New Advances in Sparse Learning, Deep Networks, and Adversarial Learning: Theory and Applications

Hongyang Zhang

May 2019
CMU-ML-19-103

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

Thesis Committee:
Maria-Florina Balcan, Co-chair
David P. Woodruff, Co-chair
Ruslan Salakhutdinov
Avrim Blum (Toyota Technological Institute at Chicago)

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

This research was sponsored by the National Science Foundation awards CCF1422910, CCF1535967, and IIS1618714 and the Office of Naval Research award N000141812562.

Keywords: Sparse Learning, Deep Networks, Adversarial Learning, Robustness, Optimization, Sample Efficiency

To my family

Abstract

Sparse learning, deep networks, and adversarial learning are new paradigms and have received significant attention in recent years due to their wide applications to various big data problems in computer vision, natural language processing, statistics, and theoretical computer science. The paradigms include learning with sparsity, learning with low-rank approximations, and learning with deep neural networks, corresponding to the assumptions that data have only a few non-zero coordinates, lie on low-rank subspaces, and lie on low-dimensional manifolds, respectively. The focus of this study is to develop algorithms which are sample-efficient, are easier to optimize, and are robust to adversarial corruptions.

Despite a large amount of work on these new paradigms, many fundamental questions remain unresolved. From the statistical aspect, understanding the *tight* sample complexity of big data problems is an important research question. Intuitively, the intrinsic dimension of structured data should be much smaller than their ambient dimension. Because the true sample complexity should be comparable to the intrinsic dimension rather than the ambient dimension, this implies the possibility of sub-linear sample complexity w.r.t. the ambient dimension. In this thesis, we design principled, practical and scalable algorithms for big data problems with near-optimal sample complexity. These include models of matrix completion, robust PCA, margin-based active learning, property testing, and compressed sensing.

From the computational aspects, direct formulations of these new paradigms are non-convex and NP-hard to optimize in general. Therefore, one of the long-standing questions is designing computationally efficient algorithms by taking into account the structure of the data. In this thesis, we develop new paradigms toward global optimality of non-convex optimization in polynomial time. In particular, we design algorithms and understand the landscape (e.g., duality gap) for the problems of (1-bit) compressed sensing, deep neural networks, GAN, and matrix factorization.

From the robustness aspects, models such as deep networks are vulnerable to adversarial examples. Although the problem has been widely studied empirically, much remains unknown concerning the theory underlying designing defense methods. There are two types of adversarial examples: training-time adversarial examples, such as data positioning, and inference-time adversarial examples. We discuss both types of adversarial examples in this thesis, for the problems of (1-bit) compressed sensing and robust PCA, as well as the problems of deep networks by adversarial learning.

Beyond theoretical contributions, our work also has significant practical impact. For example, inspired by our theoretical analysis, we design a new defense method, TRADES, against inference-time adversarial examples. Our proposed algorithm is the winner of the NeurIPS 2018 Adversarial Vision Challenge in which we won the 1st place out of 1,995 submissions, surpassing the runner-up approach by 11.41% in terms of mean ℓ_2 perturbation distance.

Acknowledgments

Time passes very quickly, often so quickly that you are surprised. I still remember four years ago, when I first time landed on the Pittsburgh airport, 11,717 kilometers away from my home town Nanjing, having two suitcases with me almost as large as myself; everything just looked unfamiliar to me. Four years later, I am going to graduate, with my advisors, my friends, my classmates by my side; CMU is my second home now. I would like to thank many people.

Firstly, I would like to thank my advisors Prof. Nina Balcan and Prof. David P. Woodruff. Nina is the first professor that I worked with in U.S. I not only learn a lot from her academic knowledge, but also learn how to be an excellent researchers. I want to thank David as well. I really enjoy every personal meeting with him. Thanks, Nina and David, for teaching me so much.

Secondly, I would like to thank my committee members Prof. Avrim Blum and Prof. Ruslan Salakhutdinov. Their valuable comments on my thesis and constructive suggestions on my oral defense significantly improve the presentation of this thesis.

I would like to thank Prof. Avrim Blum and Prof. Greg Shakhnarovich in Toyota Technological Institute at Chicago (TTIC) for providing me the best Postdoc position so that I can continue my research. Looking forward to working with them very soon.

I also want to thank my collaborators. They are (ordered alphabetically) Pranjal Awasthi, Maria-Florina Balcan, Thierry Bouwmans, Chen Dan, Travis Dick, Artur Dubrawski, Laurent El Ghaoui, Nika Haghtalab, Sajid Javed, Jiantao Jiao, Michael I. Jordan, Xin Li, Yi Li, Yingyu Liang, Zhouchen Lin, Kyle Miller, Wenlong Mou, Vasileios Nakos, Ricardo Otazo, Ruslan Salakhutdinov, Xiaofei Shi, Junru Shao, Aarti Singh, Zhao Song, Ruosong Wang, David P. Woodruff, Pengtao Xie, Eric P. Xing, Chao Xu, Susu Xu, Yichong Xu, Lin Yang, Shan You, Yaodong Yu, Chao Zhang, and Peilin Zhong. It has been really nice to work with you.

I would like to thank my friends and classmates at CMU. They are (ordered alphabetically) Ian Char, Chen Dan, Travis Dick, Simon Du, Hanzhang Hu, Juyong Kim, Mikhail Khodak, Leqi Liu, Yusha Liu, Yangyi Lu, Calvin McCarter, Adarsh Prasad, Dravyansh Sharma, Xiaofei Shi, Xiaoting Sun, Ellen Vitercik, Keyang Xu, Yichong Xu, Ruosong Wang, Yining Wang, Yu-Xiang Wang, Colin White, Yifan Wu, Yuexin Wu, Shuxin Yao, Fan Yang, Han Zhao, and Xun Zheng. Hope our friendship will last forever.

I also want to thank various professors/staffs in the Machine Leaning Department. They are (ordered alphabetically) Jian Ma, Barnabas Poczos, Pradeep Ravikumar, Aarti Singh, Alex Smola, Diane Stidle, Ryan Tibshirani, Larry Wasserman, and Eric P. Xing. Thank you for offering the greatest courses in the world from which I really learned a lot.

Finally, I would like to send my warmest acknowledgements to my family. Without their support (both spiritually and financially), I could not have completed my Ph.D. degree. Thank you.

The thesis is dedicated to all of you. Hope you can feel the same happiness as me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Why Sparse Learning, Deep Networks, and Adversarial Learning?

We are now in an era of big data as well as high dimensional data. Fortunately, high dimensional data are not unstructured. Usually, they lie near low dimensional manifolds. This is the basis of linear and nonlinear dimensionality reduction. As simple yet effective approximations, there are three typical assumptions on the structure of data:

- **Sparsity.** Sparsity is probably one of the most popular low-dimensional structures for vector-type data. That is, only a few entries of data vectors are non-zero. This assumption is the foundation of compressive sensing techniques.
- **Low rank.** Linear subspaces are widely used to model the data distribution. Because low dimensional subspaces correspond to low rank data matrices, the rank minimization problem, which models the real problem into an optimization problem by minimizing the rank in the objective function, is now widely used in machine learning and data recovery. Actually, rank is regarded as a sparsity measure for matrices. So low rank recovery problems are studied in parallel with the compressed sensing theories for sparse vector recovery. Typical rank minimization problems include matrix completion, which aims at completing the entire matrix from a small sample of its entries, and robust principal component analysis, which recovers the ground truth data from sparsely corrupted elements.
- **Low-dimensional manifolds.** The structure of low-dimensional manifolds has received significant attention in recent years as a more realistic assumption beyond the linearity of the data. Such a data structure is typically characterized by deep neural networks.

One of the commonalities among these data assumptions is that direct formulations of them lead to sparse learning, deep networks, or Adversarial Learning. For example, sparsity-induced data results in the ℓ_0-norm minimization problem, and low-rank-induced data assumption results in the rank minimization problem, all of which belong to sparse learning. The manifold-induced data assumption serves as the foundation for the recent popularity of deep neural networks and adversarial learning.

1.2 ROSE: Robustness, Optimization, and Sample Efficiency

Despite a large amount of work on these data assumptions, many fundamental questions remain unresolved:

From the statistical aspect, understanding the *tight* sample complexity of big-data problems under the above-mentioned assumptions is an important research question. Intuitively, the intrinsic dimensionality of data should be much smaller than the ambient dimension, and the *tight* sample complexity should be comparable to the intrinsic dimension, rather than the ambient dimension.

- **Contribution 1 (Sample efficiency).** We design principled, practical and scalable algorithms for big data problems with near-optimal sample complexity. These include models of matrix completion, robust PCA, margin-based active learning, property testing, compressed sensing, etc.

From the computational aspects, direct formulations of the above-mentioned approximate recovery problems are *non-convex* and might be NP-hard to optimize in general. Therefore, one of the long-standing questions is designing computationally efficient noise-tolerant learning algorithms that can approximate the unknown target parameters to any arbitrary accuracy. We address this problem in this thesis.

- **Contribution 2 (Optimization).** One of our focuses is to develop new paradigms toward global optimality of non-convex optimization in polynomial time. In particular, we design algorithms and provide an understanding of the landscape (e.g., duality gap) for the problems of (1-bit) compressed sensing, deep neural networks, GAN, matrix factorization, etc.

From the robustness aspects, models of non-convex learning such as deep neural networks, active learning, as well as low-rank models (matrix completion and PCA) might be vulnerable to adversarial examples. Although these problems have been widely studied empirically, much remains unknown concerning the theory of designing robust methods against adversarial corruptions.

- **Contribution 3 (Robustness).** We design new algorithms to improve the robustness of non-convex learning models. For example, we investigate the performance of active learning under adversarial noise model, and show that active learning works well under this challenging noise model. We also analyze the robustness of robust PCA to sparse adversarial corruption. For learning by deep neural networks, we identify a trade-off between robustness and accuracy that serves as a guiding principle in the design of defenses against adversarial examples.

1.3 Organization of This Thesis

This thesis makes the above three contributions and consists of the following three components.

- **Learning with sparsity.** Learning with sparsity involves solving non-convex optimization with ℓ_0 constraints or regularization. In Chapter 2, we will present works on how to efficiently approximate the unknown target vector $\mathbf{w}^*$ to arbitrary accuracy for the problem of (1-bit) compressed sensing [176] and active learning under Massart noise and adversarial noise models [17]. We apply the technique of solving a sequence of carefully-designed

convex surrogate problems. We are able to get arbitrarily close to the target vector, although the original problem itself is highly non-convex. We achieve exponential improvements in label complexity over passive learning approaches.

- **Learning with low-rank approximations.** Learning with low-rank approximation involves solving non-convex optimization with low-rank constraints or regularization. In Chapter 3, we will discuss the problems of matrix completion and robust PCA via strong duality [26, 27], and design efficient algorithms for property testing of matrix rank [25]. Our algorithms enjoy nearly optimal sample complexity.
- **Learning with deep neural networks.** Deep neural networks are more challenging non-convex problems as they involves non-linear activation functions beyond low-rank approximation. In Chapter 4, we will try to understand the landscape of deep neural networks [258] and GANs [257] via architecture designs. Our main results show that the multi-branch deep neural networks and GANs enjoy smaller duality gap. A smaller duality gap in relative value typically implies that the problem is less non-convex, and thus is easier to optimize. The results shed light on better understanding the power of over-parametrization where increasing the network width tends to make the loss surface less non-convex. We will also study how to improve the robustness of deep neural networks via new loss design [259].

Chapter 2

Learning with Sparsity

2.1 Learning of Halfspaces and One-Bit Compressed Sensing

2.1.1 Introduction

Linear models are a central object of study in machine learning, statistics, signal processing, and many other domains [69, 84, 133, 230, 231, 246, 256]. In machine learning and statistics, study of such models has led to significant advances in both the theory and practice of prediction and regression problems. In signal processing, linear models are used to recover sparse signals via a few linear measurements. This is known as compressed sensing or sparse recovery. In both cases, the problem can be stated as approximately recovering a vector $w^* \in \mathbb{R}^d$ given information about $w^* \cdot x_i$, where the x_i's are drawn from a distribution. The feedback typically comes in the form of the value of $w^* \cdot x_i$ or just the sign of the value. The focus of this work is on the latter setting known as classification or 1-bit compressed sensing in the respective communities. That is, given noisy 1-bit measurements of the form $\mathrm{sign}(w^* \cdot x_i)$, how to efficiently recover a vector w that is a good approximation to $w^* \in \mathbb{R}^d$, in terms of the value $\|w - w^*\|_2$. Furthermore, in the context of 1-bit compressed sensing, where w^* is t-sparse, we must use a number of measurements x_i's that scale polynomially in t and only polylogarithmically in d, the ambient dimension.

Despite a large amount of work on linear models, many fundamental questions remain unresolved. In learning theory, one of the long-standing questions is designing efficient noise-tolerant learning algorithms that can approximate the unknown target vector w^* to any arbitrary accuracy. Here noise corresponds to the corruption in the observations $\mathrm{sign}(w^* \cdot x_i)$. In the absence of noise, the recovery problem can be solved efficiently via linear programming. Several other algorithms such as Support Vector Machines [231], Perceptron [170] and Winnow [159] exist that provide better guarantees when the target vector has low L_2 or L_1 norm. This problem becomes more challenging in the context of 1-bit compressed sensing, as in addition to computational efficiency, one has to approximately recover w^* given a number of measurements $\mathrm{poly}(t, log(d))$. In the absence of noise, methods of this type are known only for Gaussian marginal distribution [97, 185, 186] or when the data has a large L_1 margin. However, this problem is left open for general distributions even in the absence of noise.

When measurements are noisy, this problem becomes more challenging in both its classification and 1-bit compressed sensing forms. This is due to the fact that direct formulations of

the approximate recovery problem are non-convex and are NP-hard to optimize [102]. There is significant evidence to indicate that without assumptions on the noise and the distribution of x_i, such recovery might not be computationally possible [71, 136]. When no assumptions are made on the nature of the noise (agnostic model), the best known result shows that when the distribution is uniform over the unit ball, one can achieve an $O(\nu) + \epsilon$ approximation, where ν is the fraction of the noisy labels [14]. An exciting work of [185] considers 1-bit compressed sensing under the challenging agnostic noise model and provides the best known result in approximately recovering a t-sparse w^* efficiently with a number of samples $\mathrm{poly}(t \log d)$, albeit with an approximation factor $(11\nu\sqrt{\log \frac{e}{\nu}} + \epsilon\sqrt{\log \frac{e}{\epsilon}})^{1/2}$ that does not match that of its non-sparse counterpart [14].

Due to the difficulty of the most general form of the problem, most positive results for obtaining arbitrarily good approximation have focused on the case of symmetric noise. A noise process is called *symmetric* if the probability that $\mathrm{sign}(w^* \cdot x_i)$ is corrupted only depends on the magnitude $|w^* \cdot x_i|$ [185]. Symmetric noise has many structural properties that one can exploit. For instance, when samples x_i's are generated from a symmetric distribution, it can be shown that the sign weighted average of the samples is enough to approximate w^*. This is the main insight behind some existing works on classification and 1-bit compressed sensing algorithms that are concerned with symmetric noise, such as [185, 206]. When 1-bit compressed sensing is considered, the more challenging aspect is to show that the number of samples scale linearly with the sparsity of w^*. Even when x_i's are not generated from a "nice" distribution, one can show that the weighted average is not far from w^*.[1] However, these observations and techniques break down when the noise is not symmetric.

2.1.2 Our results on label efficiency, optimization, and robustness

Our work tackles the problem of approximate recovery under highly asymmetric noise and advances the state-of-the-art results in multiple aspects. We first study a natural asymmetric noise model known as the *bounded noise (a.k.a Massart noise)* model. In this model, the probability of corrupting the $\mathrm{sign}(w^* \cdot x_i)$ is upper bounded by a constant $\frac{1}{2} - \frac{\beta}{2}$, i.e., an adversary flips the label of each point x_i with probability $\eta(x_i) \leq \frac{1}{2} - \frac{\beta}{2}$. This is a natural generalization of the well known *random classification noise model* of [133], where the probability of flipping the label of each example is $\eta = \frac{1}{2} - \frac{\beta}{2}$. Bounded noise model has been widely studied in statistical learning theory [42] in the context of achieving improved convergence rate. However, except for very simple classes with constant VC dimension, computationally efficient results in this space had remained unknown until recently.[2] We provide the first polynomial time algorithm for approximate recovery to arbitrary accuracy in this model for *any constant noise level* and a broad class of data distributions. Our work improves over that of [15] that required β to be very close to 1 (noise of order 10^{-7}). In this work, we introduce a novel algorithm that goes beyond this value of β and efficiently approximates linear separators to arbitrary accuracy ϵ for any constant value of

[1] This needs additional assumption on the nature of noise. The most widely studied among them is the *random classification* noise model where the sign of each observation is flipped i.i.d. with probability $\eta < \frac{1}{2}$. This can then be boosted in polynomial time to obtain a vector that is arbitrarily close [40].

[2] A variant of bounded noise, where the flipping probability for each point is either $\eta(x) = 0$ or $\eta(x) = \eta$ has been also considered as an important open problem in learning theory with the hope that understanding the complexities involved in this type of noise could shed light on the problem of learning disjunctions in the presence of noise.

$\beta > 0$ in time $\mathrm{poly}(d, \frac{1}{\epsilon})$, when the marginal distribution is isotropic log-concave in $\mathbb{R}^d$. We also introduce an attribute-efficient variant of this algorithm and perform 1-bit compressed sensing with number of samples scaling only polynomially in the sparsity parameter and polylogarithmic in the ambient dimension. This is the first such result demonstrating that efficient 1-bit compressed sensing to any desired level of accuracy is possible under highly asymmetric noise. Throughout this section, we assume $\|\mathbf{w}\|_2 = 1$. Below, we state our main theorems informally:

Theorems 2 and 3 (informal). *Let $x_1, x_2, \ldots x_m \in \mathbb{R}^d$ be generated i.i.d. from an isotropic log-concave distribution. Let $y_1, y_2, \ldots y_m$ be the corresponding labels generated as $\mathcal{N}_\beta(\mathrm{sign}(w^* \cdot x_i))$, where $\mathcal{N}_\beta$ is an arbitrary Massart noise process with a constant β. (a) There is an efficient algorithm that for any $\epsilon > 0$, runs in time polynomial in $m, d, \frac{1}{\epsilon}$, and with probability $1-\delta$ outputs a vector w such that $\|w - w^*\|_2 \leq \epsilon$, provided that $m \geq \mathrm{poly}(d, \frac{1}{\epsilon}, \log(\frac{1}{\delta}))$. (b) Furthermore, if w^* is t-sparse then the algorithm only needs $m \geq \mathrm{poly}(t, \log(d), \frac{1}{\epsilon})$.*

We also consider a more challenging noise model known as *adversarial (a.k.a. agnostic)* noise. Here, no assumptions are made about the nature of the noise and as a result, even information theoretically, approximate recovery within arbitrarily small error is not possible [132]. However, one can still recover w such that $\|w - w^*\|_2 \leq c\nu + \epsilon$, where $\epsilon > 0$ can be arbitrarily small and ν is the fraction of examples that are adversarially corrupted. One would like to keep c as small as possible, ideally a constant[3]. We provide a polynomial time algorithm that can approximately recover w^* in this model with $c = O(1)$ and the dependence on the number of samples is $O(\frac{t}{\epsilon^3}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$. Below, we state our main theorems informally:

Theorem 4 (informal). *Let $x_1, x_2, \ldots x_m \in \mathbb{R}^d$ be generated i.i.d. from an isotropic log-concave distribution. Let w^* be a t-sparse vector and $y_1, y_2, \ldots y_m$ be the measurements generated by $\mathcal{N}_{adversarial}(\mathrm{sign}(w^* \cdot x_i))$, where $\mathcal{N}_{adversarial}$ is the adversarial noise process that corrupts a ν fraction of the measurements. There is an efficient algorithm that for any $\epsilon > 0$, runs in time polynomial in $m, d, \frac{1}{\epsilon}$, and with probability $1-\delta$ outputs a vector w such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$, provided that $m = \Omega(\frac{t}{\epsilon^3}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$ or the number of actively labeled samples is $\Omega(\frac{t}{\epsilon^2}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$.*

1-bit compressed sensing under adversarial noise is also considered under a stronger requirement of *uniformity*, where the approximate recovery guarantee is required to hold with high probability over all sparse signals w^* and all possible corruption of ν fraction of the samples. In other words, in the non-uniform case (Theorem 4) an unknown sparse target vector w^* and a noisy distribution $\tilde{D}$ are fixed in advance before the samples (x_i, y_i) are drawn from $\tilde{D}$, while in the uniform case, the adversary first observes x_i's and then chooses a w^* and noisy labels y_i's. In the uniform case, one typically needs more samples to achieve the same accuracy as in the non-uniform case. In this work, when uniformity is considered our algorithm returns w such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$ when the number of samples is $O(\frac{t}{\epsilon^4}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$.

Theorem 5 (informal). *Let $x_1, x_2, \ldots x_m \in \mathbb{R}^d$ be generated i.i.d. from an isotropic log-concave distribution. With probability $1-\delta$ the following holds. For any signal w^* such that $\|w^*\|_0 \leq t$ and measurements $y_1, y_2, \ldots y_m$ generated by $\mathcal{N}_{adversarial}(\mathrm{sign}(w^* \cdot x_i))$, where $\mathcal{N}_{adversarial}$ is the adversarial noise process that corrupts a ν fraction of the measurements, there is an efficient*

[3]This is the information theoretic limit.

algorithm that for any ϵ, such that $\nu \in O(\epsilon/\log(d/\epsilon)^2)$, runs in time polynomial in $m, d, \frac{1}{\epsilon}$ and outputs a vector w such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$, provided that $m = \Omega(\frac{t}{\epsilon^4}\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$.

Our work on 1-bit compressed sensing provides the first result in non-uniform 1-bit compressed sensing under adversarial noise. Under the uniform case when ν is small, we considerably improve the best known approximation results of [185] from $\|w - w^*\|_2 \leq (11\nu\sqrt{\log\frac{e}{\nu}} + \epsilon\sqrt{\log\frac{e}{\epsilon}})^{1/2}$ to $\|w - w^*\|_2 \leq O(\nu) + \epsilon$. Furthermore, we improve the dependence of the sample complexity on ϵ from $\frac{1}{\epsilon^6}$ in the case of the results of [185] to $\frac{1}{\epsilon^4}$. While prior work on 1-bit compressed sensing only handles the special case when the distribution is Gaussian, our results hold when the distribution of x_i is any isotropic log-concave distribution.

Hardness. We now study the hardness of the above-mentioned problems. We show that one-shot minimization does not work for a large family of loss functions that include any continuous loss with a natural property that points at the same distance from the separator have the same loss. This generalizes the result of [15] who showed that one-shot minimization of hinge loss does not lead to an arbitrarily small 0/1 error even under bounded noise with small flipping probability, and justifies why minimizing a sequence of carefully designed losses, as we will do in Section 2.1.4, is indispensable to achieving an arbitrarily small excess error.

Without loss of generality, we discuss the lower bound in $\mathbb{R}^2$. Formally, let $\mathcal{P}_\beta$ be the class of noisy distribution $\tilde{D}$ with uniform marginal over the unit ball, and let (z_w, φ_w) represent the polar coordinate of a point P in the instance space, where φ_w represents the angle between the linear separator h_w and the vector from origin to P, and z_w is the L_2 distance of the point P and the origin. Let $\ell_+^w(z_w, \varphi_w)$ and $\ell_-^w(z_w, \varphi_w)$ denote the loss functions on point P with correct and incorrect classification by h_w, respectively. The loss functions we study here satisfy the following properties.

Definition 1. *Continuous loss functions $\ell_+^w(z_w, \varphi_w)$ and $\ell_-^w(z_w, \varphi_w)$ are called* proper*, if and only if*

1. *$\ell_+^w(z_w, \varphi_w) = \ell_+^w(z_w, k\pi \pm \varphi_w)$ and $\ell_-^w(z_w, \varphi_w) = \ell_-^w(z_w, k\pi \pm \varphi_w)$, for $k \in N$;*
2. *For $z_w > 0$, $\ell_-^w(z_w, \varphi_w) \geq \ell_+^w(z_w, \varphi_w)$; The equality holds if and only if $\varphi_w = k\pi$, $\forall k \in N$.*

Note that all losses that are functions of the distance to the classifier, e.g. the hinge-loss and logistic loss, etc., satisfy Property 1, since the distance of a point to classifier w is $|z_w \sin\varphi_w| = |z_w \sin(k\pi \pm \varphi_w)|$. However, Property 1 only requires the symmetry of the loss w.r.t. the linear separator, and is not limited to distance-based losses, that is, the losses on the points with the same distance can be different. Moreover, this property does not require the loss to be monotonically increasing in the distance. Property 2 is a very natural assumption since to achieve low error, it is desirable to penalize misclassification more. Note that we equally penalize correct and incorrect classifications if and only if points lie exactly on the linear separator.

In fact, most of the commonly used loss functions [32] satisfy our two properties in Definition 1, e.g., the (normalized) hinge loss, logistic loss, square loss, exponential loss, and truncated quadratic loss, because they are all functions of the distance to classifier. Furthermore, we highlight that Definition 1 covers the loss even with regularized term on w. A concrete example is 1-bit compressed sensing, with loss function formulated as $\ell_+(z_w, \varphi_w) = -|z_w \sin\varphi_w| + \lambda_1\|w\|_1 + \lambda_2\|w\|_2$ and $\ell_-(z_w, \varphi_w) = |z_w \sin\varphi_w| + \lambda_1\|w\|_1 + \lambda_2\|w\|_2$. Thus our lower bound demonstrates that one-shot 1-bit compressed sensing cannot always achieve arbitrarily small excess error under the Massart noise.

Our lower bound for any proper function is stated as follows.

Theorem 1. *For every bounded noise parameter $0 \leq \beta < 1$, there exists a distribution $\tilde{D}_\beta \in \mathcal{P}_\beta$ (that is, a distribution over $\mathbb{R}^2 \times \{+1, -1\}$, where the marginal distribution on $\mathbb{R}^2$ is uniform over the unit ball, and the labels $\{+1, -1\}$ satisfies bounded noise condition with parameter β) such that any proper loss minimization is not consistent on $\tilde{D}_\beta$ w.r.t. the class of halfspaces. That is, there exists an $\epsilon \geq 0$ and a sample size $m(\epsilon)$ such that any proper loss minimization will output a classifier of excess error larger than ϵ by a high probability over sample size at least $m(\epsilon)$.*

2.1.3 Our techniques

In this section, we discuss the techniques used for achieving our results.

Iterative polynomial regression: Our algorithm follows a localization technique inspired by the work of [22] and building on [14, 16]. Our algorithm is initialized by a classifier w_0 with a $0/1$ error that is at most an appropriate small constant more than the error of w^* w.r.t. the observed labels. This difference is known as the *excess error*. The algorithm then proceeds in rounds, aiming to cut down the excess error by half in each round. By the properties of bounded noise and the log-concave distribution, excess error of a classifier is a linear function of its angle to w^*. Therefore, our algorithm aims to cut the angle by half at each round and eventually will output a w that is close to w^*.

Consider w_{k-1} with angle $\leq \alpha_k$ to w^*. It can be shown that for a band of width $\gamma_{k-1} = \Theta(\alpha_k)$ around the separator w_{k-1}, w_{k-1} makes most of its error in this band. Therefore, improving the accuracy of w_{k-1} in the band significantly improves the accuracy of w_{k-1} overall. When considering vectors that are at angle $\leq \alpha_k$ to w_{k-1}, it can be shown that any vector w_k that achieves *a small enough constant excess error with respect to the distribution in the band*, indeed, enjoys a much stronger guarantee of having *excess error that is half of w_{k-1} overall.* Therefore, if such a vector w_k can be found efficiently in the presence of bounded noise, a classifier of excess error ϵ can be learned in $O(\log(\frac{1}{\epsilon}))$ steps. In order to make the above method work we need to achieve two goals: a) achieve a constant excess error while tolerating noise rate of $\frac{1}{2} - \frac{\beta}{2}$ and b) the hypothesis output should be a halfspace.

On one hand, efficient proper learning methods, such as surrogate loss minimization in the band, readily achieve goal (b). However, convex surrogate loss functions are only a good approximation of the 0/1 loss when the noise is small enough. Since the noise in the band can be as high as $\frac{1}{2} - \frac{\beta}{2}$, this directly restricts the noise rate of bounded noise that can be tolerated with such methods. Indeed, [15] demonstrated that when hinge-loss minimization is used in the band, such a method only works if the probability of flipping the label is as small as $\approx 10^{-6}$, i.e., when β is very close to 1. On the other hand, the polynomial regression approach of [129] learns linear separators to an arbitrary excess error of ϵ with runtime $\mathrm{poly}(d, \exp(\mathrm{poly}(\frac{1}{\epsilon}))$ when the marginal distribution is log-concave, requiring no additional assumption on noise. Since the distribution in the band is also log-concave, this method can achieve *an arbitrarily small constant* excess error in the band thereby achieving goal (a). However, this algorithm outputs the sign of a polynomial $p(\cdot)$ as a hypothesis, which is not necessarily a halfspace.

Instead, our algorithm takes a novel two-step approach to find w_k for *any amount of noise*. This is done by first finding a polynomial p_k that has a small constant excess error in the band. To

obtain such a polynomial, we choose $\mathrm{poly}(d, \log(\frac{\log(1/\epsilon)}{\delta}))$ labeled samples from the distribution in the band and use the algorithm by [128] to find a polynomial with a small enough but, importantly, *a constant excess error*, e_{KKMS}, in the band. Note that at this point p_k already satisfies goal (a) but it does not satisfy goal (b) as it is not a halfspace. At a high level, since p_k has a small excess error with respect to w^* in the band, using a structural property of bounded noise that connects the excess error and disagreement of a classifier with respect to w^*, we can show that p_k is also close in classification to w^*. Therefore, it suffices to agnostically learn a halfspace w_k to a constant error for samples in the band that are labeled based on $\mathrm{sign}(p(\cdot))$. To achieve this, we use localized hinge loss minimization in the band over a set of samples that are labeled based on predictions of p_k to find w_k. Therefore, w_k is close in classification to p_k in the band, which is in turn close to w^* in the band. As a result, w_k also has a small error in the band as desired [4].

1-bit compressed sensing: Notice that the techniques mentioned above involve minimizing a convex loss function over a suitably chosen convex set, i.e., the band. When the target vector is sparse, we show that it is enough to perform the minimization task over the set of separators (or polynomials) of small L_1 norm. Since we focus on a smaller candidate set than that of the general case, we can hope to achieve tighter concentration bounds and thus obtain better sample complexity.

Specifically, in the case of Massart noise we extend the polynomial regression algorithm of [128] to the sparse case by adding L_1 constraint for polynomials. The target polynomial can then be found using L_1 regression over the convex set of low degree polynomials with small L_1 norm. To prove the correctness of this algorithm, we show that when w^* is sparse, there exists a low degree polynomial of small L_1 norm that approximates w^*. This is due to the fact that the target polynomial can be represented by a linear combination of sparse Hermite polynomials. To derive the sample complexity, we use a concentration result of [261] on the covering number of linear functions of L_1-constrained vectors that satisfy a certain margin property. We analyze such margin property by extending the random thresholding argument of [128]. The sample complexity of our method follows by combining the two techniques together.

For non-uniform 1-bit compressed in presence of adversarial noise, we build on the algorithm of [14] for learning halfspaces. Similarly as in the previous procedure, this algorithm relies on hinge loss minimization in the band for computing a halfspace of a constant error. However, this algorithm does not use the polynomial regression as an intermediate step, rather, it directly minimizes the hinge loss on a set of points drawn from the noisy distribution in the band. To make this algorithm attribute-efficient, we constrain the hinge loss minimization step to the set of vectors with L_1 norm of at most $\sqrt{t}$. The challenge here is to derive the sample complexity under L_1 constraint. To do this, we use tools from Rademacher theory that exploit the L_1 bound of the linear separators. The improved sample complexity follows from stronger upper bounds on the L_∞ norm of x_i's and the value of hinge loss.

In the uniform case, we build on the techniques described above and show that for a larger number of samples, the analysis would hold *uniformly over all possible noisy measurements on the samples obtained from a choice of sparse w^* and any ν fraction of points corrupted.* First,

[4]The recent work of [72] also combines the margin-based approach with polynomial regression. However, in [72] polynomial regression is only used once in the end as opposed to the iterative application of polynomial regression used in this work.

we show that when the number of samples is $m = \Omega(\frac{t}{\epsilon^4}\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$, then *every band* that could be considered by the algorithm has a sufficient number of samples. This can be proved using covering number and uniform convergence bounds for a class of bands around halfspaces whose L_1 norm is bounded by $\sqrt{t}$. Next, we show that at round k, the empirical hinge loss is concentrated around its expectation uniformly over all choices of w^*, w_{k-1}, and a ν fraction of the samples whose labels differ from the labels of w^*. We note that w_{k-1} is uniquely determined by the labeled samples used by the algorithm in the previous rounds. Therefore, by arguing about the number of possible labelings that can be produced by a sparse w^* and adversarial noise only on the samples that have been used by the algorithm, we can derive a concentration bound that holds uniformly.

2.1.4 Our algorithms

In this section, we introduce efficient algorithms for recovering the true classifier in the presence of bounded noise for any constant β. We first consider the non-sparse case and show how our algorithm can return a classifier that is arbitrarily close to w^*. Building on this, we introduce an attribute-efficient variation of this algorithm that is applicable to 1-bit compressed sensing and recovers a t-sparse w^* from few measurements.

Algorithm for the general case

Here, we describe an efficient algorithm for learning in the presence of bounded noise for any constant β. At a high level, our algorithm proceeds in $\log(\frac{1}{\epsilon})$ rounds and returns a linear separator w_k at round k whose disagreement with respect to w^* is halved at every step. By induction, consider w_{k-1} whose disagreement with w^* is at most $\Pr[\text{sign}(w^* \cdot x) \neq \text{sign}(w_{k-1} \cdot x)] \leq \frac{\alpha_k}{\pi}$. First, we draw samples from the distribution of points that are at distance at most γ_{k-1} to w_{k-1}. We call this region *the band* at round k and indicate it by $S_{w_{k-1},\gamma_{k-1}}$. Next we apply the polynomial regression algorithm of [128] to get a polynomial $p(\cdot)$ of error a constant e_{KKMS} in the band. We draw additional samples from the band, label them based on $\text{sign}(p(\cdot))$, and minimize hinge loss with respect to these labels to get w_k. We then show that w_k that is obtained using this procedure has disagreement at most $\frac{\alpha_{k+1}}{\pi}$ with the target classifier. We can then use w_k as the classifier for the next iteration. The detailed procedure is presented in Algorithm 1. The main result of this section is that Algorithm 1 efficiently learns halfspaces under log-concave distributions in the presence of bounded noise for any constant parameter β that is independent of the dimension. The small excess error implies arbitrarily small approximation rate to the optimal classifier w^* under bounded noise model.

Theorem 2. *Let the optimal Bayes classifier be a halfspace denoted by w^*. Assume that the bounded noise condition holds for some constant $\beta \in (0, 1]$. For any $\epsilon > 0$, $\delta > 0$, there exist absolute constants e_0, C, C_1, C_2, c_1, c_2 such that Algorithm 1 with parameters $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = Cr_k$, $\lambda = \frac{3C_1}{8CC_2}$, $e_{KKMS} = \beta(\lambda/(4c_1 + 4c_2 + 2))^4$, and $\tau_k = \lambda\, \gamma_{k-1}/(4c_1 + 4c_2 + 2)$ runs in polynomial time, proceeds in $s = O(\log \frac{1}{\epsilon})$ rounds, where in round k it takes $n_k = \text{poly}(d, \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples and $m_k = \text{poly}(d, \log(s/\delta))$ labels and with probability $1 - \delta$ returns a vector $w \in \mathbb{R}^d$ such that $\|w - w^*\|_2 \leq \epsilon$.*

For the remainder of this section, we denote by $\tilde{D}$ the noisy distribution and by D the distribution with labels corrected according to w^*. Furthermore, we refer to $\tilde{D}_{w_{k-1},\gamma_{k-1}}$ and $D_{w_{k-1}}, \gamma_{k-1}$, the noisy and clean distributions in the band, by $\tilde{D}_k$ and D_k, respectively.

Algorithm 1 LEARNING HALFSPACES UNDER ARBITRARILY BOUNDED NOISE

Input: An initial classifier w_0, a sequence of values γ_k, τ_k and r_k for $k = 1, \ldots, \log(1/\epsilon)$. An error value e_{KKMS}.

1. Let w_0 be the initial classifier.
2. For $k = 1, \ldots, \log(1/\epsilon) = s$.
   (a) Take $\text{poly}(d, \log(\frac{s}{\delta}))$ labeled samples from $\tilde{D}_k$, the conditional distribution within the band $\{x : |w_{k-1} \cdot x| \leq \gamma_{k-1}\}$, and place them in the set T. Run the polynomial regression algorithm [128] over T to find a polynomial p_k such that $\text{err}_{\tilde{D}_k}(\text{sign}(p_k)) \leq \text{err}_{\tilde{D}_k}(h_{w^*}) + e_{\text{KKMS}}$.
   (b) Take $d(d+\log(k/\delta))$ unlabeled samples from $\tilde{D}_k$ and label them according to $\text{sign}(p_k(\cdot))$. Call this set of labeled samples T'.
   (c) Find $v_k \in B(w_{k-1}, r_{k-1})$ that approximately minimizes the empirical hinge loss over T' using threshold τ_k, i.e., $L_{\tau_k}(v_k, T') \leq \min_{w \in B(w_{k-1}, r_{k-1})} L_{\tau_k}(w, T') + \frac{\lambda}{12}$.
   (d) Let $w_k = \frac{v_k}{\|v_k\|_2}$.

Output: Return w_s, which has excess error ϵ with probability $1 - \delta$.

1-bit compressed sensing in presence of bounded noise

We consider the true classifier w^* to be t-sparse and build upon our previous Algorithm 1 to return a vector w such that $\|w - w^*\|_2 \leq \epsilon$, given a number of samples $m \geq \text{poly}(t, \frac{\log(d)}{\epsilon})$. Our algorithm is in Algorithm 2. Our main result is the following:

Theorem 3 (Bounded Noise)**.** *Let the optimal Bayes classifier be a halfspace denoted by w^* such that $\|w^*\|_0 = t$. Assume that the bounded noise condition holds for some constant $\beta \in (0, 1]$. For any $\epsilon > 0$, $\delta > 0$, there exist absolute constants e_0, C, C_1, C_2, c_1, c_2 such that Algorithm 2 with parameters $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = C r_k$, $\lambda = \frac{3C_1}{8CC_2}$, $e_{KKMS} = \beta(\lambda/(4c_1 + 4c_2 + 2))^4$, and $\tau_k = \lambda\, \gamma_{k-1}/(4c_1 + 4c_2 + 2)$ runs in polynomial time, proceeds in $s = O(\log \frac{1}{\epsilon})$ rounds, where in round k it takes $n_k = \text{poly}(t \log(d), \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples and $m_k = \text{poly}(t, \log(sd/\delta), \exp(k))$ labels and with probability $1 - \delta$ returns a vector $w \in \mathbb{R}^d$ such that $\|w - w^*\|_2 \leq \epsilon$.*

1-bit compressed sensing in presence of adversarial noise

In this section, we first consider 1-bit compressed sensing of linear separators under adversarial noise. In this noise model, the adversary can choose any distribution $\tilde{D}$ over $\mathbb{R}^d \times \{+1, -1\}$ such that the marginal over $\mathbb{R}^d$ is unchanged but a ν fraction of the labels are flipped adversarially. We introduce an attribute-efficient variant of the algorithm of [14] for noise-tolerant learning that given $O(t\, \text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})/\epsilon^3)$ samples from a given $\tilde{D}$ distribution, with probability $1 - \delta$

Algorithm 2 LEARNING SPARSE HALFSPACES UNDER ARBITRARILY BOUNDED NOISE

Input: An initial classifier w_0, a sequence of values γ_k, τ_k and r_k for $k = 1, \ldots, \log(1/\epsilon)$. An error value e_{KKMS}.

1. Let w_0 be the initial classifier.
2. For $k = 1, \ldots, \log(1/\epsilon) = s$.
   (a) Take $\text{poly}(\frac{t}{\gamma_k}, \log(\frac{ds}{\delta}))$ labeled samples from $\tilde{D}_k$, the conditional distribution within the band $\{x : |w_{k-1} \cdot x| \leq \gamma_{k-1}\}$, and place them in the set T. Run the polynomial regression algorithm [128] over T to find a polynomial p_k such that $\text{err}_{\tilde{D}_k}(\text{sign}(p_k)) \leq \text{err}_{\tilde{D}_k}(h_{w^*}) + e_{\text{KKMS}}$ and $\|p\|_1 = O((\frac{t}{\epsilon})^{\text{poly}(1/e_{\text{KKMS}})})$.
   (b) Take $m_k = \Omega(\frac{t}{\tau_k^2}\text{polylog}(d, \frac{1}{\delta}, \frac{1}{\epsilon}))$ unlabeled samples from $\tilde{D}_k$ and label them according to $\text{sign}(p_k(\cdot))$. Call this set of labeled samples T'.
   (c) Find $v_k \in B(w_{k-1}, r_{k-1})$ such that $\|v_k\|_1 \leq \sqrt{t}$ and v_k approximately minimizes the empirical hinge loss over T' using threshold τ_k, i.e., $L_{\tau_k}(v_k, T') \leq \min_{w \in B(w_{k-1}, r_{k-1}) \text{ and } ||w||_1 \leq \sqrt{t}} L_{\tau_k}(w, T') + \frac{\lambda}{12}$.
   (d) Let $w_k = \frac{v_k}{\|v_k\|_2}$.

Output: Return w_s, which has excess error ϵ with probability $1 - \delta$.

returns a vector w, such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$. To the best of our knowledge this is the first result in non-uniform 1-bit compressed sensing under adversarial noise. Furthermore, the approximation factor of this result almost matches the information theoretic bound.

Theorem 4 (Adversarial Noise – Non-uniform)**.** *Assume that the noise is adversarial and let the optimal linear classifier be a halfspace denoted by w^* such that $\|w^*\|_0 = t$. Let $\nu > 0$ be the error of w^*. For any $\epsilon > 0$, $\delta > 0$, there exist absolute constants $e_0, C, C_1, C_2, c_1, c_2$ such that Algorithm 3 with parameters $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = Cr_k$, $\lambda = \frac{3C_1}{8CC_2}$, and $\tau_k = \lambda\, \gamma_{k-1}/(4c_1 + 4c_2 + 2)$ runs in polynomial time, proceeds in $s = \log \frac{1}{\epsilon}$ rounds, where in round k it takes $n_k = \text{poly}(t, \log(d), \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples and $m_k = O(t\ \text{polylog}(sd/\delta)2^{2k})$ labels and with probability $1 - \delta$ returns a vector $w \in \mathbb{R}^d$ such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$. That is the total number of unlabeled samples is $m = O(\frac{t}{\epsilon^3}\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$ and at every round $m_k \leq O(\frac{t}{\epsilon^2}\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$ labels are requested.*

We also consider the stronger requirements of *uniform* 1-bit compressed sensing. In this setting, we show that given $O(t\ \text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})/\epsilon^4)$ samples x_i, with probability $1 - \delta$, *uniformly over all possible noisy measurements on x_i's obtained from a choice of sparse w^* and any ν fraction of measurements corrupted*, the algorithm returns a vector w such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$, when ν is small with respect to ϵ. When ν is small, this result considerably improves the best known approximation results of [185] from $\|w - w^*\|_2 \leq (11\nu\sqrt{\log \frac{e}{\nu}} + \epsilon\sqrt{\log \frac{e}{\epsilon}})^{1/2}$ to $\|w - w^*\|_2 \leq O(\nu) + \epsilon$. Furthermore, we improve the dependence of the sample complexity on ϵ from $\frac{1}{\epsilon^6}$ in the case of the results of [185] to $\frac{1}{\epsilon^4}$[5]. Our result for this setting is as follows.

[5]The sample complexity of the method of [185] is expressed as $O(t\ \text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})/\epsilon^6)$ for achieving error $\left(11\nu\sqrt{\log \frac{e}{\nu}} + \epsilon\sqrt{\log \frac{e}{\epsilon}}\right)^{1/2}$. When ν is small compared to ϵ, this in fact compares to a method with sample complexity $\frac{1}{\epsilon^{12}}$ for achieving excess error $\tilde{O}(\epsilon)$ which is the regime we work in.

Algorithm 3 NON-UNIFORM 1-BIT COMPRESSED SENSING UNDER ADVERSARIAL NOISE

Input: An initial classifier w_0, a sequence of values γ_k, τ_k and r_k for $k = 1, \ldots, \log(1/\epsilon)$.

1. Let w_0 be the initial classifier.
2. For $k = 1, \ldots, \log(1/\epsilon) = s$.
   (a) Take $m_k = \Omega(\frac{t}{\epsilon^2}\text{polylog}(d, \frac{1}{\delta}, \frac{1}{\epsilon}))$ samples from $\tilde{D}_k$ and request the labels. Call this set T'.
   (b) Find $v_k \in B(w_{k-1}, r_{k-1})$ such that $\|v_k\|_1 \leq \sqrt{t}$ and v_k approximately minimizes the empirical hinge loss over T' using threshold τ_k, that is, $L_{\tau_k}(v_k, T') \leq \min_{w \in B(w_{k-1}, r_{k-1}) \text{ and } ||w||_1 \leq \sqrt{t}} L_{\tau_k}(w, T') + \frac{\lambda}{12}$.
   (c) Let $w_k = \frac{v_k}{\|v_k\|_2}$.

Output: Return w_s, which has excess error $O(\nu) + \epsilon$ with probability $1 - \delta$.

Theorem 5 (Adversarial Noise – uniform)**.** *Let $x_1, x_2, \ldots x_m \in \mathbb{R}^d$ be drawn i.i.d. from an isotropic log-concave distribution. With probability $1 - \delta$ the following holds. For all signals w^* such that $\|w^*\|_0 \leq t$ and measurements $y_1, y_2, \ldots y_m$ generated by $\mathcal{N}_{adversarial}(\text{sign}(w^* \cdot x_i))$, where $\mathcal{N}_{adversarial}$ is the adversarial noise process that corrupts a ν fraction of the measurements, and for any ϵ such that $\nu \in O(\epsilon/\log(d/\epsilon)^2)$, there exist absolute constants e_0, C, C_1, C_2, c_1, c_2 such that Algorithm 4 with parameters $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = C r_k$, $\lambda = \frac{3C_1}{8CC_2}$, and $\tau_k = \lambda\, \gamma_{k-1}/(4c_1 + 4c_2 + 2)$ runs in time $poly(d, \frac{1}{\epsilon})$ and returns a vector $w \in \mathbb{R}^d$ such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$ if $m = \Omega(\frac{t}{\epsilon^4}\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})$. Furthermore, the number of labeled samples at every round k is $m_k \leq O(\frac{t}{\epsilon^3}\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})$.*

Algorithm 4 UNIFORM 1-BIT COMPRESSED SENSING UNDER ADVERSARIAL NOISE

Input: An initial classifier w_0, a sequence of values γ_k, τ_k and r_k for $k = 1, \ldots, \log(1/\epsilon)$.

1. Let w_0 be the initial classifier.
2. For $k = 1, \ldots, \log(1/\epsilon) = s$.
3. Take $m = O(t\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})/\epsilon^4)$ unlabeled samples from $\tilde{D}$, in set S.
   (a) Take $m_k = O(t\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})/\epsilon^2)$ of the samples in $S \cap S_k$ request the labels. Call this set of labeled samples T'.
   (b) Find $v_k \in B(w_{k-1}, r_{k-1})$ such that $\|v_k\|_1 \leq \sqrt{t}$ and v_k approximately minimizes the empirical hinge loss over T' using threshold τ_k, i.e., $L_{\tau_k}(v_k, T') \leq \min_{w \in B(w_{k-1}, r_{k-1}) \text{ and } ||w||_1 \leq \sqrt{t}} L_{\tau_k}(w, T') + \frac{\lambda}{12}$.
   (c) Let $w_k = \frac{v_k}{\|v_k\|_2}$.

Output: Return w_s, which has excess error $O(\nu) + \epsilon$ with probability $1 - \delta$.

We build on the algorithm of [14] for learning halfspaces under adversarial noise. Much like our procedure for bounded noise, this algorithm also relies on hinge loss minimization in the band for computing a halfspace of a constant error. However, this algorithm does not make use of polynomial regression as an intermediate step, rather, it directly minimizes the hinge loss on a labeled set of points drawn from the noisy distribution in the band, $\tilde{D}_k$.

To make this algorithm attribute-efficient, we constrain the hinge loss minimization step to the set of vectors with L_1 norm of at most $\sqrt{t}$. See Algorithm 4. Since w^* is t-sparse, $\|w^*\|_1 \leq \sqrt{t}$ and therefore the comparison between the outcome of every step and w^* remains valid. This shows that such a change preserves the correctness of the algorithm. We prove the correctness of this algorithm and its sample complexity in Section 2.1.5.

2.1.5 Proofs of our main results

Preliminaries

We use X to denote the domain of the samples and Y to denote the label set. In this work X is $\mathbb{R}^d$ and Y is the set $\{+1, -1\}$. We define the *sign* function as $\text{sign}(x) = 1$ if $x \geq 0$ and -1 otherwise. The problem of interest in this section is approximate recovery: *Given $\epsilon > 0$ and m i.i.d. samples $x_1, x_2, \ldots x_m$ drawn from a distribution over $\mathbb{R}^d$, and labeled as $y_i = \mathcal{N}(\text{sign}(w^* \cdot x_i))$, design a polynomial time algorithm to recover a vector w such that $\|w - w^*\|_2 \leq \epsilon$. Furthermore, if $\|w^*\|_0 = t$, we require m to grow as $poly(t, \log(d), \frac{1}{\epsilon})$.* Here $\mathcal{N}$ is a noise process that corrupts the measurements/labels. We study two asymmetric noise models in this work. The first is $\mathcal{N}_\beta$, the *bounded (a.k.a Massart) noise* model. A joint distribution over (X, Y) satisfies the bounded noise condition with parameter $\beta > 0$, if

$$|\Pr(Y = +1|x) - \Pr(Y = -1|x)| \geq \beta, \forall x \in X.$$

In other words, bounded noise is equivalent to the setting where an adversary constructs the distribution by flipping the label of each point x from $\text{sign}(w^* \cdot x)$ to $-\text{sign}(w^* \cdot x)$ with a probability $\eta(x) \leq \frac{1-\beta}{2}$. As is customary, we will use *Bayes optimal classifier* to refer to w^*, the vector generating the uncorrupted measurements. The other noise model that we study is $\mathcal{N}_{adversarial}$, the *adversarial* noise model. Here the adversary can corrupt the labels in any fashion. In this model, the goal of approximate recovery will be to get a vector w such that $\|w - w^*\|_2 \leq O(\nu) + \epsilon$, where ν is the fraction of examples corrupted by the adversary.

For any halfspace w, we denote the resulting classifier $h_w = \text{sign}(w \cdot x)$. For any classifier $h : X \mapsto \{+1, -1\}$, we define the error w.r.t. distribution $\mathcal{P}$ as $\text{err}_{\mathcal{P}}(h) = \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq y]$. We define the *excess* error of h as $\text{err}_{\mathcal{P}}(h) - \text{err}_{\mathcal{P}}(h_{w^*})$. We use OPT to denote the error of the Bayes classifier, i.e., $\text{err}_{\mathcal{P}}(h_{w^*})$. When the distribution is clear from the context, we use $\text{err}(h_{w^*})$ instead of $\text{err}_{\mathcal{P}}(h_{w^*})$. The next lemma demonstrates an important relation between the excess error of a classifier h and its "closeness" to w^* in terms of classification (or its disagreement).

Lemma 1. *Given a classifier $h : X \mapsto \{+1, -1\}$ and distribution $\mathcal{P}$ satisfying bounded noise condition with parameter β, let w^* be the Bayes optimal classifier. Then we have*

$$\beta \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq h_{w^*}(x)] \leq \text{err}_{\mathcal{P}}(h) - \text{err}_{\mathcal{P}}(h_{w^*}) \leq \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq h_{w^*}(x)]. \tag{2.1}$$

Proof. Here, we prove that the following equation holds for distribution P with Massart noise parameter $\beta > 0$.

$$\beta \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq h_{w^*}(x)] \leq \text{err}_{\mathcal{P}}(h) - \text{err}_{\mathcal{P}}(h_{w^*}) \leq \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq h_{w^*}(x)].$$

The right hand side inequality holds by the following.

$$\mathrm{err}_P(h) \leq \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq h_{w^*}(x)] + \Pr_{(x,y)\sim\mathcal{P}}[h_{w^*}(x) \neq y] = \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq h_{w^*}(x)] + \mathrm{err}_{\mathcal{P}}(h_{w^*}).$$

Let $A = \{x : h(x) \neq h_{w^*}(x)\}$ be the region where h and h_{w^*} disagree in their predictions. Note that $\Pr(A) = \Pr_{(x,y)\sim\mathcal{P}}[h(x) \neq h_{w^*}(x)]$. Then,

$$\mathrm{err}_{\mathcal{P}}(h) - \mathrm{err}_{\mathcal{P}}(h_{w^*}) = \Pr(A)[\mathrm{err}_{\mathcal{P}}(h|A) - \mathrm{err}_{\mathcal{P}}(h_{w^*}|A)] + \Pr(\bar{A})[\mathrm{err}_{\mathcal{P}}(h|\bar{A}) - \mathrm{err}_{\mathcal{P}}(h_{w^*}|\bar{A})].$$

Classifiers h and h_{w^*} agree over the set $\bar{A}$, i.e., either both make mistakes or neither does, simultaneously. Hence the second term is zero. On the other hand, the two classifiers disagree over A, so exactly one of them is making an incorrect prediction. Hence, $\mathrm{err}_{\mathcal{P}}(h|A)+\mathrm{err}_{\mathcal{P}}(h_{w^*}|A) = 1$. We have

$$\mathrm{err}_{\mathcal{P}}(h) - \mathrm{err}_{\mathcal{P}}(h_{w^*}) = \Pr(A)[1 - 2\mathrm{err}_{\mathcal{P}}(h_{w^*}|A)].$$

Since the labels are each flipped with probability at most $\frac{1-\beta}{2}$, we have that $\mathrm{err}_{\mathcal{P}}(h_{w^*}|A) \leq \frac{1-\beta}{2}$. Re-arranging the above inequality proves the claim. □

We frequently examine the region within a specified margin of a given halfspace. For distribution $\mathcal{P}$, halfspace w, and margin γ, we denote by $\mathcal{P}_{w,\gamma}$ the conditional distribution over the set $S_{w,\gamma} = \{x : |w \cdot x| \leq \gamma\}$. We define the *$\tau$-hinge loss* of a halfspace w over an example-label pair (x, y) as $\ell_\tau(w, x, y) = \max\left(0, 1 - \frac{y(w\cdot x)}{\tau}\right)$. When τ is clear from the context, we simply refer to the above quantity as the hinge loss. For a given set T of examples, we use $L_\tau(w, T)$ to denote the empirical hinge loss over the set, i.e., $L_\tau(w, T) = \frac{1}{|T|}\sum_{(x,y)\in T} \ell_\tau(w, x, y)$. For a classifier $w \in \mathbb{R}^d$ and a value r, we use $B(w, r)$ to denote the set $\{v \in \mathbb{R}^d : \|w - v\|_2 \leq r\}$. Moreover, for two unit vectors u and v, we use $\theta(u, v) = \arccos(u \cdot v)$ to denote the angle between the two vectors.

In this work, we focus on distributions whose marginal over X is an *isotropic log-concave* distribution. A distribution over d-dimensional vectors $x = \{x_1, x_2, \ldots, x_d\}$ with density function $f(x)$ is log-concave if $\log f(x)$ is concave. In addition, the distribution is isotropic if it is centered at the origin, and its covariance matrix is the identity, i.e., $\mathbf{E}[x_i] = 0$, $\mathbf{E}[{x_i}^2] = 1$, $\forall i$ and $\mathbf{E}[x_i x_j] = 0$, $\forall i \neq j$. Below we state useful properties of such distributions. See [16, 20, 160] for a proof of Lemma 2.

Lemma 2. *Let $\mathcal{P}$ be an isotropic log-concave distribution in $\mathbb{R}^d$. Then there exist absolute constants C_1, C_2 and C_3 such that*

1. *All marginals of $\mathcal{P}$ are isotropic log-concave.*
2. *For any two unit vectors u and v in $\mathbb{R}^d$, $C_1\theta(v, u) \leq \Pr_{x\sim\mathcal{P}}[\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(v \cdot x)]$.*
3. *For any unit vectors w and any γ, $C_3\gamma \leq \Pr_{x\sim\mathcal{P}}[|w \cdot x| \leq \gamma] \leq C_2\gamma$.*
4. *For any constant C_4, there exists a constant C_5 such that for two unit vectors u and v in $\mathbb{R}^d$ with $\|u - v\|_2 \leq r$ and $\theta(u, v) \leq \pi/2$, we have that*

$$\Pr_{x\sim\mathcal{P}}[\mathrm{sign}(u \cdot x) \neq \mathrm{sign}(v \cdot x) \text{ and } |v \cdot x| \geq C_5 r] \leq C_4 r.$$

5. *For any constant C_6, there exists another constant C_7, such that for any unit vectors v and u in $\mathbb{R}^d$ such that $\|u - v\|_2 \leq r$ and any $\gamma \leq C_6$, $\mathbf{E}_{x\sim\mathcal{P}_{u,\gamma}}[(v \cdot x)^2] \leq C_7(r^2 + \gamma^2)$.*

Proofs of Theorem 2

We use the following upper bound on the density of isotropic log-concave distributions.

Lemma 3 ([160]). *Let $\mathcal{P}$ be a 1-dimensional isotropic log-concave distribution over $\mathbb{R}$. Then* $\Pr_{x\sim\mathcal{P}}[x \geq \alpha] \leq \exp(-\alpha + 1)$.

Lemma 4. *There exists an absolute constant c_1, such that* $\mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] \leq c_1 \frac{\tau_k}{\gamma_{k-1}}$.

Proof. Notice that w^* never makes a mistake on distribution D_k, so the hinge loss of w^* on D_k is entirely attributed to the points of D_k that are within distance τ_k from w^*. We have,

$$\begin{aligned}
\mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] &\leq \Pr_{(x,y)\sim D_k}[|w^* \cdot x| < \tau_k] \\
&= \frac{\Pr_{(x,y)\sim D}[|w^* \cdot x| < \tau_k]}{\Pr_{(x,y)\sim D}[|w_{k-1} \cdot x| \leq \gamma_{k-1}]} \\
&\leq \frac{C_2\tau_k}{C_3\gamma_{k-1}} \qquad \text{(By Part 3 of Lemma 2)} \\
&\leq c_1 \frac{\tau_k}{\gamma_{k-1}}.
\end{aligned}$$

□

The next lemma uses VC dimension tools to show that for linear classifiers that are considered in Step 2c (the ones with angle α_k to w_k), the empirical and expected hinge loss are close. Let D'_k denote the distribution D_k where the labels are predicted based on $\text{sign}(p_k(\cdot))$. Note that T' is drawn from distribution D'_k.

Lemma 5. *There is $m_k = O(d(d + \log(k/d)))$ such that for a randomly drawn set T' of m_k labeled samples from D'_k, with probability $1 - \frac{\delta}{4(k+k^2)}$, for any $w \in B(w_{k-1}, r_{k-1})$,*

$$\left|\mathbf{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w, x, y)] - \ell_{\tau_k}(w, T')\right| \leq \frac{\lambda}{12}.$$

Proof. The pseudo-dimension of the set of hinge loss values, i.e., $\{\ell_{\tau_k}(w, \cdot) : w \in \Re^d\}$ is known to be at most d. Next, we prove that for any halfspace $w \in B(w_{k-1}, r_{k-1})$ and for any point $(x, y) \sim D'_k$, $\ell_{\tau_k}(w, x, y) \in O(\sqrt{d})$. We have,

$$\begin{aligned}
\ell_{\tau_k}(w, x, y) &\leq 1 + \frac{|w \cdot x|}{\tau_k} \\
&\leq 1 + \frac{|w_{k-1} \cdot x| + \|w - w_{k-1}\|_2\|x\|_2}{\tau_k} \\
&\leq 1 + \frac{\gamma_{k-1} + r_{k-1}\|x\|_2}{\tau_k} \\
&\leq c(1 + \|x\|_2).
\end{aligned}$$

By Lemma 3, for any $(x, y) \in T'$, $\Pr_{(x,y)\sim D'_k}[\|x\|_2 > \alpha] \leq c\exp(-\alpha/\sqrt{d})$. Using union bound and setting $\alpha = \Theta(\sqrt{d}\ln(|T'|k^2/\delta))$ we have that with probability $1 - \frac{\delta}{8(k+k^2)}$, $\max_{x\in T'} \|x\|_2 \in$

$O(\sqrt{d}\ln(|T'|k^2/\delta))$. Using standard pseudo-dimension rule we have that for $|T'| > \tilde{O}(d(d + \log\frac{k}{\delta}))$, with probability $1 - \frac{\delta}{4(k+k^2)}$,

$$\left|\mathbf{E}_{(x,y)\sim D'_k}[\ell(w,x,y)] - \ell(w,T')\right| \leq \frac{\lambda}{12}.$$

□

Lemma 6. *There exists an absolute constant c_2 such that*

$$|\mathbf{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w^*,x,y)] - \mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)]| \leq c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\text{sign}(p_k(x)) \neq h_{w^*}(x)]}.$$

Proof. Let N indicate the set of points (x,y) such that p_k and h_{w^*} disagree. We have,

$$\begin{aligned}
&\left|\mathbf{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w^*,x,y)] - \mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)]\right| \\
&\quad\leq \left|\mathbf{E}_{(x,y)\sim D'_k}\left[\mathbf{1}_{x\in N}\left(\ell_{\tau_k}(w^*,x,y) - \ell_{\tau_k}(w^*,x,\text{sign}(w^*\cdot x))\right)\right]\right| \\
&\quad\leq 2\,\mathbf{E}_{(x,y)\sim D'_k}\left[\mathbf{1}_{x\in N}\left(\frac{|w^*\cdot x|}{\tau_k}\right)\right] \\
&\quad\leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim D'_k}[x\in N]}\times\sqrt{\mathbf{E}_{(x,y)\sim D'_k}[(w^*\cdot x)^2]} \qquad \text{(By Cauchy Schwarz)} \\
&\quad\leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\text{sign}(p_k(x)) \neq h_{w^*}(x)]}\times\sqrt{\mathbf{E}_{(x,y)\sim D_k}[(w^*\cdot x)^2]} \qquad \text{(By definition of } N\text{)} \\
&\quad\leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\text{sign}(p_k(x)) \neq h_{w^*}(x)]}\times\sqrt{C_7(r_{k-1}^2+\gamma_{k-1}^2)} \qquad \text{(By Lemma 5)} \\
&\quad\leq c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\text{sign}(p_k(x)) \neq h_{w^*}(x)]}.
\end{aligned}$$

□

Lemma 7. *Let c_1 and c_2 be the absolute constants from Lemmas 6 and 4, respectively. Then with probability $1 - \frac{\delta}{2(k+k^2)}$,*

$$\text{err}_{D'_k}(h_{w_k}) \leq 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\text{sign}(p_k(x)) \neq h_{w^*}(x)]} + \frac{\lambda}{2}.$$

Proof. First, we note that the true 0/1 error of w_k on any distribution is at most its true hinge loss on that distribution. So, it suffices to bound the hinge loss of w_k on D'_k. Moreover, v_k approximately minimizes the hinge loss on distribution D'_k, so in particular, it performs better than w^* on D'_k. On the other hand, Lemma 6 shows that the difference between hinge loss of w^* on D'_k and D_k is small. So, we complete the proof by using Lemma 4 and bounding the hinge of

w^* on D_k. The following equations show the process of derivation of this bound as we explained.

$$
\begin{aligned}
&\mathrm{err}_{D'_k}(h_{w_k}) \le \mathbf{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w_k, x, y)] \qquad \text{(Since hinge loss larger than 0/1 loss)}\\
&\le 2\mathbf{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(v_k, x, y)] \qquad \text{(Since } \|v_k\|_2 > 0.5\text{)}\\
&\le 2L_{\tau_k}(v_k, T') + 2(\frac{\lambda}{12}) \qquad \text{(By Lemma 5)}\\
&\le 2L_{\tau_k}(w^*, T') + 4(\frac{\lambda}{12}) \qquad (v_k \text{ was an approximate hinge loss minimizer})\\
&\le 2\mathbf{E}_{(x,y)\sim D'_k}[\ell_{\tau_k}(w^*, x, y)] + 6(\frac{\lambda}{12}) \qquad \text{(By Lemma 5)}\\
&\le 2\mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*, x, y)] + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)]} + \frac{\lambda}{2} \quad \text{(By Lemma 6)}\\
&\le 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)]} + \frac{\lambda}{2}. \qquad \text{(By Lemma 4)}
\end{aligned}
$$

□

We are now ready to prove our main theorem.

Recall that we use the following parameters in Algorithm 1: $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = Cr_k$, where we defer the choice of C to later in the proof, $\lambda = \frac{3C_1}{8CC_2}$, $e_{\text{KKMS}} = \beta(\lambda/(4c_1 + 4c_2 + 2))^4$, and $\tau_k = \lambda\gamma_{k-1}/(4c_1 + 4c_2 + 2)$. Note, that by Equation (2.1), for any classifier h the excess error of h is upper bounded by the probability that h disagrees with h_{w^*}, i.e., $\mathrm{err}_D(h)$. Here, we show that Algorithm 1 returns w_s such that $\mathrm{err}_D(h_{w_s}) = \Pr_{(x,y)\sim D}[h_{w_s}(x) \neq h_{w^*}(x)] \le \epsilon$, and in turn, the excess error of h_{w_s} is also at most ϵ.

We use induction to show that at the k^{th} step of the algorithm, $\theta(w_k, w^*) \le \frac{e_0}{C_1 2^k}$. Since Part 4 of Lemma 2 and other Lemmas that build on it require $\theta(w, w^*) \le \frac{\pi}{2}$ for any considered halfspace, we need to choose e_0 such that $\theta(w_0, w^*) \le \frac{\pi}{2}$. Using Part 2 of Lemma 2, we have that $e_0 \le \frac{\pi}{2C_1}$.

Assume by the induction hypothesis that at round $k-1$, $\mathrm{err}_D(h_{w_{k-1}}) \le e_0/2^{k-1}$. We will show that w_k, which is chosen by the algorithm at round k, also has the property that $\mathrm{err}_D(h_{w_k}) \le e_0/2^k$. Let $S_k = \{x : |w_{k-1} \cdot x| \le \gamma_{k-1}\}$ indicate the band at round k. We divide the error of w_k to two parts, error outside the band and error inside the band. That is,

$$
\mathrm{err}_D(h_{w_k}) = \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)] + \Pr_{(x,y)\sim D}[x \in S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)]. \tag{2.2}
$$

By Part 2 of Lemma 2, $\theta(w_{k-1}, w^*) \le r_{k-1}$. So, for the first part of the above inequality, which is the error of w_k outside the band, we have that

$$
\begin{aligned}
&\Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)]\\
&\quad\le \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \neq h_{w_{k-1}}(x)] + \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_{k-1}}(x) \neq h_{w^*}(x)]\\
&\quad\le 2\frac{C_1 r_{k-1}}{16} \le \frac{e_0}{4 \times 2^k},
\end{aligned} \tag{2.3}
$$

where the penultimate inequality follows from the fact that by the choice of $w_k \in B(w_{k-1}, r_{k-1})$ and the induction hypothesis, respectively, $\theta(w_{k-1}, w_k) < r_{k-1}$ and $\theta(w_{k-1}, w^*) < r_{k-1}$; By choosing large enough constant C in $\gamma_{k-1} = Cr_{k-1}$, using Part 4 of Lemma 2, the probability of disagreement outside of the band is $C_1 r_{k-1}/16$.

For the second part of Equation (2.2) we have that

$$\Pr_{(x,y)\sim D}[x \in S_k \text{ and } h_{w_k}(x) \neq h_{w^*}(x)] = \mathrm{err}_{D_k}(h_{w_k}) \Pr_{(x,y)\sim D}[x \in S_k], \tag{2.4}$$

and

$$\mathrm{err}_{D_k}(h_{w_k}) \Pr_{(x,y)\sim D}[x \in S_k] \leq \mathrm{err}_{D_k}(h_{w_k}) C_2 \gamma_{k-1} \leq \mathrm{err}_{D_k}(h_{w_k}) \frac{2C_2 C e_0}{C_1 2^k}, \tag{2.5}$$

where the penultimate inequality is based on Part 3 of Lemma 2. Therefore, by replacing Equations (2.3) and 2.5 with Equation (2.2), we see that in order to have $\mathrm{err}_D(h_{w_k}) < \frac{e_0}{2^k}$, it suffices to show that $\mathrm{err}_{D_k}(h_{w_k}) \leq \frac{3C_1}{8CC_2} = \lambda$. The rest of the analysis is contributed to proving this bound. We have $\mathrm{err}_{D_k}(h_{w_k}) = \Pr_{(x,y)\sim D_k}[h_{w_k}(x) \neq h_{w^*}(x)] \leq \Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)] + \Pr_{(x,y)\sim D_k}[h_{w_k}(x) \neq \mathrm{sign}(p_k(x))]$. For the first part, using the assumption in Equation (2.1), we have that

$$\Pr_{(x,y)\sim D_k}[\mathrm{sign}(p_k(x)) \neq h_{w^*}(x)] \leq \frac{1}{\beta}\left(\mathrm{err}_{\tilde{D}_k}(\mathrm{sign}(p_k)) - \mathrm{err}_{\tilde{D}_k}(h_{w^*})\right) \leq \frac{e_{\text{KKMS}}}{\beta}. \tag{2.6}$$

For the second part, using Lemma 7, we have

$$\Pr_{(x,y)\sim D_k}[h_{w_k}(x) \neq \mathrm{sign}(p_k(x))] = \mathrm{err}_{D'_k}(h_{w_k}) \leq 2c_1 \frac{\tau_k}{\gamma_{k-1}} + 2c_2 \frac{\gamma_{k-1}}{\tau_k} \sqrt{\frac{e_{\text{KKMS}}}{\beta}} + \frac{\lambda}{2}.$$

Therefore, by the choice of parameter $\tau_k = \lambda\gamma_{k-1}/(4c_1 + 4c_2 + 2) = \gamma_{k-1}\left(e_{\text{KKMS}}/\beta\right)^{1/4}$, we have

$$\begin{aligned}
\mathrm{err}_{D_k}(h_{w_k}) &\leq \frac{e_{\text{KKMS}}}{\beta} + 2c_1 \frac{\tau_k}{\gamma_{k-1}} + 2c_2 \frac{\gamma_{k-1}}{\tau_k} \sqrt{\frac{e_{\text{KKMS}}}{\beta}} + \frac{\lambda}{2} \\
&\leq \frac{e_{\text{KKMS}}}{\beta} + 2c_1 \left(\frac{e_{\text{KKMS}}}{\beta}\right)^{1/4} + 2c_2 \left(\frac{e_{\text{KKMS}}}{\beta}\right)^{1/4} + \frac{\lambda}{2} \\
&\leq (2c_1 + 2c_2 + 1)\left(\frac{e_{\text{KKMS}}}{\beta}\right)^{1/4} + \frac{\lambda}{2} \leq \frac{\lambda}{2} + \frac{\lambda}{2} \leq \lambda.
\end{aligned}$$

Sample complexity and runtime: To get error of e_{KKMS} with probability $1 - \frac{s}{\delta}$ at every round, we need a labeled set of size $\mathrm{poly}(d, \log\frac{s}{\delta})$. The sample set T' is labeled based on p_k, so it does not contribute to the label complexity. So, at each round, we need $m_k = \mathrm{poly}(d, \log(\frac{\log(1/\epsilon)}{\delta}))$ labels. At each round, to get $\mathrm{poly}(d, \log(\frac{\log(1/\epsilon)}{\delta}))$ labels for the polynomial regression algorithm in the band of S_k we need $O(2^k m_k)$ samples from $\tilde{D}$. To get $d(d + \log(k/\delta))$ unlabeled samples in the band for Step 2b, we need $O(2^k(d(d + \log(k/\delta))) = \mathrm{poly}(d, \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples. So, overall, we need $n_k = \mathrm{poly}(d, \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples at each round. The running time is dominated by the polynomial regression algorithm which takes time $d^{\exp(\frac{1}{\beta^4})}$. However, since β is a constant, this is a polynomial in d.

Proofs of Theorem 3

In this section, we prove Theorem 3 for efficiently learning halfspaces under isotropic log-concave distributions in presence of bounded noise with parameter β that is independent of the dimension. We will assume that the target vector w^* is t-sparse.

We will first argue the proof of correctness and then the sample complexity. To argue correctness, we need to show that the new polynomial regression based algorithm in Step 2a of Algorithm 2 will indeed output a polynomial of excess error at most e_{KKMS}. Secondly, we need to argue that the hinge loss minimization w.r.t. the polynomial $p(\cdot)$ will output a vector v_k that is close to $p(\cdot)$. The second part is easy to see, since the vector w^* itself has L_1 norm at most $\sqrt{t}$. By restricting to vectors of small L_1 norm we still have to find a v_k with L_1 norm at most $\sqrt{t}$ that does well in the class (in comparison to w^*) in learning labels of $p(\cdot)$. For the first part, we prove the following extension of [128].

Theorem 6. *Let (X, Y) be drawn from a distribution over $\mathbb{R}^d \times \{+1, -1\}$ with isotropic log-concave marginal, constrained to the set $\{x : |w \cdot x| \leq \gamma\}$ for some w and γ. Let OPT be the error of the best t-sparse halfspace, i.e., $OPT = \min_{w \in \mathbb{R}^d, \|w\|_0 \leq t} \Pr_{(x,y)\sim D}[\text{sign}(w \cdot x) \neq y]$. Then, for every $\epsilon > 0$, there is an algorithm that runs in time $d^{\text{poly}(\frac{1}{\epsilon})}$ and uses $m = O_\epsilon\left((\frac{t}{\gamma})^{\text{poly}(\frac{1}{\epsilon})}\text{polylog}(d)\right)$ samples from the distribution and outputs a polynomial $p(\cdot)$ such that $\text{err}(p) \leq OPT + \epsilon$. Here, $\text{err}(p) = \Pr_{(x,y)}[\text{sign}(p(x)) \neq y]$. Furthermore, the polynomial $p(\cdot)$ satisfies $\|p\|_1 \leq (\frac{t}{\gamma})^{\text{poly}(\frac{1}{\epsilon})}$.*

Note that the claimed sample complexity of our approach is an immediate consequence of the above theorem, since we require error of e_{KKMS} in the band and the subsequent hinge loss minimization step of our algorithm only uses examples labeled by $p(\cdot)$ and, therefore, does not affect the overall sample complexity of our algorithm. In order to prove the theorem, we need the following result about approximation of sign of halfspaces by polynomials.

Theorem 7. *Let w^* be a halfspace in $\mathbb{R}^d$. Then, for every log-concave distribution over $\mathbb{R}^d$, there exists a degree $\frac{1}{\epsilon^2}$ polynomial $p(\cdot)$ such that $\mathbf{E}[(p(x) - \text{sign}(w^* \cdot x))^2] \leq \epsilon$. Here the expectation is over a random x drawn from the distribution.*

Proof of Theorem 6. First, consider an isotropic log-concave distribution. Notice that if w^* is t-sparse, then the polynomial $p(\cdot)$ referred to in Theorem 7 will have support size at most $t^{\frac{1}{\epsilon^2}}$. This is due to the fact that the isotropicity and log-concavity of the distribution is preserved when considering the projection of the instance space on the relevant t variables. Since there are only t^{1/ϵ^2} monomials in the lower dimension of degree at most $\frac{1}{\epsilon^2}$, the $\frac{1}{\epsilon^2}$-degree polynomial $p(\cdot)$ that satisfies the theorem in this lower dimension also satisfies the requirement in the original space and is t^{1/ϵ^2}-sparse. The analysis of [128] also shows that $p(\cdot)$ is $\sum_{i=0}^{deg} c_i \bar{H}_i(\cdot)$, the linear combination of up to degree $deg = \frac{1}{\epsilon^2}$ normalized Hermite polynomials, where $\sum_{i=0}^{deg} c_i^2 < 1$ and $\bar{H}_i(x) = H_i(x)/\sqrt{2^i i!}$ refers to the normalized Hermite polynomial with degree i. By a naïve bound of $\sqrt{i!2^i}$ on the coefficients of $\bar{H}_i(x)$ and the fact that $i < \frac{1}{\epsilon^2}$, we know that the L_1 norm of each of the Hermite polynomials is bounded by $O_\epsilon(t^{1/\epsilon^2})$, where O_ϵ considers ϵ to be a constant. Moreover, since $\sum_{i=0}^{deg} c_i \leq \sqrt{deg}\sum_{i=0}^{deg} c_i^2 < \sqrt{deg}$, the L_1 norm of p is also bounded by $t^{O(\frac{1}{\epsilon^2})}$.

This holds when the distribution is isotropic log-concave. However, the distributions we consider are conditionals of isotropic log-concave distribution over $\{|w \cdot x| \leq \gamma_k\}$. These

distributions are log-concave but not isotropic. To put them in the isotropic position, we transform each instance x to x' by a factor $O(\frac{1}{\gamma})$ along the direction of w. Then applying the above procedure on the transformed distribution we get a polynomial $p'(x') = \sum_{i=0}^{deg} p'_i \prod_{j=1}^{d}(x'_j)^{a_j}$. Since $x'_i \leq O(\frac{1}{\gamma})x_i$ for every i, this polynomial can be formed in terms of x as $p(x) = \sum_{i=0}^{deg} p_i \prod_{j=1}^{d}(x_j)^{a_j}$, where $p_i \leq O((\frac{1}{\gamma})^i)p'_i$. Therefore, for such distributions, the coefficients of the polynomial blow up by a factor of $O((\frac{1}{\gamma})^{\text{poly}(1/\epsilon)})$ and as a result $\|p\|_1 \leq O((\frac{t}{\gamma})^{\text{poly}(1/\epsilon)})$. Thus, by enforcing that the polynomial $p(\cdot)$ belongs to $S = \{q : \|q\|_1 = O((\frac{t}{\gamma})^{\text{poly}(1/\epsilon)}) \text{ and degree}(q) \leq \text{poly}(1/\epsilon)\}$, we only need to argue about polynomials in the set S as opposed to general $\text{poly}(\frac{1}{\epsilon})$-degree polynomials. Hence, as in [128], we run the L_1 regression algorithm, but we also ensure that the L_1 norm of the induced polynomial is bounded by $\|q\|_1 = O((\frac{t}{\gamma})^{\text{poly}(1/\epsilon)})$. This can be done via constrained L_1 norm minimization. The analysis of this algorithm is similar as that of [128]. For the self-completeness of the paper, we show a complete proof here. Denote by $\mathcal{Z} = (x^1, y^1), \cdots, (x^m, y^m)$ the samples. Firstly, we have

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m} I(q(x^j)y^j < \gamma) &= \frac{1}{m}\sum_{j=1}^{m} I(\text{sign}(q(x^j)) \neq y^j) + \frac{1}{m}\sum_{j=1}^{m} I(\text{sign}(q) = y^j \ \&\ q(x^j)y^j < \gamma) \\ &\leq \frac{1}{2m}\sum_{j=1}^{m} |y^j - p(x^j)| + \frac{\gamma}{2},\end{aligned} \tag{2.7}$$

where $q(x) = p(x) - T$. The above inequality holds because of a standard argument on the randomized threshold T: Note that $\text{sign}(q(x^j)) \neq y^j$ iff the threshold T lies between $p(x^j)$ and y^j; Similarly, $\text{sign}(q(x^j)) = y^j \ \&\ q(x^j)y^j < \gamma$ iff the threshold T lies between $p(x^j)$ and $p(x^j) - y^j\gamma$. So if we choose T uniformly at random on $[-1, 1]$, Equation (2.7) holds in expectation. Since we select T to minimize the LHS of Equation (2.7), the inequality holds with certainty. Then by the L_1 polynomial regression algorithm which fits the labels by polynomial in the sense of L_1 norm, we have

$$\frac{1}{m}\sum_{j=1}^{m} |y^j - p(x^j)| \leq \frac{1}{m}\sum_{j=1}^{m} |y^j - p^*(x^j)| \leq \frac{1}{m}\sum_{j=1}^{m} |y^j - c(x^j)| + |c(x^j) - p^*(x^j)|,$$

where c is the optimal classifier and p^* is a polynomial satisfying Theorem 7. Thus

$$\mathbb{E}_{\mathcal{Z}}\left[\frac{1}{m}\sum_{j=1}^{m} I(q(x^j)y^j < \gamma)\right] \leq OPT + \frac{\epsilon}{2} + \frac{\gamma}{2}.$$

Let $S = \{q : \text{degree}(q) \leq \frac{1}{\epsilon^2}, \|q\|_1 \leq (\frac{t}{\gamma})^{O(\frac{1}{\epsilon^2})}\}$ and let $\hat{L}(q) = \frac{1}{m}\sum_{(x^j,y^j)} I(q(x^j)y^j < \gamma)$ be the empirical $0/1$ loss of the polynomial q with margin γ. In order to complete the proof, we need to argue that if m is large enough then for all $q \in S$, we have, with high probability, $|\hat{L}(q) - err(q)| \leq \epsilon/4$. To see this, we need the following lemma of [261].

Lemma 8 ([261]). *Let the instance space be bounded as $\|x\|_\infty \leq X_\infty$, and consider the class of hyperplane w such that $\|w\|_1 \leq W_1$. Denote by $err(w)$ the expected 0/1 error of w. Then there is a constant C such that with probability $1-\delta$, for all γ, we have*

$$err(w) \leq \frac{1}{m}\sum_{j=1}^{m} I(y^j(w\cdot x^j) < \gamma) + \sqrt{\frac{C}{m}\left(\frac{X_\infty^2 W_1^2(\log d+1)}{\gamma^2}\log m + \log\frac{1}{\delta}\right)}.$$

Setting γ as $\epsilon/2$, W_1 as $\left(\frac{t}{\gamma}\right)^{O(\frac{1}{\epsilon^2})}$, and X_∞ as $O\left((\log(md))^{O(\frac{1}{\epsilon^2})}\right)$ (see Lemma 10), viewing the polynomial q as a $d^{O(1/\epsilon^2)}$-dimensional vector, Lemma 8 gives the desired sample complexity $m = O_\epsilon\left((\frac{t}{\gamma})^{\mathrm{poly}(\frac{1}{\epsilon})}\mathrm{polylog}(d)\right)$. □

In the above, we explicitly suppressed the dependence on ϵ, because for the purpose of our algorithm, we use a constant value e_{KKMS} for the desired value of the error in Theorem 6. Moreover, the distribution at every round is restricted to the set $\{x : |w_{k-1}\cdot x| \leq \gamma_{k-1}\}$. Since $\gamma_k \geq \epsilon$, for all k, we use the value of $\gamma = \epsilon$ in Theorem 6 and achieve the results of Theorem 3 as a consequence.

Proofs of Theorem 4

In this section, we first consider the case of non-uniform 1-bit compressed sensing under adversarial noise and provide a proof of Theorem 4. Then, we discuss an extension of our analysis that holds for uniform 1-bit compressed sensing under adversarial noise and provide a proof of Theorem 5.

We start with the following result of [14].

Theorem 8 ([14]). *Let (x,y) be drawn from a distribution over $\mathbb{R}^d \times \{+1,-1\}$ such that the marginal over x is isotropic log-concave. Let OPT be the $0/1$ error of the best halfspace, i.e., $OPT = \min_{w:\|w\|_2=1}\Pr[\mathrm{sign}(w\cdot x)\neq y]$ and let w^* be the halfspace that achieves OPT. Then, there exists an algorithm that, for every $\epsilon > 0$, runs in time polynomial in d and $\frac{1}{\epsilon}$ and outputs a halfspace w such that $\|w - w^*\|_2 \leq O(OPT) + \epsilon$.*

We extend the algorithm of [14] for 1-bit compressed sensing. The main difference between our algorithm and the algorithm of [14] is in the hinge-loss minimization step and the sample complexity. In this case, when minimizing hinge loss at each step, we restrict the search to vectors of L_1 norm bounded by $\sqrt{t}$. Note that this does not affect the correctness of the algorithm, as w^* itself is t-sparse and $\|w^*\|_1 \leq \sqrt{t}$. The crux of the argument is in showing that when $\|w\|_1 \leq \sqrt{t}$, the empirical hinge loss of w is nicely concentrated around its expectation. This is proved in Lemma 12. Using this new concentration results, the proof of Theorem 4 follows immediately by the analysis of [14]. For completeness, here we provide a complete proof of Theorem 4.

To achieve desirable concentration result, we use the tools from VC and Rademacher complexity theory to obtain a sample complexity that is polynomial in t and only logarithmic in the ambient dimension d. The following lemma helps us in achieving such concentration result.

Lemma 9 ([207]). *Let $\mathcal{F}$ be the class of linear predictors with the L_1 norm of the weights bounded by W_1. Assume that the infinity norm of all instances is bounded by X_∞. Then for the ρ-Lipschitz*

loss ℓ such that $\max_{w\cdot x\in[-W_1X_\infty,W_1X_\infty]}|\ell(w,x,y)|\le U$ and the choice of an i.i.d. sample T of size m,

$$\forall w,\ s.t.,\ \|w\|_1\le W_1,\ \Pr\left[|\mathbf{E}\ell(w,x,y)-\ell(w,T)|\ge 2\rho W_1X_\infty\sqrt{\frac{2\log(2d)}{m}}+s\right]\le 2\exp\left(-\frac{ms^2}{2U^2}\right).$$

In preparation to use Lemma 9, we bound the infinity norm of the instances used by our algorithm in the next lemma.

Lemma 10. *Let S be the set of all (unlabeled) samples drawn from D. With probability $1-\delta$ for all $x\in S$, $\|x\|_\infty\le O(\log\frac{|S|d}{\delta})$.*

Proof. Since D is an isotropic log-concave distribution, the marginal distribution on any coordinate is a one-dimensional isotropic log-concave distribution. Therefore, by concentration results of [160], we have

$$\Pr_{x\sim D}\left[\|x\|_\infty\ge c'\log\frac{d}{\delta}\right]\le\sum_{i\in[d]}\Pr_{x\sim D}\left[x_i\ge c'\log\frac{d}{\delta}\right]\le\delta.$$

Taking union bound over all elements of S, with probability $1-\delta$, $\|x\|_\infty\le O(\log\frac{|S|d}{\delta})$. □

Next, we bound the value of hinge loss on any instance (x,y) used by our algorithm. Let H be a class of halfspaces w, with $\|w\|_1\le\sqrt{t}$ and $\|w\|_2=1$.

Lemma 11. *For a given k and $v\in H$, let T' be the set of m_k samples drawn from $\tilde{D}_{v,\gamma_k}$. For any halfspace u such that $\|u\|_2=1$, $\|u\|_1\le\sqrt{t}$ and $u\in B(v,r_k)$, with probability $1-\delta$, for all $x\in T'$, $\ell_{\tau_k}(u,x,y)\le O(\log\frac{m_k}{\gamma_k\delta})$.*

Proof. We have

$$\ell_{\tau_k}(u,x,y)\le 1+\frac{|u\cdot x|}{\tau_k}\le 1+\frac{|v\cdot x|}{\tau_k}+\frac{|(u-v)\cdot x|}{\tau_k}.$$

By the choice of $x\sim D_{v,\gamma_k}$, we know that $v\cdot x\le\gamma_k$. Therefore, $\frac{|v\cdot x|}{\tau_k}\le O(1)$. For the second part of the inequality, $|(u-v)\cdot x|$, first consider all $x\sim D$. Since, D is an isotropic log-concave distribution and $\|u-v\|\le r_k$, without loss of generality, we can assume that $u-v=(r,0,\ldots,0)$ for some $r\le r_k$. Moreover, $(u-v)\cdot x=r|x_1|$, and x_1 is a one-dimensional isotropic log-concave distribution. Therefore,

$$\Pr_{x\sim D}\left[|(u-v)\cdot x|\ge r_k(1+\log\frac{1}{\delta})\right]\le\Pr_{x\sim D}\left[r|x_1|\ge r_k(1+\log\frac{1}{\delta})\right]\le\Pr_{x\sim D}\left[|x_1|\ge 1+\log\frac{1}{\delta}\right]\le\delta.$$

So,

$$\begin{aligned}\Pr_{x\sim D_k}\left[|(u-v)\cdot x|\ge r_k(1+\log\frac{1}{\gamma_k\delta})\right]&=\frac{\Pr_{x\sim D}\left[|(u-v)\cdot x|\ge r_k(1+\log\frac{1}{\gamma_k\delta})\ \&\ \ |v\cdot x|\le\gamma_k\right]}{\Pr_{x\sim D}[|v\cdot x|\le\gamma_k]}\\&\le\frac{\Pr_{x\sim D}\left[|(u-v)\cdot x|\ge r_k(1+\log\frac{1}{\gamma_k\delta})\right]}{\Pr_{x\sim D}[|v\cdot x|\le\gamma_k]}\\&\le\Theta(\frac{1}{\gamma_k})\Pr_{x\sim D}\left[|(u-v)\cdot x|\ge r_k(1+\log\frac{1}{\gamma_k\delta})\right]\\&\le\delta.\end{aligned}$$

So for a fixed v and k, and for all m_k samples T' with probability $1-\delta$, $\frac{|(u-v)\cdot x|}{\tau_k} \leq \frac{r_k}{\tau_k} \log \frac{m_k}{\gamma_k \delta} \leq O(\log \frac{m_k}{\gamma_k \delta})$.

