\|\mathbf{A}+\mathbf{N}\|_F^2$. Setting ϵ to be a small constant suffices. $\square$

Next, we provide intuition for why uniformly sampling columns of $\mathbf{R}$ does not suffice for obtaining a sketch that spans a good low rank approximation. For simplicity, we assume there is no noise ($\eta = 0$ and $\phi_{\max} = 1$) and show that our techniques to bound the column norms of $\mathbf{R}$ results in an estimate that is too large. We note that this lemma is not required for proving our result, and is just present for intuition.

Lemma 8.4.7. *Let $\eta = 0$. Let $\mathbf{R} \in \mathbb{R}^{t\times t}$ be a row projection-cost preserving sketch for $\mathbf{C}$ as described in Theorem 182. For all $j \in [t]$, with probability at least $1 - 1/n^c$,*

$$\|\mathbf{R}_{*,j}\|_2^2 \leqslant O\left(\log(n)\max\left\{\|\mathbf{C}_{*,j}\|_2^2, \frac{n\|\mathbf{A}\|_F^2}{t^2}\right\}\right) = O\left(\log(n)\max\left\{\frac{\sqrt{n}\|\mathbf{A}\|_F^2}{t}, \frac{n\|\mathbf{A}\|_F^2}{t^2}\right\}\right)$$

where c is a fixed constant.

Proof. Observe, $\|\mathbf{R}_{*,j}\|_2^2 = \sum_{i\in[t]}\mathbf{R}_{i,j}^2$, where $\mathbf{R}_{i,j}^2 = \frac{\mathrm{Tr}[\mathbf{A}]}{t\mathbf{A}_{i',i'}}\mathbf{C}_{i',j}^2$ with probability $\frac{\mathbf{A}_{i',i'}}{\mathrm{Tr}[\mathbf{A}]}$ for all $i' \in [n]$. Then, $\mathbb{E}[\|\mathbf{R}_{*,j}\|_2^2] = \sum_{i=1}^n \mathbf{C}_{i,j}^2 = \|\mathbf{C}_{*,j}\|_2^2$. Next, we compute the variance of $\|\mathbf{R}_{*,j}\|_2^2$. $\mathbf{Var}\left[\|\mathbf{R}_{*,j}\|_2^2\right] = t\mathbf{Var}\left[\mathbf{R}_{i,j}\right] \leqslant t\,\mathbb{E}\left[\mathbf{R}_{i,j}^4\right]$. Then,

$$\begin{aligned}
t\,\mathbb{E}\left[\mathbf{R}_{i,j}^4\right] = \sum_{i'\in[n]}\frac{1}{tp_{i'}}\mathbf{C}_{i',j}^4 &\leqslant \sum_{i'\in[n]}\frac{\mathrm{Tr}\left[\mathbf{A}\right]^2}{t^2\mathbf{A}_{i',i'}\mathbf{A}_{j,j}}\mathbf{A}_{i',j}^2\mathbf{A}_{i',i'}\mathbf{A}_{j,j} \\
&\leqslant \frac{\mathrm{Tr}\left[\mathbf{A}\right]^2}{t^2}\|\mathbf{A}_{*,j}\|_2^2 \\
&= \frac{\mathrm{Tr}\left[\mathbf{A}\right]}{t}\|\mathbf{C}_{*,j}\|_2^2 \\
&\leqslant \left(\frac{\mathrm{Tr}\left[\mathbf{A}\right]}{t}\right)^2 + \|\mathbf{C}_{*,j}\|_2^4 \qquad \text{[AM-GM]}
\end{aligned}$$

where we use $\mathbf{A}_{i,j}^2 \leqslant \mathbf{A}_{i,i}\mathbf{A}_{j,j}$, which follows from applying Cauchy-Schwarz to $\langle \mathbf{A}_{i,*}^{1/2}, \mathbf{A}_{j,*}^{1/2}\rangle$.

Similarly, we bound $\mathbf{R}_{i,j}^2 \leqslant \frac{\mathrm{Tr}[\mathbf{A}]}{t}$. Applying Bernstein's inequality,

$$\begin{aligned}
&\mathbf{Pr}\left[\right] |\|\mathbf{R}_{*,j}\|_2^2 - \|\mathbf{C}_{*,j}\|_2^2 \geqslant \eta \max\left\{\|\mathbf{C}_{*,j}\|_2^2, \frac{\mathrm{Tr}\left[\mathbf{A}\right]}{t}\right\} \\
&\leqslant 2\exp\left(-\frac{\eta^2 \max\left\{\|\mathbf{C}_{*,j}\|_2^4, \left(\frac{\mathrm{Tr}[\mathbf{A}]}{t}\right)^2\right\}}{\left(\frac{\mathrm{Tr}[\mathbf{A}]}{t}\right)^2 + \|\mathbf{C}_{*,j}\|_2^4 + \eta \max\left\{\frac{\mathrm{Tr}[\mathbf{A}]\|\mathbf{C}_{*,j}\|_2^2}{t}, \left(\frac{\mathrm{Tr}[\mathbf{A}]}{t}\right)^2\right\}}\right) \\
&\leqslant 2\exp\left(-\frac{\eta \max\left\{\|\mathbf{C}_{*,j}\|_2^4, \left(\frac{\mathrm{Tr}[\mathbf{A}]}{t}\right)^2\right\}}{c'\left(\frac{\mathrm{Tr}[\mathbf{A}]}{t}\right)^2 + c'\|\mathbf{C}_{*,j}\|_2^4}\right)
\end{aligned}$$

where we use the AM-GM inequality on $\frac{\mathrm{Tr}[\mathbf{A}]}{t}\|\mathbf{C}_{*,j}\|_2^2$ repeatedly. Setting $\eta = \Omega(\log(n))$ completes the proof. Finally, observe, for any $j \in [t]$, $\|\mathbf{C}_{*,j}\|_2^2 = \frac{\mathrm{Tr}[\mathbf{A}]}{t\mathbf{A}_{j',j'}}\|\mathbf{A}_{*,j'}\|_2^2$ for some $j' \in [n]$. We then use $\mathrm{Tr}\left[\mathbf{A}\right] \leqslant \sqrt{n}\|\mathbf{A}\|_F$. □

It is well-known that to recover a low-rank approximation for $\mathbf{R}$, one can sample rows of $\mathbf{R}$ proportional to row norm estimates, denoted by $\mathbf{Y}_i$ [FKV04b]. As shown in [IVWW19] the following two conditions are a relaxation of those required in [FKV04b], and suffice to obtain an additive error low-rank approximation :

1. For all $i \in [t]$, $\mathcal{Y}_i \geqslant \|\mathbf{R}_{i,*}\|_2^2$.
2. $\sum_{i\in[t]} \mathcal{Y}_i \leqslant \frac{n}{t}\|\mathbf{R}\|_F^2$

To satisfy the first condition, we need to obtain overestimates for each row. Since it is not immediately clear how to obtain overestimates for row norms of $\mathbf{R}$, a naïve approach would be to bias the estimate for each row by an upper bound on the row norm. However, by Lemma 8.4.7, a row norm could be as large as $\sqrt{n}\|\mathbf{A}\|_F^2/t$. Observe, we cannot afford to bias the estimator of each row, $\mathcal{Y}_i$, by this amount since $\sum_{i\in[t]} \mathcal{Y}_i \geqslant \sqrt{n}\|\mathbf{A}\|_F^2 \geqslant \sqrt{n}\|\mathbf{R}\|_F^2$. Therefore, we would have to sample $\sqrt{n}k/\epsilon$ rows of $\mathbf{R}$, resulting in us querying $\Omega(nk^3/\epsilon^3)$ entries in $\mathbf{A}$, even when $\eta = 0$.

An alternative strategy would be to bias the estimator for each row by $n\|\mathbf{R}\|_F^2/t^2$, as this would satisfy condition 2 above. We can now hope to detect rows of $\mathbf{R}$ that have norm larger than $n\|\mathbf{R}\|_F^2/t^2$ by sampling $\epsilon^3 t/k^3$ entries in each row of $\mathbf{R}$, uniformly at random. Note, this way we can construct an unbiased estimator for the ℓ_2^2 norm of each row. Ideally, we would want to show a high probability statement for concentration of our row norm estimates around the expectation. We could then union bound, and obtain concentration for all i simultaneously.

However, this is not possible since it may be the case that a row of $\mathbf{R}$ is $\log(n)$-sparse with each entry being large in magnitude. In this case, uniformly querying the row would not observe any non-zero with good probability and thus cannot distinguish between such a row and an empty row. Instead, we settle for a weaker statement, that shows our estimate is accurate with $o(1)$ probability. All subsequent statements hold for $\eta > 0$.

Lemma 8.4.8. *(Estimating large row norms.) Let $\mathbf{R} \in \mathbb{R}^{t\times t}$ be the row PCP output by Step 2 of Algorithm 12. For all $i \in [t]$ let $\mathcal{X}_i = \sum_{j\in[\epsilon^3 t/k^3]} \mathcal{X}_{i,j}$ such that $\mathcal{X}_{i,j} = \frac{k^3\mathbf{R}^2_{i,j'}}{\epsilon^3}$ with probability $\frac{1}{t}$, for all $j' \in [t]$. Then, for all $i \in [t]$, $\mathcal{X}_i = \left(1 \pm \frac{1}{10}\right)\|\mathbf{R}_{i,*}\|_2^2$ with probability at least $\frac{\|\mathbf{R}_{i,*}\|_2^2 k}{\epsilon n}$.*

Proof. Observe, $\mathcal{X}_i$ is an unbiased estimator of $\|\mathbf{R}_{i,*}\|_2^2$:

$$\mathbb{E}\left[\mathcal{X}_i\right] = \frac{\epsilon^3 t}{k^3}\mathbb{E}\left[\mathcal{X}_{i,j}\right] = \frac{\epsilon^3 t}{k^3}\sum_{j'\in[t]}\frac{k^3}{\epsilon^3 n}\mathbf{R}^2_{i,j'} = \|\mathbf{R}_{i,*}\|_2^2$$

Next, we compute the variance of $\mathcal{X}_i$.

$$\begin{aligned}
\mathbf{Var}\left[\mathcal{X}_i\right] = \frac{\epsilon^3 t}{k^3}\mathbf{Var}\left[\mathcal{X}_{i,j}\right] &\leqslant \sum_{j\in[t]}\frac{1}{\epsilon^3}\mathbf{R}^4_{i,j} \\
&\leqslant \sum_{j\in[t]}\frac{k^3}{\epsilon^3}\mathbf{R}^2_{i,j}\left(\frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]^2}{t^2(\mathbf{A}+\mathbf{N})_{i,i}(\mathbf{A}+\mathbf{N})_{j,j}}(\mathbf{A}+\mathbf{N})^2_{i,j}\right) \\
&\leqslant \sum_{j\in[t]}\frac{k^3}{\epsilon^3}\mathbf{R}^2_{i,j}\left(\frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]^2}{t^2(\mathbf{A}+\mathbf{N})_{i,i}(\mathbf{A}+\mathbf{N})_{j,j}}(\mathbf{A}_{i,i}\mathbf{A}_{j,j}+\mathbf{N}^2_{i,j})\right) \\
&\leqslant \sum_{j\in[t]}\frac{k^3}{\epsilon^3}\mathbf{R}^2_{i,j}\left(\frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]^2\phi^2_{\max}}{t^2}+\frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]^2\phi^2_{\max}(\mathbf{A}+\mathbf{N})_{i,i}(\mathbf{A}+\mathbf{N})_{j,j}}{t^2(\mathbf{A}+\mathbf{N})_{i,i}(\mathbf{A}+\mathbf{N})_{j,j}}\right) \\
&\leqslant \sum_{j\in[t]}\frac{k^3}{\epsilon^3}\mathbf{R}^2_{i,j}\left(\frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]^2\phi^2_{\max}}{t^2}\right) \leqslant O\left(\frac{\epsilon\|\mathbf{A}\|_F^2}{k}\|\mathbf{R}_{i,*}\|_2^2\right)
\end{aligned} \tag{8.81}$$

Here, we use that $\mathbf{N}^2_{i,j} \leqslant \phi^2_{\max}(\mathbf{A}+\mathbf{N})_{i,i}(\mathbf{A}+\mathbf{N})_{j,j}$, which follows from our truncation procedure. Further, using $t = \phi_{\max}\sqrt{n}k^2/\epsilon^2$ and $\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right] \leqslant \sqrt{n}\|\mathbf{A}\|_F + \sqrt{\eta n}\|\mathbf{A}\|_F$, we can bound

$$\frac{\mathrm{Tr}\left[\mathbf{A}+\mathbf{N}\right]^2\phi^2_{\max}}{t^2} \leqslant O\left(\frac{\epsilon\|\mathbf{A}\|_F^2}{k}\right)$$

Further, using the same argument as above

$$\mathcal{X}_{i,j} \leqslant \frac{k^3}{\epsilon^3}\mathbf{R}_{i,j}^2 \leqslant O\left(\frac{\epsilon\|\mathbf{A}\|_F^2}{k}\right) \tag{8.82}$$

Using Equations 8.81 and 8.82 in Bernstein's inequality,

$$\begin{aligned}\mathbf{Pr}\left[\left|\mathcal{X}_i - \mathbb{E}[\mathcal{X}_i]\right| \geqslant \delta\,\mathbb{E}\left[\mathcal{X}_i\right]\right] &\leqslant \exp\left(-\frac{\delta^2\|\mathbf{R}_{i,*}\|_2^4}{\frac{\epsilon\|\mathbf{A}\|_F^2}{k}\|\mathbf{R}_{i,*}\|_2^2 + \frac{\delta\epsilon\|\mathbf{A}\|_F^2}{3k}\|\mathbf{R}_{i,*}\|_2^2}\right)\\ &\leqslant \exp\left(-\frac{\delta^2\|\mathbf{R}_{i,*}\|_2^2 k\log^2(n)}{\epsilon\|\mathbf{R}\|_F^2}\right)\end{aligned}$$

where we use that $\|\mathbf{A}\|_F^2 = \Theta(\|\mathbf{R}\|_F^2)$. Setting $\eta = \frac{1}{10}$, $\mathcal{X}_i = \left(1 \pm \frac{1}{10}\right)\|\mathbf{R}_{i,*}\|_2^2$ with probability at least $1 - \exp\left(-\frac{\|\mathbf{R}_{i,*}\|_2^2 k\log^2(n)}{\epsilon\|\mathbf{R}\|_F^2}\right)$. Let ξ_i be the event that $\mathcal{X}_i = \left(1 \pm \frac{1}{10}\right)\|\mathbf{R}_{i,*}\|_2^2$. Then, union bounding over $t \leqslant n$ such events ξ_i, simultaneously for all i, ξ_i is true with probability at least

$$1 - \exp\left(-\frac{\|\mathbf{R}_{i,*}\|_2^2 k\log(n)}{\epsilon\|\mathbf{R}\|_F^2}\right) \geqslant \frac{\|\mathbf{R}_{i,*}\|_2^2 k\log(n)}{\epsilon\|\mathbf{R}\|_F^2}$$ □

We now have two major challenges: first, the probability with which the estimators are accurate is too small to even detect all rows with norm larger than $\phi_{\max}^2 n\|\mathbf{R}\|_F^2/t^2$, and second, there is no small query certificate for when an estimator is accurate in estimating the row norms. Therefore, we cannot even identify the rows where we obtain an accurate estimate of their norm.

To address the first issue, we make the crucial observation that while we cannot estimate the norm of each row accurately, we can hope to sample the row with the same probability as Frieze-Kannan-Vempala [FKV04b]. Recall, their algorithm samples row $\mathbf{R}_{i,*}$ with probability at least $\|\mathbf{R}_{i,*}\|_2^2/\|\mathbf{R}\|_F^2$, which matches the probability in Lemma 8.4.8. Therefore, we can focus on designing a weaker notion of identifiability, that may potentially include extra rows.

We begin by partitioning rows of $\mathbf{R}$ into two sets. Let $\mathcal{H} = \left\{i \;\middle|\; \|\mathbf{R}_{i,*}\|_2^2 \geqslant \phi_{\max}^2 n\widetilde{v}^2/t^2\right\}$ be the set of heavy rows and $[t] \setminus \mathcal{H}$ be the remaining rows. Since with probability at least $1 - \frac{1}{\text{poly}(n)}$, $\|\mathbf{R}\|_F^2 = \Theta(\|\mathbf{A}\|_F^2) = \Theta(\widetilde{v}^2)$,

$$|\mathcal{H}| = O(t^2/\phi_{\max}^2 n) = O(k^4\log^4(n)/\epsilon^4)$$

Observe, every row in $\mathcal{H}$ can potentially satisfy the threshold $\tau = \phi_{\max}^2 n\widetilde{v}^2/t^2$. Therefore, even if our estimator $\mathcal{X}_i$ is $\Theta(\|\mathbf{R}_{i,*}\|_2^2)$ for all $i \in \mathcal{H}$, we include at most $\widetilde{O}(k^4/\epsilon^4)$ extra rows in $\mathbf{S}$, which

is well within our budget. Observe, we can then sample a row with probability 1 whenever the corresponding estimate is larger than τ. This sampling process ensures that we identify rows in $\mathcal{H}$ with the right probability and also doees not query more than $O(\phi_{\max}^2 nk/\epsilon)$ entries in $\mathbf{A}+\mathbf{N}$. For all the remaining rows, we know the norm is at most $O(\phi_{\max}^2 n\tilde{v}^2/t^2)$. We then modify the analysis of [FKV04b] to show that we can handle both cases separately.

Theorem 184. *(Existence [FKV04b].) Let $\mathbf{R}$ be a row projection-cost preserving sketch output by Step 2 of Algorithm 12. For all $i \in [t]$, let $\mathcal{X}_i$ be estimate for $\|\mathbf{R}_{i,*}\|_2^2$ as described in Step 3 of Algorithm 12. Let $\mathbf{S}$ be a subset of $s = O(\phi_{\max}^2 nk/\epsilon t)$ columns of $\mathbf{R}$ sampled according to distribution $r = \{r_1, r_2, \ldots r_t\}$ such that r_i is the probability of sampling the i-th row. Then, with probability at least $99/100$, there exists a $t \times k$ matrix $\mathbf{U}$ in the column span of $\mathbf{S}$ such that*

$$\|\mathbf{R} - \mathbf{U}\mathbf{U}^T\mathbf{R}\|_F^2 \leqslant \|\mathbf{R} - \mathbf{R}_k\|_F^2 + \epsilon\|\mathbf{R}\|_F^2$$

Proof. We follow the proof strategy of [FKV04b] and show how to directly bound the variance in our setting as opposed to reducing to the two conditions above. Let $\mathbf{R} = \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^T = \sum_{\ell\in[t]} \mathbf{\Sigma}_{\ell,\ell}\mathbf{P}_{\ell,*}\mathbf{Q}_{\ell,*}^T = \sum_{\ell\in[t]} \sigma_\ell \mathbf{P}_\ell \mathbf{Q}_\ell^T$. Recall, $\mathbf{R}_k = \sum_{\ell\in[k]} \mathbf{A}\mathbf{Q}_\ell\mathbf{Q}_\ell^T$. For $\ell \in [t]$, let $\mathbf{W}_\ell = \frac{1}{s}\sum_{i'\in[s]} \mathbf{Y}_{i'}$ where $\mathbf{Y}_{i'} = \frac{\mathbf{P}_{i,\ell}}{r_i}\mathbf{R}_{i,*}$ with probability r_i, for all $i \in [t]$. Then,

$$\mathbb{E}[\mathbf{W}_\ell] = \mathbb{E}[\mathbf{Y}_{i'}] = \sum_{j\in[t]} \frac{\mathbf{P}_{i,\ell}}{r_i}\mathbf{R}_{i,*}r_i = \sigma_\ell\mathbf{Q}_\ell \tag{8.83}$$

Therefore, in expectation the span of the rows contain a good low-rank solution. Next, we bound the variance. Recall, here we consider the rows in $\mathcal{H}$ and its complement separately. From Lemma For all $i \in \mathcal{H}$, we know that $\mathcal{X}_i = \Theta(\|\mathbf{R}_{i,*}\|_2^2)$ with probability at least $\|\mathbf{R}_{i,*}\|_2^2 k\log(n)/\epsilon\|\mathbf{R}\|_F^2$. Since for all such i, $\|\mathbf{R}_{i,*}\|_2^2 \geqslant \tau$, the corresponding $r_i \geqslant \frac{\|\mathbf{R}_{i,*}\|_2^2 k\log(n)}{\epsilon\|\mathbf{R}\|_F^2}$, since every time we pass the threshold we sample the row. For all $i \notin \mathcal{H}$, $r_i \geqslant 1/t$ since there can be at most t such i, and we sample each such row with uniform probability. Once we have a lower bound on r_i in both cases, we open up the analysis of the variance bound in [FKV04b] and show that our lower bounds suffice.