□

Lemma 12. *Let $m_k = \Omega(\frac{t}{\epsilon^2}\text{polylog}(d, \frac{1}{\delta}, \frac{1}{\epsilon}))$ and T' be the samples drawn from $\tilde{D}_k$ and T to be the corresponding samples when their labels are corrected based on w^*. With probability $1-\delta$,*

$$\sup_w \left| \mathbf{E}_{(x,y)\sim \tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T') \right| \leq \frac{\lambda}{12},$$

and

$$\sup_w \left| \mathbf{E}_{(x,y)\sim \tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T) \right| \leq \frac{\lambda}{12}.$$

where $w \in B(w_{k-1}, r_{k-1})$ such that $\|w\|_1 \leq \sqrt{t}$.

Proof. Using Lemma 9 we have that

$$\Pr\left[\sup_w \left| \mathbf{E}_{(x,y)\sim \tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T') \right| \geq 2\rho W_1 X_\infty \sqrt{\frac{2\log(2d)}{m_k}} + s \right] \leq 2\exp\left(-\frac{m_k s^2}{2U^2}\right), \tag{2.8}$$

where U, ρ, W_1 and X_∞ are defined as Lemma 9, and the supremum is taken over all w in $K = \{w \in \mathbb{R}^d : \|w\|_1 \leq W_1, \|w\|_2 \leq 1\}$. Note that $W_1 \leq \sqrt{t}$ and $\rho = \frac{1}{\tau_k} \leq \frac{1}{\epsilon}$ and by Lemma 10 and 11 for any δ, with probability δ, $X_\infty \leq O(\log \frac{md}{\delta})$ and $U \leq O(\log \frac{m_k}{\gamma_k \delta})$.

Assume that these bounds hold for X_∞ and U. For $m = \Theta(\frac{t}{\epsilon^3}\text{polylog}(d, \frac{1}{\delta}, \frac{1}{\epsilon}))$ and $m_k \geq \Omega(\frac{t}{\epsilon^2}\log(md/\delta)\log d)$, and for appropriate choice of constant s, with probability at most δ,

$$\sup_w \left| \mathbf{E}_{(x,y)\sim \tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T') \right| \geq 2\frac{\sqrt{t}}{\epsilon}\log(\frac{md}{\delta})\sqrt{\frac{2\log(2d)}{m_k}} + s \geq \lambda/12.$$

The proof for the case of T is similar to the above. □

We would need the following lemmas:

Lemma 13. *There exists an absolute constant c_1 such that in round k of Algorithm 3, we have $\mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)] \leq c_1 \frac{\tau_k}{\gamma_{k-1}}$.*

Proof. Notice that w^* never makes a mistake on distribution D_k, so the hinge loss of w^* on D_k is entirely attributed to the points of D_k that are within distance τ_k from w^*. We have,

$$\begin{aligned}
\mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)] &\leq \Pr_{(x,y)\sim D_k}[|w^*\cdot x| < \tau_k] \\
&= \frac{\Pr_{(x,y)\sim D}[|w^*\cdot x| < \tau_k]}{\Pr_{(x,y)\sim D}[|w_{k-1}\cdot x| \leq \gamma_{k-1}]} \\
&\leq \frac{C_2\tau_k}{C_3\gamma_{k-1}} \qquad \text{(By Part 3 of Lemma 2)} \\
&\leq c_1\frac{\tau_k}{\gamma_{k-1}}.
\end{aligned}$$

□

Lemma 14. *There exists an absolute constant c_2 such that*

$$\left|\mathbf{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w^*,x,y)] - \mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)]\right| \leq c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{C2^k\nu}.$$

Proof. Let N indicate the set of points where w^* makes a mistake. We have,

$$\begin{aligned}
&\left|\mathbf{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w^*,x,y)] - \mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)]\right| \\
&\quad\leq \left|\mathbf{E}_{(x,y)\sim\tilde{D}_k}\left[\mathbf{1}_{x\in N}\left(\ell_{\tau_k}(w^*,x,y) - \ell_{\tau_k}(w^*,x,\mathrm{sign}(w^*\cdot x))\right)\right]\right| \\
&\quad\leq 2\,\mathbf{E}_{(x,y)\sim\tilde{D}_k}\left[\mathbf{1}_{x\in N}\left(\frac{|w^*\cdot x|}{\tau_k}\right)\right] \\
&\quad\leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim\tilde{D}_k}[x\in N]}\times\sqrt{\mathbf{E}_{(x,y)\sim\tilde{D}_k}[(w^*\cdot x)^2]} \qquad \text{(By Cauchy Schwarz)} \\
&\quad\leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[x\in N]}\times\sqrt{\mathbf{E}_{(x,y)\sim D_k}[(w^*\cdot x)^2]} \qquad \text{(By definition of } N\text{)} \\
&\quad\leq \frac{2}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[x\in N]}\times\sqrt{C_7(r_{k-1}^2+\gamma_{k-1}^2)} \qquad \text{(By Lemma 5)} \\
&\quad\leq c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{\Pr_{(x,y)\sim D_k}[x\in N]} \\
&\quad\leq c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{C2^k\nu}. \qquad \text{(The noise rate within the band can go up by a factor of } 2^k\text{)}
\end{aligned}$$

□

Lemma 15. *Let c_1 and c_2 be the absolute constants from Lemmas 13 and 14, respectively. Then with probability $1-\frac{\delta}{2(k+k^2)}$,*

$$\mathrm{err}_{\tilde{D}_k}(h_{w_k}) \leq 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{C\nu 2^k} + \frac{\lambda}{2}.$$

Proof. First, we note that the true 0/1 error of w_k on any distribution is at most its true hinge loss on that distribution. So, it suffices to bound the hinge loss of w_k on $\tilde{D}_k$. Moreover, v_k approximately minimizes the hinge loss on distribution $\tilde{D}_k$, so in particular, it performs better than w^* on $\tilde{D}_k$. On the other hand, Lemma 14 shows that the difference between hinge loss of w^* on $\tilde{D}_k$ and D_k is small. So, we complete the proof by using Lemma 13 and bounding the hinge of

w^* on D_k. The following equations show the process of derivation of this bound as we explained.

$$\begin{aligned}
&\mathrm{err}_{\tilde{D}_k}(h_{w_k}) \le \mathbf{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w_k,x,y)] \qquad \text{(Since hinge loss larger than 0-1 loss)}\\
&\le 2\mathbf{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(v_k,x,y)] \qquad \text{(Since } \|v_k\|_2 > 0.5\text{)}\\
&\le 2L_{\tau_k}(v_k,T') + 2(\frac{\lambda}{12}) \qquad \text{(By Lemma 12)}\\
&\le 2L_{\tau_k}(w^*,T') + 4(\frac{\lambda}{12}) \qquad (v_k \text{ was an approximate hinge loss minimizer})\\
&\le 2\mathbf{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w^*,x,y)] + 6(\frac{\lambda}{12}) \qquad \text{(By Lemma 12)}\\
&\le 2\mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w^*,x,y)] + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{C2^k\nu} + \frac{\lambda}{2} \quad \text{(By Lemma 14)}\\
&\le 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{C2^k\nu} + \frac{\lambda}{2}. \qquad \text{(By Lemma 13)}
\end{aligned}$$

□

We are now ready to prove our main theorem.

Proof of Theorem 4. We use induction to show that at the k^{th} step of the algorithm, $\theta(w_k,w^*) \le \frac{e_0}{C_1 2^k}$ where e_0 is the initial error of w_0. Assume by the induction hypothesis that at round $k-1$, $\mathrm{err}_D(h_{w_{k-1}}) \le e_0/2^{k-1}$. We will show that w_k, which is chosen by the algorithm at round k, also has the property that $\mathrm{err}_D(h_{w_k}) \le e_0/2^k$. Let $S_k = \{x : |w_{k-1}\cdot x| \le \gamma_{k-1}\}$ indicate the band at round k. We divide the error of w_k to two parts, error outside the band and error inside the band. That is,

$$\mathrm{err}_D(h_{w_k}) = \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \ne h_{w^*}(x)] + \Pr_{(x,y)\sim D}[x \in S_k \text{ and } h_{w_k}(x) \ne h_{w^*}(x)]. \tag{2.9}$$

By Part 2 of Lemma 2, $\theta(w_{k-1},w^*) \le r_{k-1}$. So, for the first part of the above inequality, which is the error of w_k outside the band, we have that

$$\begin{aligned}
&\Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \ne h_{w^*}(x)]\\
&\quad\le \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_k}(x) \ne h_{w_{k-1}}(x)] + \Pr_{(x,y)\sim D}[x \notin S_k \text{ and } h_{w_{k-1}}(x) \ne h_{w^*}(x)]\\
&\quad\le 2\frac{C_1 r_{k-1}}{16} \le \frac{e_0}{4\times 2^k},
\end{aligned}$$

where the penultimate inequality follows from the fact that by the choice of $w_k \in B(w_{k-1},r_{k-1})$ and the induction hypothesis, respectively, $\theta(w_{k-1},w_k) < r_{k-1}$ and $\theta(w_{k-1},w^*) < r_{k-1}$; By choosing large enough constant in $\gamma_{k-1} = Cr_{k-1}$, using Part 4 of Lemma 2, the probability of disagreement outside of the band is $C_1 r_{k-1}/16$.

For the second part of Equation 2.9, using the same derivation as in Lemma 15 we get that

$$\mathrm{err}_{D_k}(h_{w_k}) \le 2c_1\frac{\tau_k}{\gamma_{k-1}} + 2c_2\frac{\gamma_{k-1}}{\tau_k}\sqrt{C\nu 2^k} + \frac{\lambda}{2}.$$

By our choice of parameters, we know that the ratio of τ_k and γ_{k-1} is bounded by $\leq \frac{\lambda}{12}$. Hence, for the sum to be bounded by λ, we need $C2^k\nu$ to be bounded by a constant. But this is true since $k \geq \log \frac{1}{c\nu+\epsilon}$ for an appropriate constant c. □

Proofs of Theorem 5

Next, we provide a proof for Theorem 5. We extend our analysis from the previous section to hold for the case of uniform 1-bit compressed sensing. The main difference between the results of this section and the analysis of the previous section is that we need to obtain a concentration result that holds uniformly over all choice of underlying noisy distribution. In other words, they hold uniformly over the choice of w^* and the ν fraction of the samples whose labels differ from the labels of w^*.

First, we introduce Lemma 16 that shows that for a large enough number of unlabeled samples, *every band* around a halfspace that can be considered by the algorithm has sufficient samples. In contrast, the results of the previous section only show that the bands around $w_1, \ldots, w_k$, which are uniquely determined by the samples and the fixed (but unknown) distribution $\tilde{D}$, have sufficient samples. Next, we build on the concentration results from the previous section and show that the hinge loss is concentrated around its expectation uniformly over all choice of all w^* and ν fraction of the samples whose labels differ from the labels of w^*. Using this new concentration result, the proof of Theorem 5 follows immediately by the analysis of the non-uniform case.

Note that at every step of the algorithm, vector v_k that is chosen by the hinge loss minimization step is such that $\|v_k\|_1 \leq \sqrt{t}$. As [14] argue, $\|v_k\|_2 \geq 1/2$. Therefore, the outcome of step 3c also satisfies $\|w_k\|_1 \leq O(\sqrt{t})$. The following lemma shows that when the number of unlabeled samples is large enough, every possible band around every such w_k considered by the algorithm contains a number of points that is at least a multiplicative approximation to the number of points expected to be in that band. Therefore, in every step of the algorithm, there is a sufficient number of samples in the band.

Lemma 16. *Let S be a set of $m \geq \frac{t}{\epsilon^4}\operatorname{polylog}(d)\log(\frac{1}{\delta})$ samples drawn from D. With probability $1-\delta$ for all $\gamma \in \Omega(\epsilon)$ and w such that $\|w\|_1 \leq O(\sqrt{t})$ and $\|w\|_2 = 1$,*

$$\frac{1}{m}\Big|\{x \mid x \in S \text{ and } |w \cdot x| \leq \gamma\}\Big| \geq c\gamma,$$

where c is a constant.

Proof. Let H be a class of (non-homogeneous) halfspaces w, with $\|w\|_1 \leq O(\sqrt{t})$ and $\|w\|_2 = 1$. Let B be a class of hypothesis defined by bands around homogeneous halfspaces, w, such that $\|w\|_1 \leq O(\sqrt{t})$ and $\|w\|_2 = 1$ with arbitrary width.

The covering number of H is at most $\log N(\gamma, H) = O\left(t \operatorname{polylog}(d)/\gamma^2\right)$ [185]. Since every band is an intersection of two halfspaces, each band of B can be represented by the intersection of two halfspaces from H. Therefore, $\log N(\gamma, B) = O\left(t \operatorname{polylog}(d)/\gamma^2\right)$. Furthermore, by Lemma 2, $\mathbf{E}\left[\frac{1}{m}\left|\{x \mid x \in S \text{ and } |w \cdot x| \leq \gamma\}\right|\right] = \Theta(\gamma)$. Therefore, by the uniform convergence

results for covering number, we have that

$$\Pr\left[\sup_{w,\gamma}\ \frac{1}{m}\Big|\{x \mid x \in S \text{ and } |w \cdot x| \le \gamma\}\Big| - \Theta(\gamma) \le \gamma\right] \le N(\gamma, H)e^{-\frac{\gamma^2 m}{8}}$$
$$\le e^{-\frac{\epsilon^2 m}{8} + \frac{t\ \mathrm{polylog}(d)}{\epsilon^2}}$$
$$\le \delta.$$

□

Lemma 17. *For $\nu \in O(\epsilon/\log(\frac{d}{\epsilon})^2)$ and S of size $m = \Theta(\frac{t}{\epsilon^4}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$, with probability $1-\delta$,*

$$\sup_{w^*, \{y_i\}_{i=1}^m, w} \left|\mathbf{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \le \frac{\lambda}{12},$$

and

$$\sup_{w^*, \{y_i\}_{i=1}^m, w} \left|\mathbf{E}_{(x,y)\sim D_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T)\right| \le \frac{\lambda}{12},$$

where w^* is a t-sparse halfspace, $\{y_i\}_{i=1}^m$ are the labels of the set of samples S such at most ν fraction of them differs from the labels of w^*, and $w_{k-1} \in H$ is the unique halfspace determined by the outcome of step k of the algorithm given w^* and $\{y_k\}_{k=1}^m$ (labels used in the previous round), and $w \in B(w_{k-1}, r_{k-1})$ such that $\|w\|_1 \le \sqrt{t}$.

Proof. Using Lemma 9 we have that

$$\Pr\left[\sup_w \left|\mathbf{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \ge 2\rho W_1 X_\infty \sqrt{\frac{2\log(2d)}{m_k}} + s\right] \le 2\exp\left(-\frac{m_k s^2}{2U^2}\right), \tag{2.10}$$

where U, ρ, W_1 and X_∞ are defined as Lemma 9, and the supremum is taken over all w in $K = \{w \in \mathbb{R}^d : \|w\|_1 \le W_1, \|w\|_2 \le 1\}$. Note that $W_1 = \sqrt{t}$ and $\rho = \frac{1}{\tau_k} \le \frac{1}{\epsilon}$ and by Lemma 10 and 11 for any δ, with probability δ, $X_\infty \le O(\log \frac{md}{\delta})$ and $U \le O(\log \frac{m_k}{\gamma_k \delta})$.

Assume that these bounds hold for X_∞ and U. Then for a fixed w_{k-1} considering $m_k = \frac{t}{\epsilon^3}\ \mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$ of the samples in the band around it (there are $O(m\gamma_k) \ge m_k$ such samples by Lemma 16), and for an appropriate choice of constant s, with probability at most $2\exp\left(-\frac{m_k s^2}{2U^2}\right)$,

$$\sup_w \left|\mathbf{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \ge 2\frac{\sqrt{t}}{\epsilon}\mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})\sqrt{\frac{2\log(2d)}{m_k}} + s \ge \lambda/12$$

Next, we show how to achieve a similar concentration result over all choices of w^* and choices of νm corrupted measurements and the resulted w_{k-1}. Note that w_{k-1} depends only on the samples of S and their labels used in previous steps. Since, we only use labels of $m_k = O(\frac{t}{\gamma^3}\ \mathrm{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$ points in every step, overall, Equation (2.10) only depends on the

labels of these sample points. This is uniquely determined by the choice of w^* and the ν fraction of the samples that do not agree with labels of w^*. Therefore, we can restrict our attention to the different labelings that can be produced by such w^* and adversarial corruption on the sample of size $\sum_i m_i \leq \frac{t}{\epsilon^3}\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta})$.

Let $K' = \{w \in \mathbb{R}^d : \|w\|_0 \leq t, \|w\|_2 \leq 1\}$ be the set of all possible true signals w^*. It is known that the VC dimension of the set K is $t \log d$, therefore there are $O((\sum_i m_i)^{t \log d})$ possible labeling that can be produced by some $w^* \in K'$. Moreover, because $\sum_{i \leq k} m_i = \Theta(\gamma_k m)$. Therefore, the adversary can corrupt a $\frac{\nu}{\gamma_k}$ fraction of the $\sum_{i \leq k} m_i$ samples. This is in the worst case, $(\frac{\nu}{\epsilon})\frac{t}{\epsilon^3}\text{polylog}(d, \frac{1}{\epsilon}, \frac{1}{\delta}))$. Let $m' = \sum_{i \leq k} m_i$. By taking the union bound over choices of w^* and $\frac{\nu}{\epsilon}m'$ corrupted points, we have

$$\begin{aligned}
\Pr\left[\sup_{w^*, \{y_k\}_{k=1}^{m_k}, w} \left|\mathbf{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \geq \lambda/12\right] &\leq \exp\left(-\frac{m_k s^2}{2U^2}\right) c' m'^{t \log d} \binom{m'}{\frac{\nu}{\epsilon}m'} \\
&\leq c \exp\left(-\frac{m_k s^2}{2U^2} + t \log d \log(m') + \frac{\nu}{\epsilon}\log(\frac{\epsilon}{\nu})m'\right) \\
&\leq c \exp\left(-\frac{m_k s^2}{2U^2} + t \log d \log(m') + \frac{\nu}{\epsilon}\log(\frac{\epsilon}{\nu})\log(\frac{1}{\epsilon})m_k\right) \\
&\leq \exp\left(-O(\frac{m_k}{\log \frac{m_k}{\gamma_k \delta}})\right)
\end{aligned}$$

where the last inequality follows from $\nu \in O(\epsilon/\log(\frac{d}{\epsilon})^2)$. Therefore, with probability at least $1 - \delta$,

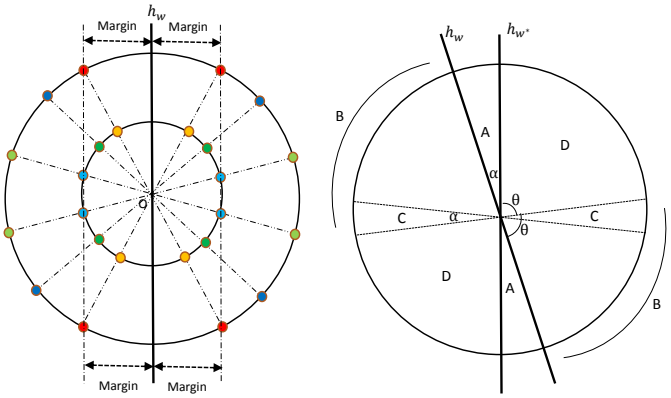
$$\sup_{w^*, \{y_k\}_{k=1}^{m_k}, w} \left|\mathbf{E}_{(x,y)\sim\tilde{D}_k}[\ell_{\tau_k}(w,x,y)] - \ell_{\tau_k}(w,T')\right| \leq \lambda/12.$$

□

Proofs of Theorem 1

Theorem 1 (restated). *For every bounded noise parameter $0 \leq \beta < 1$, there exists a distribution $\tilde{D}_\beta \in \mathcal{P}_\beta$ (that is, a distribution over $\mathbb{R}^2 \times \{+1, -1\}$, where the marginal distribution on $\mathbb{R}^2$ is uniform over the unit ball, and the labels $\{+1, -1\}$ satisfies β-bounded noise condition) such that any proper loss minimization is not consistent on $\tilde{D}_\beta$ w.r.t. the class of halfspaces. That is, there exists an $\epsilon \geq 0$ and a sample size $m(\epsilon)$ such that any proper loss minimization will output a classifier of excess error larger than ϵ by a high probability over sample size at least $m(\epsilon)$.*

Proof. We prove the theorem by constructing a distribution $\tilde{D}_\beta \in \mathcal{P}_\beta$ that is consistent with our conclusion. Since we have assumed that the marginal distribution over the instance space X is uniform over the unit ball, we now construct a noisy distribution on the label space for our purpose. To do so, given the Bayes optimal classifier h_{w^*} and some linear classifier h_w such that $\theta(h_{w^*}, h_w) = \alpha$, we first divide the instance space X into four areas A, B, C, and D, as shown in Figure 2.1(b). Namely, area A is the disagreement region between h_w and h_{w^*} with angle α, and



(a) Property 1. (b) Partition.

Figure 2.1: Demonstrating the construction for the lower bound.

the agreement region consists of areas B (points closer to h_w) and D (points closer to h_{w^*}). Area C, a wedge with an angle of α, is a part of area B. We flip the labels of all points in areas A and B with probability $\eta = (1-\beta)/2$, and retain the original labels of instances in area D. This setting naturally satisfies β-bounded noise condition. As we will show later, when the angle α is small enough, the expected value of proper loss of h_w over the whole instance space will be smaller than that of h_{w^*}. Then by the standard analysis of [15], we conclude that there exists an $\epsilon \geq 0$ and a sample size $m(\epsilon)$ such that any proper loss minimization will output a classifier of excess error larger than ϵ by a high probability over sample size at least $m(\epsilon)$.

We now show the key steps in our analysis. We consider here unit vectors w^* and w. Let cA, cB, cC, and cD be the proper loss of h_{w^*} on areas A, B, C, and D when the labels are correct, and let dA, dB, dC, and dD be the loss of h_{w^*} on areas A, B, C, and D when the labels are incorrect. By the symmetry property 1 in Definition 1, we have

$$cA = \frac{2}{\pi}\int_0^{\alpha}\int_0^1 \ell_+^{w^*}(z,\varphi)zdzd\varphi. \tag{2.11}$$

Similarly, we can calculate cB, cC, cD, dA, dB, dC, dD, and can check that

$$cA + cB = \frac{2}{\pi}\int_0^{\frac{\pi+\alpha}{2}}\int_0^1 \ell_+^{w^*}(z,\varphi)zdzd\varphi = cC + cD, \tag{2.12}$$

$$dA + dB = \frac{2}{\pi}\int_0^{\frac{\pi+\alpha}{2}}\int_0^1 \ell_-^{w^*}(z,\varphi)zdzd\varphi = dC + dD. \tag{2.13}$$

On the other side, according to the noisy distribution $\tilde{D}$ designed by us, the expected loss of h_{w^*} is

$$\mathcal{L}(h_{w^*}) = \eta(dA + dB) + (1-\eta)(cA + cB) + cD. \tag{2.14}$$

For h_w, as the role of B to h_w is the same as the role D to h_{w^*} by Property 1 in Definition 1, we have

$$\mathcal{L}(h_w) = \eta(cA + dD) + (1-\eta)(dA + cD) + cB. \tag{2.15}$$

Therefore, combining with Equations 2.12 and 2.13, we have

$$\mathcal{L}(h_w) - \mathcal{L}(h_{w^*}) = (1-\eta)(dA - cA) - \eta(dC - cC). \tag{2.16}$$

That is to say, once $\eta > \eta(\alpha) \triangleq \frac{dA-cA}{dA-cA+dC-cC}$, we will have $\mathcal{L}(h_w) < \mathcal{L}(h_{w^*})$. We now show that $\frac{dA-cA}{dA-cA+dC-cC}$ can be arbitrarily small when α approaches to zero, i.e., $\lim_{\alpha\to 0} \frac{dA-cA}{dA-cA+dC-cC} = 0$. To see this, let $f_{w^*}(z,\varphi) = \ell_-^{w^*}(z,\varphi) - \ell_+^{w^*}(z,\varphi)$, then

$$\begin{aligned}
&\lim_{\alpha\to 0} \frac{dA - cA}{dA - cA + dC - cC} \\
&= \lim_{\alpha\to 0} \frac{\frac{2}{\pi}\int_0^\alpha \int_0^1 f_{w^*}(z,\varphi) z dz d\varphi}{\frac{2}{\pi}\int_0^\alpha \int_0^1 f_{w^*}(z,\varphi) z dz d\varphi + \frac{4}{\pi}\int_{\frac{\pi-\alpha}{2}}^{\frac{\pi}{2}} \int_0^1 f_{w^*}(z,\varphi) z dz d\varphi} \\
&= \lim_{\alpha\to 0} \frac{\frac{2}{\pi}\int_0^1 f_{w^*}(z,\alpha) z dz}{\frac{2}{\pi}\int_0^1 f_{w^*}(z,\alpha) z dz + \frac{2}{\pi}\int_0^1 f_{w^*}(z,\frac{\pi-\alpha}{2}) z dz} && \text{(By L'Hospital's rule)} \\
&= \frac{\lim_{\alpha\to 0} \frac{2}{\pi}\int_0^1 f_{w^*}(z,\alpha) z dz}{\lim_{\alpha\to 0} \frac{2}{\pi}\int_0^1 f_{w^*}(z,\alpha) z dz + \frac{2}{\pi}\int_0^1 f_{w^*}(z,\frac{\pi-\alpha}{2}) z dz} && \text{(By existence of the limit)} \\
&= \frac{\int_0^1 f_{w^*}(z,0) z dz}{\int_0^1 f_{w^*}(z,0) z dz + \int_0^1 f_{w^*}(z,\frac{\pi}{2}) z dz} && \left(\text{By continuity of } \int_0^1 f_{w^*}(z,\alpha) z dz\right) \\
&= \frac{0}{0 + \int_0^1 f_{w^*}(z,\frac{\pi}{2}) z dz} \\
&= 0. && \left(\text{Since } \int_0^1 f_{w^*}\left(z,\frac{\pi}{2}\right) z dz > 0, \text{see Lemma 18}\right)
\end{aligned} \tag{2.17}$$

The following lemma guarantees the denominator of the last equation is non-zero:

Lemma 18. *For any continuous function $f_{w^*}(z,\varphi)$, we have*

$$\int_0^1 f_{w^*}\left(z,\frac{\pi}{2}\right) z dz > 0. \tag{2.18}$$

Proof. In the close interval $[1/2, 1]$, since function $f_{w^*}(z,\frac{\pi}{2})$ is continuous, by extreme value theorem, there exists $\xi \in [1/2,1]$ such that $\min_z f_{w^*}(z,\frac{\pi}{2}) z = f_{w^*}(\xi,\frac{\pi}{2})\xi > 0$ (By Property 2 in Definition 1). So

$$\begin{aligned}
\int_0^1 f_{w^*}\left(z,\frac{\pi}{2}\right) z dz &= \int_0^{\frac{1}{2}} f_{w^*}\left(z,\frac{\pi}{2}\right) z dz + \int_{\frac{1}{2}}^1 f_{w^*}\left(z,\frac{\pi}{2}\right) z dz \\
&\geq \int_{\frac{1}{2}}^1 f_{w^*}\left(z,\frac{\pi}{2}\right) z dz \\
&\geq \frac{1}{2}\min_z f_{w^*}\left(z,\frac{\pi}{2}\right) z \\
&\geq \frac{1}{2} f_{w^*}\left(\xi,\frac{\pi}{2}\right)\xi \\
&> 0.
\end{aligned} \tag{2.19}$$

□

This completes our proof.

□

2.2 Adaptive Compressed Sensing

2.2.1 Introduction

Compressed sensing, also known as sparse recovery, is a central object of study in data stream algorithms, with applications to monitoring network traffic [111], analysis of genetic data [127, 210], and many other domains [174]. The problem can be stated as recovering an underlying signal $x \in \mathbb{R}^n$ from *measurements* $A_1 \cdot x, ..., A_m \cdot x$ with the C-approximate ℓ_p/ℓ_q recovery guarantee being

$$\|x - \hat{x}\|_p \leq C \min_{k\text{-sparse } x'} \|x - x'\|_q, \tag{2.20}$$

where the A_i are drawn from a distribution and $m \ll n$. The focus of this work is on *adaptive compressed sensing*, in which the measurements are chosen in rounds, and the choice of measurement in each round depends on the outcome of the measurements in previous rounds.

Adaptive compressed sensing has been studied in a number of different works [4, 59, 112, 113, 121, 124, 168, 188] in theoretical computer science, machine learning, image processing, and many other domains [17, 121, 188]. In theoretical computer science and machine learning, adaptive compressed sensing serves as an important tool to obtain sublinear algorithms for active learning in both time and space [17, 92, 121, 188]. In image processing, the study of adaptive compressed sensing has led to compressed acquisition of sequential images with various applications in celestial navigation and attitude determination [101].

Despite a large amount of works on adaptive compressed sensing, the power of adaptivity remains a long-standing open problem. Indyk, Price, and Woodruff [121] were the first to show that without any assumptions on the signal x, one can obtain a number m of measurements which is a $\log(n)/\log\log(n)$ factor smaller than what can be achieved in the non-adaptive setting. Specifically, for $p = q = 2$ and $C = 1 + \epsilon$, they show that $m = \mathcal{O}(\frac{k}{\epsilon}\log\log(n))$ measurements suffice to achieve guarantee (2.20), whereas it is known that any non-adaptive scheme requires $k = \Omega(\frac{k}{\epsilon}\log(\frac{n}{k}))$ measurements, provided $\epsilon > \sqrt{\frac{k\log n}{n}}$ (Theorem 4.4 of [187], see also [18]). Improving the sample complexity as much as possible is desired, as it might correspond to, e.g., the amount of radiation a hospital patient is exposed to, or the amont of time a patient must be present for diagnosis.

The ℓ_1/ℓ_1 problem was studied in [187], for which perhaps surprisingly, a better dependence on ϵ was obtained than is possible for ℓ_2/ℓ_2 schemes. Still, the power of adaptivity for the ℓ_1/ℓ_1 recovery problem over its non-adaptive counterpart has remained unclear. An $O(\frac{k}{\sqrt{\epsilon}}\log n\log^3(\frac{1}{\epsilon}))$ non-adaptive bound was shown in [187], while an adaptive lower bound of $\Omega(\frac{k}{\sqrt{\epsilon}}/\log\frac{k}{\sqrt{\epsilon}})$ was shown in [188]. Recently several works [173, 212] have looked at other values of p and q, even those for which $0 < p, q < 1$, which do not correspond to normed spaces. The power of adaptivity for such error measures is also unknown.

Table 2.1: The sample complexity of adaptive compressed sensing. Results without any citation given correspond to our new results.

C, Guarantees	Upper Bounds	Rounds	Lower Bounds
$1+\epsilon$, ℓ_1/ℓ_1	$\mathcal{O}(\frac{k}{\sqrt{\epsilon}}\text{loglog}(n)\log^{\frac{5}{2}}(\frac{1}{\epsilon}))$	$\mathcal{O}(\text{loglog}(n))$	$\Omega(\frac{k}{\sqrt{\epsilon}\log(k/\sqrt{\epsilon})})$ [188]
$1+\epsilon$, ℓ_p/ℓ_p	$\mathcal{O}(\frac{k}{\epsilon^{p/2}}\text{loglog}(n)\text{poly}(\log(\frac{1}{\epsilon})))$	$\mathcal{O}(\text{loglog}(n))$	$\Omega(\frac{k}{\epsilon^{p/2}}\frac{1}{\log^2(k/\epsilon)})$
$\sqrt{\frac{1}{k}}$, ℓ_∞/ℓ_2	$\mathcal{O}(k\text{loglog}(n)+k\log(k))$	$\mathcal{O}(\text{loglog}(n))$	-
$1+\epsilon, \ell_2/\ell_2$	$\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(\frac{n\epsilon}{k}))$ [121] $\mathcal{O}(k\text{loglog}(\frac{n}{k})+\frac{k}{\epsilon}\text{loglog}(\frac{1}{\epsilon}))$ $\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(\frac{n\log(n\epsilon)}{k}))$	$\mathcal{O}(\log^*(k)\text{loglog}(\frac{n\epsilon}{k}))$ [121] $\mathcal{O}(\log^*(k)\text{loglog}(\frac{n}{k}))$ $\mathcal{O}(\text{loglog}(n\log(\frac{n\epsilon}{k}))$	$\Omega(\frac{k}{\epsilon}+\text{loglog}(n))$ [188]

2.2.2 Our results on optimization and sample efficiency

Our work studies the problem of adaptive compressed sensing by providing affirmative answers to the above-mentioned open questions. We improve over the best known results for $p=q=2$, and then provide novel adaptive compressed sensing guarantees for $0<p=q<2$ for every p and q. See Table 2.1 for a comparison of results.

For ℓ_1/ℓ_1, we design an adaptive algorithm which requires only $\mathcal{O}(\frac{k}{\sqrt{\epsilon}}\text{loglog}(n)\log^{\frac{5}{2}}(\frac{1}{\epsilon}))$ measurements for the ℓ_1/ℓ_1 problem. More generally, we study the ℓ_p/ℓ_p problem for $0<p<2$. One of our main theorems is the following.

Theorem 9 (ℓ_p/ℓ_p Recovery Upper Bound)**.** *Let $x\in\mathbb{R}^n$ and $0<p<2$. There exists a randomized algorithm that performs $\mathcal{O}(\frac{k}{\epsilon^{p/2}}\text{loglog}(n)\text{poly}(\log(\frac{1}{\epsilon})))$ adaptive linear measurements on x in $\mathcal{O}(\text{loglog}(n))$ rounds, and with probability $2/3$, returns a vector $\hat{x}\in\mathbb{R}^n$ such that $\|x-\hat{x}\|_p\le(1+\epsilon)\|x_{-k}\|_p$.*

Theorem 9 improves the previous sample complexity upper bound for the case of $C=1+\epsilon$ and $p=q=1$ from $\mathcal{O}(\frac{k}{\sqrt{\epsilon}}\log(n)\log^3(\frac{1}{\epsilon}))$ to $\mathcal{O}(\frac{k}{\sqrt{\epsilon}}\text{loglog}(n)\log^{\frac{5}{2}}(\frac{1}{\epsilon}))$. Compared with the non-adaptive $(1+\epsilon)$-approximate ℓ_1/ℓ_1 upper bound of $\mathcal{O}(\frac{k}{\sqrt{\epsilon}}\log(n)\log^3(\frac{1}{\epsilon}))$, we show that adaptivity exponentially improves the sample complexity w.r.t. the dependence on n over non-adaptive algorithms while retaining the improved dependence on ϵ of non-adaptive algorithms. Furthermore, Theorem 9 extends the working range of adaptive compressed sensing from $p=1$ to general values of $p\in(0,2)$.

We also state a complementary lower bound to formalize the hardness of the above problem.

Theorem 10 (ℓ_p/ℓ_p Recovery Lower Bound)**.** *Fix $0<p<2$, any $(1+\epsilon)$-approximate ℓ_p/ℓ_p recovery scheme with sufficiently small constant failure probability must make $\Omega(\frac{k}{\epsilon^{p/2}}/\log^2(\frac{k}{\epsilon}))$ measurements.*

Theorem 10 shows that our upper bound in Theorem 9 is tight up to the $\log(k/\epsilon)$ factor.

We also study the case when $p\neq q$. In particular, we focus on the case when $p=\infty, q=2$ and $C=\sqrt{\frac{1}{k}}$, as in the following theorem.

Theorem 11 (ℓ_∞/ℓ_2 Recovery Upper Bound)**.** *Let $x\in\mathbb{R}^n$. There exists a randomized algorithm*

that performs $\mathcal{O}(k\log(k)+k\text{loglog}(n))$ linear measurements on x in $\mathcal{O}(\text{loglog}(n))$ rounds, and with probability $1-1/\text{poly}(k)$ returns a vector $\hat{x}$ such that $\|x-\hat{x}\|_\infty^2 \leq \frac{1}{k}\|x_{-k}\|_2^2$, where $x_{-k} \in \mathbb{R}^n$ is the vector with the largest $n-k$ coordinates (in the sense of absolute value) being zeroed out.

We also provide an improved result for $(1+\epsilon)$-approximate ℓ_2/ℓ_2 problems.

Theorem 12 (ℓ_2/ℓ_2 Sparse Recovery Upper Bounds)**.** *Let $x \in \mathbb{R}^n$. There exists a randomized algorithm that*

- *uses $\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(\frac{1}{\epsilon})+k\text{loglog}(\frac{n}{k}))$ linear measurements on x in $\mathcal{O}(\text{loglog}(\frac{n}{k})\cdot\log^*(k))$ rounds;*
- *uses $\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(\frac{n\log(n\epsilon)}{k}))$ linear measurements on x in $\mathcal{O}(\text{loglog}(\epsilon n\log(\frac{n}{k})))$ rounds;*

and with constant probability returns a vector $\hat{x}$ such that $\|x-\hat{x}\|_2 \leq (1+\epsilon)\|x_{-k}\|_2$.

Previously the best known tradeoff was $\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(\frac{n\epsilon}{k}))$ samples and $\mathcal{O}(\log^*(k)\text{loglog}(\frac{n\epsilon}{k}))$ rounds for $(1+\epsilon)$-approximation for the ℓ_2/ℓ_2 problem [121]. Our result improves both the sample complexity (the first result) and the number of rounds (the second result). We summarize our results in Table 2.1.

2.2.3 Our techniques

ℓ_∞/ℓ_2 sparse recovery. Our ℓ_∞/ℓ_2 sparse recovery scheme hashes every $i \in [n]$ to $\text{poly}(k)$ buckets, and then proceeds by finding all the buckets that have ℓ_2 mass at least $\Omega(\frac{1}{\sqrt{k}}\|x_{-\Omega(k)}\|_2)$. We then find a set of buckets that contain all heavy coordinates, which are isolated from each other due to hashing. Then, we run a 1-sparse recovery in each bucket in parallel in order to find all the heavy coordinate. However, since we have $\mathcal{O}(k)$ buckets, we cannot afford to take a union bound over all one-sparse recovery routines called. Instead, we show that most buckets succeed and hence we can substract from x the elements returned, and then run a standard CountSketch algorithm to recover everything else. This algorithm obtains an optimal $\mathcal{O}(\text{loglog}(n))$ number of rounds and $\mathcal{O}(k\log(k)+k\text{loglog}(n))$ number of measurements, while succeeding with probability at least $1-1/\text{poly}(k)$.

We proceed by showing an algorithm for ℓ_2/ℓ_2 sparse recovery with $\mathcal{O}(\frac{k}{\epsilon}\text{loglog}(n))$ measurements and $\mathcal{O}(\text{loglog}(n))$ rounds. This will be important for our more general ℓ_p/ℓ_p scheme, saving a $\log^*(k)$ factor from the number of rounds, achieving optimality with respect to this quantity. For this scheme, we utilize the ℓ_∞/ℓ_2 scheme we just developed, observing that for small $k < \mathcal{O}(\log(n))$, the measurement complexity is $\mathcal{O}(k\text{loglog}(n))$. The algorithm hashes to $k/(\epsilon\log(n))$ buckets, and in each bucket runs ℓ_∞/ℓ_2 with sparsity k/ϵ. The ℓ_∞/ℓ_2 algorithm in each bucket succeeds with probability $1-1/\text{polylog}(n))$; this fact allows us to argue that all but a $1/\text{polylog}(n)$ fraction of the buckets will succeed, and hence we can recover all but a $k/\text{polylog}(n))$ fraction of the heavy coordinates. The next step is to subtract these coordinates from our initial vector, and then run a standard ℓ_2/ℓ_2 algorithm with decreased sparsity.

ℓ_p/ℓ_p sparse recovery. Our ℓ_p/ℓ_p scheme, $0 < p < 2$, is based on carefully invoking several ℓ_2/ℓ_2 schemes with different parameters. We focus our discussion on $p = 1$, then mention extensions to general p. A main difficulty of adapting the ℓ_1/ℓ_1 scheme of [187] is that it relies upon an ℓ_∞/ℓ_2 scheme, and all known schemes, including ours, have at least a $k\log k$ dependence on the number of measurements, which is too large for our overall goal.

A key insight in [187] for ℓ_1/ℓ_1 is that since the output does not need to be exactly k-sparse, one can compensate for mistakes on approximating the top k entries of x by accurately outputting enough smaller entries. For example, if $k = 1$, consider two possible signals $x = (1, \epsilon, \ldots, \epsilon)$ and $x' = (1 + \epsilon, \epsilon, \ldots, \epsilon)$, where ϵ occurs $1/\epsilon$ times in both x and x'. One can show, using known lower bound techniques, that distinguishing x from x' requires $\Omega(1/\epsilon)$ measurements. Moreover, $x_1 = (1, 0, \ldots, 0)$ and $x'_1 = (1 + \epsilon, 0, \ldots, 0)$, and any 1-sparse approximation to x or x' must therefore distinguish x from x', and so requires $\Omega(1/\epsilon)$ measurements. An important insight though, is that if one does not require the output signal y to be 1-sparse, then one can output $(1, \epsilon, 0, \ldots, 0)$ in both cases, without actually distinguishing which case one is in!

As another example, suppose that $x = (1, \epsilon, \ldots, \epsilon)$ and $x' = (1 + \epsilon^c, \epsilon, \ldots, \epsilon)$ for some $0 < c < 1$. In this case, one can show that one needs $\Omega(1/\epsilon^c)$ measurements to distinguish x and x', and as before, to output an exactly 1-sparse signal providing a $(1 + \epsilon)$-approximation requires $\tilde{\Theta}(1/\epsilon^c)$ measurements. In this case if one outputs a signal y with $y_1 = 1$, one cannot simply find a single other coordinate ϵ to "make up" for the poor approximation on the first coordinate. However, if one were to output $1/\epsilon^{1-c}$ coordinates each of value ϵ, then the ϵ^c "mass" lost by poorly approximating the first coordinate would be compensated for by outputting $\epsilon \cdot 1/\epsilon^{1-c} = \epsilon^c$ mass on these remaining coordinates. It is not clear how to find such remaining coordinates though, since they are much smaller; however, if one randomly subsamples an ϵ^c fraction of coordinates, then roughly $1/\epsilon^{1-c}$ of the coordinates of value ϵ survive and these could all be found with a number of measurements proportional to $1/\epsilon^{1-c}$. Balancing the two measurement complexities of $1/\epsilon^c$ and $1/\epsilon^{1-c}$ at $c = 1/2$ gives roughly the optimal $1/\epsilon^{1/2}$ dependence on ϵ in the number of measurements.

To extend this to the adaptive case, a recurring theme of the above examples is that the top k, while they need to be found, they do not need to be approximated very accurately. Indeed, they do need to be found, if, e.g., the top k entries of x were equal to an arbitrarily large value and the remaining entries were much smaller. We accomplish this by running an ℓ_2/ℓ_2 scheme with parameters $k' = \Theta(k)$ and $\epsilon' = \Theta(\sqrt{\epsilon})$, as well as an ℓ_2/ℓ_2 scheme with parameters $k' = \Theta(k/\sqrt{\epsilon})$ and $\epsilon' = \Theta(1)$ (up to logarithmic factors in $1/\epsilon$). Another theme is that the mass in the smaller coordinates we find to compensate for our poor approximation in the larger coordinates also does not need to be approximated very well, and we find this mass by subsampling many times and running an ℓ_2/ℓ_2 scheme with parameters $k' = \Theta(1)$ and $\epsilon' = \Theta(1)$. This technique is surprisingly general, and does not require the underlying error measure we are approximating to be a norm. It just uses scale-invariance and how its rate of growth compares to that of the ℓ_2-norm.

ℓ_2/ℓ_2 sparse recovery. Our last algorithm, which concerns ℓ_2/ℓ_2 sparse recovery, achieves $\mathcal{O}(k\mathrm{loglog}(n) + \frac{k}{\epsilon}\mathrm{loglog}(1/\epsilon))$ measurements, showing that ϵ does not need to multiply $\mathrm{loglog}(n)$. The key insight lies in first solving the 1-sparse recovery task with $\mathcal{O}(\mathrm{loglog}(n) + \frac{1}{\epsilon}\mathrm{loglog}(1/\epsilon))$ measurements, and then extending this to the general case. To achieve this, we hash to $\mathrm{polylog}(1/\epsilon)$ buckets, then solve ℓ_2/ℓ_2 with constant sparsity on a new vector, where coordinate j equals the ℓ_2 norm of the jth bucket; this steps requires only $\mathcal{O}(\frac{1}{\epsilon}\mathrm{loglog}(1/\epsilon))$ measurements. Now, we can run standard 1-sparse recovery in each of these buckets returned. Extending this idea to the general case follows by plugging this sub-routine in the iterative algorithm of [121], while ensuring that sub-sampling does not increase the number of measurements. For that we also need to sub-sample at a slower rate, slower roughly by a factor of ϵ.

Notation: For a vector $x \in \mathbb{R}^n$, we define $H_k(x)$ to be the set of its largest k coordinates in absolute value. For a set S, denote by x_S the vector with every coordinate $i \notin S$ being zeroed out. We also define $x_{-k} = x_{[n]\setminus H_k(x)}$ and $H_{k,\epsilon}(x) = \{i \in [n] : |x_i| \geq \frac{\epsilon}{k}\|x_{-k}\|_2^2\}$, where $[n]$ represents the set $\{1, 2, ..., n\}$. For a set S, let $|S|$ be the cardinality of S.

2.2.4 Proofs of our main results

Proofs of Theorem 9

This section is devoted to proving Theorem 9. Our algorithm for ℓ_p/ℓ_p recovery is in Algorithm 16.

Let $f = \epsilon^{p/2}$, $r = 2/(p\log(1/f))$ and $q = \max\{p - \frac{1}{2}, 0\} = (p - \frac{1}{2})^+$. We will invoke the following ℓ_2/ℓ_2 oracle frequently throughout the paper.

Oracle 1 ($\text{ADAPTIVESPARSERECOVERY}_{\ell_p/\ell_q}(x, k, \epsilon)$)**.** *The oracle is fed with (x, k, ϵ) as input parameters, and outputs a set of coordinates $i \in [n]$ of size $\mathcal{O}(k)$ which corresponds to the support of vector $\hat{x}$, where $\hat{x}$ can be any vector for which $\|x - \hat{x}\|_p \leq (1+\epsilon)\min_{\mathcal{O}(k)\text{-sparse } x'} \|x - x'\|_q$.*

Existing algorithms can be applied to construct Oracle 1 for the ℓ_2/ℓ_2 case, such as [121]. Without loss of generality, we assume that the coordinates of x are ranked in decreasing value, i.e., $x_1 \geq x_2 \geq \cdots \geq x_n$.

Algorithm 5 Adaptive ℓ_p/ℓ_p Recovery

1: $A \leftarrow \text{ADAPTIVESPARSERECOVERY}_{\ell_2/\ell_2}(x, 2k/f, 1/10)$.
2: $B \leftarrow \text{ADAPTIVESPARSERECOVERY}_{\ell_2/\ell_2}(x, 4k, f/r^2)$.
3: $S \leftarrow A \cup B$.
4: **For** $j = 1 : r$
5: Uniformly sample the entries of x with probability $2^{-j}f/k$ for $k/(2f(r+1)^q)$ times.
6: Run the adaptive $\text{ADAPTIVESPARSERECOVERY}_{\ell_2/\ell_2}(x, 2, 1/(4(r+1))^{\frac{2}{p}})$ algorithm on each of the $k/(2f(r+1)^q)$ subsamples to obtain sets $A_{j,1}, A_{j,2}, \ldots, A_{j,k/(2f(r+1)^q)}$.
7: Let $S_j \leftarrow \cup_{t=1}^{k/(2f(r+1)^q)} A_{j,t} \setminus \cup_{t=0}^{j-1} S_t$.
8: **End For**
9: Request the entries of x with coordinates $S_0, ..., S_r$.
10: **Output:** $\hat{x} = x_{S_0 \cup \cdots \cup S_r}$.

Lemma 19. *Suppose we subsample x with probability p and let y be the subsampled vector formed from x. Then with failure probability $e^{-\Omega(k)}$, $\|y_{-2k}\|_2 \leq \sqrt{2p}\left\|x_{-k/p}\right\|_2$.*

Proof. Let T be the set of coordinates in the subsample. Then $\mathbb{E}\left[\left|T \cap \left[\frac{3k}{2p}\right]\right|\right] = \frac{3k}{2}$. So by the Chernoff bound, $\Pr\left[\left|T \cap \left[\frac{3k}{2p}\right]\right| > 2k\right] \leq e^{-\Omega(k)}$. Thus $\left|T \cap \left[\frac{3k}{2p}\right]\right| \leq 2k$ holds with high probability. Let $Y_i = x_i^2$ if $i \in T$ $Y_i = 0$ if $i \in [n] \setminus T$. Then $\mathbb{E}\left[\sum_{i > \frac{3k}{2p}} Y_i\right] = p\left\|x_{-\frac{3k}{2p}}\right\|_2^2 \leq p\left\|x_{-k/p}\right\|_2^2$. Notice that there are at least $\frac{k}{2p}$ elements in $x_{-k/p}$ with absolute value larger than $\left|x_{\frac{3k}{2p}}\right|$. Thus for $i > \frac{3k}{2p}$, $Y_i \leq \left|x_{\frac{3k}{2p}}\right|^2 \leq \frac{2p}{k}\left\|x_{-k/p}\right\|_2^2$. Again by a Chernoff

bound, Pr $\left[\sum_{i>\frac{3k}{2p}} Y_i \geq \frac{4p}{3}\left\|x_{-k/p}\right\|_2^2\right] \leq e^{-\Omega(k)}$. Conditioned on the latter event not happening, $\|y_{-2k}\|_2^2 \leq \sum_{i>\frac{3k}{2p}} Y_i \leq \frac{4p}{3}\left\|x_{-k/p}\right\|_2^2 \leq 2p\left\|x_{-k/p}\right\|_2^2$. By a union bound, with failure probability $e^{-\Omega(k)}$, we have $\|y_{-2k}\|_2 \leq \sqrt{2p}\left\|x_{-k/p}\right\|_2$. □

Lemma 20. *Let $\hat{x}$ be the output of the ℓ_2/ℓ_2 scheme on x with parameters $(k, \epsilon/2)$. Then with small constant failure probability, $\left\|x_{[k]}\right\|_p^p - \|\hat{x}\|_p^p \leq k^{1-\frac{p}{2}}\epsilon^{\frac{p}{2}}\|x_{-k}\|_2^p$.*

Proof. Notice that with small constant failure probability, the ℓ_2/ℓ_2 guarantee holds and we have

$$\left\|x_{[k]}\right\|_2^2 - \|\hat{x}\|_2^2 = \|x-\hat{x}\|_2^2 - \|x_{-k}\|_2^2 \leq (1+\epsilon)\|x_{-k}\|_2^2 - \|x_{-k}\|_2^2 = \epsilon\|x_{-k}\|_2^2.$$

Let $S \subset [n]$ be such that $x_S = \hat{x}$, and define $y = x_{[k]\setminus S}$, $z = x_{S\setminus[k]}$. Then if $\|y\|_p^p \leq k^{1-\frac{p}{2}}\epsilon^{\frac{p}{2}}\|x_{-k}\|_2^p$ we are done. Otherwise, let $1 \leq k' \leq k$ denote the size of $[k]\setminus S$, and define $c = \|y\|_2/\sqrt{k'}$.

$$\begin{aligned}\left\|x_{[k]}\right\|_p^p - \|\hat{x}\|_p^p = \|y\|_p^p - \|z\|_p^p \leq k'^{1-\frac{p}{2}}\|y\|_2^p - \|z\|_p^p = \frac{\|y\|_2^2}{c^{2-p}} - \|z\|_p^p \\ \leq \frac{\|y\|_2^2 - \|z\|_2^2}{c^{2-p}} = \frac{\left\|x_{[k]}\right\|_2^2 - \|\hat{x}\|_2^2}{c^{2-p}} \leq \frac{\epsilon\|x_{-k}\|_2^2}{c^{2-p}}.\end{aligned}$$

Since $c \geq \frac{\|y\|_p}{k'^{\frac{1}{p}}} \geq \frac{\|y\|_p}{k^{\frac{1}{p}}} \geq \sqrt{\frac{\epsilon}{k}}\|x_{-k}\|_2$, we have $\left\|x_{[k]}\right\|_p^p - \|\hat{x}\|_p^p \leq k^{\frac{2-p}{2}}\epsilon^{1-\frac{2-p}{2}}\|x_{-k}\|_2^{2-(2-p)} = k^{1-\frac{p}{2}}\epsilon^{\frac{p}{2}}\|x_{-k}\|_2^p$. □

Theorem 13. *Fix $0 < p < 2$. For $x \in \mathbb{R}^n$, there exists a $(1+\epsilon)$-approximation algorithm that performs $\mathcal{O}(\frac{k}{\epsilon^{p/2}}\log\log(n)\log^{\frac{2}{p}+1-(p-\frac{1}{2})^+}(\frac{1}{\epsilon}))$ adaptive linear measurements in $\mathcal{O}(\log\log(n))$ rounds, and with probability at least $2/3$, we can find a vector $\hat{x} \in \mathbb{R}^n$ such that*

$$\|x-\hat{x}\|_p \leq (1+\epsilon)\|x_{-k}\|_p. \tag{2.21}$$

Proof. The algorithm is stated in Algorithm 16. We first consider the difference $\left\|x_{[k]}\right\|_p^p - \|x_{S_0}\|_p^p$. Let $i^*(0)$ be the smallest integer such that for any $l > i^*(0)$, $|x_l| \leq \|x_{-2k/f}\|_2/\sqrt{k}$.

Case 1. $i^*(0) > 4k$

Then for all $k < j \leq 4k$, we have $|x_j| > \|x_{-2k/f}\|_2/\sqrt{k}$. Hence x_{S_0} must contain at least $1/2$ of these indices; if not, the total squared loss is at least $1/2\cdot 3k\|x_{-2k/f}\|_2^2/k \geq (3/2)\|x_{-2k/f}\|_2^2$, a contradiction to $\epsilon' = 1/10$. It follows that $\|x_{S_0\cap\{k+1,\ldots,4k\}}\|_p^p \geq \frac{3}{2}k\left[\frac{\|x_{-2k/f}\|_2}{\sqrt{k}}\right]^p = \frac{3}{2}k^{1-\frac{p}{2}}\|x_{-2k/f}\|_2^p$. On the other hand, $\left\|x_{[k]}\right\|_p^p - \|x_{S_0}\|_p^p$ is at most $1.1k^{1-\frac{p}{2}}\|x_{-2k/f}\|_2^p$, since by the ℓ_2/ℓ_2 guarantee

$$\|x_{[k]}\|_p^p - \|x_{S_0\cap[k]}\|_p^p \leq k^{1-\frac{p}{2}}\|x_{[k]} - x_{S_0\cap[k]}\|_2^p \leq k^{1-\frac{p}{2}}\|x - x_{S_0}\|_2^p \leq \frac{11}{10}k^{1-\frac{p}{2}}\|x_{-2k/f}\|_2^p.$$

It follows that $\|x_{[k]}\|_p^p - \|x_{S_0}\|_p^p = \|x_{[k]}\|_p^p - \|x_{S_0\cap[k]}\|_p^p - \|x_{S_0\cap\{k+1,\ldots,4k\}}\|_p^p \leq \frac{11}{10}k^{1-\frac{p}{2}}\|x_{-2k/f}\|_2^p - \frac{3}{2}k^{1-\frac{p}{2}}\|x_{-2k/f}\|_2^p \leq 0$.

Case 2. $i^*(0) \le 4k$, and $\sum_{j=i^*(0)+1}^{2k/f} x_j^2 \ge 4\|x_{-2k/f}\|_2^2$.
We claim that x_{S_0} must contain at least a $5/8$ fraction of coordinates in $\{i^*(0)+1, ..., 2k/f\}$; if not, then the cost for missing at least a $3/8$ fraction of the ℓ_2-norm of $x_{\{i^*(0)+1,...,2k/f\}}$ will be at least $(3/2)\|x_{-2k/f}\|_2^2$, contradicting the ℓ_2/ℓ_2 guarantee. Since all coordinates x_j's for $j > i^*(0)$ have value at most $\|x_{-2k/f}\|_2/\sqrt{k}$, it follows that the p-norm of coordinates corresponding to $\{i^*(0) + 1, ..., 2k/f\} \cap S_0$ is at least $\left\|x_{\{i^*(0)+1,...,2k/f\}\cap S_0}\right\|_p^p \ge \frac{5}{2}k^{\frac{2-p}{2}} \frac{\|x_{-2k/f}\|_2^2}{\|x_{-2k/f}\|_2^{2-p}} = \frac{5}{2}k^{1-\frac{p}{2}}\|x_{-2k/f}\|_2^p$. Then

$$\begin{aligned}\|x_{[k]}\|_p^p - \|x_{S_0}\|_p^p &\le \frac{11}{10}k^{1-\frac{p}{2}}\|x_{-2k/f}\|_2^p + k\left(\frac{\|x_{-2k/f}\|_2}{\sqrt{k}}\right)^p - \|x_{\{i^*(0)+1,...,2k/f\}\cap S_0}\|_p^p \\ &\le \frac{21}{10}k^{1-\frac{p}{2}}\|x_{-2k/f}\|_2^p - \frac{5}{2}k^{1-\frac{p}{2}}\|x_{-2k/f}\|_2^p \le 0.\end{aligned}$$

Case 3. $i^*(0) \le 4k$, and $\sum_{j=i^*(0)+1}^{2k/f} x_j^2 \le 4\|x_{-2k/f}\|_2^2$.
With a little abuse of notation, let x_{S_0} denote the output of the ℓ_2/ℓ_2 with parameters $(4k, f/r^2)$. Notice that there are at most $8k$ non-zero elements in x_{S_0}, and $\|x_{-4k}\|_2^2 \le \|x_{-i^*(0)}\|_2^2 = \sum_{j=i^*(0)+1}^{2k/f} x_j^2 + \|x_{-2k/f}\|_2^2 \le 5\|x_{-2k/f}\|_2^2$. By Lemma 20, we have $\left\|x_{[k]}\right\|_p^p - \|x_{S_0}\|_p^p \le \left\|x_{[4k]}\right\|_p^p - \|x_{S_0}\|_p^p \le (4k)^{1-\frac{p}{2}}\frac{f^{\frac{p}{2}}}{r^p}\|x_{-4k}\|_2^p \le \mathcal{O}\left(\frac{1}{r^p}\right)k^{1-\frac{p}{2}}f^{\frac{p}{2}}\|x_{-2k/f}\|_2^p$. According to the above three cases, we conclude that $\|x_{[k]}\|_p^p - \|x_{S_0}\|_p^p \le \mathcal{O}\left(\frac{1}{r^p}\right)k^{1-\frac{p}{2}}f^{\frac{p}{2}}\|x_{-2k/f}\|_2^p$. Thus with failure probability at most $1/6$,

$$\|x-\hat{x}\|_p^p - \|x_{-k}\|_p^p = \|x_{[k]}\|_p^p - \sum_{j=0}^{r}\|x_{S_j}\|_p^p \le \mathcal{O}\left(\frac{1}{r^p}\right)k^{1-\frac{p}{2}}f^{\frac{p}{2}}\|x_{-2k/f}\|_2^p - \sum_{j=1}^{r}\left\|x_{S_j}\right\|_p^p. \quad (2.22)$$

In order to convert the first term on the right hand side of (2.22) to a term related to the ℓ_p norm (which is a semi-norm if $0 < p < 1$), we need the following inequalities: for every u and s, by splitting into chunks of size s, we have

$$s^{1-\frac{p}{2}}\|u_{-2s}\|_2^p \le \|u_{-s}\|_p^p, \qquad \text{and} \qquad \|u_{[\bar{s}]\cap[2s]}\|_2 \le \sqrt{s}\,|u_s|.$$

Define $c = (r+1)^{\min\{p,1\}}$. This gives us that, for $0 < p < 2$ $\frac{1}{(r+1)^p}k^{1-\frac{p}{2}}f^{\frac{p}{2}}\left\|x_{-2k/f}\right\|_2^p \le \frac{k^{1-\frac{p}{2}}f^{\frac{p}{2}}}{c}\left\|x_{-2k/f^{1+\frac{2}{p}}}\right\|_2^p + \frac{k^{1-\frac{p}{2}}f^{\frac{p}{2}}}{c}\sum_{j=1}^{r}\|\nabla\dashv\backslash\}\rceil x 2^j k/f 2^{j+1}k/f\|_2^p \le \frac{f^{(1-\frac{p}{2})(1+\frac{2}{p})+\frac{p}{2}}}{c}\left\|x_{-k/f^{1+\frac{2}{p}}}\right\|_p^p + \frac{1}{c}\sum_{j=1}^{r}k2^{pj/2}\left|x_{2^jk/f}\right|^p$. Therefore,

$$\begin{aligned}\|\hat{x}-x\|_p^p - \|x_{-k}\|_p^p &\le \mathcal{O}\left(\frac{1}{c}\right)f^{\frac{2}{p}}\left\|x_{-k/f^{1+\frac{2}{p}}}\right\|_p^p + \sum_{j=1}^{r}\mathcal{O}\left(\frac{1}{c}\right)k2^{pj/2}|x_{2^jk/f}|^p - \sum_{j=1}^{r}\|x_{S_j}\|_p^p \\ &\le \mathcal{O}\left(\frac{1}{c}\right)f^{\frac{2}{p}}\left\|x_{-k/f}\right\|_p^p + \sum_{j=1}^{r}\mathcal{O}\left(\frac{1}{c}\right)k2^{pj/2}|x_{2^jk/f}|^p - \sum_{j=1}^{r}\|x_{S_j}\|_p^p.\end{aligned} \quad (2.23)$$

Let $y = x_T$ denote an independent subsample of x with probability $f/(2^jk)$, and $\hat{y}$ be the output of the ℓ_2/ℓ_2 algorithm with parameter $\mathrm{s}(2, 1/(4(r+1))^{\frac{2}{p}})$. Notice that $|S_j| \le 2k/(r+1)f$ by the

adaptive ℓ_2/ℓ_2 guarantee. Define $Q = [2^j k/f] \setminus (S_0 \cup \cdots \cup S_{j-1})$. There are at least $2^j k/(2f)$ elements in Q, and every element in Q has absolute value at least $\left|x_{2^j k/f}\right|$. In each subsample, notice that $\mathbb{E}[|T \cap Q|] = \frac{1}{2}$. Thus with sufficiently small constant failure probability there exists at least 1 element in y with absolute value at least $|x_{2^j k/f}|$. On the other hand, by Lemma 20 and Lemma 19,

$$\left\|y_{[1]}\right\|_p^p - \|\hat{y}\|_p^p \leq \left\|y_{[2]}\right\|_p^p - \|\hat{y}\|_p^p \leq \frac{2^{1-\frac{p}{2}}}{4(r+1)} \|y_{-2}\|_2^p \leq \frac{1}{2(r+1)} \left(\frac{f}{2^j k}\right)^{\frac{p}{2}} \left\|x_{-2^j k/f}\right\|_2^p, \quad (2.24)$$

with sufficiently small constant failure probability given by the union bound. For the $k/(2f(r+1)^q)$ independent copies of subsamples, by a Chernoff bound, a $1/4$ fraction of them will have the largest absolute value in Q and (2.24) will also hold, with the overall failure probability being $e^{-\Omega(k/(fr^q))}$. Therefore, since $k/f > 2^{pj/2}k$,

$$\begin{aligned}
\left\|x_{S_j}\right\|_p^p &\geq \frac{2^{pj/2}k}{8(r+1)^q} \left[\left|x_{2^j k/f}\right|^p - \frac{1}{2(r+1)} \left(\frac{f}{2^j k}\right)^{\frac{p}{2}} \left\|x_{-2^j k/f}\right\|_2^p\right] \\
&\geq \frac{2^{pj/2}k}{8(r+1)^q} \left|x_{2^j k/f}\right|^p - \frac{k^{1-\frac{p}{2}} f^{\frac{p}{2}}}{16(r+1)^{q+1}} \left\|x_{-2k/f}\right\|_2^p,
\end{aligned}$$

and by the fact that $0 < q < p < 2$,

$$\begin{aligned}
\|x - \hat{x}\|_p^p - \|x_{-k}\|_p^p &\leq \mathcal{O}(\frac{1}{r^p}) k^{1-\frac{p}{2}} f^{\frac{p}{2}} \|x_{-2k/f}\|_2^p - \sum_{j=1}^{r} \left\|x_{S_j}\right\|_p^p \\
&\leq \left[\mathcal{O}\left(\frac{1}{r^p}\right) + \frac{r}{16(r+1)^{q+1}}\right] k^{1-\frac{p}{2}} f^{\frac{p}{2}} \|x_{-2k/f}\|_2^p - \sum_{j=1}^{r} \frac{2^{pj/2}k}{8(r+1)^q} \left|x_{2^j k/f}\right|^p \\
&\leq \mathcal{O}\left(\frac{1}{c}\right) f^{\frac{2}{p}} \left\|x_{-k/f}\right\|_p^p + \left[\mathcal{O}\left(\frac{1}{c}\right) + \frac{1}{16(r+1)^q} - \frac{1}{8(r+1)^q}\right] \sum_{j=1}^{r} k 2^{pj/2} \left|x_{2^j k/f}\right|^p \\
&\leq f^{\frac{2}{p}} \left\|x_{-k/f}\right\|_p^p \leq \epsilon \|x_{-k}\|_p^p.
\end{aligned}$$

The total number of measurements will be at most

$$\mathcal{O}\left(\frac{k}{f}\text{loglog}(n) + \frac{4kr^2}{f}\text{loglog}(n) + \frac{kr}{2fr^q} r^{\frac{2}{p}} \text{loglog}(n)\right) = \mathcal{O}\left(\frac{k}{\epsilon^{\frac{p}{2}}}\text{loglog}(n) \log^{\frac{2}{p}+1-(p-\frac{1}{2})^+}\left(\frac{1}{\epsilon}\right)\right),$$

while the total failure probability given by the union bound is $1/6 + e^{-\Omega(k/(fr^q))} < 1/3$, which completes the proof. □

Proofs of Theorem 10

We will first briefly introduce the definition and lower bound on the communication complexity of $\mathsf{Ind}\ell_\infty$, a two-party communication problem that is defined and studied in [187]. Then we will show how to use an adaptive $(1+\epsilon)$-approximate ℓ_p/ℓ_p sparse recovery scheme $\mathcal{A}$ to solve the communication problem $\mathsf{Ind}\ell_\infty$. We obtain a lower bound on the number of measurements required of an adaptive $(1+\epsilon)$-approximate ℓ_p/ℓ_p sparse recovery scheme.

Direct sum for distributional ℓ_∞ Consider two-party randomized communication complexity. There are two parties, Alice and Bob, with input vectors x and y respectively, and their goal is to solve a promise problem $f(x, y)$. The parties have private randomness. The communication cost of a protocol is its maximum transcript length, over all possible inputs and random coin tosses. The randomized communication complexity $R_\delta(f)$ is the minimum communication cost of a randomized protocol Π which for every input (x, y) outputs $f(x, y)$ with probability at least $1 - \delta$ (over the random coin tosses of the parties). We also study the distributional complexity of f, in which the parties are deterministic and the inputs (x, y) are drawn from distribution μ, and a protocol is correct if it succeeds with probability at least $1 - \delta$ in outputting $f(x, y)$, where the probability is now taken over $(x, y) \sim \mu$. We define $D_{\mu,\delta}(f)$ to be the minimum communication cost of a correct protocol Π.

We consider the following promise problem $\mathsf{Gap}\ell_\infty^B$, where B is a parameter, which was studied in [30, 200]. The inputs are pairs (x, y) of m-dimensional vectors, with $x_i, y_i \in \{0, 1, 2, \ldots, B\}$ for all $i \in [m]$, with the promise that (x, y) is one of the following types of instance:

- NO instance: for all i, $|x_i - y_i| \in \{0, 1\}$, or
- YES instance: there is a unique i for which $|x_i - y_i| = B$, and for all $j \neq i$, $|x_j - y_j| \in \{0, 1\}$.

The goal of a protocol is to decide which of the two cases (NO or YES) the input is in. Consider the distribution σ: for each $j \in [m]$, choose a random pair $(Z_j, P_j) \in \{0, 1, 2, \ldots, B\} \times \{0, 1\} \setminus \{(0, 1), (B, 0)\}$. If $(Z_j, P_j) = (z, 0)$, then $X_j = z$ and Y_j is uniformly distributed in $\{z, z + 1\}$; if $(Z_j, P_j) = (z, 1)$, then $Y_j = z$ and X_j is uniformly distributed on $\{z - 1, z\}$. Let $Z = (Z_1, \ldots, Z_m)$ and $P = (P_1, \ldots, P_m)$. Next choose a random coordinate $S \in [m]$. For coordinate S, replace (X_S, Y_S) with a uniform element of $\{(0, 0), (0, B)\}$. Let $X = (X_1, \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_m)$.

In [187], they define a problem, $\mathsf{Ind}\ell_\infty^{r,B}$, which involves solving r copies of $\mathsf{Gap}\ell_\infty^B$, and relate the ℓ_1/ℓ_1 recovery scheme with $\mathsf{Ind}\ell_\infty^{r,B}$ in order to get a lower bound. Here we introduce the definition of $\mathsf{Ind}\ell_\infty^{r,B}$ and present their results on the studies of communication complexity.

Definition 2 (Indexed $\mathsf{Ind}\ell_\infty^{r,B}$ Problem)**.** *There are r pairs of inputs $(x^1, y^1), (x^2, y^2), \ldots, (x^r, y^r)$ such that every pair (x^i, y^i) is a legal instance of the $\mathsf{Gap}\ell_\infty^B$ problem. Alice is given $x^1, \ldots, x^r$. Bob is given an index $I \in [r]$ and $y^1, \ldots, y^r$. The goal is to decide whether (x^I, y^I) is a NO or a YES instance of $\mathsf{Gap}\ell_\infty^B$.*

Let η be the distribution $\sigma^r \times U_r$, where U_r is the uniform distribution on $[r]$. We bound $D_{\eta,\delta}^{1-way}(\mathsf{Ind}\ell_\infty)^{r,B}$ as follows. For a function f, let f^r denote the problem of computing r instances of f. For a distribution ζ on instances of f, let $D_{\zeta^r,\delta}^{1-way,*}(f^r)$ denote the minimum communication cost of a deterministic protocol computing a function f with error probability at most δ *in each* of the r copies of f, where the inputs come from ζ^r.

Theorem 14. *For δ less than a sufficiently small constant, $D_{\eta,\delta}^{1-way}(\mathsf{Ind}\ell_\infty^{r,B}) = \Omega(\delta^2 rm/(B^2 \log r))$.*

Lemma 21. *Let $R = [s, cs]$ for some constant c and parameter s. Let X be a permutation independent distribution over $\{0, 1\}^n$ with $\|x\|_1 \in R$ with probability p. If y satisfies $\|x - y\|_1 \leq (1 - \epsilon)\|x\|_1$ with probability p' with $p' - (1 - p) = \Omega(1)$, then $I(x; y) = \Omega(\epsilon s \log(n/s))$.*

Lemma 22. *A lower bound of $\Omega(b)$ bits for such an adaptive ℓ_p/ℓ_p sparse recovery bit scheme with $p \leq 2$ implies a lower bound of $\Omega(b/((1 + c + d)\log n))$ bits for regular $(1 + \epsilon)$-approximate sparse recovery with failure probability $\delta - 1/n$.*

The overall lower bound The proof of the adaptive lower bound of ℓ_p/ℓ_p scheme is similar to the proof of the non-adaptive lower bound for ℓ_1/ℓ_1 sparse recovery given in [187]. Fix parameters $B = \Theta(1/\epsilon^{1/2})$, $r = k$, $m = 1/\epsilon^{(2+p)/2}$, and $n = k/\epsilon^3$. Given an instance $(x^1, y^1), \ldots, (x^r, y^r)$ of $\mathsf{Ind}\ell_\infty^{r,B}$ we define the input signal z to a sparse recovery problem. We allocate a set S^i of m disjoint coordinates in a universe of size n for each pair (x^i, y^i), and on these coordinates place the vector $y^i - x^i$. The locations turn out to be essential for the proof of Lemma 23 below, and are placed uniformly at random among the n total coordinates (subject to the constraint that the S^i are disjoint). Let ρ be the induced distribution on z.

Fix an ℓ_p/ℓ_p recovery multiround bit scheme $\mathcal{A}$ that uses b bits and succeeds with probability at least $1 - \delta_1/2$ over $z \sim \rho$. Let S be the set of top k coordinates in z. As shown in equation (14) of [187], $\mathcal{A}$ has the guarantee that if $v = \mathcal{A}(z)$, then

$$\|(v - z)_S\|_p^p + \|(v - z)_{[n]\setminus S}\|_p^p \leq (1 + 2\epsilon)\|z_{[n]\setminus S}\|_p^p. \tag{2.25}$$

Next is our generalization of Lemma 6.8 of [187].

Lemma 23. *For $B = \Theta(1/\epsilon^{1/2})$ sufficiently large, suppose that $\Pr_{z\sim\rho}[\|(v - z)_S\|_p^p \leq 10\epsilon \cdot \|z_{[n]\setminus S}\|_p^p] \geq 1 - \delta$. Then $\mathcal{A}$ requires $b = \Omega(k/(\epsilon^{p/2}\log k))$.*

Proof. We need to show how to use $\mathcal{A}$ to solve instances of $\mathsf{Ind}\ell_\infty^{r,B}$ with probability at least $1 - C$ for some small C, where the probability is over input instances to $\mathsf{Ind}\ell_\infty^{r,B}$ distributed according to η, inducing the distribution ρ. Since $\mathcal{A}$ is a deterministic sparse recovery bit scheme, it receives a sketch $f(z)$ of the input signal z and runs an arbitrary recovery algorithm g on $f(z)$ to determine its output $v = \mathcal{A}(z)$.

Given $x^1, \ldots, x^r$, for each $i = 1, 2, \ldots, r$, Alice places $-x^i$ on the appropriate coordinates in the block S^i used in defining z, obtaining a vector z_{Alice}, and transmits $f(z_{Alice})$ to Bob. Bob uses his inputs $y^1, \ldots, y^r$ to place y^i on the appropriate coordinate in S^i. He thus creates a vector z_{Bob} for which $z_{Alice} + z_{Bob} = z$. Given $f(z_{Alice})$, Bob computes $f(z)$ from $f(z_{Alice})$ and $f(z_{Bob})$, then $v = \mathcal{A}(z)$. We assume all coordinates of v are rounded to the real interval $[0, B]$, as this can only decrease the error.

We say that S^i is *bad* if either

- there is no coordinate j in S^i for which $|v_j| \geq \frac{B}{2}$ yet (x^i, y^i) is a YES instance of $\mathsf{Gap}\ell_\infty^{r,B}$, or
- there is a coordinate j in S^i for which $|v_j| \geq \frac{B}{2}$ yet either (x^i, y^i) is a NO instance of $\mathsf{Gap}\ell_\infty^{r,B}$ or j is not the unique j^* for which $y^i_{j^*} - x^i_{j^*} = B$

For B sufficiently large, the ℓ_p-error incurred by a bad block is at least $B/4$. Hence, if there are t bad blocks, the total error to the p-th power is at least $tB^p/4^p$, which must be smaller than $10\epsilon \cdot \|z_{[n]\setminus S}\|_p^p$ with probability $1 - \delta$. Condition on this, we would like to bound t. All coordinates in $z_{[n]\setminus S}$ have value in the set $\{0, 1\}$. Hence, $\|z_{[n]\setminus S}\|_p^p \leq rm$. So $t \leq 4^p 10\epsilon rm/B^p \leq 160\epsilon rm/B^p$. Plugging in r, m and B, $t \leq Ck$, where $C > 0$ is a constant that can be made arbitrarily small by increasing $B = \Theta(1/\epsilon^{1/2})$.

If a block S^i is not bad, then it can be used to solve $\mathsf{Gap}\ell_\infty^{r,B}$ on (x^i, y^i) with probability 1. Bob declares that (x^i, y^i) is a YES instance if and only if there is a coordinate j in S^i for which $|v_j| \geq B/2$.

Since Bob's index I is uniform on the m coordinates in $\mathsf{Ind}_\infty^{r,B}$, with probability at least $1-C$ the players solve $\mathsf{Ind}_\infty^{r,B}$ given that the ℓ_p error is small. Therefore they solve $\mathsf{Ind}_\infty^{r,B}$ with probability $1-\delta-C$ overall. By Theorem 14, for C and δ sufficiently small, $\mathcal{A}$ requires $\Omega(mr/(B^2\log r)) = \Omega(k/(\epsilon^{p/2}\log k))$ bits. □

Lemma 24. *Suppose* $\Pr_{z\sim\rho}[\|(v-z)_{[n]\setminus S}\|_p^p] \le (1-8\epsilon)\cdot\|z_{[n]\setminus S}\|_p^p] \ge \delta/2$. *Then* $\mathcal{A}$ *requires* $b = \Omega(\frac{1}{\epsilon^{p/2}}k\log(1/\epsilon))$.

Proof. The distribution ρ consists of $B(mr,1/2)$ ones placed uniformly throughout the n coordinates, where $B(mr,1/2)$ denotes the binomial distribution with mr events of $1/2$ probability each. Therefore with probability at least $1-\delta/4$, the number of ones lies in $[\delta mr/8,(1-\delta/8)mr]$. Thus by Lemma 21, $I(v;z)\ge\Omega(\epsilon mr\log(n/(mr)))$. Since the mutual information only passes through a b-bit string, $b=\Omega(\epsilon mr\log(n/(mr))) = \Omega(\frac{1}{\epsilon^{p/2}}k\log(1/\epsilon))$ as well. □

Theorem 15. *Any adaptive* $(1+\epsilon)$*-approximate* ℓ_p/ℓ_p *recovery scheme with sufficiently small constant failure probability* δ *must make* $\Omega(\frac{1}{\epsilon^{p/2}}k/\log^2(k/\epsilon))$ *measurements.*

Proof. We will lower bound any ℓ_p/ℓ_p sparse recovery bit scheme $\mathcal{A}$. If $\mathcal{A}$ succeeds, then in order to satisfy inequality (2.25), we must either have $\|(v-z)_S\|_p^p \le 10\epsilon\|z_{[n]\setminus S}\|_p^p$ or we must have $\|(v-z)_{[n]\setminus S}\|_p^p \le (1-8\epsilon)\|z_{[n]\setminus S}\|_p^p$. Since $\mathcal{A}$ succeeds with probability at least $1-\delta$, it must either satisfy the hypothesis of Lemma 23 or the hypothesis of Lemma 24. But by these two lemmas, it follows that $b=\Omega(\frac{1}{\epsilon^{p/2}}k/\log k)$. Therefore by Lemma 22, any $(1+\epsilon)$-approximate ℓ_p/ℓ_p sparse recovery algorithm requires $\Omega(\frac{1}{\epsilon^{p/2}}k/\log^2(k/\epsilon))$ measurements. □

Proofs of Theorem 11

In this section, we will prove Theorem 11. Our algorithm first approximates $\|x_{-k}\|_2$. The goal is to compute a value V which is not much smaller than $\frac{1}{k}\|x_{-k}\|_2^2$, and also at least $\Omega(\frac{1}{k})\|x_{-\Omega(k)}\|_2^2$. This value will be used to filter out coordinates that are not large enough, while ensuring that heavy coordinates are included. We need the following lemma, which for example can be found in Section 4 of [148].

Lemma 25. *Using* $\log(1/\delta)$ *non-adaptive measurements we can find with probability* $1-\delta$ *a value* V *such that* $\frac{1}{C_1k}\|x_{-C_2k}\|_2^2 \le V \le \frac{1}{k}\|x_{-k}\|_2^2$, *where* C_1, C_2 *are absolute constants larger than* 1.

We use the aforementioned lemma with $\Theta(\log k)$ measuremenents to obtain such a value V with probability $1-1/\mathrm{poly}(k)$. Now let c be an absolute constant and let $g:[n]\to[k^c]$ be a random hash function. Then, with probability at least $1-\frac{1}{\mathrm{poly}(k)}$ we have that for every $i,j\in H_k(x)$, $g(i)\ne g(j)$. By running PARTITIONCOUNTSKETCH$(x, 2C_1k, \{g^{-1}(1), g^{-1}(2),\dots g^{-1}(k^c)\}$, we get back an estimate w_j for every $j\in[k^c]$; here C_1 is an absolute constant. Let γ' be an absolute constant to be chosen later. We set $S=\{j\in[k^c] : w_j^2\ge\gamma' V\}$ and $T=\bigcup_{j\in S}g^{-1}(j)$. We prove the following lemma.

Lemma 26. *Let* C' *be an absolute constant. With probability at least* $1-1/\mathrm{poly}(k)$ *the following holds.*

1. $|S| = \mathcal{O}(k)$.
2. *Every* $j\in[k^c]$ *such that there exists* $i\in H_k(x)\cap g^{-1}(j)$, *will be present in* S.

3. *For every $j \in S$, there exists exactly one coordinate $i \in g^{-1}(j)$ with $x_i^2 \geq \frac{1}{C'k}\|x_{-C_2k}\|_2^2$.*
4. *For every $j \in S$, $\|x_{g^{-1}(j)\setminus H_k(x)}\|_2^2 \leq \frac{1}{k^2}\|x_{-k}\|_2^2$.*

Proof. Let C_0 be an absolute constant larger than 1. Note that with probability $1 - C_0^2 \cdot k^{6-c}$, all $i \in H_{C_0k^3}(x)$ (and, hence, also in $H_{C_0k^3,1/k^3}(x)$) are isolated under g. Fix $j \in [k^c]$ and, for $i \in [n]$, define the random variable $Y_i = 1_{g(x_i)=j}x_i^2$. Now observe that

$$\mathbb{E}\left[\sum_{i\in g^{-1}(j)\setminus H_{C_0k^3,1/k^3}(x)} Y_i\right] = \frac{1}{k^c}\|x_{-C_0k^3}\|_2^2.$$

Applying Bernstein's inequality to the variables Y_i with

$$K = \frac{1}{C_0k^3}\|x_{-C_0k^3}\|_2^2, \qquad \text{and} \qquad \sigma^2 < \frac{1}{k^{c+3}}\|x_{-C_0k^3}\|_2^4,$$

we have that

$$\Pr\left[\sum_{i\in g^{-1}(j)\setminus H_{C_0k^3,1/k^3}(x)} x_i^2 \geq 1/k^2\|x_{-C_0k^2}\|_2^2\right] \leq e^{-k},$$

where c is an absolute constant. This allows us to conclude that the above statement holds for all different k^c possible values j, by a union-bound. We now prove the bullets one by one. We remind the reader that PARTITIONCOUNTSKETCH aproximates the value of every $\|x_{g^{-1}(j)}\|_2^2$ with a multiplicate error in $[1-\gamma, 1+\gamma]$ and additive error $\frac{1}{C_0k}\|x_{-k}\|_2^2$.

1. Since there are at most $\frac{1}{\gamma'(1+\gamma)}C_2k + C_2k$ indices j with $(1+\gamma)\|x_{g^{-1}(j)}\|_2^2 \geq \frac{\gamma'}{k}\|x_{-k}\|_2^2 \geq \gamma'V$, the algorithm can output at most $\mathcal{O}(k)$ indices.

2. The estimate for such a j will be at least $(1-\gamma)\frac{1}{k}\|x_{-k}\|_2^2 - \frac{1}{2C_1k}\|x_{-C_2k}\|_2^2 \geq \gamma'V$, for some suitable choice of γ'. This implies that j will be included in S.

3. Because of the guarantee for V and the guarantee of PARTITIONCOUNTSKETCH, we have that all j that are in S satisfy $(1+\gamma)\|x_{g^{-1}(j)}\|_2^2 + \frac{1}{k}\|x_{-2C_1k}\|_2^2 \geq \frac{\gamma'}{k}\|x_{-C_2k}\|_2^2$, and since

$$\sum_{i\in g^{-1}(j)\setminus H_{C_0k^3}(x)} x_i^2 \leq \frac{1}{k^2}\|x_{-k}\|_2^2,$$

this implies that there exists $i \in H_{C_0k^3}(x) \cap g^{-1}(j)$. But since all $i \in H_{C_0k^3}(x)$ are perfectly hashed under g, this implies that this i should satisfy $x_i^2 \geq \frac{1}{C'k}\|x_{-C_2k}\|_2^2$, from which the claim follows.

4. Because elements in $H_{C_0k^3}(x)$ are perfectly hashed, we have that

$$\|x_{g^{-1}(j)\setminus H_k(x)}\|_2^2 = \|x_{g^{-1}(j)\setminus H_{C_0k^3}}(x)\|_2^2 \leq \frac{1}{k^2}\|x_{-k}\|_2^2$$

for C_0 large enough. □

Given S, we proceed in the following way. For every $j \in S$, we run the algorithm guaranteed by Lemma 15 from the full version [6] to obtain an index i_j, using $\mathcal{O}(k\text{loglog}n)$ measurements. Then we observe directly x_{i_j} using another $\mathcal{O}(k)$ measurements, and form vector $z = x - x_{\{i_j\}_{j\in S}}$. We need the following lemma.

Lemma 27. *With probability* $1 - 1/\text{poly}(k)$, $|H_k(x) \setminus \{i_j\}_{j\in S}| \leq \frac{k}{\log^2 n}$.

Proof. Let us consider the calls to the 1-sparse recovery routine in j for which there exists $i \in H_k(x) \cap g^{-1}(j)$. Since the 1-sparse recovery routine succeeds with probability $1 - 1/\text{poly}(\log n)$, then the probability that we have more than $\frac{k}{\log^2 n}$ calls that fail, is

$$\binom{k}{\frac{k}{\log^2 n}} \left(\frac{1}{\text{poly}(\log n)}\right)^{k/\log^2 n} \leq \frac{1}{\text{poly}(k)}.$$

This gives the proof of the lemma. □

For the last step of our algorithm, we run PARTITIONCOUNTSKETCH$(z_T, k/\log(n), [n])$ to estimate the entries of z. We then find the coordinates with the largest $2k$ estimates, and observe them directly. Since

$$\frac{\log n}{k}\|(z_T)_{-k/\log n}\|_2^2 \leq \frac{\log n}{k} \cdot \frac{1}{k^2}\|x_{-k}\|_2^2 = \frac{\log n}{k^3}\|x_{-k}\|_2^2,$$

every coordinate will be estimated up to additive error $\frac{\log n}{k^3}\|x_{-k}\|_2^2$, which shows that every coordinate in $T \cap H_{k,1/k}(x)$ will be included in the top $2k$ coordinates. Putting everything together, we obtain the desired result.

Proofs of Theorem 12

In this section, we give an algorithm for ℓ_2/ℓ_2 compressed sensing using $\mathcal{O}(\text{loglog}n)$ rounds, instead of $\mathcal{O}(\log^* k \cdot \text{loglog}n)$ rounds. Specifically, we firstly prove the first bullet of Theorem 12. We call this algorithm ADAPTIVESPARSERECOVERY$_{\ell_\infty/\ell_2}$.

We proceed with the design and the analysis of the algorithm. We note that for $k/\epsilon = \mathcal{O}(\log^5 n)$[7], ℓ_∞/ℓ_2 gives already the desired result. So, we focus on the case of $k/\epsilon = \Omega(\log^5 n)$. We pick a hash function $h : [n] \to [B]$, where $B = ck/(\epsilon \log n)$ for some constant c large enough. The following follows by an application of Bernstein's Inequality and the Chernoff Bound, similarly to ℓ_∞/ℓ_2.

Lemma 28. *With probability* $1 - 1/\text{poly}(n)$*, the following holds:*

$$\forall j \in [B] : |H_{k/\epsilon}(x) \cap h^{-1}(j)| \leq \log n, \qquad \textit{and} \qquad \left|\sum_{i\in h^{-1}(j)\setminus H_{k/\epsilon}(x)} x_i^2\right| \leq \frac{\epsilon}{k}\|x_{-k}\|_2^2.$$

[6] see https://arxiv.org/pdf/1804.09673.pdf

[7] the constant 5 is arbitrary

We now run the ℓ_∞/ℓ_2 algorithm for the previous section on vectors $x_{h^{-1}(1)}, x_{h^{-1}(2)}, \ldots, x_{h^{-1}(B)}$ with sparsity parameter $\mathcal{O}(\log n)$, to obtain vectors $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_B$. The number of rounds is $\mathcal{O}(\text{loglog}(n))$, since we can run the algorithm in every bucket in parallel. By the definition of the ℓ_∞/ℓ_2 algorithm, one can see that $|\text{supp}(\hat{x}_j)| \leq \mathcal{O}(\log n)$. We set $S = \cup_{j \in B} |\text{supp}(x_j)|$, and observe that $|S| = ck/(\epsilon \log n) \cdot \mathcal{O}(\log n) = \mathcal{O}(k/\epsilon)$. The number of measurements equals $ck/(\epsilon \log n) \cdot \mathcal{O}(\log n \cdot \text{loglog}(n \log(n/k))) = \mathcal{O}((k/\epsilon) \cdot \text{loglog}(n \log(n/k)))$.

Lemma 29. *With probability* $1 - 1/\text{poly}(n)$*, we have that* $|S \setminus H_{k/\epsilon}(x)| \leq \frac{k}{\epsilon \log^2 n}$.