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{W}_\ell - \sigma_\ell \mathbf{Q}_\ell\|_2^2\right] = \frac{1}{s}\left(\sum_{j\in[t]} \frac{\mathbf{P}_{i,\ell}^2}{r_i}\|\mathbf{R}_{i,*}\|_2^2\right) - \frac{\sigma_\ell^2}{s} &\leqslant \frac{1}{s}\left(\sum_{j\in\mathcal{H}} \frac{\mathbf{P}_{i,\ell}^2}{r_i}\|\mathbf{R}_{i,*}\|_2^2 + \sum_{i\in[t]\setminus\mathcal{H}} \frac{\mathbf{P}_{i,\ell}^2}{r_i}\|\mathbf{R}_{i,*}\|_2^2\right) \\
&\leqslant \frac{1}{s}\left(\sum_{j\in\mathcal{H}} \frac{k\log(n)\mathbf{P}_{i,\ell}^2\|\mathbf{R}\|_F^2}{\epsilon} + \sum_{j\in[t]\setminus\mathcal{H}} t\mathbf{P}_{i,\ell}^2\|\mathbf{R}_{i,*}\|_2^2\right) \\
&\leqslant \frac{1}{s}\left(\sum_{j\in\mathcal{H}} \frac{k\log(n)\mathbf{P}_{i,\ell}^2\|\mathbf{R}\|_F^2}{\epsilon} + \sum_{j\in[t]\setminus\mathcal{H}} \frac{n}{t}\mathbf{P}_{i,\ell}^2\widetilde{v}^2\right) \\
&\leqslant \frac{1}{s}\left(\frac{k\log(n)}{\epsilon} + \frac{\phi_{\max}^2 n}{t}\right)\|\mathbf{R}\|_F^2
\end{aligned} \tag{8.84}$$

Now, we can repeat the argument from [FKV04b] and it suffices to set $s = \left(\frac{\phi_{\max}^2 n}{t} + \frac{k}{\epsilon}\right)\frac{k}{\epsilon} = O(\frac{\phi_{\max}^2 nk}{\epsilon t})$. For completeness, we present the rest of the proof here. For all $\ell \in [t]$, let $\mathbf{Y}_\ell = \frac{1}{\sigma_\ell}\mathbf{W}_\ell$. Let $\mathcal{V} = \text{span}(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_k)$. Let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots \mathbf{Z}_t$ be an orthonormal basis for $\mathcal{R}^t$ such that $\mathcal{V} = \text{span}(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_{k'})$, where $k' = \dim(\mathcal{V})$. Let $\mathbf{S} = \mathbf{R}\sum_{\ell\in[k]} \mathbf{Z}_\ell\mathbf{Z}_\ell^T$ be the candidate low-rank approximation approximation. Then,

$$\begin{aligned}
\|\mathbf{R} - \mathbf{S}\|_F^2 &= \sum_{\ell\in[t]} \|(\mathbf{R} - \mathbf{S})\mathbf{Z}_\ell\|_2^2 \\
&= \sum_{\ell\in[k'+1,t]} \|\mathbf{R}\mathbf{Z}_\ell\|_2^2 \\
&= \sum_{\ell\in[k'+1,t]} \left\|\left(\left(\mathbf{R} - \mathbf{R}\sum_{\ell'\in[k]} \mathbf{Q}_{\ell'}\mathbf{Y}_{\ell'}^T\right)\mathbf{Z}_\ell\right\|_2^2 \right. \\
&\leqslant \left\|\mathbf{R} - \mathbf{R}\sum_{\ell'\in[k]} \mathbf{Y}_{\ell'}\mathbf{Y}_{\ell'}^T\right\|_F^2
\end{aligned} \tag{8.85}$$

where the first equality follows from $\|\mathbf{Z}_\ell\|_2^2 = 1$, the seconds follows from $\mathbf{Z}_{\ell'}^T\mathbf{Z}_\ell = 0$ for $\ell' \neq \ell$, the third follows from $\langle \mathbf{Y}_{\ell'}, \mathbf{Z}_\ell\rangle = 0$ for all $\ell' \leqslant k$ and $\ell > k'$. Let $\widehat{\mathbf{S}} = \mathbf{R}\sum_{\ell'\in[k]} \mathbf{Y}_{\ell'}\mathbf{Y}_{\ell'}^T$. Since $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_t$ forms an orthonormal basis

$$\begin{aligned}
\left\|\mathbf{R} - \widehat{\mathbf{S}}\right\|_F^2 &\leqslant \sum_{\ell\in[t]} \left\|\mathbf{P}_\ell\left(\mathbf{R} - \widehat{\mathbf{S}}\right)\right\|_2^2 \\
&= \sum_{\ell\in[k]} \|\sigma_\ell\mathbf{Q}_\ell - \mathbf{W}_\ell\|_2^2 + \sum_{\ell\in[k+1,t]} \sigma_\ell^2
\end{aligned} \tag{8.86}$$

Taking expectations on both sides of equation 8.86, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{R}-\widehat{\mathbf{S}}\right\|_F^2\right] &\leqslant \mathbb{E}\left[\sum_{\ell\in[k]}\|\sigma_\ell\mathbf{Q}_\ell-\mathbf{W}_\ell\|_2^2\right]+\|\mathbf{R}-\mathbf{R}_k\|_F^2 \\ &\leqslant \frac{k}{s}\left(\frac{k\log(n)}{\epsilon}+\frac{\phi_{\max}^2 n}{t}\right)\|\mathbf{R}\|_F^2+\|\mathbf{R}-\mathbf{R}_k\|_F^2\end{aligned} \tag{8.87}$$

Since $\widehat{\mathbf{S}}$ is a rank k matrix and $\mathbf{R}_k$ is the best rank k approximation to $\mathbf{R}$, $\|\mathbf{R}-\widehat{\mathbf{S}}\|_F^2-\|\mathbf{R}-\mathbf{R}_k\|_F^2$ is a non-negative random variable. Thus, using Markov's inequality and Equation 8.85,

$$\mathbf{Pr}\left[\|\mathbf{R}-\mathbf{S}\|_F^2-\|\mathbf{R}-\mathbf{R}_k\|_F^2\geqslant\frac{100nk}{st}\|\mathbf{R}\|_F^2\right]\leqslant\frac{1}{100}$$

Therefore, it suffices to sample $s=O\left(\frac{\phi_{\max}^2 nk}{\epsilon t}\right)$ columns, read all of them and compute a low rank approximation for $\mathbf{R}$ with probability at least $\frac{99}{100}$. Observe, the total entries read by this algorithm is $O\left(\frac{\phi_{\max}^2 nk}{\epsilon t}\cdot t\right)=O\left(\frac{\phi_{\max}^2 nk}{\epsilon}\right)$. □

It remains to show that we can now recover a low-rank approximation for $\mathbf{A}$, in factored form, from the low-rank approximation for $\mathbf{R}$. Here, we follow the approach of [CMM17],[MW17c] and [BW18], where we set up two regression problems, and use the sketch and solve paradigm to compute an approximate solution. We use the following Lemma from [BW18] that relates a good low-rank approximation of an additive error project-cost preserving sketch to a low-rank approximation of the original matrix. A similar guarantee for relative error appears in [CMM17] and [MW17c].

Lemma 8.4.9. *(Lemma 4.4 in [BW18].) Let $\mathbf{C}$ be a column PCP for $\mathbf{A}$ satisfying the guarantee of Theorem 180. Let $\mathbf{P}_{\mathbf{C}}^*=\arg\min_{rank(\mathbf{X})\leqslant k}\|\mathbf{C}-\mathbf{X}\mathbf{C}\|_F^2$ and $\mathbf{P}_{\mathbf{A}}^*=\arg\min_{rank(\mathbf{X})\leqslant k}\|\mathbf{A}-\mathbf{X}\mathbf{A}\|_F^2$. Then, for any rank k projection matrix $\mathbf{P}$ such that $\|\mathbf{C}-\mathbf{P}\mathbf{C}\|_F^2\leqslant\|\mathbf{C}-\mathbf{P}_{\mathbf{C}}^*\mathbf{C}\|_F^2+(\epsilon+\sqrt{\eta})\|\mathbf{C}\|_F^2$, with probability at least $99/100$,*

$$\|\mathbf{A}-\mathbf{P}\mathbf{A}\|_F^2\leqslant\|\mathbf{A}-\mathbf{P}_{\mathbf{A}}^*\mathbf{A}\|_F^2+(\epsilon+\sqrt{\eta})\|\mathbf{A}\|_F^2$$

A similar guarantee holds for a row PCP of $\mathbf{A}$.

Note, while $\mathbf{R}\mathbf{S}\mathbf{S}^T$ is an approximate rank-k solution for $\mathbf{R}$, it does not have the same dimensions as $\mathbf{A}$. If we do not consider running time, we could construct a low-rank approximation to $\mathbf{A}$ as follows: since projecting $\mathbf{R}$ onto $\mathbf{S}^T$ is approximately optimal, it follows from Lemma 8.4.9 that

with probability $99/100$,

$$\|\mathbf{C} - \mathbf{C}\mathbf{S}\mathbf{S}^T\|_F^2 = \|\mathbf{C} - \mathbf{C}_k\|_F^2 \pm (\epsilon + \sqrt{\eta})\|\mathbf{C}\|_F^2 \tag{8.88}$$

Let $\mathbf{C}_k = \mathbf{U}'\mathbf{V}'^T$ be such that $\mathbf{U}'$ has orthonormal columns. Then, $\|\mathbf{C} - \mathbf{U}'\mathbf{U}'^T\mathbf{C}\|_F^2 = \|\mathbf{C} - \mathbf{C}_k\|_F^2$ and by Lemma 8.4.9 it follows that with probability $98/100$, $\|\mathbf{A} - \mathbf{U}'\mathbf{U}'^T\mathbf{A}\|_F^2 \leqslant \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2$. However, even approximately computing a column space $\mathbf{U}'$ for $\mathbf{C}_k$ using an input-sparsity time algorithm, such as [CW13], could require $\Omega(nt)$ queries. To get around this issue, we observe that an approximate solution for $\mathbf{R}$ lies in the row space of $\mathbf{S}^T$ and therefore, an approximately optimal solution for $\mathbf{C}$ lies in the row space of $\mathbf{S}^T$. We then set up the following regression problem:

$$\min_{\text{rank}(\mathbf{X}) \leqslant k} \|\mathbf{C} - \mathbf{X}\mathbf{S}^T\|_F^2 \tag{8.89}$$

Note, this regression problem is still too large to be solved in sublinear time. Therefore, we sketch it by sampling columns of $\mathbf{C}$ to set up a smaller regression problem. Observe, since $\mathbf{S}$ has orthonormal columns, the leverage scores are simply ℓ_2^2 norms of rows of $\mathbf{S}$. Now, using Lemma 8.3.3, approximately solving this regression problem requires sampling $\Omega(k/\epsilon)$ rows of $\mathbf{C}$, which in turn requires $\Omega(\frac{nk}{\epsilon})$ queries to $\mathbf{A} + \mathbf{N}$. Note, the above theorem applied to Equation 8.89 can take $O(nk + \text{poly}(k, \epsilon^{-1}))$ time and thus is a lower order term. Since $\mathbf{S}^T$ has orthonomal rows, the leverage scores are precomputed. With probability at least $99/100$, we can compute $\mathbf{X}_\mathbf{C} = \arg\min_\mathbf{X} \|\mathbf{C}\mathbf{E} - \mathbf{X}\mathbf{S}^T\mathbf{E}\|_F^2$, where $\mathbf{E}$ is a leverage score sketching matrix with $O\left(\frac{k}{\epsilon}\right)$ columns, as shown in Lemma 8.3.3, and thus requires $O\left(\frac{nk}{\epsilon}\right)$ queries to $\mathbf{A}$. Then,

$$\begin{aligned}
\|\mathbf{C} - \mathbf{X}_\mathbf{C}\mathbf{S}^T\|_F^2 &\leqslant (1+\epsilon)\min_\mathbf{X} \|\mathbf{C} - \mathbf{X}\mathbf{S}^T\|_F^2 \\
&\leqslant (1+\epsilon)\|\mathbf{C} - \mathbf{C}\mathbf{S}\mathbf{S}^T\|_F^2 \\
&= \|\mathbf{C} - \mathbf{C}_k\|_F^2 \pm (\epsilon + \sqrt{\eta})\|\mathbf{C}\|_F^2
\end{aligned} \tag{8.90}$$

where the last two inequalities follow from equation 8.88. Let $\mathbf{X}_\mathbf{C}\mathbf{S}^T = \mathbf{U}'\mathbf{V}'^T$ be such that $\mathbf{U}'$ has orthonormal columns. Then, the column space of $\mathbf{U}'$ contains an approximately optimal solution for $\mathbf{A}$, since $\|\mathbf{C} - \mathbf{U}'\mathbf{V}'^T\|_F^2 = \|\mathbf{C} - \mathbf{C}_k\|_F^2 \pm \epsilon\|\mathbf{C}\|_F^2$ and $\mathbf{C}$ is a column PCP for $\mathbf{A}$. It follows from Lemma 8.4.9 that with probability at least $98/100$,

$$\|\mathbf{A} - \mathbf{U}'\mathbf{U}'^T\mathbf{A}\|_F^2 \leqslant \|\mathbf{A} - \mathbf{A}_k\|_F^2 + (\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F \tag{8.91}$$

Therefore, there exists a good solution for $\mathbf{A}$ in the column space of $\mathbf{U}'$. Since we cannot compute

this explicitly, we set up the following regression problem:

$$\min_{\mathbf{X}} \|\mathbf{A} - \mathbf{U}'\mathbf{X}\|_F^2 \tag{8.92}$$

Again, we sketch the regression problem above by sampling columns of $\mathbf{A}$ and apply Lemma 8.3.3. We can then compute $\mathbf{X}_\mathbf{A} = \arg\min_\mathbf{X} \|\mathbf{E}'\mathbf{A} - \mathbf{E}'\mathbf{U}'\mathbf{X}\|_F^2$ with probability at least $99/100$, where $\mathbf{E}'$ is a sketching matrix with $\left(\frac{k}{\epsilon}\right)$ rows and $O\left(\frac{nk}{\epsilon}\right)$ queries to $\mathbf{A}$. Then,

$$\begin{aligned}
\|\mathbf{A} - \mathbf{U}'\mathbf{X}_\mathbf{A}\|_F^2 &\leqslant (1+\epsilon)\min_\mathbf{X} \|\mathbf{A} - \mathbf{U}'\mathbf{X}\|_F^2 \\
&\leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{U}'\mathbf{U}'^T\mathbf{A}\|_F^2 \\
&\leqslant \|\mathbf{A} - \mathbf{A}_k\|_F^2 + O(\epsilon + \sqrt{\eta})\|\mathbf{A}\|_F^2
\end{aligned} \tag{8.93}$$

where the second inequality follows from $\mathbf{X}$ being the minimizer and $\mathbf{U}'^T\mathbf{A}$ being some other matrix, and the last inequality follows from equation 8.91. Recall, $\mathbf{U}'$ is an $n \times k$ matrix and the time taken to solve the regression problem is $O\left(nk + \text{poly}(k, \epsilon^{-1})\right)$.

Therefore, we observe that $\mathbf{U}'\mathbf{X}_\mathbf{A}$ suffices and we output it in factored form by setting $\mathbf{M} = \mathbf{U}'$ and $\mathbf{N} = \mathbf{X}_\mathbf{A}^T$. Union bounding over the probabilistic events, and rescaling ϵ, with probability at least $9/10$, Algorithm 12 outputs $\mathbf{M} \in \mathbf{R}^{n\times k}$ and $\mathbf{N}^T \in \mathbf{R}^{n\times k}$ such that the total number of entries queried in $\mathbf{A}$ are $\widetilde{O}\left(\frac{\phi_{\max}^2 nk}{\epsilon}\right)$. This concludes the proof of Theorem 183.

Correlation Matrices. We introduce low-rank approximation of correlation Matrices, a special case of PSD matrices where the diagonal is all 1s. Correlation matrices are well studied in numerical linear algebra, statistics and finance since an important statistic of n random variables $X_1, X_2, \ldots X_n$ is given by computing the pairwise correlation coefficient, $\text{corr}(X_i, X_j) = \text{cov}(X_i, X_j)/\sqrt{\text{var}(X_i) \cdot \text{var}(X_j)}$. A natural matrix representation of correlation coefficients results in a $n \times n$ correlation matrix $\mathbf{A}$ such that $\mathbf{A}_{i,j} = \text{corr}(X_i, X_j)$.

Definition 8.4.10. *(Correlation Matrices.)* $\mathbf{A}$ *is an* $n \times n$ *correlation matrix if* $\mathbf{A}$ *is PSD and* $\mathbf{A}_{i,i} = 1$*, for all* $i \in [n]$.

Often, in practice the correlation matrices obtained are close to being PSD, but corrupted by noise in the form of missing or asynchronous observations, stress testing or aggregation. Here the goal is to query few entries of the corrupted matrix and recover a rank-k matrix close to the underlying correlation matrix, assuming that the underlying matrix is also close to low rank to begin with.

Here we observe that since correlation matrices have all diagonal entries equal to 1, we can compute $\phi_{\max}$ by simply reading the diagonal entries of $\mathbf{A}+\mathbf{N}$. However, we can do even better since we can discard the diagonal entries of $\mathbf{A}+\mathbf{N}$. The main insight here is that for correlation matrices, our algorithm simply uniformly samples columns and rows to construct our row and column PCPs, since we know what the true diagonals should be. In this case, no matter what the adversary does to the diagonal, $\phi_{\max}=1$ and we obtain a $\widetilde{O}(nk/\epsilon)$ query algorithm.

Corollary 8.4.11. *(Robust LRA for Correlation Matrices.) Let k be an integer and $1>\epsilon>\eta>0$. Given $\mathbf{A}+\mathbf{N}$, where $\mathbf{A}$ is a correlation matrix and $\mathbf{N}$ is a corruption term such that $\|\mathbf{N}\|_F^2 \leqslant \eta\|\mathbf{A}\|_F^2$ and for all $i\in[n]$ $\|\mathbf{N}_{i,*}\|_2^2 \leqslant c\|\mathbf{A}_{i,*}\|_2^2$ for a fixed constant c, there exists an algorithm that samples $\widetilde{O}\left(nk/\epsilon\right)$ entries in $\mathbf{A}+\mathbf{N}$ and with probability at least $99/100$, computes a rank k matrix $\mathbf{B}$ such that*

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant \|\mathbf{A}-\mathbf{A}_k\|_F^2+(\epsilon+\sqrt{\eta})\|\mathbf{A}\|_F^2$$

Note, the sample complexity of this algorithm is optimal, since there is an $\Omega(nk/\epsilon)$ query lower bound for additive-error low-rank approximation of correlation matrices, even when there is no corruption (see Corollary 8.4.13).

Additive-Error PSD Low-Rank Approximation. In the limit where $\eta=0$, $\phi_{max}=1$, and we obtain an algorithm with query complexity $\widetilde{O}(nk/\epsilon)$. While this guarantee is already implied by our algorithm for *relative-error* low-rank approximation, our additive-error algorithm is simpler to implement, since the sampling probabilities can be computed *exactly* by simply reading the diagonal.

Corollary 8.4.12. *(Sample-Optimal Additive-Error LRA.) Given a PSD matrix $\mathbf{A}$, rank parameter k, and $\epsilon>0$, there exists an algorithm that samples $\widetilde{O}(nk/\epsilon)$ entries in $\mathbf{A}$ and outputs a rank-k matrix $\mathbf{B}$ such that with probability at least $99/100$,*

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \leqslant \|\mathbf{A}-\mathbf{A}_k\|_F^2+\epsilon\|\mathbf{A}\|_F^2$$

We show a matching lower bound on the query complexity of additive-error low-rank approximation of PSD matrices. Here, we simply observe that the lower bound construction introduced by [MW17c] of $\Omega\left(\frac{nk}{\epsilon}\right)$ also holds for additive error. As a consequence our algorithm is optimal in the setting where there is no corruption.

Corollary 8.4.13. *(Correlation Matrix Lower Bound, Theorem 13 [MW17c].) Let* $\mathbf{A}$ *be a PSD matrix,* $k \in \mathbb{Z}$ *and* $\epsilon > 0$ *be such that* $\frac{nk}{\epsilon} = o(n^2)$*. Any randomized algorithm,* $\mathcal{A}$*, that with probability at least* $2/3$*, computes a rank* k *matrix* $\mathbf{B}$ *such that*

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leqslant \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \epsilon\|\mathbf{A}\|_F^2$$

must read $\Omega\left(\frac{nk}{\epsilon}\right)$ *entries of* $\mathbf{A}$ *on some input, possibly adaptively, in expectation.*

Proof. We observe that in the lower bound construction of [MW17c], the matrix $\mathbf{A}$ is binary, with all 1s on a the diagonal, and k off-diagonal blocks of all 1s, each of size $\sqrt{\frac{2\epsilon n}{k}} \times \sqrt{\frac{2\epsilon n}{k}}$. Therefore, $\mathbf{A}$ is a correlation matrix and $\|\mathbf{A}\|_F^2 = (1 + 2\epsilon)n$. Further, the optimal rank-k cost, $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \Theta(n)$. To compute an additive-error approximation, any algorithm must caputure $\epsilon\|\mathbf{A}\|_F^2 = \epsilon n$ mass among the off-diagonal entries of $\mathbf{A}$. Note, the remaining proof is identical to Theorem 13 in [MW17c]. □

Chapter 9

Learning a Latent Simplex in Truly Input-Sparsity Time

9.1 Introduction

We study the problem of learning k vertices $\mathbf{M}_{*,1}, \ldots, \mathbf{M}_{*,k}$ of a latent k-dimensional simplex $\mathcal{K}$ in $\mathbb{R}^d$ using n data points generated from $\mathcal{K}$ and then possibly perturbed by a stochastic, deterministic, or adversarial source before given to the algorithm. In particular, the resulting points observed as input data could be heavily perturbed so that the initial points may no longer be discernible or they could be outside the simplex $\mathcal{K}$. Recent work of Bhattacharyya and Kannan [BK20c] unifies several stochastic models for unsupervised learning problems, including k-means clustering [CG92, GH+96, Web03, WT10, Dua20], topic models [BJ03, SG07, BL06a, Ble12, AGH+13a], mixed membership stochastic block models [ABFX08, MJG09, XFS+10, FSX09, ABEF14, LAW16, FXC16] and Non-negative Matrix Factorization [AGH+13b, GV14, Gil20] under the problem of learning a latent simplex. In general, identifying the latent simplex can be computationally intractable. However many special applications do not require the full generality. For example, in a mixture model like Gaussian mixtures, the data is assumed to be generated from a convex combination of density functions. Thus, it may be possible to efficiently approximately learn the latent simplex given certain distributional properties in these models.

Indeed, Bhattacharyya and Kannan showed that given certain reasonable geometric assumptions that are typically satisfied for real-world instances of Latent Dirichlet Allocation, Stochastic

Block Models and Clustering, there exists an $\widetilde{O}(k \cdot \mathsf{nnz}(\mathbf{A}))$ [1] time algorithm for recovering the vertices of the underlying simplex. We show that, given an additional natural assumption, we can remove the dependency on k and obtain a true input sparsity time algorithm. We begin by defining the model along with our new assumption:

Definition 9.1.1 (Latent Simplex Model)**.** *Let $\mathbf{M}$ be a $d \times k$ matrix such that $\mathbf{M}_{*,1}, \mathbf{M}_{*,2}, \ldots, \mathbf{M}_{*,k} \in \mathbb{R}^d$ denote the vertices of a k-simplex, $\mathcal{K}$. Let $\mathbf{P}$ be a $d \times n$ matrix such that $\mathbf{P}_{*,1}, \mathbf{P}_{*,2} \ldots \mathbf{P}_{*,n} \in \mathbb{R}^d$ are n points in the convex hull of $\mathcal{K}$. Given $\sigma > 0$, we observe a $d \times n$ matrix $\mathbf{A}$, such that $\|\mathbf{A} - \mathbf{P}\|_2 \leqslant \sigma\sqrt{n}$. Further, we make the following assumptions on the data generation process:*

1. ***Well-Separateness.*** *For all $\ell \in [k]$, $\mathbf{M}_{*,\ell}$ has non-trivial mass in the orthogonal complement of the span of the remaining vectors, i.e., for all $\ell \in [k]$, $|Proj(\mathbf{M}_{*,\ell}, Null(\mathbf{M} \setminus \mathbf{M}_{*,\ell}))| \geqslant \alpha \max_\ell \|\mathbf{M}_{*,\ell}\|_2$ where $Proj(x, U)$ denotes the orthogonal projection of x to the subspace U and $\mathbf{M} \setminus \mathbf{M}_{*,\ell}$ is the matrix $\mathbf{M}$ with the ℓ-th column removed.*

2. ***Proximate Latent Points.*** *Given $\delta \in (0, 1)$, for all $\ell \in [k]$, there exists a set $\mathcal{S}_\ell \subseteq [n]$ such that $|\mathcal{S}_\ell| \geqslant \delta n$ and for all $j \in \mathcal{S}_\ell$, $\|\mathbf{M}_{*,\ell} - \mathbf{P}_{*,j}\|_2 \leqslant 4\sigma/\delta$.*

3. ***Spectrally Bounded Perturbation.*** *The spectrum of $\mathbf{A} - \mathbf{P}$ is bounded, i.e., for a sufficiently large constant c, $\sigma/\sqrt{\delta} \leqslant \alpha^2 \min_\ell \|\mathbf{M}_{*,\ell}\|_2 / ck^9$.*

4. ***Significant Singular Values.*** *Let $\mathbf{A} = \sum_{i \in [d]} \sigma_i u_i v_i^T$ be the singular value decomposition and let $0 < \phi \leqslant \mathsf{nnz}(\mathbf{A})/(n \cdot poly(k))$. We assume that for all $i \in [k]$, $\sigma_i > \phi \cdot \sigma_{k+1}$ and $\|\mathbf{A} - \mathbf{A}_k\|_F^2 \leqslant \phi \|\mathbf{A} - \mathbf{A}_k\|_2^2$.*

These assumptions are natural across many interesting applications; see Section 9.2 for more details. [BK20c] introduced the Well-Separateness (1), Proximate Latent Points (2) and Spectrally Bounded Perturbation (3) assumptions. We include an additional Significant Singular Values assumption (4), which is crucial for obtaining a faster running time; we discuss this in more detail below. Our main algorithmic result can then be stated as follows:

Theorem 185 (Learning a Latent Simplex in Input-Sparsity Time)**.** *Given $k \geqslant 2$ and $\mathbf{A} \in \mathbb{R}^{d \times n}$ from the Latent Simplex Model (Definition 9.1.1), there exists an algorithm that runs in $\widetilde{O}\left(\mathsf{nnz}(\mathbf{A}) + (n + d) poly(k/\phi)\right)$ time to output subsets $\mathbf{A}_{\mathcal{R}_1}, \ldots, \mathbf{A}_{\mathcal{R}_k}$ such that upon permuting the columns of $\mathbf{M}$, with probability at least $1 - 1/\Omega(\sqrt{k})$, for all $\ell \in [k]$, we have $\|\mathbf{A}_{\mathcal{R}_\ell} - \mathbf{M}_{*,\ell}\|_2 \leqslant 300k^4\sigma/(\alpha\sqrt{\delta})$.*

[1] Throughout the paper we use the notation $\widetilde{O}$ to suppress poly-logarithmic factors.

Our result implies faster algorithms for various stochastic models that can be formulated as special cases of the Latent Simplex Model, including Latent Dirichlet Allocation for Topic Modeling, Mixed Membership Stochastic Block Models and Adversarial Clustering. We summarize the connections to these applications below. We describe our algorithm and provide an outline to our analysis; we defer all formal proofs to the supplementary material.

9.2 Connection to Stochastic Models

We first formalize the connection between the Latent Simplex Model (Definition 9.1.1) and numerous stochastic models. In particular, we show that topic models like Latent Dirichlet Allocation (LDA), Stochastic Block Models and Adversarial Clustering can be viewed as special cases of the Latent Simplex Model. We also show how our assumptions are natural in each of these applications.

9.2.1 Topic Models

Probabilistic Topic Models attempt to identify abstract topics in a collection of documents by discovering latent semantic structure [BJ03, BL06b, HBB10, ZAX12, Ble12]. Each document in the corpus is represented by a bag-of-words vectorization with the corresponding word frequencies. The standard statistical assumption is that the generative process for the corpus is a joint probability distribution over both the observed and hidden random variables. The hidden random variables can be interpreted as representative documents for each topic. The goal is to then design algorithms that can learn the underlying topics. The topics can be viewed geometrically as k latent vectors $\mathbf{M}_{*,1}, \mathbf{M}_{*,2}, \ldots, \mathbf{M}_{*,k} \in \mathbb{R}^d$, where d is the size of the dictionary and $\mathbf{M}_{i,\ell}$ is the expected frequency of word i in topic ℓ. Since each vector $\mathbf{M}_{*,\ell}$ represents a probability distribution, $\sum_i \mathbf{M}_{i,\ell} = 1$. Let $\mathbf{M}$ be the corresponding $d \times k$ matrix. One important stochastic model is Latent Dirichlet Allocation (LDA) [BNJ03], where each document consists of m words is generated as follows :

- For all $\ell \in [k]$, we pick topic weights $\mathbf{W}_{j,\ell} \sim \mathsf{Dir}(1/k)$, where $\mathsf{Dir}(1/k)$ is the Dirichlet distribution over the unit simplex. The topic distribution of document j is decided by the topic weights, $\mathbf{W}_{j,\ell}$, and given by $\mathbf{P}_{*,j} = \sum_{\ell \in [k]} \mathbf{W}_{j,\ell} \cdot \mathbf{M}_{*,\ell}$, where $\mathbf{P}_{*,j}$ are latent points.
- We then generate the j-th document with m words by taking i.i.d. samples from $\mathsf{Mult}(\mathbf{P}_{*,j})$, the multinomial distribution with $\mathbf{P}_{*,j}$ as the probability vector. The resulting document

observed is denoted by the vector $\mathbf{A}_{*,j}$, where for all $i \in [d]$ $\mathbf{A}_{i,j} = \frac{1}{m}\sum_{t=1}^{m} \mathbf{X}_{ij}^{(t)}$,, such that $\mathbf{X}_{ij}^{(t)} \sim \mathsf{Bern}(\mathbf{P}_{ij})$, where $\mathbf{X}_{ij}^{(t)} = 1$ if the i-th word was chosen in the t-th draw while generating the j-th document, and 0 otherwise.

The data generation process of LDA can be viewed as a special case of the Latent Simplex Model, where the j-th document is the data point $\mathbf{A}_{*,j}$ generated from the stochastic vector $\mathbf{P}_{*,j}$, a point in the simplex $\mathcal{K}$. The vertices of the simplex are the k topic vectors $\mathbf{M}_{*,1}, \ldots, \mathbf{M}_{*,k}$; the goal is then to recover the vertices of $\mathcal{K}$. [BK20c] remark that the Well-Separateness condition holds for LDA if we assume a Dirichlet prior on $\mathbf{M}$. We note that while $\mathcal{K}$ is a k-dimensional simplex, $d \ll k$ and the observed points need not lie inside the simplex. On the contrary, [BK20c] show that the data often lies significantly outside of $\mathcal{K}$. However, they show that the smoothed simplex obtained by taking the averages of all δn sized subsets of observed points results in a polytope $K_{\mathcal{S}}$ that is close to $\mathcal{K}$.

We formally justify our assumptions below.

Lemma 9.2.1 (LDA as a Latent Simplex)**.** *Given $\mathbf{A}, \mathbf{P}, \mathbf{M}$ following the LDA model as described above, such that for all $\ell \in [k]$, $\|\mathbf{M}_{*,\ell}\|_2 = \Omega(1)$, $m, n = \Omega(\text{poly}(k/\alpha))$ and $\delta = c\sigma/\sqrt{k}$, assumptions (2),(3) and (4) from Definition 9.1.1 are satisfied with high probability.*

Proof. Assumptions (2) and (3) follow from Lemma 7.1 in [BK20c]. By Claim 8.1 in [BK20c], $\sigma_k(\mathbf{A}) \geqslant c\alpha\sqrt{\delta/k}\min_\ell \mathbf{M}_{*,\ell}$. Each column of $\mathbf{A}$ sums to 1, so $\|\mathbf{A}\|_F^2 = O(n)$ and $\sigma_k(\mathbf{A}) \geqslant \alpha\sqrt{\delta/k}\|\mathbf{A}\|_F$. Since $\|\mathbf{A} - \mathbf{P}\|_2 \leqslant \sigma\sqrt{n}$ by definition of σ, and $\mathbf{P}$ consists of n point in the convex hull of k points and thus $\sigma_{k+1}(\mathbf{P}) = 0$, we have $\sigma_{k+1}(\mathbf{A}) \leqslant \sigma_{k+1}(\mathbf{P}) + \|\mathbf{A} - \mathbf{P}\|_2 \leqslant \sigma\sqrt{n} \leqslant \sigma\|\mathbf{A}\|_F$. Thus if $\sigma \leqslant \alpha\sqrt{\delta}/\text{poly}(k)$ for a large enough poly(k), our Significant Singular Values assumption holds. □

9.2.2 Mixed Membership Stochastic Block Models

The Stochastic Block Model is a well-studied stochastic model for generating random graphs, where the vertices are partitioned into k communities and edges within each community are more likely to occur than edges across communities. Given communities $C_1, C_2, \ldots C_k$, there exists a $k \times k$ symmetric latent matrix $\mathbf{B}$, where, $\mathbf{B}_{\ell_1,\ell_2}$ is the probability that there exists an edge between vertices in C_{ℓ_1} and C_{ℓ_2}. The MMBM can be formalized as the following stochastic process:

- For $j \in [n]$, vertex j picks a probability vector $\mathbf{W}_{*,j} \in \mathbb{R}^k$ representing community membership probabilities that sum to 1, i.e., $\mathbf{W}_{i,j} \sim \mathsf{Dir}(1/k)$ for all $i \in [k]$.

- For all pairs $(j_1, j_2) \in [n]$, vertex j_1 picks a community ℓ_1 proportional to $\mathsf{Mult}(\mathbf{W}_{*,j_1})$ and j_2 picks a community ℓ_2 proportional to $\mathsf{Mult}(\mathbf{W}_{*,j_2})$. The edge (j_1, j_2) is included in the graph with probability $\mathbf{B}_{\ell_1,\ell_2}$. Since $\sum_{\ell_1,\ell_2} \mathbf{W}_{\ell_1,j_1}\mathbf{B}_{\ell_1,\ell_2}\mathbf{W}_{\ell_2,j_2}$ represents the edge probability of the edge (j_1, j_2), the latent variable matrix $\mathbf{P}$ of edge probabilities can be represented as $\mathbf{P} = \mathbf{W}^T\mathbf{B}\mathbf{W}^T$.

However, our reduction is not straightforward since now $\mathbf{P}$ depends quadratically on $\mathbf{W}$ and the only polynomial time algorithms for $\mathbf{B}$ directly rely on semidefinite programming. Further, they require non-degeneracy assumptions in order to compute a tensor decomposition provably in polynomial time [AGHK14b, HS17]. However, we can pose the problem of recovery of the k underlying communities differently and first pick at random a subset $V_1 \subset [n]$ of d vertices and represent the ℓ-th community by a d-dimensional vector that represents the probabilities of vertices in $[n] \setminus V_1$ belonging to community ℓ and having an edge with each of the d vertices in V_1. We now define $\mathbf{W}_{(1)}$ to be a $k \times d$ matrix representing the fractional membership of weights of vertices in V_1 and $\mathbf{W}_{(2)}$ to be the analogous $k \times n$ matrix for vertices in $[n] \setminus V_1$. Observe that the probability matrix $\mathbf{P}$ can now be represented as $\mathbf{W}_{(1)}^T\mathbf{B}\mathbf{W}_{(2)}$.

The reduction to the Latent Simplex Model can now be stated as follows: given a data matrix $\mathbf{A}$ which is the adjacency matrix of the community graph, and the latent variable matrix $\mathbf{P}$, recover the simplex $\mathbf{M} = \mathbf{W}_{(1)}^T\mathbf{B}$. Further, [ABFX08] assumes that each column of $\mathbf{W}_{(2)}$ is picked from the Dirichlet distribution with parameter $1/k$. Combined with tools from random matrix theory [Ver10a], [BK20c] (Lemma 7.2) shows that the Proximate Latent Points and Spectrally Bounded assumptions hold for Stochastic Block Models. As for the Significant Singular Values assumption, it is satisfied when σ is a small enough polynomial in k.

Justifying Significant Singular Values. We give the following further justification for assumption (4) in Section 9.5: a faster algorithm only using the assumptions appearing in [BK20c] would imply an algorithmic breakthrough for spectral low-rank approximation and partially resolve the first open question of [Woo14b].

Theorem 186 (Spectral LRA and Learning a Simplex (informal))**.** *There exists a distribution over instances such that learning a latent simplex in $o(nnz(\mathbf{A}) \cdot k)$ time with good probability implies a constant factor spectral low-rank approximation algorithm in the same running time.*

9.2.3 Adversarial Clustering

We consider clustering problems that arise naturally from stochastic mixture models such as Gaussian, Mallows, categorical and so on [SK01, VW04, LB11, CSV17, DKS18, LM18b]. We can then formulate such a clustering problem in the Latent Simplex Model as follows: Given n data points $\mathbf{A}_{*,1}, \mathbf{A}_{*,2}, \ldots, \mathbf{A}_{*,n} \in \mathbb{R}^d$, such that the data is a mixture of k distinct clusters, $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k$, with means $\mathbf{M}_{*,1}, \mathbf{M}_{*,2}, \ldots, \mathbf{M}_{*,k}$, the goal is to approximately learn the means. Further, we can set the n latent vectors $\mathbf{P}_{*,j}$ to denote the mean of the cluster point $\mathbf{A}_{*,j}$ belongs to, and thus $\mathbf{P}_{*,j} \in \{\mathbf{M}_{*,1}, \mathbf{M}_{*,2}, \ldots, \mathbf{M}_{*,k}\}$. Prior work of [KK10] and [AS12] shows that if the minimum cluster size if δn and for all $\ell \neq \ell'$, $\|\mathbf{M}_{*,\ell} - \mathbf{M}_{*,\ell'}\| \geqslant ck\frac{\sigma}{\sqrt{\delta}}$ the $\mathbf{M}_{*,\ell}$ can be found within error $O(\sqrt{k}\sigma/\sqrt{\delta})$.

However, the aforementioned algorithms are not robust to adversarial perturbations. Therefore, we describe the perturbations we can handle in the Latent Simplex Model. The adversarial model is the same as the one considered in [BK20c]. The adversary is allowed to selected a subset S_ℓ of each cluster $\mathbf{C}_\ell$ of cardinality at most δn and perturb each point $\mathbf{A}_{*,j}$ for $j \in S_\ell$ by Δ_j such that :

- $\mathbf{P}_{*,j} + \Delta_j$ is still in the Convex Hull of $\mathbf{M}_{*,1}, \mathbf{M}_{*,2}, \ldots, \mathbf{M}_{*,k}$
- The norm of the perturbation is bounded, i.e., $|\Delta_j|_2 \leqslant 4\sigma/\sqrt{\delta}$.

Intuitively, the adversary can move a $1-\delta$ fraction of the data points in each cluster an arbitrary amount towards the convex hull of the means of the remaining clusters. For the remaining δn, the perturbation should have norm at most $O(\sigma/\sqrt{\delta})$. The goal is to still learn the means $\mathbf{M}_{*,\ell}$ approximately. [BK20c] shows that the aforementioned model satisfies Well-Separateness, Proximate Latent Points and Spectrally Bounded Perturbations assumptions. The proof for the Significant Singular Values assumption follows from Lemma 9.2.1. We note that there has been a flurry of recent progress on adversarial clustering in the strong contamination model, where the input data points are sampled from a mixture of Gaussians distribution and the adversary can corrupt a small fraction of the samples arbitrarily [DKS18, HL18, KSS18, DHKK20, BK20b]. In our setting, there is no distribution assumption on the data points but the adversary is constrained as the norm of the perturbation is bounded.

9.2.4 Preliminaries

We use n, d, and k to denote the number of data points, the number of dimensions of the space and the number of vertices of $\mathcal{K}$ respectively. We use the notation $\mathbf{A}_{*,j}$ to denote the j-th column of matrix $\mathbf{A}$. For $\mathbf{A} \in \mathbb{R}^{d \times n}$ with rank r, its singular value decomposition, denoted by $\texttt{SVD}(\mathbf{A}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, guarantees that $\mathbf{U}$ is a $d \times r$ matrix with orthonormal columns, $\mathbf{V}^T$ is an $r \times n$ matrix with orthonormal rows and $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix. The diagonal entries of $\mathbf{\Sigma}$ are the singular values of $\mathbf{A}$, denoted by $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_r$. Given an integer $k \leqslant r$, we define the truncated singular value decomposition of $\mathbf{A}$ that zeros out all but the top k singular values of $\mathbf{A}$, i.e., $\mathbf{A}_k = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^T$, where $\mathbf{\Sigma}_k$ has only k non-zero entries along the diagonal. It is well-known that the truncated SVD computes the best rank-k approximation to $\mathbf{A}$ under the Frobenius norm, i.e., $\mathbf{A}_k = \min_{\text{rank}(\mathbf{X}) \leqslant k} \|\mathbf{A} - \mathbf{X}\|_F$. Given an orthonormal basis $\mathbf{U}$ for a subspace, we use $\mathbf{P}_\mathbf{U} = \mathbf{U}\mathbf{U}^T$ to denote the projection matrix corresponding to the subspace. We consider the following notion of subspace distance:

Definition 9.2.2 ($\sin\Theta$ Distance)**.** *For any two subspaces $\mathbf{R}$, $\mathbf{S}$ of $\mathbb{R}^d$, the $\sin\Theta$ distance between $\mathbf{R}$ and $\mathbf{S}$ is defined as*

$$\sin\Theta(\mathbf{R}, \mathbf{S}) = \max_{u \in \mathbf{R}} \min_{v \in \mathbf{S}} \sin\theta(u, v) = \max_{u \in \mathbf{R}, |u|=1} \min_{v \in \mathbf{S}} \|u - v\|.$$

We use the notion of spectral low-rank approximation to obtain a compact representation of the input and compute matrix-vector products efficiently. We also require the notion of mixed spectral-Frobenius low-rank approximation. This guarantee is weaker than spectral-low rank approximation but admits faster algorithms and has been recently used in several sublinear time algorithms [MW17b, BCW20b].

Definition 9.2.3 (Spectral Low-rank Approximation, Spectral-Frobenius Low-rank Approximation)**.** *Given a matrix $\mathbf{A}$, an integer k and $\epsilon > 0$, a rank-k matrix $\mathbf{B}$ satisfies a relative-error spectral low-rank approximation guarantee if $\|\mathbf{A} - \mathbf{B}\|_2^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_2^2$. $\mathbf{B}$ satisfies a mixed spectral-Frobenius low-rank approximation guarantee if*

$$\|\mathbf{A} - \mathbf{B}\|_2^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_2^2 + \frac{\epsilon}{k}\|\mathbf{A} - \mathbf{A}_k\|_F^2.$$

9.3 Technical Overview

In this section, we provide an overview of our algorithmic techniques and discuss the main challenges we overcome to obtain an input-sparsity time algorithm.

Our Techniques. The starting point in [BK20c] is that the smoothened polytope, obtained by averaging points in the data matrix $\mathbf{A}$ is itself close to the latent points in the convex hull of $\mathcal{K}$ in operator norm. This fact is captured by the following lemma:

Lemma 9.3.1 (Subset Smoothing)**.** *For any $\mathcal{S} \subset [n]$, let $\mathbf{A}_{\mathcal{S}}$ be a vector obtained by averaging the columns of $\mathbf{A}$ indexed by $\mathcal{S}$ and define $\mathbf{P}_{\mathcal{S}}$ similarly. Then for $\|\mathbf{A} - \mathbf{P}\|_2 \leqslant \sigma\sqrt{n}$, we have $\|\mathbf{A}_{\mathcal{S}} - \mathbf{P}_{\mathcal{S}}\|_2 \leqslant \sigma\sqrt{n/|\mathcal{S}|}$.*

Our main insight is that we can approximately optimize a linear function on the smoothed polytope by working with a rank-k spectral approximation to $\mathbf{A}$ instead. Geometrically, this implies that while the smoothed polytope is perhaps d-dimensional, projecting it onto the k-dimensional space spanned by the top-k singular values of the data matrix $\mathbf{A}$ suffices to recover the latent k-simplex, $\mathcal{K}$. This is surprising since the data matrix can contain points significantly far from the latent polytope. Further, this approach presents several challenges: we do not have access to the left singular space of $\mathbf{A}$ and even if we are provided this subspace exactly, it is unclear why it spans a set of points that approximate vertices of $\mathcal{K}$. Finally, the points obtained by smoothing the projected polytope have no immediate relation to points in the smoothed high-dimensional polytope considered by [BK20c].

We would like to begin by computing a spectral low-rank approximation (Definition 9.2.3) for $\mathbf{A}$. Since a low-rank approximation to $\mathbf{A}$ can be represented in factored form $\mathbf{Y}\mathbf{Z}^T$, where $\mathbf{Y}$ is $d \times k$ and $\mathbf{Z}^T$ is $k \times n$, any matrix-vector product of the form $\mathbf{Y}\mathbf{Z}^T \cdot x$ only requires $(n + d)k$ time. Thus optimizing a linear function k times over a smoothed low-rank polytope requires only $(n + d)k^2$ time, circumventing the previous bound of $k \cdot \mathsf{nnz}(\mathbf{A})$. However, the best known algorithm for spectral low-rank approximation (Theorem 1 in [MM15]) requires $\tilde{O}(\mathsf{nnz}(\mathbf{A}) \cdot k/\sqrt{\epsilon})$ time and thus provides no improvement. A natural direction to pursue is then to compute a Frobenius low-rank approximation (which requires $\mathsf{nnz}(\mathbf{A})$ time) for $\mathbf{A}$ and use this as our proxy. However, a Frobenius low-rank approximation is too coarse to obtain a subspace that is close to the top-k singular vectors of $\mathbf{A}$.