Proof. Since every call to ℓ_∞/ℓ_2 fails with probability $1/\text{poly}(\log n)$, the probability that we have more than a $\frac{1}{\log n}$ fraction of the calls that fail is at most

$$\binom{B}{B/\log^2 n} \left(\frac{1}{\log n}\right)^{B/\log n} \leq (e \log^2 n)^{\log n} (\log n)^{-B/\log n} \leq \frac{1}{\text{poly}(n)}.$$

This implies that S will contain all but at most $B/\log^2 n \cdot \log n = k/(\epsilon \log^2 n)$ coordinates $i \in H_k(x)$.

□

We now observe x_S directly and form the vector $z = x - x_S$, for which $\|z_{-k/(\epsilon \log^2 n)}\|_2 \leq \|x_{-k/\epsilon}\|_2$. We now run a standard ℓ_2/ℓ_2 algorithm that fails with probability $1/\text{poly}(n)$ to obtain a vector $\hat{z}$ that approximates z (for example PARTITIONCOUNTSKETCH$(z, k/(\epsilon \log^2 n), [n])$ suffices). We then output $\hat{z} + x_S$, for which $\|\hat{z} + x_S - x\|_2 = \|\hat{z} - z\| \leq (1+\epsilon)\|z_{-k/(\epsilon \log n)}\|_2 \leq (1+\epsilon)\|x_{-k}\|_2$. The number of measurements of this step is $\mathcal{O}(\frac{1}{\epsilon} \frac{k}{\log^2 n} \cdot \log n) = o(\frac{k}{\epsilon})$. The total number of rounds is clearly $\mathcal{O}(\text{loglog}(n \log(\frac{n\epsilon}{k})))$.

We now prove the second part of Theorem 12. We first need an improved algorithm for the 1-sparse recovery problem.

Lemma 30. *Let* $x \in \mathbb{R}^n$*. There exists an algorithm* IMPROVEDONESPARSERECOVERY*, that uses* $\mathcal{O}(\text{loglog}n + \frac{1}{\epsilon}\text{loglog}(\frac{1}{\epsilon}))$ *measurements in* $\mathcal{O}(\text{loglog}(n))$ *rounds, and finds with sufficiently small constant probability an* $\mathcal{O}(1)$*-sparse vector* $\hat{x}$ *such that* $\|\hat{x} - x\|_2 \leq (1+\epsilon)\|x_{-1}\|_2$.

Proof. We pick a hash function $h : [n] \to [B]$, where $B = \lceil 1/\epsilon^h \rceil$ for a sufficiently large constant h. Observe that all elements of $H_{\sqrt{B}}(x)$ are perfectly hashed under h with constant probability, and, $\forall j \in [B]$, $\mathbb{E}\left[\left\|x_{h^{-1}(j) \setminus H_{\sqrt{B}}(x)}\right\|_2\right] \leq 1/B \|x_{-\sqrt{B}}\|_2$. As in the previous sections, invoking Bernstein's inequality we can get that with probability $1 - 1/\text{poly}(B)$, $\forall j \in [B]$, $\left\|x_{h^{-1}(j) \setminus H_{\sqrt{B}}(x)\|_2}\right\|_2^2 \leq \frac{c \log B}{B} \|x_{-\sqrt{B}}\|_2^2$, where c is some absolute constant, and the exponent in the failure probability is a function of c.

We now define the vector $z \in \mathbb{R}^B$, the j-th coordinate of which equals $z_j = \sum_{i \in h^{-1}(j)} \sigma_{i,j} x_i$. We shall invoke Khintchine inequality to obtain

$$\forall j, \ \Pr\left[\left|\sum_{i \in h^{-1}(j) \setminus H_{\sqrt{B}}(x)} \sigma_{i,j} x_i\right|^2 > \frac{c'}{\epsilon} \left\|x_{h^{-1}(j) \setminus H_{\sqrt{B}}(x)}\right\|_2^2\right] \leq e^{-\Omega(1/\epsilon^2)}$$

, for some absolute constant c'. This allows us to take a union-bound over all $B = \lceil 1/\epsilon^h \rceil$ entries of z to conclude that there exists an absolute constant ζ such that $\forall j \in [B]$, $\left|\sum_{i \in h^{-1}(j) \setminus H_{\sqrt{B}}(x)} \sigma_{i,j} x_i\right|^2 \leq \frac{c'}{\epsilon}\|x_{h^{-1}(j)\setminus H_{\sqrt{B}}(x)}\|_2^2 < \zeta\epsilon\|x_{-1}\|_2^2$, by setting h large enough. Now, for every coordinate $j \in [B]$ for which $h^{-1}(j) \cap H_{1,\epsilon}(x) = i^*$ or some $i^* \in [n]$, we have that $|z_j| \geq \left||x_{i^*}| - \sqrt{\frac{c \log B}{B} \cdot \frac{c'}{\epsilon}}\|x_{-\sqrt{B}}\|_2\right| \geq (1-\zeta)\sqrt{\epsilon}\|x_{-1}\|_2$, whereas for every $j \in [B]$ such that $h^{-1}(j) \cap H_{1,\epsilon\zeta}(x) = \emptyset$ it holds that $|z_j| \leq 2\zeta\sqrt{\epsilon}\|x_{-1}\|_2$. We note that $H_{1,\epsilon}(x) \subset H_{\sqrt{B}}(x)$, and hence all elements of $H_{1,\epsilon}(x)$ are also perfectly hashed under h. Moreover, observe that $\mathbb{E}\|z_{-1}\|_2^2 \leq \|x_{-1}\|_2^2$, and hence by Markov's inequality, we have that $\|z_{-1}\|_2^2 \leq 10\|x_{-1}\|_2^2$ holds with probability $9/10$. We run the ℓ_2/ℓ_2 algorithm of Theorem 12 for vector z with the sparsity being set to 1, and obtain vector $\hat{z}$. We then set $S = \mathrm{supp}(\hat{z})$. We now define $w = (|z_1|, |z_2|, \ldots)$, for which $\|w_{-1}\|_2 = \|z_{-1}\|_2$. Clearly, $\|z - z_S\|_2^2 \leq \|z - \hat{z}\|_2^2 \leq (1+\epsilon)\|z_{-1}\|_2^2 = (1+\epsilon)\|w_{-1}\|_2^2$. So $\|w - w_S\|_2^2 = \|z - z_S\|_2^2 \leq (1+\epsilon)\|w_{-1}\|_2^2$. We now prove that $\left\|x - x_{\cup_{j\in S} h^{-1}(j)}\right\|_2 \leq (1 + \mathcal{O}(\epsilon))\|x_{-1}\|_2$. Let i^* be the largest coordinate in magnitude of x, and $j^* = h(i^*)$. If $j^* \in S$, then it follows easily that $\|x - x_{\cup_{j\in S} h^{-1}(j)}\|_2 \leq \|x_{-1}\|_2$. Otherwise, since $\sum_{j \neq j^*} w_j^2 = \|w_{-1}\|_2^2$, and $\sum_{j \notin S} w_j^2 \leq (1+\epsilon)\|w_{-1}\|_2^2$, it must be the case that $\left|w_{j^*}^2 - \|w_S\|_2^2\right| \leq \epsilon\|w_{-1}\|_2^2 \leq 10\epsilon\|x_{-1}\|_2^2$. The above inequality, translates to $\sum_{i \in h^{-1}(j^*)} x_i^2 \leq |S|\zeta\epsilon\|x_{-1}\|_2^2 + \zeta\epsilon\|x_{-1}\|_2^2 + 10\epsilon\|x_{-1}\|_2^2 + \sum_{j\in S}\sum_{i\in h^{-1}(j)} x_j^2 = \mathcal{O}(\epsilon)\|x_{-1}\|_2^2 + \sum_{j\in S}\sum_{i\in h^{-1}(j)} x_j^2$. This gives $\left\|x - x_{\cup_{j\in S} h^{-1}(j)}\right\|_2 = \sum_{i\in h^{-1}(j^*)} x_i^2 + \sum_{j \notin S\cup\{j^*\}}\sum_{i\in h^{-1}(j)} x_i^2 \leq \mathcal{O}(\epsilon)\|x_{-1}\|_2^2 + \mathcal{O}(1)\zeta\epsilon\|x_{-1}\|_2^2 + \sum_{j\in S}\sum_{i\in h^{-1}(j)} x_j^2 + \sum_{j\notin S\cup\{j^*\}}\sum_{i\in h^{-1}(j)} x_i^2 + \leq (1 + \mathcal{O}(\epsilon))\|x_{-1}\|_2^2$.

Given S, we run the 1-sparse recovery routine on vectors x_j for $j \in S$, with a total of $\mathcal{O}(\mathrm{loglog}n)$ measurements and $\mathcal{O}(\mathrm{loglog}n)$ rounds. We then output $\{x_{i_j}\}_{j\in S}$. Let i_j be the index returned for $j \in S$ by the 1-sparse recovery routine. Since we have a constant number of calls to the 1-sparse recovery routine (because S is of constant size), all our 1-sparse recovery routines will succeed. We now have that $\|x - x_{\cup_{j\in S} i_j}\|_2 \leq \|x_{\bar{S}}\|_2 + \sum_{j\in S}\|x_{h^{-1}(j)} - x_{i_j}\|_2 \leq \|x_{\bar{S}}\|_2 + \sum_{j\in S}(1+\epsilon)\|x_{h^{-1}(j)\setminus H_1(x)}\|_1 \leq (1 + \mathcal{O}(\epsilon))\|x_{-1}\|_2$. Rescaling ϵ, we get the desired result. □

The algorithm for general k is similar to [121], apart from the fact that we subsample at a slower rate, and also use our new 1-sparse recovery algorithm as a building block. In the algorithm below, R_r is the universe we are restricting our attention on at the rth round. Moreover, J is the set of coordinates that we have detected so far. We are now ready to prove Theorem 12.

Proof. The number of measurements is bounded in the exact same way as in Theorem 3.7 from [121].

We fix a round r and $i \in H_{k_r,\epsilon_r}(x^{(r)})$. Then the call to SUBSAMPLE$(R_r, 1/(C_0 k_r))$ yields

$$\Pr\left[|H_{k_r,\epsilon_r}(x - x^{(r)}) \cap S_t| = \{i\}\right] \geq \frac{1}{C_0 k_r}, \quad \mathbb{E}\left[\|x_{S_t \setminus H_{k_r,\epsilon_i}(x^{(r)})}\|_2^2\right] = \frac{1}{C_0 k_r}\|x_{-k_r}\|_2^2.$$

Setting C_0 to be large enough and combining Markov's inequality with the guarantee of Lemma 30, we get that the probability that the call to IMPROVEDONESPARSERECOVERY(x_{S_t}) returns i is $\Theta(1/k_r)$. Because we repeat $k_r \log(1/(f_r\delta_r))$, the probability that i or a set S_i of size $\mathcal{O}(1)$ such that $\|x_{\{i\}} - x_{S_i}\|_2 \leq \epsilon_i\|x_{-k_r}\|_2^2$, is not added in J is at most $(1 - 1/k_r)^{k_r \log(1/(f_r\delta_r))} = f_r\delta_r$.

Algorithm 6 Adaptive ℓ_2/ℓ_2 Sparse Recovery

$R_0 \leftarrow [n]$.
$x_0 \leftarrow \vec{0}$.
$\delta_0 \leftarrow \delta/2, \epsilon_0 \leftarrow \epsilon/e, f_0 \leftarrow 1/32, k_0 \leftarrow k$.
$J \leftarrow \emptyset$.
For $r = 0$ to $\mathcal{O}(\log^* k)$ **do**
    For $t = 0$ to $\Theta(k_r \log(1/(\delta_r f_r)))$ **do**
        $S_t \leftarrow \text{SUBSAMPLE}(x - x^{(r)}, R_r, 1/(C_0 k_r))$.
        $J \leftarrow J \cup \text{IMPROVEDONESPARSERECOVERY}((x - x^{(r)})_{S_t})$.
    End For
    $R_{r+1} \leftarrow [n] \setminus J$.
    $\delta_{r+1} \leftarrow \delta_r/8$.
    $\epsilon_{r+1} \leftarrow \epsilon_r/2$.
    $f_{r+1} \leftarrow 1/2^{1/(4^{i+r} f_r)}$.
    $k_{r+1} \leftarrow f_r k_r$.
    $R_{r+1} \leftarrow [n] \setminus J$.
End For
$\hat{x} \leftarrow x^{(r+1)}$.
Return $\hat{x}$.

Given the above claim, the number of measurements is $\mathcal{O}((k\text{loglog}n + k/\epsilon\text{loglog}(1/\epsilon)\log(1/\delta))$ and the analysis of the iterative loop proceeds almost identically to Theorem 3.7 of [121].

□

Chapter 3

Learning with Low-Rank Approximations

3.1 Matrix Completion and Robust PCA

3.1.1 Introduction

Non-convex matrix factorization problems have been an emerging object of study in theoretical computer science [8, 106, 123, 172, 193, 214, 215, 220], optimization [209, 239], machine learning [37, 87, 88, 122, 155, 238], and many other domains. In theoretical computer science and optimization, the study of such models has led to significant advances in provable algorithms that converge to local minima in linear time [2, 5, 106, 123, 220]. In machine learning, matrix factorization serves as a building block for large-scale prediction and recommendation systems, e.g., the winning submission for the Netflix prize [138]. The matrix factorization problems can be stated as finding a target matrix $\mathbf{X}^*$ in the form of $\mathbf{X}^* = \mathbf{AB}$, by minimizing the objective function $H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2$ or $H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2$ over factor matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times n_2}$ with a known value of $r \ll \min\{n_1, n_2\}$, where $H(\cdot)$ is some function that characterizes the desired properties of $\mathbf{X}^*$. Two prototypical examples are matrix completion and robust Principal Component Analysis (PCA).

This work develops a novel framework to analyze a class of non-convex matrix factorization problems and show their strong duality, which leads to exact recoverability for matrix completion and robust PCA via the solutions to convex optimization problems. Strong duality is well understood for convex optimization, but very few non-convex problems were known to have this property. The results in this work thus significantly expand the set of non-convex problems with strong duality. Furthermore, our framework also shows exact recoverability of the two prototypical examples matrix completion and robust PCA with nearly-optimal sample complexity.

Our work is motivated by several promising areas where our analytical framework for non-convex matrix factorizations is applicable. The first area is low-rank matrix completion. It has been shown that a low-rank matrix can be exactly recovered by finding a solution of the form $\mathbf{AB}$ that is consistent with the observed entries (assuming that it is incoherent) [88, 123, 220]. This problem has received a tremendous amount of attention due to its important role in optimization and its wide applicability in many areas such as quantum information theory and collaborative filtering [21, 106, 255]. The second area is robust PCA, a fundamental problem of interest in data processing. It aims at recovering both the low-rank and the sparse components exactly from their

superposition [57, 100, 177, 247, 253, 255], where the low-rank component corresponds to the product of $\mathbf{A}$ and $\mathbf{B}$ while the sparse component is captured by a proper choice of function $H(\cdot)$, e.g., the ℓ_1 norm [17, 57]. Besides these two areas, we believe that our analytical framework can be potentially applied to other non-convex problems more broadly, e.g., matrix sensing [225], dictionary learning [219], weighted low-rank approximation [155, 193], and deep linear neural network [131], which may be of independent interest.

Without assumptions on the structure of the objective function, direct formulations of matrix factorization problems are NP-hard to optimize in general [110, 251]. With standard assumptions on the structure of the problem and with sufficiently many samples, these optimization problems can be solved efficiently, e.g., by convex relaxation [55, 63]. Some other methods run local search algorithms given an initialization close enough to the global solution in the basin of attraction [87, 106, 123, 126, 220]. However, these methods have sample complexity significantly larger than the information-theoretic lower bound; see Table 3.1 for a comparison. The problem becomes even more challenging when the number of samples is small enough that the sample-based initialization can be far from the desired solution, in which case the algorithm can run into a local minimum or a saddle point.

Another line of work has focused on studying the loss surface of matrix factorization problems, providing positive results for approximately achieving global optimality. One nice property in this line of research is that there is no spurious local minima for specific applications such as matrix completion [88], matrix sensing [37], dictionary learning [219], phase retrieval [218], linear deep neural networks [131], etc. However, these results are based on concrete forms of objective functions. Also, even when any local minimum is guaranteed to be globally optimal, in general it remains NP-hard to escape high-order saddle points [6], and additional arguments are needed to show the achievement of a local minimum. Most importantly, all existing results rely on strong assumptions on the sample size.

3.1.2 Our results on sample efficiency, optimization, and robustness

Our work studies a variety of non-convex matrix factorization problems, and the goal is to provide a unified framework to analyze a large class of matrix factorization problems and to provide efficient algorithms to achieve global optimum. Our main results show that although matrix factorization problems are hard to optimize in general, *under certain dual conditions the duality gap is zero*, and thus the problem can be converted to an equivalent convex program.

To state the main theorem of our framework, recall that a function $H(\cdot)$ is closed if for each $\alpha \in \mathbb{R}$, the sub-level set $\{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2} : H(\mathbf{X}) \leq \alpha\}$ is a closed set. Also, recall the nuclear norm (a.k.a. trace norm) of a matrix $\mathbf{X}$ is $\|\mathbf{X}\|_* = \sum_{i=1}^{r} \sigma_i(\mathbf{X})$. Define the r^*-norm to be $\|\mathbf{X}\|_{r*} = \max_{\mathbf{M}} \langle \mathbf{M}, \mathbf{X} \rangle - \frac{1}{2}\|\mathbf{M}\|_r^2$ where $\|\mathbf{M}\|_r^2 = \sum_{i=1}^{r} \sigma_i^2(\mathbf{M})$ is the sum of the first r largest squared singular values. Note that both $\|\mathbf{X}\|_*$ and $\|\mathbf{X}\|_{r*}$ are convex functions. Our main results are as follows.

Theorem 16 (Strong Duality. Informal)**.** *Under certain dual conditions, strong duality holds for the non-convex optimization problem*

$$(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = \operatorname*{argmin}_{\mathbf{A} \in \mathbb{R}^{n_1 \times r}, \mathbf{B} \in \mathbb{R}^{r \times n_2}} H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2, \tag{3.1}$$

where $H(\cdot)$ is convex and closed. In other words, problem (3.1) and its bi-dual problem

$$\widetilde{\mathbf{X}} = \underset{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}}{\operatorname{argmin}} \; H(\mathbf{X}) + \|\mathbf{X}\|_{r*}, \tag{3.2}$$

have exactly the same optimal solutions in the sense that $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \widetilde{\mathbf{X}}$.

Similarly, under certain dual conditions, strong duality holds for the non-convex optimization problem

$$(\bar{\mathbf{A}}, \bar{\mathbf{B}}) = \underset{\mathbf{A}\in\mathbb{R}^{n_1\times r}, \mathbf{B}\in\mathbb{R}^{r\times n_2}}{\operatorname{argmin}} \; H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2, \tag{3.3}$$

where $H(\cdot)$ is convex and closed. In other words, problem (3.1) and its bi-dual problem

$$\bar{\mathbf{X}} = \underset{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}}{\operatorname{argmin}} \; H(\mathbf{X}) + \|\mathbf{X}\|_{*}, \tag{3.4}$$

have exactly the same optimal solutions in the sense that $\bar{\mathbf{A}}\bar{\mathbf{B}} = \bar{\mathbf{X}}$.

Description of dual conditions. Intuitively, the dual conditions in the above-mentioned theorems state that the angle between $\partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$ and the row and column spaces of $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$ is small. In other words, there is a matrix in the sub-differential set $\partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$ which has almost the same row and column spaces as matrix $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$. For example, we have $\partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) = \Omega$ for the matrix completion problem, where Ω represents the subspace of matrices supported on the observed indices. Then the dual conditions require that there is a matrix which is supported on the observed indices and shares almost the same row and column spaces as $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$.

Theorem 20 connects the non-convex programs (3.1) to its convex counterpart (3.2) via strong duality; see Figure 3.1. Note that strong duality rarely holds in the non-convex optimization region: low-rank matrix approximation [182] and quadratic optimization with two quadratic constraints [33] are among the few paradigms that enjoy such a nice property. Given strong duality, the computational issues of the original problem can be overcome by solving the convex bi-dual problem (3.2).

Furthermore, Theorem 20 also connects the non-convex programs (3.3) to its convex counterpart (3.4): the theorem connects $\frac{1}{2}\|\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{B}\|_F^2$ to the nuclear norm $\|\mathbf{X}\|_*$. This gives new insights for the nuclear norm relaxation technique commonly used for optimization problems with low rank constraints from the perspective of strong duality. The theorem also connects the regularization $\frac{1}{2}\|\mathbf{AB}\|_F^2$ to the r^* norm $\|\mathbf{X}\|_{r*}$. This regularization is of special interest to many matrix factorization problems. For example, when $H(\mathbf{AB}) = \frac{1}{2}\|\mathbf{X}\|_F^2 - \langle \mathbf{X}, \mathbf{AB}\rangle$, problem (3.1) reduces to the PCA problem: $\min_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{X} - \mathbf{AB}\|_F^2$. When $H(\mathbf{AB}) = \frac{1}{2}\|\mathbf{X}\|_F^2 - \langle \mathbf{X}, \mathbf{AB}\rangle + \gamma\|\mathbf{A}\|_F^2 + \gamma\|\mathbf{B}\|_F^2$, problem (3.1) reduces to the quadratically regularized PCA problem [227]: $\min_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{X} - \mathbf{AB}\|_F^2 + \gamma\|\mathbf{A}\|_F^2 + \gamma\|\mathbf{B}\|_F^2$. Our framework of strong duality is then applicable to all these problems.

The positive result of our framework is complemented by a lower bound to formalize the hardness of the above problem in general. Assuming that the random 4-SAT problem [193] is hard, we give a strong negative result for deterministic algorithms. If also BPP = P, then the same conclusion holds for randomized algorithms succeeding with probability at least $2/3$.

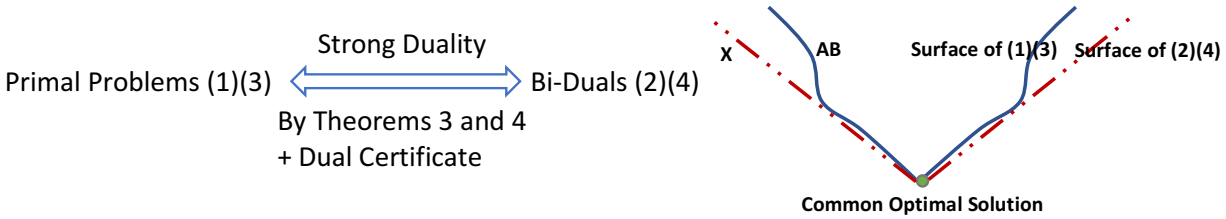


Figure 3.1: Strong duality of matrix factorizations.

Theorem 17 (Hardness Statement)**.** *Assuming that random 4-SAT is hard on average, there is a problem in the form of* (3.1) *such that any deterministic algorithm achieving* $(1+\epsilon)\mathsf{OPT}$ *in the objective function value with* $\epsilon \leq \epsilon_0$ *requires* $2^{\Omega(n_1+n_2)}$ *time, where* OPT *is the optimum and* $\epsilon_0 > 0$ *is an absolute constant. If* $\mathsf{BPP} = \mathsf{P}$*, then the same conclusion holds for randomized algorithms succeeding with probability at least* $2/3$*.*

Now we turn to the application of our framework. This only requires the verification of the dual conditions in Theorem 20. We will show that two prototypical problems, matrix completion and robust PCA, obey the conditions. They belong to the linear inverse problems of form (3.1) with a proper choice of function $H(\cdot)$, which aim at exactly recovering a hidden matrix $\mathbf{X}^*$ with $\mathsf{rank}(\mathbf{X}^*) \leq r$ given a limited number of linear observations of it.

For matrix completion, the linear measurements are of the form $\{\mathbf{X}^*_{ij} : (i,j) \in \Omega\}$, where Ω is the support set which is uniformly distributed among all subsets of $[n_1] \times [n_2]$ of cardinality m. With strong duality, we can either study the exact recoverability of the primal problem (3.1), or investigate the validity of its convex dual (or bi-dual) problem (3.2). Here we study the former with tools from geometric functional analysis. Recall that in the analysis of matrix completion, one typically requires a μ-incoherence condition for a given rank-r matrix $\mathbf{X}^*$ with skinny SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ [56, 194]:

$$\|\mathbf{U}^T\mathbf{e}_i\|_2 \leq \sqrt{\frac{\mu r}{n_1}} \quad \text{for all } i \in [n_1], \qquad \text{and} \qquad \|\mathbf{V}^T\mathbf{e}_i\|_2 \leq \sqrt{\frac{\mu r}{n_2}} \quad \text{for all } i \in [n_2], \tag{3.5}$$

$$\|\mathbf{X}^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}}\sigma_r(\mathbf{X}^*). \tag{3.6}$$

where $\mathbf{e}_i$'s are basis vectors with i-th entry equal to 1 and other entries equal to 0. The incoherence condition claims that information spreads throughout the left and right singular vectors and is standard in the matrix completion literature. Under this standard condition, we have the following results.

Theorem 18 (Matrix Completion)**.** *There exist optimization problems for matrix completion in the forms of* (3.1) *and* (3.2) *that enjoy strong duality with each other and exactly recovers* $\mathbf{X}^*$ *with high probability, provided that* $m = \mathcal{O}(\kappa^2\mu(n_1+n_2)r\log(n_1+n_2)\log_{2\kappa}(n_1+n_2))$ *or* $m = \mathcal{O}(\mu(n_1+n_2)r\log^2(n_1+n_2))$*, where* κ *is the condition number of* $\mathbf{X}^*$*. The sample complexity lower bound is* $\Omega(\mu r(n_1+n_2)\log(n_1+n_2))$*.*

To the best of our knowledge, our result is the first to connect convex matrix completion to non-convex matrix completion, two parallel lines of research that have received significant

Table 3.1: Comparison of matrix completion methods. Here $\kappa = \sigma_1(\mathbf{X}^*)/\sigma_r(\mathbf{X}^*)$ is the condition number of $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$, ϵ is the accuracy such that the output $\widetilde{\mathbf{X}}$ obeys $\|\widetilde{\mathbf{X}} - \mathbf{X}^*\|_F \leq \epsilon$, $n_{(1)} = \max\{n_1, n_2\}$ and $n_{(2)} = \min\{n_1, n_2\}$.

Work	Sample Complexity	Incoherence
[123]	$\mathcal{O}\left(\kappa^4 \mu^2 r^{4.5} n_{(1)} \log n_{(1)} \log\left(\frac{r\|\mathbf{X}^*\|_F}{\epsilon}\right)\right)$	(3.5)
[106]	$\mathcal{O}\left(\mu r n_{(1)}(r + \log\left(\frac{n_{(1)}\|\mathbf{X}^*\|_F}{\epsilon}\right)\frac{\|\mathbf{X}^*\|_F^2}{\sigma_r^2}\right)$	(3.5)
[88]	$\mathcal{O}(\max\{\mu^6\kappa^{16}r^4, \mu^4\kappa^4 r^6\} n_{(1)} \log^2 n_{(1)})$	$\|\mathbf{X}^*_{i:}\|_2 \leq \frac{\mu\|\mathbf{X}^*\|_F}{\sqrt{n_{(2)}}}$
[220]	$\mathcal{O}(r n_{(1)} \kappa^2 \max\left\{\mu \log n_{(2)}, \sqrt{\frac{n_{(1)}}{n_{(2)}}}\mu^2 r^6 \kappa^4\right\}$	(3.5)
[266]	$\mathcal{O}(\mu r^2 n_{(1)} \kappa^2 \max(\mu, \log n_{(1)}))$	(3.5)
[85]	$\mathcal{O}\left(\left(\mu^2 r^4 \kappa^2 + \mu r \log\left(\frac{\|\mathbf{X}^*\|_F}{\epsilon}\right)\right) n_{(1)} \log\left(\frac{\|\mathbf{X}^*\|_F}{\epsilon}\right)\right)$	(3.5)
[265]	$\mathcal{O}\left(\mu r^3 n_{(1)} \log n_{(1)} \log\left(\frac{1}{\epsilon}\right)\right)$	(3.5)
[134]	$\mathcal{O}\left(n_{(2)} r \sqrt{\frac{n_{(1)}}{n_{(2)}}} \kappa^2 \max\left\{\mu \log n_{(2)}, \mu^2 r \sqrt{\frac{n_{(1)}}{n_{(2)}}} \kappa^4\right\}\right)$	(3.5) and (3.6)
[98]	$\mathcal{O}(\mu r n_{(1)} \log^2 n_{(1)})$	(3.5) and (3.6)
[63]	$\mathcal{O}(\mu r n_{(1)} \log^2 n_{(1)})$	(3.5)
Ours	$\mathcal{O}(\kappa^2 \mu r n_{(1)} \log(n_{(1)}) \log_{2\kappa}(n_{(1)}))$	(3.5)
Ours	$\mathcal{O}(\mu r n_{(1)} \log^2 n_{(1)})$	(3.5)
Lower Bound[1]	$\Omega(\mu r n_{(1)} \log n_{(1)})$	(3.5)

attention in the past few years. Table 3.1 compares our results with prior results. Ours match the best known results but further provide strong duality. Also, our results are achieved by a clean framework for a class of related problems.

For robust PCA, instead of studying exact recoverability of problem (3.1) as for matrix completion, we investigate problem (3.2) directly. The robust PCA problem is to recover an incoherent low-rank component $\mathbf{X}^*$ and a sparse component $\mathbf{S}^*$ from their sum [1, 57]. We obtain the following theorem for robust PCA.

Theorem 19 (Robust PCA)**.** *There exists a convex optimization formulation for robust PCA in the form of problem* (3.2) *that exactly recovers the incoherent matrix* $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ *and* $\mathbf{S}^* \in \mathbb{R}^{n_1 \times n_2}$ *with high probability, even if* $\mathsf{rank}(\mathbf{X}^*) = \Theta\left(\frac{\min\{n_1,n_2\}}{\mu \log^2 \max\{n_1,n_2\}}\right)$ *and the size of the support of* $\mathbf{S}^*$ *is* $m = \Theta(n_1 n_2)$*, where the support set of* $\mathbf{S}^*$ *is uniformly distributed among all sets of cardinality* m*, and the incoherence parameter* μ *satisfies the incoherence condition* (3.5) *and* $\|\mathbf{X}^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}}\sigma_r(\mathbf{X}^*)$.

The bounds in Theorem 19 match the best known results in the robust PCA literature when the supports of $\mathbf{S}^*$ are uniformly sampled [57], while our assumption is arguably more intuitive. Note that our results hold even when $\mathbf{X}^*$ is close to full rank and a constant fraction of the entries have noise.

[1]This lower bound is information-theoretic [56].

Independently of our work, [89] developed a framework to analyze the loss surface of low-rank problems, and applied the framework to matrix completion and robust PCA. For matrix completion, their sample complexity is $\mathcal{O}(\kappa^6\mu^4 r^6(n_1+n_2)\log(n_1+n_2))$, significantly larger than our bound. For robust PCA, the number of the outlier entries that their method can tolerate is $\mathcal{O}\left(\frac{n_1 n_2}{\mu r \kappa^5}\right)$, but their result is for deterministic outlier entries and thus are not directly comparable to ours. [262] also studied the robust PCA problem using non-convex optimization, where the outlier entries are also deterministic and the number of outliers that their algorithm can tolerate is $\mathcal{O}\left(\frac{n_1 n_2}{r\kappa}\right)$.

3.1.3 Our techniques

Reduction to low-rank approximation. Our results are inspired by the low-rank approximation problem:

$$\min_{\mathbf{A}\in\mathbb{R}^{n_1\times r},\mathbf{B}\in\mathbb{R}^{r\times n_2}} \frac{1}{2}\|-\widetilde{\mathbf{\Lambda}}-\mathbf{AB}\|_F^2. \tag{3.7}$$

We know that all local solutions of (4.7) are globally optimal and that strong duality holds for any given matrix $-\widetilde{\mathbf{\Lambda}}\in\mathbb{R}^{n_1\times n_2}$ [99]. To extend this property to our more general problem (3.1), our main insight is to reduce problem (3.1) to the form of (4.7) using the ℓ_2-regularization term. While some prior work attempted to apply a similar reduction, their conclusions either depended on unrealistic conditions on local solutions, e.g., all local solutions are rank-deficient [99, 104], or their conclusions relied on strong assumptions on the objective functions, e.g., that the objective functions are twice-differentiable [105]. For example, the conditions that all local solutions are rank-deficient break down even for the PCA problem, and the assumptions that the objective function is twice-differential preclude $H(\cdot)$ in (3.1) and (3.3) from encoding hard constraints. Instead, our general results formulate strong duality via the existence of a dual certificate $\widetilde{\mathbf{\Lambda}}$. For concrete applications, the existence of a dual certificate is then converted to mild assumptions, e.g., that the number of measurements is sufficiently large and the positions of measurements are randomly distributed. We will illustrate the importance of randomness below.

The blessing of randomness. The desired dual certificate $\widetilde{\mathbf{\Lambda}}$ may not exist in the deterministic world. A hardness result [193] shows that for the problem of weighted low-rank approximation, which can be cast in the form of (3.1), without some randomization in the measurements made on the underlying low rank matrix, it is NP-hard to achieve a good objective value, not to mention to achieve strong duality. A similar result was shown for deterministic matrix completion [107]. Thus we should utilize randomness to analyze the existence of a dual certificate. For specific applications such as matrix completion, the assumption that the measurements are random is standard, under which, the angle between the space Ω (the space of matrices which are consistent with observations) and the space $\mathcal{T}$ (the space of matrices which are low-rank) is small with high probability, namely, $\mathbf{X}^*$ is almost the unique low-rank matrix that is consistent with the measurements. Thus, our dual certificate can be represented as another form of a convergent Neumann series concerning the projection operators on the spaces Ω and $\mathcal{T}$; otherwise, the same construction of Neumann series may diverge as the norm concerning the projection operators on the spaces Ω and $\mathcal{T}$ is larger than 1 in the deterministic worst case. The remainder of the proof

is to show that such a construction obeys the dual conditions. To show this, we use the fact that the subspace Ω and the complement space $\mathcal{T}^\perp$ are almost orthogonal when the sample size is sufficiently large. This implies the projection of our dual certificate on the space $\mathcal{T}^\perp$ has a very small norm, which exactly matches the dual conditions.

Non-convex geometric analysis. Strong duality implies that the primal problem (3.1) and its bi-dual problem (3.2) have exactly the same solutions in the sense that $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \widetilde{\mathbf{X}}$. Thus, to show exact recoverability of linear inverse problems such as matrix completion and robust PCA, it suffices to study either the non-convex primal problem (3.1) or its convex counterpart (3.2). Here we do the former analysis for matrix completion. We mention that traditional techniques [56, 61, 194] for convex optimization break down for our non-convex problem, since the subgradient of a non-convex objective function may not even exist [44]. Instead, we apply tools from geometric analysis [234] to analyze the geometry of problem (3.1). Our non-convex geometric analysis is in stark contrast to prior techniques of convex geometric analysis [236] where convex combinations of non-convex constraints were used to define the Minkowski functional (e.g., in the definition of atomic norm) while our method uses the non-convex constraint itself.

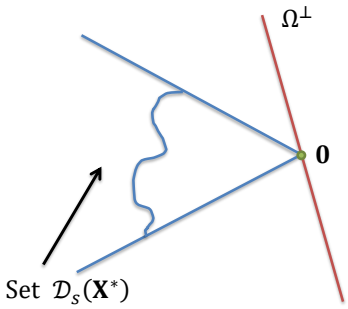


Figure 3.2: Feasibility.

For matrix completion, problem (3.1) has two hard constraints: a) the rank of the output matrix should be no larger than r, as implied by the form of $\mathbf{AB}$; b) the output matrix should be consistent with the sampled measurements, i.e., $\mathcal{P}_\Omega(\mathbf{AB}) = \mathcal{P}_\Omega(\mathbf{X}^*)$. We study the feasibility condition of problem (3.1) from a geometric perspective: $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$ is the unique optimal solution to problem (3.1) if and only if starting from $\mathbf{X}^*$, either the rank of $\mathbf{X}^* + \mathbf{D}$ or $\|\mathbf{X}^* + \mathbf{D}\|_F$ increases for all directions $\mathbf{D}$'s in the constraint set $\Omega^\perp = \{\mathbf{D} \in \mathbb{R}^{n_1 \times n_2} : \mathcal{P}_\Omega(\mathbf{X}^* + \mathbf{D}) = \mathcal{P}_\Omega(\mathbf{X}^*)\}$. This can be geometrically interpreted as the requirement that the set $\mathcal{D}_\mathcal{S}(\mathbf{X}^*) = \{\mathbf{X} - \mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\mathbf{X}) \leq r, \|\mathbf{X}\|_F \leq \|\mathbf{X}^*\|_F\}$ and the constraint set $\Omega^\perp$ must intersect uniquely at $\mathbf{0}$ (see Figure 3.2). This can then be shown by a dual certificate argument.

Putting things together. We summarize our new analytical framework with Figure 3.3.

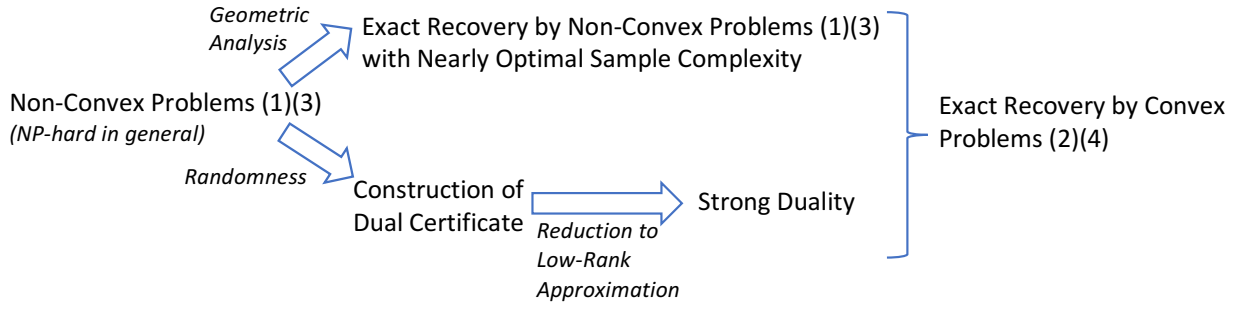


Figure 3.3: New analytical framework for non-convex matrix factorization.

Other techniques. An alternative method is to investigate the exact recoverability of problem (3.2) via standard convex analysis. We find that the sub-differential of our induced function $\|\cdot\|_{r*}$ has similar properties as that of the nuclear norm. With this observation, we prove the validity of robust PCA in the form of (3.2) by combining this property of $\|\cdot\|_{r*}$ with standard techniques from [57].

3.1.4 Experimental results

Experiments on synthetic data

We verify the exact recoverability of the r^* minimization (3.16) and the nuclear norm minimization [55] on the matrix completion problem by experiments on the synthetic data. The synthetic data are generated as follows. We construct the ground-truth matrix $\mathbf{X}^* = \mathbf{AB}$ as a product of matrices $\mathbf{A}$ of size $n \times r$ and $\mathbf{B}$ of size $r \times n$, whose entries are i.i.d. $\mathcal{N}(0, 1)$. We then uniformly sample m entries from $\mathbf{X}^*$ as the observations. For each size of the problem ($\mathbf{X}^*$ is 100×100 or 200×200), we test with different rank ratios r/n and observation ratios m/n^2. Each set of parameters is run 5 times, and the algorithm is said to succeed if $\|\widetilde{\mathbf{X}} - \mathbf{X}^*\|_F / \|\mathbf{X}^*\|_F \leq 10^{-3}$ for all five experiments, where $\widetilde{\mathbf{X}}$ is the output of the algorithms. We set the parameter r in r^* minimization (3.16) as the true rank, and use the Augmented Lagrange Multiplier Method [62] for optimization, where the proximal map of r^* norm is computed as in [99].

The two figures in Figure 3.4 plots the fraction of exact recoveries: the white region represents the exact recovery by nuclear norm minimization, the white+gray region represents the exact recovery by r^* minimization (3.16), and the black region indicates the failure for both algorithms. It is clear that both algorithms succeed for a wide range of parameters. The success region of r^* minimization is slightly larger and contains the success region of the nuclear norm minimization for both 100×100 and 200×200 matrix completion problems.

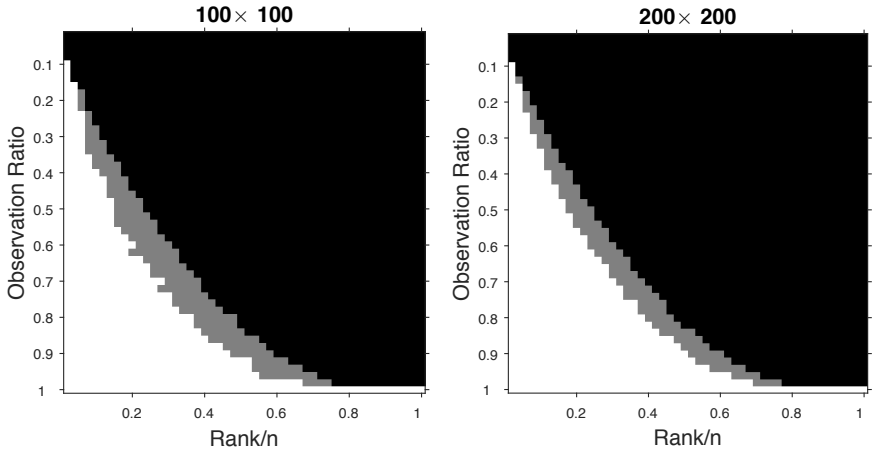


Figure 3.4: Exact recoverability of matrix completion with varying ranks and sample sizes. **White Region:** nuclear norm minimization succeeds. **White and Gray Regions:** r^* minimization succeeds. **Black Region:** both algorithms fail. It shows that the success region of r^* minimization slightly contains that of the nuclear minimization method.

Experiments on real data

Table 3.2: Relative error by matrix completion algorithms on the Hopkins 155 dataset.

		$m = 0.05n_1n_2$		$m = 0.1n_1n_2$	
#Task	Size	Nuclear	r^*	Nuclear	r^*
Average over all 155 tasks	–	0.8249	0.8114	0.5689	0.5409
#1	59×459	0.7438	0.5948	0.5115	0.5117
#2	49×482	0.8235	0.6564	0.6371	0.5919
#3	49×153	0.7803	0.9174	0.5386	0.5386
#4	49×379	0.8500	0.9583	0.7287	0.7691
#5	49×432	0.8174	0.6353	0.4476	0.4477

To verify the performance of the algorithms on real data, we conduct experiments on the Hopkins 155 dataset. This dataset consists of 155 tasks/matrices, each of which consists of multiple data points drawn from 2 or 3 moving objects. The trajectory of each object lies in a low-dimensional subspace, so the matrix for each task is supposed to be approximately low rank. We uniformly sample m entries from the matrix as our observations and run the matrix completion algorithms. The parameter r in the r^* minimization is set as the number of moving objects which is known to us in the dataset.

Table 3.2 shows the the relative errors $\|\widehat{\mathbf{X}} - \mathbf{X}^*\|_F / \|\mathbf{X}^*\|_F$ of the nuclear norm minimization and r^* minimization (3.16). On average, r^* minimization slightly outperforms the competitor, while sometimes the nuclear norm minimization is better. Table 3.2 also shows the errors on the first five tasks in the dataset. It shows that when the number of observations is relatively large (10% observations or higher), the performance of the two algorithms are competitive to each other. When the number of observations is small (5% observed entries), there is a larger variance, but on average r^* minimization has an slight advantage.

3.1.5 Proofs of our main results

New framework of strong duality

We consider the problem

$$(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = \underset{\mathbf{A}\in\mathbb{R}^{n_1\times r}, \mathbf{B}\in\mathbb{R}^{r\times n_2}}{\operatorname{argmin}} H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2,$$

We first consider an easy case where $H(\mathbf{AB}) = \frac{1}{2}\|\widehat{\mathbf{Y}}\|_F^2 - \langle\widehat{\mathbf{Y}}, \mathbf{AB}\rangle$ for a fixed $\widehat{\mathbf{Y}}$, leading to the objective function $\frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{AB}\|_F^2$. For this case, we establish the following lemma.

Lemma 31. *For any given matrix $\widehat{\mathbf{Y}} \in \mathbb{R}^{n_1\times n_2}$, any local minimum of $f(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{AB}\|_F^2$ over $\mathbf{A} \in \mathbb{R}^{n_1\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times n_2}(r \le \min\{n_1, n_2\})$ is globally optimal, given by $\mathsf{svd}_r(\widehat{\mathbf{Y}})$. The objective function $f(\mathbf{A}, \mathbf{B})$ around any saddle point has a negative second-order directional curvature. Moreover, $f(\mathbf{A}, \mathbf{B})$ has no local maximum.*[2]

[2]Prior work studying the loss surface of low-rank matrix approximation assumes that the matrix $\widetilde{\mathbf{\Lambda}}$ is of full

Proof. By the assumption of the lemma, $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ is a local minimizer of $L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}}) = \frac{1}{2}\| - \widetilde{\mathbf{\Lambda}} - \mathbf{AB}\|_F^2 + c(\widetilde{\mathbf{\Lambda}})$, where $c(\widetilde{\mathbf{\Lambda}})$ is a function that is independent of $\mathbf{A}$ and $\mathbf{B}$. So according to Lemma 31, $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = \operatorname{argmin}_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}})$, namely, $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ globally minimizes $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$ when $\mathbf{\Lambda}$ is fixed to $\widetilde{\mathbf{\Lambda}}$. Furthermore, $\widetilde{\mathbf{\Lambda}} \in \partial_{\mathbf{X}} H(\mathbf{X})|_{\mathbf{X}=\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}}$ implies that $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} \in \partial_{\mathbf{\Lambda}} H^*(\mathbf{\Lambda})|_{\mathbf{\Lambda}=\widetilde{\mathbf{\Lambda}}}$ by the convexity of function H, meaning that $\mathbf{0} \in \partial_{\mathbf{\Lambda}} L(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \mathbf{\Lambda})$. So $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\mathbf{\Lambda}} L(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \mathbf{\Lambda})$ due to the concavity of $L(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \mathbf{\Lambda})$ w.r.t. variable $\mathbf{\Lambda}$. Thus $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$.

We now prove the strong duality. By the fact that $F(\mathbf{A}, \mathbf{B}) = \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$ and that $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\mathbf{\Lambda}} L(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \mathbf{\Lambda})$, we have

$$F(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = L(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{\Lambda}}) \le L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}}), \ \ \forall \mathbf{A}, \mathbf{B}.$$

where the inequality holds because $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of L. So on the one hand, we have

$$\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) = F(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \le \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}}) \le \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}).$$

On the other hand, by weak duality,

$$\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) \ge \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}).$$

Therefore, $\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) = \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$, i.e., strong duality holds. Hence,

$$\begin{aligned}
\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} &= \operatorname*{argmin}_{\mathbf{AB}} L(\mathbf{A}, \mathbf{B}, \widetilde{\mathbf{\Lambda}}) \\
&= \operatorname*{argmin}_{\mathbf{AB}} \frac{1}{2}\| - \widetilde{\mathbf{\Lambda}} - \mathbf{AB}\|_F^2 - \frac{1}{2}\|\widetilde{\mathbf{\Lambda}}\|_F^2 - H^*(\widetilde{\mathbf{\Lambda}}) \\
&= \operatorname*{argmin}_{\mathbf{AB}} \frac{1}{2}\| - \widetilde{\mathbf{\Lambda}} - \mathbf{AB}\|_F^2 \\
&= \mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}}),
\end{aligned}$$

as desired. □

Given this lemma, we can reduce $F(\mathbf{A}, \mathbf{B})$ to the form $\frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{AB}\|_F^2$ for some $\widehat{\mathbf{Y}}$ plus an extra term:

$$\begin{aligned}
F(\mathbf{A}, \mathbf{B}) &= \frac{1}{2}\|\mathbf{AB}\|_F^2 + H(\mathbf{AB}) = \frac{1}{2}\|\mathbf{AB}\|_F^2 + H^{**}(\mathbf{AB}) = \max_{\mathbf{\Lambda}} \frac{1}{2}\|\mathbf{AB}\|_F^2 + \langle \mathbf{\Lambda}, \mathbf{AB} \rangle - H^*(\mathbf{\Lambda}) \\
&= \max_{\mathbf{\Lambda}} \frac{1}{2}\| - \mathbf{\Lambda} - \mathbf{AB}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - H^*(\mathbf{\Lambda}) \triangleq \max_{\mathbf{\Lambda}} L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda}),
\end{aligned} \tag{3.8}$$

rank and does not have the same singular values [29]. In this work, we generalize this result by removing these two assumptions.

where we define $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}) \triangleq \frac{1}{2}\|-\mathbf{\Lambda}-\mathbf{AB}\|_F^2-\frac{1}{2}\|\mathbf{\Lambda}\|_F^2-H^*(\mathbf{\Lambda})$ as the Lagrangian of problem (**P**),[3] and the second equality holds because H is closed and convex w.r.t. the argument $\mathbf{AB}$. For any fixed value of $\mathbf{\Lambda}$, by Lemma 31, any local minimum of $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$ is globally optimal, because minimizing $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$ is equivalent to minimizing $\frac{1}{2}\|-\mathbf{\Lambda}-\mathbf{AB}\|_F^2$ for a fixed $\mathbf{\Lambda}$.

The remaining part of our analysis is to choose a proper $\widetilde{\mathbf{\Lambda}}$ such that $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$, so that $\min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})$ and problem (**P**) have the same optimal solution $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$. For this, we introduce the following condition, and later we will show that the condition holds with high probability.

Condition 1. *For a solution $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ to problem* (**P**)*, there exists an $\widetilde{\mathbf{\Lambda}} \in \partial_{\mathbf{X}} H(\mathbf{X})|_{\mathbf{X}=\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}}$ such that*

$$-\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T = \widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{B}}^T \quad \textit{and} \quad \widetilde{\mathbf{A}}^T(-\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) = \widetilde{\mathbf{A}}^T\widetilde{\mathbf{\Lambda}}. \tag{3.9}$$

Explanations of Condition 2. We note that

$$\nabla_{\mathbf{A}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}) = \mathbf{ABB}^T + \mathbf{\Lambda B}^T \text{ and } \nabla_{\mathbf{B}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}) = \mathbf{A}^T\mathbf{AB} + \mathbf{A}^T\mathbf{\Lambda}$$

for a fixed $\mathbf{\Lambda}$. In particular, if we set $\mathbf{\Lambda}$ to be the $\widetilde{\mathbf{\Lambda}}$ in (4.18), then $\nabla_{\mathbf{A}} L(\mathbf{A},\widetilde{\mathbf{B}},\widetilde{\mathbf{\Lambda}})|_{\mathbf{A}=\widetilde{\mathbf{A}}} = \mathbf{0}$ and $\nabla_{\mathbf{B}} L(\widetilde{\mathbf{A}},\mathbf{B},\widetilde{\mathbf{\Lambda}})|_{\mathbf{B}=\widetilde{\mathbf{B}}} = \mathbf{0}$. So Condition 2 implies that $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ is either a saddle point or a local minimizer of $L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})$ as a function of $(\mathbf{A},\mathbf{B})$ for the fixed $\widetilde{\mathbf{\Lambda}}$.

The following lemma states that if it is a local minimizer, then strong duality holds.

Lemma 32 (Dual certificate)**.** *Let $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ be a global minimizer of $F(\mathbf{A},\mathbf{B})$. If there exists a dual certificate $\widetilde{\mathbf{\Lambda}}$ satisfying Condition 2 and the pair $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ is a local minimizer of $L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})$ for the fixed $\widetilde{\mathbf{\Lambda}}$, then strong duality holds. Moreover, we have the relation $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}})$.*

Proof. By the assumption of the lemma, we can show that $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point to the Lagrangian $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$; see Appendix 3.1.5. To show strong duality, by the fact that $F(\mathbf{A},\mathbf{B}) = \max_{\mathbf{\Lambda}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$ and that $\widetilde{\mathbf{\Lambda}} = \operatorname{argmax}_{\mathbf{\Lambda}} L(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\mathbf{\Lambda})$, we have $F(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}}) = L(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{\Lambda}}) \le L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}})$, for any $\mathbf{A},\mathbf{B}$, where the inequality holds because $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}},\widetilde{\mathbf{\Lambda}})$ is a primal-dual saddle point of L. So on the one hand, $\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}) = F(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}}) \le \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}}) \le \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$. On the other hand, by weak duality, we have $\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}) \ge \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$. Therefore, $\min_{\mathbf{A},\mathbf{B}} \max_{\mathbf{\Lambda}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}) = \max_{\mathbf{\Lambda}} \min_{\mathbf{A},\mathbf{B}} L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$, i.e., strong duality holds. Therefore, $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \operatorname{argmin}_{\mathbf{AB}} L(\mathbf{A},\mathbf{B},\widetilde{\mathbf{\Lambda}}) = \operatorname{argmin}_{\mathbf{AB}} \frac{1}{2}\|\mathbf{AB}\|_F^2 + \langle\widetilde{\mathbf{\Lambda}},\mathbf{AB}\rangle - H^*(\widetilde{\mathbf{\Lambda}}) = \operatorname{argmin}_{\mathbf{AB}} \frac{1}{2}\|-\widetilde{\mathbf{\Lambda}}-\mathbf{AB}\|_F^2 = \mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}})$, as desired. □

This lemma then leads to the following theorem.

Theorem 20. *Denote by $(\widetilde{\mathbf{A}},\widetilde{\mathbf{B}})$ the optimal solution of problem* (**P**)*. Define a matrix space*

$$\mathcal{T} \triangleq \{\widetilde{\mathbf{A}}\mathbf{X}^T + \mathbf{Y}\widetilde{\mathbf{B}},\ \mathbf{X}\in\mathbb{R}^{n_2\times r},\ \mathbf{Y}\in\mathbb{R}^{n_1\times r}\}.$$

Then strong duality holds for problem (**P**)*, provided that there exists $\widetilde{\mathbf{\Lambda}}$ such that*

$$(1)\ \widetilde{\mathbf{\Lambda}} \in \partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) \triangleq \Psi, \qquad (2)\ \mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}, \qquad (3)\ \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}). \tag{3.10}$$

[3] One can easily check that $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda}) = \min_{\mathbf{M}} L'(\mathbf{A},\mathbf{B},\mathbf{M},\mathbf{\Lambda})$, where $L'(\mathbf{A},\mathbf{B},\mathbf{M},\mathbf{\Lambda})$ is the Lagrangian of the constraint optimization problem $\min_{\mathbf{A},\mathbf{B},\mathbf{M}} \frac{1}{2}\|\mathbf{AB}\|_F^2 + H(\mathbf{M})$, s.t. $\mathbf{M} = \mathbf{AB}$. With a little abuse of notation, we call $L(\mathbf{A},\mathbf{B},\mathbf{\Lambda})$ the Lagrangian of the unconstrained problem (**P**) as well.

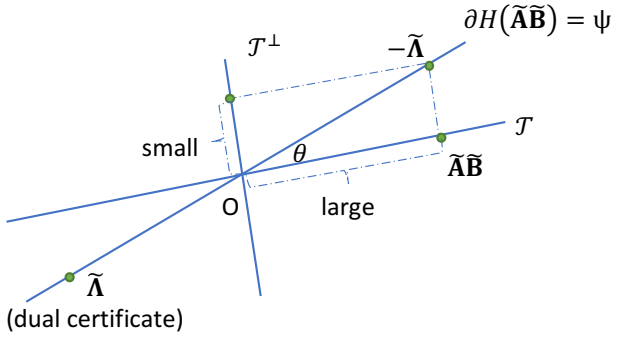


Figure 3.5: Geometry of dual condition (3.10) for general matrix factorization problems.

Proof. The proof idea is to construct a dual certificate $\widetilde{\mathbf{\Lambda}}$ so that the conditions in Lemma 67 hold. $\widetilde{\mathbf{\Lambda}}$ should satisfy the following:

$$\begin{aligned}
&\text{(a)} \quad \widetilde{\mathbf{\Lambda}} \in \partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}), \qquad \text{(by Condition 2)} \\
&\text{(b)} \quad (\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}})\widetilde{\mathbf{B}}^T = \mathbf{0} \quad \text{and} \quad \widetilde{\mathbf{A}}^T(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}}) = \mathbf{0}, \qquad \text{(by Condition 2)} \\
&\text{(c)} \quad \widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}}). \qquad \text{(by the local minimizer assumption and Lemma 31)}
\end{aligned} \tag{3.11}$$

It turns out that for any matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, $\mathcal{P}_{\mathcal{T}^\perp}\mathbf{M} = (\mathbf{I} - \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^\dagger)\mathbf{M}(\mathbf{I} - \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^\dagger)$ and so $\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{M}\| \le \|\mathbf{M}\|$, a fact that we will frequently use in the sequel. Denote by $\mathcal{U}$ the left singular space of $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$ and $\mathcal{V}$ the right singular space. Then the linear space $\mathcal{T}$ can be equivalently represented as $\mathcal{T} = \mathcal{U} + \mathcal{V}$. Therefore, $\mathcal{T}^\perp = (\mathcal{U} + \mathcal{V})^\perp = \mathcal{U}^\perp \cap \mathcal{V}^\perp$. With this, we note that: (b) $(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}})\widetilde{\mathbf{B}}^T = \mathbf{0}$ and $\widetilde{\mathbf{A}}^T(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}}) = \mathbf{0}$ imply $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}} \in \mathsf{Null}(\widetilde{\mathbf{A}}^T) = \mathsf{Col}(\widetilde{\mathbf{A}})^\perp$ and $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}} \in \mathsf{Row}(\widetilde{\mathbf{B}})^\perp$ (so $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \widetilde{\mathbf{\Lambda}} \in \mathcal{T}^\perp$), and vice versa. And (c) $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}})$ implies that for an orthogonal decomposition $-\widetilde{\mathbf{\Lambda}} = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \mathbf{E}$, where $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} \in \mathcal{T}$, and $\mathbf{E} \in \mathcal{T}^\perp$, we have $\|\mathbf{E}\| < \sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$. Conversely, $\|\mathbf{E}\| < \sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$ and condition (b) imply $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathsf{svd}_r(-\widetilde{\mathbf{\Lambda}})$. Therefore, the dual conditions in (4.19) are equivalent to (1) $\widetilde{\mathbf{\Lambda}} \in \partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) \triangleq \Psi$; (2) $\mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$; (3) $\|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$. □

To show the dual condition in Theorem 20, intuitively, we need to show that the angle θ between subspace $\mathcal{T}$ and Ψ is small (see Figure 3.5) for a specific function $H(\cdot)$. In the following, we will demonstrate applications that, with randomness, obey this dual condition with high probability.

Proofs of matrix completion

In matrix completion, there is a hidden matrix $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ with rank r. We are given measurements $\{\mathbf{X}^*_{ij} : (i,j) \in \Omega\}$, where $\Omega \sim \mathsf{Uniform}(m)$, i.e., Ω is sampled uniformly at random from all subsets of $[n_1] \times [n_2]$ of cardinality m. The goal is to exactly recover $\mathbf{X}^*$ with high probability. Here we apply our unified framework in the last section to matrix completion, by setting $H(\cdot) = \mathbf{I}_{\{\mathbf{M}:\mathcal{P}_\Omega(\mathbf{M}) = \mathcal{P}_\Omega(\mathbf{X}^*)\}}(\cdot)$.

A quantity governing the difficulties of matrix completion is the incoherence parameter μ. Intuitively, matrix completion is possible only if the information spreads evenly throughout the

low-rank matrix. This intuition is captured by the incoherence conditions. Formally, denote by $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ the skinny SVD of a fixed $n_1 \times n_2$ matrix $\mathbf{X}$ of rank r. Candès et al. [55, 57, 194, 255] introduced the μ-incoherence condition (3.5) to the low-rank matrix $\mathbf{X}$. For conditions (3.5), it can be shown that $1 \leq \mu \leq \frac{n_{(1)}}{r}$. The condition holds for many random matrices with incoherence parameter μ about $\sqrt{r \log n_{(1)}}$ [134].

We first propose a non-convex optimization problem whose unique solution is indeed the ground truth $\mathbf{X}^*$, and then apply our framework to show that strong duality holds for this non-convex optimization and its bi-dual optimization problem.

Theorem 21 (Uniqueness of solution)**.** *Let $\Omega \sim \mathsf{Uniform}(m)$ be the support set uniformly distributed among all sets of cardinality m. Suppose that $m \geq c\kappa^2 \mu n_{(1)} r \log n_{(1)} \log_{2\kappa} n_{(1)}$ for an absolute constant c and $\mathbf{X}^*$ obeys μ-incoherence (3.5). Then $\mathbf{X}^*$ is the unique solution of non-convex optimization*

$$\min_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{A}\mathbf{B}\|_F^2, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{A}\mathbf{B}) = \mathcal{P}_\Omega(\mathbf{X}^*), \tag{3.12}$$

with probability at least $1 - n_{(1)}^{-10}$.

Proof. We note that a recovery result under the Bernoulli model automatically implies a corresponding result for the uniform model [57]. So in the following, we assume the Bernoulli model.

Consider the feasibility of the matrix completion problem:

$$\text{Find a matrix } \mathbf{X} \in \mathbb{R}^{n_1 \times n_2} \text{ such that } \quad \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*), \quad \|\mathbf{X}\|_F \leq \|\mathbf{X}^*\|_F, \quad \mathsf{rank}(\mathbf{X}) \leq r. \tag{3.13}$$

Note that if $\mathbf{X}^*$ is the unique solution of (3.13), then $\mathbf{X}^*$ is the unique solution of (3.12). We now show the former. Our proof first identifies a feasibility condition for problem (3.13), and then shows that $\mathbf{X}^*$ is the only matrix that obeys this feasibility condition when the sample size is large enough. We denote by

$$\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) = \{\mathbf{X} - \mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\mathbf{X}) \leq r,\ \|\mathbf{X}\|_F \leq \|\mathbf{X}^*\|_F\},$$

and

$$\mathcal{T} = \{\mathbf{U}\mathbf{X}^T + \mathbf{Y}\mathbf{V}^T,\ \mathbf{X} \in \mathbb{R}^{n_2 \times r},\ \mathbf{Y} \in \mathbb{R}^{n_1 \times r}\},$$

where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the skinny SVD of $\mathbf{X}^*$.

Before proceeding, we first study a property of sub-gradient of the $r*$ norm.

Lemma 33. *Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the skinny SVD of matrix $\mathbf{X}^*$ of rank r. The subdifferential of $\|\cdot\|_{r*}$ evaluated at $\mathbf{X}^*$ is given by*

$$\partial\|\mathbf{X}^*\|_{r*} = \{\mathbf{X}^* + \mathbf{W} : \mathbf{U}^T\mathbf{W} = \mathbf{0}, \mathbf{W}\mathbf{V} = \mathbf{0}, \|\mathbf{W}\| \leq \sigma_r(\mathbf{X}^*)\}.$$

Proof. Note that for any fixed function $f(\cdot)$, the set of all optimal solutions of the problem

$$f^*(\mathbf{X}^*) = \max_{\mathbf{Y}} \langle \mathbf{X}^*, \mathbf{Y} \rangle - f(\mathbf{Y}) \tag{3.14}$$

form the subdifferential of the conjugate function $f^*(\cdot)$ evaluated at $\mathbf{X}^*$. Set $f(\cdot)$ to be $\frac{1}{2}\|\cdot\|_r^2$ and notice that the function $\frac{1}{2}\|\cdot\|_r^2$ is unitarily invariant. By Von Neumann's trace inequality, the optimal solutions to problem (3.14) are given by

$$[\mathbf{U}, \mathbf{U}^\perp]\mathsf{diag}([\sigma_1(\mathbf{Y}), ..., \sigma_r(\mathbf{Y}), \sigma_{r+1}(\mathbf{Y}), ..., \sigma_{n_{(2)}}(\mathbf{Y})])[\mathbf{V}, \mathbf{V}^\perp]^T,$$

where $\{\sigma_i(\mathbf{Y})\}_{i=r+1}^{n_{(2)}}$ can be any value no larger than $\sigma_r(\mathbf{Y})$ and $\{\sigma_i(\mathbf{Y})\}_{i=1}^r$ are given by the optimal solution to the problem

$$\max_{\{\sigma_i(\mathbf{Y})\}_{i=1}^r} \sum_{i=1}^r \sigma_i(\mathbf{X}^*)\sigma_i(\mathbf{Y}) - \frac{1}{2}\sum_{i=1}^r \sigma_i^2(\mathbf{Y}).$$

The solution is unique such that $\sigma_i(\mathbf{Y}) = \sigma_i(\mathbf{X}^*)$, $i = 1, 2, ..., r$. The proof is complete. □

We have the following proposition for the feasibility of problem (3.13).

Proposition 1 (Feasibility condition)**.** $\mathbf{X}^*$ *is the unique feasible solution to problem* (3.13) *if* $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$.

Proof. Notice that problem (3.13) is equivalent to another feasibility problem

$$\text{Find a matrix } \mathbf{D} \in \mathbb{R}^{n_1 \times n_2} \text{ such that } \quad \mathsf{rank}(\mathbf{X}^* + \mathbf{D}) \le r, \ \ \|\mathbf{X}^* + \mathbf{D}\|_F \le \|\mathbf{X}^*\|_F, \ \ \mathbf{D} \in \Omega^\perp.$$

Suppose that $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$. Since $\mathsf{rank}(\mathbf{X}^* + \mathbf{D}) \le r$ and $\|\mathbf{X}^* + \mathbf{D}\|_F \le \|\mathbf{X}^*\|_F$ are equivalent to $\mathbf{D} \in \mathcal{D}_{\mathcal{S}}(\mathbf{X}^*)$, and note that $\mathbf{D} \in \Omega^\perp$, we have $\mathbf{D} = \mathbf{0}$, which means $\mathbf{X}^*$ is the unique feasible solution to problem (3.13). □

The remainder of the proof is to show $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$. To proceed, we note that

$$\begin{aligned}
\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) &= \left\{\mathbf{X} - \mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2} : \mathsf{rank}(\mathbf{X}) \le r, \ \frac{1}{2}\|\mathbf{X}\|_F^2 \le \frac{1}{2}\|\mathbf{X}^*\|_F^2\right\} \\
&\subseteq \{\mathbf{X} - \mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2} : \|\mathbf{X}\|_{r*} \le \|\mathbf{X}^*\|_{r*}\} \quad \left(\text{since } \frac{1}{2}\|\mathbf{Y}\|_F^2 = \|\mathbf{Y}\|_{r*}\right) \\
&\triangleq \mathcal{D}_{\mathcal{S}_*}(\mathbf{X}^*).
\end{aligned}$$

We now show that

$$\mathcal{D}_{\mathcal{S}_*}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}, \tag{3.15}$$

when $m \ge c\kappa^2 \mu r n_{(1)} \log_{2\kappa}(n_{(1)}) \log(n_{(1)})$, which will prove $\mathcal{D}_{\mathcal{S}}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$ as desired.

By Lemma 35, there exists a $\mathbf{\Lambda}$ such that

$$\begin{aligned}
&(1) \qquad \mathbf{\Lambda} \in \Omega, \\
&(2) \qquad \mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) = \mathbf{X}^*, \\
&(3) \qquad \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}\| < \frac{2}{3}\sigma_r(\mathbf{X}^*).
\end{aligned}$$

Consider any $\mathbf{D} \in \Omega^\perp$ such that $\mathbf{D} \neq \mathbf{0}$. By Lemma 33, for any $\mathbf{W} \in \mathcal{T}^\perp$ and $\|\mathbf{W}\| \leq \sigma_r(\mathbf{X}^*)$,

$$\|\mathbf{X}^* + \mathbf{D}\|_{r*} \geq \|\mathbf{X}^*\|_{r*} + \langle \mathbf{X}^* + \mathbf{W}, \mathbf{D} \rangle.$$

Since $\langle \mathbf{W}, \mathbf{D} \rangle = \langle \mathcal{P}_{\mathcal{T}^\perp} \mathbf{W}, \mathbf{D} \rangle = \langle \mathbf{W}, \mathcal{P}_{\mathcal{T}^\perp} \mathbf{D} \rangle$, we can choose $\mathbf{W}$ such that

$$\langle \mathbf{W}, \mathbf{D} \rangle = \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_*.$$

Then

$$\begin{aligned}
\|\mathbf{X}^* + \mathbf{D}\|_{r*} &\geq \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathbf{X}^*, \mathbf{D} \rangle \\
&= \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathbf{X}^* + \mathbf{\Lambda}, \mathbf{D} \rangle \quad \text{(since } \mathbf{\Lambda} \in \Omega \text{ and } \mathbf{D} \in \Omega^\perp\text{)} \\
&= \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathbf{X}^* + \mathcal{P}_{\mathcal{T}} \mathbf{\Lambda}, \mathbf{D} \rangle + \langle \mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}, \mathbf{D} \rangle \\
&= \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}, \mathbf{D} \rangle \quad \text{(by condition (2))} \\
&= \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathcal{P}_{\mathcal{T}^\perp} \mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}, \mathbf{D} \rangle \\
&= \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* + \langle \mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}, \mathcal{P}_{\mathcal{T}^\perp} \mathbf{D} \rangle \\
&\geq \|\mathbf{X}^*\|_{r*} + \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* - \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{\Lambda}\| \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* \quad \text{(by Hölder's inequality)} \\
&\geq \|\mathbf{X}^*\|_{r*} + \frac{1}{3} \sigma_r(\mathbf{X}^*) \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{D}\|_* \quad \text{(by condition (3))}.
\end{aligned}$$

So if $\mathcal{T} \cap \Omega^\perp = \{\mathbf{0}\}$, since $\mathbf{D} \in \Omega^\perp$ and $\mathbf{D} \neq \mathbf{0}$, we have $\mathbf{D} \notin \mathcal{T}$. Therefore,

$$\|\mathbf{X}^* + \mathbf{D}\|_{r*} > \|\mathbf{X}^*\|_{r*}$$

which then leads to $\mathcal{D}_{\mathcal{S}_*}(\mathbf{X}^*) \cap \Omega^\perp = \{\mathbf{0}\}$.

The rest of proof is to show that $\mathcal{T} \cap \Omega^\perp = \{\mathbf{0}\}$. We have the following lemma.

Lemma 34. *Assume that $\Omega \sim \mathsf{Ber}(p)$ and the incoherence condition (3.5) holds. Then with probability at least $1 - n_{(1)}^{-10}$, we have $\|\mathcal{P}_{\Omega^\perp} \mathcal{P}_{\mathcal{T}}\| \leq \sqrt{1 - p + \epsilon p}$, provided that*

$$p \geq C_0 \epsilon^{-2} (\mu r \log n_{(1)}) / n_{(2)},$$

where C_0 is an absolute constant.

Proof. If $\Omega \sim \mathsf{Ber}(p)$, we have, by Theorem 23, that with high probability

$$\|\mathcal{P}_{\mathcal{T}} - p^{-1} \mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega} \mathcal{P}_{\mathcal{T}}\| \leq \epsilon,$$

provided that $p \geq C_0 \epsilon^{-2} \frac{\mu r \log n_{(1)}}{n_{(2)}}$. Note, however, that since $\mathcal{I} = \mathcal{P}_{\Omega} + \mathcal{P}_{\Omega^\perp}$,

$$\mathcal{P}_{\mathcal{T}} - p^{-1} \mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega} \mathcal{P}_{\mathcal{T}} = p^{-1} (\mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega^\perp} \mathcal{P}_{\mathcal{T}} - (1 - p) \mathcal{P}_{\mathcal{T}})$$

and, therefore, by the triangle inequality

$$\|\mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega^\perp} \mathcal{P}_{\mathcal{T}}\| \leq \epsilon p + (1 - p).$$

Since $\|\mathcal{P}_{\Omega^\perp} \mathcal{P}_{\mathcal{T}}\|^2 \leq \|\mathcal{P}_{\mathcal{T}} \mathcal{P}_{\Omega^\perp} \mathcal{P}_{\mathcal{T}}\|$, the proof is completed. □

We note that $\|\mathcal{P}_{\Omega^\perp}\mathcal{P}_{\mathcal{T}}\| < 1$ implies $\Omega^\perp \cap \mathcal{T} = \{\mathbf{0}\}$. The proof is completed. □

Given the non-convex problem, we are ready to state our main theorem for matrix completion.

Theorem 22 (Efficient matrix completion)**.** *Let $\Omega \sim \mathsf{Uniform}(m)$ be the support set uniformly distributed among all sets of cardinality m. Suppose $\mathbf{X}^*$ has condition number $\kappa = \sigma_1(\mathbf{X}^*)/\sigma_r(\mathbf{X}^*)$. Then there are absolute constants c and c_0 such that with probability at least $1 - c_0 n_{(1)}^{-10}$, the output of the convex problem*

$$\widetilde{\mathbf{X}} = \operatorname*{argmin}_{\mathbf{X}} \|\mathbf{X}\|_{r*}, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{X}^*), \tag{3.16}$$

is unique and exact, i.e., $\widetilde{\mathbf{X}} = \mathbf{X}^*$, provided that $m \geq c\kappa^2 \mu r n_{(1)} \log_{2\kappa}(n_{(1)}) \log(n_{(1)})$ and $\mathbf{X}^*$ obeys μ-incoherence (3.5). Namely, strong duality holds for problem (3.12).

Proof. We have shown in Theorem 21 that the problem

$$(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) = \operatorname*{argmin}_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{AB}\|_F^2, \quad \text{s.t.} \quad \mathcal{P}_\Omega(\mathbf{AB}) = \mathcal{P}_\Omega(\mathbf{X}^*),$$

exactly recovers $\mathbf{X}^*$, i.e., $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$, with small sample complexity. So if strong duality holds, this non-convex optimization problem can be equivalently converted to the convex program (3.16). Then Theorem 22 is straightforward from strong duality.

It now suffices to apply our unified framework in the beginning of this subsection to prove the strong duality. We show that the dual condition in Theorem 20 holds with high probability by the following arguments. Let $(\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ be a global solution to problem (3.16). For $H(\mathbf{X}) = \mathbf{I}_{\{\mathbf{M}\in\mathbb{R}^{n_1\times n_2}:\, \mathcal{P}_\Omega\mathbf{M}=\mathcal{P}_\Omega\mathbf{X}^*\}}(\mathbf{X})$, we have

$$\begin{aligned}\Psi = \partial H(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) &= \{\mathbf{G} \in \mathbb{R}^{n_1\times n_2} : \langle \mathbf{G}, \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\rangle \geq \langle \mathbf{G}, \mathbf{Y}\rangle, \text{ for any } \mathbf{Y} \in \mathbb{R}^{n_1\times n_2} \text{ s.t. } \mathcal{P}_\Omega\mathbf{Y} = \mathcal{P}_\Omega\mathbf{X}^*\} \\ &= \{\mathbf{G} \in \mathbb{R}^{n_1\times n_2} : \langle \mathbf{G}, \mathbf{X}^*\rangle \geq \langle \mathbf{G}, \mathbf{Y}\rangle, \text{ for any } \mathbf{Y} \in \mathbb{R}^{n_1\times n_2} \text{ s.t. } \mathcal{P}_\Omega\mathbf{Y} = \mathcal{P}_\Omega\mathbf{X}^*\} = \Omega,\end{aligned}$$

where the third equality holds since $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$. Then we only need to show

$$(1)\ \widetilde{\mathbf{\Lambda}} \in \Omega, \qquad (2)\ \mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}, \qquad (3)\ \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \frac{2}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}). \tag{3.17}$$

It is interesting to see that dual condition (3.17) can be satisfied if the angle θ between subspace Ω and subspace $\mathcal{T}$ is very small; see Figure 3.5. When the sample size $|\Omega|$ becomes larger and larger, the angle θ becomes smaller and smaller (e.g., when $|\Omega| = n_1 n_2$, the angle θ is zero as $\Omega = \mathbb{R}^{n_1\times n_2}$). We show that the sample size $m = \Omega(\kappa^2 \mu r n_{(1)} \log_{2\kappa}(n_{(1)}) \log(n_{(1)}))$ is a sufficient condition for condition (3.17) to hold.

Let $\widetilde{\mathbf{A}} \in \mathbb{R}^{n_1\times r}$ and $\widetilde{\mathbf{B}} \in \mathbb{R}^{r\times n_2}$ such that $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$. Then we have the following lemma.

Lemma 35. *Let $\Omega \sim \mathsf{Uniform}(m)$ be the support set uniformly distributed among all sets of cardinality m. Suppose that $m \geq c\kappa^2 \mu n_{(1)} r \log n_{(1)} \log_{2\kappa} n_{(1)}$ for an absolute constant c and $\mathbf{X}^*$ obeys μ-incoherence (3.5). Then there exists $\widetilde{\mathbf{\Lambda}}$ such that*

$$\begin{aligned} &(1) \quad \widetilde{\mathbf{\Lambda}} \in \Omega, \\ &(2) \quad \mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}, \\ &(3) \quad \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| < \frac{2}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}). \end{aligned} \tag{3.18}$$

with probability at least $1 - n_{(1)}^{-10}$.

The rest of the section is denoted to the proof of Lemma 35. We begin with the following lemma.

Lemma 36. *If we can construct an* $\mathbf{\Lambda}$ *such that*

$$\begin{aligned} &\text{(a)} \quad \mathbf{\Lambda} \in \Omega, \\ &\text{(b)} \quad \|\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_F \leq \sqrt{\frac{r}{3n_{(1)}^2}}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}), \\ &\text{(c)} \quad \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}\| < \frac{1}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}), \end{aligned} \tag{3.19}$$

then we can construct an $\widetilde{\mathbf{\Lambda}}$ *such that Eqn. (3.18) holds with probability at least* $1 - n_{(1)}^{-10}$.

Proof. To prove the lemma, we first claim the following theorem.

Theorem 23 ([55], Theorem 4.1)**.** *Assume that* Ω *is sampled according to the Bernoulli model with success probability* $p = \Theta(\frac{m}{n_1 n_2})$*, and incoherence condition (3.5) holds. Then there is an absolute constant* C_R *such that for* $\beta > 1$*, we have*

$$\|p^{-1}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}} - \mathcal{P}_{\mathcal{T}}\| \leq C_R\sqrt{\frac{\beta\mu n_{(1)} r \log n_{(1)}}{m}} \triangleq \epsilon,$$

with probability at least $1 - 3n^{-\beta}$ *provided that* $C_R\sqrt{\frac{\beta\mu n_{(1)} r \log n_{(1)}}{m}} < 1$.

Suppose that Condition (3.19) holds. Let $\mathbf{Y} = \widetilde{\mathbf{\Lambda}} - \mathbf{\Lambda} \in \Omega$ be the perturbation matrix between $\mathbf{\Lambda}$ and $\widetilde{\mathbf{\Lambda}}$ such that $\mathcal{P}_{\mathcal{T}}(-\widetilde{\mathbf{\Lambda}}) = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$. Such a $\mathbf{Y}$ exists by setting $\mathbf{Y} = \mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}(\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}})^{-1}(\mathcal{P}_{\mathcal{T}}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$. So $\|\mathcal{P}_{\mathcal{T}}\mathbf{Y}\|_F \leq \sqrt{\frac{r}{3n_{(1)}^2}}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$. We now prove Condition (3) in Eqn. (3.18). Observe that

$$\begin{aligned} \|\mathcal{P}_{\mathcal{T}^\perp}\widetilde{\mathbf{\Lambda}}\| &\leq \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}\| + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\| \\ &\leq \frac{1}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\|. \end{aligned} \tag{3.20}$$

So we only need to show $\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\| \leq \frac{1}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$.

Before proceeding, we begin by introducing a normalized version $\mathcal{Q}_{\Omega} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ of $\mathcal{P}_{\Omega}$:

$$\mathcal{Q}_{\Omega} = p^{-1}\mathcal{P}_{\Omega} - \mathcal{I}.$$

With this, we have

$$\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}} = p\mathcal{P}_{\mathcal{T}}(\mathcal{I} + \mathcal{Q}_{\Omega})\mathcal{P}_{\mathcal{T}}.$$

Note that for any operator $\mathcal{P} : \mathcal{T} \to \mathcal{T}$, we have

$$\mathcal{P}^{-1} = \sum_{k \geq 0}(\mathcal{P}_{\mathcal{T}} - \mathcal{P})^k \text{ whenever } \|\mathcal{P}_{\mathcal{T}} - \mathcal{P}\| < 1.$$

So according to Theorem 23, the operator $p(\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega\mathcal{P}_\mathcal{T})^{-1}$ can be represented as a *convergent* Neumann series

$$p(\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega\mathcal{P}_\mathcal{T})^{-1} = \sum_{k\geq 0}(-1)^k(\mathcal{P}_\mathcal{T}\mathcal{Q}_\Omega\mathcal{P}_\mathcal{T})^k,$$

because $\|\mathcal{P}_\mathcal{T}\mathcal{Q}_\Omega\mathcal{P}_\mathcal{T}\| \leq \epsilon < \frac{1}{2}$ once $m \geq C\mu n_{(1)} r \log n_{(1)}$ for a sufficiently large absolute constant C. We also note that

$$p(\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_\Omega\mathcal{P}_\mathcal{T}) = \mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_\Omega\mathcal{P}_\mathcal{T},$$

because $\mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_\mathcal{T} = 0$. Thus

$$\begin{aligned}
\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}\| &= \|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}(\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega\mathcal{P}_\mathcal{T})^{-1}(\mathcal{P}_\mathcal{T}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}))\| \\
&= \|\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_\Omega\mathcal{P}_\mathcal{T} p(\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega\mathcal{P}_\mathcal{T})^{-1}((\mathcal{P}_\mathcal{T}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}))\| \\
&= \|\sum_{k\geq 0}(-1)^k\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_\Omega(\mathcal{P}_\mathcal{T}\mathcal{Q}_\Omega\mathcal{P}_\mathcal{T})^k((\mathcal{P}_\mathcal{T}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}))\| \\
&\leq \sum_{k\geq 0}\|(-1)^k\mathcal{P}_{\mathcal{T}^\perp}\mathcal{Q}_\Omega(\mathcal{P}_\mathcal{T}\mathcal{Q}_\Omega\mathcal{P}_\mathcal{T})^k((\mathcal{P}_\mathcal{T}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}))\|_F \\
&\leq \|\mathcal{Q}_\Omega\|\sum_{k\geq 0}\|\mathcal{P}_\mathcal{T}\mathcal{Q}_\Omega\mathcal{P}_\mathcal{T}\|^k\|\mathcal{P}_\mathcal{T}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}))\|_F \\
&\leq \frac{4}{p}\|\mathcal{P}_\mathcal{T}(-\mathbf{\Lambda}) - \widetilde{\mathbf{A}}\widetilde{\mathbf{B}})\|_F \\
&\leq \Theta\left(\frac{n_1 n_2}{m}\right)\sqrt{\frac{r}{3n_{(1)}^2}}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) \\
&\leq \frac{1}{3}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})
\end{aligned}$$

with high probability. The proof is completed. □

It thus suffices to construct a dual certificate $\mathbf{\Lambda}$ such that all conditions in (3.19) hold. To this end, partition $\Omega = \Omega_1 \cup \Omega_2 \cup ... \cup \Omega_b$ into b partitions of size q. By assumption, we may choose

$$q \geq \frac{128}{3}C\beta\kappa^2\mu r n_{(1)}\log n_{(1)} \quad \text{and} \quad b \geq \frac{1}{2}\log_{2\kappa}\left(24^2 n_{(1)}^2\kappa^2\right)$$

for a sufficiently large constant C. Let $\Omega_j \sim \mathsf{Ber}(q)$ denote the set of indices corresponding to the j-th partitions. Define $\mathbf{W}_0 = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$ and set $\mathbf{\Lambda}_k = \frac{n_1 n_2}{q}\sum_{j=1}^k \mathcal{P}_{\Omega_j}(\mathbf{W}_{j-1})$, $\mathbf{W}_k = \widetilde{\mathbf{A}}\widetilde{\mathbf{B}} - \mathcal{P}_\mathcal{T}(\mathbf{\Lambda}_k)$ for $k = 1, 2, ..., b$. Then by Theorem 23,

$$\begin{aligned}
\|\mathbf{W}_k\|_F &= \left\|\mathbf{W}_{k-1} - \frac{n_1 n_2}{q}\mathcal{P}_\mathcal{T}\mathcal{P}_{\Omega_k}(\mathbf{W}_{k-1})\right\|_F = \left\|\left(\mathcal{P}_\mathcal{T} - \frac{n_1 n_2}{q}\mathcal{P}_\mathcal{T}\mathcal{P}_{\Omega_k}\mathcal{P}_\mathcal{T}\right)(\mathbf{W}_{k-1})\right\|_F \\
&\leq \frac{1}{2\kappa}\|\mathbf{W}_{k-1}\|_F.
\end{aligned}$$

So it follows that $\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} - \mathcal{P}_\mathcal{T}(\mathbf{\Lambda}_b)\|_F = \|\mathbf{W}_b\|_F \leq (2\kappa)^{-b}\|\mathbf{W}_0\|_F \leq (2\kappa)^{-b}\sqrt{r}\sigma_1(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}) \leq \sqrt{\frac{r}{24^2 n_{(1)}^2}}\sigma_r(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})$.

The following lemma together implies the strong duality of (3.16) straightforwardly.

Lemma 37. *Under the assumptions of Theorem 22, the dual certification $\mathbf{\Lambda}_b$ obeys the dual condition (3.19) with probability at least $1 - n_{(1)}^{-10}$.*

Proof. It is well known that for matrix completion, the Uniform model $\Omega \sim \mathsf{Uniform}(m)$ is equivalent to the Bernoulli model $\Omega \sim \mathsf{Ber}(p)$, where each element in $[n_1] \times [n_2]$ is included with probability $p = \Theta(m/(n_1 n_2))$ independently. By the equivalence, we can suppose $\Omega \sim \mathsf{Ber}(p)$.

To prove Lemma 37, as a preliminary, we need the following lemmas.

Lemma 38 ([63], Lemma 2)**.** *Suppose $\mathbf{Z}$ is a fixed matrix. Suppose $\Omega \sim \mathsf{Ber}(p)$. Then with high probability,*

$$\|(\mathcal{I} - p^{-1}\mathcal{P}_\Omega)\mathbf{Z}\| \le C_0' \left(\frac{\log n_{(1)}}{p} \|\mathbf{Z}\|_\infty + \sqrt{\frac{\log n_{(1)}}{p}} \|\mathbf{Z}\|_{\infty,2} \right),$$

where $C_0' > 0$ is an absolute constant and

$$\|\mathbf{Z}\|_{\infty,2} = \max \left\{ \max_i \sqrt{\sum_b \mathbf{Z}_{ib}^2}, \max_j \sqrt{\sum_a \mathbf{Z}_{aj}^2} \right\}.$$

Lemma 39 ([57], Lemma 3.1)**.** *Suppose $\Omega \sim \mathsf{Ber}(p)$ and $\mathbf{Z}$ is a fixed matrix. Then with high probability,*

$$\|\mathbf{Z} - p^{-1}\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega\mathbf{Z}\|_\infty \le \epsilon \|\mathbf{Z}\|_\infty,$$

provided that $p \ge C_0 \epsilon^{-2} (\mu r \log n_{(1)})/n_{(2)}$ for some absolute constant $C_0 > 0$.

Lemma 40 ([63], Lemma 3)**.** *Suppose that $\mathbf{Z}$ is a fixed matrix and $\Omega \sim \mathsf{Ber}(p)$. If $p \ge c_0 \mu r \log n_{(1)}/n_{(2)}$ for some c_0 sufficiently large, then with high probability,*

$$\|(p^{-1}\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega - \mathcal{P}_\mathcal{T})\mathbf{Z}\|_{\infty,2} \le \frac{1}{2}\sqrt{\frac{n_{(1)}}{\mu r}} \|\mathbf{Z}\|_\infty + \frac{1}{2}\|\mathbf{Z}\|_{\infty,2}.$$

Observe that by Lemma 39,

$$\|\mathbf{W}_j\|_\infty \le \left(\frac{1}{2}\right)^j \|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty,$$

and by Lemma 40,

$$\|\mathbf{W}_j\|_{\infty,2} \le \frac{1}{2}\sqrt{\frac{n_{(1)}}{\mu r}} \|\mathbf{W}_{j-1}\|_\infty + \frac{1}{2}\|\mathbf{W}_{j-1}\|_{\infty,2}.$$

So

$$\begin{aligned} &\|\mathbf{W}_j\|_{\infty,2} \\ &\le \left(\frac{1}{2}\right)^j \sqrt{\frac{n_{(1)}}{\mu r}} \|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + \frac{1}{2}\|\mathbf{W}_{j-1}\|_{\infty,2} \\ &\le j \left(\frac{1}{2}\right)^j \sqrt{\frac{n_{(1)}}{\mu r}} \|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + \left(\frac{1}{2}\right)^j \|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_{\infty,2}. \end{aligned}$$

Therefore,

$$\begin{aligned}
&\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}_b\| \\
&\leq \sum_{j=1}^{b} \|\frac{n_1 n_2}{q}\mathcal{P}_{\mathcal{T}^\perp}\mathcal{P}_{\Omega_j}\mathbf{W}_{j-1}\| \\
&= \sum_{j=1}^{b} \|\mathcal{P}_{\mathcal{T}^\perp}(\frac{n_1 n_2}{q}\mathcal{P}_{\Omega_j}\mathbf{W}_{j-1} - \mathbf{W}_{j-1})\| \\
&\leq \sum_{j=1}^{b} \|(\frac{n_1 n_2}{q}\mathcal{P}_{\Omega_j} - \mathcal{I})(\mathbf{W}_{j-1})\|.
\end{aligned}$$

Let p denote $\Theta\left(\frac{q}{n_1 n_2}\right)$. By Lemma 38,

$$\begin{aligned}
&\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}_b\| \\
&\leq C_0'\frac{\log n_{(1)}}{p}\sum_{j=1}^{b}\|\mathbf{W}_{j-1}\|_\infty + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sum_{j=1}^{b}\|\mathbf{W}_{j-1}\|_{\infty,2} \\
&\leq C_0'\frac{\log n_{(1)}}{p}\sum_{j=1}^{b}\left(\frac{1}{2}\right)^j\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sum_{j=1}^{b}\left[j\left(\frac{1}{2}\right)^j\sqrt{\frac{n_{(1)}}{\mu r}}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + \left(\frac{1}{2}\right)^j\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_{\infty,2}\right] \\
&\leq C_0'\frac{\log n_{(1)}}{p}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + 2C_0'\sqrt{\frac{\log n_{(1)}}{p}}\sqrt{\frac{n_{(1)}}{\mu r}}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_\infty + C_0'\sqrt{\frac{\log n_{(1)}}{p}}\|\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}\|_{\infty,2}.
\end{aligned}$$

Setting $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \mathbf{X}^*$, we note the facts that (we assume WLOG $n_2 \geq n_1$)

$$\|\mathbf{X}^*\|_{\infty,2} = \max_i \|\mathbf{e}_i^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\|_2 \leq \max_i \|\mathbf{e}_i^T\mathbf{U}\|\sigma_1(\mathbf{X}^*) \leq \sqrt{\frac{\mu r}{n_1}}\sigma_1(\mathbf{X}^*) \leq \sqrt{\frac{\mu r}{n_1}}\kappa\sigma_r(\mathbf{X}^*),$$

and that

$$\begin{aligned}
\|\mathbf{X}^*\|_\infty &= \max_{ij}\langle\mathbf{X}^*, \mathbf{e}_i\mathbf{e}_j^T\rangle = \max_{ij}\langle\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \mathbf{e}_i\mathbf{e}_j^T\rangle = \max_{ij}\langle\mathbf{e}_i^T\mathbf{U}\mathbf{\Sigma}, \mathbf{e}_j^T\mathbf{V}\rangle \\
&\leq \max_{ij}\|\mathbf{e}_i^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\|_2\|\mathbf{e}_j^T\mathbf{V}\|_2 \leq \max_j\|\mathbf{X}^*\|_{\infty,2}\|\mathbf{e}_j^T\mathbf{V}\|_2 \leq \frac{\mu r\kappa}{\sqrt{n_1 n_2}}\sigma_r(\mathbf{X}^*).
\end{aligned}$$

Substituting $p = \Theta\left(\frac{\kappa^2\mu r n_{(1)}\log(n_{(1)})\log_{2\kappa}(n_{(1)})}{n_1 n_2}\right)$, we obtain $\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{\Lambda}_b\| < \frac{1}{3}\sigma_r(\mathbf{X}^*)$. The proof is completed. □

□

This positive result matches a lower bound from prior work up to a logarithmic factor, which shows that the sample complexity in Theorem 21 is nearly optimal.