Instead we compute a mixed spectral-Frobenius low-rank approximation (see Definition 9.2.3) that runs in $O(\mathsf{nnz}(A) + dk^2)$ time, but the resulting error guarantee is weaker. In particular, it

incurs an additive $\epsilon\|\mathbf{A}-\mathbf{A}_k\|_F^2/k$ term. Here, we use the assumption we introduced (the Significant Singular Value assumption) to show that the low-rank matrix obtained from this algorithm also satisfies a *relative-error* spectral low-rank approximation guarantee. The next challenge is that the aforementioned guarantee only bounds the spectral norm of $\mathbf{A}-\mathbf{Y}\mathbf{Z}^T$ in terms of the $(k+1)$-st singular value of $\mathbf{A}$. This guarantee does not relate how close the subspaces spanned by the columns and rows of the low-rank approximation are to the top-k singular space of $\mathbf{A}$.

A key technical contribution of our work is thus to prove that the subspaces obtained via spectral low-rank approximation are close to the true left and right top-k singular space in angular ($\sin\Theta$) distance. We note that such a guarantee is crucial to approximately optimize a linear function over $\mathbf{A}$. Further, this result provides an intriguing connection between spectral low-rank approximation and power iteration. It is well known that power iteration suffices to obtain a subspace that is close to the top-k subspace of a matrix in $\sin\Theta$ distance, which at first glance appears much stronger than spectral low-rank approximation. However, our work implies that it suffices to compute a spectral low-rank approximation, which provides a succinct representation of the data matrix and can be computed faster than power iteration in several natural settings.

Algorithm 13 : Learning a Latent k-Simplex in Input Sparsity Time

Input: A matrix $\mathbf{A}\in\mathbb{R}^{d\times n}$, integer k, and $\epsilon>0$.

1. Using the algorithm from Lemma 9.4.1, compute rank-k matrices $\mathbf{Y},\mathbf{Z}$ such that $\mathbf{Y}\mathbf{Z}^T$ is a spectral low-rank approximation to $\mathbf{A}$, i.e., $\|\mathbf{A}-\mathbf{Y}\mathbf{Z}^T\|_2^2\leqslant(1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_2^2$.
2. Let $\mathcal{S}=\{\emptyset\}$. For each $t\in[k]$,
   (a) Let $\mathbf{U}_t$ be an orthonormal basis for the vectors in $\mathcal{S}$.
   (b) Compute the projection matrix $\mathbf{P}_t=\mathbf{U}_t\mathbf{U}_t^T$ that projects onto the row span of $\mathcal{S}$.
   (c) Let $g\sim\mathcal{N}(0,\mathbf{I}_k)$ and let $\mathbf{u}_t=g\mathbf{Y}^T(\mathbf{I}_d-\mathbf{P}_t)\mathbf{Y}\mathbf{Z}^T$ be a random vector in $\mathbb{R}^n$. Compute $\mathcal{R}_t\subset[n]$, a subset of δn indices corresponding to the largest coordinates of $\mathbf{u}_t$ in absolute value.
   (d) Let $\mathbf{A}_{\mathcal{R}_t}$ be the average of the columns of $\mathbf{A}$ indexed by $\mathcal{R}_t$. Update $\mathcal{S}=\mathcal{S}\cup\mathbf{A}_{R_t}$.

Output: The set of vectors $\mathbf{A}_{\mathcal{R}_1},\mathbf{A}_{\mathcal{R}_2},\ldots,\mathbf{A}_{\mathcal{R}_k}$ as our approximation to the vertices of the latent k-simplex $\mathcal{K}$.

In the context of learning the latent simplex, given a spectral low-rank approximation, $\mathbf{Y}\mathbf{Z}^T$, we first restrict to the column span of $\mathbf{Y}$, which w.l.o.g. has orthonormal columns, and iteratively

generate k vectors in this subspace. In the first iteration, we generate a random vector $g\mathbf{Y}^T$ and compute $g\mathbf{Y}^T\mathbf{Y}\mathbf{Z}^T$. We then consider the largest δn indices of $g\mathbf{Y}^T\mathbf{Y}\mathbf{Z}^T$. While the resulting vector does not have strong provable guarantees, we show that averaging the columns of $\mathbf{A}$ corresponding to these indices results in a vector, $\mathbf{A}_{\mathcal{R}_1}$, which intuitively corresponds to efficiently optimizing a linear function over a low-rank approximation to the smoothened polytope, where the smoothened polytope is obtained by averaging over all subsets of δn data points. Our next contribution is to show that $\mathbf{A}_{\mathcal{R}_1}$ obtained by the aforementioned algorithmic process is indeed close to a vertex of $\mathcal{K}$.

To obtain an approximation to the remaining vertices of $\mathcal{K}$, we consider the following iterative process: in the t-th iteration, consider the subspace $\mathbf{Y}^T(\mathbf{I}-\mathbf{P}_t)$, where $(\mathbf{I}-\mathbf{P}_t)$ is the projection onto the orthogonal complement of the span of $\mathbf{A}_{\mathcal{R}_1}, \mathbf{A}_{\mathcal{R}_2} \ldots \mathbf{A}_{\mathcal{R}_{t-1}}$. Then generate a random vector $g\mathbf{Y}^T(\mathbf{I}-\mathbf{P}_t)$, and compute the largest δn coordinates of $g\mathbf{Y}^T(\mathbf{I}-\mathbf{P}_t)\mathbf{Y}\mathbf{Z}^T$. Average the corresponding columns of $\mathbf{A}$ to obtain $\mathbf{A}_{\mathcal{R}_t}$ and output this vector. We prove that after iterating k times, the vectors $\mathbf{A}_{\mathcal{R}_1}, \mathbf{A}_{\mathcal{R}_1}, \ldots \mathbf{A}_{\mathcal{R}_k}$ approximate all the vertices of the latent simplex $\mathcal{K}$ within the desired accuracy and running time.

In contrast, prior work of [BK20c] uses power iteration to approximate the left top-k singular space $\mathbf{U}_k$ of $\mathbf{A}$ using a subspace $\widehat{\mathbf{V}}$ that is poly(α/k) close in $\sin\Theta$ distance. Each step of the power iteration uses $O(\mathsf{nnz}(\mathbf{A}) + dk^2)$ time and is repeated $\log(d)$ times. Next, they pick a random vector u_1 in the subspace spanned $\widehat{\mathbf{V}}$ and compute $\mathbf{A}_{\mathcal{R}_1} = \arg\max_{\mathcal{S}:|\mathcal{S}|=\delta n} |u_1 \cdot \mathbf{A}_{\mathcal{S}}|$, using the resulting vector as an approximation to some vertex $\mathbf{M}_{*,1}$.

They then repeat the above algorithm k times and in the i-th iteration, they pick u_i to be a uniformly random direction in the $k-i$ dimensional subspace constructed as follows: let $\widetilde{\mathbf{V}}_{i-1}$ be an orthonormal basis for $\mathbf{A}_{\mathcal{R}_1}, \mathbf{A}_{\mathcal{R}_2}, \ldots, \mathbf{A}_{\mathcal{R}_{i-1}}$. Intuitively, this corresponds to sampling a random vector from the subspace orthogonal to the set of vertex approximations picked thus far. The resulting k vectors $\mathbf{A}_{\mathcal{R}_1}, \ldots, \mathbf{A}_{\mathcal{R}_k}$ are the approximation to the vertices of the latent simplex. Since they directly optimize over the smoothened polytope, the correctness analysis is more straightforward.

However, each iteration of the algorithm requires optimizing a linear function over the smoothened polytope and in particular requires computing $u_i \cdot \mathbf{A}$, and thus, the overall running time is dominated by $k \cdot \mathsf{nnz}(\mathbf{A})$. Since the latent simplex satisfies the Well-Separateness condition, the inner product with a random direction is maximized by a unique vertex. Intuitively, it appears necessary to project away from the set of vectors obtained up to the i-th iteration in order to learn new vertices of $\mathcal{K}$. The inherently iterative nature of the algorithm combined with matrix-vector product lower bounds indicates that the new algorithmic ideas we introduce are in fact necessary.

9.4 Full Analysis

In this section, we analyze Algorithm 13 and show that it outputs a set of k vectors that approximate the vertices of the latent simplex K. Formally, the main theorem we prove is as follows:

Theorem 185 *(Restated.)* Given input data $\mathbf{A}$ from the Latent Simplex Model, there exists Algorithm 13 that takes $\widetilde{O}\left(\mathsf{nnz}(\mathbf{A}) + (n+d)\mathrm{poly}(k)\right)$ time to output k vectors $\mathcal{R}_1, \ldots, \mathcal{R}_k$ such that upon permuting the columns of $\mathbf{M}$, for all $\ell \in [k]$, we have

$$\|\mathcal{R}_\ell - \mathbf{M}_{*,\ell}\|_2 \leqslant \frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}},$$

with probability at least $1 - \frac{1}{\Omega(\sqrt{k})}$.

We start with a spectral low-rank approximation for $\mathbf{A}$. We then use the right factor as an approximation to $\mathbf{\Sigma}_k\mathbf{V}_k^T$ and the left factor as an approximation to $\mathbf{U}_k$.

Lemma 9.4.1. *(Input-Sparsity Spectral LRA [CEM+15, CMM17].) Given a matrix $\mathbf{A} \in \mathbb{R}^{d\times n}$, $\epsilon, \delta > 0$ and $k \in \mathbb{N}$, there exists an algorithm that outputs matrices $\mathbf{Y}, \mathbf{Z}$, such that with probability at least $1-\delta$, $\|\mathbf{A} - \mathbf{Y}\mathbf{Z}^T\|_2^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_2^2 + \frac{\epsilon}{k}\|\mathbf{A} - \mathbf{A}_k\|_F^2$, in time $\widetilde{O}\left(\mathit{nnz}(\mathbf{A}) + (n+d)\mathit{poly}(k/\epsilon\delta)\right)$.*

Under the Significant Singular Values condition (4), setting $\epsilon = \phi$ in Lemma 9.4.1 implies with probability $99/100$,

$$\frac{1}{\mathrm{poly}(k)}\sum_{i=k+1}^{n}\sigma_i^2 = \frac{1}{\mathrm{poly}(k)}\|\mathbf{A} - \mathbf{A}_k\|_F^2 \leqslant \sigma_{k+1}^2 = \|\mathbf{A} - \mathbf{A}_k\|_2^2 \tag{9.1}$$

and thus $\|\mathbf{A} - \mathbf{Y}\mathbf{Z}^T\|_2^2 \leqslant 2\|\mathbf{A} - \mathbf{A}_k\|_2^2$. Further, the aforementioned lemma implies such a matrix $\mathbf{Y}\mathbf{Z}^T$ can be computed in $\widetilde{O}\left(\mathsf{nnz}(\mathbf{A}) + (n+d)\mathrm{poly}(k/\phi)\right)$ time. Thus the Well-Separateness condition immediately implies that the algorithm from Lemma 9.4.1 is a spectral low-rank approximation.

Next, we show that if $\mathbf{Y}\mathbf{Z}^T$ is a good rank k spectral approximation to $\mathbf{A}$, then the subspace spanned by the columns of $\mathbf{Y}$ must be close to the column span of $\mathbf{U}_k$, the top-k left singular vectors of $\mathbf{A}$. In fact, the subspace $\mathbf{Y}$ obtained via spectral low-rank approximation is a good approximation to the subspace $\mathbf{U}_k$ in angular distance. The appropriate measure of angular distance between subspaces can be formalized as the principal angle between the subspaces and

the corresponding $\sin\Theta$ function. Wedin [Wed72] bounded the $\sin\Theta$ between the SVD subspace of a matrix and the SVD subspace of a slight perturbation of the matrix.

Theorem 187 (Wedin's $\sin\Theta$ theorem [Wed72])**.** *Let $\mathbf{R}, \mathbf{S} \in \mathbb{R}^{d\times n}$ and $0 < m \leqslant \ell$ be integers. Let $\mathbf{R}_m$ and $\mathcal{S}\sigma_\ell$ denote the subspaces spanned by the top m singular vectors of $\mathbf{R}$ and top ℓ singular vectors of $\mathbf{S}$, respectively. Suppose $\gamma = \sigma_m(\mathbf{R}) - \sigma_{\ell+1}(\mathbf{S})$. Then*

$$\sin\Theta(\mathbf{R}_m, \mathcal{S}\sigma_\ell) \leqslant \frac{\|\mathbf{R}-\mathbf{S}\|_2}{\gamma}.$$

Bhattacharyya and Kannan [BK20c] use Wedin's $\sin\Theta$ theorem to measure the distance between the subspace $\mathbf{U}_k$ spanned by the top k left singular vectors of $\mathbf{A}$ and the subspace returned by their iterative subspace power method. Since we create the sketch $\mathbf{Y}$ for $\mathbf{U}_k$, we would instead like to argue that $\mathbf{Y}$ and $\mathbf{U}_k$ are close in $\sin\Theta$ distance.

Lemma 9.4.2 (Proximity of Subspace Projections)**.** *Let $\mathbf{Y}$ be defined as in Algorithm 13 and let $\mathbf{U}_k$ be the subspace spanned by the top k left singular vectors of $\mathbf{A}$. Let $\mathbf{P}_\mathbf{Y}$ and $\mathbf{P}_{\mathbf{U}_k}$ be the $d \times d$ projection matrices onto the row span of $\mathbf{Y}$ and $\mathbf{U}_k$. Then $\|\mathbf{P}_\mathbf{Y} - \mathbf{P}_{\mathbf{U}_k}\|_2 \leqslant \frac{1}{1000k^{10}}$.*

Proof. Suppose by way of contradiction that $\|\mathbf{P}_\mathbf{Y} - \mathbf{P}_{\mathbf{U}_k}\|_2 \geqslant \frac{1}{1000k^{10}}$. Note that since $\mathbf{Y}$ and $\mathbf{U}_k$ are each orthonormal matrices with rank k, then

$$\|\mathbf{U}_k\mathbf{U}_k^T - \mathbf{Y}\mathbf{Y}^T\|_F^2 \geqslant \|\mathbf{U}_k\mathbf{U}_k^T - \mathbf{Y}\mathbf{Y}^T\|_2^2 \geqslant \frac{1}{(1000k^{10})^2}$$

so that

$$\begin{aligned}\|\mathbf{U}_k\mathbf{U}_k^T - \mathbf{Y}\mathbf{Y}^T\|_F^2 &= \|\mathbf{U}_k\|_F^2 + \|\mathbf{Y}\|_F^2 - 2\|\mathbf{U}_k\mathbf{Y}^T\|_F^2 \\ &= 2k - 2\|\mathbf{U}_k\mathbf{Y}^T\|_F^2 \geqslant \frac{1}{(1000k^{10})^2}\end{aligned}$$

Hence, $\|\mathbf{U}_k\mathbf{Y}^T\|_F^2 \leqslant k - \frac{1}{(1000k^{10})^2}$. Now we would like to show for the sake of contradiction that $\|\mathbf{A} - \mathbf{P}_\mathbf{Y}\mathbf{A}\|_2$ is large. Thus, for the singular value decomposition $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$, we write

$$\begin{aligned}\|\mathbf{A} - \mathbf{P}_\mathbf{Y}\mathbf{A}\|_2 &= \|\mathbf{U}^T\Sigma - \mathbf{Y}\mathbf{Y}^T\mathbf{U}^T\Sigma\|_2 \\ &\geqslant \|\mathbf{U}_k\mathbf{U}^T\Sigma - \mathbf{U}_k\mathbf{Y}\mathbf{Y}^T\mathbf{U}^T\Sigma\|_2\end{aligned}$$

since $\|\mathbf{U}_k\|_2 \leqslant \|\mathbf{U}\|_2 \leqslant 1$. Thus, there exist matrices $\mathbf{C}_1, \mathbf{C}_2$ such that

$$\mathbf{U}_k\mathbf{U}^T\Sigma - \mathbf{U}_k\mathbf{Y}\mathbf{Y}^T\mathbf{U}^T\Sigma = \begin{bmatrix}\mathbf{C}_1 & \mathbf{C}_2\end{bmatrix}\begin{bmatrix}\Sigma_k & 0 \\ 0 & \Sigma_{n-k}\end{bmatrix},$$

where Σ_k is the diagonal matrix consisting of the top k singular values of $\mathbf{A}$ and Σ_{n-k} is the diagonal matrix consisting of the bottom $n-k$ singular values of $\mathbf{A}$. Now we know that one of the top k eigenvalues of $\mathbf{U}_k^T\mathbf{Y}\mathbf{Y}^T\mathbf{U}_k$ is at most $1 - \frac{1}{(1000k^{10})^2}$. Thus, one of the top k eigenvalues of $\mathbf{I}_k - \mathbf{C}_1$ is at least $\frac{1}{(1000k^{10})^2}$. In particular, let λ be such an eigenvalue and let $\mathbf{x}$ be the corresponding unit eigenvector of $\mathbf{I} - \mathbf{C}_1$. Then we have

$$\|\mathbf{U}_k\mathbf{U}^T\Sigma - \mathbf{U}_k\mathbf{Y}\mathbf{Y}^T\mathbf{U}^T\Sigma\|_2 \geqslant \|(\mathbf{I} - \mathbf{C}_1)\Sigma_k\mathbf{x}\|_2 \geqslant \sigma_k(\mathbf{A})\lambda \geqslant \frac{1}{(1000k^{10})^2}\sigma_k(\mathbf{A}).$$

Since the Significant Singular Values assumption implies that $\frac{1}{(1000k^{10})^2}\sigma_k(\mathbf{A}) > (1+\epsilon)\sigma_{k+1}(\mathbf{A})$, this implies that $\|\mathbf{A} - \mathbf{P}_\mathbf{Y}\mathbf{A}\|_2 > (1+\epsilon)\sigma_{k+1}(\mathbf{A})$, which contradicts the assumption that $\mathbf{Y}$ is a good low-rank approximation to $\mathbf{A}$. Thus we have $\|\mathbf{P}_\mathbf{Y} - \mathbf{P}_{\mathbf{U}_k}\|_2 \leqslant \frac{1}{1000k^{10}}$, as desired. □

Our analysis proceeds via induction on the number of iterations performed by the algorithm. Suppose our algorithm has selected t points from our approximation of the top k subspace and these points are reasonably close to i points of the k-simplex. In the $(t+1)$-st iteration, we again bound the $\sin\Theta$ distance between $\mathbf{Y}^T(\mathbf{I}-\mathbf{P}_t)$, which corresponds to our approximation of the top k subspace projected away from the selected vectors, and the actual k-simplex projected away from the corresponding points closest to our selected vectors. This argues that we can continue selecting random vectors in the subspace spanned by $\mathbf{Y}^T(\mathbf{I} - \mathbf{P}_t)$ as a close approximation to random vectors in $\mathbf{M}(\mathbf{I} - \mathbf{P}_t)$.

We first bound the k-th singular values of the simplex vertices ($\mathbf{M}$) and latent variables ($\mathbf{P}$), leveraging the Well-Separateness and Spectrally Bounded Perturbations assumptions.

Lemma 9.4.3 (Claim 8.1 in [BK20c])**.** *If the underlying points* $\mathbf{M}$ *follow the Well-Separateness and Spectrally Bounded Perturbation assumptions, then*

$$\sigma_k(\mathbf{M}) \geqslant \frac{1000k^{8.5}}{\alpha^2}\frac{\sigma}{\sqrt{\delta}}, \qquad \sigma_k(\mathbf{P}) \geqslant \frac{995k^{8.5}\sqrt{n}}{\alpha^2}\sigma.$$

We can then upper bound $\sin\Theta$ distance between $\mathbf{Y}$ and $\mathbf{U}_k$ as follows:

Corollary 9.4.4. *Let* $\mathbf{Y}$ *be defined as in Algorithm 13 and let* $\mathbf{U}_k$ *be the subspace spanned by*

the top k left singular vectors of $\mathbf{A}$. Then $\sin\Theta(\mathbf{Y},\mathbf{U}_k) \leqslant \frac{1}{1000k^{10}}$.

Proof. By setting $m = k = \ell$ in Theorem 187, we have

$$\begin{aligned}\sin\Theta(\mathbf{Y},\mathbf{U}_k) &= \sin\Theta(\mathbf{P}_{\mathbf{Y}},\mathbf{P}_{\mathbf{U}_k}) \\ &\leqslant \frac{\|\mathbf{P}_{\mathbf{Y}} - \mathbf{P}_{\mathbf{U}_k}\|_2}{\sigma_k(\mathbf{Y}) - \sigma_{k+1}(\mathbf{U}_k)}.\end{aligned}$$

By definition of σ, we have that $\|\mathbf{A}-\mathbf{P}\|_2 \leqslant \sigma\sqrt{n}$. Thus, Lemma 9.4.3 implies that $\sigma_k(\mathbf{A}) \gg 1$. Since $\mathbf{Y}$ has rank k, we have $\sigma_{k+1}(\mathbf{Y}) = 0$. By Lemma 9.4.2, $\sin\Theta(\mathbf{Y},\mathbf{U}_k) \leqslant \|\mathbf{P}_{\mathbf{Y}} - \mathbf{P}_{\mathbf{U}_k}\|_2 \leqslant \frac{1}{1000k^{10}}$. □

They also showed that vectors in $\mathbf{U}_k$ are close to the subspace $\mathbf{M}$:

Lemma 9.4.5. *[BK20c] Let $\mathbf{U}_k$ be the subspace spanned by the top k left singular vectors of $\mathbf{A}$ and let $\mathbf{R}$ be any k-dimensional subspace of $\mathbb{R}^d$ with $\sin\Theta(\mathbf{U}_k,\mathbf{R}) \leqslant \frac{\alpha^2}{1001k^9}$. Let $\mathbf{M}$ be the underlying latent k-simplex. Then for each unit vector $\mathbf{x} \in \mathbf{R}$, there exists a vector $\mathbf{y} \in$ **Span** $(\mathbf{M})$ with $\|\mathbf{x}-\mathbf{y}\|_2 \leqslant \frac{\alpha^2}{500k^{8.5}}$.*

Since we have $\sin\Theta(\mathbf{Y},\mathbf{U}_k) \leqslant \frac{1}{1000k^{10}}$ from Corollary 9.4.4, then it follows from Lemma 9.4.5 and the triangle inequality of $\sin\Theta$ distance that vectors in $\mathbf{Y}_k$ are close to the subspace $\mathbf{M}$:

Corollary 9.4.6. *Let $\mathbf{Y}$ be defined as in Algorithm 13 and let $\mathbf{R}$ be any k-dimensional subspace of $\mathbb{R}^d$ with*

$$\sin\Theta(\mathbf{Y},\mathbf{R}) \leqslant \frac{\alpha^2}{1000k^9}.$$

*Let $\mathbf{M}$ be the underlying latent k-simplex. Then for each unit vector $\mathbf{x} \in \mathbf{R}$, there exists a vector $\mathbf{y} \in$ **Span** $(\mathbf{M})$ with $\|\mathbf{x}-\mathbf{y}\|_2 \leqslant \frac{\alpha^2}{500k^{8.5}}$.*

We then use the following structural result between the first r points selected by Algorithm 13 and the closest r points in the latent k-simplex $\mathbf{M}$.