Theorem 24 (Information-theoretic lower bound. [56], Theorem 1.7)**.** *Denote by* $\Omega \sim \mathsf{Uniform}(m)$ *the support set uniformly distributed among all sets of cardinality* m*. Suppose that* $m \leq c\mu n_{(1)} r\log n_{(1)}$ *for an absolute constant* c*. Then there exist infinitely many* $n_1 \times n_2$ *matrices* $\mathbf{X}'$ *of rank at most* r *obeying* μ*-incoherence* (3.5) *such that* $\mathcal{P}_\Omega(\mathbf{X}') = \mathcal{P}_\Omega(\mathbf{X}^*)$*, with probability at least* $1 - n_{(1)}^{-10}$*.*

Proofs of Robust PCA

Theorem 19 (Robust PCA. Restated). *Suppose $\mathbf{X}^*$ is an $n_1 \times n_2$ matrix of rank r, and obeys incoherence (3.5) and (3.6). Assume that the support set Ω of $\mathbf{S}^*$ is uniformly distributed among all sets of cardinality m. Then with probability at least $1 - cn_{(1)}^{-10}$, the output of the optimization problem*

$$(\widetilde{\mathbf{X}}, \widetilde{\mathbf{S}}) = \underset{\mathbf{X},\mathbf{S}}{\operatorname{argmin}} \|\mathbf{X}\|_{r*} + \lambda\|\mathbf{S}\|_1, \quad \text{s.t.} \quad \mathbf{D} = \mathbf{X} + \mathbf{S}, \tag{3.21}$$

with $\lambda = \frac{\sigma_r(\mathbf{X}^*)}{\sqrt{n_{(1)}}}$ is exact, namely, $\widetilde{\mathbf{X}} = \mathbf{X}^*$ and $\widetilde{\mathbf{S}} = \mathbf{S}^*$, provided that

$$\mathsf{rank}(\mathbf{X}^*) \le \rho_r \frac{n_{(2)}}{\mu \log^2 n_{(1)}} \text{ and } m \le \rho_s n_1 n_2$$

, where c, ρ_r, and ρ_s are all positive absolute constants, and function $\|\cdot\|_{r*}$ is given by $\|\mathbf{X}\|_{r*} := \max_{\mathbf{M}} \langle \mathbf{X}, \mathbf{M}\rangle - \frac{1}{2}\|\mathbf{M}\|_r^2$ and $\|\mathbf{M}\|_r^2 = \sum_{i=1}^r \sigma_i^2(\mathbf{M})$.

Dual certificates We first show the dual certificates.

Lemma 41. *Assume that $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \le 1/2$ and $\lambda < \sigma_r(\mathbf{X}^*)$. Then $(\mathbf{X}^*, \mathbf{S}^*)$ is the unique solution to problem (19) if there exists $(\mathbf{W}, \mathbf{F}, \mathbf{K})$ for which*

$$\mathbf{X}^* + \mathbf{W} = \lambda(\operatorname{sign}(\mathbf{S}^*) + \mathbf{F} + \mathcal{P}_\Omega \mathbf{K}),$$

where $\mathbf{W} \in \mathcal{T}^\perp$, $\|\mathbf{W}\| \le \frac{\sigma_r(\mathbf{X}^*)}{2}$, $\mathbf{F} \in \Omega^\perp$, $\|\mathbf{F}\|_\infty \le \frac{1}{2}$, and $\|\mathcal{P}_\Omega \mathbf{K}\|_F \le \frac{1}{4}$.

Proof. Let $(\mathbf{X}^* + \mathbf{H}, \mathbf{S}^* - \mathbf{H})$ be any optimal solution to problem (3.21). Denote by $\mathbf{X}^* + \mathbf{W}^*$ an arbitrary subgradient of the $r*$ function at $\mathbf{X}^*$ (see Lemma 33), and $\operatorname{sign}(\mathbf{S}^*) + \mathbf{F}^*$ an arbitrary subgradient of the ℓ_1 norm at $\mathbf{S}^*$. By the definition of the subgradient, the inequality follows

$$\begin{aligned}
\|\mathbf{X}^* + \mathbf{H}\|_{r*} + \lambda\|\mathbf{S}^* - \mathbf{H}\|_1 &\ge \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \langle \mathbf{X}^* + \mathbf{W}^*, \mathbf{H}\rangle - \lambda\langle \operatorname{sign}(\mathbf{S}^*) + \mathbf{F}^*, \mathbf{H}\rangle \\
&= \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \langle \mathbf{X}^* - \lambda \operatorname{sign}(\mathbf{S}^*), \mathbf{H}\rangle + \langle \mathbf{W}^*, \mathbf{H}\rangle - \lambda\langle \mathbf{F}^*, \mathbf{H}\rangle \\
&= \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \langle \mathbf{X}^* - \lambda \operatorname{sign}(\mathbf{S}^*), \mathbf{H}\rangle + \sigma_r(\mathbf{X}^*)\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1 \\
&= \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \langle \lambda\mathbf{F} + \lambda\mathcal{P}_\Omega\mathbf{K} - \mathbf{W}, \mathbf{H}\rangle + \sigma_r(\mathbf{X}^*)\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1 \\
&\ge \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \frac{\sigma_r(\mathbf{X}^*)}{2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_* + \frac{\lambda}{2}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1 - \frac{\lambda}{4}\|\mathcal{P}_\Omega\mathbf{H}\|_F,
\end{aligned}$$

where the third line holds by picking $\mathbf{W}^*$ such that $\langle \mathbf{W}^*, \mathbf{H}\rangle = \sigma_r(\mathbf{X}^*)\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_*$ and $\langle \mathbf{F}^*, \mathbf{H}\rangle = -\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1$.[4] We note that

$$\begin{aligned}
\|\mathcal{P}_\Omega\mathbf{H}\|_F &\le \|\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\mathbf{H}\|_F + \|\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_F \\
&\le \frac{1}{2}\|\mathbf{H}\|_F + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_F \\
&\le \frac{1}{2}\|\mathcal{P}_\Omega\mathbf{H}\|_F + \frac{1}{2}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_F + \|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_F,
\end{aligned}$$

[4]For instance, $\mathbf{F}^* = -\operatorname{sign}(\mathcal{P}_{\Omega^\perp}\mathbf{H})$ is such as matrix. Also, by the duality between the nuclear norm and the operator norm, there is a matrix obeying $\|\mathbf{W}\| = \sigma_r(\mathbf{X}^*)$ such that $\langle \mathbf{W}, \mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\rangle = \sigma_r(\mathbf{X}^*)\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_*$. We pick $\mathbf{W}^* = \mathcal{P}_{\mathcal{T}^\perp}\mathbf{W}$ here.

which implies that $\frac{\lambda}{4}\|\mathcal{P}_\Omega \mathbf{H}\|_F \le \frac{\lambda}{4}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_F + \frac{\lambda}{2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_F \le \frac{\lambda}{4}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1 + \frac{\lambda}{2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_*$. Therefore,

$$\begin{aligned}\|\mathbf{X}^* + \mathbf{H}\|_{r*} + \lambda\|\mathbf{S}^* - \mathbf{H}\|_1 &\ge \|\mathbf{X}^*\|_{r*} + \lambda\|\mathbf{S}^*\|_1 + \frac{\sigma_r(\mathbf{X}^*) - \lambda}{2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_* + \frac{\lambda}{4}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1 \\ &\ge \|\mathbf{X}^* + \mathbf{H}\|_{r*} + \lambda\|\mathbf{S}^* - \mathbf{H}\|_1 + \frac{\sigma_r(\mathbf{X}^*) - \lambda}{2}\|\mathcal{P}_{\mathcal{T}^\perp}\mathbf{H}\|_* + \frac{\lambda}{4}\|\mathcal{P}_{\Omega^\perp}\mathbf{H}\|_1,\end{aligned}$$

where the second inequality holds because $(\mathbf{X}^* + \mathbf{H}, \mathbf{S}^* - \mathbf{H})$ is optimal. Thus $\mathbf{H} \in \mathcal{T} \cap \Omega$. Note that $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| < 1$ implies $\mathcal{T} \cap \Omega = \{0\}$ and thus $\mathbf{H} = 0$. This completes the proof. □

According to Lemma 41, to show the exact recoverability of problem (3.21), it is sufficient to find an appropriate $\mathbf{W}$ for which

$$\begin{cases} \mathbf{W} \in \mathcal{T}^\perp, \\ \|\mathbf{W}\| \le \frac{\sigma_r(\mathbf{X}^*)}{2}, \\ \|\mathcal{P}_\Omega(\mathbf{X}^* + \mathbf{W} - \lambda \text{sign}(S^*))\|_F \le \frac{\lambda}{4}, \\ \|\mathcal{P}_{\Omega^\perp}(\mathbf{X}^* + \mathbf{W})\|_\infty \le \frac{\lambda}{2}, \end{cases} \tag{3.22}$$

under the assumptions that $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \le 1/2$ and $\lambda < \sigma_r(\mathbf{X}^*)$. We note that $\lambda = \frac{\sigma_r(\mathbf{X}^*)}{\sqrt{n_{(1)}}} < \sigma_r(\mathbf{X}^*)$. To see $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \le 1/2$, we have the following lemma.

Lemma 42 ([57], Cor 2.7)**.** *Suppose that $\Omega \sim \mathsf{Ber}(p)$ and incoherence (3.5) holds. Then with probability at least $1 - n_{(1)}^{-10}$, $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\|^2 \le p + \epsilon$, provided that $1 - p \ge C_0 \epsilon^{-2} \mu r \log n_{(1)}/n_{(2)}$ for an absolute constant C_0.*

Setting p and ϵ as small constants in Lemma 42, we have $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \le 1/2$ with high probability.

Dual certification by least squares and the golfing scheme The remainder of the proof is to construct $\mathbf{W}$ such that the dual condition (3.22) holds true. Before introducing our construction, we assume $\Omega \sim \mathsf{Ber}(p)$, or equivalently $\Omega^\perp \sim \mathsf{Ber}(1-p)$, where p is allowed be as large as an absolute constant. Note that $\Omega^\perp$ has the same distribution as that of $\Omega_1 \cup \Omega_2 \cup ... \cup \Omega_{j_0}$, where the Ω_j's are drawn independently with replacement from $\mathsf{Ber}(q)$, $j_0 = \lceil \log n_{(1)} \rceil$, and q obeys $p = (1-q)^{j_0}$ ($q = \Omega(1/\log n_{(1)})$ implies $p = \mathcal{O}(1)$). We construct $\mathbf{W}$ based on such a distribution.

Our construction separates $\mathbf{W}$ into two terms: $\mathbf{W} = \mathbf{W}^L + \mathbf{W}^S$. To construct $\mathbf{W}^L$, we apply the golfing scheme introduced by [98, 194]. Specifically, $\mathbf{W}^L$ is constructed by an inductive procedure:

$$\begin{aligned} \mathbf{Y}_j = \mathbf{Y}_{j-1} + q^{-1}\mathcal{P}_{\Omega_j}\mathcal{P}_\mathcal{T}(\mathbf{X}^* - \mathbf{Y}_{j-1}),\ \mathbf{Y}_0 = \mathbf{0}, \\ \mathbf{W}^L = \mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}_{j_0}. \end{aligned} \tag{3.23}$$

To construct $\mathbf{W}^S$, we apply the method of least squares by [57], which is

$$\mathbf{W}^S = \lambda \mathcal{P}_{\mathcal{T}^\perp} \sum_{k \ge 0} (\mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)^k \text{sign}(\mathbf{S}^*). \tag{3.24}$$

Note that $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \leq 1/2$. Thus $\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega\| \leq 1/4$ and the Neumann series in (3.24) is well-defined. Observe that $\mathcal{P}_\Omega \mathbf{W}^S = \lambda(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)(\mathcal{P}_\Omega - \mathcal{P}_\Omega \mathcal{P}_\mathcal{T} \mathcal{P}_\Omega)^{-1} \mathrm{sign}(\mathbf{S}^*) = \lambda \mathrm{sign}(\mathbf{S}^*)$. So to prove the dual condition (3.22), it suffices to show that

$$\begin{aligned} &\text{(a)} \quad \|\mathbf{W}^L\| \leq \frac{\sigma_r(\mathbf{X}^*)}{4}, \\ &\text{(b)} \quad \|\mathcal{P}_\Omega(\mathbf{X}^* + \mathbf{W}^L)\|_F \leq \frac{\lambda}{4}, \\ &\text{(c)} \quad \|\mathcal{P}_{\Omega^\perp}(\mathbf{X}^* + \mathbf{W}^L)\|_\infty \leq \frac{\lambda}{4}, \end{aligned} \tag{3.25}$$

$$\begin{aligned} &\text{(d)} \quad \|\mathbf{W}^S\| \leq \frac{\sigma_r(\mathbf{X}^*)}{4}, \\ &\text{(e)} \quad \|\mathcal{P}_{\Omega^\perp} \mathbf{W}^S\|_\infty \leq \frac{\lambda}{4}. \end{aligned} \tag{3.26}$$

Proof of dual conditions Since we have constructed the dual certificate $\mathbf{W}$, the remainder is to show that $\mathbf{W}$ obeys dual conditions (3.25) and (3.26) with high probability. We have the following.

Lemma 43. *Assume $\Omega_j \sim \mathsf{Ber}(q)$, $j = 1, 2, ..., j_0$, and $j_0 = 2\lceil \log n_{(1)} \rceil$. Then under the other assumptions of Theorem 19, $\mathbf{W}^L$ given by (3.23) obeys dual condition (3.25).*

Proof. Let $\mathbf{Z}_j = \mathcal{P}_\mathcal{T}(\mathbf{X}^* - \mathbf{Y}_j) \in \mathcal{T}$. Then we have

$$\mathbf{Z}_j = \mathcal{P}_\mathcal{T} \mathbf{Z}_{j-1} - q^{-1} \mathcal{P}_\mathcal{T} \mathcal{P}_{\Omega_j} \mathcal{P}_\mathcal{T} \mathbf{Z}_{j-1} = (\mathcal{P}_\mathcal{T} - q^{-1} \mathcal{P}_\mathcal{T} \mathcal{P}_{\Omega_j} \mathcal{P}_\mathcal{T}) \mathbf{Z}_{j-1},$$

and $\mathbf{Y}_j = \sum_{k=1}^{j} q^{-1} \mathcal{P}_{\Omega_k} \mathbf{Z}_{k-1} \in \Omega^\perp$. We set $q = \Omega(\epsilon^{-2} \mu r \log n_{(1)} / n_{(2)})$ with a small constant ϵ.

Proof of (a). It holds that

$$\begin{aligned} \|\mathbf{W}^L\| = \|\mathcal{P}_{\mathcal{T}^\perp} \mathbf{Y}_{j_0}\| &\leq \sum_{k=1}^{j_0} \|q^{-1} \mathcal{P}_{\mathcal{T}^\perp} \mathcal{P}_{\Omega_k} \mathbf{Z}_{k-1}\| \\ &= \sum_{k=1}^{j_0} \|\mathcal{P}_{\mathcal{T}^\perp}(q^{-1} \mathcal{P}_{\Omega_k} \mathbf{Z}_{k-1} - \mathbf{Z}_{k-1})\| \\ &\leq \sum_{k=1}^{j_0} \|q^{-1} \mathcal{P}_{\Omega_k} \mathbf{Z}_{k-1} - \mathbf{Z}_{k-1}\| \\ &\leq C_0' \left(\frac{\log n_{(1)}}{q} \sum_{k=1}^{j_0} \|\mathbf{Z}_{k-1}\|_\infty + \sqrt{\frac{\log n_{(1)}}{q}} \sum_{k=1}^{j_0} \|\mathbf{Z}_{k-1}\|_{\infty,2} \right). \quad \text{(by Lemma 38)} \end{aligned}$$

We note that by Lemma 39,

$$\|\mathbf{Z}_{k-1}\|_\infty \leq \left(\frac{1}{2}\right)^{k-1} \|\mathbf{Z}_0\|_\infty,$$

and by Lemma 40,

$$\|\mathbf{Z}_{k-1}\|_{\infty,2} \le \frac{1}{2}\sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{Z}_{k-2}\|_\infty + \frac{1}{2}\|\mathbf{Z}_{k-2}\|_{\infty,2}.$$

Therefore,

$$\begin{aligned}\|\mathbf{Z}_{k-1}\|_{\infty,2} &\le \left(\frac{1}{2}\right)^{k-1}\sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{Z}_0\|_\infty + \frac{1}{2}\|\mathbf{Z}_{k-2}\|_{\infty,2} \\ &\le (k-1)\left(\frac{1}{2}\right)^{k-1}\sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{Z}_0\|_\infty + \left(\frac{1}{2}\right)^{k-1}\|\mathbf{Z}_0\|_{\infty,2},\end{aligned}$$

and so we have

$$\begin{aligned}&\|\mathbf{W}^L\| \\ &\le C_0'\left[\frac{\log n_{(1)}}{q}\sum_{k=1}^{j_0}\left(\frac{1}{2}\right)^{k-1}\|\mathbf{Z}_0\|_\infty + \sqrt{\frac{\log n_{(1)}}{q}}\sum_{k=1}^{j_0}\left((k-1)\left(\frac{1}{2}\right)^{k-1}\sqrt{\frac{n_{(1)}}{\mu r}}\|\mathbf{Z}_0\|_\infty + \left(\frac{1}{2}\right)^{k-1}\|\mathbf{Z}_0\|_{\infty,2}\right)\right] \\ &\le 2C_0'\left[\frac{\log n_{(1)}}{q}\|\mathbf{X}^*\|_\infty + \sqrt{\frac{n_{(1)}\log n_{(1)}}{q\mu r}}\|\mathbf{X}^*\|_\infty + \sqrt{\frac{\log n_{(1)}}{q}}\|\mathbf{X}^*\|_{\infty,2}\right] \\ &\le \frac{1}{16}\left[\frac{n_{(2)}}{\mu r}\|\mathbf{X}^*\|_\infty + \frac{\sqrt{n_{(1)}n_{(2)}}}{\mu r}\|\mathbf{X}^*\|_\infty + \sqrt{\frac{n_{(2)}}{\mu r}}\|\mathbf{X}^*\|_{\infty,2}\right] \quad \text{(since } q = \Omega(\mu r \log n_{(1)})/n_{(2)}) \\ &\le \frac{\sigma_r(\mathbf{X}^*)}{4}, \quad \text{(by incoherence (3.6))}\end{aligned}$$

where we have used the fact that

$$\|\mathbf{X}^*\|_{\infty,2} \le \sqrt{n_{(1)}}\|\mathbf{X}^*\|_\infty \le \sqrt{\frac{\mu r}{n_{(2)}}}\sigma_r(\mathbf{X}^*).$$

Proof of (b). Because $\mathbf{Y}_{j_0} \in \Omega^\perp$, we have $\mathcal{P}_\Omega(\mathbf{X}^* + \mathcal{P}_{\mathcal{T}^\perp}\mathbf{Y}_{j_0}) = \mathcal{P}_\Omega(\mathbf{X}^* - \mathcal{P}_{\mathcal{T}}\mathbf{Y}_{j_0}) = \mathcal{P}_\Omega\mathbf{Z}_{j_0}$. It then follows from Theorem 23 that for a properly chosen t,

$$\begin{aligned}\|\mathbf{Z}_{j_0}\|_F &\le t^{j_0}\|\mathbf{X}^*\|_F \\ &\le t^{j_0}\sqrt{n_1 n_2}\|\mathbf{X}^*\|_\infty \\ &\le t^{j_0}\sqrt{n_1 n_2}\sqrt{\frac{\mu r}{n_1 n_2}}\sigma_r(\mathbf{X}^*) \\ &\le \frac{\lambda}{8}. \quad (t^{j_0} \le e^{-2\log n_{(1)}} \le n_{(1)}^{-2})\end{aligned}$$

Proof of (c). By definition, we know that $\mathbf{X}^* + \mathbf{W}^L = \mathbf{Z}_{j_0} + \mathbf{Y}_{j_0}$. Since we have shown

$\|\mathbf{Z}_{j_0}\|_F \le \lambda/8$, it suffices to prove $\|\mathbf{Y}_{j_0}\|_\infty \le \lambda/8$. We have

$$\begin{aligned}
\|\mathbf{Y}_{j_0}\|_\infty &\le q^{-1}\sum_{k=1}^{j_0}\|\mathcal{P}_{\Omega_k}\mathbf{Z}_{k-1}\|_\infty \\
&\le q^{-1}\sum_{k=1}^{j_0}\epsilon^{k-1}\|\mathbf{X}^*\|_\infty \quad \text{(by Lemma 39)} \\
&\le \frac{n_{(2)}\epsilon^2}{C_0\mu r\log n_{(1)}}\sqrt{\frac{\mu r}{n_{(1)}n_{(2)}}}\sigma_r(\mathbf{X}^*) \quad \text{(by incoherence (3.6))} \\
&\le \frac{\lambda}{8},
\end{aligned}$$

if we choose $\epsilon = C\left(\frac{\mu r(\log n_{(1)})^2}{n_{(2)}}\right)^{1/4}$ for an absolute constant C. This can be true once the constant ρ_r is sufficiently small. □

We now prove that $\mathbf{W}^S$ given by (3.24) obeys dual condition (3.26). We have the following.

Lemma 44. *Assume $\Omega \sim \mathsf{Ber}(p)$. Then under the other assumptions of Theorem 19, $\mathbf{W}^S$ given by (3.24) obeys dual condition (3.26).*

Proof. According to the standard de-randomization argument [57], it is equivalent to studying the case when the signs δ_{ij} of $\mathbf{S}^*_{ij}$ are independently distributed as

$$\delta_{ij} = \begin{cases} 1, & \text{w.p. } p/2, \\ 0, & \text{w.p. } 1-p, \\ -1, & \text{w.p. } p/2. \end{cases}$$

Proof of (d). Recall that

$$\begin{aligned}
\mathbf{W}^S &= \lambda\mathcal{P}_{\mathcal{T}^\perp}\sum_{k\ge 0}(\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega)^k\mathrm{sign}(\mathbf{S}^*) \\
&= \lambda\mathcal{P}_{\mathcal{T}^\perp}\mathrm{sign}(\mathbf{S}^*) + \lambda\mathcal{P}_{\mathcal{T}^\perp}\sum_{k\ge 1}(\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega)^k\mathrm{sign}(\mathbf{S}^*).
\end{aligned}$$

To bound the first term, we have $\|\mathrm{sign}(\mathbf{S}^*)\| \le 4\sqrt{n_{(1)}p}$ [235]. So $\|\lambda\mathcal{P}_{\mathcal{T}^\perp}\mathrm{sign}(\mathbf{S}^*)\| \le \lambda\|\mathrm{sign}(\mathbf{S}^*)\| \le 4\sqrt{p}\sigma_r(\mathbf{X}^*) \le \sigma_r(\mathbf{X}^*)/8$.

We now bound the second term. Let $\mathcal{G} = \sum_{k\ge 1}(\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}}\mathcal{P}_\Omega)^k$, which is self-adjoint, and denote by N_{n_1} and N_{n_2} the $\frac{1}{2}$-nets of $\mathbb{S}^{n_1-1}$ and $\mathbb{S}^{n_1-1}$ of sizes at most 6^{n_1} and 6^{n_2}, respectively [145]. We know that [[235], Lemma 5.4]

$$\begin{aligned}
\|\mathcal{G}(\mathrm{sign}(\mathbf{S}^*))\| &= \sup_{\mathbf{x}\in\mathbb{S}^{n_2-1},\mathbf{y}\in\mathbb{S}^{n_1-1}}\langle\mathcal{G}(\mathbf{y}\mathbf{x}^T),\mathrm{sign}(\mathbf{S}^*)\rangle \\
&\le 4\sup_{\mathbf{x}\in N_{n_2},\mathbf{y}\in N_{n_1}}\langle\mathcal{G}(\mathbf{y}\mathbf{x}^T),\mathrm{sign}(\mathbf{S}^*)\rangle.
\end{aligned}$$

Consider the random variable $X(\mathbf{x},\mathbf{y}) = \langle \mathcal{G}(\mathbf{y}\mathbf{x}^T), \mathrm{sign}(\mathbf{S}^*)\rangle$ which has zero expectation. By Hoeffding's inequality, we have

$$\Pr(|X(\mathbf{x},\mathbf{y})| > t) \le 2\exp\left(-\frac{t^2}{2\|\mathcal{G}(\mathbf{x}\mathbf{y}^T)\|_F^2}\right) \le 2\exp\left(-\frac{t^2}{2\|\mathcal{G}\|^2}\right).$$

Therefore, by a union bound,

$$\Pr(\|\mathcal{G}(\mathrm{sign}(\mathbf{S}^*))\| > t) \le 2\times 6^{n_1+n_2}\exp\left(-\frac{t^2}{8\|\mathcal{G}\|^2}\right).$$

Note that conditioned on the event $\{\|\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\| \le \sigma\}$, we have $\|\mathcal{G}\| = \left\|\sum_{k\ge 1}(\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega)^k\right\| \le \frac{\sigma^2}{1-\sigma^2}$. So

$$\begin{aligned}&\Pr(\lambda\|\mathcal{G}(\mathrm{sign}(\mathbf{S}^*))\| > t)\\ &\le 2\times 6^{n_1+n_2}\exp\left(-\frac{t^2}{8\lambda^2}\left(\frac{1-\sigma^2}{\sigma^2}\right)^2\right)\Pr(\|\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\| \le \sigma) + \Pr(\|\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\| > \sigma).\end{aligned}$$

Lemma 42 guarantees that event $\{\|\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\| \le \sigma\}$ holds with high probability for a very small absolute constant σ. Setting $t = \frac{\sigma_r(\mathbf{X}^*)}{8}$, this completes the proof of (d). □

Proof of (e). Recall that $\mathbf{W}^S = \lambda\mathcal{P}_{\mathcal{T}^\perp}\sum_{k\ge 0}(\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega)^k\mathrm{sign}(\mathbf{S}^*)$ and so

$$\begin{aligned}\mathcal{P}_{\Omega^\perp}\mathbf{W}^S &= \lambda\mathcal{P}_{\Omega^\perp}(\mathcal{I}-\mathcal{P}_\mathcal{T})\sum_{k\ge 0}(\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega)^k\mathrm{sign}(\mathbf{S}^*)\\ &= -\lambda\mathcal{P}_{\Omega^\perp}\mathcal{P}_\mathcal{T}\sum_{k\ge 0}(\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega)^k\mathrm{sign}(\mathbf{S}^*).\end{aligned}$$

Then for any $(i,j)\in\Omega^\perp$, we have

$$\mathbf{W}^S_{ij} = \langle \mathbf{W}^S, \mathbf{e}_i\mathbf{e}_j^T\rangle = \left\langle \lambda\mathrm{sign}(\mathbf{S}^*), -\sum_{k\ge 0}(\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega)^k\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}(\mathbf{e}_i\mathbf{e}_j^T)\right\rangle.$$

Let $X(i,j) = -\sum_{k\ge 0}(\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}\mathcal{P}_\Omega)^k\mathcal{P}_\Omega\mathcal{P}_\mathcal{T}(\mathbf{e}_i\mathbf{e}_j^T)$. By Hoeffding's inequality and a union bound,

$$\Pr\left(\sup_{ij}|\mathbf{W}^S_{ij}| > t\right) \le 2\sum_{ij}\exp\left(-\frac{2t^2}{\lambda^2\|X(i,j)\|_F^2}\right).$$

We note that conditioned on the event $\{\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \le \sigma\}$, for any $(i,j) \in \Omega^\perp$,

$$\begin{aligned}
\|X(i,j)\|_F &\le \frac{1}{1-\sigma^2}\sigma \|\mathcal{P}_\mathcal{T}(\mathbf{e}_i \mathbf{e}_j^T)\|_F \\
&\le \frac{1}{1-\sigma^2}\sigma \sqrt{1 - \|\mathcal{P}_{\mathcal{T}^\perp}(\mathbf{e}_i \mathbf{e}_j^T)\|_F^2} \\
&= \frac{1}{1-\sigma^2}\sigma \sqrt{1 - \|(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{e}_i\|_2^2 \|(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{e}_j\|_2^2} \\
&\le \frac{1}{1-\sigma^2}\sigma \sqrt{1 - \left(1 - \frac{\mu r}{n_{(1)}}\right)\left(1 - \frac{\mu r}{n_{(2)}}\right)} \\
&\le \frac{1}{1-\sigma^2}\sigma \sqrt{\frac{\mu r}{n_{(1)}} + \frac{\mu r}{n_{(2)}}}.
\end{aligned}$$

Then unconditionally,

$$\begin{aligned}
&\Pr\left(\sup_{ij} |\mathbf{W}_{ij}^S| > t\right) \\
&\le 2n_{(1)}n_{(2)} \exp\left(-\frac{2t^2}{\lambda^2}\frac{(1-\sigma^2)^2 n_{(1)} n_{(2)}}{\sigma^2 \mu r (n_{(1)} + n_{(2)})}\right) \Pr(\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| \le \sigma) + \Pr(\|\mathcal{P}_\Omega \mathcal{P}_\mathcal{T}\| > \sigma).
\end{aligned}$$

By Lemma 42 and setting $t = \lambda/4$, the proof of (e) is completed.

Proof of Theorem 17

Our computational lower bound for problem (**P**) assumes the hardness of random 4-SAT.

Conjecture 1 (Random 4-SAT)**.** *Let $c > \ln 2$ be a constant. Consider a random 4-SAT formula on n variables in which each clause has 4 literals, and in which each of the $16n^4$ clauses is picked independently with probability c/n^3. Then any algorithm which always outputs 1 when the random formula is satisfiable, and outputs 0 with probability at least $1/2$ when the random formula is unsatisfiable, must run in $2^{c'n}$ time on some input, where $c' > 0$ is an absolute constant.*

Based on Conjecture 1, we have the following computational lower bound for problem (**P**). We show that problem (**P**) is in general hard for deterministic algorithms. If we additionally assume BPP = P, then the same conclusion holds for randomized algorithms with high probability.

Theorem 17 (Computational Lower Bound. Restated)**.** *Assume Conjecture 1. Then there exists an absolute constant $\epsilon_0 > 0$ for which any algorithm that achieves $(1+\epsilon)$OPT in objective function value for problem* (**P**) *with $\epsilon \le \epsilon_0$, and with constant probability, requires $2^{\Omega(n_1+n_2)}$ time, where* OPT *is the optimum. If in addition,* BPP = P*, then the same conclusion holds for randomized algorithms succeeding with probability at least $2/3$.*

Proof. Theorem 17 is proved by using the hypothesis that random 4-SAT is hard to show hardness of the Maximum Edge Biclique problem for deterministic algorithms.

Definition 3 (Maximum Edge Biclique)**.** *The problem is*

Input: An n-by-n bipartite graph G.

Output: A k_1-by-k_2 complete bipartite subgraph of G, such that $k_1 \cdot k_2$ is maximized.

[94] showed that under the random 4-SAT assumption there exist two constants $\epsilon_1 > \epsilon_2 > 0$ such that no efficient deterministic algorithm is able to distinguish between bipartite graphs $G(U, V, E)$ with $|U| = |V| = n$ which have a clique of size $\geq (n/16)^2(1+\epsilon_1)$ and those in which all bipartite cliques are of size $\leq (n/16)^2(1+\epsilon_2)$. The reduction uses a bipartite graph G with at least tn^2 edges with large probability, for a constant t.

Given a given bipartite graph $G(U, V, E)$, define $H(\cdot)$ as follows. Define the matrix $\mathbf{Y}$ and $\mathbf{W}$: $\mathbf{Y}_{ij} = 1$ if edge $(U_i, V_j) \in E$, $\mathbf{Y}_{ij} = 0$ if edge $(U_i, V_j) \notin E$; $\mathbf{W}_{ij} = 1$ if edge $(U_i, V_j) \in E$, and $\mathbf{W}_{ij} = \text{poly}(n)$ if edge $(U_i, V_j) \notin E$. Choose a large enough constant $\beta > 0$ and let $H(\mathbf{AB}) = \beta \sum_{ij} \mathbf{W}_{ij}^2 (\mathbf{Y}_{ij} - (\mathbf{AB})_{ij})^2$. Now, if there exists a biclique in G with at least $(n/16)^2(1+\epsilon_2)$ edges, then the number of remaining edges is at most $tn^2 - (n/16)^2(1+\epsilon_1)$, and so the solution to $\min H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2$ has cost at most $\beta[tn^2 - (n/16)^2(1+\epsilon_1)] + n^2$. On the other hand, if there does not exist a biclique that has more than $(n/16)^2(1+\epsilon_2)$ edges, then the number of remaining edges is at least $(n/16)^2(1+\epsilon_2)$, and so any solution to $\min H(\mathbf{AB}) + \frac{1}{2}\|\mathbf{AB}\|_F^2$ has cost at least $\beta[tn^2 - (n/16)^2(1+\epsilon_2)]$. Choose β large enough so that $\beta[tn^2 - (n/16)^2(1+\epsilon_2)] > \beta[tn^2 - (n/16)^2(1+\epsilon_1)] + n^2$. This combined with the result in [94] completes the proof for deterministic algorithms.

To rule out randomized algorithms running in time $2^{\alpha(n_1+n_2)}$ for some function α of n_1, n_2 for which $\alpha = o(1)$, observe that we can define a new problem which is the same as problem (**P**) except the input description of H is padded with a string of 1s of length $2^{(\alpha/2)(n_1+n_2)}$. This string is irrelevant for solving problem (**P**) but changes the input size to $N = \text{poly}(n_1, n_2) + 2^{(\alpha/2)(n_1+n_2)}$. By the argument in the previous paragraph, any deterministic algorithm still requires $2^{\Omega(n)} = N^{\omega(1)}$ time to solve this problem, which is super-polynomial in the new input size N. However, if a randomized algorithm can solve it in $2^{\alpha(n_1+n_2)}$ time, then it runs in $\text{poly}(N)$ time. This contradicts the assumption that BPP = P. This completes the proof. □

3.2 Property Testing of Matrix Rank

3.2.1 Introduction

Data intrinsic dimensionality is a central object of study in compressed sensing, sketching, numerical linear algebra, machine learning, and many other domains [68, 139, 158, 241, 251, 252, 254]. In compressed sensing and sketching, the study of intrinsic dimensionality has led to significant advances in compressing the data to a size that is far smaller than the ambient dimension while still preserving useful properties of the signal [17, 176]. In numerical linear algebra and machine learning, understanding intrinsic dimensionality serves as a necessary condition for the success of various subspace recovery problems [108], e.g., matrix completion [106, 123, 220, 255] and robust PCA [26, 43, 253]. The focus of this work is on the intrinsic dimensionality of matrices, such as the rank, stable rank, Schatten-p norms, and SVD entropy. The stable rank is defined to be the squared ratio of the Frobenius norm and the largest singular value, and the Schatten-p norm is the ℓ_p norm of the singular values. We study these quantities in the framework of non-adaptive property testing [60, 74, 183]: given non-adaptive query access to the unknown matrix $\mathbf{A} \in \mathbb{F}^{n \times n}$ over a field $\mathbb{F}$, our goal is to determine whether $\mathbf{A}$ is of dimension d (where dimension depends

on the specific problem), or is ϵ-far from having this property. The latter means that at least an ϵ-fraction of entries of $\mathbf{A}$ should be modified in order to have dimension d. Query access typically comes in the form of reading a single entry of the matrix, though we will also discuss sensing models where a query returns the value $\langle \mathbf{X}_i, \mathbf{A} \rangle := \mathsf{tr}(\mathbf{X}_i^\top \mathbf{A})$ for a given $\mathbf{X}_i$. Without making assumptions on $\mathbf{A}$, we would like to choose our sample pattern or set $\{\mathbf{X}_i\}$ of query matrices so that the query complexity is as small as possible.

Despite a large amount of work on testing matrix rank, many fundamental questions remain open. In the rank testing problem in the sampling model, one such question is to design an efficient algorithm that can distinguish rank-d vs. ϵ-far from rank-d with optimal sample complexity. The best-known sampling upper bound for non-adaptive rank testing for general d is $\mathcal{O}(d^2/\epsilon^2)$, which is achieved simply by sampling an $\mathcal{O}(d/\epsilon) \times \mathcal{O}(d/\epsilon)$ submatrix uniformly at random [139]. For arbitrary fields $\mathbb{F}$, only an $\Omega((1/\epsilon)\log(1/\epsilon))$ lower bound for constant d is known [153].

Besides the rank problem above, testing many numerical properties of real matrices has yet to be explored. For example, it is unknown what the query complexity is for the stable rank, which is a natural relaxation of rank in applications. Other examples for which previously we had no bounds are the Schatten-p norms and SVD entropy. We discuss these problems in a new property testing framework that we call the *bounded entry model*. This model has many realistic applications in the Netflix challenge [138], where each entry of the matrix corresponds to the rating from a customer to a movie, ranging from 1 to 5. Understanding the query complexity of testing numerical properties in the bounded entry model is an important problem in recommendation systems and applications of matrix completion, where often entries are bounded.

3.2.2 Our results on sample efficiency

Our work has two parts: (1) we resolve the query complexity of non-adaptive matrix rank testing, a well-studied problem in this model, and (2) we develop a new framework for testing numerical properties of real matrices, including the stable rank, the Schatten-p norms and the SVD entropy. Our results are summarized in Table 3.3. We use $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ notation to hide polylogarithmic factors in the arguments inside. For the rank testing results, the hidden polylogarithmic factors depend only on d and $1/\epsilon$ and do not depend on n; for the other problems, they may depend on n.

Rank testing. We first study the rank testing problem when we can only non-adaptively query entries. The goal is to design a sampling scheme on the entries of the unknown matrix $\mathbf{A}$ and an algorithm so that we can distinguish whether $\mathbf{A}$ is of rank d, or at least an ϵ-fraction of entries of $\mathbf{A}$ should be modified in order to reduce the rank to d. This problem was first proposed by Krauthgamer and Sasson in [139] with a sample complexity upper bound of $\mathcal{O}(d^2/\epsilon^2)$. In this work, we improve this to $\widetilde{\mathcal{O}}(d^2/\epsilon)$ for every d and ϵ, and complement this with a matching lower bound, showing that any algorithm with constant success probability requires at least $\widetilde{\Omega}(d^2/\epsilon)$ samples:

Theorem 25. *For any matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ over any field, there is a randomized non-adaptive sampling algorithm which reads $\widetilde{\mathcal{O}}(d^2/\epsilon)$ entries and runs in $\mathsf{poly}(d/\epsilon)$ time, and with high probability correctly solves the rank testing problem. Further, any non-adaptive algorithm with constant success probability requires $\widetilde{\Omega}(d^2/\epsilon)$ samples over $\mathbb{R}$ or any finite field.*

Table 3.3: Query complexity results for non-adaptive testing of the rank, stable rank, Schatten-p norms, and SVD entropy. The testing of the stable rank, Schatten p-norm and SVD entropy are considered in the bounded entry model.

Problems	Rank	Stable Rank	Schatten-p Norm	Entropy
Sampling	$\widetilde{\mathcal{O}}(d^2/\epsilon)$ (all fields) $\widetilde{\Omega}(d^2/\epsilon)$ (finite fields and $\mathbb{R}$)	$\widetilde{\mathcal{O}}(d^3/\epsilon^4)$ $\widetilde{\Omega}(d^2/\epsilon^2)^\dagger$	$\widetilde{\mathcal{O}}(1/\epsilon^{4p/(p-2)})$ $(p > 2)$ $\Omega(n)$ $(p \in [1, 2))$	$\Omega(n)^\dagger$
Sensing	$\mathcal{O}(d^2)$ (all fields) $\widetilde{\Omega}(d^2)$ (finite fields)	$\widetilde{\mathcal{O}}(d^{2.5}/\epsilon^2)$ $\widetilde{\Omega}(d^2/\epsilon^2)^\dagger$		

† The lower bound involves a reparameterization of the testing problem.

Our non-adaptive sample complexity bound of $\widetilde{\mathcal{O}}(d^2/\epsilon)$ matches what is known with adaptive queries [153], and thus we show the best known upper bound might as well be non-adaptive.

New framework for testing matrix properties. Testing rank is only one of many tasks in determining if a matrix has low intrinsic dimensionality. In several applications, we require a less fragile measure of the collinearity of rows and columns, which is known as the stable rank [223]. We introduce what we call the *bounded entry model* as a new framework for studying such problems through the lens of property testing. In this model, we require all entries of a matrix to be bounded by 1 in absolute value. Boundedness has many natural applications in recommendation systems, e.g., the user-item matrix of preferences for products by customers has bounded entries in the Netflix challenge [138]. Indeed, there are many user rating matrices, etc., which naturally have a small number of discrete values, and therefore fit into a bounded entry model. The boundedness of entries also avoids trivialities in which one can modify a matrix to have a property by setting a single entry to be arbitrarily large, which, e.g., could make the stable rank arbitrarily close to 1.

Our model is a generalization of previous work in which stable rank testing was done in a model for which all rows had to have bounded norm [153], and the algorithm is only allowed to change entire rows at a time. As our non-adaptive rank testing algorithm will illustrate, one can sometimes do better by only reading certain carefully selected entries in rows and columns. Indeed, this is precisely the source of our improvement over prior work. Thus, the restriction of having to read an entire row is often unnatural, and further motivates our bounded entry model. We first informally state our main theorems on stable rank testing in this model.

Theorem 26. *There is a randomized algorithm for the stable rank testing problem to decide whether a matrix is of stable rank at most d or is ϵ-far from stable rank at most d, with failure probability at most $1/3$, and which reads $\widetilde{\mathcal{O}}(d^3/\epsilon^4)$ entries.*

Theorem 26 relies on a new $(1 \pm \tau)$-approximate non-adaptive estimator of the largest singular value of a matrix, which may be of independent interest.

Theorem 27. *Suppose that $\mathbf{A} \in \mathbb{R}^{n \times n}$ has stable rank $\mathcal{O}(d)$ and $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$. Then in the bounded entry model, there is a randomized non-adaptive sampling algorithm which reads $\widetilde{\mathcal{O}}(d^2/\tau^4)$ entries and with probability at least 0.9, outputs a $(1 \pm \tau)$-approximation to the largest singular value of $\mathbf{A}$.*

We remark that when the stable rank is constant and the singular value gap $\sigma_1(\mathbf{A})/\sigma_2(\mathbf{A}) =$

$(1/\tau)^\gamma$ for an arbitrary constant $\gamma > 0$, the operator norm can be estimated up to a $(1 \pm \tau)$-factor by querying $\mathcal{O}(1/\tau^2)$ entries non-adaptively.

Other measures of intrinsic dimensionality include matrix norms, such as the Schatten-p norm $\|\cdot\|_{\mathcal{S}_p}$, which measures the central tendency of the singular values. Familiar special cases are $p = 1$, 2 and ∞, which have applications in differential privacy [24, 109] and non-convex optimization [26, 76] for $p = 1$, and in numerical linear algebra [167] for $p \in \{2, \infty\}$. Matrix norms have been studied extensively in the streaming literature [149, 150, 151, 152], though their study in property testing models is lacking.

We study non-adaptive algorithms for these problems in the bounded entry model. We consider distinguishing whether $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ is at least cn^p for $p > 2$ (at least $cn^{1+1/p}$ for $p < 2$), or at least an ϵ-fraction of entries of $\mathbf{A}$ should be modified in order to have this property, where c is a constant (depending only on p). We choose the threshold n^p for $p > 2$ and $n^{1+1/p}$ for $p < 2$ because they are the largest possible value of $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ for $\mathbf{A}$ under the bounded entry model. When $p > 2$, $\|\mathbf{A}\|_{\mathcal{S}_p}$ is maximized when $\mathbf{A}$ is of rank 1, and so this gives us an alternative "measure" of how close we are to a rank-1 matrix. Testing whether $\|\mathbf{A}\|_{\mathcal{S}_p}$ is large in sublinear time allows us to quickly determine whether $\mathbf{A}$ can be well approximated by a low-rank matrix, which could save us from running more expensive low-rank approximation algorithms. In contrast, when $p < 2$, $\|\mathbf{A}\|_{\mathcal{S}_p}$ is maximized when $\mathbf{A}$ has a flat spectrum, and so is a measure of how well-conditioned $\mathbf{A}$ is. A fast tester could save us from running expensive pre-conditioning algorithms. We state our main theorems informally below.

Theorem 28. *For constant $p > 2$, there is a randomized algorithm for the Schatten-p norm testing problem with failure probability at most $1/3$ which reads $\widetilde{\mathcal{O}}(1/\epsilon^{4p/(p-2)})$ entries.*

Results for sensing algorithms. We also consider a more powerful query oracle known as the *sensing model*, where query access comes in the form of $\langle \mathbf{X}_i, \mathbf{A} \rangle := \mathsf{tr}(\mathbf{X}_i^\top \mathbf{A})$ for some sensing matrices $\mathbf{X}_i$ of our choice. These matrices are chosen non-adaptively. We show differences in the complexity of the above problems in this and the above sampling model. For the testing and the estimation problems above, we have the following results in the sensing model:

Theorem 29. *Over an arbitrary finite field, any non-adaptive algorithm with constant success probability for the rank testing problem in the sensing model requires $\widetilde{\Omega}(d^2)$ queries.*

Theorem 30. *There is a randomized algorithm for the stable rank testing problem with failure probability at most $1/3$ in the sensing model with $\widetilde{\mathcal{O}}(d^{2.5}/\epsilon^2)$ queries. Further, any algorithm with constant success probability requires $\widetilde{\Omega}(d^2/\epsilon^2)$ queries.*

Theorem 31. *For $p \in [1, 2)$, any algorithm for the Schatten-p norm testing problem with failure probability at most $1/3$ requires $\Omega(n)$ queries.*

Theorem 32. *Suppose that $\mathbf{A} \in \mathbb{R}^{n \times n}$ has stable rank $\mathcal{O}(d)$ and $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$. In the bounded entry model, there is a randomized sensing algorithm with sensing complexity $\widetilde{\mathcal{O}}(d^2/\tau^2)$ which outputs a $(1 \pm \tau)$-approximation to the largest singular value with probability at least 0.9. This sensing complexity is optimal up to polylogarithmic factors.*

We also provide an $\Omega(n)$ query lower bound for the SVD entropy testing in the sensing model.

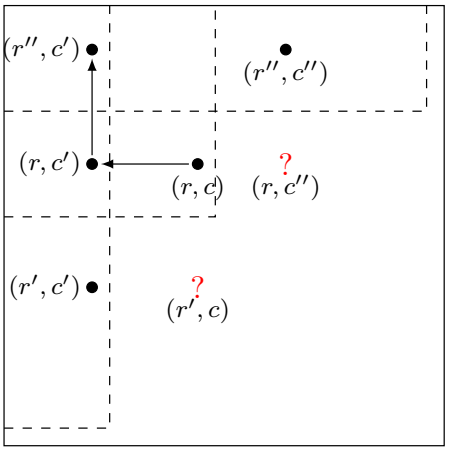


Figure 3.6: Our sampling scheme (the region enclosed by the dotted lines modulo permutation of rows and columns) and our path of augmenting a 1×1 submatrix. The whole region is the $\mathcal{O}(d/\epsilon) \times \mathcal{O}(d/\epsilon)$ submatrix sampled from the $n \times n$ matrix.

3.2.3 Our techniques

We now discuss the techniques in more detail, starting with the rank testing problem.

Prior to the work of [153], the only known algorithm for $d = 1$ was to sample an $\mathcal{O}(1/\epsilon) \times \mathcal{O}(1/\epsilon)$ submatrix. In contrast, for rank 1 an algorithm in [153] samples $\mathcal{O}(\log(1/\epsilon))$ blocks of varying shapes "within a random $\mathcal{O}(1/\epsilon) \times \mathcal{O}(1/\epsilon)$ submatrix" and argues that these shapes are sufficient to expose a rank-2 submatrix. For $d = 1$ the goal is to augment a 1×1 matrix to a full-rank 2×2 matrix. One can show that with good probability, one of the shapes "catches" an entry that enlarges the 1×1 matrix to a full-rank 2×2 matrix. For instance, in Figure 3.6, (r, c) is our 1×1 matrix and the leftmost vertical block catches an "augmentation element" (r', c') which makes $\begin{bmatrix} (r,c') & (r,c) \\ (r',c') & (r',c) \end{bmatrix}$ a full-rank 2×2 matrix. Hereby, the "augmentation element" means the entry by adding which we augment a $r \times r$ matrix to a $(r+1) \times (r+1)$ matrix. In [153], an argument was claimed for $d = 1$, though we note an omission in their analysis. Namely, the "augmentation entry" (r', c') can be the 1×1 matrix we begin with (meaning that $\mathbf{A}_{r',c'} \neq 0$, which might not be true), and since one can show that both (r, c) and (r', c') fall inside the same sampling block with good probability, the 2×2 matrix would be fully observed and the algorithm would thus be able to determine that it has rank 2. However, it is possible that $\mathbf{A}_{r',c'} = 0$ and (r', c') would not be a starting point (i.e., a 1×1 rank-1 matrix), and in this case, (r', c) may not be observed, as illustrated in Figure 3.6. In this case the algorithm will not be able to determine whether the augmented 2×2 matrix is of full rank. For $d > 1$, nothing was known. One issue is that the probability of fully observing a $d \times d$ submatrix within these shapes is very small. To overcome this, we propose what we call *rebasing* and *transformation to a canonical structure*. These arguments allow us to tolerate unobserved entries and conveniently obtain an algorithm for every d, completing the analysis of [153] for $d = 1$ in the process.

Rebasing argument + canonical structure. The best previous result for the rank testing problem uniformly samples an $\mathcal{O}(d/\epsilon) \times \mathcal{O}(d/\epsilon)$ submatrix and argues that one can find a $(d+1) \times (d+1)$ full-rank submatrix within it when $\mathbf{A}$ is ϵ-far from rank-d [139]. In contrast, our algorithm

follows from subsampling an $\mathcal{O}(\epsilon)$-fraction of entries in this $\mathcal{O}(d/\epsilon) \times \mathcal{O}(d/\epsilon)$ submatrix. Let $\mathcal{R}_1 \subseteq \cdots \subseteq \mathcal{R}_m$ and $\mathcal{C}_1 \supseteq \cdots \supseteq \mathcal{C}_m$ be the indices of subsampled rows and columns, respectively, with $m = \mathcal{O}(\log(1/\epsilon))$. We choose these indices uniformly at random such that $|\mathcal{R}_i| = \widetilde{\mathcal{O}}(d2^i)$ and $|\mathcal{C}_i| = \widetilde{\mathcal{O}}(d/(2^i\epsilon))$, and sample the entries in all m blocks determined by the $\{\mathcal{R}_i, \mathcal{C}_i\}$ (see Figure 3.6, where our sampled regions are enclosed by the dotted lines). Since there are $\widetilde{\mathcal{O}}(\log(1/\epsilon))$ blocks and in each block we sample $\widetilde{\mathcal{O}}(d^2/\epsilon)$ entries, the sample complexity of our algorithm is as small as $\widetilde{\mathcal{O}}(d^2/\epsilon)$.

The correctness of our algorithm for $d = 1$ follows from what we call a rebasing argument. Starting from an empty matrix, our goal is to maintain and augment the matrix to a 2×2 full-rank matrix when $\mathbf{A}$ is ϵ-far from rank-d. By a level-set argument, we show an oracle lemma which states that *we can augment any $r \times r$ full-rank matrix to an $(r+1) \times (r+1)$ full-rank matrix by an augmentation entry in the sampled region*, as long as $r \leq d$ and $\mathbf{A}$ is ϵ-far from rank-d. Therefore, as a first step we successfully find a 1×1 full-rank matrix, say with index (r, c), in the sampled region. We then argue that we can either (a) find a 2×2 fully-observed full-rank submatrix or a 2×2 submatrix which is not fully observed but we know must be of full rank, or (b) move our maintained 1×1 full-rank submatrix upwards or leftwards to a new 1×1 *full-rank* submatrix and repeat checking whether case (a) happens or not; if not, we implement case (b) again and repeat the procedure. To see case (a), by the oracle lemma, if the augmented entry is (r'', c') (see Figure 3.6), then we fully observe the submatrix determined by (r'', c') and (r, c) and so the algorithm is correct in this case. On the other hand, if the augmented entry is (r', c'), then we fail to see the entry at (r', c). In this case, when $\mathbf{A}_{r,c'} = 0$, then we must have $\mathbf{A}_{r',c'} \neq 0$; otherwise, (r', c') is not an augment of (r, c), which leads to a contradiction with the oracle lemma. Thus we find a 2×2 matrix with structure

$$\begin{bmatrix} \mathbf{A}_{r,c'} & \mathbf{A}_{r,c} \\ \mathbf{A}_{r',c'} & \mathbf{A}_{r',c} \end{bmatrix} = \begin{bmatrix} 0 & \neq 0 \\ \neq 0 & ? \end{bmatrix}, \tag{3.27}$$

which must be of rank 2 despite an unobserved entry, and the algorithm therefore is correct in this case. The remaining case of the analysis above is when $\mathbf{A}_{r,c'} \neq 0$. Instead of trying to augment $\mathbf{A}_{r,c}$, we augment $\mathbf{A}_{r,c'}$ in the next step. Note that the index (r, c') is to the left of (r, c). This leads to case (b). In the worst case, we move the 1×1 non-zero matrix to the uppermost left corner,[5] e.g., (r'', c'). Fortunately, since (r'', c') is in the uppermost left corner, we can, as guaranteed by the oracle lemma, augment it to a 2×2 *fully-observed* full-rank matrix. Again the algorithm outputs correctly in this case.

The analysis becomes more challenging for general d, since the number of unobserved entries (i.e., those entries marked as "?") may propagate as we augment an $r \times r$ submatrix ($r = 1, 2, ..., d$) in each round. To resolve the issue, we maintain a structure (modulo elementary transformations)

[5]The upper-left corner refers to the intersection of all sampled blocks, namely, $\mathcal{R}_1 \times \mathcal{C}_m$; it does not mean the top-left entry.

similar to structure (3.27) for the $r \times r$ submatrix, that is,

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & \cdots & 0 & \neq 0 \\ 0 & 0 & \cdots & 0 & \cdots & \neq 0 & ? \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & \neq 0 & \cdots & ? & \cdots & ? & ? \\ \neq 0 & ? & \cdots & ? & \cdots & ? & ? \end{bmatrix}. \tag{3.28}$$

Since the proposed structure has non-zero determinant, the submatrix is always of full rank. Similar to the case for $d = 1$, we show that we can either (a) augment the $r \times r$ submatrix to an $(r+1) \times (r+1)$ submatrix with the same structure (3.28) (modulo elementary transformations); or (b) find another $r \times r$ submatrix of structure (3.28) that is closer to the upper-left corner than the original $r \times r$ matrix. Hence the algorithm is correct for general d.

Pivot-node assignment. Our rank testing lower bound under the sampling model over a finite field $\mathbb{F}$ follows from distinguishing two hard instances $\mathbf{U}\mathbf{V}^\top$ vs. $\mathbf{W}$, where $\mathbf{U}, \mathbf{V} \in \mathbb{F}^{t \times d}$ and $\mathbf{W} \in \mathbb{F}^{t \times t}$ have i.i.d. entries that are uniform over $\mathbb{F}$. For an observed subset $\mathcal{S}$ of entries with $|\mathcal{S}| = \mathcal{O}(d^2)$, we bound the total variation distance between the distributions of the observed entries in the two cases by a small constant. In particular, we show that the probability $\Pr[(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}]$ is large for any observation $\mathbf{x} \in \mathbb{F}^{|\mathcal{S}|}$, by a *pivot-node assignment* argument, as follows. We reformulate our problem as a bipartite graph assignment problem $G = (L \cup R, E)$, where L corresponds to the rows of $\mathbf{U}$, R the rows of $\mathbf{V}$ and each edge of E one entry in $\mathcal{S}$. We want to assign each node a vector/affine subspace, meaning that the corresponding row in $\mathbf{U}$ or $\mathbf{V}$ will be that vector or in that affine subspace, such that they agree with our observation, i.e., $(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}$. Since $\mathbf{U}, \mathbf{V}$ are random matrices, we assign random vectors to nodes adaptively, one at a time, and try to maintain consistency with the fact that $(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}$. Note that the order of the assignment is important, as a bad choice for an earlier node may invalidate any assignment to a later node. To overcome this issue, we choose nodes of large degrees as *pivot nodes* and assign each non-pivot node adaptively in a careful manner so as to guarantee that the incident pivot nodes will always have valid assignments (which in fact form an affine subspace). In the end we assign the pivot node vectors from their respective affine subspaces. We employ a counting argument for each step in this assignment procedure to lower bound the number of valid assignments, and thus lower bound the probability $\Pr[(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}]$.

The above analysis gives us an $\Omega(d^2)$ lower bound for constant ϵ since $\mathbf{W}$ is constant-far from being of rank d. The desired $\Omega(d^2/\epsilon)$ lower bound follows from planting $\mathbf{U}\mathbf{V}^\top$ vs. $\mathbf{W}$ with $t = \sqrt{\epsilon}n$ into an $n \times n$ matrix at uniformly random positions, and padding zeros everywhere else.

New analytical framework for stable rank, Schatten-p norm, and entropy testing. We propose a new analytical framework by reducing the testing problem to a sequence of estimation problems *without involving* $\mathsf{poly}(n)$ *in the sample complexity*. There is a two-stage estimation in our framework: (1) a constant-approximation to some statistic X of interest (e.g., stable rank) which enables us to distinguish $X \le d$ vs. $X \ge 10d$ for the threshold parameter d of interest. If $X \ge 10d$, we can safely output "$\mathbf{A}$ is far from $X \le d$"; otherwise, the statistic is at most $10d$, and (2) we show that X has a $(1 \pm \epsilon)$-factor difference between "$X \le d$" and "far from $X \le d$", and so we implement a more accurate $(1 \pm \epsilon)$-approximation to distinguish the two cases. The sample

complexity does not depend on n polynomially because (1) the first estimator is "rough" and gives only a constant-factor approximation and (2) the second estimator operates under the condition that $X \leq 10d$ and thus $\mathbf{A}$ has a low intrinsic dimension. We apply the proposed framework to the testing problems of the stable rank and the Schatten-p norm by plugging in our estimators in Theorem 27 and Theorem 32. This analytical framework may be of independent interest to other property testing problems more broadly.

In a number of these problems, a key difficulty is arguing about spectral properties of a matrix $\mathbf{A}$ when it is ϵ-far from having a property, such as having stable rank at most d. Because of the fact that the entries must always be bounded by 1 in absolute value, it becomes non-trivial to argue, for example, that if $\mathbf{A}$ is ϵ-far from having stable rank at most d, that its stable rank is even slightly larger than d. A natural approach is to argue that you could change an ϵ-fraction of rows of $\mathbf{A}$ to agree with a multiple of the top left or right singular vector of $\mathbf{A}$, and since we are still guaranteed to have stable rank at least d after changing such entries, it means that the operator norm of $\mathbf{A}$ must have been small to begin with (which says something about the original stable rank of $\mathbf{A}$, since its Frobenius norm can also be estimated). The trouble is, if the top singular vector has some entries that are very large, and others that are small, one cannot scale the singular vector by a large amount since then we would violate the boundedness criterion of our model. We get around this by arguing there either needs to exist a left or a right singular vector of large ℓ_1-norm (in some cases such vectors may only be right singular vectors, and in other cases only left singular vectors). The ℓ_1-norm is a natural norm to study in this context, since it is dual to the ℓ_∞-norm, which we use to capture the boundedness property of the matrix.

Our lower bounds for the above problems follow from the corresponding sketching lower bounds for the estimation problem in [150, 154], together with rigidity-type results [229] for the hard instances regarding the respective statistic of interest.

3.2.4 Proofs of our main results

Proofs of Theorem 25

Useful lemmas In this section, we study the following problem of testing low-rank matrices.

Problem 1 (Rank Testing with Parameter (n, d, ϵ) in the Sampling Model)**.** *Given a field $\mathbb{F}$ and a matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ which has one of promised properties:*

- H0. *$\mathbf{A}$ has rank at most d;*
- H1. *$\mathbf{A}$ is ϵ-far from having rank at most d, meaning that $\mathbf{A}$ requires changing at least an ϵ-fraction of its entries to have rank at most d.*

The problem is to design a property testing algorithm that outputs H0 *with probability* 1 *if* $\mathbf{A} \in$ H0*, and output* H1 *with probability at least 0.99 if* $\mathbf{A} \in$ H1*, with the least number of queried entries.*

Positive Results Below we provide a non-adaptive algorithm for the rank testing problem under the sampling model with $\widetilde{\mathcal{O}}(\frac{d^2}{\epsilon})$ queries when $\epsilon \leq \frac{1}{e}$. Let $\eta \in (0, \frac{1}{2})$ be such that $\eta \log(\frac{1}{\eta}) = \epsilon$ and let $m = \lceil \log(\frac{1}{\eta}) \rceil$.

We note that the number of entries that Algorithm 7 queries is

$$\mathcal{O}(k \cdot [\log d + \log\log(1/\eta)]^2 d^2 \log^2(1/\eta)/\eta) = \widetilde{\mathcal{O}}(d^2/\epsilon).$$

Algorithm 7 Robust non-adaptive testing of matrix rank

1: Choose $\mathcal{R}_1, \dots, \mathcal{R}_m$ and $\mathcal{C}_1, \dots, \mathcal{C}_m$ from $[n]$ uniformly at random such that

$$\mathcal{R}_1 \subseteq \cdots \subseteq \mathcal{R}_m, \qquad \mathcal{C}_1 \supseteq \cdots \supseteq \mathcal{C}_m,$$

and

$$|\mathcal{R}_i| = c[\log d + \log\log(1/\eta)]d\log(1/\eta)2^i, \qquad |\mathcal{C}_i| = c[\log d + \log\log(1/\eta)]d\log(1/\eta)/(2^i\eta),$$

where $c > 0$ is an absolute constant. To impose containment for $\mathcal{R}_i$'s, $\mathcal{R}_i$ can be formed by appending to $\mathcal{R}_{i-1}$ uniformly random $|\mathcal{R}_i| - |\mathcal{R}_{i-1}|$ rows. The containment for $\mathcal{C}_i$'s can be imposed similarly.

2: Query the entries in $\mathcal{Q} = \bigcup_{i=1}^m (\mathcal{R}_i \times \mathcal{C}_i)$. Note that the entries in $(\mathcal{R}_m \times \mathcal{C}_1) \setminus \mathcal{Q}$ are unobserved. The algorithm solves the following minimization problem by filling in those entries of $\mathbf{A}_{(\mathcal{R}_m \times \mathcal{C}_1)\setminus\mathcal{Q}}$ given input $\mathbf{A}_{\mathcal{Q}}$.

$$r := \min_{\mathbf{A}_{(\mathcal{R}_m \times \mathcal{C}_1)\setminus\mathcal{Q}}} \mathsf{rank}(\mathbf{A}_{\mathcal{R}_m, \mathcal{C}_1}). \tag{3.29}$$

3: Output "$\mathbf{A}$ is ϵ-far from having rank d" if $r > d$; otherwise, output "$\mathbf{A}$ is of rank at most d".

We now prove the correctness of Algorithm 7. Before proceeding, we reproduce the definitions *augment set* and *augment pattern* i and relevant lemmata from [153] as follows.

Definition 4 (Augment)**.** *For $n \times n$ fixed matrix $\mathbf{A}$, we call (r, c) an augment for $\mathcal{R} \times \mathcal{C} \subseteq [n] \times [n]$ if $r \in [n]\setminus\mathcal{R}$, $c \in [n]\setminus\mathcal{C}$ and $\mathsf{rank}(\mathbf{A}_{\mathcal{R}\cup\{r\},\mathcal{C}\cup\{c\}}) > \mathsf{rank}(\mathbf{A}_{\mathcal{R},\mathcal{C}})$. We denote by $\mathsf{aug}(\mathcal{R}, \mathcal{C})$ the set of all the augments for $\mathcal{R} \times \mathcal{C}$, namely,*

$$\mathsf{aug}(\mathcal{R}, \mathcal{C}) = \{(r, c) \in ([n]\setminus\mathcal{R}) \times ([n]\setminus\mathcal{C}) \mid \mathsf{rank}(\mathbf{A}_{\mathcal{R}\cup\{r\},\mathcal{C}\cup\{c\}}) > \mathsf{rank}(\mathbf{A}_{\mathcal{R},\mathcal{C}})\}.$$

Definition 5 (Augment Pattern)**.** *For fixed $\mathcal{R}$, $\mathcal{C}$ and $\mathbf{A}$, define count_r (where $r \in [n]\setminus\mathcal{R}$) to be the number of c's such that $(r, c) \in \mathsf{aug}(\mathcal{R}, \mathcal{C})$. Let $\{\mathsf{count}_i^*\}_{i\in[n-|\mathcal{R}|]}$ the non-increasing reordering of the sequence $\{\mathsf{count}_i\}_{i\in[n]\setminus\mathcal{R}}$, and $\mathsf{count}_i^* = 0$ for $i > n - |\mathcal{R}|$. We say that $(\mathcal{R}, \mathcal{C})$ has augment pattern i on $\mathbf{A}$ if and only if $\mathsf{count}_{n/2^i}^* \geq 2^{i-1}\eta n$.*

Lemma 45. *Let $\mathbf{A}_{\mathcal{R},\mathcal{C}}$ be a $t \times t$ full-rank matrix. If $\mathbf{A}$ is ϵ-far from having rank d and $\mathsf{rank}(\mathbf{A}_{\mathcal{R},\mathcal{C}}) = t \leq d$, then*

$$|\mathsf{aug}(\mathcal{R}, \mathcal{C})| = \sum_{r\in[n]\setminus\mathcal{R}} \mathsf{count}_r = \sum_{i=1}^{n-|\mathcal{R}|} \mathsf{count}_i^* \geq \frac{\epsilon n^2}{3}.$$

Proof. Let $\mathcal{S}$ be the set of entries (r, c) in $\mathcal{R}^c \times \mathcal{C}^c$ such that $\mathsf{rank}(\mathbf{A}_{\mathcal{R}\cup\{r\},\mathcal{C}\cup\{c\}}) > \mathsf{rank}(\mathbf{A}_{\mathcal{R},\mathcal{C}})$, i.e., $\mathcal{S} = \mathsf{aug}(\mathcal{R}, \mathcal{C})$. We will show that $|\mathcal{S}| \geq \epsilon n^2/3$.

Let $\mathcal{T}$ be the complement of $\mathcal{S}$ inside the set $\mathcal{R}^c \times \mathcal{C}^c$. For any $(r, c) \in \mathcal{S}$, we discuss the following two cases.

Case (i). There is $c' \in \mathcal{C}^c$ such that $(r, c') \in \mathcal{T}$ or $r' \in \mathcal{R}^c$ such that $(r', c) \in \mathcal{T}$

In the former case, the row vector $\mathbf{A}_{r,\mathcal{C}\cup\{c'\}}$ is a linear combination of the rows of $\mathbf{A}_{\mathcal{R},\mathcal{C}\cup\{c'\}}$. So we can change the value of $\mathbf{A}_{r,c}$ so that $\mathbf{A}_{r,\mathcal{C}\cup\{c\}}$ is a linear combination of $\mathbf{A}_{\mathcal{R},\mathcal{C}\cup\{c\}}$ with the same representation coefficients as that of $\mathbf{A}_{\mathcal{R},\mathcal{C}\cup\{c'\}}$. Therefore, augmenting $\mathbf{A}_{R,C}$ by the pair (r,c) would not increase $\mathsf{rank}(\mathbf{A}_{\mathcal{R},\mathcal{C}})$. Similarly, if there is $r' \in \mathcal{R}^c$ such that $(r',c) \in T$, we can change the value of $\mathbf{A}_{r,c}$ so that augmenting $\mathbf{A}_{\mathcal{R},\mathcal{C}}$ by the pair (r,c) would not increase $\mathsf{rank}(\mathbf{A}_{\mathcal{R},\mathcal{C}})$. We change at most $|\mathcal{S}|$ entries for both cases combined.

Case (ii). $(r,c') \in \mathcal{S}$ **for all** $c' \in \mathcal{C}^c$ **and** $(r',c) \in \mathcal{S}$ **for all** $r' \in \mathcal{R}^c$

In this case, we can change the entire r-th row and c-th column of $\mathbf{A}$ so that $\mathsf{rank}(\mathbf{A}_{\mathcal{R},\mathcal{C}})$ does not increase by augmenting it with any pair in $(\mathcal{R}^c \times \{c\}) \cup (\{r\} \times \mathcal{C}^c)$. Recall that $n \geq 2d$ and $t \leq d$. It follows that $n \leq 2(n-t)$. Therefore, this specific pair (r,c) would lead to the change of at most $2n \leq 2(n-t) + 2(n-t) \leq 2(|\mathcal{R}^c| + |\mathcal{C}^c|)$ entries. For all such (r,c)'s, we change at most $2|\mathcal{S}|$ entries in this case.

In summary, we can change at most $3|\mathcal{S}|$ entries of $\mathbf{A}$ so that $\mathsf{rank}(\mathbf{A}_{\mathcal{R},\mathcal{C}})$ cannot increase by augmenting $\mathbf{A}_{\mathcal{R},\mathcal{C}}$ with any pair $(r,c) \in \mathcal{R}^c \times \mathcal{C}^c$. Since $\mathbf{A}$ is ϵ-far from being rank d, we must have $3|\mathcal{S}| \geq \epsilon n^2$. Namely, $|\mathsf{aug}(\mathcal{R},\mathcal{C})| = |\mathcal{S}| \geq \epsilon n^2/3$. □

Lemma 46. *Let $\mathbf{A}_{\mathcal{R},\mathcal{C}}$ be a $t \times t$ full-rank matrix. If $\mathbf{A}$ is ϵ-far from being rank d and $\mathsf{rank}(\mathbf{A}_{\mathcal{R},\mathcal{C}}) = t \leq d$, then there exists i such that $(\mathcal{R},\mathcal{C})$ has augment pattern i.*

Proof. Suppose that $(\mathcal{R},\mathcal{C})$ does not have any augment pattern in $[\log(1/\eta)]$. That is

$$\mathsf{count}^*_{n/2^i} < 2^{i-1}\eta n, \quad i = 1,2,...,\log(1/\eta).$$

Therefore,

$$\begin{aligned}
\sum_i \mathsf{count}^*_i &= \sum_{i=\frac{n}{2}+1}^{n} \mathsf{count}^*_i + \sum_{i=\frac{n}{4}+1}^{\frac{n}{2}} \mathsf{count}^*_i + ... + \sum_{i=\frac{n}{2^{\log(1/\eta)}}+1}^{\frac{n}{2^{\log(1/\eta)-1}}} \mathsf{count}^*_i + \sum_{i=1}^{\eta n} \mathsf{count}^*_i \\
&\leq \frac{n}{2}\mathsf{count}^*_{\frac{n}{2}+1} + \frac{n}{4}\mathsf{count}^*_{\frac{n}{4}+1} + \cdots + \frac{n}{2^{\log(1/\eta)}}\mathsf{count}^*_{\frac{n}{2^{\log(1/\eta)}}+1} + \eta n \mathsf{count}^*_1 \\
&< \frac{n}{2}\eta n + \frac{n}{4}2\eta n + ... + \eta n 2^{\log(1/\eta)-1}\eta n + \eta n^2 \\
&= \frac{\eta n^2}{2}(\log(1/\eta) + 2) \\
&\leq \frac{\epsilon n^2}{3},
\end{aligned}$$

which leads to a contradiction with Lemma 45. □

Lemma 47. *For fixed $(\mathcal{R},\mathcal{C})$, suppose that $(\mathcal{R},\mathcal{C})$ has augment pattern i on $\mathbf{A}$. Let $\mathcal{R}',\mathcal{C}' \subseteq [n]$ be uniformly random such that $|\mathcal{R}'| = c2^i$, $|\mathcal{C}'| = c/(2^i\eta)$. Then the probability that $(\mathcal{R}',\mathcal{C}')$ contains at least one augment of $(\mathcal{R},\mathcal{C})$ on $\mathbf{A}$ is at least $1 - 2e^{-c/2}$.*

Proof. Since $(\mathcal{R},\mathcal{C})$ has augment pattern i on matrix $\mathbf{A}$, the probability that $\mathcal{R}'$ (and $\mathcal{C}'$) does not hit row (and column) of any augment is $(1-2^{-i})^{c2^i}$ (and $(1-2^{i-1}\eta)^{c/(2^i\eta)}$). Therefore, the probability that $(\mathcal{R}',\mathcal{C}')$ hits at least one augment is given by

$$\left(1-(1-2^{-i})^{c2^i}\right)\left(1-(1-2^{i-1}\eta)^{c/(2^i\eta)}\right) \geq 1 - \frac{2}{e^{c/2}}.$$ □

Warm-up: the case of $d = 1$. Without loss of generality, we may permute the rows and columns of $\mathbf{A}$ and assume that $\mathcal{R}_i = \{1, \ldots, |\mathcal{R}_i|\}$ and $\mathcal{C}_i = \{1, \ldots, |\mathcal{C}_i|\}$ for all $i \le \lceil \log \frac{1}{\eta} \rceil$.

Theorem 33. *Let $\epsilon \le 1/e$ and $d = 1$. For any matrix $\mathbf{A}$, the probability that Algorithm 7 fails is at most $1/3$.*

Proof. If $\mathbf{A}$ is of rank at most d, then the algorithm will never make mistake; so we assume that $\mathbf{A}$ is ϵ-far from being rank d in the proof below.

Lemma 46 shows that $(\emptyset, \emptyset)$ has some augment pattern s and by Lemma 47, with probability at least $1 - 2e^{-c/2}$ there exists $(r, c) \in (\mathcal{R}_s, \mathcal{C}_s)$ such that $(r, c) \in \mathsf{aug}(\emptyset, \emptyset)$, i.e., $\mathbf{A}_{(r,c)} \neq 0$. We now argue that the rank-1 submatrix $\mathbf{A}_{(r,c)}$ can be augmented to a rank-2 submatrix.

Again by Lemma 46, $(\{r\}, \{c\})$ has an augment pattern j; otherwise, $\mathbf{A}$ is not ϵ-far from being rank-d, and with probability at least $1 - 2e^{-c/2}$ there exists $(r', c') \in (\mathcal{R}_j, \mathcal{C}_j)$ such that $(r', c') \in \mathsf{aug}(\{r\}, \{c\})$. We now discuss three cases based on the position of (r', c') in relation to (r, c).

Case (i). $(r', c') \in R_s \times C_s$.

By Lemma 47, with probability at least $1 - 2e^{-c/2}$, $\mathcal{R}_j \times \mathcal{C}_j$ contains an argument for (r, c), denoted by (r', c'). By construction of $\{\mathcal{R}_j\}$ and $\{\mathcal{C}_j\}$, (r, c') and (r', c) are also queried (See Figure 3.7(a)). Thus we find a 2×2 non-singular matrix. The algorithm answers correctly with probability at least $1 - 4e^{-c_0/4} > 2/3$ in this case.

Case (ii). $r' \notin \mathcal{R}_s$ or $c' \notin \mathcal{C}_s$.

In this case, we show that starting from $\mathbf{A}_{r,c}$, we can always find a path for the non-singular 1×1 submatrix $\mathbf{A}_{*,*}$ such that the index $(*, *)$ always moves to the left or above, so we make progress towards case (i): we note that the non-zero element in the most upper left corner can always be augmented with three queried elements in the same augment pattern (i.e., Case (i)), because the uppermost left corner belongs to all $(\mathcal{R}_i, \mathcal{C}_i)$'s by construction. We now show how to find the path (Please refer to Figure 3.7(b) for the following proofs).

For index (r, c) such that $\mathbf{A}_{r,c} \neq 0$, if $r' \notin \mathcal{R}_s$ or $c' \notin \mathcal{C}_s$ (say $r' \notin \mathcal{R}_s$ at the moment), then by Lemma 47, there exists an index $(r', c') \in (\mathcal{R}_j, \mathcal{C}_j)$ such that (r, c) can be augmented by (r', c'). However, we cannot observe $\mathbf{A}_{r',c}$ so we do not find a 2×2 submatrix at the moment. To make progress, we further discuss two cases.

Case (ii.1). $\mathbf{A}_{r,c'} = 0$ and $\mathbf{A}_{r',c'} = 0$.

This case is impossible; otherwise, (r', c') cannot be an augment of (r, c).

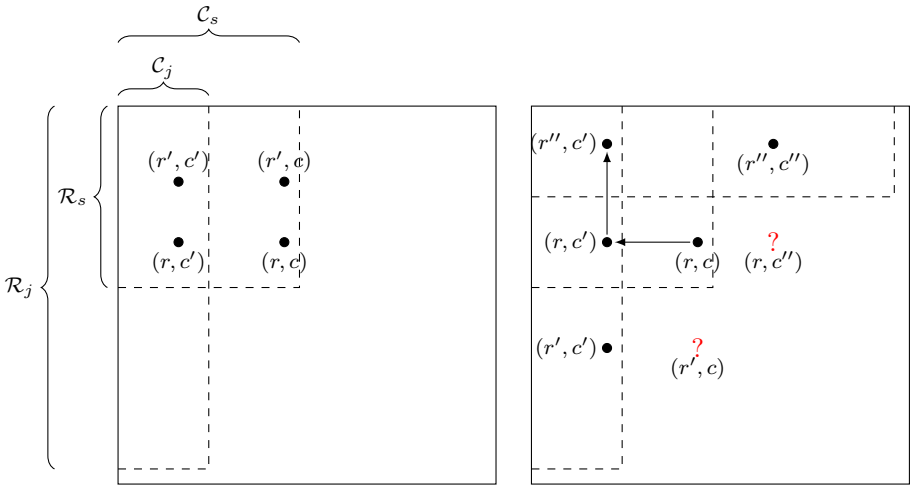
Case (ii.2). $\mathbf{A}_{r,c'} = 0$ and $\mathbf{A}_{r',c'} \neq 0$.

Since $\mathbf{A}_{r,c} \neq 0$, $\mathbf{A}_{r',c'} \neq 0$ and $\mathbf{A}_{r,c'} = 0$, no matter what $\mathbf{A}_{r,c'}$ is, the 2×2 submatrix

$$\begin{bmatrix} \mathbf{A}_{r,c'} & \mathbf{A}_{r,c} \\ \mathbf{A}_{r',c'} & \mathbf{A}_{r',c} \end{bmatrix} = \begin{bmatrix} 0 & \neq 0 \\ \neq 0 & ? \end{bmatrix}$$

is always non-singular (Denote by ? the entry which can be observed or unobserved, meaning that the specific value of the entry is unimportant for our purpose). So the algorithm answers correctly with probability at least $1 - 4e^{-c_0/4} > 2/3$.

Case (ii.3). $\mathbf{A}_{r,c'} \neq 0$. Instead of augmenting (r, c), we shall pick (r, c') to be our new base entry (1×1 matrix) and try to augment it to a 2×2 matrix. In this way, we have moved our base 1×1 matrix towards the upper-left corner. We can repeat the preceding arguments of different cases.



(a) Case (i). (b) Case (ii).

Figure 3.7: Finding an augmentation path $(d = 1)$, where the whole region is the $\mathcal{O}(d/\epsilon) \times \mathcal{O}(d/\epsilon)$ submatrix uniformly sampled from the original $n \times n$ matrix.

If Case (i) happens for (r, c'), we immediately have a 2×2 rank-2 submatrix and the algorithm answers correctly with a good probability. If Case (i) does not happen, we shall demonstrate that we can make further progress. Suppose that (r'', c'') is an augment of $(\{r\}, \{c'\})$ and $c'' \notin \mathcal{C}_s \cup \mathcal{C}_j$. We intend to look at the submatrix

$$\begin{bmatrix} \mathbf{A}_{r'',c'} & \mathbf{A}_{r'',c''} \\ \mathbf{A}_{r,c'} & \mathbf{A}_{r,c''} \end{bmatrix}$$

Here we cannot observe $\mathbf{A}_{r,c''}$. We know that $\mathbf{A}_{r'',c'}$ and $\mathbf{A}_{r'',c''}$ cannot be both 0, otherwise (r'', c'') would not be an augment for (r, c'). If $\mathbf{A}_{r'',c'} = 0$ and $\mathbf{A}_{r'',c''} \neq 0$, this 2×2 matrix is nonsingular regardless of the value of $\mathbf{A}_{r,c''}$ and the algorithm will answer correctly. If $\mathbf{A}_{r'',c'} \neq 0$, we can rebase our 1×1 base matrix to be (r'', c') and try to augment it. Since (r'', c') is above (r', c), we have again moved towards the upper-left corner.

Note that there are at most $\log(1/\eta)$ different augment patterns and each time we rebase, $\mathbf{A}_{*,*}$ moves from one $(\mathcal{R}_t, \mathcal{C}_t)$ to another for some t. Hence, after repeating the argument above at most $2\log(1/\eta)$ times, the algorithm is guaranteed to observe a 2×2 non-singular submatrix. Since the failure probability in each round is at most $4e^{-c_0/4}$, by union bound over $2\log(1/\eta)$ rounds, the overall failure probability is at most $8\log(1/\eta)e^{-c_0/4} \leq 1/3$, provided that $c_0 = \mathcal{O}(\log\log(\frac{1}{\eta}))$.

In summary, the overall probability is at least $2/3$ that the algorithm answers correctly in all cases by finding a submatrix of rank 2, when $\mathbf{A}$ is ϵ-far from being rank-1. □

Extension to general rank d. We now extend the analysis to the general rank d.

Theorem 34. *Let $\epsilon \leq 1/e$ and $d \geq 1$. For any matrix $\mathbf{A}$, the probability that Algorithm 7 fails is at most $1/\mathsf{poly}(d\log(\frac{1}{\epsilon}))$.*

Proof. If $\mathbf{A}$ is of rank at most d, then the algorithm will never make mistake, so we assume that $\mathbf{A}$ is ϵ-far from being rank d in the proof below.

The idea is that, we start with the base case of an empty matrix, and augment it to a full-rank $r \times r$ matrix in r rounds, where in each round we increase the dimension of the matrix by exactly one. Each round may contain several steps in which we move the intermediate $j \times j$ matrix ($j \le r$) towards the upper-left corner without augmenting it; here, moving the matrix towards the upper-left corner means changing $\mathbf{A}_{\mathcal{R},\mathcal{C}}$ to $\mathbf{A}_{\mathcal{R}',\mathcal{C}'}$, of the same rank, with $|\mathcal{R}'| = |\mathcal{R}| = |\mathcal{C}'| = |\mathcal{C}| = j$ and $\mathcal{R}' \preceq \mathcal{R}$ and $\mathcal{C}' \preceq \mathcal{C}$, where $\mathcal{R}' \preceq \mathcal{R}$ means that, suppose that $r'_1 < r'_2 < \cdots < r'_j$ are the (sorted) elements in $\mathcal{R}'$ and $r_1 < r_2 < \cdots < r_j$ are the (sorted) elements in $\mathcal{R}$, it holds that $r'_i \le r_i$ for all $1 \le i \le j$, and $\mathcal{C}' \preceq \mathcal{C}$ has a similar meaning.

The challenge is that those unobserved entries ?'s may propagate as we augment the submatrix in each round. Our goal is to prove that starting from a *structural* $(r-1) \times (r-1)$ full-rank submatrix which might have ?'s as its entries, no matter what values of *all* ?'s are, with the augment operator we either (1) make progress for $(r-1) \times (r-1)$ submatrix, or (2) obtain an $r \times r$ full-rank submatrix *with the same structure*. Let us first condition on the event that Lemma 47 holds true. Regarding the structure, we have the following claim.

Claim 1. *There exists a searching path for $r \times r$ full-rank submatrices with non-decreasing r which has the following lower triangular form modulo an elementary transformation*

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & \cdots & 0 & \neq 0 \\ 0 & 0 & \cdots & 0 & \cdots & \neq 0 & ? \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \neq 0 & \cdots & ? & ? \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & \neq 0 & \cdots & ? & \cdots & ? & ? \\ \neq 0 & ? & \cdots & ? & \cdots & ? & ? \end{bmatrix}, \tag{3.30}$$

where $\neq 0$ denotes the known entry which is non-zero, and ? denotes an entry which can be either observed or unobserved.

Proof of Claim 1. Without loss of generality, we assume that all ?'s are unobserved, which is the most challenging case; otherwise, the proof degenerates to the discussion of *central submatrix* in Case (iii) which we shall specify later. We prove the claim by induction. The base case $r = 1$ is true by Theorem 33. Suppose the claims holds for $r-1$. We now argue the correctness for r.

Let (p, q) be the augment. Denote the augment row by

$$\begin{bmatrix} y_1 & \cdots & y_b & \mathbf{A}_{p,q} & y_{b+2} & \cdots & y_r \end{bmatrix},$$

and the augment column by

$$\begin{bmatrix} x_1 & \cdots & x_a & \mathbf{A}_{p,q} & x_{a+2} & \cdots & x_r \end{bmatrix}^\top.$$

We now discuss three cases based on the relation between $a + b$ and r.

Case (i). $a + b = r - 1$ **($\mathbf{A}_{p,q}$ is on the antidiagonal of $r \times r$ submatrix).**

In this case, $y_{b+2}, \ldots, y_r$ and $x_{a+2}, \ldots, x_r$ are all ?'s. We argue that $x_1 = x_2 = \cdots = x_a = 0$ and $y_1 = y_2 = \cdots = y_b = 0$; otherwise, we can make progress. First consider y_i for $1 \le i \le b$.

If some $y_i \neq 0$, we can delete the $(r-i)$-th row in the $(r-1)\times(r-1)$ submatrix and insert the augment row (without the augment entry $\mathbf{A}_{p,q}$), which is above the deleted row. Thus we obtain a new $(r-1)\times(r-1)$ submatrix towards the upper-left corner, and furthermore, the new submatrix exhibits the structure in (3.30). The same argument applies for $x_1, x_2, ..., x_a$. Therefore, if no progress is made, it must hold that $x_1 = x_2 = ... = x_a = 0$ and $y_1 = y_2 = ... = y_b = 0$. In this case, $\mathbf{A}_{p,q} \neq 0$; otherwise, (p,q) is not an augment. Therefore we obtain an $r \times r$ full-rank matrix of the form (3.30).

Case (ii). $a + b < r - 1$ **($\mathbf{A}_{p,q}$ is above the antidiagonal of $r \times r$ submatrix)**.

In this case, $y_{r-a+1}, \ldots, y_r$ and $x_{r-b+1}, \ldots, x_r$ are all ?'s. Similarly to Case (i), we shall argue that $x_1 = \cdots = x_a = x_{a+2} = \cdots = x_{r-b} = 0$ and $y_1 = \cdots = y_b = y_{b+2} = \cdots = y_{r-a} = 0$; otherwise, we can make progress. To see this, consider first y_i for $1 \leq i \leq b$ and then for $b+2 \leq i \leq r-a$. If $y_i \neq 0$ for some $i \leq b$, we can delete the $(r-i)$-th row in the $(r-1)\times(r-1)$ submatrix and insert the augment row (without the augment entry $\mathbf{A}_{p,q}$), which is above the deleted row, and so we make progress. Now assume that $y_1 = \cdots = y_b = 0$. If $y_i \neq 0$ for some i such that $b+2 \leq i \leq r-a$, we can delete the $(r-i+1)$-st row in the $(r-1)\times(r-1)$ submatrix of the last step and insert the augment row (without the augment entry $\mathbf{A}_{p,q}$), which is above the deleted row. So we make progress towards the most upper left corner. The same argument applies to $x_1, \ldots, x_a, x_{a+2}, \ldots, x_{r-b}$. Therefore, $x_1 = \cdots = x_a = x_{a+2} = ... = x_{r-b} = 0$ and $y_1 = \cdots = y_b = y_{b+2} = \cdots = y_{r-a} = 0$. In this case, $\mathbf{A}_{p,q} \neq 0$; otherwise, (p,q) is not an augment since all possible choices of ?'s cannot make the $r \times r$ submatrix non-singular. By exchanging the $(a+1)$-st row and the $(r-b)$-th row of the $r \times r$ submatrix or exchanging the $(b+1)$-st column and the $(r-a)$-th column, we obtain an $r \times r$ submatrix of the form (3.30).

Case (iii). $a + b > r - 1$ **($\mathbf{A}_{p,q}$ is below the antidiagonal of $r \times r$ submatrix)**.

In this case, we argue that $x_i = y_j = 0$ for all $i \leq r-b-1$ and $j \leq r-a-1$; otherwise we can make progress as Cases (i) and (ii) for y_j. To see this, let us discuss from $j = 1$ to $r-a-1$. If $y_j \neq 0$ ($j = 1, 2, \ldots, r-a-1$), we can delete the $(r-j)$-th row in the $(r-1)\times(r-1)$ submatrix and insert the augment row (without the augment entry $\mathbf{A}_{p,q}$), which is above the deleted row. So we make progress. The same argument applies to $x_1, \ldots, x_{r-b-1}$. So $x_i = y_j = 0$ for all $i \leq r-b-1$ and $j \leq r-a-1$.

Given that there is only one non-zero entry in the first $r-b-1$ rows and the first $r-a-1$ columns of the $r \times r$ submatrix (i.e., the Laplace expansion of the determinant), we only need to focus on a minor corresponding to a $\min\{a,b\} \times \min\{a,b\}$ *central submatrix*, which decides whether the determinant of the $r \times r$ submatrix is zero and is fully-observed because the augment (p,q) is at the lower right corner of the central submatrix (see the red part in Eqn. (3.31)). Since it is fully-observed, the minor must be non-zero; otherwise, (p,q) cannot be an augment for all choices of ?'s. Therefore, we can do an elementary transformation to make the central submatrix a lower triangular matrix with non-zero antidiagonal entries. More importantly, such an elementary transformation also transforms the $r \times r$ matrix to a lower triangular matrix with non-zero antidiagonal entries, because all the entries to the left and above of the central matrix are 0's, and all the entries to the right and below of the central matrix are ?'s. Hence any elementary transformation keeps 0's and ?'s unchanged, and we obtain therefore an $r \times r$ submatrix of the form (3.30).

$$\begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 & \cdots & 0 & \cdots & 0 & \neq 0 \\ 0 & 0 & \cdots & \cdots & 0 & \cdots & 0 & \cdots & \neq 0 & ? \\ \vdots & \vdots & & & \vdots & & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & \neq 0 & \cdots & \text{known} & \cdots & ? & ? \\ \vdots & \vdots & & & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & \text{known} & \cdots & \text{augment } \mathbf{A}_{p,q} & \cdots & ? & ? \\ \vdots & \vdots & & & \vdots & & \vdots & & \vdots & \vdots \\ 0 & \neq 0 & \cdots & \cdots & ? & \cdots & ? & \cdots & ? & ? \\ \neq 0 & ? & \cdots & \cdots & ? & \cdots & ? & \cdots & ? & ? \end{bmatrix}. \tag{3.31}$$

□

Now we are ready to prove Theorem 34. Note that Lemma 47 works only for *fixed* $(\mathcal{R}, \mathcal{C})$. To make the lemma applicable "for all" $(\mathcal{R}, \mathcal{C})$ throughout the augmentation process, we shall take a union bound by choosing $|\mathcal{R}|$ and $|\mathcal{C}|$ large enough. Specifically, for each i, we divide $\mathcal{R}_i = \bigcup_{k=1}^{\ell} \mathcal{R}_i^{(k)}$ uniformly at random into $\ell = d + d\log(\frac{1}{\eta})$[6] even parts $\mathcal{R}_i^{(1)}, \mathcal{R}_i^{(2)}, \ldots, \mathcal{R}_i^{(d)}$, where each $|\mathcal{R}_i^{(k)}| = c[\log(d) + \log\log(\frac{1}{\eta})]2^i$, and divide $\mathcal{C}_i = \bigcup_{k=1}^{d} \mathcal{C}_i^{(k)}$ uniformly at random into ℓ even parts $\mathcal{C}_i^{(1)}, \mathcal{C}_i^{(2)}, \ldots, \mathcal{C}_i^{(\ell)}$, where each $|\mathcal{C}_i^{(k)}| = c[\log(d) + \log\log(\frac{1}{\eta})]/(2^i\eta)$ for every k. We note that $\{\mathcal{R}_i^{(k)}\}_k$ (and $\{\mathcal{C}_i^{(k)}\}_k$) are independent of each other. It follows that the event in Lemma 47 holds with probability at least $1 - \frac{1}{\text{poly}(d\log(1/\eta))}$. By a union bound over all $\ell^2 = \Theta(d^2\log^2(\frac{1}{\eta}))$ possible choices of $\{\mathcal{R}_i^{(k)}\} \times \{\mathcal{C}_i^{(k)}\}$ and Claim 1, with probability at least $1 - 1/\text{poly}(d\log(\frac{1}{\epsilon}))$, Algorithm 7 answers correctly, when $\mathbf{A}$ is ϵ-far from having rank d. □

A computationally efficient algorithm We now show how to implement Algorithm 7 efficiently, for which we only need to give a polynomial-time algorithm to solve the minimization problem (3.29) in Algorithm 8. We have the following theorem.

Theorem 35. *Algorithm 8 correctly solves the minimization problem* (3.29) *in* $\text{poly}(\frac{d}{\epsilon})$ *time.*

Proof. Without loss of generality, we may permute the rows and columns of $\mathbf{A}$ and assume that $\mathcal{R}_i = \{1, \ldots, |\mathcal{R}_i|\}$ and $\mathcal{C}_i = \{1, \ldots, |\mathcal{C}_i|\}$ for all $i \le \lceil \log \frac{1}{\eta} \rceil$. Our goal is to complete the submatrix $\mathbf{A}_{(\mathcal{R}_m \times \mathcal{C}_1)}$ such that its rank is minimal. Denote by $\mathcal{R}^{(i)}$ the set of sampled indices in the i-th column of $\mathbf{A}$. We start from an empty matrix $\mathbf{S} = [\,]$. We will extend $\mathbf{S}$ as we process the columns of $\mathbf{A}$ that are not in the block $(\mathcal{R}_k, \mathcal{C}_k)$ from left to right. We will maintain the following two invariants:

- The minimal rank of matrix completion is always equal to the number of columns of $\mathbf{S}$.

[6] In the number of parts $d + d\log(\frac{1}{\eta})$, the first term follows from the operation of augmenting 1×1 submatrix to $d \times d$. The second term follows from moving the submatrix towards the upper left corner (from the lower-right corner in the worst case).

Algorithm 8 Solving problem (3.29) in the polynomial time

Input: $\mathcal{R}_1, \ldots, \mathcal{R}_k$ and $\mathcal{C}_1, \ldots, \mathcal{C}_k$. Denote by $\mathcal{R}^{(i)}$ the set of sampled indices in the i-th column of $\mathbf{A}$, and let $\mathcal{R}_\perp^{(i)} = \mathcal{R}_k \setminus \mathcal{R}^{(i)}$.

Output: the solution r to the minimization problem (3.29).

$\mathbf{S} \leftarrow [\,]$ is an empty matrix.
$r \leftarrow 0$.
For $i = 1, \ldots, |\mathcal{C}_1|$
  If there exists $\mathbf{x}$ such that $\mathbf{S}_{\mathcal{R}^{(i)},:}\mathbf{x} = \mathbf{A}_{\mathcal{R}^{(i)},i}$
    $\mathbf{A}_{\mathcal{R}_\perp^{(i)},i} \leftarrow \mathbf{S}_{\mathcal{R}_\perp^{(i)},:}\mathbf{x}$.
  Else
    $\mathbf{A}_{\mathcal{R}_\perp^{(i)},i} \leftarrow \mathbf{1}$.
    $\mathbf{S} \leftarrow [\mathbf{S}, \mathbf{A}_{\mathcal{R}_k,i}]$.
    $r \leftarrow r + 1$.
  End If
End For

- After processing the i-th column of $\mathbf{A}$, the restricted column $\mathbf{A}_{\mathcal{R}^{(i)},i}$ is in the column space of $\mathbf{S}_{\mathcal{R}^{(i)},:}$.

Note that both invariants hold in the base case. For the i-th column $\mathbf{A}_{:,i}$ that we encounter, if $\mathbf{A}_{\mathcal{R}^{(i)},i}$ is in the column space of $\mathbf{S}_{\mathcal{R}^{(i)},:}$, then we use a linear combination on the first $|\mathcal{R}^{(i)}|$ coordinates of vectors given by the columns of $\mathbf{S}_{\mathcal{R}^{(i)},:}$ to extend $\mathbf{A}_{\mathcal{R}^{(i)},i}$ from a vector in $|\mathcal{R}^{(i)}|$ dimensions to a vector in $|\mathcal{R}_m|$ dimensions, and we do not change $\mathbf{S}$. Notice that the two invariants are preserved in this case.

Otherwise, $\mathbf{A}_{\mathcal{R}^{(i)},i}$ is not in the column space of $\mathbf{S}_{\mathcal{R}^{(i)},:}$. If $\mathbf{A}_{\mathcal{R}^{(i)},i}$ were in the column space of $\mathbf{A}_{\mathcal{R}^{(i)},1:(i-1)}$, we would, by the second invariant above, have that $\mathbf{A}_{\mathcal{R}^{(i)},i}$ is in the column space of $\mathbf{S}_{\mathcal{R}^{(i)},:}$, a contradiction. Therefore, $\mathbf{A}_{\mathcal{R}^{(i)},i}$ is not in the column space of $\mathbf{A}_{\mathcal{R}^{(i)},[i-1]}$. In this case, $\mathbf{A}_{:,i}$ must be linearly independent of all previous columns $\mathbf{A}_{:,[i-1]}$. We can thus append to $\mathbf{S}$ on the right the vector $[\mathbf{A}_{\mathcal{R}^{(i)},i}; \mathbf{1}]$ (The vector $\mathbf{1}$ can be replaced with any $(|\mathcal{R}_k| - |\mathcal{R}^{(i)}|)$-dimensional vector), which increases the size and rank of $\mathbf{S}$ by 1, and we maintain our two invariants.

The time complexity of Algorithm 8 is $\mathsf{poly}(\frac{d}{\epsilon})$. □

Lower bounds over finite fields in the sampling model According to Yao's minimax principle, it suffices to provide a distribution on $n \times n$ input matrices $\mathbf{A}$ for which any deterministic testing algorithm fails with significant probability over the choice of $\mathbf{A}$. Before proceeding, we first state a hardness result that we want to reduce from.

Lemma 48. *Let $G = (L \cup R, E)$ be a bipartite graph such that $|L| = |R| = n$ and $|E| < \gamma^2 d^2$ for $d \le n/\gamma$. Then Algorithm 9 returns a partition $E = E_1 \cup E_2 \cup \cdots \cup E_t$, where $t \le \gamma^2 d^2$ and $|E_i| \le \gamma d$ for all i.*

Proof. We first show that Algorithm 9 can be executed correctly, that is, whenever $E \neq \emptyset$ there always exists v such that $1 \le \mathsf{deg}(v) \le \gamma d$. We note that $1 \le \mathsf{deg}(v)$ is obvious because $E \neq \emptyset$. If all vertices with non-zero degree have degree at least γd, the total number of edges would be at

Algorithm 9 Decomposing edges E

Input: A bipartite graph $G = (L \cup R, E)$.
Output: Partition of $E = E_1 \cup \cdots \cup E_t$ and the set of pivot nodes $\{w_t\}$.

$t \leftarrow 0$.
While $E \neq \emptyset$
  Find v such that $1 \leq \mathsf{deg}(v) \leq \gamma d$.
  $t \to t + 1$.
  $E_t \leftarrow$ edges between v and all its neighbours.
  $w_t \leftarrow v$.
  $E \leftarrow E \setminus E_t$.
End While
return $E = E_1 \cup \cdots \cup E_t$ and $\{w_t\}$.

least $\gamma dn \geq d^2\gamma^2$, contradicting our assumption on the size of E. When the algorithm terminates, it is clear that each E_i generates at most γd edges and the E_i's are disjoint and so $t \leq \gamma^2 d^2$. $\square$

Lemma 49. *Suppose that there are t groups of (fixed) vectors $\{\mathbf{v}_1^{(k)}, \ldots, \mathbf{v}_{s_k}^{(k)}\}_{k\in[t]} \subset \mathbb{F}^d$ such that the vectors in each group are linearly independent (denoted by $\perp$). Let $\mathbf{w}_1, \ldots, \mathbf{w}_r$ be random vectors in $\mathbb{F}^d$ such that each $\mathbf{w}_i$ is chosen uniformly at random from some set $\mathcal{S}_i \subseteq \mathbb{F}^d$ with $|\mathcal{S}_i| \geq |\mathbb{F}|^{(1-\gamma)d}$. Let $s = \max_k s_k$. When $s + r \leq \gamma d$ for all k and $t \leq \gamma^2 d^2$, it holds that*

$$\Pr_{\mathbf{w}_1,\ldots,\mathbf{w}_r}\left\{\mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_r \textit{ for all } k \in [t]\right\} \geq 1 - \frac{\gamma^3 d^3}{|\mathbb{F}|^{(1-2\gamma)d}}.$$

Proof. For fixed $\mathbf{w}_1, \ldots, \mathbf{w}_{i-1}$ such that $\mathbf{v}_1^{(k)}, \ldots, \mathbf{v}_{s_k}^{(k)}, \mathbf{w}_1, \ldots, \mathbf{w}_{i-1}$ are linearly independent for all $k \in [t]$, the probability that $\mathbf{w}_i \in \mathcal{S}_i$ is linearly independent of $\mathbf{v}_1^{(k)}, \ldots, \mathbf{v}_{s_k}^{(k)}, \mathbf{w}_1, \ldots, \mathbf{w}_{i-1}$ for all $k \in [t]$ is at least $1 - \frac{t|\mathbb{F}|^{s_k+i-1}}{|\mathcal{S}_i|} \geq 1 - \frac{t|\mathbb{F}|^{s_k+i-1}}{|\mathbb{F}|^{(1-\gamma)d}} = 1 - \frac{t}{|\mathbb{F}|^{(1-\gamma)d-(s_k+i-1)}} \geq 1 - \frac{t}{|\mathbb{F}|^{(1-\gamma)d-(s+i-1)}}$.

Therefore, for all $k \in [t]$ with $t \le \gamma^2 d^2$, we have

$$\begin{aligned}
&\Pr_{\mathbf{w}_1,\dots,\mathbf{w}_r}\{\mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_r \text{ for all } k\} \\
&= \prod_{i=2}^{r} \Pr_{\mathbf{w}_i}\{\mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_i \text{ for all } k \mid \mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_{i-1} \text{ for all } k\} \\
&\quad \times \Pr_{\mathbf{w}_1}\{\mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \text{ for all } k\} \\
&\ge \prod_{i=1}^{r} \left(1 - \frac{t}{|\mathbb{F}|^{(1-\gamma)d-(s+i-1)}}\right) \\
&\ge \prod_{i=1}^{r} \left(1 - \frac{\gamma^2 d^2}{|\mathbb{F}|^{(1-2\gamma)d}}\right) \quad \text{(by } s+i-1 \le \gamma d \text{ and } t \le \gamma^2 d^2\text{)} \\
&\ge 1 - \frac{r\gamma^2 d^2}{|\mathbb{F}|^{(1-2\gamma)d}} \quad \text{(by } (1-x)^t \ge 1 - tx \text{ for } x \in (0,1)\text{)} \\
&\ge 1 - \frac{\gamma^3 d^3}{|\mathbb{F}|^{(1-2\gamma)d}}. \quad \text{(since } r \le \gamma d\text{).}
\end{aligned}$$

□

When $|\mathcal{S}_i| = |\mathbb{F}|^{d-d_i}$ for $d_i \le \gamma d$, it follows from Lemma 49 that the number of choices of the event

$$\begin{aligned}
&\left|\left\{(\mathbf{w}_1,\dots,\mathbf{w}_r) \in \mathcal{S}_1 \times \cdots \times \mathcal{S}_r : \mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_r \text{ for all } k\right\}\right| \\
&= \Pr_{\mathbf{w}_1,\dots,\mathbf{w}_r}\left\{\mathbf{v}_1^{(k)} \perp \cdots \perp \mathbf{v}_{s_k}^{(k)} \perp \mathbf{w}_1 \perp \cdots \perp \mathbf{w}_r \text{ for all } k\right\} \cdot \prod_{i=1}^{r} |\mathcal{S}_i| \\
&\ge \left(1 - \frac{\gamma^3 d^3}{|\mathbb{F}|^{(1-2\gamma)d}}\right) \cdot \prod_{i=1}^{r} |\mathcal{S}_i| \quad \text{(by Lemma 49)} \\
&= \left(1 - \frac{\gamma^3 d^3}{|\mathbb{F}|^{(1-2\gamma)d}}\right) |\mathbb{F}|^{rd-\sum_{i=1}^{r} d_i}. \quad \text{(recall that } |\mathcal{S}_i| = |\mathbb{F}|^{d-d_i}\text{)}
\end{aligned} \tag{3.32}$$

Based on this result, we have the following lemma.

Lemma 50. *Let $\mathbf{U}, \mathbf{V} \sim \mathcal{U}_{\mathbb{F}}(n,d)$, where $\mathcal{U}_{\mathbb{F}}(m,n)$ represents $m \times n$ i.i.d. uniform matrix over a finite field $\mathbb{F}$. Denote by $\mathcal{S}$ any subset of $[n] \times [n]$ such that $|\mathcal{S}| < \gamma^2 d^2$ for $\gamma \in (0, 1/4)$ and $d \le n/\gamma$. It holds that for any $\mathbf{x} \in \mathbb{F}^{|\mathcal{S}|}$,*

$$\Pr[(\mathbf{U}\mathbf{V}^T)|_{\mathcal{S}} = \mathbf{x}] - \frac{1}{|\mathbb{F}|^{|\mathcal{S}|}} \ge -\frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d+|\mathcal{S}|}}.$$

Proof. Consider a bipartite graph $G = (L \cup R, E)$ where $|L| = |R| = n$ and $(i,j) \in E$ if and only if $(i,j) \in \mathcal{S}$. We run Algorithm 9 on graph G. By Lemma 48, we obtain a sequence of edge sets $E_1 \dots, E_t$ with $w_1, \dots, w_t$ (called *pivot nodes*), such that

1. $\{E_i, \dots, E_t\}$ forms a partition of E;
2. $|E_i| \le \gamma d$ for all i.

Algorithm 10 Path for assigning subspace H_v and random vector $\mathbf{x}_v$ to each node v

Input: Bipartite graph $G = (L \cup R, E)$, partition $E = E_1 \cup \cdots \cup E_t$ and pivot nodes $\{w_t\}$ by Algorithm 9, observed entries $\mathbf{x}|_E$.

Output: An affine space H_v of vectors for every node v and a vector $\mathbf{x}_v \in H_v$ for every node v.

1: $H_v \leftarrow \mathbb{F}^d$ for all v.
2: Set all nodes v unassigned.
3: **For** $i \leftarrow t$ **down to** 1
4: Let $v_1^{(i)}, \ldots, v_{|E_i|}^{(i)}$ be the non-pivot nodes in E_i (i.e., the edges in E_i are $(w_i, v_j^{(i)})$).
5: **For** $j \leftarrow 1$ **to** $|E_i|$
6: **If** $v_j^{(i)}$ is unassigned
7: $W_{v_j^{(i)}} \leftarrow H_{v_j^{(i)}} \backslash \bigcup_{k \le i : v_j^{(i)} \neq w_k} \mathrm{span}\{\mathbf{x}_{v_1}^{(i)}, \ldots, \mathbf{x}_{v_{j-1}}^{(i)}, \text{previously assigned non-pivot nodes in } E_k\}$.
8: Choose $w_{v_j^{(i)}}$ uniformly at random from $H_{v_j^{(i)}}$.
9: **If** $w_{v_j^{(i)}} \notin W_{v_j^{(i)}}$
10: **abort**.
11: **End If**
12: Set $v_j^{(i)}$ to be assigned.
13: **End If**
14: **End For**
15: Let H_{w_i} be the solution set to the linear system (w.r.t. $\mathbf{x}_{w_i}$): $\mathbf{x}_{w_i}^\top [\mathbf{x}_{v_1}^{(i)}, \cdots, \mathbf{x}_{v_{|E_i|}}^{(i)}] = (\mathbf{x}|_{E_i})^\top$.
16: **End For**
17: Choose $\mathbf{x}_{w_s}$ uniformly from H_{w_s} of dimension $d - |E_s|$ for all $s \in \mathcal{S}_0 = \{p \in [t] \mid w_p \text{ is unassigned}\}$.
18: **return** $\{H_v\}$ and $\{\mathbf{x}_v\}$.

Since there is a one-by-one correspondence between the edges and the entries in $\mathcal{S}$, we will not distinguish edges and entries in the rest of the proof.

We associate each node v of G with an affine space $H_v \subseteq \mathbb{F}^d$ and a random vector $\mathbf{x}_v \in H_v$ as in Algorithm 10. Basically, Algorithm 10 first assigns the non-pivot nodes (to determine the affine subspace H_{w_i}) from the E_t down to the E_1), and in the end assigns all unassigned pivot nodes.

In the following argument, we number the for-loop iterations in Algorithm 10 backwards, i.e., the for-loop starts with the t-th iteration and goes down to the first iteration. In the i-th iteration, let r_i denote the number of nodes $v_j^{(i)}$ that are unassigned at the runtime of Line 6 and let $\#\mathcal{E}_i$ denote the number of good choices (which do not trigger abortion) of Step 8 over all r_i nodes to be assigned. Let $\#\mathcal{G}$ be the number of possible choices of Step 17 of Algorithm 10 and $s_0 = |\mathcal{S}_0|$ be the number of assigned pivot nodes by Step 17. Note that by the construction of Algorithm 9, the non-pivot nodes of E_i cannot be the pivot nodes of E_j for $j < i$. So Algorithm 10, if terminated successfully, can find an assignment such that $(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}$. We now lower bound the success probability.

Let $d_j^{(i)} = d - \dim(H_{v_j^{(i)}})$, which is either 0 or $|E_k|$ for some $k > i$. For any given realization $\mathbf{x}$, we have the following:

$$
\begin{aligned}
&\Pr\{(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}\} \\
&\geq \frac{\#\mathcal{E}_t \cdot \#\mathcal{E}_{t-1} \cdots \#\mathcal{E}_1}{|\mathbb{F}|^{d(r_t+\cdots+r_1)}} \frac{\#\mathcal{G}}{|\mathbb{F}|^{ds_0}} \quad \text{(by rule of product and definition of } \#\mathcal{E}_i\text{)} \\
&\geq \prod_{i=1}^{t} \frac{1}{|\mathbb{F}|^{d_1^{(i)}+\cdots+d_{r_i}^{(i)}}} \left(1 - \frac{\gamma^3 d^3}{|\mathbb{F}|^{(1-2\gamma)d}}\right)^t \cdot \frac{\#\mathcal{G}}{|\mathbb{F}|^{ds_0}} \quad \text{(by Eqn. (3.32))} \\
&\geq \frac{1}{|\mathbb{F}|^{\sum_{i=1}^{t}\sum_{j=1}^{r_i} d_j^{(i)}}} \left(1 - \frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d}}\right) \cdot \frac{\#\mathcal{G}}{|\mathbb{F}|^{d\times s_0}} \quad \text{(by } (1-x)^t \geq 1 - tx \text{ for } x \in (0,1) \text{ and } t \leq \gamma^2 d^2\text{)} \\
&= \frac{1}{|\mathbb{F}|^{\sum_{i=1}^{t}\sum_{j=1}^{r_i} d_j^{(i)}}} \left(1 - \frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d}}\right) \cdot \frac{1}{|\mathbb{F}|^{\sum_{s\in\mathcal{S}_0}|E_s|}} \quad \text{(by definition of } \#\mathcal{G}\text{)} \\
&\geq \frac{1}{|\mathbb{F}|^{|E_1|+\cdots+|E_t|}} \left(1 - \frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d}}\right),
\end{aligned}
$$

where the last inequality holds because $|E_1| + \cdots + |E_t| = \sum_{j=1}^{t}\sum_{i=1}^{r_j} d_i^{(j)} + \sum_{s\in\mathcal{S}_0}|E_s|$ as every pivot and non-pivot node must be assigned exactly once by Algorithm 10 upon successful termination. (Recall that $d_i^{(j)}$ is either equal to 0 when $v_j^{(i)}$ is non-pivotal, or equal to $|E_k|$ when $v_j^{(i)} = w_k$.) □

Denote by $\mathcal{S} \subset [n] \times [n]$ a set of indices of an $n \times n$ matrix. For any distribution $\mathcal{L}$ over $\mathbb{F}^{n\times n}$, define $\mathcal{L}(\mathcal{S})$ on $\mathbb{F}^{|\mathcal{S}|}$ as the marginal distribution of $\mathcal{L}$ on the entries of $\mathcal{S}$, namely,

$$
(\mathbf{X}_{p_1,q_1}, \mathbf{X}_{p_2,q_2}, \ldots, \mathbf{X}_{p_{|\mathcal{S}|},q_{|\mathcal{S}|}}) \sim \mathcal{L}(\mathcal{S}), \qquad \mathbf{X} \sim \mathcal{L}.
$$

Now we are ready to show a lower bound of robust testing problem over any finite field.

Theorem 36. *Suppose that $\mathbb{F}$ is a finite field and $\gamma \in (0, 1/4)$ is an absolute constant. Let $\mathbf{U}, \mathbf{V} \sim \mathcal{U}_{\mathbb{F}}(n, d)$ and $\mathbf{W} \sim \mathcal{U}_{\mathbb{F}}(n, n)$, where $\mathcal{U}_{\mathbb{F}}(m, n)$ represents $m \times n$ i.i.d. uniform matrix over a finite field $\mathbb{F}$. Consider two distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ over $\mathbb{F}^{n\times n}$ defined by $\mathbf{U}\mathbf{V}^\top$ and $\mathbf{W}$, respectively. Let $\mathcal{S} \subset [n] \times [n]$. When $|\mathcal{S}| < \gamma^2 d^2$, it holds that*

$$
d_{TV}(\mathcal{L}_1(\mathcal{S}), \mathcal{L}_2(\mathcal{S})) \leq C d^5 |\mathbb{F}|^{-cd},
$$

where $C, c > 0$ are constants depending on γ, and $d_{TV}(\cdot, \cdot)$ represents the total variation distance between two distributions.

Proof. Let

$$
\mathcal{X} = \left\{ \mathbf{x} \in \mathbb{F}^{|\mathcal{S}|} \,\middle|\, \Pr\left[(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}\right] < \frac{1}{|\mathbb{F}|^{|\mathcal{S}|}} \right\}.
$$

It follows from the definition of total variation distance that

$$
d_{TV}(\mathcal{L}_1(\mathcal{S}), \mathcal{L}_2(\mathcal{S})) = \sum_{\mathbf{x}\in\mathcal{X}} \left[\frac{1}{|\mathbb{F}|^{|\mathcal{S}|}} - \Pr[(\mathbf{U}\mathbf{V}^\top)|_{\mathcal{S}} = \mathbf{x}]\right] \leq \sum_{\mathbf{x}\in\mathcal{X}} \frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d}} \frac{1}{|\mathbb{F}|^{|\mathcal{S}|}} \leq \frac{\gamma^5 d^5}{|\mathbb{F}|^{(1-2\gamma)d}},
$$

where the last inequality holds since $|\mathcal{X}| \leq |\mathbb{F}|^{|\mathcal{S}|}$. □

Based on the above theorem, we have the following lower bound for the rank testing problem over finite field.

Theorem 37. *Let $d \leq \sqrt{\epsilon}n$. Any non-adaptive algorithm for Problem 1 over any finite field $\mathbb{F}$ requires $\Omega(d^2/\epsilon)$ queries.*

Proof. We first show that for constant ϵ, any non-adaptive algorithm for Problem 1 over finite field $\mathbb{F}$ requires $\Omega(d^2)$ queries. Note that $\mathbf{W} \sim \mathcal{U}_{\mathbb{F}}(n, n)$ is ϵ-far from having rank less than d. It follows immediately from the preceding theorem that any algorithm which solves the matrix rank testing problem over a finite field must read $\Omega(d^2)$ entries; otherwise when d is large enough, it will hold that $d_{TV}(\mathcal{L}_1(\mathcal{S}), \mathcal{L}_2(\mathcal{S})) < 1/4$, contradicting the correctness of the algorithm on distinguishing $\mathcal{L}_1$ from $\mathcal{L}_2$.

We now prove the case for arbitrary ϵ. Denote by $\mathbf{A}$ and $\mathbf{B}$ the two hard instances in Theorem 36. We construct two hard instances $\mathbf{C}$ and $\mathbf{D}$ by uniformly at random planting the above-mentioned hard instances $\mathbf{A}$ and $\mathbf{B}$ of dimension $\sqrt{\epsilon}n \times \sqrt{\epsilon}n$, respectively, and padding zeros everywhere else. Note that $\mathbf{D}$ being ϵ-far from rank d is equivalent to $\mathbf{B}$ being constant-far from rank d. Suppose that we can request cd^2/ϵ queries with a small absolute constant c to distinguish the ranks of the hard instances $\mathbf{C}$ and $\mathbf{D}$, then in expectation (and with high probability by a Markov bound) we can request cd^2 queries of the hard instances $\mathbf{A}$ and $\mathbf{B}$ to distinguish their ranks, which leads to a contradiction. □

Lower bounds over real field under the sampling model The *rigidity* of a matrix $\mathbf{A}$ over a field $\mathbf{F}$, denoted by $\mathcal{R}^{\mathbf{F}}_{\mathbf{A}}(r)$, is the least number of entries of $\mathbf{A}$ that must be changed in order to reduce the rank of $\mathbf{A}$ to a value at most r: $\mathcal{R}^{\mathbf{F}}_{\mathbf{A}}(r) := \min\{\|\mathbf{C}\|_0 \mid \mathsf{rank}_{\mathbf{F}}(\mathbf{A}+\mathbf{C}) \leq r\}$. We first cite the following lemma and theorem.

Lemma 51 (Matrix rigidity, Theorem 6.4, [229])**.** *The real $n \times n$ i.i.d. Gaussian matrix $\mathbf{G}$ is of rigidity $\mathcal{R}^{\mathbb{R}}_{\mathbf{G}}(r) = \Omega((n-r)^2)$ with probability* 1.

Theorem 38 (Theorem 3.5, [152])**.** *Let $\mathbf{U}, \mathbf{V} \sim \mathcal{G}(n, d)$ and $\mathbf{G} \sim \mathcal{G}(n, n)$. Consider two distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ over $\mathbb{R}^{n\times n}$ defined by $\mathbf{U}\mathbf{V}^\top$ and $\mathbf{U}\mathbf{V}^\top + n^{-14}\mathbf{G}$, respectively. Let $\mathcal{S} \subset [n] \times [n]$. Whenever $|\mathcal{S}| \leq d^2$, it holds that*

$$d_{TV}(\mathcal{L}_1(\mathcal{S}), \mathcal{L}_2(\mathcal{S})) \leq C|S|(n^{-2} + dc^d),$$

where $C > 0$ and $0 < c < 1$ are absolute constants.

Now we are ready to prove the sample complexity lower bound of rank testing over the reals in the sampling model.

Theorem 39. *Let $d \leq \sqrt{\epsilon}n$. Any non-adaptive algorithm for Problem 1 over $\mathbb{R}$ requires $\Omega(d^2/\epsilon)$ queries.*

Proof. We first show that for constant ϵ, any non-adaptive algorithm for Problem 1 over $\mathbb{R}$ requires $\Omega(d^2)$ queries. Note that Theorem 38 provides two hard instances for distinguishing a rank-d matrix (of the form $\mathbf{A} = \mathbf{U}\mathbf{V}^\top$) from a rank-$n$ matrix (of the form $\mathbf{B} = \mathbf{U}\mathbf{V}^\top + n^{-14}\mathbf{G}$), where $\mathbf{U}, \mathbf{V} \sim \mathcal{G}(n, d)$ and $\mathbf{G} \sim \mathcal{G}(n, n)$. For our purpose, we only need to show that the rank-n matrix $\mathbf{B} = \mathbf{U}\mathbf{V}^\top + n^{-14}\mathbf{G}$ has rigidity $\mathcal{R}^{\mathbb{R}}_{\mathbf{B}}(d) = \Omega(n^2)$. Denote by $\mathsf{rank}_\ell(\mathbf{B}) = \min_{\|\mathbf{S}\|_0=\ell} \mathsf{rank}(\mathbf{B}+\mathbf{S})$.

We note that

$$\begin{aligned} d &\geq \mathsf{rank}_{\mathcal{R}^{\mathbb{R}}_{\mathbf{B}}(d)}(\mathbf{B}) \\ &= \min_{\|\mathbf{S}\|_0=\mathcal{R}^{\mathbb{R}}_{\mathbf{B}}(d)} \mathsf{rank}(\mathbf{U}\mathbf{V}^\top + n^{-14}\mathbf{G} + \mathbf{S}) \\ &\geq \min_{\|\mathbf{S}\|_0=\mathcal{R}^{\mathbb{R}}_{\mathbf{B}}(d)} \mathsf{rank}(n^{-14}\mathbf{G} + \mathbf{S}) - \mathsf{rank}(\mathbf{U}\mathbf{V}^\top) \\ &\geq \min_{\|\mathbf{S}\|_0=\mathcal{R}^{\mathbb{R}}_{\mathbf{B}}(d)} \mathsf{rank}(n^{-14}\mathbf{G} + \mathbf{S}) - d. \end{aligned}$$

Therefore, $\min_{\|\mathbf{S}\|_0=\mathcal{R}^{\mathbb{R}}_{\mathbf{B}}(d)} \mathsf{rank}(n^{-14}\mathbf{G} + \mathbf{S}) \leq 2d$, i.e., $\mathcal{R}^{\mathbb{R}}_{\mathbf{B}}(d) \geq \mathcal{R}^{\mathbb{R}}_{n^{-14}\mathbf{G}}(2d)$. By Lemma 51, we have $\mathcal{R}^{\mathbb{R}}_{n^{-14}\mathbf{G}}(2d) = \Omega(n-2d)^2 = \Omega(n^2)$. So $\mathcal{R}^{\mathbb{R}}_{\mathbf{B}}(d) = \Omega(n^2)$.

We now prove the case for arbitrary ϵ. We construct two hard instances $\mathbf{C}$ and $\mathbf{D}$ by uniformly at random planting the above-mentioned hard instances $\mathbf{A}$ and $\mathbf{B}$ of dimension $\sqrt{\epsilon}n \times \sqrt{\epsilon}n$, respectively, and padding zeros everywhere else. Note that $\mathbf{D}$ being ϵ-far from rank d is equivalent to $\mathbf{B}$ being constant-far from rank d. Suppose that we can request cd^2/ϵ queries with a small absolute constant c to distinguish the ranks of the hard instances $\mathbf{C}$ and $\mathbf{D}$, then in expectation (and with high probability by a Markov bound) we can request cd^2 queries of the hard instances $\mathbf{A}$ and $\mathbf{B}$ to distinguish their ranks, which leads to a contradiction. □

Proofs of Theorem 29

In this section, we provide a lower bound for the rank testing problem in the sensing model over any finite field $\mathbb{F}$. The sensing problem can query the underlying matrix $\mathbf{A}$ in the form of $\langle \mathbf{A}, \mathbf{X}_i \rangle$ for any sequence of (randomized or deterministic) sensing matrices $\{\mathbf{X}_i\}$. The algorithms for querying entries of $\mathbf{A}$ are a special case of matrix sensing problem if we set $\mathbf{X}_i = \mathbf{e}_p \mathbf{e}_q^\top$ for some (p, q). The problem can be stated more formally as follows:

Problem 2 (Rank Testing with Parameter (n, d, ϵ) in the Sensing Model)**.** *Given a field $\mathbb{F}$ and a matrix $\mathbf{A} \in \mathbb{F}^{n\times n}$ which has one of promised properties:*

- H0. *$\mathbf{A}$ has rank at most d;*
- H1. *$\mathbf{A}$ is ϵ-far from having rank at most d, meaning that $\mathbf{A}$ requires changing at least an ϵ-fraction of its entries to have rank at most d.*

The problem is to design a property testing algorithm that outputs H0 *with probability* 1 *if $\mathbf{A} \in$* H0*, and output* H1 *with probability at least 0.99 if $\mathbf{A} \in$* H1*, with the least number of queries of the form $\langle \mathbf{A}, \mathbf{X}_i \rangle$, where $\{\mathbf{X}_i\}$ is a sequence of sensing matrices.*

Definition 6 (Ruzsa-Szemerédi Graph)**.** *A graph G is an (r, t)-*Ruzsa-Szemerédi graph (RS graph for short)*, if and only if the set of edges of G consists of t pairwise disjoint induced matchings $M_1, \ldots, M_t$, each of which is of size r.*

Definition 7 (Boolean Hidden Hypermatching, $\mathsf{BHH}_{n,p}$)**.** *The* Boolean Hidden Hypermatching *problem is a one-way communication problem where Alice is given a boolean vector $\mathbf{x} \in \{0,1\}^n$ such that $n = 2kp$ for some integer $k \geq 1$, and Bob is given a boolean vector $\mathbf{w}$ of length n/p and a perfect p-hypermatching $\mathcal{M}$ on n vertices such that each hyperedge contains p vertices. Denote by $\mathcal{M}\mathbf{x}$ the length n/p boolean vector $(\bigoplus_{1\leq i\leq p} \mathbf{x}_{\mathcal{M}_{1,i}}, \bigoplus_{1\leq i\leq p} \mathbf{x}_{\mathcal{M}_{2,i}}, \ldots, \bigoplus_{1\leq i\leq p} \mathbf{x}_{\mathcal{M}_{n/p,i}})$ where $\{\mathcal{M}_{1,1}, \ldots, \mathcal{M}_{1,p}\}, \ldots, \{\mathcal{M}_{n/p,1}, \ldots, \mathcal{M}_{n/p,p}\}$ are the hyperedges of $\mathcal{M}$. It is promised that*

either $\mathcal{M}\mathbf{x} = \mathbf{w}$ *or* $\mathcal{M}\mathbf{x} = \overline{\mathbf{w}}$. *The goal of the problem is for Bob to output* **YES** *when* $\mathcal{M}\mathbf{x} = \mathbf{w}$ *and* **NO** *when* $\mathcal{M}\mathbf{x} = \overline{\mathbf{w}}$ *($\oplus$ stands for addition modulo 2).*

For our purpose, it is more convenient to focus on a special case of Boolean Hidden Hypermatching problem, namely, $\mathsf{BHH}^0_{n,p}$ where the vector $\mathbf{w} = \mathbf{0}^{n/p}$ (p is an even integer) and Bob's task is to output **YES** if $\mathcal{M}\mathbf{x} = \mathbf{0}^{n/p}$ and output **NO** if $\mathcal{M}\mathbf{x} = \mathbf{1}^{n/p}$. It is known that we can reduce any instance of $\mathsf{BHH}_{n,p}$ to an instance of $\mathsf{BHH}^0_{2n,p}$ deterministically without any communication between Alice and Bob [54, 149, 233], by the following reduction.

Reduction from $\mathsf{BHH}_{n,p}$ to $\mathsf{BHH}^0_{2n,p}$. We reduce any instance of $\mathsf{BHH}_{n,p}$ to an instance of $\mathsf{BHH}^0_{2n,p}$ ($n = 2kp$ for some integer k). Let $\mathcal{M}$ be a perfect p-hypermatching and $\mathbf{x} \in \{0,1\}^n$ in $\mathsf{BHH}_{n,p}$. Denote by $\mathbf{x}' = [\mathbf{x}; \overline{\mathbf{x}}]$ the concatenation of $\mathbf{x}$ and $\overline{\mathbf{x}}$, where $\overline{\mathbf{x}}$ is the bitwise negation of $\mathbf{x}$. Let $\mathcal{M}'$ be the p-hypermatching in $\mathsf{BHH}^0_{2n,p}$. Denote by $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\} \in \mathcal{M}$ the l-th hyperedge of $\mathcal{M}$ ($l \in [n/p]$). We add two hyperedges to $\mathcal{M}'$ as follows. If $\mathbf{w}_l = 0$, we add $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\}$ and $\{\overline{\mathbf{x}_1}, \overline{\mathbf{x}_2}, \ldots, \overline{\mathbf{x}_p}\}$ to $\mathcal{M}'$; Otherwise, we add $\{\overline{\mathbf{x}_1}, \mathbf{x}_2, \ldots, \mathbf{x}_p\}$ and $\{\mathbf{x}_1, \overline{\mathbf{x}_2}, \ldots, \overline{\mathbf{x}_p}\}$ to $\mathcal{M}'$. Note that we flip an even number of bits when $\mathbf{w}_l = 0$ and an odd number of bits when $\mathbf{w}_l = 1$. This does not change the parity of each set as p is even. Thus $\mathcal{M}\mathbf{x} = \mathbf{w}$ implies $\mathcal{M}'\mathbf{x}' = \mathbf{0}^{2n/p}$, and $\mathcal{M}\mathbf{x} = \overline{\mathbf{w}}$ implies $\mathcal{M}'\mathbf{x}' = \mathbf{1}^{2n/p}$.

Previous papers [12, 54, 82] used the $\mathsf{BHH}^0_{n,p}$ problem to prove lower bounds for estimating matching size in the data stream: given an instance $(\mathbf{x}, \mathcal{M})$ in $\mathsf{BHH}^0_{n,p}$ (Denote by $\mathcal{D}_{\mathsf{BHH}}$ the hard distribution of $\mathsf{BHH}^0_{n,p}$), we create a graph $G(V \cup W, E)$ with $|V| = |W| = n$ via the following algorithm.

Algorithm 11 Reduction from $\mathsf{BHH}^0_{n,p}$ to the problem of estimating matching size in the data stream

---

Input: An instance from $\mathsf{BHH}^0_{n,p}$.
Output: A graph $G = (V \cup W, E)$.

1: For any $\mathbf{x}_i = 1$, Alice adds an edge between v_i and w_i to E.
2: Bob adds to E a clique between the vertices w_i that belongs to the same hyperedge $\mathbf{e}$ in the p-hypermatching $\mathcal{M}$.

We shall use the graph created by Algorithm 11 to build a hard distribution for our *rank testing* problem. The following claim guarantees the correctness of this reduction from $\mathsf{BHH}^0_{n,p}$ to the problem of estimating matrix rank in a data stream.

Lemma 52. *Let $G(V \cup W, E)$ be the graph derived from an instance $(\mathbf{x}, \mathcal{M})$ of $\mathsf{BHH}^0_{n,p}$ (for even integers p and n) with the property that $\|\mathbf{x}\|_0 = n/2$ (see Algorithm 11). Denote by $\mathbf{A}$ the $2n \times 2n$ adjacency matrix of G. Then with probability at least $1 - e^{-n/p^4}$, we have*

- *if $\mathcal{M}\mathbf{x} = \mathbf{0}^{n/p}$ (i.e., YES case), then $\mathsf{rank}(\mathbf{A}) \geq \frac{3n}{2} - \frac{n}{2p^2}$;*
- *if $\mathcal{M}\mathbf{x} = \mathbf{1}^{n/p}$ (i.e., NO case), then $\mathsf{rank}(\mathbf{A}) \leq \frac{3n}{2} - \frac{3n}{2p^2}$.*

Proof. According to Algorithm 11, the graph consists of n vertices $v_1, \ldots, v_n$ and n/p cliques, together with edges which connect v_i's with the cliques according to $\mathbf{x} \in \{0,1\}^n$. We call these latter edges 'tentacles'.

Let $\mathbf{A}$ be the adjacency matrix of G where both the rows and columns are indexed by the nodes in G. The diagonals of $\mathbf{A}$ are all zeros. For each pair w, u of clique nodes in G, we have $\mathbf{A}_{w,u} = 1$. For each 'tentacle' pair (v, w), $\mathbf{A}_{v,w} = 1$. All other entries of $\mathbf{A}$ are zeros. Then $\mathbf{A}$ is an $n \times n$ block diagonal matrix, where each block $\mathbf{A}_{q_i}$ ($\mathbf{A}_{q_i}$ represents the block with q_i 'tentacles') is of the following form modulo permutations of rows and columns (The red rows and columns represent 'tentacles'):

$$\mathbf{A}_{q_i} = \begin{bmatrix} 0 & 1 & \cdots & 1 & 1 & 1 & 0 & \cdots \\ 1 & 0 & \cdots & 1 & 1 & 0 & 1 & \cdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \\ 1 & 1 & \cdots & 0 & 1 & 0 & 0 & \cdots \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 & \cdots \\ 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

According to the reduction from $\mathsf{BHH}_{n/2,p}$ to $\mathsf{BHH}^0_{n,p}$, the hypermatching in the hard distribution of $\mathsf{BHH}^0_{n,p}$ can be divided into $n/(2p)$ groups. Each group consists of two hyperedges such that the sum of the number of 'tentacles' connecting to these two hyperedges is p for every group, i.e., $(q_i, p - q_i)$ where q_i is the number of 'tentacles' connecting to one of hyperedges, which is either even (YES case) or odd (NO case) according to the promise. Moreover, the q_i's are independent across the n/p groups, because we can process each group one by one and after processing each group, the number of remaining 'tentacles' decreases by p.

Let $r_{q_i} = \mathsf{rank}(\mathbf{A}_{q_i})$. Denote by $A = \mathbb{E}_{\mathsf{YES}}(r_{q_i} + r_{p-q_i})$ and $B = \mathbb{E}_{\mathsf{NO}}(r_{q_i} + r_{p-q_i})$, where A and B will be calculated later. Summing up $n/(2p)$ *independent* groups and by the Chernoff bound, with probability at least $1 - e^{-\delta^2 \frac{n}{2p} A/2}$ and $1 - e^{-\delta^2 \frac{n}{2p} B/3}$, respectively, $\mathsf{rank}(\mathbf{A}) \geq (1-\delta)\frac{n}{2p}A$ in the even case and $\mathsf{rank}(\mathbf{A}) \leq (1+\delta)\frac{n}{2p}B$ in the odd case, where $\delta > 0$ is an absolute constant. We note that $A = 3p$ and $B = 3p - 4/p$. Therefore, $\mathsf{rank}(\mathbf{A}) \geq (1-\delta)3n/2$ in the even case and $\mathsf{rank}(\mathbf{A}) \leq (1+\delta)(3n/2 - 2n/p^2)$ in the odd case. Choosing $\delta = \frac{1}{3p^2}$ finishes the proof. □

In the following we shall set $\epsilon = \Theta(1/\log n)$ and $p = \Theta(\log n)$. Denote by $\mathsf{Matching}_{n,k,\epsilon}$ the *k-player simultaneous communication* problem of estimating the size of maximum matching up to a factor of $(1 \pm \epsilon)$, where the edges of an n-vertex input graph are partitioned across the k players and the referee. For our purpose, we reduce from the problem of $\mathsf{Matching}_{n,k,\epsilon}$ to our problem of *rank testing*. We use the hard distribution $\mathcal{D}_{\mathsf{M}}$ in Algorithm 12 for $\mathsf{Matching}_{n,k,\epsilon}$. Notice that the hard instance of $\mathsf{BHH}^0_{r,p}$ in Step 2 is reduced from that of $\mathsf{BHH}_{r/2,p}$ as we did before in this section.

Claim 2. *Let I_{BHH} be the embedded $\mathsf{BHH}^0_{r,p}$ instance $(\mathbf{x}^{(i)}, \mathcal{M})$ in Algorithm 12. The adjacency matrix $\mathbf{A} \in \mathbb{F}^{n \times n}$ of the graph that is drawn from distribution $\mathcal{D}_{\mathsf{M}}$ (Algorithm 12) obeys*

1. *If I_{BHH} is a* YES *instance, then $\mathsf{rank}(\mathbf{A}) \geq k(\frac{3r}{2} - \frac{r}{2p^2})$;*
2. *If I_{BHH} is a* NO *instance, then $\mathsf{rank}(\mathbf{A}) \leq k(\frac{3r}{2} - \frac{3r}{2p^2}) + N - 2r$,*

with probability at least $1 - ke^{-n/p^4}$.

Algorithm 12 A construction of a hard distribution $\mathcal{D}_{\mathsf{M}}$ for $\mathsf{Matching}_{n,k,\epsilon}$

Input: $r = N^{1-o(1)}$, $t = \frac{\binom{N}{2}-o(N^2)}{r}$, $k = \frac{N}{\epsilon r}$, $n = N + 2r(k-1)$, and $p = \lfloor \frac{1}{8\epsilon} \rfloor$.

1: Fix an (r,t)-RS graph $G^{\mathcal{R}}$ on N vertices.
2: Pick $j^* \in [t]$ uniformly at random and draw a $\mathsf{BHH}^0_{r,p}$ instance $(\mathbf{x}^{(j^*)}, \mathcal{M})$ from the distribution $\mathcal{D}_{\mathsf{BHH}}$.
3: **For** each player $P^{(i)}$ independently
4: (a) Let G_i be the input graph of $P^{(i)}$, initialized by a copy of $G^{\mathcal{R}}$ with vertices $V_i = [N]$.
5: (b) Let V_i^* be the set of vertices matched in the j^*-th induced matching of G_i. Change the induced matching $M^{\mathcal{R}}_{j^*}$ of G_i to $M_{j^*} := M^{\mathcal{R}}_{j^*}|_{\mathbf{x}^{(j^*)}}$.
6: (c) For any $j \in [t]\backslash\{j^*\}$, draw a vector $\mathbf{x}^{(i,j)} \in \{0,1\}^r$ from the distribution $\mathcal{D}_{\mathsf{BHH}}$ for $\mathsf{BHH}^0_{r,p}$, and change the induced matching $M^{\mathcal{R}}_j$ of G_i to $M_j := M^{\mathcal{R}}_j|_{\mathbf{x}^{(i,j)}}$.
7: (d) Create the family of p-cliques of $\mathcal{M}$ on the vertices $R(M^{\mathcal{R}}_{j^*})$, and give the edges of the p-clique family to the referee.
8: **End For**
9: Choose a random permutation σ of $[n]$. For each player $P^{(i)}$, relabel v to $\sigma(j)$ for each vertex v in $V_i\backslash V_i^*$ with label $j \in [N-2r]$. Enumerate the vertices in V_i^* from the one with the smallest label to the one with the largest label, and relabel the j-th vertex to $\sigma(N + (i-2)2r + j)$. Finally, let the vertices with the same label correspond to the same vertex.

Proof. Note that by construction, the adjacency matrix of the graph drawn from $\mathcal{D}_{\mathsf{M}}$ is a k-block-diagonal matrix together with some 'junk' (area of size $(N-2r) \times n$ union $n \times (N-2r)$) outside the block area such that each block is an independent sample of the matrix $\mathbf{A}$ in Lemma 52. The claim then is a straightforward result of Lemma 52. □

Reduction from $\mathsf{Matching}_{n,k,\epsilon}$ to problem 2. Given a hard graph instance G of $\mathsf{Matching}_{n,k,\epsilon}$, we can estimate the maximum matching size of G by testing the rank of the adjacency matrix $\mathbf{A}_G$ of G: If we can distinguish out $\mathsf{rank}(\mathbf{A}_G) \geq k(\frac{3r}{2} - \frac{r}{2p^2})$, we ouput that the matching size is strictly larger than $\frac{3N}{\epsilon}$; If we can distinguish out $\mathsf{rank}(\mathbf{A}_G) \leq k(\frac{3r}{2} - \frac{3r}{2p^2}) + N - 2r$, we output that the matching size is smaller than $\frac{3N}{\epsilon} - 3N$. The correctness for the reduction follows from Claim 2, the construction that the hard distributions of $\mathsf{Matching}_{n,k,\epsilon}$ and Problem 2 are derived from the same graph, and the fact that the matching size is strictly larger than $\frac{3N}{\epsilon}$ when I_{BHH} is a YES instance and is smaller than $\frac{3N}{\epsilon} - 3N$ when I_{BHH} is a NO instance (see Claim 6.3, [12]).

The hardness of $\mathsf{Matching}_{n,k,\epsilon}$ by the construction in Algorithm 12 was proved in [12].

Theorem 40 (Theorem 10, [12])**.** *For any sufficiently large n and sufficiently small $\epsilon < \frac{1}{2}$, there exists some $k = n^{o(1)}$ such that the distribution $\mathcal{D}_{\mathsf{M}}$ for $\mathsf{Matching}_{n,k,\epsilon}$ in Algorithm 12 satisfies*

$$\mathsf{IC}^{\delta}_{\mathsf{SMP},\mathcal{D}_{\mathsf{M}}}(\mathsf{Matching}_{n,k,\epsilon}) = n^{2-\mathcal{O}(\epsilon)},$$

where $\mathsf{IC}^{\delta}_{\mathsf{SMP},\mathcal{D}_{\mathsf{M}}}(\mathsf{Matching}_{n,k,\epsilon})$ is the information complexity of $\mathsf{Matching}_{\mathsf{n,k,\epsilon}}$ in the multi-party number-in-hand simultaneous message passing model (SMP).

The following theorem summarizes the results in this section, providing a lower bound for Problem 2.

Theorem 41. *Any non-adaptive algorithm for Problem 2 over* $\mathsf{GF}(p)$ *requires* $\Omega(d^2/\log p)$ *queries.*

Proof. We first discuss the case when $d = \Omega(n)$, where we will give an $\Omega(n^2)$ lower bound. Let $\mathbf{A}_G$ be the hard instance given by Algorithm 12. We want to find an $n \times n$ random matrix $\mathbf{H}'$ such that: (1) $\mathsf{rank}(\mathbf{H}'\mathbf{A}) = \mathsf{rank}(\mathbf{A})$ (Multiplying $\mathbf{H}'$ does not change the rank of $\mathbf{A}$ so that testing $\mathbf{A}$ is equivalent to testing $\mathbf{H}'\mathbf{A}$); (2) $\mathbf{M} = \mathbf{H}'\mathbf{A}$ is rigid (Multiplying $\mathbf{H}'$ makes matrix $\mathbf{A}$ rigid). We now show how to do this. Let $\mathbf{B}$ be a random matrix such that we want to distinguish rank n v.s. rank $n - n/\log^2 n$ for matrix $\mathbf{A} := \mathbf{A}_G + \mathbf{B}$. Let $k = \mathsf{rank}(\mathbf{A})$, $\mathbf{H}$ be a $3nk/\delta \times n$ uniformly sampled matrix over $\mathsf{GF}(p)^{3nk/\delta \times n}$ and $\mathbf{H}'$ be the first n rows of $\mathbf{H}$. One can see that any subset of at most n rows of $\mathbf{H}$ has full rank with a large probability.

Proof of (1). We note that $\mathsf{rank}(\mathbf{H}'\mathbf{A}) \le k$. We will show that $\mathbf{H}'\mathbf{A}$ has rank k with probability at least $1 - \delta$. We will use the following lemma.

Lemma 53 (Lemma 5.3, [67])**.** *If* $\mathcal{L} \subseteq \mathsf{GF}(p)^n$ *is a* j*-dimensional linear subspace, and* $\mathbf{A}$ *has rank* $k \ge j$*, then the dimension of* $\mathcal{L}_{\mathbf{A}} := \{\mathbf{w} \in \mathsf{GF}(p)^n \mid \mathbf{w}^\top \mathbf{A} \in \mathcal{L}\}$ *is at most* $n - k + j$*.*

For $j < n$, consider the linear subspace $\mathcal{L}_j$ spanned by the first j rows of $\mathbf{HA}$. By the above lemma, the dimension of the subspace $\mathcal{L}'_j := \{\mathbf{w} \in \mathbb{R}^n \mid \mathbf{w}^\top \mathbf{A} \in \mathcal{L}_j\}$ is at most $n - k + j$. Given that the rows of $\mathbf{H}$ are linearly independent with high probability, at most $n - k + j$ of them can be in $\mathcal{L}'_j$. Thus the probability that $\mathbf{H}'_{(j+1),:}\mathbf{A}$ is not in $\mathcal{L}_j$ is at least $1 - (n - k + j)/(3nk/\delta - j)$, and the probability that all such events hold, for $j = 0, \ldots, k-1$, is at least

$$\left(1 - \frac{n}{3nk/\delta - k}\right)^k = \left(1 - \frac{1}{k}\frac{\delta/3}{1 - \delta/(3n)}\right)^k \ge 1 - \frac{\delta}{2}$$

for small δ. All such independence events occur if and only if $\mathbf{H}'_{1:k,:}\mathbf{A}$ has rank k. Therefore, the probability that $\mathbf{H}'\mathbf{A}$ is of rank k is at least $1 - \delta/2$.

Proof of (2). We need the following result on matrix rigidity.

Lemma 54 (Matrix rigidity, Theorem 6.4, [229])**.** *The fraction of matrices over* $\mathsf{GF}(p)^{n\times n}$ *with matrix rigidity* $\mathcal{R}^{\mathsf{GF}(p)}(r) = \Omega((n-r)^2/\log_p n)$ *is at least* 0.99*, for* $r < n - \sqrt{2n\log_p 2 + \log n}$*.*

For uniform matrix $\mathbf{H}'$, we note that $\mathbf{H}'\mathbf{A}$ is uniform as well: for any given matrix $\mathbf{T}$ in $\mathsf{GF}(p)^{n\times n}$

$$\Pr_{\mathbf{H}'\sim\mathsf{Unif}}[\mathbf{H}'\mathbf{A} = \mathbf{T}] = \Pr_{\mathbf{H}'\sim\mathsf{Unif}}[\mathbf{H}' = \mathbf{T}\mathbf{A}^{-1}] = \left(\frac{1}{p}\right)^{kn}.$$

Then by Lemma 54, $\mathcal{R}^{\mathsf{GF}(p)}_{\mathbf{H}'\mathbf{A}}(n - n/\log n) = \Omega(n^2/\log^2 n)$ with high probability.

Now we are ready to prove the hardness of Problem 2 with parameter $(n, n - \frac{n}{\log^2(n)}, \frac{1}{\log^4(n)\log_p(n)})$. For any non-adaptive algorithm $\mathcal{A}_{test}$ for Problem 2 with $\epsilon = 1/\log n$ and $d = n - n/\log n$, assume that the required number of queries is q. We use such algorithms to estimate the maximum matching size by our reduction. Given a graph G with maximum matching size $\ge 3N/\epsilon$ v.s. $\le 3N/\epsilon - 3N$. We know that the rank of $\mathbf{A} := \mathbf{A}_G + \mathbf{B}$ is of rank n v.s. $n - n/\log^2 n$. By left multiplying matrix $\mathbf{A}$ with above-mentioned $\mathbf{H}'$, the rank of resulting matrix $\mathbf{H}'\mathbf{A}$ remains the same and is of rigidity $\mathcal{R}^{\mathsf{GF}(p)}_{\mathbf{B}}(n - n/\log^2 n) = \Omega(n^2/(\log^4(n)\log_p(n)))$ according to properties (1)

and (2) that we have proven. By assumption, $\mathcal{A}_{test}$ can distinguish rank n from rank $n - n/\log^2 n$ for matrix $\mathbf{H}'\mathbf{A}$ in q queries with high probability. So $\mathcal{A}_{test}$ can be used to compute the maximum matching size with $(1 \pm 1/\log n)$-approximation rate with $\mathcal{O}(q \log p)$ bits of communication. By Theorem 40, we have $q \log p = \Omega(n^2)$, which implies that $q = \Omega(n^2/\log p)$.

We now prove the lower bound for arbitrary d. Let $\mathbf{1} \in \mathsf{GF}(p)^{\frac{n(1-1/\log d)}{d} \times \frac{n(1-1/\log d)}{d}}$ be the all-ones matrix and $\mathbf{A} \in \mathsf{GF}(p)^{\frac{d}{1-1/\log d} \times \frac{d}{1-1/\log d}}$ be the above hard instance. We do the Kronecker product to generate matrix $\mathbf{C} = \mathbf{1} \otimes \mathbf{A} \in \mathsf{GF}(p)^{n \times n}$. If there exists a non-adaptive algorithm $\mathcal{A}_{test}$ that can correctly test whether $\mathbf{C}$ has rank at most d or is far from having rank d with $cd^2/\log p$ queries and high probability for an absolute constant c, the algorithm $\mathcal{A}_{test}$ can also test whether $\mathbf{A}$ has rank at most d or is far from having rank d with $cd^2/\log p$ queries by outputting the same result as testing $\mathbf{C}$. This leads to a contradiction. □

Theorem 41 is tight up to a logarithmic factor. Indeed, there is an $\mathcal{O}(d^2)$ upper bound for every field, independent of ϵ, as follows. If $\mathbf{A}$ is an (unknown) $n \times n$ matrix and has rank at least $d+1$, the matrix $\mathbf{SAT}$ will have rank at least $d+1$ with high probability for random $\mathbf{S}$ of $d+1$ rows and $\mathbf{T}$ of $d+1$ columns; furthermore, this matrix product can be computed in the matrix sensing model because $(\mathbf{SAT})_{i,j}$ can be written as $\langle \mathbf{A}, \mathbf{S}_{i,:}\mathbf{T}_{:,j} \rangle_{i,j}$, which is in the form of matrix sensing. Computing $\mathbf{SAT}$ uses only $(d+1)^2$ measurements instead of the d^2/ϵ we need for reading entries.

Proofs of Theorems 26 and 30

In this section and onwards, we study the problem of non-adaptively testing numerical properties of real-valued matrices. They can be studied under a unified framework in this section.

Roughly, our analytical framework reduces the testing problem to a sequence of estimation problems *without involving* $\mathsf{poly}(n)$ *in the sample complexity*. Our framework consists of two levels of estimation: (1) a constant-factor approximation to the statistic X of interest (e.g., stable rank), and (2) a more accurate $(1 \pm \tau)$-approximation to X.

Definition 8 (Stable rank)**.** *The stable rank of* $\mathbf{A}$ *is defined by* $\mathsf{srank}(\mathbf{A}) = \|\mathbf{A}\|_F^2/\|\mathbf{A}\|^2$*, where* $\|\mathbf{A}\|_F$ *is the Frobenius norm and* $\|\mathbf{A}\|$ *the spectral norm (largest singular value).*

Problem 3 (Stable rank testing in the entry Model)**.** *Let* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *be a matrix which satisfies* $\|\mathbf{A}\|_\infty \le 1$ *and has one of promised properties:*

- H0. $\mathbf{A}$ *has stable rank at most* d*;*
- H1. $\mathbf{A}$ *is* ϵ/d*-far from having stable rank at most* d*, meaning that* $\mathbf{A}$ *requires changing at least an* ϵ/d*-fraction of its entries to have stable rank at most* d*.*

The problem is to design a property testing algorithm that outputs H0 *with probability at least* 0.99 *if* $\mathbf{A} \in$ H0*, and output* H1 *with probability at least 0.99 if* $\mathbf{A} \in$ H1*, with the least number of queried entries.*

Upper bounds We first prove the upper bounds.

Lemma 55 ([197, Theorem 1.8])**.** *Let* $\mathbf{A}$ *be an* $n \times n$ *matrix. Let* $\mathcal{Q}$ *be a uniformly random subset*

of $\{1, 2, \ldots, n\}$ *of expected cardinality* q *with replacement. Then*

$$\mathbb{E}\,\|\mathbf{A}|_{\mathcal{Q}}\| \lesssim \sqrt{\frac{q}{n}}\|\mathbf{A}\| + \sqrt{\log q}\|\mathbf{A}\|_{(n/q)},$$

where $\mathbf{A}|_{\mathcal{Q}} = (\mathbf{A}_{i,j})_{i\in\mathcal{Q}, j\le n}$ *is a random row-submatrix of* $\mathbf{A}$*, and* $\|\mathbf{A}\|_{(n/q)}$ *is the average of* n/q *biggest Euclidean lengths of the columns of* $\mathbf{A}$*.*

Lemma 56. *Let* $\mathbf{x} \sim \mathsf{Unif}(\mathbb{S}^{n-1})$*. Then with probability at least* $1 - n^{-2}$*, we have* $\|\mathbf{x}\|_\infty \le \sqrt{\frac{2\log n}{n}}$*.*

Algorithm 13 Algorithm for stable rank testing under sampling/sensing model

▷ Lines 1-2 estimate the Frobenius norm of $\mathbf{A}$.
1: Uniformly sample $q_0 = \mathcal{O}(\frac{\sqrt{d}}{\epsilon^{2.5}})$ entries $\mathbf{A}$, forming vector $\mathbf{y}$.
2: $X \leftarrow \frac{n^2}{q_0}\|\mathbf{y}\|_2^2$. ▷ X is an estimator of $\|\mathbf{A}\|_F^2$.
3: **If** $X \le \frac{9}{10}(1 - \frac{1}{d})\epsilon n^2$
4: Output "stable rank $\le d$".
5: **Else**
6: Uniformly sample a $q \times q$ submatrix $\widetilde{\mathbf{A}}'$ with $q = \mathcal{O}(\frac{d\log n}{\epsilon})$.
7: **If** $\|\widetilde{\mathbf{A}}'\| \le C_0\frac{\sqrt{X}}{\sqrt{c_1 d}}\frac{q}{n}$
8: Output "ϵ/d-far from being stable rank $\le d$".
9: **Else**
10: Run Algorithm 14 (with $\tau = \Theta(\epsilon/d^{1/4})$) for the sampling model or Algorithm 16 (with $\tau = \Theta(\epsilon/(d^{1/4}\sqrt{\log n}))$) for the sensing model to obtain an operator norm estimate Z.
11: **If** $Z^2 \ge \frac{X}{d}$
12: Output "stable rank $\le d$".
13: **Else**
14: Output "ϵ/d-far from being stable rank $\le d$".
15: **End If**
16: **End If**
17: **End If**

Theorem 42. *Suppose that* $d = \Omega((1/\epsilon)^{1/3})$*. Then (a) Algorithm 13 is a correct algorithm for the stable rank testing problem with failure probability at most* $1/3$ *under the sampling model, and it reads* $\mathcal{O}(\frac{d^3}{\epsilon^4}\log^2 n)$ *entries; (b) Algorithm 13 is a correct algorithm for the stable rank testing problem with failure probability at most* $1/3$ *under the sensing model, and it makes* $\mathcal{O}(\frac{d^{2.5}}{\epsilon^2}\log n)$ *sensing queries.*

Proof. When $\mathbf{A}$ is ϵ/d-far from being stable rank at most d, we claim that $\|\mathbf{A}\|_F^2 \ge \epsilon n^2(1 - \frac{1}{d})$. Otherwise, replacing any $\frac{\epsilon n}{d}$ rows of $\mathbf{A}$ with all-one row vectors $\mathbf{1}^\top$'s results in a new matrix $\mathbf{B}$ such that $\|\mathbf{B}\|^2 \ge \frac{\epsilon n^2}{d}$ and $\|\mathbf{B}\|_F^2 = \|\mathbf{A}\|_F^2 + \frac{\epsilon n^2}{d} \le \epsilon n^2(1 - \frac{1}{d}) + \frac{\epsilon n^2}{d} = \epsilon n^2$, leading to $\mathsf{srank}(\mathbf{B}) \le d$, a contradiction. We note that by sampling q_0 entries from $\mathbf{A}$ and stacking them as vector $\mathbf{y}$, the resulting estimator $X = \frac{n^2}{q_0^2}\|\mathbf{y}\|_2^2$ satisfies $\mathbb{E}[X] = \|\mathbf{A}\|_F^2$ and $\mathsf{Var}[X] \le$

$n^2(n^4/q_0^2)(q_0/n^2) = n^4/q_0$. So by the Chebyshev's inequality, when $q_0 = \mathcal{O}(n^4/(\tau^2\|\mathbf{A}\|_F^4))$ we have

$$\Pr\left[|X - \|\mathbf{A}\|_F^2| > \tau\|\mathbf{A}\|_F^2\right] \le \frac{n^4/q_0}{\tau^2\|\mathbf{A}\|_F^4} \le \frac{1}{3}.$$

Thus we have

$$(1-\tau)\|\mathbf{A}\|_F^2 \le X \le (1+\tau)\|\mathbf{A}\|_F^2, \tag{3.33}$$

where τ will be specified later multiple times. So the algorithm is correct in Step 4 with constant τ (although we over-sample entries here for later purpose).

Let $c_1 > 1$ be an absolute constant to be specified later. We discuss two separate cases.

Case (i). $\mathsf{srank}(\mathbf{A}) > c_1 d$ when $\mathbf{A}$ is far from $\mathsf{srank}(\mathbf{A}) \le d$.

We first discuss the case when $\mathbf{A}$ is far from $\mathsf{srank}(\mathbf{A}) \le d$. Let $\mathbf{U}$ be a uniformly random $n \times n$ orthogonal matrix and let $\mathbf{A}'_{\mathsf{row}}$ be the matrix after uniform row sampling of $\mathbf{A}$ of expected cardinality q. Note that $\|\mathbf{A}'_{\mathsf{row}}\| = \|\mathbf{A}'_{\mathsf{row}}\mathbf{U}\|$, and $(\mathbf{A}\mathbf{U})_{i,:} = \mathbf{A}_{i,:}\mathbf{U}$ uniformly distributes on $\|\mathbf{A}_{i,:}\|_2 \cdot \mathbb{S}^{n-1}$. So $\|\mathbf{A}_{i,:}\mathbf{U}\|_\infty^2 \le 2\|\mathbf{A}_{i,:}\mathbf{U}\|_2^2 \log(n)/n$ for any fixed i with probability at least $1 - 1/n^2$ by Lemma 56. Therefore, with probability at least $1 - 1/n$ by a union bound over all rows, $\|\mathbf{A}\mathbf{U}\|_{\mathsf{col}}^2 \le 2\|\mathbf{A}\|_F^2 \log(n)/n$, where $\|\mathbf{A}\|_{\mathsf{col}}$ represents the maximum ℓ_2 norm among all columns of $\mathbf{A}$. By Lemma 55,

$$\mathbb{E}\|\mathbf{A}'_{\mathsf{row}}\| \le C_1'\sqrt{\frac{q}{n}}\|\mathbf{A}\| + C_2'\sqrt{\log q}\sqrt{\frac{\log n}{n}}\|\mathbf{A}\|_F$$

for absolute constants C_1' and C_2', and by the Markov bound, with probability at least 0.9,

$$\begin{aligned}
\|\mathbf{A}'_{\mathsf{row}}\| &\le C_1\sqrt{\frac{q}{n}}\|\mathbf{A}\| + C_2\sqrt{\log q}\sqrt{\frac{\log n}{n}}\|\mathbf{A}\|_F \\
&\le C_1\sqrt{\frac{q}{n}}\frac{\|\mathbf{A}\|_F}{\sqrt{c_1 d}} + C_2\sqrt{\log q}\sqrt{\frac{\log n}{n}}\|\mathbf{A}\|_F \quad (\text{since } \mathsf{srank}(\mathbf{A}) > c_1 d) \\
&\le \frac{1}{\sqrt{1-\tau}}\left(C_1\sqrt{\frac{q}{c_1 d}} + C_2\sqrt{\log q \log n}\right)\sqrt{\frac{X}{n}} \quad (\text{by Eqn. (3.33)})
\end{aligned}$$

for absolute constants C_1 and C_2. By the Markov bound, we also have with constant probability that

$$\|\mathbf{A}'_{\mathsf{row}}\|_F^2 \le c'\frac{q}{n}\|\mathbf{A}\|_F^2 \le c\frac{q}{n}X.$$

Conditioning on this event, by applying the same argument on the column sampling of $\mathbf{A}'_{\mathsf{row}}$, we have

$$\begin{aligned}
\|\widetilde{\mathbf{A}}'\| &\le C_1\sqrt{\frac{q}{n}}\|\mathbf{A}'_{\mathsf{row}}\| + C_2\sqrt{\log q}\sqrt{\frac{\log n}{n}}\|\mathbf{A}'_{\mathsf{row}}\|_F \\
&\le C_1\sqrt{\frac{q}{n}}\frac{1}{\sqrt{1-\tau}}\left(C_1\sqrt{\frac{q}{c_1 d}} + C_2\sqrt{\log q \log n}\right)\sqrt{\frac{X}{n}} + C_2\sqrt{\log q}\sqrt{\frac{\log n}{n}}\sqrt{qc}\sqrt{\frac{X}{n}} \\
&\le C_0\frac{1}{\sqrt{1-\tau}}\frac{\sqrt{X}}{\sqrt{c_1 d}}\frac{q}{n} \qquad (\text{because the first term dominates as } q \gg d) \\
&\le C_0'\sqrt{\frac{1+\tau}{1-\tau}}\frac{\|\mathbf{A}\|_F}{\sqrt{c_1 d}}\frac{q}{n}, \qquad (\text{by Eqn. (3.33)})
\end{aligned}$$

where $\widetilde{\mathbf{A}}'$ is the matrix after the column sampling of $\mathbf{A}'_{\mathsf{row}}$, and C_0, C_0' are absolute constants.

On the other hand, when $\mathsf{srank}(\mathbf{A}) \le d$ and $q = \mathcal{O}(\frac{d \log n}{\epsilon})$, we have with high probability that

$$\|\widetilde{\mathbf{A}}'\| \ge C\frac{q}{n}\|\mathbf{A}\| = C\frac{q}{n}\frac{\|\mathbf{A}\|_F}{\sqrt{\mathsf{srank}(\mathbf{A})}} \ge C\frac{q}{n}\frac{\|\mathbf{A}\|_F}{\sqrt{d}},$$

where the first inequality holds by applying Lemma 60 twice on row and column sampling (set $\beta = \Theta(\epsilon)$ and $\tau = \Theta(1)$ there). By setting c_1 as a large absolute constant, we have

$$C_0'\sqrt{\frac{1+\tau}{1-\tau}}\frac{\|\mathbf{A}\|_F}{\sqrt{c_1 d}}\frac{q}{n} < C\frac{q}{n}\frac{\|\mathbf{A}\|_F}{\sqrt{d}}.$$

Thus we can distinguish (a) $\mathsf{srank}(\mathbf{A}) \le d$ from (b) $\mathsf{srank}(\mathbf{A})$ ϵ/d-far from being at most d by checking $\|\widetilde{\mathbf{A}}'\|$ in Case (i).

Case (ii). $\mathsf{srank}(\mathbf{A}) \le c_1 d$ when $\mathbf{A}$ is far from $\mathsf{srank}(\mathbf{A}) \le d$.

We now show that we can distinguish the two cases of $\mathsf{srank}(\mathbf{A}) \le d$ from $\mathsf{srank}(\mathbf{A})$ being ϵ/d-far from at most d, suppose we have an accurate estimator to estimate the stable rank.

Let $\mathbf{u} \in \mathbb{S}^{n-1}$ be a unit vector such that $\|\mathbf{A}\| = \|\mathbf{A}\mathbf{u}\|_2$, i.e., $\mathbf{u}$ is a right singular vector corresponding to the largest singular value. First we claim that we can drop off coordinates in $\mathbf{u}$ that are at most $\theta/\sqrt{n}$ for some small constant θ without affecting $\|\mathbf{A}\mathbf{u}\|_2$ by too much.

Let $\mathbf{u}'$ be the vector obtained from $\mathbf{u}$ by zeroing out the coordinates of $\mathbf{u}$ which are *at least* $\theta/\sqrt{n}$, then

$$\|\mathbf{A}\mathbf{u}'\|_2^2 \le \|\mathbf{A}\|^2\|\mathbf{u}'\|_2^2 \le \|\mathbf{A}\|^2 n\left(\frac{\theta}{\sqrt{n}}\right)^2 \le \theta^2\|\mathbf{A}\|^2,$$

and thus

$$\|\mathbf{A}(\mathbf{u}-\mathbf{u}')\|_2 \ge \|\mathbf{A}\mathbf{u}\|_2 - \|\mathbf{A}\mathbf{u}'\|_2 \ge (1-\theta)\|\mathbf{A}\|.$$

Let $\mathbf{u}'' = \mathbf{u} - \mathbf{u}'$ and $\mathbf{v} = \mathbf{A}\mathbf{u}''/\|\mathbf{A}\mathbf{u}''\|_2$, then $(1-\theta)\|\mathbf{A}\| \le \langle \mathbf{A}\mathbf{u}'', \mathbf{v}\rangle$. Next we show similarly that we can drop off coordinates in $\mathbf{v}$ that are at most $\theta/\sqrt{n}$. Similarly we let $\mathbf{v}'$ be the vector obtained from $\mathbf{v}$ by zeroing out the coordinates of $\mathbf{v}$ which are at least $\theta/\sqrt{n}$, then $\|\mathbf{v}'\|_2 \le \theta$, hence

$$\langle \mathbf{A}\mathbf{u}'', \mathbf{v}-\mathbf{v}'\rangle \ge (1-\theta)\|\mathbf{A}\| - \langle \mathbf{A}\mathbf{u}'', \mathbf{v}'\rangle \ge (1-\theta)\|\mathbf{A}\| - \|\mathbf{A}\mathbf{u}''\|_2\|\mathbf{v}'\|_2 \ge (1-2\theta)\|\mathbf{A}\|.$$

Let $\mathbf{v}'' = \mathbf{v} - \mathbf{v}'$. Observe that

$$(1-2\theta)\frac{\|\mathbf{A}\|_F}{\sqrt{c_1 d}} \le (1-2\theta)\|\mathbf{A}\| \le \langle \mathbf{A}\mathbf{u}'', \mathbf{v}''\rangle \le \|\mathbf{A}\mathbf{u}''\|_\infty\|\mathbf{v}''\|_1 \le \|\mathbf{u}''\|_1\|\mathbf{v}''\|_1,$$

where we used the fact that $|\mathbf{A}_{ij}| \le 1$ in the last inequality. This implies that at least one of $\|\mathbf{u}''\|_1$ and $\|\mathbf{v}''\|_1$ is at least $\sqrt{(1-2\theta)\|\mathbf{A}\|_F}/(c_1 d)^{1/4} = c\sqrt{\|\mathbf{A}\|_F}/d^{1/4}$ for some constant $c = \sqrt{1-2\theta}/c_1^{1/4}$.

Without loss of generality, assume that $\|\mathbf{u}''\|_1 \ge c\sqrt{\|\mathbf{A}\|_F}/d^{1/4}$. Next we shall argue that we can drop large coordinates from $\mathbf{u}''$ by affecting $\|\mathbf{u}''\|_1$ by at most a constant factor. To see this, let $I = \{i : |\mathbf{u}''_i| \ge \kappa\}$ for some κ to be determined later. It follows that $|I| \le 1/\kappa^2$ and

$$\|\mathbf{u}''_I\|_1 \le \sqrt{|I|}\,\|\mathbf{u}''_I\|_2 \le \frac{1}{\kappa} = \frac{c}{2}\frac{\sqrt{\|\mathbf{A}\|_F}}{d^{1/4}},$$

provided that

$$\kappa = \frac{2d^{1/4}}{c\sqrt{\|\mathbf{A}\|_F}}.$$

Let $\hat{\mathbf{x}} = \mathbf{u}'' - \mathbf{u}''_I$, we see that $\|\hat{\mathbf{x}}\|_1 \geq \frac{1}{2}\|\mathbf{u}''\|_1$. For notational simplicity let $S = \mathrm{supp}(\hat{\mathbf{x}})$.

Suppose that $\mathbf{A}$ is ϵ/d-far from being stable rank at most d, and we reorder the rows of $\mathbf{A}$ such that $|\langle \mathbf{A}_{1,:}, \mathbf{u}\rangle| \leq |\langle \mathbf{A}_{2,:}, \mathbf{u}\rangle| \leq \cdots \leq |\langle \mathbf{A}_{n,:}, \mathbf{u}\rangle|$. Let $m = \frac{\epsilon n^2}{d|S|}$ (we shall verify that $m \leq n$ later). For $i = 1, \ldots, m$, change $\mathbf{A}_{i,j}$ to $\mathrm{sgn}(\hat{\mathbf{x}}_j)$ for all $j \in S$ if $\langle \mathbf{A}_{i,S^c}, \mathbf{u}\rangle \geq 0$, and change $\mathbf{A}_{i,j}$ to $-\mathrm{sgn}(\hat{\mathbf{x}}_j)$ for all $j \in S$ if $\langle \mathbf{A}_{i,S^c}, \mathbf{u}\rangle < 0$, yielding a matrix $\mathbf{B}$ and we know that $\mathrm{srank}(\mathbf{B}) > d$.

Now we verify that $m \leq n$ so that the aforementioned change is valid. It is clear that $|S| \geq \|\hat{\mathbf{x}}\|_1/\kappa$, and so

$$m \leq \frac{\epsilon n^2}{d \cdot \|\hat{\mathbf{x}}\|_1/\kappa} \leq \frac{\epsilon n^2}{d} \cdot \frac{4\sqrt{d}}{c^2\|\mathbf{A}\|_F} \leq \frac{8\sqrt{\epsilon}}{c^2\sqrt{d}} n < n,$$

provided that $\epsilon \leq \epsilon_0$ for some absolute constant ϵ_0 small enough.

We observe that

$$\|\mathbf{B}\|_F^2 \leq \|\mathbf{A}\|_F^2 + m|S|, \tag{3.34}$$

and

$$\begin{aligned} \|\mathbf{B}\|^2 \geq \|\mathbf{B}\mathbf{u}\|_2^2 &\geq \sum_{i=m+1}^{n} \langle \mathbf{A}_{i,:}, \mathbf{u}\rangle^2 + m\|\hat{\mathbf{x}}\|_1^2 \\ &\geq \left(1 - \frac{m}{n}\right)\|\mathbf{A}\mathbf{u}\|_2^2 + m\|\hat{\mathbf{x}}\|_1^2 \\ &= \left(1 - \frac{m}{n}\right)\|\mathbf{A}\|^2 + m\|\hat{\mathbf{x}}\|_1^2. \end{aligned} \tag{3.35}$$

It follows from $\mathsf{srank}(\mathbf{B}) > d$ that

$$d < \mathsf{srank}(\mathbf{B}) = \frac{\|\mathbf{B}\|_F^2}{\|\mathbf{B}\|^2} \leq \frac{\|\mathbf{A}\|_F^2 + m|S|}{(1 - \frac{m}{n})\|\mathbf{A}\|^2 + m\|\hat{\mathbf{x}}\|_1^2},$$

or,

$$d\left(1 - \frac{m}{n}\right)\|\mathbf{A}\|^2 < \|\mathbf{A}\|_F^2\left(1 - \frac{m(d\|\hat{\mathbf{x}}\|_1^2 - |S|)}{\|\mathbf{A}\|_F^2}\right). \tag{3.36}$$

Next we claim that it holds under certain assumptions

$$d\|\hat{\mathbf{x}}\|_1^2 \geq \frac{1}{\eta_1}|S|. \tag{3.37}$$

Observe that

$$\frac{d\|\hat{\mathbf{x}}\|_1^2}{|S|} = d\|\hat{\mathbf{x}}\|_1 \frac{\|\hat{\mathbf{x}}\|_1}{|S|} \geq d \cdot \frac{\|\mathbf{u}''\|_1}{2} \cdot \frac{\theta}{\sqrt{n}} \geq d^{\frac{3}{4}} c\theta \sqrt{\frac{\|\mathbf{A}\|_F}{n}}, \tag{3.38}$$

which is at least $1/\eta_1$, provided that

$$\|\mathbf{A}\|_F \geq \frac{n}{d^{\frac{3}{2}} c^2 \theta^2 \eta_1^2}. \tag{3.39}$$

This holds when $d = \Omega((1/\epsilon)^{1/3})$ since we know that $\|\mathbf{A}\|_F^2 = \Omega(\epsilon n^2)$.

Hence under the assumption (3.39) it follows from (3.36) that

$$d\left(1-\frac{m}{n}\right)\|\mathbf{A}\|^2 < \|\mathbf{A}\|_F^2\left(1-(1-\eta_1)\frac{md\|\hat{\mathbf{x}}\|_1^2}{\|\mathbf{A}\|_F^2}\right). \tag{3.40}$$

Note that

$$(1-\eta_1)\frac{md\|\hat{\mathbf{x}}\|_1^2}{\|\mathbf{A}\|_F^2} \geq (1-\eta_1)md\cdot\frac{\frac{c^2}{4}\cdot\frac{\|\mathbf{A}\|_F}{\sqrt{d}}}{\|\mathbf{A}\|_F^2} = \frac{(1-\eta_1)c^2}{4}m\cdot\frac{\sqrt{d}}{\|\mathbf{A}\|_F} \geq \frac{1}{\eta_2}\cdot\frac{m}{n},$$

provided that

$$\|\mathbf{A}\|_F \leq \frac{\eta_2(1-\eta_1)c^2}{4}n\sqrt{d}. \tag{3.41}$$

Combining (3.39) and (3.41) leads to that

$$d \geq \frac{2}{c^2\theta\eta_1\sqrt{(1-\eta_1)\eta_2}} = \frac{2\sqrt{c_1}}{\theta(1-2\theta)\eta_1\sqrt{(1-\eta_1)\eta_2}}. \tag{3.42}$$

Now, under both assumptions (3.39) and (3.42), it follows from (3.40) that

$$\begin{aligned}\frac{\|\mathbf{A}\|_F^2}{d\|\mathbf{A}\|^2} &\geq 1+(1-\eta_1-\eta_2)\frac{md\|\hat{\mathbf{x}}\|_1^2}{\|\mathbf{A}\|_F^2}\\ &\geq 1+(1-\eta_1-\eta_2)\frac{\epsilon n^2}{d\|\mathbf{A}\|_F^2}\cdot\frac{d\|\hat{\mathbf{x}}\|_1^2}{|S|}\\ &\geq 1+(1-\eta_1-\eta_2)c\theta\frac{\epsilon n^{3/2}}{d^{1/4}\|\mathbf{A}\|_F^{3/2}}, \qquad \text{(by (3.38))}\end{aligned}$$

Choosing $\theta = 1/4$, $\eta_1+\eta_2 < 1$, we see from (3.42) that we shall need $d = \Omega(\sqrt{c_1})$. It is also easy to verify that (3.41) is satisfied for such d. Overall, we see that we shall need $\tau = \Theta\left(\frac{\epsilon n^{3/2}}{d^{1/4}\|\mathbf{A}\|_F^{3/2}}\right)$ in (3.33).

We are now ready to prove Theorem 26.

Result (a): In fact, We have an accurate estimator to estimate the stable rank by reading an $\mathcal{O}(d^{1.5}\log(n)/\epsilon^2)\times\mathcal{O}(d^{1.5}\log(n)/\epsilon^2)$ submatrix: combining Theorem 27 with Eqn. (3.33) yields an accurate estimator of the stable rank of $\mathbf{A}$:

$$(1-\Theta(\tau))\cdot\mathsf{srank}(\mathbf{A}) \leq \frac{X}{\|\widetilde{\mathbf{A}}\|^2} \leq (1+\Theta(\tau))\cdot\mathsf{srank}(\mathbf{A}).$$

Setting τ as $\Theta(\frac{\epsilon}{d^{1/4}})$ gives the claimed result immediately.

Result (b): It follows from setting $\tau = \Theta(\frac{\epsilon}{d^{1/4}})$ in Theorem 32 on sketching complexity. □

Lower bounds We then prove the lower bounds.