Lemma 9.4.7 (Equation 10.21 in [BK20c])**.** *For $r \in [k]$ let $\mathcal{R}_1,\ldots,\mathcal{R}_k \in \mathbb{R}^d$ be points such that there exist distinct $\ell_1,\ldots,\ell_r \subseteq [n]$ with*

$$\|\mathcal{R}_i - \mathbf{M}_{*,\ell_i}\|_2 \leqslant \frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}$$

Let $\widehat{\mathbf{A}} = \mathcal{R}_1 \circ \ldots \circ \mathcal{R}_t$ *and* $\widehat{\mathbf{M}} = \mathbf{M}_{*,\ell_1} \circ \ldots \circ \mathbf{M}_{*,\ell_r}$. *Then*

$$\|\widehat{\mathbf{M}} - \widehat{\mathbf{A}}\|_2 \leqslant \frac{k^{4.5}}{\alpha}\frac{\sigma}{\sqrt{\delta}}.$$

Proof. Note that the claim follows immediately from the hypothesis and applying the Cauchy-Schwarz inequality. □

We first bound the $\sin\Theta$ distance between $\mathbf{Span}\,(\mathbf{M}) \cap \text{Null}(\widehat{\mathbf{M}})$ and $\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)$. This essentially says that we can work in the subspace $\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)$ rather than $\mathbf{Span}\,(\mathbf{M}) \cap \text{Null}(\widehat{\mathbf{M}})$ and we will not incur too much error.

Next, we prove our lemma relating angular distance of the subspace obtained in the i-th iteration of the algorithm ($\mathbf{Y}(\mathbf{I} - \mathbf{P}_i)$) to the optimal subspace ($\mathbf{M}(\mathbf{I} - \mathbf{P}_i)$).

Lemma 9.4.8 (Angular Distance between Subspaces.)**.** *For some* $r \in [k]$, *let* $\widehat{\mathbf{M}} = \mathbf{M}_{*,\ell_1} \circ \ldots \circ \mathbf{M}_{*,\ell_r}$ *be the matrix with* r *columns corresponding to vertices of the latent* k*-simplex* $\mathbf{M}$ *closest to the first* r *points selected by Algorithm 13,* $\mathbf{A}_{\mathcal{R}_1}, \ldots, \mathbf{A}_{\mathcal{R}_r}$, *respectively. Suppose* $\|\mathbf{A}_{\mathcal{R}_i} - \mathbf{M}_{*,\ell_i}\|_2 \leqslant \frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}$ *for each* $i \in [r]$. *Let* $\mathbf{P}_r$ *be the projection matrix orthogonal to* $\mathbf{A}_{\mathcal{R}_1}, \ldots, \mathbf{A}_{\mathcal{R}_r}$. *Then,*

$$\sin\Theta\left(\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r), \textbf{\textit{Span}}\,(\mathbf{M}) \cap \textit{Null}(\widehat{\mathbf{M}})\right) \leqslant \frac{\alpha}{100k^4}$$
$$\sin\Theta\left(\textbf{\textit{Span}}\,(\mathbf{M}) \cap \textit{Null}(\widehat{\mathbf{M}}), \mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)\right) \leqslant \frac{\alpha}{100k^4}.$$

Proof. Let $\mathbf{y} \in \mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)$ be a unit vector. By Corollary 9.4.6, there exists $\mathbf{x} \in \mathbf{Span}\,(\mathbf{M})$ with

$$\|\mathbf{x} - \mathbf{y}\|_2 \leqslant \frac{\alpha^2}{500k^{8.5}}. \tag{9.2}$$

Let $\mathbf{z} = \mathbf{x} - \widehat{\mathbf{M}}\widehat{\mathbf{M}}^\dagger\mathbf{x}$ be the component of $\mathbf{x}$ in $\text{Null}(\widehat{\mathbf{M}})$. Note that $\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\dagger$ is a projection matrix and thus $\|\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\dagger\|_2 \leqslant 1$. Then we have

$$\begin{aligned}\|\mathbf{x} - \mathbf{z}\|_2 &\leqslant \|\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\dagger(\mathbf{x} - \mathbf{y})\|_2 + \|\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\dagger\mathbf{y}\|_2 \\ &\leqslant \|\mathbf{x} - \mathbf{y}\|_2 + \|\widehat{\mathbf{M}}(\widehat{\mathbf{M}}^T\widehat{\mathbf{M}})^{-1}(\widehat{\mathbf{M}}^T - \widehat{\mathbf{A}}^T)\mathbf{y}\|_2\end{aligned}$$

where $\widehat{\mathbf{A}} = \mathcal{R}_1 \circ \ldots \circ \mathcal{R}_t$ so that $\widehat{\mathbf{A}}^T\mathbf{y} = 0$ since $\mathbf{P}_r$ projects away from $\widehat{\mathbf{A}}$. We also have

$\|\widehat{\mathbf{M}}(\widehat{\mathbf{M}}^T\widehat{\mathbf{M}})^{-1}\|_2 = \frac{1}{\sigma_r(\widehat{\mathbf{M}})}$. Thus by (9.2) and Lemma 9.4.7, we have

$$\begin{aligned}\|\mathbf{x}-\mathbf{z}\|_2 &\leqslant \|\mathbf{x}-\mathbf{y}\|_2 + \frac{1}{\sigma_r(\widehat{\mathbf{M}})}\|(\widehat{\mathbf{M}}^T - \widehat{\mathbf{A}}^T)\mathbf{y}\|_2 \\ &\leqslant \frac{\alpha^2}{500k^{8.5}} + \frac{k^{4.5}\sigma}{\alpha\sqrt{\delta}\sigma_k(\widehat{\mathbf{M}})}.\end{aligned}$$

Hence by the triangle inequality and Lemma 9.4.3, we have $\|\mathbf{y}-\mathbf{z}\|_2 \leqslant \frac{\alpha}{100k^4}$. Since $\mathbf{y} \in \mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)$ and $\mathbf{z} \in \textbf{Span}\,(\mathbf{M}) \cap \text{Null}(\widehat{\mathbf{M}})$, then by definition of the $\sin\Theta$ distance, it follows that $\sin\Theta\left(\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r), \textbf{Span}\,(\mathbf{M}) \cap \text{Null}(\widehat{\mathbf{M}})\right) \leqslant \frac{\alpha}{100k^4}$, proving the first part of the claim.

To prove the second half of the claim, it suffices to show that the dimension of $\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)$ is $k-r$, since $\textbf{Span}\,(\mathbf{M}) \cap \text{Null}(\widehat{\mathbf{M}})$ has dimension $k-r$ and the $\sin\Theta$ distance is symmetric between two subspaces of the same dimension. By construction, $\mathbf{Y}$ has dimension k so that $\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)$ has dimension at least $k-r$. But if $\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)$ has dimension larger than $k-r$, then there exists a set of orthonormal vectors $\mathbf{u}_1, \ldots, \mathbf{u}_{k-r+1} \in \mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)$. By the first part of the claim and the definition of the $\sin\Theta$ distance, there exists a set of corresponding vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{k-r+1} \in \textbf{Span}\,(\mathbf{M}) \cap \text{Null}(\widehat{\mathbf{M}})$ such that $\|\mathbf{u}_i - \mathbf{v}_j\|_2 < \frac{\alpha}{100k^4}$. But then for $a \neq b$, we have by the triangle inequality and the fact that $\mathbf{u}_a \cdot \mathbf{u}_b = 0$,

$$\begin{aligned}|\mathbf{v}_a \cdot \mathbf{v}_b| &\leqslant |\mathbf{u}_a \cdot \mathbf{u}_b| + |(\mathbf{v}_a - \mathbf{u}_a)\cdot \mathbf{u}_b| + |\mathbf{v}_a \cdot (\mathbf{v}_b - \mathbf{u}_b)| \\ &\leqslant \frac{\alpha}{50k^4}\end{aligned}$$

Similarly, since $\mathbf{u}_a \cdot \mathbf{u}_a = 1$, we have

$$\begin{aligned}|\mathbf{v}_a \cdot \mathbf{v}_a| &\geqslant |\mathbf{u}_a \cdot \mathbf{u}_a| - |(\mathbf{v}_a - \mathbf{u}_a)\cdot \mathbf{u}_a| - |\mathbf{v}_a \cdot (\mathbf{v}_a - \mathbf{u}_a)| \\ &\geqslant 1 - \frac{\alpha}{50k^4}.\end{aligned}$$

Thus if $\mathbf{V} = \mathbf{v}_1 \circ \ldots \circ \mathbf{v}_{k-r+1} \in \mathbb{R}^{d \times k-r+1}$ is formed by concatenating the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{k-r+1}$, then $\mathbf{V}^T\mathbf{V}$ is diagonally-dominant. Hence, $\mathbf{V}^T\mathbf{V}$ is nonsingular, so $\mathbf{v}_1, \ldots, \mathbf{v}_{k-r+1}$ must be linearly independent vectors in $\textbf{Span}\,(\mathbf{M}) \cap \text{Null}(\widehat{\mathbf{M}})$, which contradicts the fact that its dimension is $k-r$. Therefore, the dimension of $\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)$ must be $k-r$, and so $\sin\Theta\left(\textbf{Span}\,(\mathbf{M}) \cap \text{Null}(\widehat{\mathbf{M}}), \mathbf{Y}(\mathbf{I}_d - \mathbf{P}_r)\right) \leqslant \frac{\alpha}{100k^4}$. □

We now recall a structural lemma from [BK20c].

Lemma 9.4.9 (Claim 10.1 in [BK20c])**.** *Let $a, b \notin \{\ell_1, \ldots, \ell_r\}$ be distinct indices. Then*

$$\|\mathit{Proj}(\mathbf{M}_{*,a} - \mathbf{M}_{*,b}, \mathit{Null}(\widehat{\mathbf{M}})\|_2 \geqslant \alpha \max_{\ell} \|\mathbf{M}_{*,\ell}\|_2.$$

Now we need to show that our algorithm is (1) well-defined and (2) preserves the invariant that the $(i+1)$-st point sampled from $\mathbf{Y}^T(\mathbf{I} - \mathbf{P}_i)$ will also be reasonably close to some different point of the k-simplex. We show the selected procedure is well-defined in Lemma 9.4.10 by arguing that there exists a unique solution to the maximization problem.

Lemma 9.4.10 (Optimization is Well-Defined)**.** *Let $\mathbf{u} \in \mathbb{R}^d$ be a random unit vector in the space of $\mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r)$, where $\mathbf{P}_r$ is the orthogonal projection to $\mathbf{A}_{\mathcal{R}_1}, \ldots, \mathbf{A}_{\mathcal{R}_r}$. Then there exists a constant $c > 0$ so that with probability at least $1 - c/k^{1.5}$:*

1. *For all distinct $a, b \notin \{\ell_1, \ldots, \ell_r\}$, then $|\mathbf{u} \cdot (\mathbf{M}_{*,a} - \mathbf{M}_{*,b})| \geqslant \frac{0.097}{k^4} \alpha \max_\ell \|\mathbf{M}_{*,\ell}\|_2$.*
2. *For all $a \notin \{\ell_1, \ldots, \ell_r\}$, then $|\mathbf{u} \cdot \mathbf{M}_{*,a}| \geqslant \frac{0.0989}{k^4} \alpha \max_\ell \|\mathbf{M}_{*,\ell}\|_2$.*

Proof. For $a \notin \{\ell_1, \ldots, \ell_r\}$, let $\mathbf{p}_a$ be the projection of $\mathbf{M}_{*,a}$ onto $\mathrm{Null}(\widehat{\mathbf{M}})$ and $\mathbf{q}_a$ be the projection of $\mathbf{M}_{*,a}$ onto $\textbf{Span}\left(\widehat{\mathbf{M}}\right)$. By the Well-Separateness assumption, we have $\|\mathbf{p}_a\|_2 \geqslant \alpha \max_\ell \|\mathbf{M}_{*,\ell}\|_2$. Let $\mathbf{w}_a$ be defined so that $\mathbf{q}_a = \widehat{\mathbf{M}}\mathbf{w}_a$. Since $\|\mathbf{q}_a\|_2 \leqslant \|\mathbf{M}_{*,a}\|_2$ and $\sigma_r(\widehat{\mathbf{M}}) \leqslant \sigma_k(\mathbf{M})$, then Lemma 9.4.3 gives

$$\|\mathbf{w}_a\|_2 \leqslant \frac{\|\mathbf{q}_a\|_2}{\sigma_r(\widehat{\mathbf{M}})} \leqslant \frac{\|\mathbf{M}_{*,a}\|_2 \alpha^2}{1000k^{8.5}} \frac{\sqrt{\delta}}{\sigma}. \tag{9.3}$$

Since $\widehat{\mathbf{A}}\mathbf{u} = 0$, we can also write

$$\begin{aligned}\mathbf{u} \cdot \mathbf{M}_{*,a} &= \mathbf{u} \cdot \mathbf{p}_a + \mathbf{u} \cdot \mathbf{q}_a \\ &= \mathbf{u} \cdot \mathrm{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r)) + \mathbf{u}^T(\widehat{\mathbf{M}} - \widehat{\mathbf{A}})\mathbf{w}_a.\end{aligned}$$

By Lemma 9.4.7, (9.3), and normalizing so that $\|\mathbf{u}\|_2 = 1$, we have

$$\begin{aligned}|\mathbf{u} \cdot (\mathbf{M}_{*,a} - \cdot\mathrm{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r)))| &\leqslant \|\widehat{\mathbf{M}} - \widehat{\mathbf{A}}\|_2 \|\mathbf{w}_a\|_2 \\ &\leqslant \frac{\alpha \|\mathbf{M}_{*,a}\|_2}{1000k^4}.\end{aligned} \tag{9.4}$$

The same holds for $\mathbf{u} \cdot (\mathbf{M}_{*,a} - \mathbf{M}_{*,b})$, so that

$$\begin{aligned} &|\mathbf{u} \cdot (\mathbf{M}_{*,a} - \mathbf{M}_{*,b}) - \mathbf{u} \cdot \text{Proj}(\mathbf{p}_a - \mathbf{p}_b, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))| \\ &\leqslant \frac{\|\mathbf{M}_{*,a} - \mathbf{M}_{*,b}\|_2 \alpha}{1000k^4}. \end{aligned} \tag{9.5}$$

Let $\mathcal{E}$ be the event that:

1. For all a, $|\mathbf{u} \cdot \text{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))| \geqslant \frac{1}{10k^4}\|\text{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))\|_2$.
2. For all $a \neq b$, $|\mathbf{u} \cdot \text{Proj}(\mathbf{p}_a - \mathbf{p}_b, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))| \geqslant \frac{1}{10k^4}\|\text{Proj}(\mathbf{p}_a - \mathbf{p}_b, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))\|_2$.

Note that $|\mathbf{u} \cdot \text{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))| \geqslant \frac{1}{10k^4}\|\text{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))\|_2$ holds as long as $\mathbf{u} \cdot \text{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r)) \neq 0$. Since the volume of the set $\{\mathbf{x} \in \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r) : \mathbf{u} \cdot \mathbf{x} = 0\}$ is at most $\sqrt{k}$ times the volume of the unit ball $\{\mathbf{x} \in \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r) : \|\mathbf{x}\|_2 = 1\}$, then by taking a union bound over at most k^2 indices, it follows that $\mathcal{E}$ holds with probability at least $1 - \frac{1}{k^{1.5}}$.

By Lemma 9.4.8, there exists $\mathbf{p}'_a \in \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r)$ such that $\|\mathbf{p}'_a - \mathbf{p}_a\|_2 \leqslant \frac{\alpha\|\mathbf{p}_a\|_2}{100k^4}$. Hence for $k \geqslant 2$, $\|\mathbf{p}_a - \text{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))\|_2 \leqslant \frac{\alpha\|\mathbf{p}_a\|_2}{100k^4} \leqslant \frac{\|\mathbf{p}_a\|_2}{1600}$. This implies $\|\text{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))\|_2 \geqslant 0.999\|\mathbf{p}_a\|_2$. Then conditioning on $\mathcal{E}$,

$$\begin{aligned} |\mathbf{u} \cdot \text{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))| &\geqslant \frac{\|\text{Proj}(\mathbf{p}_a, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r))\|_2}{10k^4} \\ &\geqslant \frac{0.999\|\mathbf{p}_a\|_2}{10k^4} \\ &\geqslant \frac{0.999 \max_\ell \|\mathbf{M}_{*,\ell}\|_2}{10k^4}, \end{aligned}$$

where the last inequality follows since $\|\mathbf{p}_a\|_2 \geqslant \|\text{Proj}(\mathbf{M}_{*,a}, \text{Null}(\mathbf{M} \setminus \mathbf{M}_{*,a}))\|_2 \geqslant \alpha \max_\ell \mathbf{M}_{*,\ell}$ by the Well-Separateness assumption. Hence by (9.4), it follows that for all $a \notin \{\ell_1, \ldots, \ell_r\}$,

$$\begin{aligned} |\mathbf{u} \cdot \mathbf{M}_{*,a}| &\geqslant |\mathbf{u} * \text{Proj}(\mathbf{u}, \mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r)) - \frac{\alpha\|\mathbf{M}_{*,a}\|_2}{1000k^4} \\ &\geqslant \frac{0.0989\alpha \max_\ell \|\mathbf{M}_{*,\ell}\|_2}{k^4}, \end{aligned}$$

which proves the second half of the claim.

To prove the first half of the claim, note that conditioned on $\mathcal{E}$, then (9.5) implies

$$\begin{aligned}|\mathbf{u}\cdot(\mathbf{M}_{*,a}-\mathbf{M}_{*,b})| &\geqslant |\mathbf{u}\cdot\text{Proj}(\mathbf{p}_a-\mathbf{p}_b,\mathbf{Y}^T(\mathbf{I}_d-\mathbf{P}_r))|\\ &\quad-\frac{\|\mathbf{M}_{*,a}-\mathbf{M}_{*,b}\|_2\alpha}{1000k^4}\\ &\geqslant \frac{\|\text{Proj}(\mathbf{p}_a-\mathbf{p}_b,\mathbf{Y}^T(\mathbf{I}_d-\mathbf{P}_r))\|_2}{10k^4}\\ &\quad-\frac{\|\mathbf{M}_{*,a}-\mathbf{M}_{*,b}\|_2\alpha}{1000k^4}.\end{aligned}$$

By Lemma 9.4.8, there exists $\mathbf{v}\in\mathbf{Y}^T(\mathbf{I}_d-\mathbf{P}_r)$ such that $\|\mathbf{v}-(\mathbf{p}_a-\mathbf{p}_b)\|_2\leqslant\frac{\alpha\|\mathbf{p}_a-\mathbf{p}_b\|_2}{100k^4}$. Thus, $\|\text{Proj}(\mathbf{p}_a-\mathbf{p}_b,\mathbf{Y}^T(\mathbf{I}_d-\mathbf{P}_r))\|_2\geqslant 0.99\|\mathbf{p}_a\|_2\geqslant 0.99\alpha\max_\ell\|\mathbf{M}_{*,\ell}\|_2$, by Lemma 9.4.9. Since $\frac{\|\mathbf{M}_{*,a}-\mathbf{M}_{*,b}\|_2\alpha}{1000k^4}\leqslant\frac{2\alpha\max_\ell\|\mathbf{M}_{*,\ell}\|_2}{1000k^4}$, it follows that $|\mathbf{u}\cdot(\mathbf{M}_{*,a}-\mathbf{M}_{*,b})|\geqslant\frac{0.097}{k^4}\alpha\max_\ell\|\mathbf{M}_{*,\ell}\|_2$. □

We next show that the selected index is not among the previously selected indices. Thus, we obtain a new index at each iteration, which implies that we only need k iterations.

Lemma 9.4.11. *Let $\widehat{\mathbf{M}}=\mathbf{M}_{*,\ell_1}\circ\ldots\circ\mathbf{M}_{*,\ell_r}$ be the r points in the latent k-simplex $\mathbf{M}$ closest to the first r points selected by Algorithm 13, $\mathcal{R}_1,\ldots,\mathcal{R}_r$, respectively. Suppose*

$$\|\mathcal{R}_i-\mathbf{M}_{*,\ell_i}\|_2\leqslant\frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}$$

for each $i\in[r]$. Let $\mathbf{u}\in\mathbb{R}^d$ be a random unit vector in the space of $\mathbf{Y}^T(\mathbf{I}_d-\mathbf{P}_r)$, where $\mathbf{P}_r$ is the orthogonal projection to $\mathcal{R}_1,\ldots,\mathcal{R}_r$. Let

$$\ell_{r+1}=\begin{cases}\arg\max_\ell\mathbf{u}\cdot\mathbf{M}_{*,\ell} & \text{if } \mathbf{u}\cdot\mathcal{R}_{r+1}\geqslant 0\\ \arg\min_\ell\mathbf{u}\cdot\mathbf{M}_{*,\ell} & \text{if } \mathbf{u}\cdot\mathcal{R}_{r+1}<0\end{cases}.$$

Then $\ell_{r+1}\notin\{\ell_1,\ldots,\ell_r\}$.