Lemma 57 (Corollary 5.35, [235]). *Let $\mathbf{A}$ be an $m \times n$ ($m > n$) matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$ and fixed $\mathbf{v} \in \mathbb{R}^n$, it holds with probability at least $1 - 2\exp(-t^2/2)$ that*

$$\sqrt{m} - \sqrt{n} - t \leq \sigma_{\min}(\mathbf{A}) \leq \sigma_{\max}(\mathbf{A}) \leq \sqrt{m} + \sqrt{n} + t.$$

Lemma 58 (Lemma 1, [142]). *Let $X \sim \chi^2(k)$. Then we have the tail bound*

$$\Pr[k - 2\sqrt{kx} \leq X \leq k + 2\sqrt{kx} + 2x] \geq 1 - 2e^{-x}.$$

Lemma 59 (Theorem 4, [150]). *Let $\mathbf{u}_1, ..., \mathbf{u}_r$ be i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ vectors and $\mathbf{v}_1, ..., \mathbf{v}_r$ be i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ vectors and further suppose that $\{\mathbf{u}_i\}$ and $\{\mathbf{v}_i\}$ are independent. Let $\mathcal{D}_1 = \mathcal{G}(m, n)$ and $\mathcal{D}_2 = \mathcal{G}(m, n) + \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$, where $\mathbf{s} = [s_1, ..., s_r]^\top$ and $\mathcal{G}(m, n)$ represents $m \times n$ i.i.d. standard Gaussian matrix over $\mathbb{R}$. Denote by $\mathcal{L}_1$ and $\mathcal{L}_2$ the corresponding distribution of the linear sketch of size k on $\mathcal{D}_1$ and $\mathcal{D}_2$. Then there exists an absolute constant $c > 0$ such that $d_{TV}(\mathcal{L}_1, \mathcal{L}_2) \leq 1/10$ whenever $k \leq c/\|\mathbf{s}\|_2^4$, where $d_{TV}(\cdot, \cdot)$ represents the total variation distance between two distributions.*

Theorem 43. *Let $\epsilon \in (0, 1/3)$ and let $d \geq 4$. For $\mathbf{A} \in \mathbb{R}^{(d/\epsilon^2) \times d}$, any algorithm that distinguishes "$\mathsf{srank}(\mathbf{A}) \leq d_0$" from "$\mathbf{A}$ being ϵ_0/d_0-far from stable rank $\leq d_0$" with error probability at most $1/6$ requires measurements $\Omega(d^2/(\epsilon^2 \log(d/\epsilon)))$ for any linear sketch, where $d_0 = \frac{d}{1+\Theta(\epsilon)}$ and $\epsilon_0 = \Theta(\frac{\epsilon}{\log^2(d/\epsilon)})$.*

Remark 1. *Theorem 43 can be generalized to the $(d/\epsilon^2) \times (d/\epsilon^2)$ matrix by concatenating the columns of two hard instances in Theorem 43 $(1/\epsilon^2)$ times. This scales up all singular values in our $(d/\epsilon^2) \times d$ hard instances by a factor of $1/\epsilon$, and thus the stable rank remains the same. Observe that the bounds on $\|\mathbf{G}\|$, $\|\mathbf{G}\|_F$ and $\|\mathbf{S}\|_F$ in (3.45) and (3.46) in the proof below are also scaled up by $1/\epsilon$. The concatenated matrix is therefore ϵ_0/d_0-far from having stable rank at most d_0 following the same argument.*

Proof. Let $m = d/\epsilon^2$ and $n = d$. We will apply Lemma 59 with $r = 1$. To this end, we need to justify that

$$\frac{C}{\log(d/\epsilon)}\mathbf{G} \qquad \text{and} \qquad \frac{C}{\log(d/\epsilon)}(\mathbf{G}_0 + s_1 \mathbf{u}\mathbf{v}^\top)$$

differ in the stable rank (i.e., $\mathsf{srank}(\mathbf{G}) > d_0 \geq \mathsf{srank}(\mathbf{G}_0 + s_1\mathbf{u}\mathbf{v}^\top)$) and that $\mathbf{G}$ is rigid (i.e., changing ϵ_0/d_0-fraction of entries of $\mathbf{G}$ would not change the stable rank of $\mathbf{G}$ to be less than d_0), where the multiplicative factor $C/\log(d/\epsilon)$ is to keep the maximum absolute value of entries in the two hard instances less than 1, $\mathbf{G}, \mathbf{G}_0 \sim \mathcal{G}(m, n)$, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and $s_1 = 3\sqrt{\epsilon/d}$. Note that by Lemma 59, we cannot distinguish $\mathbf{G}$ from $\mathbf{G}' := \mathbf{G}_0 + s_1\mathbf{u}\mathbf{v}^\top$ with $\Omega(d^2/\epsilon^2)$ samples. So if the stable ranks of $\mathbf{G}$ and $\mathbf{G}'$ have a gap, we cannot detect the gap either.

For the operator norm, on one hand, it follows from Lemma 57 that with probability at least $1 - 2\exp(-d/2)$,

$$(1 - 1.1\epsilon)\frac{\sqrt{d}}{\epsilon} \leq \sigma_{\min}(\mathbf{G}) \leq \sigma_{\max}(\mathbf{G}) \leq (1 + 1.1\epsilon)\frac{\sqrt{d}}{\epsilon},$$

for an absolute constant $C_0 > 0$. On the other hand, with probability $\geq 1 - \exp(-\Omega(d))$ we have

$$\begin{aligned}
\left\|\mathbf{G}_0 + s_1\mathbf{u}\mathbf{v}^\top\right\|^2 &= \sup_{\mathbf{x}\in\mathbb{S}^{n-1}} \left\|\mathbf{G}_0\mathbf{x} + s_1\mathbf{u}\mathbf{v}^\top\mathbf{x}\right\|_2^2 \\
&\geq \frac{\left\|\mathbf{G}_0\mathbf{v} + s_1\mathbf{u}\mathbf{v}^\top\mathbf{v}\right\|_2^2}{\|\mathbf{v}\|_2^2} \\
&= \frac{\|\mathbf{G}_0\mathbf{v}\|_2^2}{\|\mathbf{v}\|_2^2} + s_1^2\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2 + 2\langle\mathbf{G}_0\mathbf{v}, s_1\mathbf{u}\rangle \\
&\geq \left((1-1.1\epsilon)\sqrt{\frac{d}{\epsilon^2}}\right)^2 + 0.9^2 s_1^2\frac{d^2}{\epsilon^2} - \mathcal{O}\left(\frac{d}{\sqrt{\epsilon}}\right) \\
&\geq \left((1-1.1\epsilon)\sqrt{\frac{d}{\epsilon^2}}\right)^2 + 0.9^2 s_1^2\frac{d^2}{\epsilon^2} - \mathcal{O}\left(\frac{d}{\sqrt{\epsilon}}\right) \\
&\geq ((1-1.1\epsilon)^2 + 7.29\epsilon)\frac{d}{\epsilon^2} - \mathcal{O}\left(\frac{d}{\sqrt{\epsilon}}\right) \\
&\geq ((1-1.1\epsilon)^2 + 7.29\epsilon)\frac{d}{\epsilon^2} - \mathcal{O}\left(\frac{d}{\sqrt{\epsilon}}\right) \\
&\geq (1+2\epsilon)^2\frac{d}{\epsilon^2},
\end{aligned}$$

where the second inequality (line 4) follows from the concentration of the quadratic form (see, e.g., [198])

$$\Pr_{\mathbf{u},\mathbf{v}}\{|\mathbf{v}^T\mathbf{G}_0\mathbf{u}| > t\} \leq 2\exp\left(-c\min\left\{\frac{t}{\|\mathbf{G}_0\|}, \frac{t^2}{\|\mathbf{G}_0\|_F^2}\right\}\right) \tag{3.43}$$

for fixed $\mathbf{G}_0$; since $\|\mathbf{G}_0\| \simeq \sqrt{d}/\epsilon$ and $\|\mathbf{G}_0\|_F^2 \simeq d^2/\epsilon^2$ with high probability, we can take $t = \Theta(d^{3/2}/\epsilon)$. For the Frobenius norm, we note that

$$\left\|\mathbf{G}_0 + 3\sqrt{\frac{\epsilon}{d}}\mathbf{u}\mathbf{v}^\top\right\|_F^2 = \|\mathbf{G}_0\|_F^2 + 9\frac{\epsilon}{d}\|\mathbf{u}\mathbf{v}^\top\|_F^2 + 6\sqrt{\frac{\epsilon}{d}}\langle\mathbf{G}_0, \mathbf{u}\mathbf{v}^\top\rangle.$$

Observe that $\|\mathbf{G}_0\|_F^2 \sim \chi^2(\frac{d^2}{\epsilon^2})$ so $\|\mathbf{G}_0\|_F^2 = (1 \pm \Theta(\frac{\epsilon}{d}))\frac{d^2}{\epsilon^2}$ with probability ≥ 0.9 by Lemma 58, and $9\frac{\epsilon}{d}\|\mathbf{u}\mathbf{v}^\top\|_F^2 = 9\frac{\epsilon}{d}\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2 = \Theta(\frac{d}{\epsilon})$ with high probability. And also, setting $t = \Theta(d/\epsilon)$ in (3.43), we have with probability at least 0.9 that $|\langle\mathbf{G}_0, \mathbf{u}\mathbf{v}^\top\rangle| = \mathcal{O}(\frac{d}{\epsilon})$ and thus $6\sqrt{\frac{\epsilon}{d}}|\langle\mathbf{G}_0, \mathbf{u}\mathbf{v}^\top\rangle| = \mathcal{O}(\sqrt{\frac{d}{\epsilon}})$. Therefore,

$$\left(1 - \Theta\left(\frac{\epsilon}{d}\right)\right)\|\mathbf{G}_0\|_F^2 \leq \left\|\mathbf{G}_0 + 3\sqrt{\frac{\epsilon}{d}}\mathbf{u}\mathbf{v}^\top\right\|_F^2 \leq \left(1 + \Theta\left(\frac{\epsilon}{d}\right)\right)\|\mathbf{G}_0\|_F^2.$$

As a result,

$$\mathsf{srank}(\mathbf{G}) = \frac{\|\mathbf{G}\|_F^2}{\|\mathbf{G}\|^2} \geq \frac{d}{(1+1.2\epsilon)^2},$$

and

$$\mathsf{srank}(\mathbf{G}') = \frac{\|\mathbf{G}'\|_F^2}{\|\mathbf{G}'\|^2} \le \frac{d}{(1+1.9\epsilon)^2}.$$

By Lemma 59, it is therefore hard to distinguish

$$\mathsf{srank}\left(\frac{C}{\log(d/\epsilon)}\mathbf{G}'\right) = \mathsf{srank}(\mathbf{G}') = \frac{\|\mathbf{G}'\|_F^2}{\|\mathbf{G}'\|^2} \le \frac{d}{(1+1.9\epsilon)^2} \triangleq d_0$$

from

$$\mathsf{srank}\left(\frac{C}{\log(d/\epsilon)}\mathbf{G}\right) = \mathsf{srank}(\mathbf{G}) = \frac{\|\mathbf{G}\|_F^2}{\|\mathbf{G}\|^2} \ge \frac{d}{(1+1.2\epsilon)^2} \ge (1+1.3\epsilon)d_0, \tag{3.44}$$

with sample size $\mathcal{O}(d^2/\epsilon^2)$, provided that ϵ is sufficiently small.

We now show that $\frac{C}{\log(d/\epsilon)}\mathbf{G}$ is rigid, i.e., changing ϵ_0/d_0-fraction of entries of $\frac{C}{\log(d/\epsilon)}\mathbf{G}$ will not make $\mathsf{srank}\left(\frac{C}{\log(\frac{d}{\epsilon})}\mathbf{G}\right) \le d_0$. For any $\mathbf{S} \in \mathbb{R}^{(d/\epsilon^2)\times d}$ such that $\|\mathbf{S}\|_0 = \frac{\theta d}{\epsilon \log^2(\frac{d}{\epsilon})}$ (where $0 < \theta < 1$) and $\left\|\frac{C}{\log(\frac{d}{\epsilon})}\mathbf{G} + \mathbf{S}\right\|_\infty \le 1$ (thus $\mathbf{S}$ has an $\frac{\epsilon_0}{d_0}$-fraction of non-zero entries and $\|\mathbf{S}\|_\infty \le 2$), we have

$$\begin{aligned}
&\left\|\frac{C}{\log(d/\epsilon)}\mathbf{G} + \mathbf{S}\right\|^2 \\
&= \sup_{\|\mathbf{u}\|_2=1, \|\mathbf{v}\|_2=1} \left\langle \left(\frac{C}{\log(d/\epsilon)}\mathbf{G} + \mathbf{S}\right)\mathbf{u}, \mathbf{v}\right\rangle^2 \\
&\le \frac{C^2}{\log^2(d/\epsilon)} \sup_{\substack{\|\mathbf{u}\|_2=1 \\ \|\mathbf{v}\|_2=1}} \langle \mathbf{G}\mathbf{u}, \mathbf{v}\rangle^2 + \sup_{\substack{\|\mathbf{u}\|_2=1 \\ \|\mathbf{v}\|_2=1}} \langle \mathbf{S}\mathbf{u}, \mathbf{v}\rangle^2 + \frac{2C}{\log(d/\epsilon)} \sup_{\substack{\|\mathbf{u}\|_2=1 \\ \|\mathbf{v}\|_2=1}} \langle \mathbf{G}^\top\mathbf{S}\mathbf{u}, \mathbf{v}\rangle \\
&= \frac{C^2}{\log^2(d/\epsilon)}\|\mathbf{G}\|^2 + \|\mathbf{S}\|^2 + \frac{2C}{\log(d/\epsilon)}\|\mathbf{G}^\top\mathbf{S}\| \\
&\le \frac{C^2}{\log^2(d/\epsilon)}\|\mathbf{G}\|^2 + \|\mathbf{S}\|^2 + \frac{2C}{\log(d/\epsilon)}\|\mathbf{G}^\top\|\|\mathbf{S}\| \\
&\le \frac{C^2}{\log^2(d/\epsilon)}\|\mathbf{G}\|^2 + \|\mathbf{S}\|_F^2 + \frac{2C}{\log(d/\epsilon)}\|\mathbf{G}^\top\|\|\mathbf{S}\|_F
\end{aligned}$$

Now, observe that

$$\|\mathbf{S}\|_F^2 \le \frac{4\theta d}{\epsilon \log^2(d/\epsilon)}$$

and that, by setting $t = O(\sqrt{d}/\epsilon)$ in Lemma 57,

$$\|\mathbf{G}\| = \mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}\right)$$

with probability at least $1 - 2\exp(-\Omega(d/\epsilon^2))$, we have that

$$\left\|\frac{C}{\log(d/\epsilon)}\mathbf{G} + \mathbf{S}\right\|^2 \le \frac{C^2}{\log^2(d/\epsilon)}\|\mathbf{G}\|^2 + \mathcal{O}\left(\frac{(\theta+\sqrt{\theta})d}{\epsilon\log^2(d/\epsilon)}\right) \le \left(1 + c_1\sqrt{\theta}\epsilon\right)\frac{C^2}{\log^2(d/\epsilon)}\|\mathbf{G}\|^2, \tag{3.45}$$

where $c_1 > 0$ is an absolute constant.

On the other hand, with probability at least $1 - c' \exp(-\Omega(d/\epsilon))$ it holds that

$$\begin{aligned} \left\| \frac{C}{\log(d/\epsilon)} \mathbf{G} + \mathbf{S} \right\|_F^2 &\geq \left(\frac{C}{\log(d/\epsilon)} \|\mathbf{G}\|_F - \|\mathbf{S}\|_F \right)^2 \\ &\geq \left(1 - c_2 \sqrt{\frac{\epsilon\theta}{d}} \right) \frac{C^2}{\log^2(d/\epsilon)} \|\mathbf{G}\|_F^2, \end{aligned} \tag{3.46}$$

where $c_2 > 0$ is an absolute constant.

Therefore,

$$\mathsf{srank}\left(\frac{C}{\log(d/\epsilon)} \mathbf{G} + \mathbf{S} \right) \geq \left(1 - c_3 \sqrt{\theta} \epsilon \right) \mathsf{srank}\left(\frac{C}{\log(d/\epsilon)} \mathbf{G} \right) > d_0,$$

where $c_3 > 0$ is an absolute constant, and θ is small enough such that the last inequality holds.

We conclude that $\frac{C}{\log(d/\epsilon)} \mathbf{G}$ is ϵ_0/d_0-far from having stable rank $\leq d_0$. The proof is complete. □

Proofs of Theorems 27 and 32

In this section, we develop new $(1 \pm \tau)$-approximation estimators to the operator norm in sampling and sensing models.

Sampling algorithms. We first discuss the sampling algorithms which are only allowed to read the entries of a matrix.

Estimation without eigengap. Before proceeding, we first cite the following result from [165].

Lemma 60 (Theorem 20, [165])**.** *Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ have rows $\{\mathbf{A}_{t,:}\}_{t=1}^n$. Independently sample q rows $\mathbf{A}_{t_1,:}, \ldots, \mathbf{A}_{t_q,:}$ with replacement from $\mathbf{A}$ according to the probabilities:*

$$p_t \geq \beta \frac{\|\mathbf{A}_{t,:}\|_2^2}{\|\mathbf{A}\|_F^2}$$

for $\beta < 1$. Let

$$\mathbf{A}_0 = \begin{bmatrix} \frac{\mathbf{A}_{t_1,:}}{\sqrt{q p_{t_1}}} \\ \vdots \\ \frac{\mathbf{A}_{t_q,:}}{\sqrt{q p_{t_q}}} \end{bmatrix}.$$

Then if $q \geq \frac{4\mathsf{srank}(\mathbf{A})}{\beta \tau^2} \log \frac{2n}{\delta}$, with probability at least $1 - \delta$, we have

$$\|\mathbf{A}^\top \mathbf{A} - \mathbf{A}_0^\top \mathbf{A}_0\| \leq \tau \|\mathbf{A}\|^2.$$

Algorithm 14 The sampling algorithm to estimate $\|\mathbf{A}\|$ up to $(1 \pm \tau)$ relative error

▷ Lines 1-5 estimates the row norms of $\mathbf{A}$ and then sample rows non-uniformly.

1: Sample each row of $\mathbf{A}$ by Bernoulli distribution with probability $\mathcal{O}(\frac{1}{n\tau})$. Denote by $\mathcal{S}_{\mathsf{row}}$ the sampled set and $q = |\mathcal{S}_{\mathsf{row}}|$.
2: **For** $i \leftarrow 1$ **to** q
3: Uniformly sample $\mathcal{O}(\frac{1}{\tau})$ entries from $\mathbf{A}_{\mathcal{S}_{\mathsf{row}}(i),:}$, forming vector $\mathbf{x}$.
4: $r_i \leftarrow \max\{\tau n \|\mathbf{x}\|_2^2, \tau n\}$.
5: **End For**
6: Sample $q_{\mathsf{row}} = \mathcal{O}(\frac{d \log n}{\tau^2})$ indices in $\mathcal{S}_{\mathsf{row}}$ independently with replacement according to the probability $p_i = \frac{r_i}{r}$, where $r = \sum_{j=1}^{q} r_j$. Denote by $\mathcal{I}_{\mathsf{row}}$ the sampled row indices.

▷ Lines 6-10 estimates the column norms of $\mathbf{A}$ and then sample columns non-uniformly.

7: Sample each row with probability $\mathcal{O}(\frac{1}{n\tau})$. Repeat the procedure n times with replacement. Denote the sampled set by $\mathcal{S}_{\mathsf{col}}$ and $q' = |\mathcal{S}_{\mathsf{col}}|$.
8: **For** $i \leftarrow 1$ **to** q'
9: Uniformly sample $\mathcal{O}(\frac{1}{\tau})$ entries from $\mathbf{A}_{\mathcal{I}_{\mathsf{row}},\mathcal{S}_{\mathsf{col}}(i)}$, forming vector $\mathbf{x}$.
10: $r'_i \leftarrow \max\{\tau q \|\mathbf{x}\|_2^2, \tau q\}$.
11: **End For**
12: Sample $q_{\mathsf{col}} = \mathcal{O}(\frac{d \log n}{\tau^2})$ indices in $\mathcal{S}_{\mathsf{col}}$ independently with replacement according to the probability $p'_i = \frac{r'_i}{r'}$, where $r' = \sum_{j=1}^{q'} r'_j$. Denote by $\mathcal{I}_{\mathsf{col}}$ the sampled row indices.
13: $\widetilde{\mathbf{A}} \leftarrow \mathbf{A}_{\mathcal{I}_{\mathsf{row}},\mathcal{I}_{\mathsf{col}}}$. Rescale the rows of $\widetilde{\mathbf{A}}$ by $\left\{\sqrt{\frac{q}{p_i q_{\mathsf{row}}}}\right\}$ and the columns of $\widetilde{\mathbf{A}}$ by $\left\{\sqrt{\frac{q'}{p'_i q_{\mathsf{col}}}}\right\}$.
14: **return** index sets $\mathcal{I}_{\mathsf{row}}, \mathcal{I}_{\mathsf{col}}$, scaling factors $\left\{\sqrt{\frac{q}{p_i q_{\mathsf{row}}}}\right\}$, $\left\{\sqrt{\frac{q'}{p'_i q_{\mathsf{col}}}}\right\}$, $\widetilde{\mathbf{A}}$, and $\|\widetilde{\mathbf{A}}\|$.

Remark 2. *Lemma 60 implies that*

$$(1-\tau)\|\mathbf{A}\|^2 \le \|\mathbf{A}_0\|^2 \le (1+\tau)\|\mathbf{A}\|^2,$$

because

$$\left|\|\mathbf{A}\|^2 - \|\mathbf{A}_0\|^2\right| = \left|\|\mathbf{A}^\top\mathbf{A}\| - \|\mathbf{A}_0^\top\mathbf{A}_0\|\right| \le \|\mathbf{A}^\top\mathbf{A} - \mathbf{A}_0^\top\mathbf{A}_0\| \le \tau\|\mathbf{A}\|^2.$$

Theorem 27 (restated). *Suppose that* $\mathbf{A}$ *is an* $n \times n$ *matrix satisfying that* $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$, $\|\mathbf{A}\|_\infty \le 1$ *and* $\mathsf{srank}(\mathbf{A}) = \mathcal{O}(d)$. *Then with probability at least* 0.9, *the output of Algorithm 14 satisfies* $(1-\tau)\|\mathbf{A}\| \le \|\widetilde{\mathbf{A}}\| \le (1+\tau)\|\mathbf{A}\|$. *The sample complexity is* $\mathcal{O}(d^2\log^2(n)/\tau^4)$.

Proof of Theorem 27. We note that for any row $\mathbf{A}_{i,:}$ such that $|\mathbf{A}_{i,j}| \le 1$ and $\eta \le \|\mathbf{A}_{i,:}\|_2^2 \le n$, uniformly sampling $\Theta(\frac{n}{\eta})$ entries of $\mathbf{A}_{i,:}$ suffices to estimate $\|\mathbf{A}_{i,:}\|_2^2$ within a constant multiplicative factor. To see this, we use Chebyshev's inequality. Let $s = \Theta(\frac{n}{\eta})$ be the number of sampled entries, Z_j be the square of the j-th sampled entry $\mathbf{A}_{i,l(j)}$ of vector $\mathbf{A}_{i,:}$, and $Z = \frac{n}{s}\sum_{j=1}^s Z_j$. So Z is an unbiased estimator:

$$\mathbb{E}[Z] = \frac{n}{s}s\mathbb{E}[Z_1] = n\sum_{j=1}^n \frac{1}{n}\mathbf{A}_{i,l(j)}^2 = \|\mathbf{A}_{i,:}\|_2^2.$$

For the variance, we have

$$\begin{aligned}
\mathsf{Var}[Z] = \frac{n^2}{s^2}\sum_{j=1}^s \mathsf{Var}[Z_j] &\le \frac{n^2}{s^2}\sum_{j=1}^s \mathbb{E}[Z_j^2] = \frac{n^2}{s}\mathbb{E}[Z_1^2] = \frac{n^2}{s}\sum_{j=1}^n \frac{1}{n}\mathbf{A}_{i,j}^4 \\
&\le \frac{n}{s}\sum_{j=1}^n \mathbf{A}_{i,j}^2 \qquad (\text{since } |\mathbf{A}_{i,j}| \le 1) \\
&= \Theta(\eta)\|\mathbf{A}_{i,:}\|_2^2 \\
&\le \Theta(\|\mathbf{A}_{i,:}\|_2^4). \qquad (\text{since } \eta \le \|\mathbf{A}_{i,:}\|_2^2)
\end{aligned}$$

Therefore, by Chebyshev's inequality, we have

$$\Pr\left[\left|Z - \|\mathbf{A}_{i,:}\|_2^2\right| \ge 10\|\mathbf{A}_{i,:}\|_2^2\right] \le \frac{1}{3}.$$

Note that in Step 6 of Algorithm 14, in total we sample $q_{\mathsf{row}} = \mathcal{O}(\frac{d\log n}{\tau^2})$ row indices, obeying the conditions in Lemma 60 for a constant β. By concentration, with high probability $r = \mathcal{O}(\frac{\|\mathbf{A}\|_F^2}{\tau n})$ in Step 6, because in expectation we sample $\mathcal{O}(\frac{1}{\tau^2})$ entries to estimate r and we scale $\|\mathbf{x}\|_2^2$ by a τn factor in Steps 3 and 4, and that $\|\mathbf{A}\|_F^2$ is as large as $\Omega(\tau n^2)$. The probability that any given row i is sampled is equal to $\frac{1}{n\tau} \times \frac{r_i}{r} = \Omega(\frac{r_i}{\|\mathbf{A}\|_F^2})$. Suppose first that $\|\mathbf{A}_{i,:}\|_2^2 \le \tau n$. Then we have $r_i = \tau n$. Consequently, for such i, the probability of sampling row i is at least $\Omega(\frac{\tau n}{\|\mathbf{A}\|_F^2}) \ge \Theta(\frac{\|\mathbf{A}_{i,:}\|_2^2}{\|\mathbf{A}\|_F^2})$, just as in Lemma 60. Suppose next that $\|\mathbf{A}_{i,:}\|_2^2 \ge \tau n$. Then we have $r_i = \Theta(\|\mathbf{A}_{i,:}\|_2^2)$. Consequently, for such i, the probability of sampling row i is at least $\Omega(\frac{\|\mathbf{A}_{i,:}\|_2^2}{\|\mathbf{A}\|_F^2})$, just as in Lemma 60. Therefore, in the followings we can set β in Lemma 60 as an absolute constant.

It follows from Lemma 60 that with probability at least 0.9,

$$(1-\tau)\|\mathbf{A}\|^2 \le \|\mathbf{A}_{\mathsf{row}}\|^2 \le (1+\tau)\|\mathbf{A}\|^2,$$

where $\mathbf{A}_{\mathsf{row}}$ is the *scaled* row sampling of $\mathbf{A}$ as in Lemma 60. Conditioning on this event, by applying Lemma 60 again to the column sampling of $\mathbf{A}_{\mathsf{row}}$, we have with high probability,

$$(1-\tau)^2\|\mathbf{A}\|^2 \le (1-\tau)\|\mathbf{A}_{\mathsf{row}}\|^2 \le \|\widetilde{\mathbf{A}}\|^2 \le (1+\tau)\|\mathbf{A}_{\mathsf{row}}\|^2 \le (1+\tau)^2\|\mathbf{A}\|^2, \tag{3.47}$$

where we have used the fact that $\mathsf{srank}(\mathbf{A}_{\mathsf{row}}) = \mathcal{O}(d)$. The statement $\mathsf{srank}(\mathbf{A}_{\mathsf{row}}) = \mathcal{O}(d)$ holds because $\mathbb{E}\|\mathbf{A}_{\mathsf{row}}\|_F^2 = \|\mathbf{A}\|_F^2$ and by the Markov bound, we have with constant probability that

$$\|\mathbf{A}_{\mathsf{row}}\|_F^2 \le c\|\mathbf{A}\|_F^2,$$

so

$$\mathsf{srank}(\mathbf{A}_{\mathsf{row}}) = \frac{\|\mathbf{A}_{\mathsf{row}}\|_F^2}{\|\mathbf{A}_{\mathsf{row}}\|^2} \le \frac{c\|\mathbf{A}\|_F^2}{(1-\tau)\|\mathbf{A}\|^2} \le C\mathsf{srank}(\mathbf{A}) \le C'd. \qquad \square$$

Estimation with eigengap. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$. Suppose that $p = 2q$. We define a cycle σ to be an ordered pair of a sequence of length q: $\lambda = ((i_1, ..., i_q), (j_1, ..., j_q))$ such that $i_r, j_r \in [k]$ for all r. Now we associate with λ a scalar

$$\mathbf{A}_\lambda = \prod_{\ell=1}^{q} \mathbf{A}_{i_\ell, j_\ell}\mathbf{A}_{i_{\ell+1}, j_\ell}, \tag{3.48}$$

where for convention we define that $i_{q+1} = i_1$. Denote by

$$Z = \frac{1}{N}\sum_{i=1}^{N} \mathbf{A}_{\lambda_i}. \tag{3.49}$$

Our goal is to estimate $\sigma_1(\mathbf{A})$ up to $(1 \pm \tau)$ relative error, which is an $(1 \pm \tau)$ approximation to $\|\mathbf{A}\|$.

Algorithm 15 Estimate $\|\mathbf{A}\|$ up to $(1 \pm \tau)$ relative error

Input: Cycle length q, matrix size n.
Output: $(1 \pm \tau)$-approximation estimator.

1: **For** $i = 1$ **to** N
2: Uniformly sample a cycle λ_i of length q.
3: Compute $\mathbf{A}_{\lambda_i}$ by Eqn. (3.48).
4: **End For**
5: Compute Z as defined in (3.49).
6: **return** $Z^{1/(2q)}n$.

Theorem 44. *Let $\tau \in (0, \frac{1}{2})$ be the accuracy parameter and suppose that the input matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ satisfies*

- $\|\mathbf{A}\|_\infty \le 1$;
- $\|\mathbf{A}\|_F \ge cn$ *for some absolute constant* $c > 0$;
- $\sigma_2(\mathbf{A})/\sigma_1(\mathbf{A}) \le \tau^\gamma$ *for some absolute constant* $\gamma > 0$;
- $\mathsf{srank}(\mathbf{A}) = \mathcal{O}(1)$.

Let $N = \frac{C_1}{\tau^2}\exp(\frac{c_1}{\gamma})$ *and* $q = \frac{C_2}{\gamma}$ *for some large constants* $C_1, C_2 > 0$ *and some small constant* $c_1 > 0$. *Then with probability at least* 0.9, *the estimator returned by Algorithm 15 satisfies* $(1-\tau)\|\mathbf{A}\| \le Z^{1/(2q)}n \le (1+\tau)\|\mathbf{A}\|$. *The sample complexity is* $\Theta(Nq) = \Theta\left(\frac{1}{\gamma\tau^2}\exp(\frac{c_1}{\gamma})\right)$.

Proof of Theorem 44. We show that the cycle estimator approximates $\|\mathbf{A}\|$ within a $(1 \pm \tau)$ relative error. Let $\lambda = (\{i_s\}, \{j_s\})$ which is chosen uniformly with replacement. Recall that

$$\mathbf{A}_\lambda = \prod_{\ell=1}^{q} \mathbf{A}_{i_\ell, j_\ell}\mathbf{A}_{i_{\ell+1}, j_\ell}.$$

Hence

$$\mathbb{E}\mathbf{A}_\lambda = \mathbb{E}\left[\prod_{\ell=1}^{q} \mathbf{A}_{i_\ell, j_\ell}\mathbf{A}_{i_{\ell+1}, j_\ell}\right] = \frac{1}{n^{2q}}\left[\sum_{i_1, i_2, \ldots, i_q, j_1, j_2, \ldots, j_q} \prod_{\ell=1}^{q} \mathbf{A}_{i_\ell, j_\ell}\mathbf{A}_{i_{\ell+1}, j_\ell}\right].$$

Note that (see, e.g., [149])

$$\sum_{i_1, i_2, \ldots, i_q, j_1, j_2, \ldots, j_q} \prod_{\ell=1}^{q} \mathbf{A}_{i_\ell, j_\ell}\mathbf{A}_{i_{\ell+1}, j_\ell} = \|\mathbf{A}\|_{2q}^{2q},$$

and by the assumption on the singular values and the stable rank,

$$\sigma_1(\mathbf{A})^{2q} \le \|\mathbf{A}\|_{2q}^{2q} \le (1+\tau)\sigma_1(\mathbf{A})^{2q},$$

provided that $q \ge \frac{1}{2\gamma}(\frac{\log \mathsf{srank}(\mathbf{A})}{\log(1/\tau)} + 1)$, and thus it suffices to take $q = \Theta(\frac{1}{\gamma})$.

Therefore, noting that $\mathbb{E}[Z] = \mathbb{E}[\mathbf{A}_\lambda]$,

$$\mathbb{E}[Z] \le \frac{1+\tau}{n^{2q}}\sigma_1(\mathbf{A})^{2q} \le 1 + \tau, \tag{3.50}$$

$$\mathbb{E}[Z] \ge \frac{1}{n^{2q}}\sigma_1(\mathbf{A})^{2q} \ge \frac{1}{n^{2q}}\left(\frac{\|\mathbf{A}\|_F^2}{\mathsf{srank}(\mathbf{A})}\right)^q \ge \left(\frac{c^2}{\mathsf{srank}(\mathbf{A})}\right)^q = \exp\left(\frac{c_1}{\gamma}\right). \tag{3.51}$$

We now bound the variance of $\mathbf{A}_\lambda$. Observe that

$$\mathsf{Var}[\mathbf{A}_\lambda] \le \mathbb{E}[\mathbf{A}_\lambda^2] \le 1,$$

because $|\mathbf{A}_{i,j}| \le 1$ for all $i, j \in [n]$. Thus by repeating the procedure $N = \frac{C_1}{\tau^2}\exp\left(\frac{2c_1}{\gamma}\right)$ times, we have

$$\mathsf{Var}[Z] = \frac{1}{N}\mathsf{Var}[\mathbf{A}_\lambda] \le \frac{1}{10}\tau^2\exp\left(\frac{2c_1}{\gamma}\right),$$

by choosing C_1 sufficiently large. It follows from the Chebyshev inequality that

$$\Pr\left[|\mathbb{E}[Z] - Z| > \tau\mathbb{E}[Z]\right] \le \frac{\mathsf{Var}[Z]}{\tau^2\mathbb{E}[Z]^2} \le \frac{1}{10},$$

where we have used the lower bound (3.51). This together with (3.50) and (3.51) implies that

$$\Pr\left[(1-\tau)\frac{1}{n^{2q}}\sigma_1(\mathbf{A})^{2q} \le Z \le (1+\tau)^2\frac{1}{n^{2q}}\sigma_1(\mathbf{A})^{2q}\right] > \frac{9}{10}.$$

So

$$\Pr\left[(1-\tau)\,\sigma_1(\mathbf{A}) \le Z^{1/(2q)}n \le (1+\tau)\,\sigma_1(\mathbf{A})\right] > \frac{9}{10}.$$

□

Sensing algorithms. We now discuss the sensing algorithms.

Theorem 32 (restated). *Suppose that $\mathbf{A}$ is an $n \times n$ matrix such that $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$, $\|\mathbf{A}\|_\infty \le 1$ and $\mathsf{srank}(\mathbf{A}) = \mathcal{O}(d)$. Then Algorithm 16 outputs a value Z, which satisfies $(1-\tau)\|\mathbf{A}\| \le Z \le (1+\tau)\|\mathbf{A}\|$ with probability at least 0.9. The sketching complexity is $\mathcal{O}(\max\{\log^2(d\log(n)/\tau), d^2\log(n)\}/\tau^2)$.*

Remark 3. *The optimality of Theorem 32 follows from the hard instance in the proof of Theorem 43.*

Algorithm 16 The sketching/sensing algorithm to estimate $\|\mathbf{A}\|$ up to $(1\pm\tau)$ relative error

1: Obtain indices $\mathcal{I}_{\mathsf{row}}$, $\mathcal{I}_{\mathsf{col}}$ and scaling factors $\left\{\sqrt{\frac{q}{p_i q_{\mathsf{row}}}}\right\}$, $\left\{\sqrt{\frac{q'}{p'_i q_{\mathsf{col}}}}\right\}$ by Algorithm 14 with $|\mathcal{I}_{\mathsf{row}}| = |\mathcal{I}_{\mathsf{col}}| = \mathcal{O}(d\log(n)/\tau^2)$.
2: Let $\mathbf{G}$ and $\mathbf{H}$ be $\Theta(\frac{\max\{\log(d\log(n)/\tau), d\}}{\tau}) \times \mathcal{O}(\frac{d\log n}{\tau^2})$ matrices with i.i.d. $\mathcal{N}(0,1)$ entries. Scale the columns of $\mathbf{G}$ by $\left\{\sqrt{\frac{q}{p_i q_{\mathsf{row}}}}\right\}$ and the columns of $\mathbf{H}$ by $\left\{\sqrt{\frac{q'}{p'_i q_{\mathsf{col}}}}\right\}$.
3: Maintain $\mathbf{G}\mathbf{A}_{\mathcal{I}_{\mathsf{row}},\mathcal{I}_{\mathsf{col}}}\mathbf{H}^\top$.
4: Compute Y defined in Eqn. (3.53).
5: **return** $Y^{\tau/(2\log(d\log(n)/\tau^2))}$.

Before proving Theorem 32, we introduce a new estimator of operator norm under the sensing model, which approximates the operator norm by the Schatten-p norm of large p.

Specifically, let $\mathbf{A}$ be an $n \times n$ matrix. We define a cycle σ to be an ordered pair of a sequence of length q with $p = 2q$: $\lambda = ((i_1,\ldots,i_q),(j_1,\ldots,j_q))$ such that $i_r, j_r \in [k]$ for all r, $i_r \ne i_s$ and $j_r \ne j_s$ for $r \ne s$. Now we associate with λ a scalar

$$\mathbf{A}_\lambda = \prod_{\ell=1}^{q} \mathbf{A}_{i_\ell,j_\ell}\mathbf{A}_{i_{\ell+1},j_\ell}, \tag{3.52}$$

where for convention we define that $i_{q+1} = i_1$. Denote by $\mathcal{C}$ the set of cycles. We define

$$Y = \frac{1}{|\mathcal{C}|}\sum_{\lambda\in\mathcal{C}}(\mathbf{G}\mathbf{A}\mathbf{H}^\top)_\lambda \tag{3.53}$$

for even p, where $\mathbf{G} \sim \mathcal{G}(k,n)$, $\mathbf{H} \sim \mathcal{G}(k,n)$, and $k \geq q$. This estimator, akin to that in [152], approximates the Schatten-p and thus the operator norm, as we shall show below.

Lemma 61. *Suppose that $\mathbf{A}$ is a $n \times n$ matrix of stable rank at most d. Let $k = \Theta(\max\{\sqrt{nd}, \log n\})$ and Y be the estimator defined in (3.53). With probability at least 0.9, it holds that $(1-\tau)\|\mathbf{A}\| \leq Y^{\tau/(2\log(n))} \leq (1+\tau)\|\mathbf{A}\|$. The sketching complexity is $\mathcal{O}(k^2) = \mathcal{O}(\max\{nd, \log^2 n\})$.*

Proof of Lemma 61. We first show that $\|\mathbf{A}\|_{\mathcal{S}_p}$ and $\|\mathbf{A}\|$ differ at most a $(1 \pm \tau)$ factor for $p = 2\lceil \log(n)/\tau \rceil$. To see this,

$$1 \leq \frac{\|\mathbf{A}\|_{\mathcal{S}_p}^p}{\|\mathbf{A}\|^p} = \frac{\sigma_1^p(\mathbf{A}) + \sigma_2^p(\mathbf{A}) + \cdots + \sigma_n^p(\mathbf{A})}{\sigma_1^p(\mathbf{A})} \leq n,$$

and therefore

$$1 \leq \frac{\|\mathbf{A}\|_{\mathcal{S}_p}}{\|\mathbf{A}\|} \leq n^{1/p} \leq 1 + \frac{1}{2}\tau.$$

We now show that the cycle estimator $Y^{1/p}$ approximates $\|\mathbf{A}\|_{\mathcal{S}_p}$ within a $(1 \pm \frac{1}{2}\tau)$ relative error. We say that two cycles $\lambda = (\{i\}, \{j\})$ and $\tau = (\{i'\}, \{j'\})$ are (a_1, a_2)-disjoint if $|i \Delta i'| = 2a_1$ and $|j \Delta j'| = 2a_2$, denoted by $|\lambda \Delta \tau| = (a_1, a_2)$. Here Δ is the symmetric difference. Denote by $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ the skinny SVD of $\mathbf{A}$. Let $\mathbf{G}$ and $\mathbf{H}$ be random matrices with i.i.d. $\mathcal{N}(0,1)$ entries. Note that $\mathbf{G}\mathbf{A}\mathbf{H}^\top$ is identically distributed as $\mathbf{G}\mathbf{\Sigma}\mathbf{H}^\top$ by rotational invariance. Let $\widetilde{\mathbf{A}}$ be the $k \times k$ matrix $\mathbf{G}\mathbf{\Sigma}\mathbf{H}^\top$, where $k \geq q$. It is clear that

$$\widetilde{\mathbf{A}}_{s,t} = \sum_{i=1}^{n} \sigma_i \mathbf{G}_{s,i} \mathbf{H}_{t,i}.$$

Define

$$Y = \frac{1}{|\mathcal{C}|} \sum_{\lambda \in \mathcal{C}} \widetilde{\mathbf{A}}_\lambda.$$

Let $\lambda = (\{i_s\}, \{j_s\})$. Then

$$\widetilde{\mathbf{A}}_\lambda = \sum_{\substack{\ell_1 \in [n], \ldots, \ell_q \in [n] \\ m_1 \in [n], \ldots, m_q \in [n]}} \prod_{s=1}^{q} \sigma_{\ell_s} \sigma_{m_s} \mathbf{G}_{i_s, \ell_s} \mathbf{H}_{j_s, \ell_s} \mathbf{G}_{i_{s+1}, m_s} \mathbf{H}_{j_s, m_s}.$$

We note that

$$\mathbb{E} Y = \mathbb{E} \widetilde{\mathbf{A}}_\lambda = \sum_{i=1}^{n} \sigma_i^{2q} = \|\mathbf{A}\|_{\mathcal{S}_p}^p.$$

We now bound the variance of Y. Let $\tau = (\{i'_s\}, \{j'_s\})$. Observe that

$$\mathbb{E} Y^2 = \frac{1}{|\mathcal{C}|^2} \sum_{a_1=0}^{q} \sum_{a_2=0}^{q} \sum_{\substack{\lambda, \tau \in \mathcal{C} \\ |\lambda \Delta \tau| = (a_1, a_2)}} \mathbb{E}(\widetilde{\mathbf{A}}_\lambda \widetilde{\mathbf{A}}_\tau),$$

where

$$\mathbb{E}(\widetilde{\mathbf{A}}_\lambda \widetilde{\mathbf{A}}_\tau) = \sum_{\substack{\ell_1\in[n],\dots,\ell_q\in[n]\\ \ell'_1\in[n],\dots,\ell'_q\in[n]\\ m_1\in[n],\dots,m_q\in[n]\\ m'_1\in[n],\dots,m'_q\in[n]}} \left(\prod_{i=1}^{q}\sigma_{\ell_i}\sigma_{m_i}\sigma_{\ell'_i}\sigma_{m'_i}\right)\mathbb{E}\left(\prod_{s=1}^{q}\mathbf{G}_{i_s,\ell_s}\mathbf{G}_{i_{s+1},m_s}\mathbf{G}_{i'_s,\ell'_s}\mathbf{G}_{i'_{s+1},m'_s}\right) \times \mathbb{E}\left(\prod_{s=1}^{q}\mathbf{H}_{j_s,\ell_s}\mathbf{H}_{j_s,m_s}\mathbf{H}_{j'_s,\ell'_s}\mathbf{H}_{j'_s,m'_s}\right). \tag{3.54}$$

For any fixed cycles $\lambda = (\{i_s\},\{j_s\})$ and $\tau = (\{i'_s\},\{j'_s\})$ such that $|\lambda\Delta\tau| = (a_1,a_2)$, we notice that

$$\mathbb{E}(\widetilde{\mathbf{A}}_\lambda \widetilde{\mathbf{A}}_\tau) \le (2cnd)^p \|\mathbf{A}\|_{\mathcal{S}_p}^{2p}, \tag{3.55}$$

for an absolute constant c. To see this, we observe that for the expectation $\mathbb{E}(\widetilde{\mathbf{A}}_\lambda \widetilde{\mathbf{A}}_\tau)$ to be non-zero, we must have that each appeared $\mathbf{G}$ and $\mathbf{H}$ in Eqn. (3.54) repeats an even number of times. Though there are totally n^{4q} many of configurations for $\{\ell_s\}$, $\{\ell'_s\}$, $\{m_s\}$ and $\{m'_s\}$, there are at most $n^{2q}3^q$ non-zero terms among the summation in Eqn. (3.54). This is because each $\mathbf{G}$ and $\mathbf{H}$ must have power 2 or 4 by the construction of the cycle. We know that for each fixed configuration of blocks there are at most n^{2q} free variables, and there are at most 16^q different kinds of configurations of blocks because the size of each block is at most 4. So the number of non-zero terms is at most $(4n)^{2q}$. This is true no matter whether there exists some i_r, i'_s or j_r, j'_s such that $i_r = i'_s$ or $j_r = j'_s$. We also claim that for each non-zero term in the summation of Eqn. (3.54),

$$\mathbb{E}\left(\prod_{s=1}^{q}\mathbf{G}_{i_s,\ell_s}\mathbf{G}_{i_{s+1},m_s}\mathbf{G}_{i'_s,\ell'_s}\mathbf{G}_{i'_{s+1},m'_s}\right)\cdot\mathbb{E}\left(\prod_{s=1}^{q}\mathbf{H}_{j_s,\ell_s}\mathbf{H}_{j_s,m_s}\mathbf{H}_{j'_s,\ell'_s}\mathbf{H}_{j'_s,m'_s}\right) \le 25^q.$$

This is because $\mathbb{E}\mathbf{G}^2 = \mathbb{E}\mathbf{H}^2 = 1$ and $\mathbb{E}\mathbf{G}^4 = \mathbb{E}\mathbf{H}^4 = 3$. Therefore, for a certain configuration in which $p_1,\dots,p_w$ are free variables with multiplicity $r_1,\dots,r_w \ge 2$, the summation in Eqn. (3.54) is bounded by

$$4n^{2q}100^q \sum_{p_1,\dots,p_w} \sigma_{p_1}^{r_1}\cdots\sigma_{p_w}^{r_w} \le (2n)^p \|\mathbf{A}\|_{\mathcal{S}_{r_1}}^{r_1}\cdots\|\mathbf{A}\|_{\mathcal{S}_{r_w}}^{r_w} \le (2nd)^p\|\mathbf{A}\|_{\mathcal{S}_p}^{2p},$$

where the last inequality follows from the facts that $\sum_{i=1}^{w} r_i = 2p$ and, by the assumption $\mathsf{srank}(A) \le d$, that $\|\mathbf{A}\|_{\mathcal{S}_r} \le \|\mathbf{A}\|_F \le \sqrt{d}\|\mathbf{A}\|_{\mathcal{S}_p}$ for any $r \ge 2$. Thus we obtain Eqn. (3.55).

We now bound $\mathbb{E}Y^2$. Note that $|\mathcal{C}| = \Theta(k^p)$ and there are

$$\binom{k}{q}\binom{q}{q-a_1}\binom{k-(q-a_1)}{a_1}\binom{k}{q}\binom{q}{q-a_2}\binom{k-(q-a_2)}{a_2}$$

pairs of (a_1,a_2)-disjoint cycles, which can be upper bounded by $\mathcal{O}(10^q)$. Hence

$$\mathbb{E}Y^2 = \frac{1}{|\mathcal{C}|^2}\sum_{a_1=0}^{q}\sum_{a_2=0}^{q}\sum_{\substack{\lambda,\tau\in\mathcal{C}\\ |\lambda\Delta\tau|=(a_1,a_2)}} \mathbb{E}(\widetilde{\mathbf{A}}_\lambda\widetilde{\mathbf{A}}_\tau) \le C'\frac{1}{k^{2p}}q^2 10^q (2nd)^p\|\mathbf{A}\|_{\mathcal{S}_p}^{2p} \le \|\mathbf{A}\|_{\mathcal{S}_p}^{2p},$$

by the assumption that $k = \Omega(\sqrt{nd})$.

It follows that

$$\mathsf{Var}[Y] \le \mathbb{E}Y^2 \le \|\mathbf{A}\|_{\mathcal{S}_p}^{2p}.$$

Then by the Chebyshev inequality,

$$\Pr\left[\left|\|\mathbf{A}\|_{\mathcal{S}_p}^p - Y\right| > \frac{1}{2}\|\mathbf{A}\|_{\mathcal{S}_p}^p\right] \le \frac{\mathsf{Var}[Y]}{4\|\mathbf{A}\|_{\mathcal{S}_p}^{2p}} \le \frac{1}{10},$$

namely,

$$\Pr\left[\left(1 - \frac{1}{2}\tau\right)\|\mathbf{A}\|_{\mathcal{S}_p} \le Y^{1/p} \le \left(1 + \frac{1}{2}\tau\right)\|\mathbf{A}\|_{\mathcal{S}_p}\right] > \frac{9}{10}.$$

This together with the fact that $\|\mathbf{A}\| \le \|\mathbf{A}\|_{\mathcal{S}_p} \le (1 + \frac{1}{2}\tau)\|\mathbf{A}\|$ implies that

$$\Pr\left[(1-\tau)\|\mathbf{A}\| \le Y^{1/p} \le (1+\tau)\|\mathbf{A}\|\right] > \frac{9}{10},$$

as desired. This completes the proof of Lemma 61. □

We are now ready to prove Theorem 32. Recall that we have shown that by focusing on an $\mathcal{O}(\frac{d\log n}{\tau^2}) \times \mathcal{O}(\frac{d\log n}{\tau^2})$ submatrix (without sampling it), we can achieve guarantee (3.47) when $\|\mathbf{A}\|_F^2 = \Omega(\tau n^2)$ and $\|\mathbf{A}\|_\infty \le 1$. Letting $d \leftarrow c_1 d$ and $n \leftarrow \mathcal{O}(\frac{d\log n}{\tau^2})$ in Lemma 61 concludes the proof of Theorem 32.

Proofs of Theorems 28 and 31

We study the problem of testing Schatten-p norms in this section.

Upper bounds. We first prove the upper bound for $p > 2$.

Problem 4 (Schatten-p Norm Testing in the Bounded Entry Model for $p > 2$)**.** *Let $p > 2$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a matrix such that $\|\mathbf{A}\|_\infty \le 1$. For an absolute constant c, the matrix $\mathbf{A}$ satisfies one of the promised properties:*

- H0. *$\|\mathbf{A}\|_{\mathcal{S}_p}^p \ge cn^p$;*
- H1. *$\mathbf{A}$ is ϵ-far from $\|\mathbf{A}\|_{\mathcal{S}_p}^p \ge cn^p$, meaning that it requires changing at least an ϵ-fraction of the entries of $\mathbf{A}$ such that that $\|\mathbf{A}\|_{\mathcal{S}_p}^p \ge cn^p$.*

The problem is to design a non-adaptive property testing algorithm that outputs H0 *with probability at least 0.9 if $\mathbf{A} \in$ H0, and output* H1 *with probability at least 0.99 if $\mathbf{A} \in$ H1, with the least number of queried entries.*

First we prove a lemma showing that H0 and H1 can be distinguished by the Schatten p-norm.

Lemma 62. *Suppose that $p > 2$ is a constant. There exist constants $c = c(p)$, $C = C(p)$ and $\epsilon_0 = \epsilon(p)$ such that for any $\epsilon \in [C/n, \epsilon_0]$, when $\mathbf{A} \in$ H1 it holds that $\|\mathbf{A}\|_{\mathcal{S}_p}^p \le (c - c'\epsilon)n^p$ for some small constant c' that may depend on p.*

Proof. Assume that $\|\mathbf{A}\|_{\mathcal{S}_p}^p \geq c_1 n^p$ for some constant $c_1 < c$, otherwise there is already a constant-factor gap. Together with the assumption that $\|\mathbf{A}\|_\infty \leq 1$ and thus $\|\mathbf{A}\|_F^2 \leq n^2$, it must hold that $\|\mathbf{A}\| \geq c_2 n$ for $c_2 = c_1^{1/(p-1)}$.

We claim that we can find a set T of ϵn rows such that $\|\mathbf{A}_{T^c,:}\|_{\mathcal{S}_p}^p \geq (1 - C'\epsilon)\|\mathbf{A}\|_{\mathcal{S}_p}^p$ for some C' (which may depend on p) and therefore $\|\mathbf{A}_{T^c,:}\| \geq c_2' n$, where $\mathbf{A}_{T^c,:}$ stands for a submatrix of $\mathbf{A}$ with rows restricted on the set T^c. Consider a random subset T formed by including rows independently with probability ϵ, that is, let δ_i be the indicator variable whether $i \in T$ and $\mathbb{E}\delta_i = \epsilon$. Denote by $\mathbf{A}_{i,:} \in \mathbb{R}^{1\times n}$ the i-th row of $\mathbf{A}$. Then we have, by the standard symmetrization trick (see, e.g., [146, Lemma 6.3]), that

$$\begin{aligned}\mathbb{E}_{\delta_i}\|\mathbf{A}_{T,:}\|_{\mathcal{S}_p}^2 = \mathbb{E}_{\delta_i}\left\|\sum_i \delta_i \mathbf{A}_{i,:}^\top \mathbf{A}_{i,:}\right\|_{\mathcal{S}_{p/2}} &\leq \mathbb{E}_{\delta_i}\left(\left\|\sum_i (\delta_i - \epsilon)\mathbf{A}_{i,:}^\top \mathbf{A}_{i,:}\right\|_{\mathcal{S}_{p/2}} + \left\|\sum_i \epsilon \mathbf{A}_{i,:}^\top \mathbf{A}_{i,:}\right\|_{\mathcal{S}_{p/2}}\right) \\ &\leq 2\mathbb{E}_{\delta_i}\mathbb{E}_{\epsilon_i}\left\|\sum_i \epsilon_i \delta_i \mathbf{A}_{i,:}^\top \mathbf{A}_{i,:}\right\|_{\mathcal{S}_{p/2}} + \epsilon\|\mathbf{A}\|_{\mathcal{S}_p}^2,\end{aligned}$$

where ϵ_i's are i.i.d. $\{\pm 1\}$-valued Rademacher variables with $\Pr(\epsilon_i = +1) = \Pr(\epsilon_i = -1) = 1/2$. Applying the Non-Commutative Khintchine Inequality (abbreviated as NCKI) [162] yields that

$$\begin{aligned}\mathbb{E}_{\epsilon_i}\left\|\sum_i \epsilon_i \delta_i \mathbf{A}_{i,:}^\top \mathbf{A}_{i,:}\right\|_{\mathcal{S}_{p/2}} &\leq \left(\mathbb{E}_{\epsilon_i}\left\|\sum_i \epsilon_i \delta_i \mathbf{A}_{i,:}^\top \mathbf{A}_{i,:}\right\|_{\mathcal{S}_{p/2}}^{p/2}\right)^{2/p} \qquad \text{(by Jensen's inequality)} \\ &\leq C_1\sqrt{\frac{p}{2}}\left\|\left(\sum_i \delta_i (\mathbf{A}_{i,:}^\top \mathbf{A}_{i,:})^2\right)^{\frac{1}{2}}\right\|_{\mathcal{S}_{p/2}} \qquad \text{(by NCKI)} \\ &\leq C_1\sqrt{\frac{p}{2}}\left\|\max_i \|\mathbf{A}_{i,:}\|_2 \cdot \left(\sum_i \delta_i \mathbf{A}_{i,:}^\top \mathbf{A}_{i,:}\right)^{\frac{1}{2}}\right\|_{\mathcal{S}_{p/2}} \\ &\leq C_1\sqrt{\frac{p}{2}}\sqrt{n}\|\mathbf{A}_{T,:}\|_{\mathcal{S}_{p/2}} \\ &\leq C_1\sqrt{\frac{p}{2}}n^{\frac{1}{2}}|T|^{\frac{1}{p}}\|\mathbf{A}_{T,:}\|_{\mathcal{S}_p}, \qquad \text{(by Hölder's inequality)}\end{aligned}$$

where the third inequality holds since $\sum_i \delta_i (\mathbf{A}_{i,:}^\top \mathbf{A}_{i,:})^2 \preceq \max_i \|\mathbf{A}_{i,:}\|_2^2 \cdot \sum_i \delta_i \mathbf{A}_{i,:}^\top \mathbf{A}_{i,:}$. Hence, taking expectation on both sides w.r.t. δ_i,

$$\begin{aligned}\mathbb{E}\|\mathbf{A}_{T,:}\|_{\mathcal{S}_p}^2 &\leq C_1\sqrt{\frac{p}{2}}n^{\frac{1}{2}}(\mathbb{E}|T|^{\frac{2}{p}})^{\frac{1}{2}}(\mathbb{E}\|\mathbf{A}_{T,:}\|_{\mathcal{S}_p}^2)^{\frac{1}{2}} + \epsilon\|\mathbf{A}\|_{\mathcal{S}_p}^2 \qquad \text{(by Cauchy-Schwarz inequality)} \\ &\leq C_1\sqrt{\frac{p}{2}}n^{\frac{1}{2}}(\mathbb{E}|T|)^{\frac{1}{p}}(\mathbb{E}\|\mathbf{A}_{T,:}\|_{\mathcal{S}_p}^2)^{\frac{1}{2}} + \epsilon\|\mathbf{A}\|_{\mathcal{S}_p}^2 \qquad \text{(by Jensen's inequality)} \\ &\leq C_1\sqrt{\frac{p}{2}}n^{\frac{1}{2}}(\epsilon n)^{\frac{1}{p}}(\mathbb{E}\|\mathbf{A}_{T,:}\|_{\mathcal{S}_p}^2)^{\frac{1}{2}} + \epsilon\|\mathbf{A}\|_{\mathcal{S}_p}^2,\end{aligned}$$

whence we can solve that

$$\mathbb{E}\|\mathbf{A}_{T,:}\|_{\mathcal{S}_p}^2 \le C_1^2 \frac{p}{2} \epsilon^{\frac{2}{p}} n^{1+\frac{2}{p}} + 4\epsilon \|\mathbf{A}\|_{\mathcal{S}_p}^2 \le C_2 \epsilon \|\mathbf{A}\|_{\mathcal{S}_p}^2.$$

That is, we can find T such that $\|\mathbf{A}_{T,:}\|_{\mathcal{S}_p}^2 \le C_2\epsilon\|\mathbf{A}\|_{\mathcal{S}_p}^2$ and thus

$$\|\mathbf{A}_{T^c,:}\|_{\mathcal{S}_p}^2 = \|\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:}\|_{\mathcal{S}_{p/2}} = \|\mathbf{A}^\top\mathbf{A} - \mathbf{A}_{T,:}^\top \mathbf{A}_{T,:}\|_{\mathcal{S}_{p/2}} \ge (1 - C_2\epsilon)\|\mathbf{A}^\top\mathbf{A}\|_{\mathcal{S}_{p/2}} = (1-C_2\epsilon)\|\mathbf{A}\|_{\mathcal{S}_p}^2$$

as desired. A Chernoff bound shows that $|T| \ge 0.9\epsilon n$ with at least a high constant probability. We can assume that it happens, since conditioning on this event will increase $\mathbb{E}\|\mathbf{A}_{T,:}\|_{\mathcal{S}_p}^2$ just by a constant factor. When $|T| > \epsilon n$ we can just remove rows from T, which only decreases $\|\mathbf{A}_{T,:}\|_{\mathcal{S}_p}$.

Let $\mathbf{v}$ be the normalized eigenvector associated with the largest eigenvalue of $\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:}$. We shall change the rows of T all to $\tilde{\mathbf{v}} := \operatorname{sgn}(\mathbf{v})$, obtaining a matrix $\mathbf{B}$. Note that $\mathbf{B}^\top\mathbf{B} = \mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:} + |T|\mathbf{v}\mathbf{v}^\top \succeq \mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:}$ is a rank-1 PSD perturbation of $\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:}$ and $\mathbf{v}$ is the leading eigenvector of $\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:}$, we have that for the i-th eigenvalue $\lambda_i(\cdot)$,

$$\lambda_i(\mathbf{B}^\top\mathbf{B}) \ge \lambda_i(\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:}), \quad i \ge 2.$$

and the largest eigenvalue

$$\begin{aligned}
\lambda_1(\mathbf{B}^\top\mathbf{B}) &= \sup_{\mathbf{x}:\|\mathbf{x}\|_2=1} \mathbf{x}^\top(\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:} + \epsilon n \tilde{\mathbf{v}}\tilde{\mathbf{v}}^\top)\mathbf{x} \\
&\ge \mathbf{v}^\top(\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:} + |T|\tilde{\mathbf{v}}\tilde{\mathbf{v}}^\top)\mathbf{v} \\
&= \lambda_1(\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:}) + |T|\|\mathbf{v}\|_1^2.
\end{aligned}$$

Observe that

$$\lambda_1(\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:}) = \|\mathbf{A}_{T^c,:}\mathbf{v}\|_2^2 = \sum_{i\in T^c} \langle \mathbf{A}_{i,:}, \mathbf{v}\rangle^2 \le \sum_{i\in T^c} \|\mathbf{A}_{i,:}\|_\infty^2 \|\mathbf{v}\|_1^2 \le (n - |T|)\|\mathbf{v}\|_1^2.$$

Then

$$\lambda_1(\mathbf{B}^\top\mathbf{B}) \ge \left(1 + \frac{0.9\epsilon}{1-\epsilon}\right) \lambda_1(\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:})$$

and so

$$\begin{aligned}
cn^p > \|\mathbf{B}\|_{\mathcal{S}_p}^p &\ge \left(1 + \frac{0.9\epsilon}{1-\epsilon}\right)^{\frac{p}{2}} \lambda_1^{\frac{p}{2}}(\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:}) + \sum_{i\ge 2} \lambda_i^{\frac{p}{2}}(\mathbf{A}_{T^c,:}^\top \mathbf{A}_{T^c,:}) \\
&\ge \left(\left(1 + \frac{0.9\epsilon}{1-\epsilon}\right)^{\frac{p}{2}} - 1\right)(c_2' n)^p + \|\mathbf{A}_{T^c,:}\|_{\mathcal{S}_p}^p \\
&\ge c_3 p\epsilon n^p + (1 - C'\epsilon)\|\mathbf{A}\|_{\mathcal{S}_p}^p,
\end{aligned}$$

whence it follows that

$$\|\mathbf{A}\|_{\mathcal{S}_p}^p \le (c - (c_3 p - cC')\epsilon) n^p,$$

provided that c and ϵ are sufficiently small. □

Algorithm 17 Algorithm for Schatten-p norm testing ($p > 2$)

▷ Lines 1-2 estimate the Frobenius norm of $\mathbf{A}$.

1: Uniformly sample $q_0 = \mathcal{O}(\frac{1}{\epsilon^2})$ entries $\mathbf{A}$, forming vector $\mathbf{y}$.
2: $X \leftarrow \frac{n^2}{q_0}\|\mathbf{y}\|_2^2$. ▷ X is an estimator of $\|\mathbf{A}\|_F^2$.
3: Uniformly sample a $q \times q$ submatrix $\widetilde{\mathbf{A}}'$ with $q = \mathcal{O}(\frac{\log n}{\epsilon})$.
4: **If** $\|\widetilde{\mathbf{A}}'\| \le C_0\sqrt{X}\frac{q}{n}$
5: Output "$\mathbf{A} \in \mathsf{H1}$".
6: **Else**
7: Run Algorithm 14 with $\tau = \Theta(\epsilon^{p/(p-2)}/p)$ and obtain indices $\mathcal{I}_{\mathsf{row}}$ and $\mathcal{I}_{\mathsf{col}}$.
8: $\mathbf{A}_0 \leftarrow \mathbf{A}_{\mathcal{I}_{\mathsf{row}},\mathcal{I}_{\mathsf{col}}}$.
9: $\mathcal{I} \leftarrow \{i \mid \sigma_i(\mathbf{A}_0) > (1+\epsilon/(3p))n(c\epsilon/3)^{1/(p-2)}\}$.
10: **If** $\sum_{i\in\mathcal{I}}\sigma_i^p(\mathbf{A}_0) \ge cn^p$
11: Output "$\mathbf{A} \in \mathsf{H0}$".
12: **Else**
13: Output "$\mathbf{A} \in \mathsf{H1}$".
14: **End If**
15: **End If**

Theorem 45. *Let $p > 2$ be a constant, and c and ϵ be as in Lemma 62. Then Algorithm 17 is a correct algorithm for the Schatten-p norm testing problem under the sampling model with probability at least 0.99. It reads $\mathcal{O}\left(\frac{\log^2 n}{\epsilon^{4p/(p-2)}}\right)$ entries.*

Proof. When $\mathbf{A} \in \mathsf{H0}$, we claim that $\mathsf{srank}(\mathbf{A}) = \mathcal{O}(1)$ which is independent of n and $1/\epsilon$. Otherwise, suppose $\mathsf{srank}(\mathbf{A}) = f(n, 1/\epsilon)$. Then $\|\mathbf{A}\| = \|\mathbf{A}\|_F/\sqrt{\mathsf{srank}(\mathbf{A})} \le n/f(n,1/\epsilon) = o(n)$. In this case, $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ is maximized when the first r singular values are equal to $\|\mathbf{A}\|$, where $r \le n^2/\|\mathbf{A}\|^2$ in order to satisfy $\|\mathbf{A}\|_F \le n$. So the maximal $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ is $r\|\mathbf{A}\|^p \le n^2\|\mathbf{A}\|^{p-2} = o(n^p)$, which leads to a contradiction with $\mathbf{A} \in \mathsf{H0}$. That is, $\mathsf{srank}(\mathbf{A})$ is an absolute constant which is independent of n and $1/\epsilon$, say $4e^2$. Thus, when $\mathbf{A} \in \mathsf{H0}$ we have $\|\mathbf{A}\| = \Theta(n)$ and $\|\mathbf{A}\|_F = \Theta(n)$, because $n \ge \|\mathbf{A}\|_F \ge \|\mathbf{A}\|_{\mathcal{S}_p} \ge c^{1/p}n$.

We note that by sampling q_0 entries from $\mathbf{A}$ and stacking them as vector $\mathbf{y}$, the resulting estimator $X = \frac{n^2}{q_0^2}\|\mathbf{y}\|_2^2$ satisfies $\mathbb{E}[X] = \|\mathbf{A}\|_F^2$ and $\mathsf{Var}[X] \le n^2(n^4/q_0^2)(q_0/n^2) = n^4/q_0$. Taking $q_0 = \mathcal{O}(1/\epsilon^2)$, we have, by Chebyshev's inequality, that

$$\Pr\left[|X - \|\mathbf{A}\|_F^2| > \epsilon n^2\right] \le \frac{n^4/q_0}{\epsilon^2 n^4} \le 0.999.$$

Thus with constant probability, $|X - \|\mathbf{A}\|_F^2| \le \epsilon n^2$.

We argue that uniformly sampling an $\mathcal{O}(\log(n)/\epsilon) \times \mathcal{O}(\log(n)/\epsilon)$ submatrix of $\mathbf{A}$ suffices to distinguish $\mathsf{srank}(\mathbf{A}) \le 4e^2$ v.s. $\mathsf{srank}(\mathbf{A}) > 4c_1e^2$ for a large absolute constant c_1 with a constant probability. To see this, when $\mathsf{srank}(\mathbf{A}) > 4c_1e^2$, let $\mathbf{U}$ be a uniformly random $n \times n$ orthogonal matrix and let $\mathbf{A}'_{\mathsf{row}}$ be the matrix after uniform row sampling of $\mathbf{A}$ of expected cardinality q. Note that $\|\mathbf{A}'_{\mathsf{row}}\| = \|\mathbf{A}'_{\mathsf{row}}\mathbf{U}\|$, and $(\mathbf{A}\mathbf{U})_{i,:} = \mathbf{A}_{i,:}\mathbf{U}$ is uniform on $\|\mathbf{A}_{i,:}\|_2 \cdot \mathbb{S}^{n-1}$. So $\|\mathbf{A}_{i,:}\mathbf{U}\|_\infty^2 \le 2\|\mathbf{A}_{i,:}\mathbf{U}\|_2^2\log(n)/n$ for any fixed i with probability at least $1 - 1/n^2$ by Lemma 56. Therefore,

with probability at least $1 - 1/n$ by union bound over all rows, $\|\mathbf{AU}\|_{\mathsf{col}}^2 \le 2\|\mathbf{A}\|_F^2 \log(n)/n$, where $\|\mathbf{A}\|_{\mathsf{col}}$ represents the maximum ℓ_2 norm among all columns of $\mathbf{A}$. By Lemma 55,

$$\mathbb{E}\|\mathbf{A}'_{\mathsf{row}}\| \le C_1'\sqrt{\frac{q}{n}}\|\mathbf{A}\| + C_2'\sqrt{\log q}\sqrt{\frac{\log n}{n}}\|\mathbf{A}\|_F$$

for absolute constants C_1' and C_2', and by a Markov bound, with probability at least 0.999,

$$\begin{aligned}\|\mathbf{A}'_{\mathsf{row}}\| &\le C_1\sqrt{\frac{q}{n}}\|\mathbf{A}\| + C_2\sqrt{\log q}\sqrt{\frac{\log n}{n}}\|\mathbf{A}\|_F \\ &\le C_1\sqrt{\frac{q}{n}}\frac{\|\mathbf{A}\|_F}{\sqrt{4c_1e^2}} + C_2\sqrt{\log q}\sqrt{\frac{\log n}{n}}\|\mathbf{A}\|_F \quad (\text{since } \mathsf{srank}(\mathbf{A}) > 4c_1e^2)\end{aligned}$$

for absolute constants C_1 and C_2. By a Markov bound, we also have with constant probability that

$$\|\mathbf{A}'_{\mathsf{row}}\|_F^2 \le c\frac{q}{n}\|\mathbf{A}\|_F^2.$$

Conditioning on this event, by applying the same argument on the column sampling of $\mathbf{A}'_{\mathsf{row}}$, we have

$$\begin{aligned}\|\widetilde{\mathbf{A}}'\| &\le C_1\sqrt{\frac{q}{n}}\|\mathbf{A}'_{\mathsf{row}}\| + C_2\sqrt{\log q}\sqrt{\frac{\log n}{n}}\|\mathbf{A}'_{\mathsf{row}}\|_F \\ &\le C_1\sqrt{\frac{q}{n}}\left(C_1\sqrt{\frac{q}{n}}\frac{\|\mathbf{A}\|_F}{\sqrt{4c_1e^2}} + C_2\sqrt{\log q}\sqrt{\frac{\log n}{n}}\|\mathbf{A}\|_F\right) + C_2\sqrt{\log q}\sqrt{\frac{\log n}{n}}\sqrt{\frac{cq}{n}}\|\mathbf{A}\|_F \\ &\le C_0\frac{\|\mathbf{A}\|_F}{\sqrt{4c_1e^2}}\frac{q}{n}, \qquad (\text{because the first term dominates as } q \gg \text{a constant})\end{aligned}$$

where $\widetilde{\mathbf{A}}'$ is the matrix after the column sampling of $\mathbf{A}'_{\mathsf{row}}$, and C_0 is an absolute constant. On the other hand, when $\mathsf{srank}(\mathbf{A}) \le 4e^2$ and $q = \mathcal{O}(\frac{\log n}{\epsilon})$, we have with high probability that

$$\|\widetilde{\mathbf{A}}'\| \ge C\frac{q}{n}\|\mathbf{A}\| = C\frac{q}{n}\frac{\|\mathbf{A}\|_F}{\sqrt{\mathsf{srank}(\mathbf{A})}} \ge C\frac{q}{n}\frac{\|\mathbf{A}\|_F}{\sqrt{4e^2}},$$

where the first inequality holds by applying Lemma 60 twice on row and column sampling (set $\beta = \Theta(\epsilon)$ and $\tau = \Theta(1)$ there). By setting c_1 as a large absolute constant, we have

$$C_0\frac{\|\mathbf{A}\|_F}{\sqrt{4c_1e^2}}\frac{q}{n} < C\frac{q}{n}\frac{\|\mathbf{A}\|_F}{\sqrt{4e^2}}.$$

Thus we can distinguish (a) $\mathsf{srank}(\mathbf{A}) \le 4e^2$ and (b) $\mathsf{srank}(\mathbf{A}) > 4c_1e^2$ by checking $\|\widetilde{\mathbf{A}}'\|$. If we find $\mathsf{srank}(\mathbf{A}) > 4c_1e^2$, then we can safely output "$\mathbf{A} \in \mathsf{H1}$". Therefore, in the following we can assume $\mathsf{srank}(\mathbf{A}) \le 4c_1e^2$.

Recall that, according to Lemma 62, there is a multiplicative gap in the Schatten-p norm between case $\mathsf{H0}$ and case $\mathsf{H1}$. Without loss of generality, we assume $\|\mathbf{A}\|_{\mathcal{S}_p}^p = (1 \pm \epsilon)cn^p$

in the following, which represents the hardest case to distinguish H0 from H1. In that case, $\mathsf{srank}(\mathbf{A}) = \mathcal{O}(1)$ and $\|\mathbf{A}\|_F^2 = \Theta(n^2)$, as we have shown in the beginning of the proof.

We now show that with $\mathrm{poly}(1/\epsilon)$ sampled entries, we can have an estimator which approximates $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ up to $(1 \pm \epsilon)$ factor; therefore, we can distinguish H0 from H1 due to the gap of $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ in the two cases. Consider all singular values of $\mathbf{A}$ which are at most $n/\sqrt{r}$ and consider $\|\mathbf{A}\|_{\mathcal{S}_p}^p$. This is maximized when there are as many singular values as possible that are equal to $n/\sqrt{r}$. Note that there can be at most r singular values of value $n/\sqrt{r}$, since $\|\mathbf{A}\|_F^2 \le n^2$ for $\|\mathbf{A}\|_\infty \le 1$. Therefore, the total contribution of singular values which are no larger than $n/\sqrt{r}$ is at most $n^p r^{1-p/2}$. So if $r = (c\epsilon/3)^{\frac{2}{2-p}}$, this quantity is at most $c\epsilon n^p/3$. Thus all singular values less than $n(c\epsilon/3)^{\frac{1}{p-2}}$ contribute not too much, at most $c\epsilon n^p/3$. For the remaining singular values (i.e., $\sigma_i(\mathbf{A}) > n(c\epsilon/3)^{\frac{1}{p-2}}$), by Theorem 27, with $\mathcal{O}\left(\frac{p^4}{\epsilon^{4p/(p-2)}}\log^2 n\right)$ samples we have

$$\begin{aligned}\left|\sigma_i^2(\mathbf{A}) - \sigma_i^2(\mathbf{A}_0)\right| = \left|\sigma_i(\mathbf{A}^\top\mathbf{A}) - \sigma_i(\mathbf{A}_0^\top\mathbf{A}_0)\right| &\le \|\mathbf{A}^\top\mathbf{A} - \mathbf{A}_0^\top\mathbf{A}_0\| \le \frac{2\epsilon}{3p}\left(\frac{c\epsilon}{3}\right)^{\frac{2}{p-2}}\|\mathbf{A}\|^2 \\ &\le \frac{2\epsilon}{3p}\left(\frac{c\epsilon}{3}\right)^{\frac{2}{p-2}} n^2 < \frac{2\epsilon}{3p}\sigma_i^2(\mathbf{A}),\end{aligned}$$

namely, $\sigma_i^p(\mathbf{A}_0) = (1 \pm \epsilon/3)\sigma_i^p(\mathbf{A})$. Therefore,

$$\sum_{i:\ \sigma_i(\mathbf{A}) > n(c\epsilon/3)^{1/(p-2)}} \sigma_i^p(\mathbf{A}_0) = (1 \pm \epsilon/3) \sum_{i:\ \sigma_i(\mathbf{A}) > n(c\epsilon/3)^{1/(p-2)}} \sigma_i^p(\mathbf{A}) = (1 \pm 2\epsilon/3)\|\mathbf{A}\|_{\mathcal{S}_p}^p,$$

where the last $\subseteq$ holds because all singular values less than $n(c\epsilon/3)^{\frac{1}{p-2}}$ contribute at most $c\epsilon n^p/3$. Let $\mathcal{I} = \{i \mid \sigma_i(\mathbf{A}_0) > (1 + \epsilon/(3p))n(c\epsilon/3)^{1/(p-2)}\}$ and $\mathcal{J} = \{i \mid \sigma_i(\mathbf{A}) > n(c\epsilon/3)^{1/(p-2)}\}$. We note that $\mathcal{I} \subseteq \mathcal{J}$, and that all singular values of $\mathbf{A}$ less than $(1 + \epsilon/(3p))n(c\epsilon/3)^{\frac{1}{p-2}}$ contribute not too much, at most $c\epsilon n^p/2$, by a similar analysis as above. Therefore, those singular values of $\mathbf{A}$ that lie in $\mathcal{J}\backslash\mathcal{I}$ contribute at most $c\epsilon n^p/6$, and by the relation $\sigma_i^p(\mathbf{A}_0) = (1 \pm \epsilon/3)\sigma_i^p(\mathbf{A})$ for all $i \in \mathcal{J}$, those singular values of $\mathbf{A}_0$ that lie in $\mathcal{J}\backslash\mathcal{I}$ contribute at most $c\epsilon n^p/5$. Therefore

$$\sum_{i\in\mathcal{I}} \sigma_i^p(\mathbf{A}_0) = \sum_{i\in\mathcal{J}} \sigma_i^p(\mathbf{A}_0) \pm \frac{c\epsilon n^p}{5} = (1 \pm \epsilon)\|\mathbf{A}\|_{\mathcal{S}_p}^p,$$

as desired. □

Lower bounds. We then prove the lower bound for $p \in [1, 2)$.

Problem 5 (Schatten-p Norm Testing in the Bounded Entry Model for $p \in [1, 2)$)**.** *Let $p \in [1, 2)$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$ with $\|\mathbf{A}\|_\infty \le 1$. For a constant c, the matrix $\mathbf{A}$ satisfies one of the promised properties:*

- H0. $\|\mathbf{A}\|_{\mathcal{S}_p}^p \ge cn^{1+p/2}$*;*
- H1. $\mathbf{A}$ *is ϵ-far from $\|\mathbf{A}\|_{\mathcal{S}_p}^p \ge cn^{1+p/2}$, meaning that it requires changing at least an ϵ-fraction of entries of $\mathbf{A}$ such that $\|\mathbf{A}\|_{\mathcal{S}_p}^p \ge cn^{1+p/2}$.*

The problem is to design a non-adaptive property testing algorithm that outputs H0 *with probability at least* 0.9 *if $\mathbf{A} \in$* H0*, and output* H1 *with probability at least* 0.9 *if $\mathbf{A} \in$* H1*, with the least number of queried entries.*

Suppose that $\mathbf{G} \sim \mathcal{G}(n,n)$ and $\mathbf{O}$ is a random $n \times n$ orthogonal matrix. Consider two distributions $\mathcal{D}_1 = \frac{1+\eta}{\sqrt{n}}\mathbf{G}$ and $\mathcal{D}_2 = \mathbf{O} + \frac{\eta}{\sqrt{n}}\mathbf{G}$, where $\eta > 0$ is a small absolute constant. The following lemma comes from a manuscript of Li et al. [154].

Lemma 63 ([154])**.** *Consider a linear sketch of length m for random matrices drawn from $\mathcal{D}_1$ or $\mathcal{D}_2$. Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be the induced distribution of the linear sketch of $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. There exists $\alpha = \alpha(\eta) \in (0,1)$ such that whenever $m \le \alpha n$, it holds that $d_{TV}(\mathcal{L}_1, \mathcal{L}_2) < 1/10$.*

Theorem 46. *Let $p \in [1,2)$ be a constant. There exist constants $c = c(p)$ and $\epsilon_0 = \epsilon_0(p)$ such that for any $\epsilon \le \epsilon_0$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$, any non-adaptive algorithm that correctly tests* H0 *against* H1 *with probability at least 0.99 must make $\Omega(n)$ queries (i.e., the sketch size is $\Omega(n)$).*

Proof. Consider the hard distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ for Lemma 63. For $p < 2$, it is a well-known fact that with high probability over $\mathbf{G} \sim \mathcal{G}(n,n)$, it holds that $\|\frac{1}{\sqrt{n}}\mathbf{G}\| \le 2(1+o(1))$ and $\|\frac{1}{\sqrt{n}}\mathbf{G}\|_{\mathcal{S}_p}^p \le (1+o(1))c_p n$ for some constant $c_p < 1$ that depends only on p. Hence with high probability, when $\mathbf{A} \sim \mathcal{D}_1$, it holds that $\|\mathbf{A}\|_{\mathcal{S}_p}^p \le (1+o(1))(1+\eta)c_p n$. On the other hand, with high probability, when $\mathbf{A} \sim \mathcal{D}_2$, it follows from the triangle inequality that $\|\mathbf{A}\|_{\mathcal{S}_p}^p \ge (1-(1+o(1))\eta c_p^{1/p})^p n$. Therefore, when η is sufficiently small (depending on p only), there is a constant-factor multiplicative gap in $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ between $\mathcal{D}_1$ and $\mathcal{D}_2$.

Let C be a large constant to be determined. We truncate $\mathcal{D}_1$ and $\mathcal{D}_2$ by applying the map

$$x \mapsto \max\left\{\min\left\{x, \frac{C}{\sqrt{n}}\right\}, -\frac{C}{\sqrt{n}}\right\}$$

entrywise to the matrices, resulting in two new distributions $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$. We claim that with high probability, there remains a constant-factor multiplicative gap in $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ between $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$. It suffices to show that with high probability, truncation incurs only a change of cn for some small constant $c > 0$ in $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ for both $\mathcal{D}_1$ and $\mathcal{D}_2$.

Suppose that $\mathbf{A} \sim \mathcal{D}_1$, and let $\widetilde{\mathbf{A}}$ be the truncated matrix. We can write $\widetilde{\mathbf{A}} = \mathbf{A} + \frac{1}{\sqrt{n}}\mathbf{B}$, where $\mathbf{B}$ is a random matrix with i.i.d. entries following a truncated Gaussian distribution $\tilde{\mathcal{N}}_C(0,1)$ whose probability density function is

$$f_C(t) = (1-p_C)\delta(t) + \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(|t|+C)^2}{2}\right),$$

where $\delta(t)$ is the Dirac delta function and

$$p_C = \frac{2}{\sqrt{2\pi}}\int_C^\infty \exp\left(-\frac{x^2}{2}\right)dx =: \operatorname{erfc}\left(\frac{C}{\sqrt{2}}\right).$$

One can also calculate that

$$m_C := \mathbb{E}|\mathbf{B}_{ij}|^2 \tag{3.56}$$

has a subgaussian decay w.r.t. C. It follows from a Chernoff bound that $\|\mathbf{B}\|_F^2 \le 2m_C n^2$ with high probability. Since $\|\mathbf{B}\|_{\mathcal{S}_p} \le n^{\frac{1}{p}-\frac{1}{2}}\|\mathbf{B}\|_F \le \sqrt{2m_C}n^{\frac{1}{p}+\frac{1}{2}}$, we see that $\|\frac{1}{\sqrt{n}}\mathbf{B}\|_{\mathcal{S}_p} \le \sqrt{2m_C}n^{\frac{1}{p}}$, where the constant factor can be made arbitrarily small by choosing C large enough; that is,

truncating $\mathbf{A} \sim \mathcal{D}_1$ incurs only a constant factor loss (where the constant can be made arbitrarily small) in $\|\mathbf{A}\|_{\mathcal{S}_p}^p$ with probability ≥ 0.999.

Next, suppose that $\mathbf{A}' \sim \mathcal{D}_2$ and we write the truncation as $\tilde{\mathbf{A}}' = \mathbf{A}' + \mathbf{B}'$. It is a classical result [41] that

$$\lim_{n\to\infty} \Pr\{\sqrt{n}\mathbf{O}_{ij} \leq t\} = \Pr_{g\sim\mathcal{N}(0,1)}\{g \leq t\}. \tag{3.57}$$

Observe that $\mathbf{A}'_{ij} \overset{\text{dist}}{=} \mathbf{O}_{ij} + \frac{\eta}{\sqrt{n}}g'$ and $\mathbf{A}_{ij} \overset{\text{dist}}{=} \frac{1}{\sqrt{n}}g + \frac{\eta}{\sqrt{n}}g'$ with the same additive 'noise' $\frac{\eta}{\sqrt{n}}g'$, where $g, g' \sim N(0,1)$ are independent, it follows that $\Pr\{\sqrt{n}\mathbf{A}'_{ij} \leq t\} \to \Pr\{\sqrt{n}\mathbf{A}_{ij} \leq t\}$ for any (fixed) t as $n \to \infty$ (note that $\sqrt{n}\mathbf{A}_{ij} \overset{\text{dist}}{=} (1+\eta)g$ and does not depend on n). Hence each entry of $\mathbf{B}'_{ij}$ is stochastically dominated by $\frac{1}{\sqrt{n}}\tilde{\mathcal{N}}_{C/2}(0,1)$. Similarly to before, $\mathbb{E}\|\mathbf{B}'\|_{\mathcal{S}_p} \leq n^{\frac{1}{p}-\frac{1}{2}}\mathbb{E}\|\mathbf{B}'\|_F \leq \sqrt{m_{C/2}}n^{\frac{1}{p}}$, and thus by Markov's inequality, with probability ≥ 0.999 it holds that $\|\mathbf{B}'\|_{\mathcal{S}_p} \leq 1000\sqrt{m_{C/2}}n^{\frac{1}{p}}$; that is, truncating $\mathbf{A}' \sim \mathcal{D}_2$ incurs only a constant factor loss (where the constant can be made arbitrarily small) in $\|\mathbf{A}'\|_{\mathcal{S}_p}^p$ with probability ≥ 0.999.

Now, the matrices from $\frac{\sqrt{n}}{C}\widetilde{\mathcal{D}_1}$ and $\frac{\sqrt{n}}{C}\widetilde{\mathcal{D}_2}$ have entries bounded by 1, and with high probability, $\|\mathbf{A}\|_{\mathcal{S}_p} \leq c_1 n^{\frac{1}{2}+\frac{1}{p}}$ when $\mathbf{A} \sim \frac{\sqrt{n}}{C}\widetilde{\mathcal{D}_1}$ and $\|\mathbf{A}\|_{\mathcal{S}_p} \geq c_2 n^{\frac{1}{2}+\frac{1}{p}}$ when $\mathbf{A} \sim \frac{\sqrt{n}}{C}\widetilde{\mathcal{D}_2}$, for constants $c_1 < c_2$ (depending on η and C). Our result of the theorem would follow immediately from Theorem 46 once we establish that with high probability, a random matrix from $\frac{\sqrt{n}}{C}\widetilde{\mathcal{D}_1}$ is ϵ-far from having Schatten p-norm at least $c_2 n^{\frac{1}{2}+\frac{1}{p}}$. Indeed, let $\mathbf{E}$ denote the pertubation to $\mathbf{A}$ such that $\|\mathbf{E}\|_0 \leq \epsilon n^2$ and $\|\mathbf{A}+\mathbf{E}\|_\infty \leq 1$. Since $\|\mathbf{A}\|_\infty \leq 1$, it must hold that $\|\mathbf{E}\|_\infty \leq 2$. Thus $\|\mathbf{E}\|_{\mathcal{S}_p} \leq n^{\frac{1}{p}-\frac{1}{2}}\|\mathbf{E}\|_F \leq 2\sqrt{\epsilon}n^{\frac{1}{p}+\frac{1}{2}}$. When ϵ is sufficiently small, it is easy to see via triangle inequality that there remains a constant-factor gap between $\|\mathbf{A}+\mathbf{E}\|_{\mathcal{S}_p}$ and $c_2 n^{\frac{1}{2}+\frac{1}{p}}$. □

Chapter 4

Learning with Deep Neural Networks

4.1 Deep Neural Networks with Multi-Branch Architectures

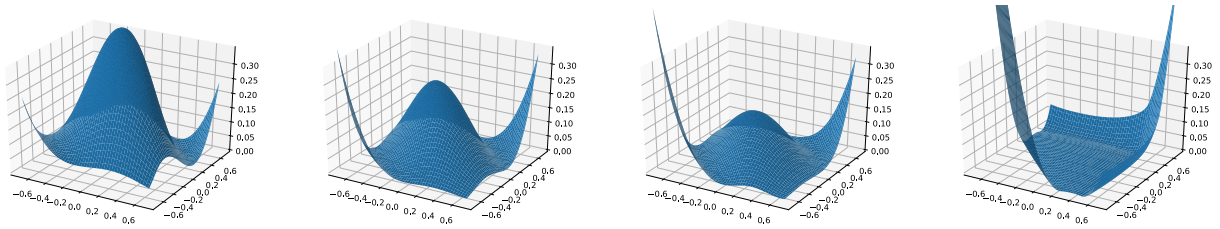
4.1.1 Introduction

Deep neural networks are a central object of study in machine learning, computer vision, and many other domains. They have substantially improved over conventional learning algorithms in many areas, including speech recognition, object detection, and natural language processing. The focus of this work is to investigate the duality gap of deep neural networks. The duality gap is the discrepancy between the optimal values of primal and dual problems. While it has been well understood for convex optimization, little is known for non-convex problems. A smaller duality gap in relative value typically implies that the problem itself is less non-convex intrinsically, and thus is easier to optimize.[1] Our results establish that: *Deep neural networks with multi-branch architecture have small duality gap in relative value.*

Our study is motivated by the computational difficulties of deep neural networks due to its non-convex nature. While many works have witnessed the power of local search algorithms for deep neural networks [49], these algorithms typically converge to a suboptimal solution in the worst cases according to various empirical observations [208]. It is reported that for a single-hidden-layer neural network, when the number of hidden units is small, stochastic gradient descent may get easily stuck at the poor local minima [90, 199]. Furthermore, there is significant evidence indicating that when the networks are deep enough, bad saddle points do exist [6] and might be hard to escape [6, 31, 39, 73].

Given the computational obstacles, several efforts have been devoted to designing new architectures to alleviate the above issues, including over-parametrization [10, 50, 156, 179, 211] and multi-branch architectures [64, 120, 222, 232, 245]. Empirically, increasing the number of hidden units of a single-hidden-layer network encourages the first-order methods to converge to a global solution, which probably supports the folklore that the loss surface of a wider network looks more "convex" (see Figure 4.1). Furthermore, several recently proposed architectures, including ResNeXt [245], Inception [222], Xception [64], SqueezeNet [120] and Wide ResNet [249] are

[1] Throughout the paper, we discuss the duality gap w.r.t. the Lagrangian function, rather than the augmented Lagrangian function as in Chapter 11 of [195] where the duality gap is always zero.



(a) $I = 10$. (b) $I = 30$. (c) $I = 70$. (d) $I = 1,000$.

Figure 4.1: The loss surface of one-hidden-layer ReLU network projected onto a 2-d plane, which is spanned by three points to which the SGD algorithm converges according to three different initialization seeds. It shows that as the number of hidden neurons I increases, the landscape becomes less non-convex.

based on having multiple branches and have demonstrated substantial improvement over many of the existing models in many applications. In this work, we show that one cause for such success is due to the fact that the loss of multi-branch network is less non-convex in terms of duality gap.

Why is duality gap a measure of *intrinsic* non-convexity? Although some highly non-convex problems such as PCA and quadratic programming may have small/zero duality gap, we argue that the duality gap is a measure of *intrinsic* non-convexity of an optimization problem. There are two reasons for such an argument. a) The optimal value of the dual problem is equal to the optimal value of the convex relaxation of the primal problem. Hereby, the convex relaxation is the problem arising by replacing the non-convex objective with its convex closure and replacing the non-convex feasible set with its closed convex hull. Therefore, the duality gap measures the discrepancy between the optimal values of primal problem and its convex relaxation (Taking convex problems as an example, the duality gap is zero in most cases). When the duality gap is small, one can solve the convex relaxation problem whose solution is guaranteed to being close to the solution of primal problem. b) We show in our main result that the duality gap is a lower bound of the discrepancy between objective and its convex relaxation[2] (see Theorem 47 for the case of $I = 1$). So a smaller duality gap implies a possibly smaller discrepancy between objective and its convex relaxation.

Notation. We will use bold capital letter to represent matrix and lower-case letter to represent scalar. Specifically, let $\mathbf{I}$ be the identity matrix and denote by $\mathbf{0}$ the all-zero matrix. Let $\{\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}} : i = 1, 2, ..., H\}$ be a set of network parameters, each of which represents the connection weights between the i-th and $(i+1)$-th layers of neural network. We use $\mathbf{W}_{:,t} \in \mathbb{R}^{n_1 \times 1}$ to indicate the t-th column of $\mathbf{W}$. We will use $\sigma_i(\mathbf{W})$ to represent the i-th largest singular value of matrix $\mathbf{W}$. Given skinny SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ of matrix $\mathbf{W}$, we denote by $\mathsf{svd}_r(\mathbf{W}) = \mathbf{U}_{:,1:r}\mathbf{\Sigma}_{1:r,1:r}\mathbf{V}^T_{:,1:r}$ the truncated SVD of $\mathbf{W}$ to the first r singular values. For matrix norms, denote by $\|\mathbf{W}\|_{\mathcal{S}_H} = \left(\sum_i \sigma_i^H(\mathbf{W})\right)^{1/H}$ the matrix Schatten-H norm. Nuclear norm and Frobenius norm are special cases of Schatten-H norm: $\|\mathbf{W}\|_* = \|\mathbf{W}\|_{\mathcal{S}_1}$ and $\|\mathbf{W}\|_F = \|\mathbf{W}\|_{\mathcal{S}_2}$. We use $\|\mathbf{W}\|$ to represent the matrix operator norm, i.e., $\|\mathbf{W}\| = \sigma_1(\mathbf{W})$, and denote by $\mathsf{rank}(\mathbf{W})$ the rank of matrix $\mathbf{W}$.

[2]Note that the convex relaxation of objective is different from the convex relaxation of primal non-convex problem which requires convexification operation on both objective and constraint.

Denote by $\mathsf{Row}(\mathbf{W})$ the span of rows of $\mathbf{W}$. Let $\mathbf{W}^\dagger$ be the Moore-Penrose pseudo-inverse of $\mathbf{W}$.

For convex matrix function $K(\cdot)$, we denote by $K^*(\mathbf{\Lambda}) = \max_{\mathbf{M}} \langle \mathbf{\Lambda}, \mathbf{M} \rangle - K(\mathbf{M})$ the conjugate function of $K(\cdot)$ and $\partial K(\cdot)$ the sub-differential. We use $\mathsf{diag}(\sigma_1, ..., \sigma_r)$ to represent a $r \times r$ diagonal matrix with diagonal entries $\sigma_1, ..., \sigma_r$. Let $d_{\min} = \min\{d_i : i = 1, 2, ..., H-1\}$, and $[I] = \{1, 2, ..., I\}$. For any two matrices $\mathbf{A}$ and $\mathbf{B}$ of matching dimensions, we denote by $[\mathbf{A}, \mathbf{B}]$ the concatenation of $\mathbf{A}$ and $\mathbf{B}$ along the row and $[\mathbf{A}; \mathbf{B}]$ the concatenation of two matrices along the column.

4.1.2 Our results on optimization

Duality gap of multi-branch neural networks

We first study the duality gap of neural networks in a classification setting. We show that the wider the network is, the smaller the duality gap becomes.

Network Setup. The output of our network follows from a multi-branch architecture (see Figure 4.2):

$$f(\mathbf{w}; \mathbf{x}) = \frac{1}{I} \sum_{i=1}^{I} f_i(\mathbf{w}_{(i)}; \mathbf{x}), \ \mathbf{w}_{(i)} \in \mathcal{W}_i,$$

where $\mathcal{W}_i$ is a convex set, $\mathbf{w}$ is the concatenation of all network parameters $\{\mathbf{w}_{(i)}\}_{i=1}^I$, $\mathbf{x} \in \mathbb{R}^{d_0}$ is the input instance, $\{\mathcal{W}_i\}_{i=1}^I$ is the parameter space, and $f_i(\mathbf{w}_{(i)}; \cdot)$ represents an $\mathbb{R}^{d_0} \to \mathbb{R}$ continuous mapping by a sub-network which is allowed to have *arbitrary* architecture such as convolutional and recurrent neural networks. As an example, $f_i(\mathbf{w}_{(i)}; \cdot)$ can be in the form of a H_i-layer feed-forward sub-network:

$$\begin{aligned} f_i(\mathbf{w}_{(i)}; \mathbf{x}) &= \mathbf{w}_i^\top \psi_{H_i}(\mathbf{W}_{H_i}^{(i)} ... \psi_1(\mathbf{W}_1^{(i)} \mathbf{x})) \in \mathbb{R}, \\ \mathbf{w}_{(i)} &= [\mathbf{w}_i; \mathsf{vec}(\mathbf{W}_1^{(i)}); ...; \mathsf{vec}(\mathbf{W}_{H_i}^{(i)})] \in \mathbb{R}^{p_i}. \end{aligned}$$

Hereby, the functions $\psi_k(\cdot), k = 1, 2, ..., H_i$ are allowed to encode *arbitrary* form of continuous element-wise non-linearity (and linearity) after each matrix multiplication, such as sigmoid, rectification, convolution, while the number of layers H_i in each sub-network can be *arbitrary* as well. When $H_i = 1$ and $d_{H_i} = 1$, i.e., each sub-network in Figure 4.2 represents one hidden unit, the architecture $f(\mathbf{w}; \mathbf{x})$ reduces to a one-hidden-layer network. We apply the so-called *τ-hinge loss* [17, 28] on the top of network output for label $y \in \{-1, +1\}$:

$$\ell_\tau(\mathbf{w}; \mathbf{x}, y) := \max\left(0, 1 - \frac{y \cdot f(\mathbf{w}; \mathbf{x})}{\tau}\right), \quad \tau > 0. \tag{4.1}$$

The τ-hinge loss has been widely applied in active learning of classifiers and margin based learning [17, 28]. When $\tau = 1$, it reduces to the classic hinge loss [50, 143, 157].

We make the following assumption on the margin parameter τ, which states that the parameter τ is sufficiently large.

Assumption 1 (Parameter τ)**.** *For sample $(\mathbf{x}, y)$ drawn from distribution $\mathcal{P}$, we have $\tau > y \cdot f(\mathbf{w}; \mathbf{x})$ for all $\mathbf{w} \in \mathcal{W}_1 \times \mathcal{W}_2 \times ... \times \mathcal{W}_I$ with probability measure* 1.

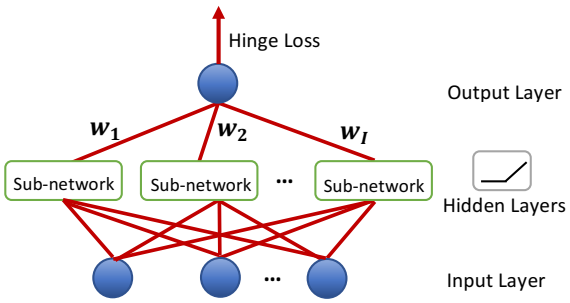


Figure 4.2: Multi-branch architecture, where the sub-networks are allowed to have arbitrary architectures, depths, and continuous activation functions. Hereby, I represents the number of branches. In the extreme case when the sub-network is chosen to have a single neuron, the multi-branch architecture reduces to a single-hidden-layer neural network and the I represents the network width.

We further empirically observe that using smaller values of the parameter τ and other loss functions support our theoretical result as well. It is an interesting open question to extend our theory to more general losses in the future.

To study how close these generic neural network architectures approach the family of convex functions, we analyze the duality gap of minimizing the risk w.r.t. the loss (4.1) with an extra regularization constraint. The normalized duality gap is a measure of intrinsic non-convexity of a given function [36]: the gap is zero when the given function itself is convex, and is large when the loss surface is far from the convexity intrinsically. Typically, the closer the network approaches to the family of convex functions, the easier we can optimize the network.

Multi-Branch Architecture. Our analysis of multi-branch neural networks is built upon tools from non-convex geometric analysis — Shapley–Folkman lemma. Basically, the Shapley–Folkman lemma states that the sum of constrained non-convex functions is close to being convex. A neural network is an ideal target to apply this lemma to: the width of network is associated with the number of summand functions. So intuitively, the wider the neural network is, the smaller the duality gap will be. In particular, we study the following non-convex problem concerning the population risk:

$$\min_{\mathbf{w}\in\mathcal{W}_1\times\ldots\times\mathcal{W}_I} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell_\tau(\mathbf{w};\mathbf{x},y)], \quad \text{s.t.} \quad \frac{1}{I}\sum_{i=1}^{I} h_i(\mathbf{w}_{(i)}) \le K, \tag{4.2}$$

where $h_i(\cdot), i \in [I]$ are convex regularization functions, e.g., the weight decay, and K can be arbitrary such that the problem is feasible. Correspondingly, the dual problem of problem (4.2) is a one-dimensional convex optimization problem:[3]

$$\begin{aligned} &\max_{\lambda\ge 0} \mathcal{Q}(\lambda) - \lambda K, \\ &\mathcal{Q}(\lambda) := \inf_{\mathbf{w}\in\mathcal{W}_1\times\ldots\times\mathcal{W}_I} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[\ell_\tau(\mathbf{w};\mathbf{x},y)] + \frac{\lambda}{I}\sum_{i=1}^{I} h_i(\mathbf{w}_{(i)}). \end{aligned} \tag{4.3}$$

[3]Although problem (4.3) is convex, it does not necessarily mean the problem can be solved easily. This is because computing $\mathcal{Q}(\lambda)$ is a hard problem. So rather than trying to solve the convex dual problem, our goal is to study the duality gap in order to understand the degree of non-convexity of the problem.

Before proceeding, we first define some notations to be used in our main results. For $\widetilde{\mathbf{w}} \in \mathcal{W}_i$, denote by

$$\widetilde{f}_i(\widetilde{\mathbf{w}}) := \inf_{a^j, \mathbf{w}_{(i)}^j \in \mathcal{W}_i} \sum_{j=1}^{p_i+2} a^j \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}} \left(1 - \frac{y \cdot f_i(\mathbf{w}_{(i)}^j; \mathbf{x})}{\tau}\right),$$
$$\text{s.t.} \quad \widetilde{\mathbf{w}} = \sum_{j=1}^{p_i+2} a^j \mathbf{w}_{(i)}^j, \sum_{j=1}^{p_i+2} a^j = 1, a^j \geq 0.$$

This represents the convex relaxation of the i-th summand term $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[1 - y \cdot f_i(\cdot; \mathbf{x})/\tau]$ in the objective, because the epigraph of $\widetilde{f}_i$ is exactly the convex hull of epigraph of $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[1 - y \cdot f_i(\cdot; \mathbf{x})/\tau]$ by the definition of $\widetilde{f}_i$. For $\widetilde{\mathbf{w}} \in \mathcal{W}_i$, we also define

$$\widehat{f}_i(\widetilde{\mathbf{w}}) := \inf_{\mathbf{w}_{(i)} \in \mathcal{W}_i} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}} \left(1 - \frac{y \cdot f_i(\mathbf{w}_{(i)}; \mathbf{x})}{\tau}\right), \quad \text{s.t.} \quad h_i(\mathbf{w}_{(i)}) \leq h_i(\widetilde{\mathbf{w}}).$$

This is a "restricted" version of the i-th summand term $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}[1 - y \cdot f_i(\mathbf{w}_{(i)}; \mathbf{x})/\tau]$ to the hard constraint $h_i(\mathbf{w}_{(i)}) \leq h_i(\widetilde{\mathbf{w}})$.

Our main results for multi-branch neural networks are as follows:

Theorem 47. *Denote by* $\inf(\mathbf{P})$ *the minimum of primal problem* (4.2) *and* $\sup(\mathbf{D})$ *the maximum of dual problem* (4.3). *Let* $\Delta_i := \sup_{\mathbf{w}\in\mathcal{W}_i} \left\{\widehat{f}_i(\mathbf{w}) - \widetilde{f}_i(\mathbf{w})\right\} \geq 0$ *and* $\Delta_{worst} := \max_{i\in[I]} \Delta_i$. *Suppose* $\mathcal{W}_i$*'s are compact and both* $f_i(\mathbf{w}_{(i)}; \mathbf{x})$ *and* $h_i(\mathbf{w}_{(i)})$ *are continuous w.r.t.* $\mathbf{w}_{(i)}$. *If there exists at least one feasible solution of problem* (**P**)*, then under Assumption 1 the duality gap w.r.t. problems* (4.2) *and* (4.3) *can be bounded by*

$$0 \leq \frac{\inf(\mathbf{P}) - \sup(\mathbf{D})}{\Delta_{worst}} \leq \frac{2}{I}.$$

Remark 4. *Note that* Δ_i *measures the divergence between the function value of* $\widehat{f}_i$ *and its convex relaxation* $\widetilde{f}_i$. *The constant* Δ_{worst} *is the maximal divergence among all sub-networks, which grows slowly with the increase of* I. *This is because* Δ_{worst} *only measures the divergence of* one branch. *The normalized duality gap* $(\inf(\mathbf{P}) - \sup(\mathbf{D}))/\Delta_{worst}$ *has been widely used before to measure the degree of non-convexity of optimization problems [36, 38, 75, 83, 226]. Such a normalization avoids trivialities in characterizing the degree of non-convexity: scaling the objective function by any constant does not change the value of normalized duality gap.*

Remark 5. *Even though Theorem 47 is in the form of population risk, the conclusion still holds for the* empirical loss *as well. This can be achieved by setting the marginal distribution* $\mathcal{P}_{\mathbf{x}}$ *as the uniform distribution on a finite set and* $\mathcal{P}_y$ *as the corresponding labels uniformly distributed on the same finite set.*

Remark 6. *Setting* K *in problem* (4.2) *infinitely large implies that Theorem 47 holds for unconstrained deep neural networks as well.*

Inspiration for architecture designs. Theorem 47 shows that the duality gap of deep network shrinks when the width I is large; when $I \to +\infty$, surprisingly, deep network is as easy as a convex

optimization, as the gap is zero. An intuitive explanation is that the large number of randomly initialized hidden units represent all possible features. Thus the optimization problem involves just training the top layer of the network, which is convex. Our result encourages a class of network architectures with multiple branches and supports some of the most successful architectures in practice, such as Inception [222], Xception [64], ResNeXt [245], SqueezeNet [120], Wide ResNet [249], Shake-Shake regularization [86] — all of which benefit from the split-transform-merge behaviour as shown in Figure 4.2.

Strong duality of linear neural networks

In this section, we show that the duality gap is zero if the activation function is linear. Deep linear neural network has received significant attention in recent years [131, 161, 204, 264] because of its simple formulation[4] and its connection to non-linear neural networks.

Network Setup. We discuss the strong duality of regularized deep linear neural networks of the form

$$(\mathbf{W}_1^*, ..., \mathbf{W}_H^*) = \underset{\mathbf{W}_1,...,\mathbf{W}_H}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right], \quad (4.4)$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d_0 \times n}$ is the given instance matrix, $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \mathbb{R}^{d_H \times n}$ is the given label matrix, and $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}, i \in [I]$ represents the weight matrix in each linear layer. We mention that (a) while the linear operation is simple matrix multiplications in problem (4.4), it can be easily extended to other linear operators, e.g., the convolutional operator or the linear operator with the bias term, by properly involving a group of kernels in the variable $\mathbf{W}_i$ [105]. (b) The regularization terms in problem (4.4) are of common interest, e.g., see [105]. When $H = 2$, our regularization terms reduce to $\frac{1}{2}\|\mathbf{W}_i\|_F^2$, which is well known as the weight-decay or Tikhonov regularization. (c) The regularization parameter γ is the same for each layer since we have no further information on the preference of layers.

Our analysis leads to the following guarantees for the deep linear neural networks.

Theorem 48. *Denote by* $\widetilde{\mathbf{Y}} := \mathbf{Y}\mathbf{X}^\dagger\mathbf{X} \in \mathbb{R}^{d_H \times n}$ *and* $d_{\min} := \min\{d_1, ..., d_{H-1}\} \leq \min\{d_0, d_H, n\}$. *Let* $0 \leq \gamma < \sigma_{\min}(\widetilde{\mathbf{Y}})$ *and* $H \geq 2$, *where* $\sigma_{\min}(\widetilde{\mathbf{Y}})$ *stands for the minimal non-zero singular value of* $\widetilde{\mathbf{Y}}$. *Then the strong duality holds for deep linear neural network* (4.4). *In other words, the optimum of problem* (4.4) *is the same as its convex dual problem*

$$\mathbf{\Lambda}^* = \underset{\mathsf{Row}(\mathbf{\Lambda}) \subseteq \mathsf{Row}(\mathbf{X})}{\operatorname{argmax}} -\frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda}\|_{d_{\min}}^2 + \frac{1}{2}\|\mathbf{Y}\|_F^2, \quad \text{s.t.} \quad \|\mathbf{\Lambda}\| \leq \gamma, \quad (4.5)$$

where $\|\cdot\|_{d_{\min}}^2 = \sum_{i=1}^{d_{\min}} \sigma_i^2(\cdot)$ *is a convex function. Moreover, the optimal solutions of primal problem* (4.4) *can be obtained from the dual problem* (4.5) *in the following way: let* $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$ *be the skinny SVD of matrix* $\mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$, *then* $\mathbf{W}_i^* = [\mathbf{\Sigma}^{1/H}, \mathbf{0}; \mathbf{0}, \mathbf{0}] \in \mathbb{R}^{d_i \times d_{i-1}}$ *for* $i = 2, 3, ..., H-1$, $\mathbf{W}_H^* = [\mathbf{U}\mathbf{\Sigma}^{1/H}, \mathbf{0}] \in \mathbb{R}^{d_H \times d_{H-2}}$ *and* $\mathbf{W}_1^* = [\mathbf{\Sigma}^{1/H}\mathbf{V}^T; \mathbf{0}]\mathbf{X}^\dagger \in \mathbb{R}^{d_1 \times d_0}$ *is a globally optimal solution to problem* (4.4).

[4] Although the expressive power of deep linear neural networks and three-layer linear neural networks are the same, the analysis of landscapes of two models are significantly different, as pointed out by [131, 161].

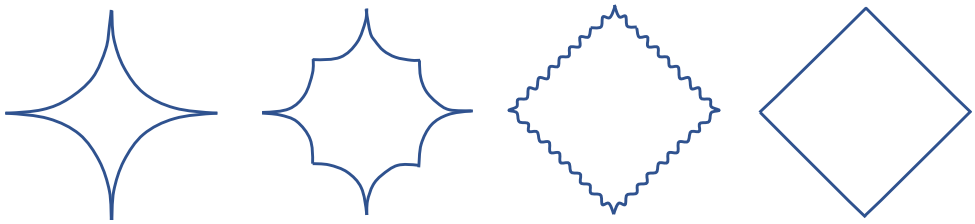


Figure 4.3: Visualization of Shapley-Folkman lemma. **The first figure:** an $\ell_{1/2}$ ball. **The second and third figures:** the averaged Minkowski sum of two and ten $\ell_{1/2}$ balls. **The fourth figure:** the convex hull of $\ell_{1/2}$ ball (the Minkowski average of infinitely many $\ell_{1/2}$ balls). It show that with the number of $\ell_{1/2}$ balls to be averaged increasing, the Minkowski average tends to be more convex.

The regularization parameter γ cannot be too large in order to avoid underfitting. Our result provides a suggested upper bound $\sigma_{\min}(\widetilde{\mathbf{Y}})$ for the regularization parameter, where oftentimes $\sigma_{\min}(\widetilde{\mathbf{Y}})$ characterizes the level of random noise. When $\gamma = 0$, our analysis reduces to the *un-regularized deep linear neural network*, a model which has been widely studied in [131, 161].

Theorem 48 implies the followig result on the landscape of deep linear neural networks: the regularized deep learning can be converted into an equivalent convex problem by dual. To the best of our knowledge, this is the first result on the strong duality of linear neural networks. We note that the strong duality rarely happens in the non-convex optimization: matrix completion [26], Fantope [182], and quadratic optimization with two quadratic constraints [33] are among the few paradigms that enjoy the strong duality. For deep networks, the effectiveness of convex relaxation has been observed empirically in [11, 264], but much remains unknown for the theoretical guarantees of the relaxation. Our work shows strong duality of regularized deep linear neural networks and provides an alternative approach to overcome the computational obstacles due to the non-convexity: one can apply convex solvers, e.g., the Douglas–Rachford algorithm,[5] for problem (4.5) and then conduct singular value decomposition to compute the weights $\{\mathbf{W}_i^*\}_{i=1}^H$ from $\mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$. In addition, our result inherits the benefits of convex analysis. The vast majority results on deep learning study the generalization error or expressive power by analyzing its complicated non-convex form [178, 250, 263]. In contrast, with strong duality one can investigate various properties of deep linear networks with much simpler convex form.

4.1.3 Our techniques

In this section, we present our techniques and proof sketches of Theorems 47 and 48.

(a) Shapley-Folkman lemma. The proof of Theorem 47 is built upon the Shapley-Folkman lemma [36, 75, 83, 216], which characterizes a convexification phenomenon concerning the average of multiple sets and is analogous to the central limit theorem in the probability theory. Consider the averaged Minkowski sum of I sets $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_I$ given by $\{I^{-1}\sum_{j\in[I]} a_j : a_j \in \mathcal{A}_j\}$. Intuitively, the lemma states that $\rho(I^{-1}\sum_{j\in[I]}\mathcal{A}_j) \to 0$ as $I \to +\infty$, where $\rho(\cdot)$ is a

[5] Grussler et al. [99] provided a fast algorithm to compute the proximal operators of $\frac{1}{2}\|\cdot\|_{d_{\min}}^2$. Hence, the Douglas–Rachford algorithm can find the global solution up to an ϵ error in function value in time $\mathsf{poly}(1/\epsilon)$ [114].

metric of the non-convexity of a set (see Figure 4.3 for visualization). We apply this lemma to the optimization formulation of deep neural networks. Denote by *augmented epigraph* the set $\{(h(\mathbf{w}), \ell(\mathbf{w})) : \text{all possible choices of } \mathbf{w}\}$, where h is the constraint and ℓ is the objective function in the optimization problem. The key observation is that the augmented epigraph of neural network loss with multi-branch architecture can be expressed as the Minkowski average of augmented epigraphs of all branches. Thus we obtain a natural connection between an optimization problem and its corresponding augmented epigraph. Applying Shapley-Folkman lemma to the augmented epigraph leads to a characteristic of non-convexity of the deep neural network.

(b) Variational form. The proof of Theorem 48 is built upon techniques (b), (c), and (d). In particular, problem (4.4) is highly non-convex due to its multi-linear form over the optimized variables $\{\mathbf{W}_i\}_{i=1}^H$. Fortunately, we are able to analyze the problem by grouping $\mathbf{W}_H\mathbf{W}_{H-1}...\mathbf{W}_1\mathbf{X}$ together and converting the original non-convex problem in terms of the separate variables $\{\mathbf{W}_i\}_{i=1}^H$ to a convex optimization with respect to the new grouping variable $\mathbf{W}_H\mathbf{W}_{H-1}...\mathbf{W}_1\mathbf{X}$. This typically requires us to represent the objective function of (4.4) as a convex function of $\mathbf{W}_H\mathbf{W}_{H-1}...\mathbf{W}_1$. To this end, we prove that $\|\mathbf{W}_H\mathbf{W}_{H-1}...\mathbf{W}_1\mathbf{X}\|_* = \min_{\mathbf{W}_1,...,\mathbf{W}_H} \frac{1}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right]$. So the objective function in problem (4.4) has an equivalent form

$$\min_{\mathbf{W}_1,...,\mathbf{W}_H} \frac{1}{2}\|\mathbf{Y} - \mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \gamma\|\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_*. \tag{4.6}$$

This observation enables us to represent the optimization problem as a convex function of the output of a neural network. Therefore, we can analyze the non-convex problem by applying powerful tools from convex analysis.

(c) Reduction to low-rank approximation. Our results of strong duality concerning problem (4.6) are inspired by the problem of low-rank matrix approximation:

$$\min_{\mathbf{W}_1,...,\mathbf{W}_H} \frac{1}{2}\|\mathbf{Y} - \mathbf{\Lambda}^* - \mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_F^2. \tag{4.7}$$

We know that all local solutions of (4.7) are globally optimal [26, 131, 161]. To analyze the more general regularized problem (4.4), our main idea is to reduce problem (4.6) to the form of (4.7) by Lagrangian function. In other words, the Lagrangian function of problem (4.6) should be of the form (4.7) for a fixed Lagrangian variable $\mathbf{\Lambda}^*$, which we will construct later in subsection (d). While some prior works attempted to apply a similar reduction, their conclusions either depended on unrealistic conditions on local solutions, e.g., all local solutions are rank-deficient [99, 105], or their conclusions relied on strong assumptions on the objective functions, e.g., that the objective functions are twice-differentiable [105], which do not apply to the non-smooth problem (4.6). Instead, our results bypass these obstacles by formulating the strong duality of problem (4.6) as the existence of a dual certificate $\mathbf{\Lambda}^*$ satisfying certain dual conditions. Roughly, the dual conditions state that the optimal solution $(\mathbf{W}_1^*, \mathbf{W}_2^*, ..., \mathbf{W}_H^*)$ of problem (4.6) is locally optimal to problem (4.7). On one hand, by the above-mentioned properties of problem (4.7), $(\mathbf{W}_1^*, ..., \mathbf{W}_H^*)$ globally minimizes the Lagrangian function when $\mathbf{\Lambda}$ is fixed to $\mathbf{\Lambda}^*$. On the other hand, by the

convexity of nuclear norm, for the fixed $(\mathbf{W}_1^*, ..., \mathbf{W}_H^*)$ the Lagrangian variable $\mathbf{\Lambda}^*$ globally optimize the Lagrangian function. Thus $(\mathbf{W}_1^*, ..., \mathbf{W}_H^*, \mathbf{\Lambda}^*)$ is a primal-dual saddle point of the Lagrangian function of problem (4.6). The desired strong duality is a straightforward result from this argument.

(d) Dual certificate. The remaining proof is to construct a dual certificate $\mathbf{\Lambda}^*$ such that the dual conditions hold true. The challenge is that the dual conditions impose several constraints simultaneously on the dual certificate, making it hard to find a desired certificate. This is why progress on the dual certificate has focused on convex programming. To resolve the issue, we carefully choose the certificate as an appropriate scaling of subgradient of nuclear norm around a low-rank solution, where the nuclear norm follows from our regularization term in technique (b). Although the nuclear norm has infinitely many subgradients, we prove that our construction of dual certificate obeys all desired dual conditions. Putting techniques (b), (c), and (d) together, our proof of strong duality is completed.

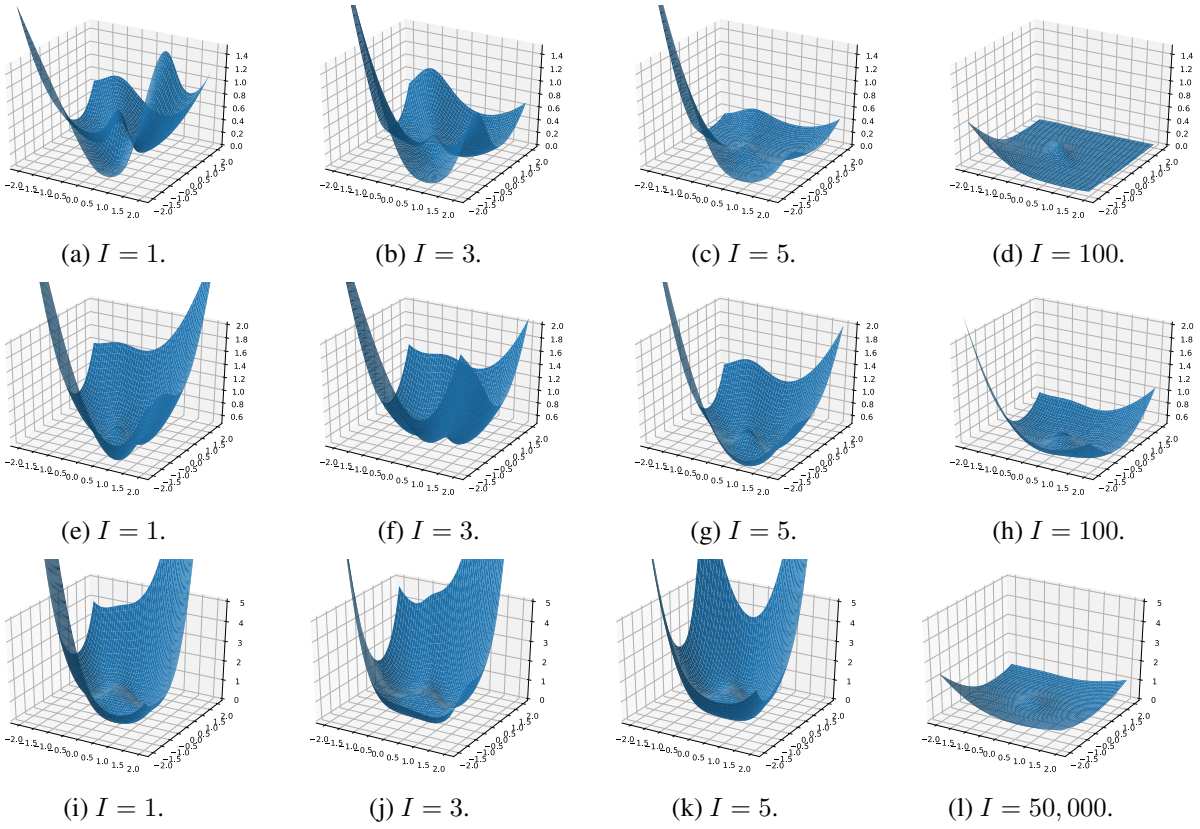
4.1.4 Experimental results

Visualization of loss landscape

Experiments on Synthetic Datasets. We first show that over-parametrization results in a less non-convex loss surface for a synthetic dataset. The dataset consists of $1,000$ examples in $\mathbb{R}^{10}$ whose labels are generated by an underlying one-hidden-layer ReLU network $f(\mathbf{x}) = \sum_{i=1}^{I} \mathbf{w}_{i,2}^* [\mathbf{W}_{i,1}^* \mathbf{x}]_+$ with 11 hidden neurons [199]. We make use of the visualization technique employed by [147] to plot the landscape, where we project the high-dimensional hinge loss ($\tau = 1$) landscape onto a 2-d plane spanned by three points. These points are found by running the SGD algorithm with three different initializations until the algorithm converges. As shown in Figure 4.1, the landscape exhibits strong non-convexity with lots of local minima in the under-parameterized case $I = 10$. But as I increases, the landscape becomes more convex. In the extreme case, when there are $1,000$ hidden neurons in the network, no non-convexity can be observed on the landscape.

Experiments on MNIST and CIFAR-10. We next verify the phenomenon of over-parametrization on MNIST [144] and CIFAR-10 [140] datasets. For both datasets, we follow the standard preprocessing step that each pixel is normalized by subtracting its mean and dividing by its standard deviation. We do not apply data augmentation. For MNIST, we consider a single-hidden-layer network defined as: $f(\mathbf{x}) = \sum_{i=1}^{I} \mathbf{W}_{i,2} [\mathbf{W}_{i,1} \mathbf{x}]_+$, where $\mathbf{W}_{i,1} \in \mathbb{R}^{h \times d}$, $\mathbf{W}_{i,2} \in \mathbb{R}^{10 \times h}$, d is the input dimension, h is the number of hidden neurons, and I is the number of branches, with $d = 784$ and $h = 8$. For CIFAR-10, in addition to considering the exact same one-hidden-layer architecture, we also test a deeper network containing 3 hidden layers of size 8-8-8, with ReLU activations and $d = 3,072$. We apply 10-class hinge loss on the top of the output of considered networks.

Figure 4.4 shows the changes of landscapes when I increases from 1 to 100 for MNIST, and from 1 to $50,000$ for CIFAR-10, respectively. When there is only one branch, the landscapes have strong non-convexity with many local minima. As the number of branches I increases, the landscape becomes more convex. When $I = 100$ for 1-hidden-layer networks on MNIST and CIFAR-10, and $I = 50,000$ for 3-hidden-layer network on CIFAR-10, the landscape is almost convex.



(a) $I = 1$. (b) $I = 3$. (c) $I = 5$. (d) $I = 100$.

(e) $I = 1$. (f) $I = 3$. (g) $I = 5$. (h) $I = 100$.

(i) $I = 1$. (j) $I = 3$. (k) $I = 5$. (l) $I = 50,000$.

Figure 4.4: **Top Row:** Landscape of one-hidden-layer network on MNIST. **Middle Row:** Landscape of one-hidden-layer network on CIFAR-10. **Bottom Row:** Landscape of three-hidden-layer, multi-branch network on CIFAR-10 dataset. From left to right, the landscape looks less non-convex.

Frequency of hitting global minimum

To further analyze the non-convexity of loss surfaces, we consider various one-hidden-layer networks, where each network was trained 100 times using different initialization seeds under the setting discussed in our synthetic experiments of Section 4.1.4. Since we have the ground-truth global minimum, we record the frequency that SGD hits the global minimum up to a small error 1×10^{-5} after $100,000$ iterations. Table 4.1 shows that increasing the number of hidden neurons results in higher hitting rate of global optimality. This further verifies that the loss surface of one-hidden-layer neural network becomes less non-convex as the width increases.

Table 4.1: Frequency of hitting global minimum by SGD with 100 different initialization seeds.

# Hidden Neurons	Hitting Rate	# Hidden Neurons	Hitting Rate
10	2 / 100	16	30 / 100
11	9 / 100	17	32 / 100
12	21 / 100	18	35 / 100
13	24 / 100	19	52 / 100
14	24 / 100	20	64 / 100
15	29 / 100	21	75 / 100

4.1.5 Proofs of our main results

Proofs of Theorem 47

The lower bound $0 \leq \frac{\inf(\mathbf{P})-\sup(\mathbf{D})}{\Delta_{worst}}$ is obvious by the weak duality. So we only need to prove the upper bound $\frac{\inf(\mathbf{P})-\sup(\mathbf{D})}{\Delta_{worst}} \leq \frac{2}{I}$.

Consider the subset of $\mathbb{R}^2$:

$$\mathcal{Y}_i := \left\{ \mathbf{y}_i \in \mathbb{R}^2 : \mathbf{y}_i = \frac{1}{I} \left[h_i(\mathbf{w}_{(i)}), \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}} \left(1 - \frac{y \cdot f_i(\mathbf{w}_{(i)}; \mathbf{x})}{\tau} \right) \right], \mathbf{w}_{(i)} \in \mathcal{W}_i \right\}, \quad i \in [I].$$

Define the vector summation

$$\mathcal{Y} := \mathcal{Y}_1 + \mathcal{Y}_2 + ... + \mathcal{Y}_I.$$

Since $f_i(\mathbf{w}_{(i)}; \mathbf{x})$ and $h_i(\mathbf{w}_{(i)})$ are continuous w.r.t. $\mathbf{w}_{(i)}$ and $\mathcal{W}_i$'s are compact, the set

$$\{(\mathbf{w}_{(i)}, h_i(\mathbf{w}_{(i)}), f_i(\mathbf{w}_{(i)}; \mathbf{x})) : \mathbf{w}_{(i)} \in \mathcal{W}_i\}$$

is compact as well. So $\mathcal{Y}$, $\mathsf{conv}(\mathcal{Y})$, $\mathcal{Y}_i$, and $\mathsf{conv}(\mathcal{Y}_i)$, $i \in [I]$ are all compact sets. According to the definition of $\mathcal{Y}$ and the standard duality argument [166], we have

$$\inf(\mathbf{P}) = \min \{w : \text{there exists } (r, w) \in \mathcal{Y} \text{ such that } r \leq K\},$$

and

$$\sup(\mathbf{D}) = \min \{w : \text{there exists } (r, w) \in \mathsf{conv}(\mathcal{Y}) \text{ such that } r \leq K\}.$$

Technique (a): Shapley-Folkman lemma. We are going to apply the following Shapley-Folkman lemma.

Lemma 64 (Shapley-Folkman, [216]). *Let $\mathcal{Y}_i, i \in [I]$ be a collection of subsets of $\mathbb{R}^m$. Then for every $\mathbf{y} \in \mathsf{conv}(\sum_{i=1}^I \mathcal{Y}_i)$, there is a subset $\mathcal{I}(\mathbf{y}) \subseteq [I]$ of size at most m such that*

$$\mathbf{y} \in \left[\sum_{i \notin \mathcal{I}(\mathbf{y})} \mathcal{Y}_i + \sum_{i \in \mathcal{I}(\mathbf{y})} \mathsf{conv}(\mathcal{Y}_i) \right].$$

We apply Lemma 73 to prove Theorem 47 with $m = 2$. Let $(\overline{r}, \overline{w}) \in \mathsf{conv}(\mathcal{Y})$ be such that

$$\overline{r} \le K, \quad \text{and} \quad \overline{w} = \sup(\mathbf{D}).$$

Applying the above Shapley-Folkman lemma to the set $\mathcal{Y} = \sum_{i=1}^{I} \mathcal{Y}_i$, we have that there are a subset $\overline{\mathcal{I}} \subseteq [I]$ of size 2 and vectors

$$(\overline{r}_i, \overline{w}_i) \in \mathsf{conv}(\mathcal{Y}_i), \;\; i \in \overline{\mathcal{I}} \qquad \text{and} \qquad \overline{\mathbf{w}}_{(i)} \in \mathcal{W}_i, \;\; i \notin \overline{\mathcal{I}},$$

such that

$$\frac{1}{I} \sum_{i \notin \overline{\mathcal{I}}} h_i(\overline{\mathbf{w}}_{(i)}) + \sum_{i \in \overline{\mathcal{I}}} \overline{r}_i = \overline{r} \le K, \tag{4.8}$$

$$\frac{1}{I} \sum_{i \notin \overline{\mathcal{I}}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}} \left(1 - \frac{y \cdot f_i(\overline{\mathbf{w}}_{(i)}; \mathbf{x})}{\tau}\right) + \sum_{i \in \overline{\mathcal{I}}} \overline{w}_i = \sup(\mathbf{D}). \tag{4.9}$$

Representing elements of the convex hull of $\mathcal{Y}_i \subseteq \mathbb{R}^2$ by Carathéodory theorem, we have that for each $i \in \overline{\mathcal{I}}$, there are vectors $\mathbf{w}_{(i)}^1, \mathbf{w}_{(i)}^2, \mathbf{w}_{(i)}^3 \in \mathcal{W}_i$ and scalars $a_i^1, a_i^2, a_i^3 \in \mathbb{R}$ such that

$$\sum_{j=1}^{3} a_i^j = 1, \quad a_i^j \ge 0, \; j = 1, 2, 3,$$

$$\overline{r}_i = \frac{1}{I} \sum_{j=1}^{3} a_i^j h_i(\mathbf{w}_{(i)}^j), \qquad \overline{w}_i = \frac{1}{I} \sum_{j=1}^{3} a_i^j \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}} \left(1 - \frac{y \cdot f_i(\mathbf{w}_{(i)}^j; \mathbf{x})}{\tau}\right).$$

Recall that we define

$$\widehat{f}_i(\widetilde{\mathbf{w}}) := \inf_{\mathbf{w}_{(i)} \in \mathcal{W}_i} \left\{ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}} \left(1 - \frac{y \cdot f_i(\mathbf{w}_{(i)}; \mathbf{x})}{\tau}\right) : h_i(\mathbf{w}_{(i)}) \le h_i(\widetilde{\mathbf{w}}) \right\}, \tag{4.10}$$

$$\widetilde{f}_i(\widetilde{\mathbf{w}}) := \inf_{a^j, \mathbf{w}_{(i)}^j \in \mathcal{W}_i} \left\{ \sum_{j=1}^{p_i+2} a^j \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}} \left(1 - \frac{y \cdot f_i(\mathbf{w}_{(i)}^j; \mathbf{x})}{\tau}\right) : \widetilde{\mathbf{w}} = \sum_{j=1}^{p_i+2} a^j \mathbf{w}_{(i)}^j, \sum_{j=1}^{p_i+2} a^j = 1, a^j \ge 0 \right\},$$

and $\Delta_i := \sup_{\mathbf{w} \in \mathcal{W}_i} \left\{ \widehat{f}_i(\mathbf{w}) - \widetilde{f}_i(\mathbf{w}) \right\} \ge 0$. We have for $i \in \overline{\mathcal{I}}$,

$$\overline{r}_i \ge \frac{1}{I} h_i \left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j \right), \quad \text{(because } h_i(\cdot) \text{ is convex)} \tag{4.11}$$

and

$$\begin{aligned} \overline{w}_i &\ge \frac{1}{I} \widetilde{f}_i \left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j \right) \quad \text{(by the definition of } \widetilde{f}_i(\cdot)) \\ &\ge \frac{1}{I} \widehat{f}_i \left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j \right) - \frac{1}{I} \Delta_i. \quad \text{(by the definition of } \Delta_i) \end{aligned} \tag{4.12}$$

Thus, by Eqns. (4.26) and (4.28), we have

$$\frac{1}{I}\sum_{i\notin\overline{\mathcal{I}}} h_i(\overline{\mathbf{w}}_{(i)}) + \frac{1}{I}\sum_{i\in\overline{\mathcal{I}}} h_i\left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j\right) \leq K, \tag{4.13}$$

and by Eqns. (4.27) and (4.29), we have

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left[\frac{1}{I}\sum_{i\notin\overline{\mathcal{I}}}\left(1-\frac{y\cdot f_i(\overline{\mathbf{w}}_{(i)};\mathbf{x})}{\tau}\right)\right] + \frac{1}{I}\sum_{i\in\overline{\mathcal{I}}}\widehat{f}_i\left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j\right) \leq \sup(\mathbf{D}) + \frac{1}{I}\sum_{i\in\overline{\mathcal{I}}}\Delta_i. \tag{4.14}$$

Given any $\epsilon > 0$ and $i \in \overline{\mathcal{I}}$, we can find a vector $\overline{\mathbf{w}}_{(i)} \in \mathcal{W}_i$ such that

$$h_i(\overline{\mathbf{w}}_{(i)}) \leq h_i\left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j\right) \text{ and } \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left(1-\frac{y\cdot f_i(\overline{\mathbf{w}}_{(i)};\mathbf{x})}{\tau}\right) \leq \widehat{f}_i\left(\sum_{j=1}^{3} a_i^j \mathbf{w}_{(i)}^j\right) + \epsilon, \tag{4.15}$$

where the first inequality holds because $\mathcal{W}_i$ is convex and the second inequality holds by the definition (4.10) of $\widehat{f}_i(\cdot)$. Therefore, Eqns. (4.30) and (4.15) impliy that

$$\frac{1}{I}\sum_{i=1}^{I} h_i(\overline{\mathbf{w}}_{(i)}) \leq K.$$

Namely, $(\overline{\mathbf{w}}_{(1)}, ..., \overline{\mathbf{w}}_{(I)})$ is a feasible solution of problem (4.2). Also, Eqns. (4.14) and (4.15) yield

$$\begin{aligned}
\inf(\mathbf{P}) &\leq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}}\left[\frac{1}{I}\sum_{i=1}^{I}\left(1-\frac{y\cdot f_i(\overline{\mathbf{w}}_{(i)};\mathbf{x})}{\tau}\right)\right] \\
&\leq \sup(\mathbf{D}) + \frac{1}{I}\sum_{i\in\overline{\mathcal{I}}}(\Delta_i + \epsilon) \\
&\leq \sup(\mathbf{D}) + \frac{2}{I}\Delta_{worst} + 2\epsilon,
\end{aligned}$$

where the last inequality holds because $|\overline{\mathcal{I}}| = 2$. Finally, letting $\epsilon \to 0$ leads to the desired result.

Proofs of Theorem 48

Let $\widetilde{\mathbf{Y}} = \mathbf{Y}\mathbf{X}^\dagger\mathbf{X}$. We note that by Pythagorean theorem, for every $\mathbf{Y}$,

$$\frac{1}{2}\|\mathbf{Y} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 = \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \underbrace{\frac{1}{2}\|\mathbf{Y}-\widetilde{\mathbf{Y}}\|_F^2}_{\text{independent of } \mathbf{W}_1,\ldots,\mathbf{W}_H}.$$

So we can focus on the following optimization problem instead of problem (4.4):

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H ... \mathbf{W}_1\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right]. \tag{4.16}$$

Technique (b): Variational form. Our work is inspired by a variational form of problem (4.16) given by the following lemma.

Lemma 65. *If* $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ *is optimal to problem*

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} F(\mathbf{W}_1, \ldots, \mathbf{W}_H) := \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + \gamma\|\mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_*, \tag{4.17}$$

then $(\mathbf{W}_1^{**}, \ldots, \mathbf{W}_H^{**})$ *is optimal to problem* (4.16)*, where* $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ *is the skinny SVD of*

$$\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X},$$

$\mathbf{W}_i^{**} = [\mathbf{\Sigma}^{1/H}, \mathbf{0}; \mathbf{0}, \mathbf{0}] \in \mathbb{R}^{d_i \times d_{i-1}}$ *for* $i = 2, 3, ..., H-1$, $\mathbf{W}_H^{**} = [\mathbf{U}\mathbf{\Sigma}^{1/H}, \mathbf{0}] \in \mathbb{R}^{d_H \times d_{H-2}}$ *and* $\mathbf{W}_1^{**} = [\mathbf{\Sigma}^{1/H}\mathbf{V}^T; \mathbf{0}]\mathbf{X}^\dagger \in \mathbb{R}^{d_1 \times d_0}$*. Furthermore, problems* (4.16) *and* (4.17) *have the same optimal objective function value.*

Proof of Lemma 65. Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the skinny SVD of matrix $\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X} =: \mathbf{Z}$. We notice that

$$\begin{aligned}
\|\mathbf{Z}\|_* &= \|\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_* \\
&\le \|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}\prod_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H} \quad \text{(by the generalized Hölder's inequality)} \\
&\le \frac{1}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right]. \quad \text{(by the inequality of mean)}
\end{aligned}$$

Hence, on one hand, for every $(\mathbf{W}_1, \ldots, \mathbf{W}_H)$,

$$\begin{aligned}
\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} F(\mathbf{W}_1, \ldots, \mathbf{W}_H) &\le \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + \gamma\|\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_* \\
&\le \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right],
\end{aligned}$$

which yields

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} F(\mathbf{W}_1, \ldots, \mathbf{W}_H) \le \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right].$$

On the other hand, suppose $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ is optimal to problem (4.17), and let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the skinny SVD of matrix $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}$. We choose $(\mathbf{W}_1^{**}, \ldots, \mathbf{W}_H^{**})$ such that

$$\mathbf{W}_H^{**} = [\mathbf{U}\mathbf{\Sigma}^{\frac{1}{H}}, \mathbf{0}],\ \mathbf{W}_1^{**}\mathbf{X} = [\mathbf{\Sigma}^{\frac{1}{H}}\mathbf{V}^T; \mathbf{0}],\ \mathbf{W}_i^{**} = [\mathbf{\Sigma}^{\frac{1}{H}}, \mathbf{0}; \mathbf{0}, \mathbf{0}],\ i = 2, \ldots, H-1.$$

We pad $\mathbf{0}$ around $\mathbf{W}_i^{**}$ so as to adapt to the dimensionality of each $\mathbf{W}_i^{**}$. Notice that

$$\begin{aligned}
\|\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}\|_* &= \|\mathbf{W}_H^{**}\mathbf{W}_{H-1}^{**}\cdots\mathbf{W}_1^{**}\mathbf{X}\|_* \\
&= \frac{1}{H}\left[\|\mathbf{W}_1^{**}\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^{H}\|\mathbf{W}_i^{**}\|_{\mathcal{S}_H}^H\right].
\end{aligned}$$

Since $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} = \mathbf{W}_H^{**}\mathbf{W}_{H-1}^{**}\cdots\mathbf{W}_1^{**}\mathbf{X}$, for every $\widetilde{\mathbf{Y}}$,

$$\|\widetilde{\mathbf{Y}} - \mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}\|_F = \|\widetilde{\mathbf{Y}} - \mathbf{W}_H^{**}\mathbf{W}_{H-1}^{**}\cdots\mathbf{W}_1^{**}\mathbf{X}\|_F.$$

Hence

$$\begin{aligned}
&\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} F(\mathbf{W}_1,\ldots,\mathbf{W}_H) = F(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*) = F(\mathbf{W}_1^{**},\ldots,\mathbf{W}_H^{**}) \\
&= \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H^{**}\cdots\mathbf{W}_1^{**}\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1^{**}\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^H \|\mathbf{W}_i^{**}\|_{\mathcal{S}_H}^H\right] \\
&\geq \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \frac{\gamma}{H}\left[\|\mathbf{W}_1\mathbf{X}\|_{\mathcal{S}_H}^H + \sum_{i=2}^H \|\mathbf{W}_i\|_{\mathcal{S}_H}^H\right],
\end{aligned}$$

which yields the other direction of the inequality and hence completes the proof. □

Technique (c): Reduction to low-rank approximation. We now reduce problem (4.17) to the classic problem of low-rank approximation of the form $\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2$, which has the following nice properties.

Lemma 66. *For any $\widehat{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$, every global minimum $(\mathbf{W}_1^*,\ldots,\mathbf{W}_H^*)$ of function*

$$f(\mathbf{W}_1,\ldots,\mathbf{W}_H) = \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2$$

obeys $\mathbf{W}_H^*\cdots\mathbf{W}_1^*\mathbf{X} = \mathsf{svd}_{d_{\min}}(\widehat{\mathbf{Y}})$. Here $\widehat{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$ means the row vectors of $\widehat{\mathbf{Y}}$ belongs to the row space of $\mathbf{X}$.

Proof of Lemma 66. Note that the optimal solution to $\min_{\mathbf{W}_H,\ldots,\mathbf{W}_1} \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2$ is equal to the optimal solution to the low-rank approximation problem $\min_{\mathsf{rank}(\mathbf{Z})\leq d_{\min}} \frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{Z}\|_F^2$ when $\widehat{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$, which has a closed-form solution $\mathsf{svd}_{d_{\min}}(\widehat{\mathbf{Y}})$.[6] □

We now reduce $F(\mathbf{W}_1,\ldots,\mathbf{W}_H)$ to the form of $\frac{1}{2}\|\widehat{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2$ for some $\widehat{\mathbf{Y}}$ plus an extra additive term that is independent of $(\mathbf{W}_1,\ldots,\mathbf{W}_H)$. To see this, denote by $K(\cdot) = \gamma\|\cdot\|_*$. We have

$$\begin{aligned}
F(\mathbf{W}_1,\ldots,\mathbf{W}_H) &= \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + K^{**}(\mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}) \\
&= \max_{\mathbf{\Lambda}} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 + \langle\mathbf{\Lambda}, \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\rangle - K^*(\mathbf{\Lambda}) \\
&= \max_{\mathbf{\Lambda}} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda} - \mathbf{W}_H\cdots\mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - K^*(\mathbf{\Lambda}) + \langle\widetilde{\mathbf{Y}}, \mathbf{\Lambda}\rangle \\
&=: \max_{\mathbf{\Lambda}} L(\mathbf{W}_1,\ldots,\mathbf{W}_H,\mathbf{\Lambda}),
\end{aligned}$$

[6]Note that the low-rank approximation problem might have non-unique solution. However, we will use in this section the abuse of language $\mathsf{svd}_{d_{\min}}(\widehat{\mathbf{Y}})$ as the non-uniqueness issue does not lead to any issue in our developments.

where we define $L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}) := \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - K^*(\mathbf{\Lambda}) + \langle \widetilde{\mathbf{Y}}, \mathbf{\Lambda}\rangle$ as the Lagrangian of problem (4.17). The first equality holds because $K(\cdot)$ is closed and convex w.r.t. the argument $\mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}$ so $K(\cdot) = K^{**}(\cdot)$, and the second equality is by the definition of conjugate function. One can check that $L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}) = \min_{\mathbf{M}} L'(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{M}, \mathbf{\Lambda})$, where $L'(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{M}, \mathbf{\Lambda})$ is the Lagrangian of the constraint optimization problem

$$\min_{\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{M}} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + K(\mathbf{M}), \text{ s.t. } \mathbf{M} = \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}.$$

With a little abuse of notation, we call $L(\mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$ the Lagrangian of the unconstrained problem (4.17) as well.

The remaining analysis is to choose a proper $\mathbf{\Lambda}^* \in \mathsf{Row}(\mathbf{X})$ such that $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*, \mathbf{\Lambda}^*)$ is a primal-dual saddle point of $L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda})$, so that the problem

$$\min_{\mathbf{W}_1, \ldots, \mathbf{W}_H} L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}^*)$$

and problem (4.17) have the same optimal solution $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$. For this, we introduce the following condition, and later we will show that the condition holds.

Condition 2. *For a solution $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ to optimization problem (4.17), there exists an*

$$\mathbf{\Lambda}^* \in \partial_{\mathbf{Z}} K(\mathbf{Z})|_{\mathbf{Z} = \mathbf{W}_H^* \cdots \mathbf{W}_1^* \mathbf{X}} \cap \mathsf{Row}(\mathbf{X})$$

such that

$$\begin{aligned} \mathbf{W}_{i+1}^{*T} \cdots \mathbf{W}_H^{*T}(\mathbf{W}_H^* \cdots \mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}})\mathbf{X}^T\mathbf{W}_1^{*T} \cdots \mathbf{W}_{i-1}^{*T} &= \mathbf{0}, \quad i = 2, \ldots, H-1, \\ \mathbf{W}_2^{*T} \cdots \mathbf{W}_H^{*T}(\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}})\mathbf{X}^T &= \mathbf{0}, \\ (\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}})\mathbf{X}^T\mathbf{W}_1^{*T} \cdots \mathbf{W}_{H-1}^{*T} &= \mathbf{0}. \end{aligned} \tag{4.18}$$

We note that if we set $\mathbf{\Lambda}$ to be the $\mathbf{\Lambda}^*$ in (4.18), then $\nabla_{\mathbf{W}_i} L(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*, \mathbf{\Lambda}^*) = \mathbf{0}$ for every i. So $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ is either a saddle point, a local minimizer, or a global minimizer of $L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}^*)$ as a function of $(\mathbf{W}_1, \ldots, \mathbf{W}_H)$ for the fixed $\mathbf{\Lambda}^*$. The following lemma states that if it is a global minimizer, then strong duality holds.

Lemma 67. *Let $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ be a global minimizer of $F(\mathbf{W}_1, \ldots, \mathbf{W}_H)$. If there exists a dual certificate $\mathbf{\Lambda}^*$ satisfying Condition 2 and the pair $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ is a global minimizer of $L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}^*)$ for the fixed $\mathbf{\Lambda}^*$, then strong duality holds. Moreover, we have the relation $\mathbf{W}_H^* \cdots \mathbf{W}_1^*\mathbf{X} = \mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$.*

Proof of Lemma 67. By the assumption of the lemma, $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ is a global minimizer of

$$L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}^*) = \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^* - \mathbf{W}_H\mathbf{W}_{H-1} \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + c(\mathbf{\Lambda}^*),$$

where $c(\mathbf{\Lambda}^*)$ is a function of $\mathbf{\Lambda}^*$ that is independent of $\mathbf{W}_i$ for all i's. Namely, $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ globally minimizes $L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda})$ when $\mathbf{\Lambda}$ is fixed to $\mathbf{\Lambda}^*$. Furthermore,

$$\mathbf{\Lambda}^* \in \partial_{\mathbf{Z}} K(\mathbf{Z})|_{\mathbf{Z} = \mathbf{W}_H^* \cdots \mathbf{W}_1^* \mathbf{X}}$$

implies that $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} \in \partial_{\mathbf{\Lambda}} K^*(\mathbf{\Lambda})|_{\mathbf{\Lambda}=\mathbf{\Lambda}^*}$ by the convexity of function $K(\cdot)$, meaning that $\mathbf{0} \in \partial_{\mathbf{\Lambda}} L(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*, \mathbf{\Lambda})$. So $\mathbf{\Lambda}^* = \operatorname{argmax}_{\mathbf{\Lambda}} L(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*, \mathbf{\Lambda})$ due to the concavity of function $L(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*, \mathbf{\Lambda})$ w.r.t. variable $\mathbf{\Lambda}$. Thus $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*, \mathbf{\Lambda}^*)$ is a primal-dual saddle point of $L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda})$.

We now prove the strong duality. By the fact that

$$F(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*) = \max_{\mathbf{\Lambda}} L(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*, \mathbf{\Lambda})$$

and that $\mathbf{\Lambda}^* = \operatorname{argmax}_{\mathbf{\Lambda}} L(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*, \mathbf{\Lambda})$, for every $\mathbf{W}_1, \ldots, \mathbf{W}_H$, we have

$$F(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*) = L(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*, \mathbf{\Lambda}^*) \leq L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}^*),$$

where the inequality holds because $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*, \mathbf{\Lambda}^*)$ is a primal-dual saddle point of L. Notice that we also have

$$\begin{aligned}\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \max_{\mathbf{\Lambda}} L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}) &= F(\mathbf{W}_1^*, \ldots \mathbf{W}_H^*) \\ &\leq \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}^*) \\ &\leq \max_{\mathbf{\Lambda}} \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}).\end{aligned}$$

On the other hand, by weak duality,

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \max_{\mathbf{\Lambda}} L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}) \geq \max_{\mathbf{\Lambda}} \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}).$$

Therefore,

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \max_{\mathbf{\Lambda}} L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}) = \max_{\mathbf{\Lambda}} \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}),$$

i.e., strong duality holds. Hence,

$$\begin{aligned}\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^* &= \operatorname*{argmin}_{\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1} L(\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{\Lambda}^*) \\ &= \operatorname*{argmin}_{\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^* - \mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}^*\|_F^2 - K^*(\mathbf{\Lambda}^*) + \langle\widetilde{\mathbf{Y}}, \mathbf{\Lambda}^*\rangle \\ &= \operatorname*{argmin}_{\mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^* - \mathbf{W}_H\mathbf{W}_{H-1}\cdots\mathbf{W}_1\mathbf{X}\|_F^2 \\ &= \mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*).\end{aligned}$$

The proof of Lemma 67 is completed. □

Technique (d): Dual certificate. We now construct dual certificate $\mathbf{\Lambda}^*$ such that all of conditions in Lemma 67 hold. We note that $\mathbf{\Lambda}^*$ should satisfy the followings by Lemma 67:

(a) $\mathbf{\Lambda}^* \in \partial K(\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X}) \cap \mathsf{Row}(\mathbf{X})$; (by Condition 2)

(b) Equations (4.18); (by Condition 2)

(c) $\mathbf{W}_H^*\mathbf{W}_{H-1}^*\cdots\mathbf{W}_1^*\mathbf{X} = \mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$. (by the global optimality and Lemma 66)

(4.19)

Before proceeding, we denote by $\widetilde{\mathbf{A}} := \mathbf{W}_H^* \cdots \mathbf{W}_{\min+1}^*$, $\widetilde{\mathbf{B}} := \mathbf{W}_{\min}^* \cdots \mathbf{W}_1^* \mathbf{X}$, where $\mathbf{W}_{\min}^*$ is a matrix among $\{\mathbf{W}_i^*\}_{i=1}^{H-1}$ which has $d_{\min}$ rows, and let

$$\mathcal{T} := \{\widetilde{\mathbf{A}}\mathbf{C}_1^T + \mathbf{C}_2\widetilde{\mathbf{B}} : \ \mathbf{C}_1 \in \mathbb{R}^{n\times d_{\min}}, \ \mathbf{C}_2 \in \mathbb{R}^{d_H \times d_{\min}}\}$$

be a matrix space. Denote by $\mathcal{U}$ the left singular space of $\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$ and $\mathcal{V}$ the right singular space. Then the linear space $\mathcal{T}$ can be equivalently represented as $\mathcal{T} = \mathcal{U} + \mathcal{V}$. Therefore, $\mathcal{T}^\perp = (\mathcal{U}+\mathcal{V})^\perp = \mathcal{U}^\perp \cap \mathcal{V}^\perp$. With this, we note that: (b) $\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}} \in \mathsf{Null}(\widetilde{\mathbf{A}}^T) = \mathsf{Col}(\widetilde{\mathbf{A}})^\perp$ and $\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}} \in \mathsf{Row}(\widetilde{\mathbf{B}})^\perp$ (so $\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}} \in \mathcal{T}^\perp$) imply Equations (4.18) since either $\mathbf{W}_{i+1}^{*T} \cdots \mathbf{W}_H^{*T}(\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}}) = \mathbf{0}$ or $(\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} + \mathbf{\Lambda}^* - \widetilde{\mathbf{Y}})\mathbf{X}^T\mathbf{W}_1^{*T} \cdots \mathbf{W}_{i-1}^{*T} = \mathbf{0}$ for all i's. And (c) for an orthogonal decomposition $\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^* = \mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} + \mathbf{E}$ where $\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} \in \mathcal{T}$ and $\mathbf{E} \in \mathcal{T}^\perp$, we have that

$$\|\mathbf{E}\| \le \sigma_{d_{\min}}(\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X})$$

and condition (b) together imply $\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} = \mathsf{svd}_{d_{\min}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)$ by Lemma 66. Therefore, the dual conditions in (4.19) are implied by

$$\begin{aligned}
&(1) \quad \mathbf{\Lambda}^* \in \partial K(\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X}) \cap \mathsf{Row}(\mathbf{X});\\
&(2) \quad \mathcal{P}_{\mathcal{T}}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*) = \mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X};\\
&(3) \quad \|\mathcal{P}_{\mathcal{T}^\perp}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)\| \le \sigma_{d_{\min}}(\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X}).
\end{aligned}$$

It thus suffices to construct a dual certificate $\mathbf{\Lambda}^*$ such that conditions (1), (2) and (3) hold, because conditions (1), (2) and (3) are stronger than conditions (a), (b) and (c). Let $r = \mathsf{rank}(\widetilde{\mathbf{Y}})$ and $\bar{r} = \min\{r, d_{\min}\}$. To proceed, we need the following lemma.

Lemma 68 ([227]). *Suppose $\widetilde{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$. Let $(\mathbf{W}_1^*, \ldots, \mathbf{W}_H^*)$ be the solution to problem (4.17) and let $\mathbf{U}\mathsf{diag}(\sigma_1(\widetilde{\mathbf{Y}}), \ldots, \sigma_r(\widetilde{\mathbf{Y}}))\mathbf{V}^T$ denote the skinny SVD of $\widetilde{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$. We have $\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X} = \mathbf{U}\mathsf{diag}((\sigma_1(\widetilde{\mathbf{Y}}) - \gamma)_+, \ldots, (\sigma_{\bar{r}}(\widetilde{\mathbf{Y}}) - \gamma)_+, 0, \ldots, 0)\mathbf{V}^T$.*

Recall that the sub-differential of the nuclear norm of a matrix $\mathbf{Z}$ is

$$\partial_{\mathbf{Z}}\|\mathbf{Z}\|_* = \{\mathbf{U}_\mathbf{Z}\mathbf{V}_\mathbf{Z}^T + \mathbf{T}_\mathbf{Z} : \mathbf{T}_\mathbf{Z} \in \mathcal{T}^\perp, \|\mathbf{T}_\mathbf{Z}\| \le 1\},$$

where $\mathbf{U}_\mathbf{Z}\mathbf{\Sigma}_\mathbf{Z}\mathbf{V}_\mathbf{Z}^T$ is the skinny SVD of the matrix $\mathbf{Z}$. So with Lemma 68, the sub-differential of (scaled) nuclear norm at optimizer $\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X}$ is given by

$$\partial(\gamma\|\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X}\|_*) = \{\gamma\mathbf{U}_{:,1:\bar{r}}\mathbf{V}_{:,1:\bar{r}}^T + \mathbf{T} : \mathbf{T} \in \mathcal{T}^\perp, \|\mathbf{T}\| \le \gamma\}. \tag{4.20}$$

To construct the dual certificate, we set

$$\mathbf{\Lambda}^* = \underbrace{\gamma\mathbf{U}_{:,1:\bar{r}}\mathbf{V}_{:,1:\bar{r}}^T}_{\text{Component in space } \mathcal{T}} + \underbrace{\mathbf{U}_{:,(\bar{r}+1):r}\mathsf{diag}(\gamma, \ldots, \gamma)\mathbf{V}_{:,(\bar{r}+1):r}^T}_{\text{Component } \mathbf{T} \text{ in space } \mathcal{T}^\perp \text{ with } \|\mathbf{T}\| \le \gamma} \in \mathsf{Row}(\mathbf{X}),$$

where $\mathbf{\Lambda}^* \in \mathsf{Row}(\mathbf{X})$ because $\mathbf{V}^T \in \mathsf{Row}(\mathbf{X})$ (This is because $\mathbf{V}^T$ is the right singular matrix of $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{Y}} \in \mathsf{Row}(\mathbf{X})$). So condition (1) is satisfied according to (4.20). To see condition (2), $\mathcal{P}_{\mathcal{T}}(\widetilde{\mathbf{Y}} -$

$\mathbf{\Lambda}^*) = \mathcal{P}_{\mathcal{T}}\widetilde{\mathbf{Y}} - \gamma \mathbf{U}_{:,1:\bar{r}}\mathbf{V}_{:,1:\bar{r}}^T = \mathbf{U}\mathsf{diag}((\sigma_1(\widetilde{\mathbf{Y}}) - \gamma)_+, \ldots, (\sigma_{\bar{r}}(\widetilde{\mathbf{Y}}) - \gamma)_+, 0, 0, \ldots, 0)\mathbf{V}^T = \mathbf{W}_H^*\mathbf{W}_{H-1}^* \ldots \mathbf{W}_1^*\mathbf{X}$, where the last equality is by Lemma 68 and the assumption $\sigma_{\min}(\widetilde{\mathbf{Y}}) > \gamma$. As for condition (3), note that

$$\left\|\mathcal{P}_{\mathcal{T}^\perp}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)\right\| = \left\|\mathbf{U}_{:,(\bar{r}+1):r}\mathsf{diag}(\sigma_{\bar{r}+1}(\widetilde{\mathbf{Y}}) - \gamma, \ldots, \sigma_r(\widetilde{\mathbf{Y}}) - \gamma)\mathbf{V}_{:,(\bar{r}+1):r}^T\right\| = \begin{cases} 0, & \text{if } \bar{r} = r, \\ \sigma_{d_{\min}+1}(\widetilde{\mathbf{Y}}) - \gamma, & \text{otherwise.} \end{cases}$$

By Lemma 68, $\sigma_{d_{\min}}(\mathbf{W}_H^*\mathbf{W}_{H-1}^* \cdots \mathbf{W}_1^*\mathbf{X}) \geq \|\mathcal{P}_{\mathcal{T}^\perp}(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}^*)\|$. So the proof of strong duality is completed, where the dual problem is given in the next section.

To see the relation between the solutions of primal and dual problems, it is a direct result of Lemmas 65 and 67.

Dual problem of deep linear neural network

In this section, we derive the dual problem of non-convex program (4.4). Denote by $G(\mathbf{W}_1, \ldots, \mathbf{W}_H)$ the objective function of problem (4.4). Let $K(\cdot) = \gamma\|\cdot\|_*$, and let $\widetilde{\mathbf{Y}} = \mathbf{Y}\mathbf{X}^\dagger\mathbf{X}$ be the projection of $\mathbf{Y}$ on the row span of $\mathbf{X}$. We note that

$$\begin{aligned}
&\min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} G(\mathbf{W}_1, \ldots, \mathbf{W}_H) - \frac{1}{2}\|\mathbf{Y} - \widetilde{\mathbf{Y}}\|_F^2 \\
&= \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\mathbf{Y} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{Y} - \widetilde{\mathbf{Y}}\|_F^2 + K(\mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}) \\
&= \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + K^{**}(\mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}) \\
&= \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 + \langle \mathbf{\Lambda}, \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\rangle - K^*(\mathbf{\Lambda}) \\
&= \min_{\mathbf{W}_1,\cdots,\mathbf{W}_H} \max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda} - \mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - K^*(\mathbf{\Lambda}) + \langle \widetilde{\mathbf{Y}}, \mathbf{\Lambda}\rangle,
\end{aligned}$$

where the second equality holds since $K(\cdot)$ is closed and convex w.r.t. the argument $\mathbf{W}_H \cdots \mathbf{W}_1\mathbf{X}$ and the third equality is by the definition of conjugate function of nuclear norm. Therefore, the dual problem is given by

$$\begin{aligned}
&\max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} \min_{\mathbf{W}_1,\ldots,\mathbf{W}_H} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda} - \mathbf{W}_H \ldots \mathbf{W}_1\mathbf{X}\|_F^2 - \frac{1}{2}\|\mathbf{\Lambda}\|_F^2 - K^*(\mathbf{\Lambda}) + \langle \widetilde{\mathbf{Y}}, \mathbf{\Lambda}\rangle + \frac{1}{2}\|\mathbf{Y} - \widetilde{\mathbf{Y}}\|_F^2 \\
&= \max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} \frac{1}{2}\sum_{i=d_{\min}+1}^{\min\{d_H,n\}} \sigma_i^2(\widetilde{\mathbf{Y}} - \mathbf{\Lambda}) - \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda}\|_F^2 - K^*(\mathbf{\Lambda}) + \frac{1}{2}\|\mathbf{Y}\|_F^2 \\
&= \max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} -\frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{\Lambda}\|_{d_{\min}}^2 - K^*(\mathbf{\Lambda}) + \frac{1}{2}\|\mathbf{Y}\|_F^2,
\end{aligned}$$

where $\|\cdot\|_{d_{\min}}^2 = \sum_{i=1}^{d_{\min}} \sigma_i^2(\cdot)$. We note that

$$K^*(\mathbf{\Lambda}) = \begin{cases} 0, & \|\mathbf{\Lambda}\| \leq \gamma; \\ +\infty, & \|\mathbf{\Lambda}\| > \gamma. \end{cases}$$

So the dual problem is given by

$$\max_{\mathsf{Row}(\mathbf{\Lambda})\subseteq\mathsf{Row}(\mathbf{X})} -\frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{\Lambda}\|_{d_{\min}}^2+\frac{1}{2}\|\mathbf{Y}\|_F^2, \quad \text{s.t.} \quad \|\mathbf{\Lambda}\|\leq\gamma. \tag{4.21}$$

Problem (4.21) can be solved efficiently due to their convexity. In particular, Grussler et al. [99] provided a computationally efficient algorithm to compute the proximal operators of functions $\frac{1}{2}\|\cdot\|_r^2$. Hence, the Douglas-Rachford algorithm can find the global minimum up to an ϵ error in function value in time $\mathsf{poly}(1/\epsilon)$ [114].

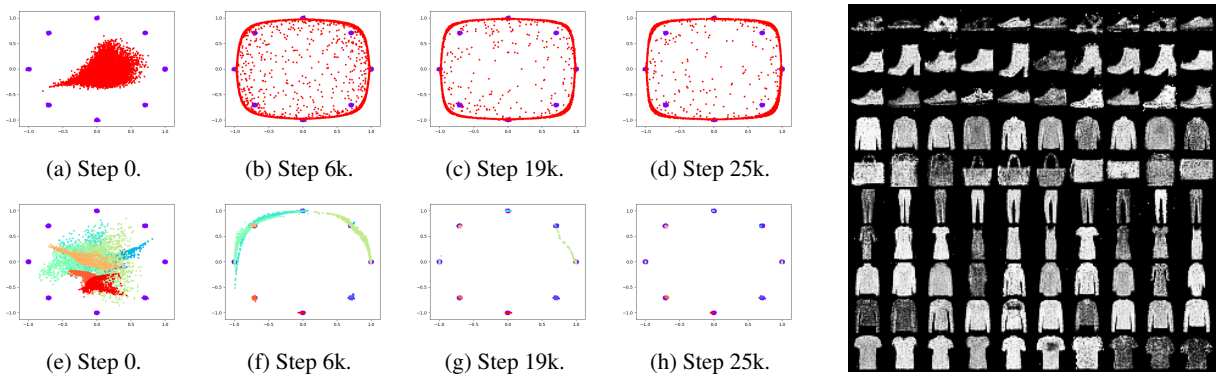
4.2 Stackelberg Generative Adversarial Nets

4.2.1 Introduction

Generative Adversarial Nets (GANs) are emerging objects of study in machine learning, computer vision, natural language processing, and many other domains. In machine learning, study of such a framework has led to significant advances in adversarial defenses [203, 242] and machine security [13, 203]. In computer vision and natural language processing, GANs have resulted in improved performance over standard generative models for images and texts [96], such as variational autoencoder [135] and deep Boltzmann machine [201]. A main technique to achieve this goal is to play a minimax two-player game between generator and discriminator under the design that the generator tries to confuse the discriminator with its generated contents and the discriminator tries to distinguish real images/texts from what the generator creates.

Despite a large amount of variants of GANs, many fundamental questions remain unresolved. One of the long-standing challenges is designing *universal, easy-to-implement* architectures that alleviate the instability issue of GANs training. Ideally, GANs are supposed to solve the minimax optimization problem [96], but in practice alternating gradient descent methods do not clearly privilege minimax over maximin or vice versa (page 35, [95]), which may lead to instability in training if there exists a large discrepancy between the minimax and maximin objective values. The focus of this work is on improving the stability of such minimax game in the training process of GANs.

To alleviate the issues caused by the large minimax gap, our study is motivated by the so-called Stackelberg competition in the domain of game theory. In the Stackelberg leadership model, the players of this game are one *leader* and multiple *followers*, where the leader firm moves first and then the follower firms move sequentially. It is known that the Stackelberg model can be solved to find a *subgame perfect Nash equilibrium*. We apply this idea of Stackelberg leadership model to the architecture design of GANs. That is, we design an improved GAN architecture with multiple generators (followers) which team up to play against the discriminator (leader). We therefore name our model *Stackelberg GAN*. Our theoretical and experimental results establish that: *GANs with multi-generator architecture have smaller minimax gap, and enjoy more stable training performances.*



(a) Step 0. (b) Step 6k. (c) Step 19k. (d) Step 25k.

(e) Step 0. (f) Step 6k. (g) Step 19k. (h) Step 25k.

Figure 4.5: **Left Figure, Top Row:** Standard GAN training on a toy 2D mixture of 8 Gaussians. **Left Figure, Bottom Row:** Stackelberg GAN training with 8 generator ensembles, each of which is denoted by one color. **Right Figure:** Stackelberg GAN training with 10 generator ensembles on fashion-MNIST dataset without cherry pick, where each row corresponds to one generator.

4.2.2 Stackelberg GANs

Before proceeding, we define some notations and formalize our model setup in this section.

Notations. We will use bold lower-case letter to represent vector and lower-case letter to represent scalar. Specifically, we denote by $\theta \in \mathbb{R}^t$ the parameter vector of discriminator and $\gamma \in \mathbb{R}^g$ the parameter vector of generator. Let $D_\theta(\mathbf{x})$ be the output probability of discriminator given input $\mathbf{x}$, and let $G_\gamma(\mathbf{z})$ represent the generated vector given random input $\mathbf{z}$. For any function $f(\mathbf{u})$, we denote by $f^*(\mathbf{v}) := \sup_{\mathbf{u}}\{\mathbf{u}^T\mathbf{v} - f(\mathbf{u})\}$ the conjugate function of f. Let $\breve{\mathsf{cl}}f$ be the convex closure of f, which is defined as the function whose epigraph is the convex closed hull of that of function f. We define $\widehat{\mathsf{cl}}f := -\breve{\mathsf{cl}}(-f)$. We will use I to represent the number of generators.

Model setup

Preliminaries. The key ingredient in the standard GAN is to play a *zero-sum two-player* game between a discriminator and a generator — which are often parametrized by deep neural networks in practice — such that the goal of the generator is to map random noise $\mathbf{z}$ to some plausible images/texts $G_\gamma(\mathbf{z})$ and the discriminator $D_\theta(\cdot)$ aims at distinguishing the real images/texts from what the generator creates.

For every parameter implementations γ and θ of generator and discriminator, respectively, denote by the payoff value

$$\phi(\gamma;\theta) := \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_d} f(D_\theta(\mathbf{x})) + \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_z} f(1 - D_\theta(G_\gamma(\mathbf{z}))),$$

where $f(\cdot)$ is some concave, increasing function. Hereby, $\mathcal{P}_d$ is the distribution of true images/texts and $\mathcal{P}_z$ is a noise distribution such as Gaussian or uniform distribution. The standard GAN thus solves the following saddle point problems:

$$\inf_{\gamma\in\mathbb{R}^g}\sup_{\theta\in\mathbb{R}^t}\phi(\gamma;\theta), \qquad \text{or} \qquad \sup_{\theta\in\mathbb{R}^t}\inf_{\gamma\in\mathbb{R}^g}\phi(\gamma;\theta). \tag{4.22}$$

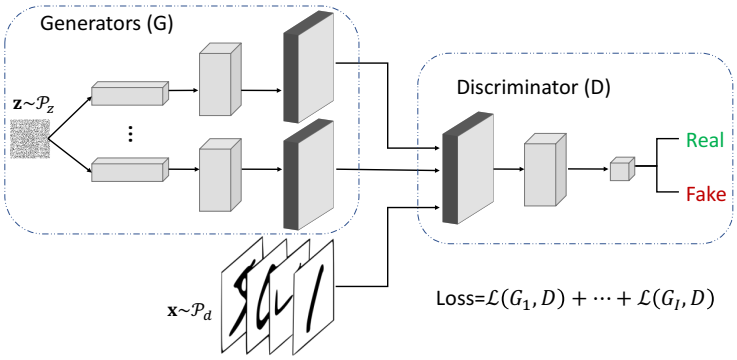


Figure 4.6: Architecture of Stackelberg GAN. We ensemble the losses of various generator and discriminator pairs with equal weights.

For different choices of function f, problem (4.22) leads to various variants of GAN. For example, when $f(t) = \log t$, problem (4.22) is the classic GAN; when $f(t) = t$, it reduces to the Wasserstein GAN. We refer interested readers to the paper of [181] for more variants of GANs.

Stackelberg GAN. Our model of Stackelberg GAN is inspired from the Stackelberg competition in the domain of game theory. Instead of playing a two-player game as in the standard GAN, in Stackelberg GAN there are $I + 1$ players with two firms — one discriminator and I generators. One can make an analogy between the discriminator (generators) in the Stackelberg GAN and the leader (followers) in the Stackelberg competition.

Stackelberg GAN is a general framework which can be built on top of all variants of standard GANs. The objective function is simply an ensemble of losses w.r.t. all possible pairs of generators and discriminator: $\Phi(\gamma_1, ..., \gamma_I; \theta) := \sum_{i=1}^{I} \phi(\gamma_i; \theta)$. Thus it is very easy to implement. The Stackelberg GAN therefore solves the following saddle point problems:

$$w^* := \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \sup_{\theta \in \mathbb{R}^t} \frac{1}{I} \Phi(\gamma_1, ..., \gamma_I; \theta), \qquad \text{or} \qquad q^* := \sup_{\theta \in \mathbb{R}^t} \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \frac{1}{I} \Phi(\gamma_1, ..., \gamma_I; \theta).$$

We term $w^* - q^*$ the *minimax (duality) gap*. We note that there are key differences between the naïve ensembling model and ours. In the naïve ensembling model, one trains multiple GAN models *independently* and averages their outputs. In contrast, our Stackelberg GAN shares a unique discriminator for various generators, thus requires *jointly training*. Figure 4.6 shows the architecture of our Stackelberg GAN.

How to generate samples from Stackelberg GAN? In the Stackelberg GAN, we expect that each generator learns only a few modes. In order to generate a sample that may come from all modes, we use a mixed model. In particular, we generate a uniformly random value i from 1 to I and use the i-th generator to obtain a new sample. Note that this procedure in independent of the training procedure.

4.2.3 Our results on optimization

In this section, we develop our theoretical contributions and compare our results with the prior work.

Minimax duality gap

We begin with studying the minimax gap of Stackelberg GAN. Our main results show that the minimax gap shrinks as the number of generators increases.

To proceed, denote by $h_i(\mathbf{u}_i) := \inf_{\gamma_i \in \mathbb{R}^g} (-\phi(\gamma_i; \cdot))^*(\mathbf{u}_i)$, where the conjugate operation is w.r.t. the second argument of $\phi(\gamma_i; \cdot)$. We clarify here that the subscript i in h_i indicates that the function h_i is derived from the i-th generator. The argument of h_i should depend on i, so we denote it by $\mathbf{u}_i$. Intuitively, h_i serves as an approximate convexification of $-\phi(\gamma_i, \cdot)$ w.r.t the second argument due to the conjugate operation. Denote by $\breve{\mathsf{cl}} h_i$ the convex closure of h_i:

$$\breve{\mathsf{cl}} h_i(\widetilde{\mathbf{u}}) := \inf_{\{a^j\}, \{\mathbf{u}_i^j\}} \left\{ \sum_{j=1}^{t+2} a^j h_i(\mathbf{u}_i^j) : \widetilde{\mathbf{u}} = \sum_{j=1}^{t+2} a^j \mathbf{u}_i^j, \sum_{j=1}^{t+2} a^j = 1, a^j \geq 0 \right\}.$$

$\breve{\mathsf{cl}} h_i$ represents the convex relaxation of h_i because the epigraph of $\breve{\mathsf{cl}} h_i$ is exactly the convex hull of epigraph of h_i by the definition of $\breve{\mathsf{cl}} h_i$. Let $\Delta_\theta^{\mathsf{minimax}} = \inf_{\gamma_1, ..., \gamma_I \in \mathbb{R}^g} \sup_{\theta \in \mathbb{R}^t} \frac{1}{I} \Phi(\gamma_1, ..., \gamma_I; \theta) - \inf_{\gamma_1, ..., \gamma_I \in \mathbb{R}^g} \sup_{\theta \in \mathbb{R}^t} \frac{1}{I} \widetilde{\Phi}(\gamma_1, ..., \gamma_I; \theta)$, and $\Delta_\theta^{\mathsf{maximin}} = \sup_{\theta \in \mathbb{R}^t} \inf_{\gamma_1, ..., \gamma_I \in \mathbb{R}^g} \frac{1}{I} \widetilde{\Phi}(\gamma_1, ..., \gamma_I; \theta) - \sup_{\theta \in \mathbb{R}^t} \inf_{\gamma_1, ..., \gamma_I \in \mathbb{R}^g} \frac{1}{I} \Phi(\gamma_1, ..., \gamma_I; \theta)$, where $\widetilde{\Phi}(\gamma_1, ..., \gamma_I; \theta) := \sum_{i=1}^I \widehat{\mathsf{cl}} \phi(\gamma_i; \theta)$ and $-\widehat{\mathsf{cl}} \phi(\gamma_i; \theta)$ is the convex closure of $-\phi(\gamma_i; \theta)$ w.r.t. argument θ. Therefore, $\Delta_\theta^{\mathsf{maximin}} + \Delta_\theta^{\mathsf{minimax}}$ measures the non-convexity of objective function w.r.t. argument θ. For example, it is equal to 0 if and only if $\phi(\gamma_i; \theta)$ is concave and closed w.r.t. discriminator parameter θ.

We have the following guarantees on the minimax gap of Stackelberg GAN.

Theorem 49. *Let $\Delta_\gamma^i := \sup_{\mathbf{u} \in \mathbb{R}^t} \{h_i(\mathbf{u}) - \breve{\mathsf{cl}} h_i(\mathbf{u})\} \geq 0$ and $\Delta_\gamma^{\mathsf{worst}} := \max_{i \in [I]} \Delta_\gamma^i$. Denote by t the number of parameters of discriminator, i.e., $\theta \in \mathbb{R}^t$. Suppose that $h_i(\cdot)$ is continuous and $\mathsf{dom} h_i$ is compact and convex. Then the duality gap can be bounded by*

$$0 \leq w^* - q^* \leq \Delta_\theta^{\mathsf{minimax}} + \Delta_\theta^{\mathsf{maximin}} + \epsilon,$$

provided that the number of generators $I > \frac{t+1}{\epsilon} \Delta_\gamma^{\mathsf{worst}}$.

Remark 7. *Theorem 49 makes mild assumption on the continuity of loss and no assumption on the model capacity of discriminator and generators. The analysis instead depends on their non-convexity as being parametrized by deep neural networks. In particular, Δ_γ^i measures the divergence between the function value of h_i and its convex relaxation $\breve{\mathsf{cl}} h_i$; When $\phi(\gamma_i; \theta)$ is convex w.r.t. argument γ_i, Δ_γ^i is exactly 0. The constant $\Delta_\gamma^{\mathsf{worst}}$ is the maximal divergence among all generators, which does not grow with the increase of I. This is because $\Delta_\gamma^{\mathsf{worst}}$ measures the divergence of only* one generator *and when each generator for example has the same architecture, we have $\Delta_\gamma^{\mathsf{worst}} = \Delta_\gamma^1 = ... = \Delta_\gamma^I$. Similarly, the terms $\Delta_\theta^{\mathsf{minimax}}$ and $\Delta_\theta^{\mathsf{maximin}}$ characterize the non-convexity of discriminator. When the discriminator is concave such as logistic regression and support vector machine, $\Delta_\theta^{\mathsf{minimax}} = \Delta_\theta^{\mathsf{maximin}} = 0$ and we have the following straightforward corollary about the minimax duality gap of Stackelberg GAN.*

Corollary 1. *Under the settings of Theorem 49, when $\phi(\gamma_i; \theta)$ is concave and closed w.r.t. discriminator parameter θ and the number of generators $I > \frac{t+1}{\epsilon} \Delta_\gamma^{\mathsf{worst}}$, we have $0 \leq w^* - q^* \leq \epsilon$.*

Existence of approximate equilibrium

The results of Theorem 49 and Corollary 1 are independent of model capacity of generators and discriminator. When we make assumptions on the expressive power of generator as in [9], we have the following guarantee (4.23) on the existence of ϵ-approximate equilibrium.

Theorem 50. *Under the settings of Theorem 49, suppose that for any $\xi > 0$, there exists a generator G such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{P}_d,\mathbf{z}\sim\mathcal{P}_\mathbf{z}}\|G(\mathbf{z})-\mathbf{x}\|_2 \leq \xi$. Let the discriminator and the generators be L-Lipschitz w.r.t. inputs and parameters, respectively. Then for any $\epsilon > 0$, there exist $I = \frac{t+1}{\epsilon}\Delta_\gamma^{\text{worst}}$ generators $G_{\gamma_1^*}, ..., G_{\gamma_I^*}$ and a discriminator D_{θ^*} such that for some value $V \in \mathbb{R}$,*

$$\begin{aligned} &\forall \gamma_1, ..., \gamma_I \in \mathbb{R}^g, \quad && \Phi(\gamma_1, ..., \gamma_I; \theta^*) \leq V + \epsilon, \\ &\forall \theta \in \mathbb{R}^t, && \Phi(\gamma_1^*, ..., \gamma_I^*; \theta) \geq V - \epsilon. \end{aligned} \tag{4.23}$$

Related work. While many efforts have been devoted to empirically investigating the performance of multi-generator GAN, little is known about how many generators are needed so as to achieve certain equilibrium guarantees. Probably the most relevant prior work to Theorem 50 is that of [9]. In particular, [9] showed that there exist $I = \frac{100t}{\epsilon^2}\Delta^2$ generators and one discriminator such that ϵ-approximate equilibrium can be achieved, provided that for *all* $\mathbf{x}$ and any $\xi > 0$, there exists a generator G such that $\mathbb{E}_{\mathbf{z}\sim\mathcal{P}_\mathbf{z}}\|G(\mathbf{z})-\mathbf{x}\|_2 \leq \xi$. Hereby, Δ is a global upper bound of function $|f|$, i.e., $f \in [-\Delta, \Delta]$. In comparison, Theorem 50 improves over this result in two aspects: a) the assumption on the expressive power of generators in [9] implies our condition $\mathbb{E}_{\mathbf{x}\sim\mathcal{P}_d,\mathbf{z}\sim\mathcal{P}_\mathbf{z}}\|G(\mathbf{z})-\mathbf{x}\|_2 \leq \xi$. Thus our assumption is weaker. b) The required number of generators in Theorem 50 is as small as $\frac{t+1}{\epsilon}\Delta_\gamma^{\text{worst}}$. We note that $\Delta_\gamma^{\text{worst}} \ll 2\Delta$ by the definition of $\Delta_\gamma^{\text{worst}}$. Therefore, Theorem 50 requires much fewer generators than that of [9].

4.2.4 Experimental results

In this section, we verify our theoretical contributions by the experimental validation.

MNIST dataset

We first show that Stackelberg GAN generates more diverse images on the MNIST dataset [144] than classic GAN. We follow the standard preprocessing step that each pixel is normalized via subtracting it by 0.5 and dividing it by 0.5.