Proof. We consider the case $\mathbf{u}\cdot\mathcal{R}_{r+1}\geqslant 0$ as the analysis for the case $\mathbf{u}\cdot\mathcal{R}_{r+1}<0$ is symmetric. Let $\ell_{r+1}=\arg\max_\ell\mathbf{u}\cdot\mathbf{M}_{*,\ell}$. Suppose by way of contradiction that $\ell_{r+1}\in\{\ell_1,\ldots,\ell_r\}$. Without loss of generality, let $\ell_{r+1}=\ell_1$. Since $\|\mathcal{R}_1-\mathbf{M}_{*,\ell_1}\|_2\leqslant\frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}$ and $\mathbf{u}\cdot\mathcal{R}_1$, then

$$\mathbf{u}\cdot\mathbf{M}_{*,\ell_i}\|_2\leqslant\mathbf{u}\cdot\mathcal{R}_i+\frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}=\frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}.$$

Since $\ell_1=\arg\max_\ell\mathbf{u}\cdot\mathbf{M}_{*,\ell}$, then $\mathbf{u}\cdot\mathbf{M}_{*,\ell}\leqslant\mathbf{u}\cdot\mathbf{M}_{*,\ell_1}$ for all ℓ. Thus $\mathbf{u}\cdot\mathbf{P}_{*,S}\leqslant\frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}$ for

any set of indices $S \subseteq [n]$ inside the convex hull of $\mathbf{M}$. In conjunction, Lemma 9.4.12 implies

$$\mathbf{u} \cdot \mathbf{A}_{*,\mathcal{R}_{r+1}} \leqslant \mathbf{u} \cdot \mathbf{P}_{*,\mathcal{R}_{r+1}} + \frac{\sigma}{\sqrt{\delta}} \leqslant \left(\frac{300k^4}{\alpha} + 1\right) \frac{\sigma}{\sqrt{\delta}}. \tag{9.6}$$

Recall that by Lemma 9.4.1, $\|\mathbf{A} - \mathbf{Y}\mathbf{Z}^T\|_2^2 \leqslant (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_2^2 + \frac{\epsilon}{k}\|\mathbf{A} - \mathbf{A}_k\|_F^2$ and thus $\|\mathbf{A} - \mathbf{Y}\mathbf{Z}^T\| \leqslant (1+2\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_2$, given the Significant Singular Values assumption. Since $\mathbf{A}_{*,\mathcal{R}_{r+1}}$ is a subset of δn columns of $\mathbf{A}$ and $\mathcal{R}_{r+1}$ is a subset of δn columns of $\mathbf{Y}$, then for $\epsilon < 1$,

$$\begin{aligned}
\mathbf{u} \cdot \mathcal{R}_{r+1} &\leqslant \mathbf{u} \cdot \mathbf{A}_{*,\mathcal{R}_{r+1}} + \mathbf{u} \cdot (\mathcal{R}_{r+1} - \mathbf{A}_{*,\mathcal{R}_{r+1}}) \\
&\leqslant \left(\frac{300k^4}{\alpha} + 1\right) \frac{\sigma}{\sqrt{\delta}} + \frac{3}{\sqrt{\delta n}}\|\mathbf{A} - \mathbf{A}_k\|_2,
\end{aligned}$$

where the last step follows from (9.6) and applying the Cauchy-Schwarz inequality and the fact that $\mathbf{u}$ is a unit vector. Since $\mathbf{P}$ has rank k and $\mathbf{A}_k$ is the best rank k approximation to $\mathbf{A}$, then $\|\mathbf{A} - \mathbf{A}_k\|_2 \leqslant \|\mathbf{A} - \mathbf{P}\|_2$ so that

$$\begin{aligned}
\mathbf{u} \cdot \mathcal{R}_{r+1} &\leqslant \left(\frac{300k^4}{\alpha} + 1\right) \frac{\sigma}{\sqrt{\delta}} + \frac{3}{\sqrt{\delta n}}\|\mathbf{A} - \mathbf{P}\|_2, \\
&\leqslant \left(\frac{300k^4}{\alpha} + 1\right) \frac{\sigma}{\sqrt{\delta}} + \frac{3\sigma}{\sqrt{\delta}} \qquad (9.7) \\
&= \left(\frac{300k^4}{\alpha} + 4\right) \frac{\sigma}{\sqrt{\delta}}, \qquad (9.8)
\end{aligned}$$

since $\|\mathbf{A} - \mathbf{P}\|_2 \leqslant \sigma\sqrt{n}$ by definition of σ. However for $t \notin \{\ell_1, \ldots, \ell_r\}$, Lemma 9.4.12 and the Proximate Latent Points assumption imply the existence of a set σ_t of δn columns such that

$$\begin{aligned}
|\mathbf{u} \cdot \mathbf{A}_{*,\sigma_t}| &\geqslant |\mathbf{u} \cdot \mathbf{P}_{*,\sigma_t}| - \frac{\sigma}{\sqrt{\delta}} \\
&\geqslant |\mathbf{u} \cdot \mathbf{M}_{*,t}| - \frac{5\sigma}{\sqrt{\delta}} \qquad (9.9) \\
&\geqslant \frac{0.0989}{k^4}\alpha \max_{\ell} \|\mathbf{M}_{*,\ell}\|_2 - \frac{5\sigma}{\sqrt{\delta}},
\end{aligned}$$

where the last step follows from Lemma 9.4.10. Moreover, σ_t has δn columns, so again by applying the Cauchy-Schwarz inequality and the fact that $\mathbf{u}$ is a unit vector, we have

$$\begin{aligned}
|\mathbf{u} \cdot (\mathbf{A}_{*,\sigma_t} - \mathbf{Y}_{*,\sigma_t})| &\leqslant \frac{1}{\sqrt{\delta n}}\|\mathbf{A} - \mathbf{A}_k\|_2 \\
&\leqslant \frac{1}{\sqrt{\delta n}}\|\mathbf{A} - \mathbf{P}\|_2 \leqslant \frac{3\sigma}{\sqrt{\delta}}.
\end{aligned} \tag{9.10}$$

where the last two inequalities come from the fact that $\mathbf{P}$ has rank k and $\|\mathbf{A}-\mathbf{P}\|_2 \leqslant \sigma\sqrt{n}$ by definition of σ.

Thus from (9.9) and (9.10),

$$\begin{aligned}|\mathbf{u}\cdot\mathbf{Y}_{*,\sigma_t}| &\geqslant |\mathbf{u}\cdot\mathbf{A}_{*,\sigma_t}| - |\mathbf{u}\cdot(\mathbf{A}_{*,\sigma_t}-\mathbf{Y}_{*,\sigma_t})| \\ &\geqslant \frac{0.0989}{k^4}\alpha\max_{\ell}\|\mathbf{M}_{*,\ell}\|_2 - \frac{8\sigma}{\sqrt{\delta}}.\end{aligned}$$

However by the Spectrally Bounded Perturbation assumption, we have $|\mathbf{u}\cdot\mathbf{Y}_{*,\sigma_t}| \geqslant \frac{2400k^5}{\alpha}\frac{\sigma}{\sqrt{\delta}} - \frac{8\sigma}{\sqrt{\delta}}$, which contradicts the maximality of $\mathcal{R}_{r+1}$ in (9.8). Therefore, it holds that $\ell_{r+1} \notin \{\ell_1,\ldots,\ell_r\}$. □

Before showing that the selected index completes the inductive step, we recall the following:

Lemma 9.4.12 (Lemma 3.1 in [BK20c])**.** *For a subset $S \subseteq [n]$, let $\mathbf{A}_{*,S} = \frac{1}{|S|}\sum_{i\in S}\mathbf{A}_{*,i}$. For all $S \subseteq [n]$, $|\mathbf{A}_{*,S} - \mathbf{P}_{*,S}| \leqslant \sigma\sqrt{n/|S|}$.*

We then show that the algorithm preserves the aforementioned invariant by showing that the unique solution $\mathbf{A}_{\mathcal{R}_i}$ cannot correspond to one of the vertices of the k-simplex that have been found in the first i rounds, thus proving that we find a solution $\mathbf{A}_{\mathcal{R}_i}$ that corresponds to a new vertex of $\mathbf{M}$. We then show $\mathbf{A}_{\mathcal{R}_i}$ is close to the new vertex of $\mathbf{M}$, preserving the inductive hypothesis.

Lemma 9.4.13 (Recovery Guarantees)**.** *Let $\widehat{\mathbf{M}} = \mathbf{M}_{*,\ell_1}\circ\ldots\circ\mathbf{M}_{*,\ell_r}$ be the r points in the latent k-simplex $\mathbf{M}$ closest to the first r points selected by Algorithm 13, $\mathcal{R}_1,\ldots,\mathcal{R}_r$, respectively. Suppose*

$$\|\mathcal{R}_i - \mathbf{M}_{*,\ell_i}\|_2 \leqslant \frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}$$

for each $i \in [r]$. Let $\mathbf{u} \in \mathbb{R}^d$ be a random unit vector in the space of $\mathbf{Y}^T(\mathbf{I}_d - \mathbf{P}_r)$, where $\mathbf{P}_r$ is the orthogonal projection to $\mathcal{R}_1,\ldots,\mathcal{R}_r$. Let

$$\ell_{r+1} = \begin{cases}\arg\max_\ell \mathbf{u}\cdot\mathbf{M}_{*,\ell} & \text{if } \mathbf{u}\cdot\mathcal{R}_{r+1} \geqslant 0 \\ \arg\min_\ell \mathbf{u}\cdot\mathbf{M}_{*,\ell} & \text{if } \mathbf{u}\cdot\mathcal{R}_{r+1} < 0\end{cases}.$$

Then

$$\|\mathcal{R}_{r+1} - \mathbf{M}_{*,\ell_{r+1}}\|_2 \leqslant \frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}.$$

Proof. We consider the case $\mathbf{u} \cdot \mathcal{R}_{r+1} \geqslant 0$ as the analysis for the case $\mathbf{u} \cdot \mathcal{R}_{r+1} < 0$ is symmetric. Let $\ell_{r+1} = \arg\max_\ell \mathbf{u} \cdot \mathbf{M}_{*,\ell}$. By Lemma 9.4.11, we have $\ell_{r+1} \notin \{\ell_1, \ldots, \ell_r\}$. Thus applying Lemma 9.4.10,

$$\mathbf{u} \cdot \mathbf{M}_{*,\ell_{r+1}} \geqslant \frac{0.0989}{k^4} \alpha \max_\ell \|\mathbf{M}_{*,\ell}\|. \tag{9.11}$$

By the Proximate Latent Points assumption, there exists a set $\sigma_{\ell_{r+1}}$ of size δn so that $\|\mathbf{P}_{*,j} - \mathbf{M}_{*,\ell_{r+1}}\|_2 \leqslant \frac{4\sigma}{\sqrt{\delta}}$ for all $j \in \sigma_{\ell_{r+1}}$ so that $\|\mathbf{P}_{*,\sigma_{\ell_{r+1}}} - \mathbf{M}_{*,\ell_{r+1}}\|_2 \leqslant \frac{4\sigma}{\sqrt{\delta}}$. Then by Lemma 9.4.12,

$$\mathbf{u} \cdot \mathbf{A}_{*,\sigma_{\ell_{r+1}}} \geqslant \mathbf{u} \cdot \mathbf{P}_{*,\sigma_{\ell_{r+1}}} - \frac{\sigma}{\sqrt{\delta}} \geqslant \mathbf{u} \cdot \mathbf{M}_{*,\ell_{r+1}} - \frac{5\sigma}{\sqrt{\delta}}.$$

By the same reasoning as 9.10, we have $\|\mathcal{R}_{r+1} - \mathbf{A}_{*,\sigma_{\ell_{r+1}}}\|_2 \leqslant \frac{3\sigma}{\sqrt{\delta}}$ and thus,

$$\mathbf{u} \cdot \mathcal{R}_{r+1} \geqslant \mathbf{u} \cdot \mathbf{M}_{*,\ell_{r+1}} - \frac{8\sigma}{\sqrt{\delta}}. \tag{9.12}$$

Now for any $a \notin \{\ell_1, \ldots, \ell_{r+1}\}$, Lemma 9.4.10 says

$$\mathbf{u} \cdot \mathbf{M}_{*,a} \leqslant \mathbf{u} \cdot \mathbf{M}_{*,\ell_{r+1}} - \frac{0.097}{k^4} \alpha \max_\ell \|\mathbf{M}_{*,\ell}\|_2. \tag{9.13}$$

Similarly, for $a \in \{\ell_1, \ldots, \ell_r\}$, we have $\|\mathcal{R}_a - \mathbf{M}_{*,a}\| \leqslant \frac{300k^4}{\alpha} \frac{\sigma}{\sqrt{\delta}}$ by the inductive hypothesis. Since $\mathbf{u} \cdot \mathcal{R}_a = 0$, then

$$\begin{aligned}
\mathbf{u} \cdot \mathbf{M}_{*,a} &\leqslant \mathbf{u} \cdot \mathcal{R}_a + \frac{300k^4}{\alpha} \frac{\sigma}{\sqrt{\delta}} = \frac{300k^4}{\alpha} \frac{\sigma}{\sqrt{\delta}} \\
&\leqslant \mathbf{u} \cdot \mathbf{M}_{*,\ell_{r+1}} - \frac{0.0989}{k^4} \alpha \max_\ell \|\mathbf{M}_{*,\ell}\| \\
&\quad + \frac{300k^4}{\alpha} \frac{\sigma}{\sqrt{\delta}}
\end{aligned}$$

by (9.11). Thus by the Spectrally Bounded Perturbation assumption,

$$\mathbf{u} \cdot \mathbf{M}_{*,a} \leqslant \mathbf{u} \cdot \mathbf{M}_{*,\ell_{r+1}} - \frac{0.097}{k^4} \alpha \max_\ell \|\mathbf{M}_{*,\ell}\| \tag{9.14}$$

Since $\mathbf{P}_{*,\mathcal{R}_{r+1}}$ is a convex combination of the columns of $\mathbf{M}$, there exists a vector $\mathbf{w}$ such that

$\mathbf{P}_{*,\mathcal{R}_{r+1}} = \mathbf{M}\mathbf{w}$. Then by the same reasoning as 9.10 and Lemma 9.4.12,

$$\begin{aligned}\mathbf{u}\cdot\mathcal{R}_{r+1} &\leqslant \mathbf{u}\cdot\mathbf{A}_{*,\mathcal{R}_{r+1}} + \frac{3\sigma}{\sqrt{\delta}} \leqslant \mathbf{u}\cdot\mathbf{P}_{*,\mathcal{R}_{r+1}} + \frac{3\sigma}{\sqrt{\delta}} + \frac{4\sigma}{\sqrt{\delta}} \\ &\leqslant w_{\ell_{r+1}}(\mathbf{u}\cdot\mathbf{M}_{*,\ell_{r+1}}) + \\ &\sum_{a\neq\ell_{r+1}} w_a\left((\mathbf{u}\cdot\mathbf{M}_{*,\ell_{r+1}} - \frac{0.097}{k^4}\alpha\max_{\ell}\|\mathbf{M}_{*,\ell}\|_2\right) \\ &+ \frac{4\sigma}{\sqrt{\delta}},\end{aligned}$$

where the last line follows from decomposing $\mathbf{M}$ and applying (9.13) and (9.14) to $\mathbf{M}_{*,a}$ for $a \neq \ell_{r+1}$. Hence,

$$\begin{aligned}\mathbf{u}\cdot\mathcal{R}_{r+1} &\leqslant \mathbf{u}\cdot\mathbf{M}_{*,\ell_{r+1}} - \frac{0.097\alpha\max_{\ell}\|\mathbf{M}_{*,\ell}\|_2(1-w_{\ell_{r+1}})}{k^4} \\ &+ \frac{4\sigma}{\sqrt{\delta}}.\end{aligned}$$

Combining with (9.12), we have

$$(1-w_{\ell_{r+1}})\max_{\ell}\|\mathbf{M}_{*,\ell}\|_2 \leqslant \frac{12\sigma}{\sqrt{\delta}}\frac{k^4}{0.097\alpha} \leqslant \frac{124k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}.$$

Thus,

$$\begin{aligned}\|\mathbf{P}_{*,\mathcal{R}_{r+1}} - \mathbf{M}_{*,\ell_{r+1}}\|_2 &= \|(w_{\ell_{r+1}}-1)\mathbf{M}_{*,\ell_{r+1}} \\ &\quad + \sum_{a\neq\ell_{r+1}} w_a\mathbf{M}_{*,a}\| \\ &\leqslant \sum_{a\neq\ell_{r+1}} w_a\|\mathbf{M}_{*,\ell_{r+1}} - \mathbf{M}_{*,a}\|_2 \\ &\leqslant 2(1-w_{\ell_{r+1}})\max_{\ell}\|\mathbf{M}_{*,\ell}\|_2 \\ &\leqslant \frac{248k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}.\end{aligned}$$

Finally from the triangle inequality and Lemma 9.4.12, we have

$$\begin{aligned}\|\mathcal{R}_{r+1} - \mathbf{M}_{*,\ell_{r+1}}\|_2 &\leqslant \|\mathcal{R}_{r+1} - \mathbf{P}_{*,\mathcal{R}_{r+1}}\|_2 \\ &\quad + \|\mathbf{P}_{*,\mathcal{R}_{r+1}} - \mathbf{M}_{*,\ell_{r+1}}\|_2 \\ &\leqslant \frac{3\sigma}{\sqrt{\delta}} + \frac{248k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}} \\ &\leqslant \frac{300k^4}{\alpha}\frac{\sigma}{\sqrt{\delta}}.\end{aligned}$$

□

9.5 Connection to Spectral Low-Rank Approximation

In this section, we show that learning a latent simplex is closely related to computing a spectral low-rank approximation. Spectral low-rank approximation is a fundamental primitive for algorithm design and numerical linear algebra and the best known algorithm for computing a $(1+\epsilon)$-approximation is $O(\mathsf{nnz}(\mathbf{A}) \cdot k)$ [MM15]. A major open question in randomized linear algebra is to determine whether the dependence on k in the running time is necessary for spectral low-rank approximation.

We show that for a candidate hard distribution over the input, determined by a Stochastic Block Model (with appropriate parameters) satisfying Well-Separateness1, Proximate Latent Points2 and Spectrally Bounded Perturbations3, an algorithm for learning a latent simplex requiring $o(\mathsf{nnz}(\mathbf{A}) \cdot k)$ time also recovers a spectral low-rank approximation for the input. One way to interpret this statement is that improving the running time for learning a latent simplex under the same assumptions as [BK20c] would likely lead to a major algorithmic breakthrough for spectral low-rank approximation.

Theorem 188 (Spectral LRA to Latent Simplex)**.** *Given $k \in [n]$, let $\mathcal{S}_1, \mathcal{S}_2 \ldots, \mathcal{S}_k$ be a partition of $[n]$ such that for all $\ell \in [k]$, $|\mathcal{S}_\ell| = n/k$. Consider a stochastic block model with k communities, $\mathcal{S}_1, \ldots, \mathcal{S}_k$ such that for all $i \in \mathcal{S}_\ell$ and $j \in \mathcal{S}_{\ell'}$, the probability of an edge (i,j) is $p = poly(k)/n^{1/8}$ when $\ell = \ell'$ and $q = p/10$ otherwise. Let $\mathbf{A}$ be a matrix drawn from the aforementioned model such that $\mathbf{A}_{i,j} = 1$ if there exists an edge between (i,j) and 0 otherwise. Then any algorithm that learns the simplex also recovers a rank k matrix $\mathbf{B}$ such that $\|\mathbf{A} - \mathbf{B}\|_2^2 \leqslant \|\mathbf{A} - \mathbf{A}_k\|_2^2 + \frac{1}{n^{1/3}}\|\mathbf{A} - \mathbf{A}_k\|_F^2$.*

Proof. Let $\mathbf{P_B}$ be the projection matrix onto the column span of the output matrix $\mathbf{B}$. We show that $\mathbf{A} - \mathbf{P_B}$ is a good mixed spectral-Frobenius low-rank approximation to $\mathbf{A}$.

$$\begin{aligned}\|\mathbf{A} - \mathbf{P_B}\mathbf{A}\|_2 &\leqslant \|\mathbf{A} - \mathbf{P} + \mathbf{P}\|_2 \|\mathbf{I} - \mathbf{P_B}\|_2 \\ &\leqslant \|\mathbf{A} - \mathbf{P}\|_2 \|\mathbf{I} - \mathbf{P_B}\|_2 + \mathbf{P}\|_2 \|\mathbf{I} - \mathbf{P_B}\|_2 \\ &\leqslant \|\mathbf{A} - \mathbf{P}\|_2 + \|\mathbf{P}\|_2 \|\mathbf{I} - \mathbf{P_B}\|_2.\end{aligned}$$

From the definition of σ, we have $\|\mathbf{A} - \mathbf{P}\|_2 \leqslant \sigma\sqrt{n}$. For the specific stochastic block model, we have $\sigma \leqslant \sqrt{p(1-p)}$, e.g., see [Awa17]. Moreover, the algorithm of [BK20c] guarantees specifically in their Theorem 7.2 that $\|\mathbf{I} - \mathbf{P_B}\|_2 \leqslant \frac{C_1 k^{4.5} d^{1/8}}{n^{1/4}}$ for some constant $C_1 > 0$. Since $\|\mathbf{P}\|_F \geqslant \|\mathbf{P}\|_2$ and $\|\mathbf{P}\|_F^2 \leqslant C_2 p^2 nd$ for some constant $C_2 > 0$ with high probability, then we have

$$\begin{aligned}\|\mathbf{A} - \mathbf{P_B}\mathbf{A}\|_2 &\leqslant \sqrt{p(1-p)n} + \frac{C_1 k^{4.5} d^{1/8} \sqrt{C_2 p^2 nd}}{n^{1/4}} \\ &\leqslant \sqrt{pn} + C_1 p k^{4.5} d^{5/8} \sqrt{C_2} n^{1/4}.\end{aligned}$$

On the other hand, we have $\|\mathbf{A} - \mathbf{A}_k\|_F^2 \geqslant \|\mathbf{A}\|_F^2 - k\|\mathbf{A}\|_2^2$. As before, we have $\|\mathbf{P}\|_2 \leqslant p\sqrt{C_2 nd}$, so that

$$\|\mathbf{A}\|_2 \leqslant \|\mathbf{P}\|_2 + \|\mathbf{A} - \mathbf{P}\|_2 \leqslant p\sqrt{C_2 nd} + \sqrt{p(1-p)n}.$$

Moreover, we have $\|\mathbf{A}\|_F \geqslant C_3\sqrt{qnd}$ for some constant $C_3 > 0$ with high probability. Hence for $q > C_4 p^2$ with a sufficiently high constant C_4, we have

$$\|\mathbf{A} - \mathbf{A}_k\|_F^2 \geqslant C_5 qnd,$$

for some $C_5 > 0$. Let $p = O(q)$ and $d = n^{1/C}$ for some constant $C \geqslant 3$ so that $k^{4.5} d^{5/8} = o(n^{1/4})$. Since $\|\mathbf{A} - \mathbf{P_B}\mathbf{A}\|_2^2 \leqslant C_6 pn$ for some constant C_6, then

$$\begin{aligned}\|\mathbf{A} - \mathbf{P_B}\mathbf{A}\|_2^2 \leqslant C_6 pn \leqslant \frac{C_5}{n^{1/C}} qnd = O\left(\frac{1}{n^{1/C}}\right) \|\mathbf{A} - \mathbf{A}_k\|_F^2 \\ \leqslant \|\mathbf{A} - \mathbf{A}_k\|_2^2 + O\left(\frac{1}{n^{1/C}}\right) \|\mathbf{A} - \mathbf{A}_k\|_F^2.\end{aligned}$$

Taking $C = 3$ gives the desired claim. □

9.6 Empirical Evaluation

In this section, we describe a series of experiments that demonstrate the advantage of our algorithm, performed in Python 3.6.9 on an Intel Core i7-8700K 3.70 GHz CPU with 12 cores and 64GB DDR4 memory, using an Nvidia Geforce GTX 1080 Ti 11GB GPU, on both synthetic and real-world data. Whereas previous work requires computing the top k subspace as a preprocessing step, our main improvement is that we only require a crude approximation. Thus we compared the running times for finding the top k subspace as required by [BK20c] to finding a mixed spectral-Frobenius approximation using an input sparsity algorithm, as required by our algorithm. For the former, we use the `svds` method from the sparse scipy linalg package optimized by LAPACK. For the latter, [CEM+15, CMM17] show that using a sparse CountSketch matrix [CW13, MM13a, NN13b], i.e., a matrix with $O(k^2)$ columns and a single nonzero entry in each row that is in a random location and is a random sign, suffices to obtain a mixed spectral-Frobenius guarantee; we evaluate such a matrix with exactly k^2 columns. Across all parameters and datasets, the input sparsity procedure used by our algorithm significantly outperforms the optimized power iteration methods required by [BK20c].

Synthetic Data. Since our theoretical results are most interesting when $k \ll d \ll n$, we set $n = 50000$, $d = 1000$, $k \in \{20, 50, 100\}$ and generate a random $d \times n$ matrix $\mathbf{A}$ that consists of independent entries that are each 1 with probability $p \in \left\{\frac{1}{500}, \frac{1}{2000}, \frac{1}{5000}\right\}$ and 0 with probability $1-p$. In Figure 9.1, we report the average running time of both algorithms, among 5 independent runs for each choice of p and k.

Mean Runtime of Algorithms across Parameters	$p = 1/500$	$p = 1/2000$	$p = 1/5000$
Top k Subspace, $k = 20$	35.056s	29.725s	16.45s
Input Sparsity Approximation, $k = 20$	0.595s	0.329s	0.83s
Top k Subspace, $k = 50$	56.146s	54.613s	53.213s
Input Sparsity Approximation, $k = 50$	0.658s	0.657s	0.434s
Top k Subspace, $k = 100$	78.420s	79.410s	71.424s
Input Sparsity Approximation, $k = 100$	0.501s	0.387s	0.440s

Figure 9.1: Mean runtime comparison of algorithms across parameters on synthetic data.

Social Networks. We also evaluate the algorithms on the `email-Eu-core network` dataset of interactions across email data between individuals from a large European research institution [YBLG17, LKF07] and the `com-Youtube` dataset of friendships on the Youtube social network [YL15], both accessed through the Stanford Network Analysis Project (SNAP). In the former, there are $n = d = 1005$ nodes in the adjacency matrix over 25571 total edges,

forming $k = 42$ communities. In the latter, there are 1134890 nodes with 8385 communities, from which we extract a $d \times n$ matrix with $n = 100000$, $d = 1000$ to represent a bipartite graph, as described in both Section 9.2.2 and [BK20c]. In Figure 9.2, we report the running time of both algorithms across each dataset among choices of $k \in \{20, 50, 100\}$. We observe that the resulting matrix has sparsity roughly 1000, which is consistent with $p \approx \frac{1}{n}$ and is much less than the sparsity parameters tested in our synthetic data.

	`email-Eu-core network`	`com-Youtube`
Top k Subspace, $k = 20$	0.387s	5.713s
Input Sparsity Approximation, $k = 20$	0.005s	0.379s
Top k Subspace, $k = 50$	0.556s	16.711s
Input Sparsity Approximation, $k = 50$	0.003s	0.373s
Top k Subspace, $k = 100$	1.281s	41.788s
Input Sparsity Approximation, $k = 100$	0.003s	0.366s

Figure 9.2: Mean runtime comparison of algorithms across parameters on real-world data.

Finally, we consider a full end-to-end implementation comparing the runtime and least squares loss of the top k subspace algorithm and our input sparsity approximation algorithm over various ranges of the parameter k and smoothening parameter δn on the `com-Youtube` dataset, from which we randomly extract an $n \times d$ matrix, with $n = 20000$ and $d = 1000$ to represent a bipartite graph. Our results in Figure 9.3 show that our algorithm not only significantly outperforms the top k subspace algorithm in runtime, but also produces solutions with lower least squared loss.

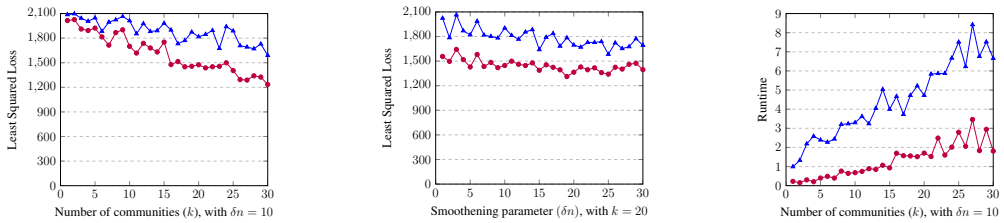

Figure 9.3: Comparison of least squares loss by power iteration algorithm (in blue triangles) and by our algorithm (in red circles), over various ranges of the parameter k with smoothening parameter $\delta n = 10$, and over various ranges of δn with $k = 20$, on the `com-Youtube` dataset. Also runtime comparison over a range of k, with $\delta n = 10$.

Bibliography

[AAK21] Naman Agarwal, Pranjal Awasthi, and Satyen Kale. A deep conditioning treatment of neural networks. In *Algorithmic Learning Theory*, pages 249–305. PMLR, 2021.

[ABB+19] Pranjal Awasthi, Ainesh Bakshi, Maria-Florina Balcan, Colin White, and David P Woodruff. Robust communication-optimal distributed clustering algorithms. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[ABEF14] Edoardo M. Airoldi, David M. Blei, Elena A. Erosheva, and Stephen E. Fienberg. Introduction to mixed membership models and methods, 2014.

[ABFX08] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, June 2008.

[ABGM14] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In *International Conference on Machine Learning*, pages 584–592, 2014.

[ABMM16] Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

[ACW17] Haim Avron, Kenneth L. Clarkson, and David P. Woodruff. Sharper bounds for regularized data fitting, 2017.

[AFKM01] Dimitris Achlioptas, Amos Fiat, Anna R Karlin, and Frank McSherry. Web search via hub synthesis. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 500–509. IEEE, 2001.

[AGGR98] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining

applications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 94–105, 1998.

[AGGR05] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery*, 11(1):5–33, 2005.

[AGGS17] Nima Anari, Leonid Gurvits, Shayan Oveis Gharan, and Amin Saberi. Simply exponential approximation of the permanent of positive semidefinite matrices. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 914–925. IEEE, 2017.

[AGH+13a] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288. PMLR, 2013.

[AGH+13b] Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 2013.

[AGHK14a] Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, 15(1):2239–2312, 2014.

[AGHK14b] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, and Sham M. Kakade. A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research*, 15(1):2239–2312, 2014.

[AGMR17] Sanjeev Arora, Rong Ge, Tengyu Ma, and Andrej Risteski. Provable learning of noisy-or networks. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1057–1066. ACM, 2017.

[AGMS12] Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ica with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In *Advances in Neural Information Processing Systems*, pages 2375–2383, 2012.

[AK05] Sanjeev Arora and Ravi Kannan. Learning mixtures of separated nonspherical gaussians. *The Annals of Applied Probability*, 15(1A):69–92, 2005.

[AK12] Koenraad MR Audenaert and Fuad Kittaneh. Problems and conjectures in matrix

and operator inequalities. *arXiv preprint arXiv:1201.5232*, 2012.

[ALN07] Sanjeev Arora, James R Lee, and Assaf Naor. Fréchet embeddings of negative type metrics. *Discrete & Computational Geometry*, 38(4):726–739, 2007.

[ALN08] Sanjeev Arora, James Lee, and Assaf Naor. Euclidean distortion and the sparsest cut. *Journal of the American Mathematical Society*, 21(1):1–21, 2008.

[AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

[AM15] Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

[AN13] Alexandr Andoni and Huy L. Nguyen. Eigenvalues of a matrix in the streaming model. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1729–1737. Society for Industrial and Applied Mathematics, 2013.

[Ara90] Huzihiro Araki. On an inequality of Lieb and Thirring. *LMaPh*, 19(2):167–170, 1990.

[ARV09] Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):5, 2009.

[AS12] Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. *CoRR*, abs/1206.3204, 2012.

[Ash] Robert B. Ash. Lecture notes 21-25 in statistics, finding the density. https://faculty.math.illinois.edu/~r-ash/Stat/StatLec21-25.pdf.

[ATV21] Pranjal Awasthi, Alex Tang, and Aravindan Vijayaraghavan. Efficient algorithms for learning depth-2 neural networks with general relu activations. *Advances in Neural Information Processing Systems*, 34:13485–13496, 2021.

[Aud08] Koenraad MR Audenaert. On a norm compression inequality for $2\times$ N partitioned block matrices. *Linear algebra and its applications*, 428(4):781–795, 2008.

[Avr10] Haim Avron. Counting triangles in large graphs using randomized matrix trace estimation. In *Workshop on Large-scale Data Mining: Theory and Applications*, volume 10, pages 10–9, 2010.

[Awa17] Pranjal Awasthi. Cs 598: Theoretical machine learning lecture notes, 2017. https://www.cs.rutgers.edu/~pa336/mlt_f17/lec-14.pdf.

[AY00] Charu C Aggarwal and Philip S Yu. Finding generalized projected clusters in high dimensional spaces. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 70–81, 2000.

[AZL19] Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *Advances in Neural Information Processing Systems*, 32, 2019.

[Bar] Boaz Barak. Proofs, beliefs, and algorithms through the lens of sum-of-squares.

[BBB+19] Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, and David P Woodruff. A PTAS for lp-low rank approximation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 747–766. SIAM, 2019.

[BBK+21a] Ainesh Bakshi, Chiranjib Bhattacharyya, Ravi Kannan, David Woodruff, and Samson Zhou. Learning a latent simplex in input sparsity time. In *International Conference on Learning Representations*, 2021.

[BBK+21b] Ainesh Bakshi, Chiranjib Bhattacharyya, Ravi Kannan, David P Woodruff, and Samson Zhou. Learning a latent simplex in input-sparsity time. *arXiv preprint arXiv:2105.08005*, 2021.

[BBV08] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 671–680, 2008.

[BCJ20] Ainesh Bakshi, Nadiia Chepurko, and Rajesh Jayaram. Testing positive semi-definiteness via random submatrices. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1191–1202. IEEE, 2020.

[BCMV14] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 594–603, 2014.

[BCPV19] Aditya Bhaskara, Aidao Chen, Aidan Perreault, and Aravindan Vijayaraghavan. Smoothed analysis in unsupervised learning via decoupling. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 582–610. IEEE, 2019.

[BCW19] Ainesh Bakshi, Nadiia Chepurko, and David P Woodruff. Weighted maximum independent set of geometric objects in turnstile streams. *arXiv preprint*

arXiv:1902.10328, 2019.

[BCW20a] Ainesh Bakshi, Nadiia Chepurko, and David P Woodruff. Robust and sample optimal algorithms for psd low rank approximation. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 506–516. IEEE, 2020.

[BCW20b] Ainesh Bakshi, Nadiia Chepurko, and David P Woodruff. Robust and sample optimal algorithms for PSD low rank approximation. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 506–516. IEEE, 2020.

[BCW22] Ainesh Bakshi, Kenneth L Clarkson, and David P Woodruff. Low-rank approximation with $1/\epsilon^{1/3}$ matrix-vector products. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1130–1143, 2022.

[BDH+20] Ainesh Bakshi, Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, Sushrut Karmalkar, and Pravesh K Kothari. Outlier-robust clustering of gaussians and other non-spherical mixtures. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 149–159. IEEE, 2020.

[BDJ+22] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M Kane, Pravesh K Kothari, and Santosh S Vempala. Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247, 2022.

[BDL18] Digvijay Boob, Santanu S Dey, and Guanghui Lan. Complexity of training relu neural network. *arXiv preprint arXiv:1809.10787*, 2018.

[BDN15] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 499–508, 2015.

[BFG96] Zhaojun Bai, Gark Fahey, and Gene Golub. Some large-scale matrix computation problems. *Journal of Computational and Applied Mathematics*, 74(1-2):71–89, 1996.

[BG17] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

[Bha13] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[BHSW20] Mark Braverman, Elad Hazan, Max Simchowitz, and Blake Woodworth. The gradient complexity of linear regression. In *Conference on Learning Theory*, pages 627–647. PMLR, 2020.

[BJ03] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.

[BJW19] Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory*, pages 195–268. PMLR, 2019.

[BK20a] Ainesh Bakshi and Pravesh Kothari. List-decodable subspace recovery via sum-of-squares. *arXiv preprint arXiv:2002.05139*, 2020.

[BK20b] Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. *arXiv preprint arXiv:2005.02970*, 2020.

[BK20c] Chiranjib Bhattacharyya and Ravindran Kannan. Finding a latent k-simplex in O^* ($k \cdot$ nnz(data)) time via subset smoothing. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms*, pages 122–140. SIAM, 2020.

[BK20d] Chiranjib Bhattacharyya and Ravindran Kannan. Finding a latent k–simplex in o*(k· nnz (data)) time via subset smoothing. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 122–140. SIAM, 2020.

[BK21] Ainesh Bakshi and Pravesh K Kothari. List-decodable subspace recovery: Dimension independent error in polynomial time. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1279–1297. SIAM, 2021.

[BKKS19] Vladimir Braverman, Robert Krauthgamer, Aditya Krishnan, and Roi Sinoff. Schatten norms in matrix streams: Hello sparsity, goodbye dimension. *arXiv preprint arXiv:1907.05457*, 2019.

[BKL02] Rajendra Bhatia, William Kahan, and Ren-Cang Li. Pinchings and norms of scaled triangular matrices. *Linear and Multilinear Algebra*, 50(1):15–21, 2002.

[BKS15] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151, 2015.

[BKS17] Boaz Barak, Pravesh K Kothari, and David Steurer. Quantum entanglement, sum of squares, and the log rank conjecture. In *Proceedings of the 49th Annual ACM*

SIGACT Symposium on Theory of Computing, pages 975–988, 2017.

[BL06a] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[BL06b] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[Ble12] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

[BM16] Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*, pages 417–445, 2016.

[BNJ03] David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[BP21] Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 102–115, 2021.

[BR92] Avrim Blum and Ronald L. Rivest. Training a 3-node neural network is np-complete. *Neural Networks*, 5(1):117–127, 1992.

[Bry12] Wlodzimierz Bryc. *The normal distribution: characterizations with applications*, volume 100. Springer Science & Business Media, 2012.

[BS15] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. *SIAM Journal on Computing*, 44(4):889–911, 2015.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[BV08] S Charles Brubaker and Santosh S Vempala. Isotropic pca and affine-invariant clustering. In *Building Bridges*, pages 241–281. Springer, 2008.

[BW18] Ainesh Bakshi and David Woodruff. Sublinear time low-rank approximation of distance matrices. In *Advances in Neural Information Processing Systems*, pages 3782–3792, 2018.

[BWZ16] Christos Boutsidis, David P Woodruff, and Peilin Zhong. Optimal principal component analysis in distributed and streaming models. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 236–249, 2016.

[BWZ19] Frank Ban, David Woodruff, and Qiuyi Zhang. Regularized weighted low rank approximation. *arXiv preprint arXiv:1911.06958*, 2019.

[BY02] Ziv Bar-Yossef. *The complexity of massive data set computations*. PhD thesis,

University of California, Berkeley, 2002.

[CAT+20] Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I Jordan, Nicolas Flammarion, and Peter L Bartlett. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020.

[CCH+20] Nadiia Chepurko, Kenneth L Clarkson, Lior Horesh, Honghao Lin, and David P Woodruff. Quantum-inspired algorithms from randomized numerical linear algebra. *arXiv preprint arXiv:2011.04125*, 2020.

[CEM+15] Michael B Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 163–172. ACM, 2015.

[CFZ99] Chun-Hung Cheng, Ada Waichee Fu, and Yi Zhang. Entropy-based subspace clustering for mining numerical data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 84–93, 1999.

[CG92] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332, 1992.

[CGKM22] Sitan Chen, Aravind Gollakota, Adam R Klivans, and Raghu Meka. Hardness of noise-free learning for two-hidden-layer neural networks. *arXiv preprint arXiv:2202.05258*, 2022.

[CGR05] Shuchi Chawla, Anupam Gupta, and Harald Räcke. Embeddings of negative-type metrics and an improved approximation to generalized sparsest cut. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 102–111. Society for Industrial and Applied Mathematics, 2005.

[CK98] João Paulo Costeira and Takeo Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

[CKM+11] Paul Christiano, Jonathan A Kelner, Aleksander Madry, Daniel A Spielman, and Shang-Hua Teng. Electrical flows, laplacian systems, and faster approximation of maximum flow in undirected graphs. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 273–282. ACM, 2011.

[CKM22] Sitan Chen, Adam R Klivans, and Raghu Meka. Learning deep relu networks is

fixed-parameter tractable. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 696–707. IEEE, 2022.

[CLL+10] Pei-Chun Chen, Kuang-Yao Lee, Tsung-Ju Lee, Yuh-Jye Lee, and Su-Yun Huang. Multiclass support vector classification via coding and regression. *Neurocomputing*, 73(7-9):1501–1512, 2010.

[CLM+15] Michael B Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 181–190. ACM, 2015.

[CLMW11] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[CLW18] Nai-Hui Chia, Han-Hsuan Lin, and Chunhao Wang. Quantum-inspired sublinear classical algorithms for solving low-rank linear systems. *arXiv preprint arXiv:1811.04852*, 2018.

[CMM17] Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 1758–1777, 2017.

[CMTV17] Michael B Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. Matrix scaling and balancing via box constrained newton's method and interior point methods. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 902–913. IEEE, 2017.

[CNW15] Michael B Cohen, Jelani Nelson, and David P Woodruff. Optimal approximate matrix product in terms of stable rank. *arXiv preprint arXiv:1507.02268*, 2015.

[Coh16] Michael B Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 278–287. SIAM, 2016.

[Com94] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[Con] Keith Conrad. Expository papers: Universal identities. http://www.math.uconn.edu/~kconrad/blurbs/linmultialg/univid.pdf.

[CP10] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[CR07] Emmanuel Candes and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.

[CR09] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[CRR+96] Ashok K Chandra, Prabhakar Raghavan, Walter L Ruzzo, Roman Smolensky, and Prasoon Tiwari. The electrical resistance of a graph captures its commute and cover times. *Computational Complexity*, 6(4):312–340, 1996.

[CSV13] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.

[CW01] Anthony Carbery and James Wright. Distributional and lq norm inequalities for polynomials over convex bodies in rn. *Mathematical research letters*, 8(3):233–248, 2001.

[CW09] Kenneth L Clarkson and David P Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 205–214. ACM, 2009.

[CW13] Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 81–90. ACM, 2013.

[CW17] Kenneth L Clarkson and David P Woodruff. Low-rank psd approximation in input-sparsity time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2061–2072. Society for Industrial and Applied Mathematics, 2017.

[Das99] Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

[Das08] A. DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer New York, 2008.

[DFK+04] Petros Drineas, Alan Frieze, Ravi Kannan, Santosh Vempala, and V Vinay. Clus-

tering large graphs via the singular value decomposition. *Machine learning*, 56(1-3):9–33, 2004.

[DG18] Simon S Du and Surbhi Goel. Improved learning of one-hidden-layer convolutional neural networks with overlaps. *arXiv preprint arXiv:1805.07798*, 2018.

[DHKK20] Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*, 2020.

[DK19] Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

[DK20] Ilias Diakonikolas and Daniel M Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 184–195. IEEE, 2020.

[DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[DKKZ20] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pages 1514–1539. PMLR, 2020.

[DKM06] Petros Drineas, Ravi Kannan, and Michael W Mahoney. Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, 2006.

[DKR02] Petros Drineas, Iordanis Kerenidis, and Prabhakar Raghavan. Competitive recommendation systems. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 82–90. ACM, 2002.

[DKS17] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of*

the 50th Annual ACM SIGACT Symposium on Theory of Computing, pages 1047–1060, 2018.

[DKS19] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.

[DL09] Michel Marie Deza and Monique Laurent. *Geometry of cuts and metrics*, volume 15. Springer, 2009.

[DRST09] Ilias Diakonikolas, Prasad Raghavendra, Rocco A Servedio, and Li-Yang Tan. Average sensitivity and noise sensitivity of polynomial threshold functions. *arXiv preprint arXiv:0909.5011*, 2009.

[DSD+13] Misha Denil, Babak Shakibi, Laurent Dinh, Nando De Freitas, et al. Predicting parameters in deep learning. In *Advances in neural information processing systems*, pages 2148–2156, 2013.

[Dua20] Leo L Duan. Latent simplex position model: High dimensional multi-view clustering with uncertainty quantification. *Journal of Machine Learning Research*, 21(38):1–25, 2020.

[DVW18] Ilias Diakonikolas, Santosh Vempala, and David Woodruff. Research vignette: Foundations of data science. *Simons Institute, Semester on Foundations of Big Data*, 2018.

[DVW19] Ilias Diakonikolas, Santosh Vempala, and David Woodruff. Research vignette: Foundations of data science, 2019.

[DZB+14] Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in neural information processing systems*, pages 1269–1277, 2014.

[ELMM20] Yonina C Eldar, Jerry Li, Cameron Musco, and Christopher Musco. Sample efficient toeplitz covariance estimation. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 378–397. SIAM, 2020.

[EV13] Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.

[FB81] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[FJK96] Alan Frieze, Mark Jerrum, and Ravi Kannan. Learning linear transformations. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 359–368. IEEE, 1996.

[FKP+19] Noah Fleming, Pravesh Kothari, Toniann Pitassi, et al. *Semialgebraic Proofs and Efficient Algorithm Design.* now the essence of knowledge, 2019.

[FKV04a] Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.

[FKV04b] Alan M. Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, 2004.

[FSS13] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1434–1453. Society for Industrial and Applied Mathematics, 2013.

[FSX09] Wenjie Fu, Le Song, and Eric P Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th annual international conference on machine learning*, pages 329–336, 2009.

[FT07] Shmuel Friedland and Anatoli Torokhti. Generalized rank-constrained matrix approximations. *SIAM Journal on Matrix Analysis and Applications*, 29(2):656–659, 2007.

[FXC16] Xuhui Fan, Richard Yi Da Xu, and Longbing Cao. Copula mixed-membership stochastic block model. In *IJCAI International Joint Conference on Artificial Intelligence*, 2016.

[Ge18] Rong Ge. Personal communication. October, 2018.

[GH+96] Zoubin Ghahramani, Geoffrey E Hinton, et al. The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

[Gil20] Nicolas Gillis. *Nonnegative Matrix Factorization*. SIAM, 2020.

[GK17] Surbhi Goel and Adam Klivans. Learning depth-three neural networks in polynomial time. *arXiv preprint arXiv:1709.06010*, 2017.

[GKKT16] Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. *arXiv preprint arXiv:1611.10258*, 2016.

[GKLW18] Rong Ge, Rohith Kuditipudi, Zhize Li, and Xiang Wang. Learning two-layer neural networks with symmetric inputs. *arXiv preprint arXiv:1810.06793*, 2018.

[GKM18] Surbhi Goel, Adam Klivans, and Raghu Meka. Learning one convolutional layer with overlapping patches. *arXiv preprint arXiv:1802.02547*, 2018.

[GKX19] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241. PMLR, 09–15 Jun 2019.

[GLF+10] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

[GLM17] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

[GLS81] Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

[GLT18] András Gilyén, Seth Lloyd, and Ewin Tang. Quantum-inspired low-rank stochastic regression with logarithmic dependence on the dimension. *arXiv preprint arXiv:1811.04909*, 2018.

[GM15] Rong Ge and Tengyu Ma. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. *arXiv preprint arXiv:1504.05287*, 2015.

[GNC99] Sanjay Goil, Harsha Nagesh, and Alok Choudhary. Mafia: Efficient and scalable subspace clustering for very large data sets. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 443, page 452. ACM, 1999.

[Gro11] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[Gru17] Marvin Gruber. *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*. Routledge, 2017.

[GSLW19] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics. In *Proceedings of the 51st Annual ACM SIGACT Symposium on The-*

ory of Computing, pages 193–204. ACM, 2019.

[Gut09] Allan Gut. An intermediate course in probability. chapter 5. Springer Publishing Company, Incorporated, 2009.

[GV14] Nicolas Gillis and Stephen A. Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714, 2014.

[GVX14] Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier pca and robust tensor decomposition. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 584–593. ACM, 2014.

[GXM+17] Shuhang Gu, Qi Xie, Deyu Meng, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. Weighted nuclear norm minimization and its applications to low level vision. *International journal of computer vision*, 121(2):183–208, 2017.

[GZZF14] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014.

[Har14] Moritz Hardt. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 651–660. IEEE, 2014.

[HBB10] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[Hig02] Nicholas J Higham. Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis*, 22(3):329–343, 2002.

[HK13] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

[HL13] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

[HL18] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

[HM13] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace

recovery. In *Conference on Learning Theory*, pages 354–375. PMLR, 2013.

[HO00] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[HP15] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760, 2015.

[HRRS11] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[HS17] Samuel B. Hopkins and David Steurer. Efficient bayesian estimation from few samples: Community detection and related problems. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 379–390, 2017.

[Hub64] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[Hub04] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[Hub11] Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer, 2011.

[HWHM06] Wei Hong, John Wright, Kun Huang, and Yi Ma. Multiscale hybrid linear models for lossy image representation. *IEEE Transactions on Image Processing*, 15(12):3655–3671, 2006.

[Hyv99] Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Ind06] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM (JACM)*, 53(3):307–323, 2006.

[IVWW19] Piotr Indyk, Ali Vakilian, Tal Wagner, and David Woodruff. Sample-optimal low-rank approximation of distance matrices. *arXiv preprint arXiv:1906.00339*, 2019.

[Jae72] Louis A Jaeckel. Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics*, pages 1449–1458, 1972.

[JMM20] Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *The Annals of Statistics*, 48(6):3619–3642, 2020.

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[JS89] Mark Jerrum and Alistair Sinclair. Approximating the permanent. *SIAM journal on computing*, 18(6):1149–1178, 1989.

[JSA14] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*, 2014.

[JSA15] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[JSV04] Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004.

[Jud88] J Stephen Judd. Neural network design and the complexity of learning. Technical report, CALIFORNIA INST OF TECH PASADENA DEPT OF COMPUTER SCIENCE, 1988.

[Kan20] Daniel M Kane. Robust learning of mixtures of gaussians. *arXiv preprint arXiv:2007.05912*, 2020.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[KK10] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. *FOCS*, 2010.

[KKK19] Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. List-decodable linear regression. In *Advances in Neural Information Processing Systems*, pages 7423–7432, 2019.

[KKM18] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. *arXiv preprint arXiv:1803.03241*, 2018.

[KKSK11] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In

Advances in Neural Information Processing Systems, pages 927–935, 2011.

[Kle99] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[KLM+17] Michael Kapralov, Yin Tat Lee, CN Musco, Christopher Paul Musco, and Aaron Sidford. Single pass spectral sparsification in dynamic streams. *SIAM Journal on Computing*, 46(1):456–477, 2017.

[KMO10] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[KMP14] Ioannis Koutis, Gary L Miller, and Richard Peng. Approaching optimality for solving sdd linear systems. *SIAM Journal on Computing*, 43(1):337–354, 2014.

[KMV10] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010.

[KOSZ13] Jonathan A Kelner, Lorenzo Orecchia, Aaron Sidford, and Zeyuan Allen Zhu. A simple, combinatorial algorithm for solving sdd systems in nearly-linear time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 911–920, 2013.

[KOTZ14] Manuel Kauers, Ryan O'Donnell, Li-Yang Tan, and Yuan Zhou. Hypercontractive inequalities via sos, and the frankl–rödl graph. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1644–1658. SIAM, 2014.

[KP16] Iordanis Kerenidis and Anupam Prakash. Quantum recommendation systems. *arXiv preprint arXiv:1603.08675*, 2016.

[KS17] Pravesh K Kothari and David Steurer. Outlier-robust moment-estimation via sum-of-squares. *arXiv preprint arXiv:1711.11581*, 2017.

[KSS18] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

[KSV05] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *International Conference on Computational Learning Theory*, pages 444–457. Springer, 2005.

[KV09] Ravi Kannan and Santosh S. Vempala. Spectral algorithms. *Found. Trends Theor. Comput. Sci.*, 4(3-4):157–288, 2009.

[LAF+12] Yi-Kai Liu, Animashree Anandkumar, Dean P Foster, Daniel Hsu, and Sham M Kakade. Two svds su ffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. In *Neural Information Processing Systems (NIPS)*, 2012.

[Las01] Jean B Lasserre. New positive semidefinite relaxations for nonconvex quadratic programs. In *Advances in Convex Analysis and Global Optimization*, pages 319–331. Springer, 2001.

[Lau09] Monique Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, pages 157–270. Springer, 2009.

[LAW16] Wenzhe Li, Sungjin Ahn, and Max Welling. Scalable mcmc for mixed membership stochastic blockmodels. In *Artificial Intelligence and Statistics*, pages 723–731, 2016.

[LB11] Tyler Lu and Craig Boutilier. Learning mallows models with pairwise preferences, 2011.

[LC15] Sergey Loyka and Charalambos D. Charalambous. Novel matrix singular value inequalities and their applications to uncertain MIMO channels. *IEEE Trans. Inf. Theory*, 61(12):6623–6634, 2015.

[LKF07] Jure Leskovec, Jon M. Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2, 2007.

[LLY+12] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012.

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 10 2000.

[LM18a] Gilad Lerman and Tyler Maunu. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8):1380–1410, 2018.

[LM18b] Allen Liu and Ankur Moitra. Efficiently learning mixtures of mallows models. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 627–638. IEEE, 2018.

[LM21] Allen Liu and Ankur Moitra. Settling the robust learnability of mixtures of gaus-

sians. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 518–531, 2021.

[LMP13] Mu Li, Gary L Miller, and Richard Peng. Iterative row sampling. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 127–136. IEEE, 2013.

[LMZ+12] Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *European conference on computer vision*, pages 347–360. Springer, 2012.

[LNW14a] Yi Li, Huy L Nguyen, and David P Woodruff. On sketching matrix norms and the top singular vector. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1562–1581. SIAM, 2014.

[LNW14b] Yi Li, Huy L. Nguyen, and David P. Woodruff. Turnstile streaming algorithms might as well be linear sketches. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 174–183, 2014.

[LRV16] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

[LS13] Jörg Liesen and Zdenek Strakos. *Krylov subspace methods: principles and analysis*. Oxford University Press, 2013.

[LS15] Yin Tat Lee and Aaron Sidford. Efficient inverse maintenance and faster algorithms for linear programming. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 230–249. IEEE, 2015.

[LSS+] Erik M Lindgren, Vatsal Shah, Yanyao Shen, Alexandros G Dimakis, and Adam Klivans. On robust learning of ising models.

[LSSS14] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

[LSW15] Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1049–1065. IEEE, 2015.

[LW16a] Yi Li and David P Woodruff. On approximating functions of the singular values

in a stream. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 726–739, 2016.

[LW16b] Yi Li and David P Woodruff. Tight bounds for sketching the operator norm, Schatten norms, and subspace embeddings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[LW17] Yi Li and David P Woodruff. Embeddings of Schatten norms with applications to data streams. In *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[LW20] Yi Li and David P. Woodruff. Input-sparsity low rank approximation in Schatten norm. *CoRR*, abs/2004.12646, 2020.

[LY17a] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

[LY17b] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 597–607. Curran Associates, Inc., 2017.

[M+11] Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[Mah90] Philip J Maher. Some operator inequalities concerning generalized inverses. *Illinois Journal of Mathematics*, 34(3):503–514, 1990.

[Mah11] Michael W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, 2011.

[McS01] Frank McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE, 2001.

[Meg88] Nimrod Megiddo. On the complexity of polyhedral separability. *Discrete & Computational Geometry*, 3(4):325–337, 1988.

[MH02] John C Mason and David C Handscomb. *Chebyshev polynomials*. CRC press, 2002.

[MJG09] Kurt Miller, Michael I Jordan, and Thomas L Griffiths. Nonparametric latent

feature models for link prediction. In *Advances in neural information processing systems*, pages 1276–1284, 2009.

[MM13a] Xiangrui Meng and Michael W Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 91–100. ACM, 2013.

[MM13b] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 91–100, 2013.

[MM14] Brian McWilliams and Giovanni Montana. Subspace clustering of high-dimensional data: a predictive approach. *Data Mining and Knowledge Discovery*, 28(3):736–772, 2014.

[MM15] Cameron Musco and Christopher Musco. Randomized block Krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems*, pages 1396–1404, 2015.

[MM17] Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. In *Advances in Neural Information Processing Systems*, pages 3833–3845, 2017.

[MM18] Marco Mondelli and Andrea Montanari. On the connection between learning two-layers neural networks and tensor decomposition. *arXiv preprint arXiv:1802.07301*, 2018.

[MMMW21] Raphael A. Meyer, Cameron Musco, Christopher Musco, and David P. Woodruff. Hutch++: Optimal stochastic trace estimation. In *4th Symposium on Simplicity in Algorithms, SOSA 2021, Virtual Conference, January 11-12, 2021*, pages 142–155, 2021.

[MR18] Pasin Manurangsi and Daniel Reichman. The computational complexity of training relu(s). *arXiv preprint arXiv:1810.04207*, 2018.

[MST15] Aleksander Madry, Damian Straszak, and Jakub Tarnawski. Fast generation of random spanning trees and the effective resistance metric. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 2019–2036. Society for Industrial and Applied Mathematics, 2015.

[MT+11] Michael McCoy, Joel A Tropp, et al. Two proposals for robust pca using semidef-

inite programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011.

[Mut05] S. Muthukrishnan. Data streams: Algorithms and applications. *Found. Trends Theor. Comput. Sci.*, 1(2), 2005.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

[MW17a] Cameron Musco and David Woodruff. Is input sparsity time possible for kernel low-rank approximation? *Advances in Neural Information Processing Systems*, 30:4435–4445, 2017.

[MW17b] Cameron Musco and David P Woodruff. Sublinear time low-rank approximation of positive semidefinite matrices. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–683. IEEE, 2017.

[MW17c] Cameron Musco and David P. Woodruff. Sublinear time low-rank approximation of positive semidefinite matrices. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 672–683, 2017.

[MW21] Arvind V Mahankali and David P Woodruff. Optimal L1 column subset selection and a fast PTAS for low rank approximation. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 560–578. SIAM, 2021.

[MZ10] Lingsheng Meng and Bing Zheng. The optimal perturbation bounds of the moore–penrose inverse under the frobenius norm. *Linear Algebra and its Applications*, 432(4):956 – 963, 2010.

[Nel11] Jelani Jelani Osei Nelson. *Sketching and streaming high-dimensional vectors*. PhD thesis, Massachusetts Institute of Technology, 2011.

[Nes00] Yurii Nesterov. Squared functional systems and optimization problems. In *High performance optimization*, pages 405–440. Springer, 2000.

[NN13a] Jelani Nelson and Huy L. Nguyen. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 117–126, 2013.

[NN13b] Jelani Nelson and Huy L. Nguyen. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 117–126. IEEE Computer Soci-

ety, 2013.

[O'D14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[Par00] Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000.

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[Pea94] Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6:147–160, 1994.

[PHL04] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm sigkdd explorations newsletter*, 6(1):90–105, 2004.

[PJAM02] Cecilia M Procopiuc, Michael Jones, Pankaj K Agarwal, and TM Murali. A monte carlo algorithm for fast projective clustering. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pages 418–427, 2002.

[PPZ+20] William Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. The Hessian penalty: A weak prior for unsupervised disentanglement. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[PSBR20] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):601–627, 2020.

[PV13] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

[PV21] Richard Peng and Santosh Vempala. Solving sparse linear systems faster than matrix multiplication. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 504–521. SIAM, 2021.

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[Riv74] Theodore J. Rivlin. The chebyshev polynomials. *Wiley*, 1974.

[Riv20] Theodore J Rivlin. *Chebyshev polynomials*. Courier Dover Publications, 2020.

[Rou84] Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

[RS00] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[RSML18] Patrick Rebentrost, Adrian Steffens, Iman Marvian, and Seth Lloyd. Quantum singular-value decomposition of nonsparse low-rank matrices. *Physical review A*, 97(1):012327, 2018.

[RSW16] Ilya Razenshteyn, Zhao Song, and David P Woodruff. Weighted low rank approximations with provable guarantees. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 250–263, 2016.

[RV07] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21, 2007.

[RV10] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

[RWYZ21] Cyrus Rashtchian, David P. Woodruff, Peng Ye, and Hanlin Zhu. Average-case communication complexity of statistical problems, 2021.

[RWZ20] Cyrus Rashtchian, David P. Woodruff, and Hanlin Zhu. Vector-matrix-vector queries for solving linear algebra, statistics, and graph problems. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2020, August 17-19, 2020, Virtual Conference*, pages 26:1–26:20, 2020.

[RY84] Peter Rousseeuw and Victor Yohai. Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer, 1984.

[RY20a] Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 161–180. SIAM, 2020.

[RY20b] Prasad Raghavendra and Morris Yau. List decodable subspace recovery. In *Conference on Learning Theory*, pages 3206–3226. PMLR, 2020.

[SA14] Hanie Sedghi and Anima Anandkumar. Provable methods for training neural networks with sparse connectivity. *arXiv preprint arXiv:1412.2693*, 2014.

[Saa81] Yousef Saad. Krylov subspace methods for solving large unsymmetric linear systems. *Mathematics of computation*, 37(155):105–126, 1981.

[Sar06] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152, 2006.

[SAR18] Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. Tight query complexity lower bounds for PCA via finite sample deformed wigner law. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1249–1259, 2018.

[SC16] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[Sch38] Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[Sch60] Robert Schatten. Norm ideals of completely continuous operators. 1960.

[SEC14] Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J Candes. Robust subspace clustering. *The annals of Statistics*, 42(2):669–699, 2014.

[Sen68] Pranab Kumar Sen. Estimates of the regression coefficient based on kendall's tau. *Journal of the American statistical association*, 63(324):1379–1389, 1968.

[SG07] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[Sho87] Naum Z Shor. Quadratic optimization problems. *Soviet Journal of Computer and Systems Sciences*, 25:1–11, 1987.

[SJ03] Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 720–727, 2003.

[SJA16] Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, pages 1223–1231, 2016.

[SK01] Arora Sanjeev and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*,

pages 247–257, 2001.

[Sol17] Mahdi Soltanolkotabi. Learning relus via gradient descent. In *Advances in Neural Information Processing Systems*, pages 2007–2017, 2017.

[SS11] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

[ST14] Daniel A Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, 35(3):835–885, 2014.

[SW19] Xiaofei Shi and David P. Woodruff. Sublinear time numerical linear algebra for structured matrices. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, pages 4918–4925, 2019.

[SWW12] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, pages 37–1, 2012.

[SWYZ19] Xiaoming Sun, David P. Woodruff, Guang Yang, and Jialin Zhang. Querying a matrix through matrix-vector products. In *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece*, pages 94:1–94:16, 2019.

[SWZ16] Zhao Song, David Woodruff, and Huan Zhang. Sublinear time orthogonal tensor decomposition. In *Advances in Neural Information Processing Systems*, pages 793–801, 2016.

[SWZ17] Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise l1-norm error. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 688–701, 2017.

[SWZ20] Zhao Song, David P Woodruff, and Peilin Zhong. Average case column subset selection for entrywise l1-norm loss. *arXiv preprint arXiv:2004.07986*, 2020.

[Tan19] Ewin Tang. A quantum-inspired classical algorithm for recommendation systems. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 217–228. ACM, 2019.

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[Tao20] Terence Tao. Notes 3a: Eigenvalues and sums of hermitian matrices, 2020.

[TD87] Paul Terwilliger and Michel Deza. The classification of finite connected hypermetric spaces. *Graphs and Combinatorics*, 3(1):293–298, 1987.

[TDSL00] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[The92] Henri Theil. A rank-invariant method of linear and polynomial regression analysis. In *Henri Theil's contributions to economics and econometrics*, pages 345–381. Springer, 1992.

[Tia17a] Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. *arXiv preprint arXiv:1703.00560*, 2017.

[Tia17b] Yuandong Tian. Symmetry-breaking convergence analysis of certain two-layered neural networks with relu nonlinearity. 2017.

[Tso08] Charalampos E Tsourakakis. Fast counting of triangles in large real networks without counting: Algorithms and laws. In *2008 Eighth IEEE International Conference on Data Mining*, pages 608–617. IEEE, 2008.

[TV17] Manolis C Tsakiris and René Vidal. Hyperplane clustering via dual principal component pursuit. In *International conference on machine learning*, pages 3472–3481. PMLR, 2017.

[Val84] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Ver10a] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Ver10b] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[VMS05] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959, 2005.

[VN37] John Von Neumann. *Some matrix-inequalities and metrization of matric space*.

1937.

[VN18] Namrata Vaswani and Praneeth Narayanamurthy. Static and dynamic robust pca and matrix completion: A review. *Proceedings of the IEEE*, 106(8):1359–1379, 2018.

[VW04] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[WA16] Yining Wang and Anima Anandkumar. Online and differentially-private tensor decomposition. In *Advances in Neural Information Processing Systems*, pages 3531–3539, 2016.

[Wai19] M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[Web03] Marcus Weber. Clustering by using a simplex structure. 2003.

[Wed72] Per-Ake Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[Wei05] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[Woo14a] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

[Woo14b] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[WT10] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

[WWZ14] Karl Wimmer, Yi Wu, and Peng Zhang. Optimal query complexity for estimating the trace of a matrix. In *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, pages 1051–1062, 2014.

[XCS10] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. *arXiv preprint arXiv:1010.4237*, 2010.

[XFS+10] Eric P Xing, Wenjie Fu, Le Song, et al. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*,

4(2):535–566, 2010.

[Yao77] Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the 18th Annual Symposium on Foundations of Computer Science*, SFCS '77, pages 222–227, Washington, DC, USA, 1977. IEEE Computer Society.

[YBLG17] Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 555–564. ACM, 2017.

[YGKM20] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney. PyHessian: Neural networks through the lens of the Hessian, 2020.

[YL15] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. *Knowl. Inf. Syst.*, 42(1):181–213, 2015.

[YP21] Chenyang Yuan and Pablo A Parrilo. Maximizing products of linear forms, and the permanent of positive semidefinite matrices. *Mathematical Programming*, pages 1–12, 2021.

[YPCC16] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

[YZ16] Ming Yuan and Cun-Hui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

[ZAX12] Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models. *Journal of Machine Learning Research*, 13(Aug):2237–2278, 2012.

[ZFIM12] Amy Zhang, Nadia Fawaz, Stratis Ioannidis, and Andrea Montanari. Guess who rated this movie: Identifying users through subspace clustering. *arXiv preprint arXiv:1208.1544*, 2012.

[ZJS19] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *arXiv preprint arXiv:1909.08755*, 2019.

[ZJS20] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Robust estimation via generalized quasi-gradients. *arXiv preprint arXiv:2005.14073*, 2020.

[ZLJ16] Yuchen Zhang, Jason D Lee, and Michael I Jordan. l1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning*, pages 993–1001, 2016.

[ZSJ+17] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.

[ZWR+18] Zhihui Zhu, Yifan Wang, Daniel Robinson, Daniel Naiman, Rene Vidal, and Manolis Tsakiris. Dual principal component pursuit: Improved analysis and efficient algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.