Figure 4.7 shows the diversity of generated digits by Stackelberg GAN with varying number of generators. When there is only one generator, the digits are not very diverse with many "1"'s and much fewer "2"'s. As the number of generators increases, the images tend to be more diverse. In particular, for 10-generator Stackelberg GAN, each generator is associated with one or two digits without any digit being missed.

Fashion-MNIST dataset

We also observe better performance by the Stackelberg GAN on the Fashion-MNIST dataset. Fashion-MNIST is a dataset which consists of 60,000 examples. Each example is a 28×28

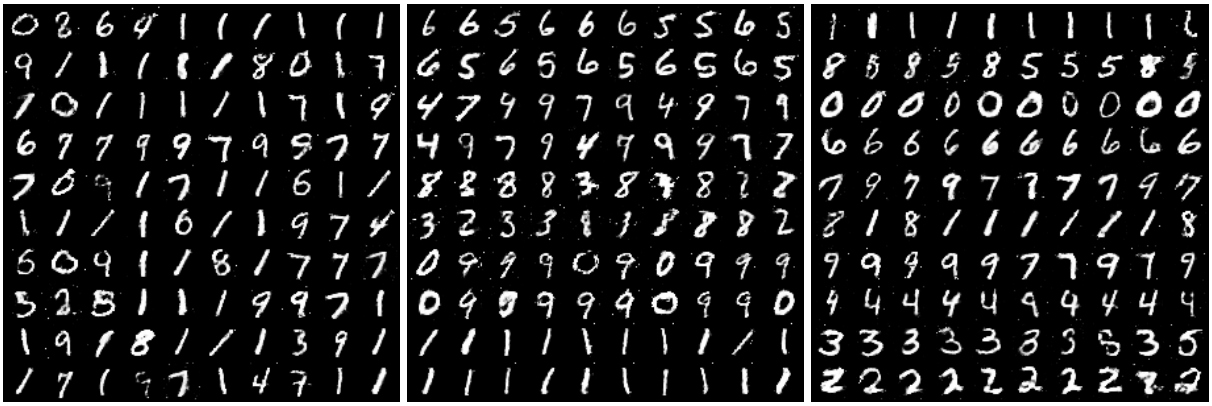


Figure 4.7: Standard GAN vs. Stackelberg GAN on the MNIST dataset without cherry pick. **Left Figure:** Digits generated by the standard GAN. It shows that the standard GAN generates many "1"'s which are not very diverse. **Middle Figure:** Digits generated by the Stackelberg GAN with 5 generators, where every two rows correspond to one generator. **Right Figure:** Digits generated by the Stackelberg GAN with 10 generators, where each row corresponds to one generator.

grayscale image associating with a label from 10 classes. We follow the standard preprocessing step that each pixel is normalized via subtracting it by 0.5 and dividing it by 0.5.

Figure 4.8 shows the diversity of generated fashions by Stackelberg GAN with varying number of generators. When there is only one generator, the generated images are not very diverse without any "bags" being found. However, as the number of generators increases, the generated images tend to be more diverse. In particular, for 10-generator Stackelberg GAN, each generator is associated with one class without any class being missed.

CIFAR-10 dataset

We then implement Stackelberg GAN on the CIFAR-10 dataset. CIFAR-10 includes 60,000 32×32 training images, which fall into 10 classes [140]). The architecture of generators and discriminator follows the design of DCGAN in [189]. We train models with 5, 10, and 20 fixed-size generators. The results show that the model with 10 generators performs the best. We also train 10-generator models where each generator has 2, 3 and 4 convolution layers. We find that the generator with 2 convolution layers, which is the most shallow one, performs the best. So we report the results obtained from the model with 10 generators containing 2 convolution layers. Figure 4.9 shows the samples produced by different generators. The samples are randomly drawn instead of being cherry-picked to demonstrate the quality of images generated by our model.

For quantitative evaluation, we use Inception score and Fréchet Inception Distance (FID) to measure the difference between images generated by models and real images.

Results of Inception score. The Inception score measures the quality of a generated image and is correlated well with human's judgment [202]. We report the Inception score obtained by our Stackelberg GAN and other baseline methods in Table 4.2. For fair comparison, we only consider the baseline models which are completely unsupervised model and do not need any

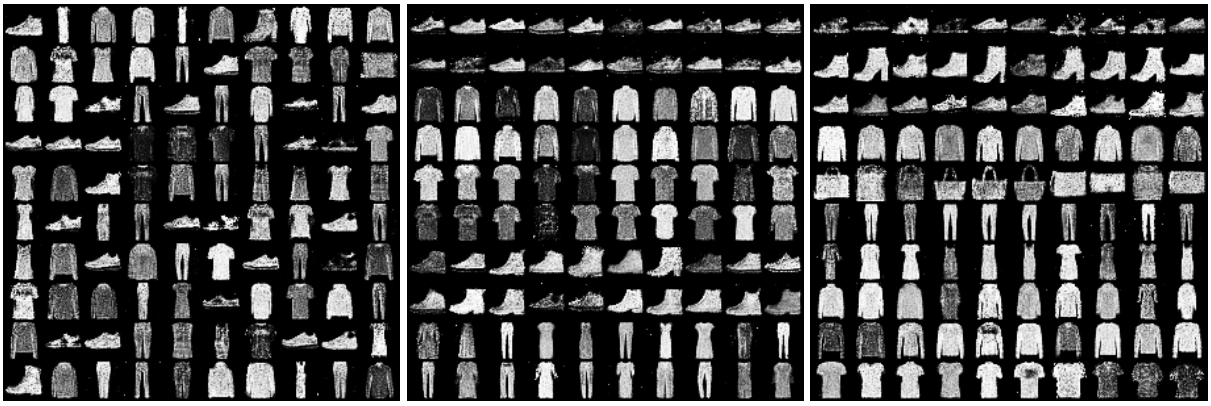


Figure 4.8: Generated samples by Stackelberg GAN on fashion-MNIST dataset without cherry pick. **Left Figure:** Examples generated by the standard GAN. It shows that the standard GAN fails to generate bags. **Middle Figure:** Examples generated by the Stackelberg GAN with 5 generators, where every two rows correspond to one generator. **Right Figure:** Examples generated by the Stackelberg GAN with 10 generators, where each row corresponds to one generator.

label information. Instead of directly using the reported Inception scores by original papers, we replicate the experiment of *MGAN* using the code, architectures and parameters reported by their original papers, and evaluate the scores based on the new experimental results. Table 4.2 shows that our model achieves a score of 7.62 in CIFAR-10 dataset, which outperforms the state-of-the-art models. For fairness, we configure our Stackelberg GAN with the same capacity as MGAN, that is, the two models have comparative number of total parameters. When the capacity of our Stackelberg GAN is as small as DCGAN, our model improves over DCGAN significantly.

Results of Fréchet Inception distance. We then evaluate the performance of models on CIFAR-10 dataset using the Fréchet Inception Distance (FID), which better captures the similarity between generated images and real ones [117]. As Table 4.2 shows, under the same capacity as DCGAN, our model reduces the FID by 20.74%. Meanwhile, under the same capacity as MGAN, our model reduces the FID by 14.61%. This improvement further indicates that our Stackelberg GAN with multiple light-weight generators help improve the quality of the generated images.

Tiny ImageNet dataset

We also evaluate the performance of Stackelberg GAN on the Tiny ImageNet dataset. The Tiny ImageNet is a large image dataset, where each image is labelled to indicate the class of the object inside the image. We resize the figures down to 32×32 following the procedure described in [65]. Figure 4.9 shows the randomly picked samples generated by 10-generator Stackelberg GAN. Each row has samples generated from one generator.

Table 4.2: Quantitative evaluation of various GANs on CIFAR-10 dataset. All results are either reported by the authors themselves or run by us with codes provided by the authors. Every model is trained *without label*. Methods with higher inception score and lower Fréchet Inception Distance are better.

Model	Inception Score	Fréchet Inception Distance
Real data	11.24 ± 0.16	-
WGAN [7]	3.82 ± 0.06	-
MIX+WGAN [9]	4.04 ± 0.07	-
Improved-GAN [202]	4.36 ± 0.04	-
ALI [78]	5.34 ± 0.05	-
BEGAN [34]	5.62	-
MAGAN [237]	5.67	-
GMAN [79]	6.00 ± 0.19	-
DCGAN [189]	6.40 ± 0.05	37.7
Ours (capacity as DCGAN)	$\mathbf{7.02 \pm 0.07}$	**29.88**
D2GAN [180]	7.15 ± 0.07	-
MAD-GAN (our run, capacity $1\times$MGAN) [91]	6.67 ± 0.07	34.10
MGAN (our run) [118]	7.52 ± 0.1	31.34
Ours (capacity $1\times$MGAN$\approx 1.8\times$DCGAN)	$\mathbf{7.62 \pm 0.07}$	**26.76**

4.2.5 Proofs of our main results

Useful lemmas

Lemma 69. *Given the function*

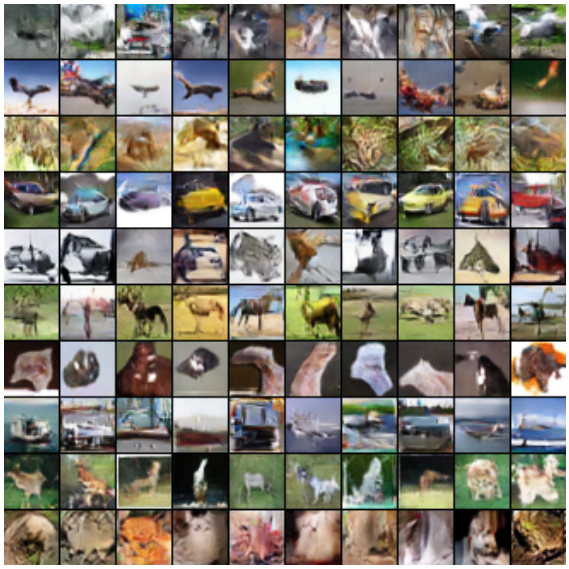
$$(f_1 + ... + f_I)(\theta) := f_1(\theta) + ... + f_I(\theta),$$

where $f_i : \mathbb{R}^t \to \mathbb{R}$, $i \in [I]$ are closed proper convex functions. Denote by $f_1^* \oplus ... \oplus f_I^*$ the infimal convolution

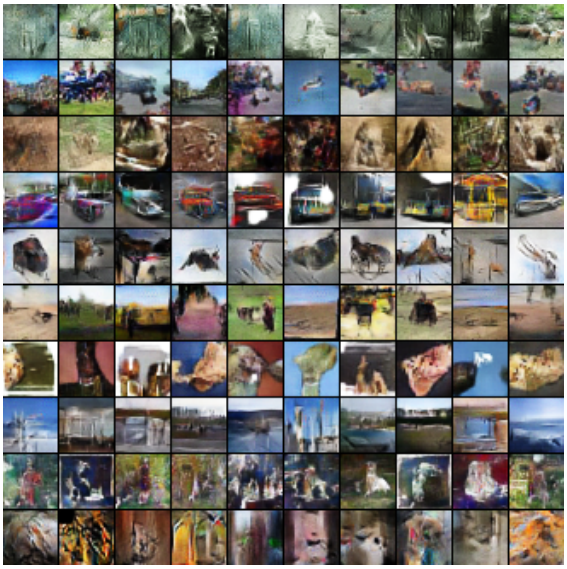
$$(f_1^* \oplus ... \oplus f_I^*)(\mathbf{u}) := \inf_{\mathbf{u}_1 + ... + \mathbf{u}_I = \mathbf{u}} \{f_1^*(\mathbf{u}_1) + ... + f_I^*(\mathbf{u}_I)\}, \qquad \mathbf{u} \in \mathbb{R}^t.$$

Provided that $f_1 + ... + f_I$ is proper, then we have

$$(f_1 + ... + f_I)^*(\mathbf{u}) = \mathsf{cl}(f_1^* \oplus ... \oplus f_I^*)(\mathbf{u}), \quad \forall \mathbf{u} \in \mathbb{R}^t.$$



(a) Samples on CIFAR-10.



(b) Samples on Tiny ImageNet.

Figure 4.9: Examples generated by Stackelberg GAN on CIFAR-10 (left) and Tiny ImageNet (right) without cherry pick, where each row corresponds to samples from one generator.

Proof. For all $\theta \in \mathbb{R}^t$, we have

$$\begin{aligned} f_1(\theta) + ... + f_I(\theta) &= \sup_{\mathbf{u}_1}\{\theta^T\mathbf{u}_1 - f_1^*(\mathbf{u}_1)\} + ... + \sup_{\mathbf{u}_I}\{\theta^T\mathbf{u}_I - f_I^*(\mathbf{u}_I)\} \\ &= \sup_{\mathbf{u}_1,...,\mathbf{u}_I}\{\theta^T(\mathbf{u}_1 + ... + \mathbf{u}_I) - f_1^*(\mathbf{u}_1) - ... - f_I^*(\mathbf{u}_I)\} \\ &= \sup_{\mathbf{u}}\sup_{\mathbf{u}_1+...+\mathbf{u}_I=\mathbf{u}}\left\{\theta^T\mathbf{u} - f_1^*(\mathbf{u}_1) - ... - f_I^*(\mathbf{u}_I)\right\} \\ &= \sup_{\mathbf{u}}\left\{\theta^T\mathbf{u} - \inf_{\mathbf{u}_1+...+\mathbf{u}_I=\mathbf{u}} f_1^*(\mathbf{u}_1) - ... - f_I^*(\mathbf{u}_I)\right\} \\ &= \sup_{\mathbf{u}}\left\{\theta^T\mathbf{u} - (f_1^* \oplus ... \oplus f_I^*)(\mathbf{u})\right\} \\ &= (f_1^* \oplus ... \oplus f_I^*)^*(\theta). \end{aligned} \tag{4.24}$$

Therefore,

$$\mathsf{cl}(f_1^* \oplus ... \oplus f_I^*)(\mathbf{u}) = \breve{\mathsf{cl}}(f_1^* \oplus ... \oplus f_I^*)(\mathbf{u}) = (f_1^* \oplus ... \oplus f_I^*)^{**}(\mathbf{u}) = (f_1 + ... + f_I)^*(\mathbf{u}),$$

where the first equality holds because $(f_1^* \oplus ... \oplus f_I^*)$ is convex, the second quality is by standard conjugate theorem, and the last equality holds by conjugating the both sides of Eqn. (4.24). □

Lemma 70 (Proposition 3.4 (b), [35])**.** *For any function* $p(\mathbf{u})$*, denote by* $q(\mu) := \inf_{\mathbf{u}\in\mathbb{R}^t}\{p(\mathbf{u}) + \mu^T\mathbf{u}\}$*. We have* $\sup_{\mu\in\mathbb{R}^t} q(\mu) = \breve{\mathsf{cl}}p(\mathbf{0})$.

Proofs of Theorem 49 and Corollary 1

Theorem 49 (restated). *Let* $\Delta_\gamma^i := \sup_{\mathbf{u}\in\mathbb{R}^t}\{h_i(\mathbf{u}) - \breve{\mathsf{cl}}h_i(\mathbf{u})\} \geq 0$ *and* $\Delta_\gamma^{\mathsf{worst}} := \max_{i\in[I]}\Delta_\gamma^i$. *Denote by* t *the number of parameters of discriminator, i.e.,* $\theta \in \mathbb{R}^t$. *Suppose that* $h_i(\cdot)$ *is continuous and* $\mathsf{dom}h_i$ *is compact and convex. Then the duality gap can be bounded by*

$$0 \leq w^* - q^* \leq \Delta_\theta^{\mathsf{minimax}} + \Delta_\theta^{\mathsf{maximin}} + \epsilon,$$

provided that the number of generators $I > \frac{t+1}{\epsilon}\Delta_\gamma^{\mathsf{worst}}$.

Proof. The statement $0 \leq w^* - q^*$ is by the weak duality. Thus it suffices to prove the other side of the inequality. All notations in this section are defined in Section 4.2.3.

We first show that

$$\inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}\sup_{\theta\in\mathbb{R}^t}\frac{1}{I}\widetilde{\Phi}(\gamma_1,...,\gamma_I;\theta) - \sup_{\theta\in\mathbb{R}^t}\inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}\frac{1}{I}\widetilde{\Phi}(\gamma_1,...,\gamma_I;\theta) \leq \epsilon.$$

Denote by

$$p(\mathbf{u}) := \inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}\sup_{\theta\in\mathbb{R}^t}\left\{\widetilde{\Phi}(\gamma_1,...,\gamma_I;\theta) - \mathbf{u}^T\theta\right\}.$$

We have the following lemma.

Lemma 71. *We have*

$$\sup_{\theta\in\mathbb{R}^t}\inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}\widetilde{\Phi}(\gamma_1,...,\gamma_I;\theta) = (\breve{\mathsf{cl}}p)(\mathbf{0}) \leq p(\mathbf{0}) = \inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}\sup_{\theta\in\mathbb{R}^t}\widetilde{\Phi}(\gamma_1,...,\gamma_I;\theta).$$

Proof. By the definition of $p(\mathbf{0})$, we have $p(\mathbf{0}) = \inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}\sup_{\theta\in\mathbb{R}^t}\widetilde{\Phi}(\gamma_1,...,\gamma_I;\theta)$. Since $(\breve{\mathsf{cl}}p)(\cdot)$ is the convex closure of function $p(\cdot)$ (a.k.a. weak duality theorem), we have $(\breve{\mathsf{cl}}p)(\mathbf{0}) \leq p(\mathbf{0})$. We now show that $\sup_{\theta\in\mathbb{R}^t}\inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}\widetilde{\Phi}(\gamma_1,...,\gamma_I;\theta) = (\breve{\mathsf{cl}}p)(\mathbf{0})$. Note that $p(\mathbf{u}) = \inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}p_{\gamma_1,...,\gamma_I}(\mathbf{u})$, where

$$p_{\gamma_1,...,\gamma_I}(\mathbf{u}) = \sup_{\theta\in\mathbb{R}^t}\{\widetilde{\Phi}(\gamma_1,...,\gamma_I;\theta) - \mathbf{u}^T\theta\} = (-\widetilde{\Phi}(\gamma_1,...,\gamma_I;\cdot))^*(-\mathbf{u}),$$

and that

$$\begin{aligned}
&\inf_{\mathbf{u}\in\mathbb{R}^t}\{p_{\gamma_1,...,\gamma_I}(\mathbf{u}) + \mathbf{u}^T\mu\}\\
&= -\sup_{\mathbf{u}\in\mathbb{R}^t}\{\mathbf{u}^T(-\mu) - p_{\gamma_1,...,\gamma_I}(\mathbf{u})\}\\
&= -(p_{\gamma_1,...,\gamma_I})^*(-\mu) \quad \text{(by the definition of conjugate function)}\\
&= -(-\widetilde{\Phi}(\gamma_1,...,\gamma_I;\cdot))^{**}(\mu) = \widetilde{\Phi}(\gamma_1,...,\gamma_I;\mu). \quad \text{(by conjugate theorem)}
\end{aligned} \tag{4.25}$$

So we have

$$\begin{aligned}
&(\breve{\mathsf{cl}}p)(\mathbf{0})\\
&= \sup_{\mu\in\mathbb{R}^t}\inf_{\mathbf{u}\in\mathbb{R}^t}\{p(\mathbf{u}) + \mathbf{u}^T\mu\} \quad \text{(by Lemma 70)}\\
&= \sup_{\mu\in\mathbb{R}^t}\inf_{\mathbf{u}\in\mathbb{R}^t}\inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}\{p_{\gamma_1,...,\gamma_I}(\mathbf{u}) + \mathbf{u}^T\mu\} \quad \text{(by the definition of } p(\mathbf{u}))\\
&= \sup_{\mu\in\mathbb{R}^t}\inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}\inf_{\mathbf{u}\in\mathbb{R}^t}\{p_{\gamma_1,...,\gamma_I}(\mathbf{u}) + \mathbf{u}^T\mu\} = \sup_{\mu\in\mathbb{R}^t}\inf_{\gamma_1,...,\gamma_I\in\mathbb{R}^g}\widetilde{\Phi}(\gamma_1,...,\gamma_I;\mu), \quad \text{(by Eqn. (4.25))}
\end{aligned}$$

□

as desired.

By Lemma 71, it suffices to show $p(\mathbf{0}) - (\breve{\mathsf{cl}}p)(\mathbf{0}) \leq (t+1)\Delta_\gamma^{\mathsf{worst}}$. We have the following lemma.

Lemma 72. *Under the assumption in Theorem 49, $p(\mathbf{0}) - (\breve{\mathsf{cl}}p)(\mathbf{0}) \leq (t+1)\Delta_\gamma^{\mathsf{worst}}$.*

Proof. We note that

$$
\begin{aligned}
p(\mathbf{u}) &:= \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \sup_{\theta \in \mathbb{R}^t} \left\{ \widetilde{\Phi}(\gamma_1, ..., \gamma_I; \theta) - \mathbf{u}^T\theta \right\} \\
&= \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \sup_{\theta \in \mathbb{R}^t} \left\{ \sum_{i=1}^I \widehat{\mathsf{cl}}\phi(\gamma_i; \theta) - \mathbf{u}^T\theta \right\} \quad \text{(by the definition of } \widetilde{\Phi}) \\
&= \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \left(\sum_{i=1}^I -\widehat{\mathsf{cl}}\phi(\gamma_i; \cdot) \right)^* (-\mathbf{u}) \quad \text{(by the definition of conjugate function)} \\
&= \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \inf_{\mathbf{u}_1+...+\mathbf{u}_I=-\mathbf{u}} \left\{ \sum_{i=1}^I (-\widehat{\mathsf{cl}}\phi(\gamma_i; \cdot))^*(\mathbf{u}_i) \right\} \quad \text{(by Lemma 69)} \\
&= \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \inf_{\mathbf{u}_1+...+\mathbf{u}_I=-\mathbf{u}} \left\{ \sum_{i=1}^I (-\phi(\gamma_i; \cdot))^*(\mathbf{u}_i) \right\} \quad \text{(by conjugate theorem)} \\
&= \inf_{\mathbf{u}_1+...+\mathbf{u}_I=-\mathbf{u}} \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \{(-\phi(\gamma_1; \cdot))^*(\mathbf{u}_1) + ... + (-\phi(\gamma_I; \cdot))^*(\mathbf{u}_I)\} \\
&=: \inf_{\mathbf{u}_1+...+\mathbf{u}_I=-\mathbf{u}} \{h_1(\mathbf{u}_1) + ... + h_I(\mathbf{u}_I)\}, \quad \text{(by the definition of } h_i(\cdot))
\end{aligned}
$$

where $\mathbf{u}_1, ..., \mathbf{u}_I, \mathbf{u} \in \mathbb{R}^t$. Therefore,

$$
p(\mathbf{0}) = \inf_{\mathbf{u}_1,...,\mathbf{u}_I \in \mathbb{R}^t} \sum_{i=1}^I h_i(\mathbf{u}_i), \quad \text{s.t.} \quad \sum_{i=1}^I \mathbf{u}_i = \mathbf{0}.
$$

Consider the subset of $\mathbb{R}^{t+1}$:

$$
\mathcal{Y}_i := \left\{ \mathbf{y}_i \in \mathbb{R}^{t+1} : \mathbf{y}_i = [\mathbf{u}_i, h_i(\mathbf{u}_i)], \mathbf{u}_i \in \mathsf{dom}h_i \right\}, \quad i \in [I].
$$

Define the vector summation

$$
\mathcal{Y} := \mathcal{Y}_1 + \mathcal{Y}_2 + ... + \mathcal{Y}_I.
$$

Since $h_i(\cdot)$ is continuous and $\mathsf{dom}h_i$ is compact, the set

$$
\{(\mathbf{u}_i, h_i(\mathbf{u}_i)) : \mathbf{u}_i \in \mathsf{dom}h_i\}
$$

is compact. So $\mathcal{Y}$, $\mathsf{conv}(\mathcal{Y})$, $\mathcal{Y}_i$, and $\mathsf{conv}(\mathcal{Y}_i)$, $i \in [I]$ are all compact sets. According to the definition of $\mathcal{Y}$ and the standard duality argument [35], we have

$$
p(\mathbf{0}) = \inf \{w : \text{there exists } (\mathbf{r}, w) \in \mathcal{Y} \text{ such that } \mathbf{r} = \mathbf{0}\},
$$

and

$$\breve{\mathsf{cl}}p(\mathbf{0}) = \inf\{w : \text{there exists } (\mathbf{r}, w) \in \mathsf{conv}(\mathcal{Y}) \text{ such that } \mathbf{r} = \mathbf{0}\}.$$

We are going to apply the following Shapley-Folkman lemma.

Lemma 73 (Shapley-Folkman, [216])**.** *Let $\mathcal{Y}_i, i \in [I]$ be a collection of subsets of $\mathbb{R}^m$. Then for every $\mathbf{y} \in \mathsf{conv}(\sum_{i=1}^I \mathcal{Y}_i)$, there is a subset $\mathcal{I}(\mathbf{y}) \subseteq [I]$ of size at most m such that*

$$\mathbf{y} \in \left[\sum_{i \notin \mathcal{I}(\mathbf{y})} \mathcal{Y}_i + \sum_{i \in \mathcal{I}(\mathbf{y})} \mathsf{conv}(\mathcal{Y}_i)\right].$$

We apply Lemma 73 to prove Lemma 72 with $m = t + 1$. Let $(\overline{\mathbf{r}}, \overline{w}) \in \mathsf{conv}(\mathcal{Y})$ be such that

$$\overline{\mathbf{r}} = \mathbf{0}, \quad \text{and} \quad \overline{w} = \breve{\mathsf{cl}}p(\mathbf{0}).$$

Applying the above Shapley-Folkman lemma to the set $\mathcal{Y} = \sum_{i=1}^I \mathcal{Y}_i$, we have that there are a subset $\overline{\mathcal{I}} \subseteq [I]$ of size $t + 1$ and vectors

$$(\overline{\mathbf{r}}_i, \overline{w}_i) \in \mathsf{conv}(\mathcal{Y}_i), \ \ i \in \overline{\mathcal{I}} \qquad \text{and} \qquad \overline{\mathbf{u}}_i \in \mathsf{dom} h_i, \ \ i \notin \overline{\mathcal{I}},$$

such that

$$\sum_{i \notin \overline{\mathcal{I}}} \overline{\mathbf{u}}_i + \sum_{i \in \overline{\mathcal{I}}} \overline{\mathbf{r}}_i = \overline{\mathbf{r}} = \mathbf{0}, \tag{4.26}$$

$$\sum_{i \notin \overline{\mathcal{I}}} h_i(\overline{\mathbf{u}}_i) + \sum_{i \in \overline{\mathcal{I}}} \overline{w}_i = \breve{\mathsf{cl}}p(\mathbf{0}). \tag{4.27}$$

Representing elements of the convex hull of $\mathcal{Y}_i \subseteq \mathbb{R}^{t+1}$ by Carathéodory theorem, we have that for each $i \in \overline{\mathcal{I}}$, there are vectors $\{\mathbf{u}_i^j\}_{j=1}^{t+2}$ and scalars $\{a_i^j\}_{j=1}^{t+2} \in \mathbb{R}$ such that

$$\sum_{j=1}^{t+2} a_i^j = 1, \quad a_i^j \geq 0, \ j \in [t+2],$$

$$\overline{\mathbf{r}}_i = \sum_{j=1}^{t+2} a_i^j \mathbf{u}_i^j =: \overline{\mathbf{u}}_i \in \mathsf{dom} h_i, \qquad \overline{w}_i = \sum_{j=1}^{t+2} a_i^j h_i(\mathbf{u}_i^j). \tag{4.28}$$

Recall that we define

$$\breve{\mathsf{cl}} h_i(\widetilde{\mathbf{u}}) := \inf_{\{a^j\}, \{\mathbf{u}_i^j\}} \left\{\sum_{j=1}^{t+2} a^j h_i(\mathbf{u}_i^j) : \widetilde{\mathbf{u}} = \sum_{j=1}^{t+2} a^j \mathbf{u}_i^j, \sum_{j=1}^{t+2} a^j = 1, a^j \geq 0\right\},$$

and $\Delta_\gamma^i := \sup_{\mathbf{u} \in \mathbb{R}^t} \{h_i(\mathbf{u}) - \breve{\mathsf{cl}} h_i(\mathbf{u})\} \geq 0$. We have for $i \in \overline{\mathcal{I}}$,

$$\begin{aligned} \overline{w}_i &\geq \breve{\mathsf{cl}} h_i \left(\sum_{j=1}^{t+2} a_i^j \mathbf{u}_i^j\right) \quad \text{(by the definition of } \breve{\mathsf{cl}} h_i(\cdot)) \\ &\geq h_i \left(\sum_{j=1}^{t+2} a_i^j \mathbf{u}_i^j\right) - \Delta_\gamma^i \quad \text{(by the definition of } \Delta_\gamma^i) \\ &= h_i\left(\overline{\mathbf{u}}_i\right) - \Delta_\gamma^i. \quad \text{(by Eqn. (4.28))} \end{aligned} \tag{4.29}$$

Thus, by Eqns. (4.26) and (4.28), we have

$$\sum_{i=1}^{I} \overline{\mathbf{u}}_i = \mathbf{0}, \quad \overline{\mathbf{u}}_i \in \mathsf{dom} h_i,\ i \in [I]. \tag{4.30}$$

Therefore, we have

$$\begin{aligned} p(\mathbf{0}) &= \sum_{i=1}^{I} h_i(\overline{\mathbf{u}}_i) \quad \text{(by Eqn. (4.30))} \\ &\leq \breve{\mathsf{cl}} p(\mathbf{0}) + \sum_{i \in \overline{\mathcal{I}}} \Delta_\gamma^i \quad \text{(by Eqns. (4.27) and (4.29))} \\ &\leq \breve{\mathsf{cl}} p(\mathbf{0}) + |\overline{\mathcal{I}}| \Delta_\gamma^{\mathsf{worst}} \\ &= \breve{\mathsf{cl}} p(\mathbf{0}) + (t+1)\Delta_\gamma^{\mathsf{worst}}, \quad \text{(by Lemma 73)} \end{aligned}$$

as desired. □

By Lemmas 71 and 72, we have proved that

$$\inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \sup_{\theta \in \mathbb{R}^t} \frac{1}{I} \widetilde{\Phi}(\gamma_1, ..., \gamma_I; \theta) - \sup_{\theta \in \mathbb{R}^t} \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \frac{1}{I} \widetilde{\Phi}(\gamma_1, ..., \gamma_I; \theta) \leq \epsilon.$$

To prove Theorem 49, we note that

$$\begin{aligned} w^* - q^* &:= \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \sup_{\theta \in \mathbb{R}^t} \frac{1}{I} \Phi(\gamma_1, ..., \gamma_I; \theta) - \sup_{\theta \in \mathbb{R}^t} \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \frac{1}{I} \Phi(\gamma_1, ..., \gamma_I; \theta) \\ &= \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \sup_{\theta \in \mathbb{R}^t} \frac{1}{I} \Phi(\gamma_1, ..., \gamma_I; \theta) - \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \sup_{\theta \in \mathbb{R}^t} \frac{1}{I} \widetilde{\Phi}(\gamma_1, ..., \gamma_I; \theta) \\ &\quad + \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \sup_{\theta \in \mathbb{R}^t} \frac{1}{I} \widetilde{\Phi}(\gamma_1, ..., \gamma_I; \theta) - \sup_{\theta \in \mathbb{R}^t} \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \frac{1}{I} \widetilde{\Phi}(\gamma_1, ..., \gamma_I; \theta) \\ &\quad + \sup_{\theta \in \mathbb{R}^t} \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \frac{1}{I} \widetilde{\Phi}(\gamma_1, ..., \gamma_I; \theta) - \sup_{\theta \in \mathbb{R}^t} \inf_{\gamma_1,...,\gamma_I \in \mathbb{R}^g} \frac{1}{I} \Phi(\gamma_1, ..., \gamma_I; \theta) \\ &\leq \Delta_\theta^{\mathsf{minimax}} + \Delta_\theta^{\mathsf{maximin}} + \epsilon, \end{aligned}$$

as desired. □

Corollary 1 (restated). *Under the settings of Theorem 49, when $\phi(\gamma_i; \theta)$ is concave and closed w.r.t. discriminator parameter θ and the number of generators $I > \frac{t+1}{\epsilon} \Delta_\gamma^{\mathsf{worst}}$, we have $0 \leq w^* - q^* \leq \epsilon$.*

Proof. When $\phi(\gamma_i; \theta)$ is concave and closed w.r.t. discriminator parameter θ, we have $\widehat{\mathsf{cl}} \phi = \phi$. Thus, $\Delta_\theta^{\mathsf{minimax}} = \Delta_\theta^{\mathsf{maximin}} = 0$ and $0 \leq w^* - q^* \leq \epsilon$. □

Proofs of Theorem 50

Theorem 50 (restated). *Under the settings of Theorem 49, suppose that for any $\xi > 0$, there exists a generator G such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{P}_d,\mathbf{z}\sim\mathcal{P}_\mathbf{z}}\|G(\mathbf{z})-\mathbf{x}\|_2 \le \xi$. Let the discriminator and generators be L-Lipschitz w.r.t. inputs and parameters, and let f be L_f-Lipschitz. Then for any $\epsilon > 0$, there exist $I = \frac{t+1}{\epsilon}\Delta_\gamma^{\text{worst}}$ generators $G_{\gamma_1^*}, ..., G_{\gamma_I^*}$ and a discriminator D_{θ^*} such that for some value $V \in \mathbb{R}$,*

$$\begin{aligned} &\forall \gamma_1, ..., \gamma_I \in \mathbb{R}^g, \quad && \Phi(\gamma_1, ..., \gamma_I; \theta^*) \le V + \epsilon, \\ &\forall \theta \in \mathbb{R}^t, && \Phi(\gamma_1^*, ..., \gamma_I^*; \theta) \ge V - \epsilon. \end{aligned}$$

Proof. We first show that the equilibrium value V is $2f(1/2)$. For the discriminator D_θ which only outputs $1/2$, it has payoff $2f(1/2)$ for all possible implementations of generators $G_{\gamma_1}, ..., G_{\gamma_I}$. Therefore, we have $V \ge 2f(1/2)$. We now show that $V \le 2f(1/2)$. We note that by assumption, for any $\xi > 0$, there exists a closed neighbour of implementation of generator G_ξ such that $\mathbb{E}_{\mathbf{x}\sim\mathcal{P}_d,\mathbf{z}\sim\mathcal{P}_z}\|G'_\xi(\mathbf{z})-\mathbf{x}\|_2 \le \xi$ for all G'_ξ in the neighbour. Such a neighbour exists because the generator is Lipschitz w.r.t. its parameters. Let the parameter implementation of such neighbour of G_ξ be Γ. The Wasserstein distance between G_ξ and $\mathcal{P}_d$ is ξ. Since the function f and the discriminator are L_f-Lipschitz and L-Lipschitz w.r.t. their inputs, respectively, we have

$$\left|\mathbb{E}_{\mathbf{z}\sim G_\xi} f(1 - D_\theta(\mathbf{z})) - \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_d} f(1 - D_\theta(\mathbf{x}))\right| \le \mathcal{O}(L_f L \xi).$$

Thus, for any fixed γ, we have

$$\begin{aligned} &\sup_{\theta\in\mathbb{R}^t} \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_d} f(D_\theta(\mathbf{x})) + \mathbb{E}_{\mathbf{z}\sim G_\xi} f(1 - D_\theta(\mathbf{z})) \\ \le\ &\mathcal{O}(L_f L\xi) + \sup_{\theta\in\mathbb{R}^t} \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_d} f(D_\theta(\mathbf{x})) + \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_d} f(1 - D_\theta(\mathbf{x})) \\ \le\ &\mathcal{O}(L_f L\xi) + 2f(1/2) \to 2f(1/2), \quad (\xi \to +0) \end{aligned}$$

which implies that $\frac{1}{I}\sup_{\theta\in\mathbb{R}^t}\Phi(\gamma_1, ..., \gamma_I; \theta) \le 2f(1/2)$ for all $\gamma_1, ..., \gamma_I \in \Gamma$. So we have $V = 2f(1/2)$. This means that the discriminator cannot do much better than a random guess.

The above analysis implies that the equilibrium is achieved when D_{θ^*} only outputs $1/2$. Denote by Θ the small closed neighbour of this θ^* such that $\Phi(\gamma_1, ..., \gamma_I; \theta)$ is concave w.r.t. $\theta \in \Theta$ for any fixed $\gamma_1, ..., \gamma_I \in \Gamma$. We thus focus on the loss in the range of $\Theta \subseteq \mathbb{R}^t$ and $\Gamma \subseteq \mathbb{R}^g$:

$$\Phi(\gamma_1, ..., \gamma_I; \theta) := \sum_{i=1}^{I} \left[\mathbb{E}_{\mathbf{x}\sim\mathcal{P}_d} f(D_\theta(\mathbf{x})) + \mathbb{E}_{\mathbf{z}\sim\mathcal{P}_z} f(1 - D_\theta(G_{\gamma_i}(\mathbf{z})))\right], \quad \theta \in \Theta,\ \gamma_1, ..., \gamma_I \in \Gamma.$$

Since $\Phi(\gamma_1, ..., \gamma_I; \theta)$ is concave w.r.t. $\theta \in \Theta$ for all $\gamma_1, ..., \gamma_I \in \Gamma$, by Corollary 1, we have

$$\inf_{\gamma_1,...,\gamma_I\in\Gamma} \sup_{\theta\in\Theta} \frac{1}{I}\Phi(\gamma_1, ..., \gamma_I; \theta) - \sup_{\theta\in\Theta} \inf_{\gamma_1,...,\gamma_I\in\Gamma} \frac{1}{I}\Phi(\gamma_1, ..., \gamma_I; \theta) \le \epsilon.$$

The optimal implementations of $\gamma_1, ..., \gamma_I$ is achieved by $\operatorname{argmin}_{\gamma_1,...,\gamma_I\in\Gamma}\sup_{\theta\in\Theta}\frac{1}{I}\Phi(\gamma_1, ..., \gamma_I; \theta)$. □

4.3 Robustness of Deep Classification Networks

4.3.1 Introduction

In response to the vulnerability of deep neural networks to small perturbations around input data [221], adversarial defenses have been an imperative object of study in machine learning [119, 259], computer vision [169, 213, 244], natural language processing [125], and many other domains. In machine learning, study of adversarial defenses has led to significant advances in understanding and defending against adversarial threat [116]. In computer vision and natural language processing, adversarial defenses serve as indispensable building blocks for a range of security-critical systems and applications, such as autonomous cars and speech recognition authorization. The problem of adversarial defenses can be stated as that of learning a classifier with high test accuracy on both natural and *adversarial examples*. The adversarial example for a given labeled data $(\mathbf{x}, y)$ is a data point $\mathbf{x}'$ that causes a classifier c to output a different label on $\mathbf{x}'$ than y, but is "imperceptibly similar" to $\mathbf{x}$. Given the difficulty of providing an operational definition of "imperceptible similarity," adversarial examples typically come in the form of *restricted attacks* such as ϵ-bounded perturbations [221], or *unrestricted attacks* such as adversarial rotations, translations, and deformations [3, 48, 80, 93, 243, 260]. The focus of this work is the former setting.

Despite a large literature devoted to improving the robustness of deep-learning models, many fundamental questions remain unresolved. One of the most important questions is how to trade off adversarial robustness against natural accuracy. Statistically, robustness can be be at odds with accuracy when no assumptions are made on the data distribution [224]. This has led to an empirical line of work on adversarial defense that incorporates various kinds of assumptions [141, 217]. On the theoretical front, methods such as *relaxation based defenses* [137, 190] provide provable guarantees for adversarial robustness. They, however, ignore the performance of classifier on the non-adversarial examples, and thus leave open the theoretical treatment of the putative robustness/accuracy trade-off.

The problem of adversarial defense becomes more challenging when considering computational issues. This is due to the fact that direct formulations of robust-classification problems involves minimizing the robust 0-1 loss

$$\max_{\mathbf{x}':\|\mathbf{x}'-\mathbf{x}\|\leq\epsilon} \mathbf{1}\{c(\mathbf{x}') \neq y\}, \tag{4.31}$$

a loss which is NP-hard to optimize [103]. This is why progress on algorithms that focus on accuracy have built on *minimum contrast methods* that minimize a surrogate of the 0–1 loss function [32], e.g., the hinge loss or cross-entropy loss. While prior work on adversarial defense replaced the 0-1 loss $\mathbf{1}(\cdot)$ in Eqn. (4.31) with a surrogate loss to defend against adversarial threat [141, 164, 228], this line of research may suffer from loose surrogate approximation to the 0-1 loss. It may thus result in degraded performance.

We begin with an illustrative example that illustrates the trade-off between accuracy and adversarial robustness, a phenomenon which has been demonstrated by [224], but without theoretical guarantees. We demonstrate that the minimal risk is achieved by a classifier with 100% accuracy on the non-adversarial examples. We refer to this accuracy as the *natural accuracy* and

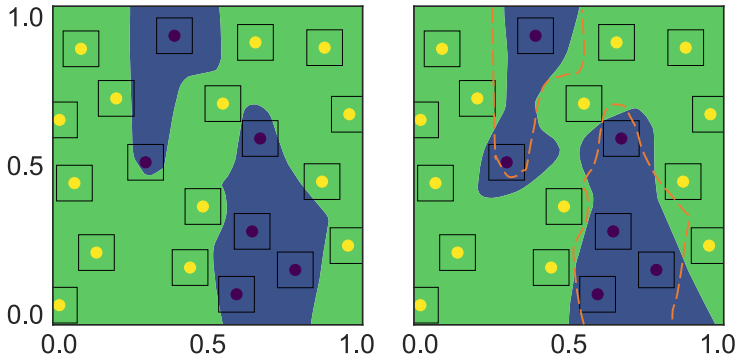


Figure 4.10: **Left figure:** decision boundary learned by natural training method. **Right figure:** decision boundary learned by our adversarial training method, where the orange dotted line represents the decision boundary in the left figure. It shows that both methods achieve zero natural training error, while our adversarial training method achieves better robust training error than the natural training method.

we similarly refer to the *natural error* or *natural risk*. In this same example, the accuracy to the adversarial examples, which we refer to as the *robust accuracy*, is as small as 0% (see Table 4.3). This motivates us to quantify the trade-off by the gap between optimal natural error and the robust error. Note that the latter is an adversarial counterpart of the former which allows a bounded worst-case perturbation before feeding the perturbed sample to the classifier.

We study this gap in the context of a differentiable surrogate loss. We show that surrogate loss minimization suffices to derive a classifier with guaranteed robustness and accuracy. Our theoretical analysis naturally leads to a new formulation of adversarial defense which has several appealing properties; in particular, it inherits the benefits of scalability to large datasets exhibited by Tiny ImageNet, and the algorithm achieves state-of-the-art performance on a range of benchmarks while providing theoretical guarantees. For example, while the defenses overviewed in [13] achieve robust accuracy no higher than ~47% under white-box attacks, our method achieves robust accuracy as high as ~57% in the same setting. The methodology is the foundation of our entry to the NeurIPS 2018 Adversarial Vision Challenge where we won first place out of 1,995 submissions, surpassing the runner-up approach by 11.41% in terms of mean ℓ_2 perturbation distance.

Summary of contributions. Our work tackles the problem of trading accuracy off against robustness and advances the state-of-the-art in multiple ways.

- Theoretically, we characterize the trade-off between accuracy and robustness for classification problems via the gap between robust error and optimal natural error. We provide an upper bound for this gap in terms of surrogate loss. The bound is *optimal* as it matches the lower bound in the worst-case scenario.
- Algorithmically, inspired by our theoretical analysis, we propose a new formulation of adversarial defense, TRADES, as optimizing a regularized surrogate loss. The loss consists of two terms: the term of empirical risk minimization encourages the algorithm to maximize the natural accuracy, while the regularization term encourages the algorithm to push the

decision boundary away from the data, so as to improve adversarial robustness (see Figure 4.10).

- Experimentally, we show that our proposed algorithm outperforms state-of-the-art methods under both black-box and white-box threat models. In particular, the methodology won the final round of the NeurIPS 2018 Adversarial Vision Challenge.

4.3.2 Preliminaries

Before proceeding, we define some notation and clarify our problem setup.

Notations. We will use *bold capital* letters such as $\mathbf{X}$ and $\mathbf{Y}$ to represent random vector, *bold lower-case* letters such as $\mathbf{x}$ and $\mathbf{y}$ to represent realization of random vector, *capital* letters such as X and Y to represent random variable, and *lower-case* letters such as x and y to represent realization of random variable. Specifically, we denote by $\mathbf{x} \in \mathcal{X}$ the sample instance, and by $y \in \{-1, +1\}$ the label, where $\mathcal{X} \subseteq \mathbb{R}^d$ indicates the instance space. $\text{sign}(x)$ represents the sign of scalar x with $\text{sign}(0) = +1$. Denote by $f : \mathcal{X} \to \mathbb{R}$ the *score function* which maps an instance to a confidence value associated with being positive. It can be parametrized, e.g., by deep neural networks. The associated binary classifier is $\text{sign}(f(\cdot))$. We will frequently use $\mathbf{1}\{\text{event}\}$, the 0-1 loss, to represent an indicator function that is 1 if an event happens and 0 otherwise. For norms, we denote by $\|\mathbf{x}\|$ a generic norm. Examples of norms include $\|\mathbf{x}\|_\infty$, the infinity norm of vector $\mathbf{x}$, and $\|\mathbf{x}\|_2$, the ℓ_2 norm of vector $\mathbf{x}$. We use $\mathbb{B}(\mathbf{x}, \epsilon)$ to represent a neighborhood of $\mathbf{x}$: $\{\mathbf{x}' \in \mathcal{X} : \|\mathbf{x}' - \mathbf{x}\| \leq \epsilon\}$. For a given score function f, we denote by $\text{DB}(f)$ the decision boundary of f; that is, the set $\{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) = 0\}$. $\mathbb{B}(\text{DB}(f), \epsilon)$ indicates the neighborhood of the decision boundary of f: $\{\mathbf{x} \in \mathcal{X} : \exists \mathbf{x}' \in \mathbb{B}(\mathbf{x}, \epsilon) \text{ s.t. } f(\mathbf{x})f(\mathbf{x}') \leq 0\}$. For a given function $\psi(\mathbf{u})$, we denote by $\psi^*(\mathbf{v}) := \sup_{\mathbf{u}}\{\mathbf{u}^T\mathbf{v} - \psi(\mathbf{u})\}$ the conjugate function of ψ, by ψ^{**} the bi-conjugate, and by ψ^{-1} the inverse function. We will frequently use $\phi(\cdot)$ to indicate the surrogate of 0-1 loss.

Robust (classification) error

In the setting of adversarial learning, we are given a set of instances $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathcal{X}$ and labels $y_1, ..., y_n \in \{-1, +1\}$. We assume that the data are sampled from an unknown distribution $(\mathbf{X}, Y) \sim \mathcal{D}$. To characterize the robustness of a score function $f : \mathcal{X} \to \mathbb{R}$, [52, 70, 205] defined *robust (classification) error* under the threat model of bounded ϵ distortion:

$$\mathcal{R}_{\text{adv}}(f) := \mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{D}}\mathbf{1}\{\exists \mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon) \text{ s.t. } f(\mathbf{X}')Y \leq 0\}.$$

This is in sharp contrast to the standard measure of classifier performance—the *natural (classification) error* $\mathcal{R}_{\text{nat}}(f) := \mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{D}}\mathbf{1}\{f(\mathbf{X})Y \leq 0\}$. We note that the two errors satisfy $\mathcal{R}_{\text{adv}}(f) \geq \mathcal{R}_{\text{nat}}(f)$ for all f; the robust error is equal to the natural error when $\epsilon = 0$.

Trade-off between natural and robust errors

Our study is motivated by the trade-off between natural and robust errors. [224] showed that training robust models may lead to a reduction of standard accuracy. To illustrate the phenomenon, we provide a toy example here.

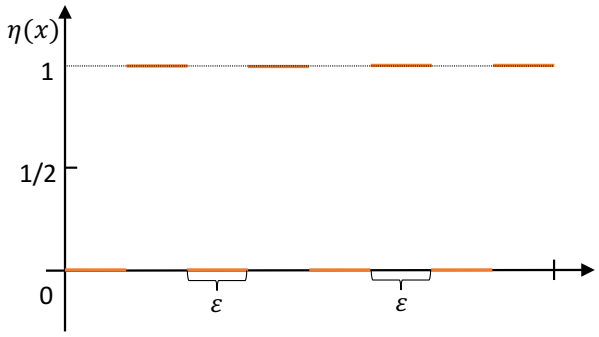


Figure 4.11: Counterexample given by Eqn. (4.32).

Table 4.3: Comparisons of natural and robust errors of Bayes optimal classifier and all-one classifier in example (4.32). The Bayes optimal classifier has the optimal natural error while the all-one classifier has the optimal robust error.

	Bayes Optimal Classifier	All-One Classifier
$\mathcal{R}_{\text{nat}}$	0 (optimal)	1/2
$\mathcal{R}_{\text{adv}}$	1	1/2 (optimal)

Example. Consider the case $(X, Y) \sim \mathcal{D}$, where the marginal distribution over the instance space is a uniform distribution over $[0, 1]$, and for $k = 0, 1, ..., \lceil \frac{1}{2\epsilon} - 1 \rceil$,

$$\begin{aligned} \eta(x) &:= \Pr(Y = 1 | X = x) \\ &= \begin{cases} 0, & x \in [2k\epsilon, (2k+1)\epsilon), \\ 1, & x \in ((2k+1)\epsilon, (2k+2)\epsilon]. \end{cases} \end{aligned} \tag{4.32}$$

See Figure 4.11 for the visualization of $\eta(x)$. We consider two classifiers: a) the Bayes optimal classifier $\text{sign}(2\eta(x) - 1)$; b) the all-one classifier which always outputs "positive." Table 4.3 displays the trade-off between natural and robust errors: the minimal natural error is achieved by the Bayes optimal classifier with large robust error, while the optimal robust error is achieved by the all-one classifier with large natural error. Despite a large literature on the analysis of robust error in terms of generalization [70, 205, 248] and computational complexity [51, 52], the trade-off between the natural error and the robust error has not been a focus of theoretical study.

Our goal. To characterize the trade-off, we aim at approximately solving a constrained problem for a score function $\widehat{f}$ with guarantee $\mathcal{R}_{\text{adv}}(\widehat{f}) \leq \mathsf{OPT} + \delta$, given a precision parameter $\delta > 0$:

$$\mathsf{OPT} := \min_f \mathcal{R}_{\text{adv}}(f), \quad \text{s.t.} \quad \mathcal{R}_{\text{nat}}(f) \leq \mathcal{R}^*_{\text{nat}} + \delta,$$

where $\mathcal{R}^*_{\text{nat}}$ represents the risk of the Bayes optimal classifier, the classifier with the minimal natural error. We note that it suffices to show $\mathcal{R}_{\text{adv}}(f) - \mathcal{R}^*_{\text{nat}} \leq \delta$. This is because a) $\mathcal{R}_{\text{nat}}(f) - \mathcal{R}^*_{\text{nat}} \leq \mathcal{R}_{\text{adv}}(f) - \mathcal{R}^*_{\text{nat}} \leq \delta$, and b) $\mathcal{R}_{\text{adv}}(f) \leq \mathcal{R}^*_{\text{nat}} + \delta \leq \mathsf{OPT} + \delta$, where the last inequality holds since $\mathcal{R}_{\text{nat}}(f) \leq \mathcal{R}_{\text{adv}}(f)$ for all f's and therefore $\min_f \mathcal{R}_{\text{nat}}(f) \leq \min_f \mathcal{R}_{\text{adv}}(f) \leq \mathsf{OPT}$. In this section, our principal goal is to provide a *tight* bound on $\mathcal{R}_{\text{adv}}(f) - \mathcal{R}^*_{\text{nat}}$, using a regularized surrogate loss which can be optimized easily.

Table 4.4: Examples of classification-calibrated loss ϕ and associated ψ-transform. Here $\psi_{\mathsf{log}}(\theta) = \frac{1}{2}(1-\theta)\log_2(1-\theta) + \frac{1}{2}(1+\theta)\log_2(1+\theta)$.

Loss	$\phi(\alpha)$	$\psi(\theta)$
Hinge	$\max\{1-\alpha, 0\}$	θ
Sigmoid	$1-\tanh(\alpha)$	θ
Exponential	$\exp(-\alpha)$	$1-\sqrt{1-\theta^2}$
Logistic	$\log_2(1+\exp(-\alpha))$	$\psi_{\mathsf{log}}(\theta)$

Classification-calibrated surrogate loss

Definition. Minimization of the 0-1 loss in the natural and robust errors is computationally intractable and the demands of computational efficiency have led researchers to focus on minimization of a tractable *surrogate loss*, $\mathcal{R}_\phi(f) := \mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{D}}\phi(f(\mathbf{X})Y)$. We then need to find quantitative relationships between the excess errors associated with ϕ and those associated with 0–1 loss. We make a weak assumption on ϕ: it is *classification-calibrated* [32]. Formally, for $\eta \in [0,1]$, define the *conditional ϕ-risk* by

$$H(\eta) := \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha) := \inf_{\alpha\in\mathbb{R}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right),$$

and define $H^-(\eta) := \inf_{\alpha(2\eta-1)\le 0} C_\eta(\alpha)$. The classification-calibrated condition requires that imposing the constraint that α has an inconsistent sign with the Bayes decision rule $\mathrm{sign}(2\eta-1)$ leads to a strictly larger ϕ-risk:

Assumption 2 (Classification-Calibrated Loss)**.** *We assume that the surrogate loss ϕ is classification-calibrated, meaning that for any $\eta \neq 1/2$, $H^-(\eta) > H(\eta)$.*

We argue that Assumption 2 is indispensable for classification problems, since without it the Bayes optimal classifier cannot be the minimizer of the ϕ-risk. Examples of classification-calibrated loss include hinge loss, sigmoid loss, exponential loss, logistic loss, and many others (see Table 4.4).

Properties. Classification-calibrated loss has many structural properties that one can exploit. We begin by introducing a functional transform of classification-calibrated loss ϕ which was proposed by [32]. Define the function $\psi : [0,1] \to [0,\infty)$ by $\psi = \widetilde{\psi}^{**}$, where $\widetilde{\psi}(\theta) := H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$. Indeed, the function $\psi(\theta)$ is the largest convex lower bound on $H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$. The value $H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$ characterizes how close the surrogate loss ϕ is to the class of non-classification-calibrated losses.

Below we state useful properties of the ψ-transform. We will frequently use the function ψ to bound $\mathcal{R}_{\mathrm{adv}}(f) - \mathcal{R}^*_{\mathrm{nat}}$.

Lemma 74 ([32])**.** *Under Assumption 2, the function ψ has the following properties: ψ is non-decreasing, continuous, convex on $[0,1]$ and $\psi(0) = 0$.*

4.3.3 Our results on robustness

In this section, we present our main theoretical contributions for binary classification and compare our results with prior literature. Binary classification problems have received significant attention

in recent years as many competitions evaluate the performance of robust models on binary classification problems [48]. We defer the discussions for multi-class problems to Section 4.3.4.

Upper bound

Our analysis leads to the following guarantee on the performance of surrogate loss minimization.

Theorem 51. *Under Assumption 2, for any non-negative loss function ϕ such that $\phi(0) \geq 1$, any measurable $f : \mathcal{X} \to \mathbb{R}$, any probability distribution on $\mathcal{X} \times \{\pm 1\}$, and any $\lambda > 0$, we have*[7]

$$\begin{aligned}
&\mathcal{R}_{\mathrm{adv}}(f) - \mathcal{R}^*_{\mathrm{nat}} \\
&\leq \psi^{-1}(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi) + \Pr[\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), c_0(\mathbf{X}) = Y] \\
&\leq \psi^{-1}(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi) + \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')f(\mathbf{X})/\lambda),
\end{aligned}$$

where $\mathcal{R}_\phi(f) := \mathbb{E}\phi(f(\mathbf{X})Y)$, $\mathcal{R}^*_\phi := \min_f \mathcal{R}_\phi(f)$ and $c_0(\cdot) := \mathrm{sign}(2\eta(\cdot) - 1)$ is the Bayes optimal classifier.

Quantity governing model robustness. Our result provides a formal justification for the existence of adversarial examples: learning models are brittle to small adversarial attacks because the probability that data lie around the decision boundary of the model, $\Pr[\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), c_0(\mathbf{X}) = Y]$, is large. As a result, small perturbations may move the data point to the wrong side of the decision boundary, leading to weak robustness of classification models.

Lower bound

We now establish a lower bound on $\mathcal{R}_{\mathrm{adv}}(f) - \mathcal{R}^*_{\mathrm{nat}}$. Our lower bound matches our analysis of the upper bound in Section 4.3.3 up to an arbitrarily small constant.

Theorem 52. *Suppose that $|\mathcal{X}| \geq 2$. Under Assumption 2, for any non-negative loss function ϕ such that $\phi(x) \to 0$ as $x \to +\infty$, any $\xi > 0$, and any $\theta \in [0, 1]$, there exists a probability distribution on $\mathcal{X} \times \{\pm 1\}$, a function $f : \mathbb{R}^d \to \mathbb{R}$, and a regularization parameter $\lambda > 0$ such that $\mathcal{R}_{\mathrm{adv}}(f) - \mathcal{R}^*_{\mathrm{nat}} = \theta$ and*

$$\begin{aligned}
\psi\left(\theta - \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')f(\mathbf{X})/\lambda)\right) &\leq \mathcal{R}_\phi(f) - \mathcal{R}^*_\phi \\
&\leq \psi\left(\theta - \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')f(\mathbf{X})/\lambda)\right) + \xi.
\end{aligned}$$

Theorem 52 demonstrates that in the presence of extra conditions on the loss function, i.e., $\lim_{x \to +\infty} \phi(x) = 0$, the upper bound in Section 4.3.3 is tight. The condition holds for all the losses in Table 4.4.

[7] We study the population form of the loss function, although we believe that our analysis can be extended to the empirical form by the uniform convergence argument. We leave this analysis as an interesting problem for future research.

4.3.4 Our algorithms

Optimization. Theorems 51 and 52 shed light on algorithmic designs of adversarial defenses. In order to minimize $\mathcal{R}_{\text{adv}}(f) - \mathcal{R}^*_{\text{nat}}$, the theorems suggest minimizing[8]

$$\min_f \mathbb{E}\Big\{ \underbrace{\phi(f(\mathbf{X})Y)}_{\text{for accuracy}} + \underbrace{\max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X},\epsilon)} \phi(f(\mathbf{X})f(\mathbf{X}')/\lambda)}_{\text{regularization for robustness}} \Big\}. \tag{4.33}$$

We name our method **TRADES** (TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization).

Intuition behind the optimization. Problem (4.33) captures the trade-off between the natural and robust errors: the first term in (4.33) encourages the natural error to be optimized by minimizing the "difference" between $f(\mathbf{X})$ and Y, while the second regularization term encourages the output to be smooth, that is, it pushes the decision boundary of classifier away from the sample instances via minimizing the "difference" between the prediction of natural example $f(\mathbf{X})$ and that of adversarial example $f(\mathbf{X}')$. This is conceptually consistent with the argument that smoothness is an indispensable property of robust models [66]. The tuning parameter λ plays a critical role on balancing the importance of natural and robust errors. To see how the hyperparameter λ affects the solution in the example of Section 4.3.2, problem (4.33) tends to the Bayes optimal classifier when $\lambda \to +\infty$, and tends to the all-one classifier when $\lambda \to 0$.

Comparisons with prior works. We compare our approach with several related lines of research in the prior literature. One of the best known algorithms for adversarial defense is based on *robust optimization* [137, 164, 190, 191, 240]. Most results in this direction involve algorithms that approximately minimize

$$\min_f \mathbb{E}\left\{ \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X},\epsilon)} \phi(f(\mathbf{X}')Y) \right\}, \tag{4.34}$$

where the objective function in problem (4.34) serves as an upper bound of the robust error $\mathcal{R}_{\text{adv}}(f)$. In complex problem domains, however, this objective function might not be tight as an upper bound of robust error, and may not capture the trade-off between natural and robust errors.

A related line of research is adversarial training by regularization [141, 196, 267]. There are several key differences between the results in this section and those of [141, 196, 267]. Firstly, the optimization formulations are different. In the previous works, the regularization term either measures the "difference" between $f(\mathbf{X}')$ and Y [141], or its gradient [196]. In contrast, our regularization term measures the "difference" between $f(\mathbf{X})$ and $f(\mathbf{X}')$. While [267] generated the adversarial example $\mathbf{X}'$ by adding random Gaussian noise to $\mathbf{X}$, our method simulates the adversarial example by solving the inner maximization problem in Eqn. (4.33). Secondly, we note that the losses in [141, 196, 267] lack of theoretical guarantees. Our loss, with the presence of the second term in problem (4.33), makes our theoretical analysis significantly more subtle. Moreover, our algorithm takes the same computational resources as *adversasrial training at scale* [141],

[8] There is correspondence between the λ in problem (4.33) and the λ in the right hand side of Theorem 51, because ψ^{-1} is a non-decreasing function. Therefore, in practice we do not need to involve function ψ^{-1} in the optimization formulation.

which makes our method scalable to large-scale datasets. We defer the experimental comparisons of various regularization based methods to Table 4.7.

Differences with Adversarial Logit Pairing. We also compare TRADES with Adversarial Logit Pairing (ALP) [81, 130]. The algorithm of ALP works as follows: given a fixed network f in each round, the algorithm firstly generates an adversarial example $\mathbf{X}'$ by solving $\operatorname{argmax}_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')Y)$; ALP then updates the network parameter by solving a minimization problem

$$\min_f \mathbb{E}\left\{\alpha\phi(f(\mathbf{X}')Y) + (1-\alpha)\phi(f(\mathbf{X})Y) + \|f(\mathbf{X}) - f(\mathbf{X}')\|_2/\lambda\right\},$$

where $0 \leq \alpha \leq 1$ is a regularization parameter; the algorithm finally repeats the above-mentioned procedure until it converges. We note that there are fundamental differences between TRADES and ALP. While ALP simulates adversarial example $\mathbf{X}'$ by the FGSM^k attack, TRADES simulates $\mathbf{X}'$ by solving $\operatorname{argmax}_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X})f(\mathbf{X}')/\lambda)$. Moreover, while ALP uses the ℓ_2 loss between $f(\mathbf{X})$ and $f(\mathbf{X}')$ to regularize the training procedure without theoretical guarantees, TRADES uses the classification-calibrated loss according to Theorems 51 and 52.

Heuristic algorithm. In response to the optimization formulation (4.33), we use two heuristics to achieve more general defenses: a) extending to multi-class problems by involving multi-class calibrated loss; b) approximately solving the minimax problem via alternating gradient descent. For multi-class problems, a surrogate loss is *calibrated* if minimizers of the surrogate risk are also minimizers of the 0-1 risk [184]. Examples of multi-class calibrated loss include cross-entropy loss. Algorithmically, we extend problem (4.33) to the case of multi-class classifications by replacing ϕ with a multi-class calibrated loss $\mathcal{L}(\cdot, \cdot)$:

$$\min_f \mathbb{E}\left\{\mathcal{L}(f(\mathbf{X}), \mathbf{Y}) + \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \mathcal{L}(f(\mathbf{X}), f(\mathbf{X}'))/\lambda\right\}, \tag{4.35}$$

where $f(\mathbf{X})$ is the output vector of learning model (with softmax operator in the top layer for the cross-entropy loss $\mathcal{L}(\cdot, \cdot)$), $\mathbf{Y}$ is the label-indicator vector, and $\lambda > 0$ is the regularization parameter. The pseudocode of adversarial training procedure, which aims at minimizing the empirical form of problem (4.35), is displayed in Algorithm 18.

The key ingredient of the algorithm is to approximately solve the linearization of inner maximization in problem (4.35) by the *projected gradient descent* (see Step 7). We note that $\mathbf{x}_i$ is a global minimizer with zero gradient to the objective function $g(\mathbf{x}') := \mathcal{L}(f(\mathbf{x}_i), f(\mathbf{x}'))$ in the inner problem. Therefore, we initialize $\mathbf{x}_i'$ by adding a small, random perturbation around $\mathbf{x}_i$ in Step 5 to start the inner optimizer. More exhaustive approximations of the inner maximization problem in terms of either optimization formulations or solvers would lead to better defense performance.

4.3.5 Experimental results

In this section, we verify the effectiveness of TRADES by numerical experiments. We denote by $\mathcal{A}_{\text{adv}}(f) := 1 - \mathcal{R}_{\text{adv}}(f)$ the robust accuracy, and by $\mathcal{A}_{\text{nat}}(f) := 1 - \mathcal{R}_{\text{nat}}(f)$ the natural accuracy on test dataset. The pixels of input images are normalized to $[0, 1]$.

Algorithm 18 Adversarial training by TRADES

Input: Step sizes η_1 and η_2, batch size m, number of iterations K in inner optimization, network architecture parametrized by θ
Output: Robust network f_θ
Randomly initialize network f_θ, or initialize network with pre-trained configuration
Repeat until convergence
  Read mini-batch $B = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ from training set
  For $i = 1, ..., m$ (in parallel)
  $\mathbf{x}_i' \leftarrow \mathbf{x}_i + 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is the Gaussian distribution with zero mean and identity variance
    For $k = 1, ..., K$
      $\mathbf{x}_i' \leftarrow \Pi_{\mathbb{B}(\mathbf{x}_i, \epsilon)}(\eta_1 \text{sign}(\nabla_{\mathbf{x}_i'} \mathcal{L}(f_\theta(\mathbf{x}_i), f_\theta(\mathbf{x}_i'))) + \mathbf{x}_i')$, where Π is the projection operator
    End For
  End For
$\theta \leftarrow \theta - \eta_2 \sum_{i=1}^m \nabla_\theta [\mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i) + \mathcal{L}(f_\theta(\mathbf{x}_i), f_\theta(\mathbf{x}_i'))/\lambda]/m$

Optimality of Theorem 51

We verify the tightness of the established upper bound in Theorem 51 for binary classification problem on MNIST dataset. The negative examples are '1' and the positive examples are '3'. Here we use a Convolutional Neural Network (CNN) with two convolutional layers, followed by two fully-connected layers. The output size of the last layer is 1. To learn the robust classifier, we minimize the regularized surrogate loss in Eqn. (4.33), and use the hinge loss in Table 4.4 as the surrogate loss ϕ, where the associated ψ-transform is $\psi(\theta) = \theta$.

To verify the tightness of our upper bound, we calculate the left hand side in Theorem 51, i.e.,

$$\Delta_{\text{LHS}} = \mathcal{R}_{\text{adv}}(f) - \mathcal{R}_{\text{nat}}^*,$$

and the right hand side, i.e.,

$$\Delta_{\text{RHS}} = (\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*) + \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')f(\mathbf{X})/\lambda).$$

As we cannot have access to the unknown distribution $\mathcal{D}$, we approximate the above expectation terms by test dataset. We first use natural training method to train a classifier so as to approximately estimate $\mathcal{R}_{\text{nat}}^*$ and $\mathcal{R}_\phi^*$, where we find that the naturally trained classifier can achieve natural error $\mathcal{R}_{\text{nat}}^* = 0\%$, and loss value $\mathcal{R}_\phi^* = 0.0$ for the binary classification problem. Next, we optimize problem (4.33) to train a robust classifier f. We take perturbation $\epsilon = 0.1$, number of iterations $K = 20$ and run 30 epochs on the training dataset. Finally, to approximate the second term in Δ_{RHS}, we use FGSM^k (white-box) attack (a.k.a. PGD attack) [141] with 20 iterations to approximately calculate the worst-case perturbed data $\boldsymbol{X}'$.

The results in Table 4.5 show the tightness of our upper bound in Theorem 51. It shows that the differences between Δ_{RHS} and Δ_{LHS} under various λ's are very small.

Table 4.5: Theoretical verification on the optimality of Theorem 51.

λ	$\mathcal{A}_{\text{adv}}(f)$ (%)	$\mathcal{R}_{\phi}(f)$	$\Delta = \Delta_{\text{RHS}} - \Delta_{\text{LHS}}$
2.0	99.43	0.0006728	0.006708
3.0	99.41	0.0004067	0.005914
4.0	99.37	0.0003746	0.006757
5.0	99.34	0.0003430	0.005860

Table 4.6: Sensitivity of regularization hyperparameter λ on MNIST and CIFAR10 datasets.

	MNIST		CIFAR10	
$1/\lambda$	$\mathcal{A}_{\text{adv}}(f)$ (%)	$\mathcal{A}_{\text{nat}}(f)$ (%)	$\mathcal{A}_{\text{adv}}(f)$ (%)	$\mathcal{A}_{\text{nat}}(f)$ (%)
0.1	91.09 ± 0.0385	99.41 ± 0.0235	26.53 ± 1.1698	91.31 ± 0.0579
0.2	92.18 ± 0.0450	99.38 ± 0.0094	37.71 ± 0.6743	89.56 ± 0.2154
0.4	93.21 ± 0.0660	99.35 ± 0.0082	41.50 ± 0.3376	87.91 ± 0.2944
0.6	93.87 ± 0.0464	99.33 ± 0.0141	43.37 ± 0.2706	87.50 ± 0.1621
0.8	94.32 ± 0.0492	99.31 ± 0.0205	44.17 ± 0.2834	87.11 ± 0.2123
1.0	94.75 ± 0.0712	99.28 ± 0.0125	44.68 ± 0.3088	87.01 ± 0.2819
2.0	95.45 ± 0.0883	99.29 ± 0.0262	48.22 ± 0.0740	85.22 ± 0.0543
3.0	95.57 ± 0.0262	99.24 ± 0.0216	49.67 ± 0.3179	83.82 ± 0.4050
4.0	95.65 ± 0.0340	99.16 ± 0.0205	50.25 ± 0.1883	82.90 ± 0.2217
5.0	95.65 ± 0.1851	99.16 ± 0.0403	50.64 ± 0.3336	81.72 ± 0.0286

Sensitivity of regularization hyperparameter λ

The regularization parameter λ is an important hyperparameter in our proposed method. We show how the regularization parameter affects the performance of our robust classifiers by numerical experiments on two datasets, MNIST and CIFAR10. For both datasets, we minimize the loss in Eqn. (4.35) to learn robust classifiers for multi-class problems, where we choose $\mathcal{L}$ as the cross-entropy loss.

MNIST setup. We use the CNN which has two convolutional layers, followed by two fully-connected layers. The output size of the last layer is 10. We set perturbation $\epsilon = 0.1$, perturbation step size $\eta_1 = 0.01$, number of iterations $K = 20$, learning rate $\eta_2 = 0.01$, batch size $m = 128$, and run 50 epochs on the training dataset. To evaluate the robust error, we apply FGSM^k (white-box) attack with 40 iterations and 0.005 step size. The results are in Table 4.6.

CIFAR10 setup. We apply ResNet-18 [115] for classification. The output size of the last layer is 10. We set perturbation $\epsilon = 0.031$, perturbation step size $\eta_1 = 0.007$, number of iterations $K = 10$, learning rate $\eta_2 = 0.1$, batch size $m = 128$, and run 100 epochs on the training dataset. To evaluate the robust error, we apply FGSM^k (white-box) attack with 20 iterations and the step size is 0.003. The results are in Table 4.6.

We observe that as the regularization parameter $1/\lambda$ increases, the natural accuracy $\mathcal{A}_{\text{nat}}(f)$ decreases while the robust accuracy $\mathcal{A}_{\text{adv}}(f)$ increases, which verifies our theory on the trade-off between robustness and accuracy. Note that for MNIST dataset, the natural accuracy does not decrease too much as the regularization term $1/\lambda$ increases, which is different from the results

of CIFAR10. This is probably because the classification task for MNIST is easier. Meanwhile, our proposed method is not very sensitive to the choice of λ. Empirically, when we set the hyperparameter $1/\lambda$ in $[1, 10]$, our method is able to learn classifiers with both high robustness and high accuracy.

Adversarial defenses under various attacks

Previously, [13] showed that 7 defenses in ICLR 2018 which relied on obfuscated gradients may easily break down. In this section, we verify the effectiveness of our method with the same experimental setup under both white-box and black-box threat models.

MNIST setup. We use the CNN architecture in [58] with four convolutional layers, followed by three fully-connected layers. We set perturbation $\epsilon = 0.3$, perturbation step size $\eta_1 = 0.01$, number of iterations $K = 40$, learning rate $\eta_2 = 0.01$, batch size $m = 128$, and run 100 epochs on the training dataset.

CIFAR10 setup. We use the same neural network architecture as [164], i.e., the wide residual network WRN-34-10 [249]. We set perturbation $\epsilon = 0.031$, perturbation step size $\eta_1 = 0.007$, number of iterations $K = 10$, learning rate $\eta_2 = 0.1$, batch size $m = 128$, and run 100 epochs on the training dataset.

White-box attacks We summarize our results in Table 4.7 together with the results from [13]. We also implement methods in [141, 196, 267] on the CIFAR10 dataset as they are also regularization based methods. For MNIST dataset, we apply FGSM^k (white-box) attack with 40 iterations and the step size is 0.01. For CIFAR10 dataset, we apply FGSM^k (white-box) attack with 20 iterations and the step size is 0.003, under which the defense model in [164] achieves 47.04% robust accuracy. Table 4.7 shows that our proposed defense method can significantly improve the robust accuracy of models, which is able to achieve robust accuracy as high as 56.61%. We also evaluate our robust model on MNIST dataset under the same threat model as in [203] (C&W white-box attack [58]), and the robust accuracy is 99.46%. See appendix for detailed information of models in Table 4.7.

In addition, we also evaluate our models by using FGSM^k with more perturbation steps. For MNIST dataset, we use FGSM^k (white-box) attack with $1,000$ iterations and the step size is 6×10^{-5}. For CIFAR10 dataset, we use FGSM^k (white-box) attack with $1,000$ iterations and the step size is 6×10^{-4}. The results are in Table 4.7. It shows that the performance of TRADES under various FGSM^k attacks of varying k's does not differ too much.

Black-box attacks We verify the robustness of our models under black-box attacks. We first train models without using adversarial training on the MNIST and CIFAR10 datasets. We use the same network architectures that are specified in the beginning of this section, i.e., the CNN architecture in [58] and the WRN-34-10 architecture in [249]. We denote these models by naturally trained models (*Natural*). The accuracy of the naturally trained CNN model is 99.50% on the MNIST dataset. The accuracy of the naturally trained WRN-34-10 model is 95.29% on the CIFAR10 dataset. We also implement the method proposed in [164] on both datasets. We denote these models by Madry's models (*Madry*). The accuracy of [164]'s CNN model is 99.36% on the MNIST dataset. The accuracy of [164]'s WRN-34-10 model is 85.49% on the CIFAR10 dataset.

Table 4.7: Comparisons of TRADES with prior defense models under white-box attacks.

Defense	Under which attack	Dataset	Distance	$\mathcal{A}_{\text{nat}}(f)$	$\mathcal{A}_{\text{adv}}(f)$
[53]	[13]	CIFAR10	0.031 (ℓ_∞)	-	0%
[163]	[13]	CIFAR10	0.031 (ℓ_∞)	-	5%
[77]	[13]	CIFAR10	0.031 (ℓ_∞)	-	0%
[213]	[13]	CIFAR10	0.031 (ℓ_∞)	-	9%
[175]	[13]	CIFAR10	0.015 (ℓ_∞)	-	15%
[240]	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	27.07%	23.54%
[164]	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	87.30%	**47.04%**
[267]	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	94.64%	0.15%
[141]	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	85.25%	45.89%
[196]	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	95.34%	0%
TRADES ($1/\lambda = 1$)	FGSM^{1,000} (PGD)	CIFAR10	0.031 (ℓ_∞)	88.64%	48.90%
TRADES ($1/\lambda = 6$)	FGSM^{1,000} (PGD)	CIFAR10	0.031 (ℓ_∞)	84.92%	**56.43%**
TRADES ($1/\lambda = 1$)	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	88.64%	49.14%
TRADES ($1/\lambda = 6$)	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	84.92%	**56.61%**
TRADES ($1/\lambda = 1$)	DeepFool (ℓ_∞)	CIFAR10	0.031 (ℓ_∞)	88.64%	59.10%
TRADES ($1/\lambda = 6$)	DeepFool (ℓ_∞)	CIFAR10	0.031 (ℓ_∞)	84.92%	61.38%
TRADES ($1/\lambda = 1$)	LBFGSAttack	CIFAR10	0.031 (ℓ_∞)	88.64%	84.41%
TRADES ($1/\lambda = 6$)	LBFGSAttack	CIFAR10	0.031 (ℓ_∞)	84.92%	81.58%
TRADES ($1/\lambda = 1$)	MI-FGSM	CIFAR10	0.031 (ℓ_∞)	88.64%	51.26%
TRADES ($1/\lambda = 6$)	MI-FGSM	CIFAR10	0.031 (ℓ_∞)	84.92%	57.95%
TRADES ($1/\lambda = 1$)	C&W	CIFAR10	0.031 (ℓ_∞)	88.64%	84.03%
TRADES ($1/\lambda = 6$)	C&W	CIFAR10	0.031 (ℓ_∞)	84.92%	81.24%
[203]	[13]	MNIST	0.005 (ℓ_2)	-	55%
[164]	FGSM^{40} (PGD)	MNIST	0.3 (ℓ_∞)	99.36%	96.01%
TRADES ($1/\lambda = 6$)	FGSM^{1,000} (PGD)	MNIST	0.3 (ℓ_∞)	99.48%	95.60%
TRADES ($1/\lambda = 6$)	FGSM^{40} (PGD)	MNIST	0.3 (ℓ_∞)	99.48%	96.07%
TRADES ($1/\lambda = 6$)	C&W	MNIST	0.005 (ℓ_2)	99.48%	99.46%

Table 4.8: Comparisons of TRADES with prior defenses under black-box FGSM^{40} attack on the MNIST dataset. The models inside parentheses are source models which provide gradients to adversarial attackers. The defense model 'Madry' is the same model as in the antepenultimate line of Table 4.7. The defense model 'TRADES' is the same model as in the penultimate line of Table 4.7.

Defense Model	Robust Accuracy $\mathcal{A}_{\text{adv}}(f)$	
Madry	97.43% (Natural)	97.38% (Ours)
TRADES	**97.63%** (Natural)	**97.66%** (Madry)

Table 4.9: Comparisons of TRADES with prior defenses under black-box FGSM^{20} attack on the CIFAR10 dataset. The models inside parentheses are source models which provide gradients to adversarial attackers. The defense model 'Madry' is implemented based on [164] and defined in Section 4.3.5, and the defense model 'TRADES' is the same model as in the 11th line of Table 4.7.

Defense Model	Robust Accuracy $\mathcal{A}_{\text{adv}}(f)$	
Madry	84.39% (Natural)	66.00% (Ours)
TRADES	**87.60%** (Natural)	**70.14%** (Madry)

For both datasets, we use FGSM^k (black-box) method to attack various defense models. For MNIST dataset, we set perturbation $\epsilon = 0.3$ and apply FGSM^k (black-box) attack with 40 iterations and the step size is 0.01. For CIFAR10 dataset, we set $\epsilon = 0.031$ and apply FGSM^k (black-box) attack with 20 iterations and the step size is 0.003. Note that the setup is the same as the setup specified in Section 4.3.5. We summarize our results in Table 4.8 and Table 4.9. In both tables, we use two source models (noted in the parentheses) to generate adversarial perturbations: we compute the perturbation directions according to the gradients of the source models on the input images. It shows that our models are more robust against black-box attacks transfered from naturally trained models and [164]'s models. Moreover, our models can generate stronger adversarial examples for black-box attacks compared with naturally trained models and [164]'s models.

Interpretability

We show that models trained by TRADES have strong interpretability.

Adversarial examples on MNIST and CIFAR10 datasets We show adversarial examples on MNIST and CIFAR10. We apply **`foolbox`**[9] [192] to generate adversarial examples, which is able to return the smallest adversarial perturbations under the ℓ_∞-norm distance. The adversarial examples are generated by using FGSM^k (white-box) attack on the models described in Section 4.3.5, including *Natural* models, *Madry*'s models and *TRADES* models. Note that the FGSM^k attack is `foolbox.attacks.LinfinityBasicIterativeAttack` in **`foolbox`**. See Figure 4.12 and Figure 4.13 for the adversarial examples of different models on the MNIST and CIFAR10 datasets.

Adversarial examples on Bird-or-Bicycle dataset We find that the robust models trained by TRADES have strong interpretability. To see this, we apply a (spatial-tranformation-invariant) variant of TRADES to train ResNet-50 models in response to the unrestricted adversarial examples in the Bird-or-Bicycle competition [48]. The dataset in the competition is *Bird-or-Bicycle*, which consists of 30,000 pixel-224×224 images with label either 'bird' or 'bicycle'. The unrestricted threat models include structural perturbations, rotations, translations, resizing, 17+ common corruptions, etc. Please refer to [48] for more detailed setup of the competition.

[9]Link: `https://foolbox.readthedocs.io/en/latest/index.html`

We show in Figures 4.14 and 4.15 the adversarial examples by the boundary attack with random spatial transformation on our robust model trained by the variant of TRADES. The boundary attack [46] is a black-box attack method which searches for data points near the decision boundary and attack robust models by these data points. Therefore, the adversarial images obtained by boundary attack characterize the images around the decision boundary of robust models. We attack our model by boundary attack with random spatial transformations, a baseline in the competition. The classification accuracy on the adversarial test data is as high as 95% (at 80% coverage), even though the adversarial corruptions are perceptible to human. We observe that the robust model trained by TRADES has strong interpretability: in Figure 4.14 all of adversarial images have obvious feature of 'bird', while in Figure 4.15 all of adversarial images have obvious feature of 'bicycle'. This shows that images around the decision boundary of truly robust model have features of both classes.

4.3.6 Case study: NeurIPS 2018 Adversarial Vision Challenge

Competition settings. In the NeurIPS 2018 Adversarial Vision Challenge [47], the adversarial attacks and defenses are under the black-box setting. The dataset in this challenge is Tiny ImageNet, which consists of 550,000 data (with our data augmentation) and 200 classes. The robust models only return label predictions instead of explicit gradients and confidence scores. The task for robust models is to defend against adversarial examples that are generated by the top-5 submissions in the un-targeted attack track. The score for each defense model is evaluated by the smallest perturbation distance that makes the defense model fail to output correct labels.

Competition results. The methodology in this section was applied to the competition, where our entry ranked the 1st place in the robust model track. We implemented our method to train ResNet models. We report the mean ℓ_2 perturbation distance of the top-6 entries in Figure 4.16. It shows that our method outperforms other approaches with a large margin. In particular, we surpass the runner-up submission by 11.41% in terms of mean ℓ_2 perturbation distance.

4.3.7 Proofs of our main results

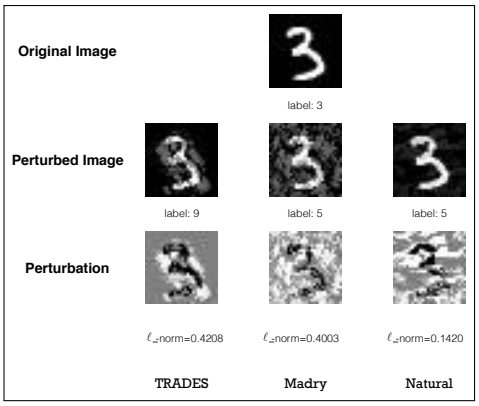
In this section, we provide the proofs of our main results.

Proofs of Theorem 51

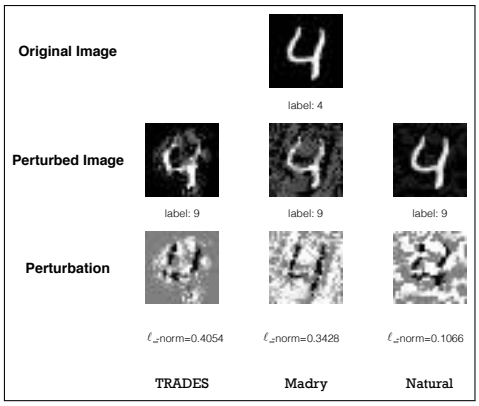
We denote by $f^*(\cdot) := 2\eta(\cdot) - 1$ the Bayes decision rule throughout the proofs.

Lemma 75. *For any classifier f, we have*

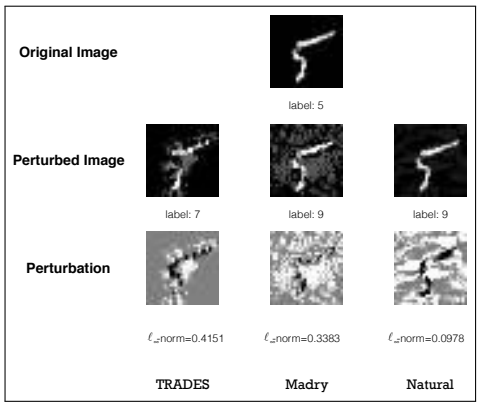
$$\begin{aligned}
&\mathcal{R}_{\mathrm{adv}}(f) - \mathcal{R}^*_{\mathrm{nat}} \\
=&\mathbb{E}[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(f^*(\mathbf{X})), \mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)^{\perp}\}|2\eta(\mathbf{X}) - 1|] \\
&+ \Pr[\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), \mathrm{sign}(f^*(\mathbf{X})) = Y].
\end{aligned}$$



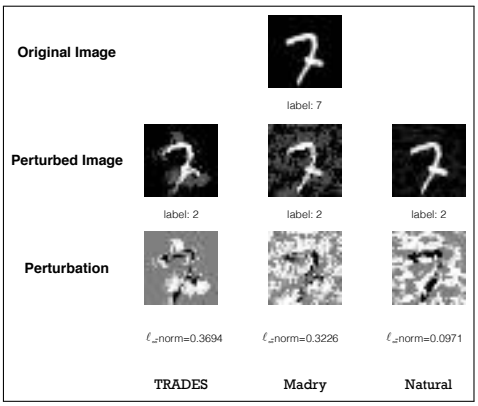
(a) adversarial examples of class '3'



(b) adversarial examples of class '4'

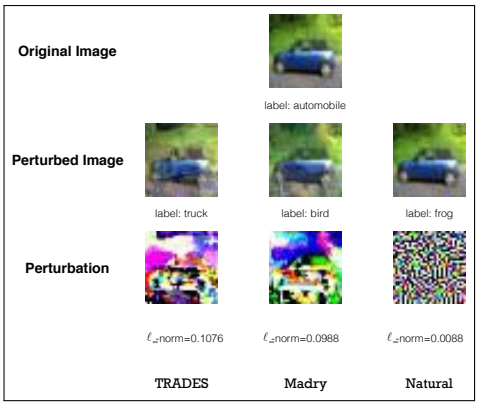


(c) adversarial examples of class '5'

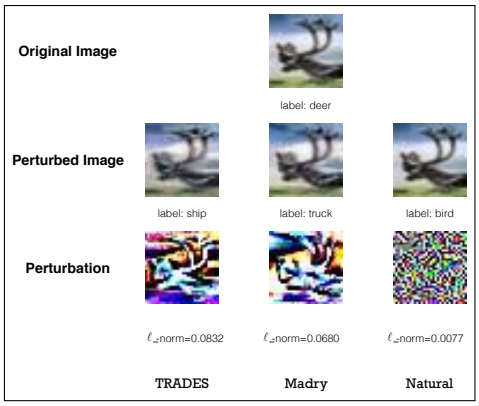


(d) adversarial examples of class '7'

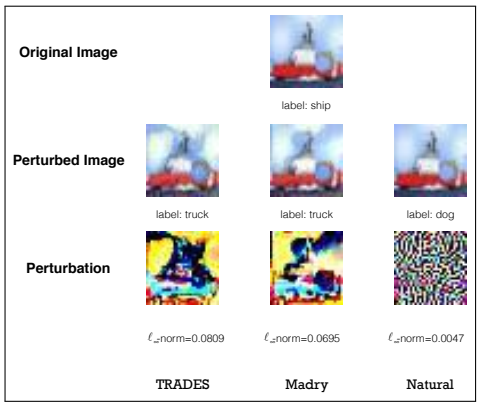
Figure 4.12: Adversarial examples on MNIST dataset. In each subfigure, the image in the first row is the original image and we list the corresponding correct label beneath the image. We show the perturbed images in the second row. The differences between the perturbed images and the original images, i.e., the perturbations, are shown in the third row. In each column, the perturbed image and the perturbation are generated by FGSM^k (white-box) attack on the model listed below. The labels beneath the perturbed images are the predictions of the corresponding models, which are different from the correct labels. We record the smallest perturbations in terms of ℓ_∞ norm that make the models predict a wrong label.



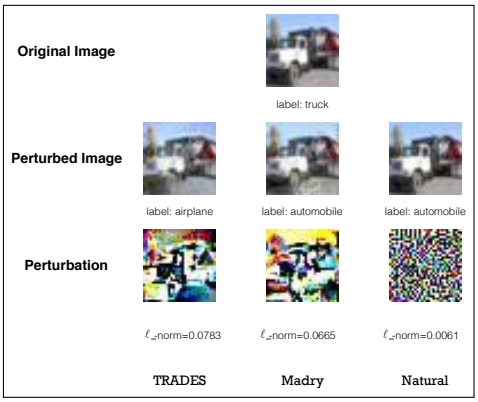
(a) adversarial examples of class 'automobile'



(b) adversarial examples of class 'deer'



(c) adversarial examples of class 'ship'



(d) adversarial examples of class 'truck'

Figure 4.13: Adversarial examples on CIFAR10 dataset. In each subfigure, the image in the first row is the original image and we list the corresponding correct label beneath the image. We show the perturbed images in the second row. The differences between the perturbed images and the original images, i.e., the perturbations, are shown in the third row. In each column, the perturbed image and the perturbation are generated by FGSM^k (white-box) attack on the model listed below. The labels beneath the perturbed images are the predictions of the corresponding models, which are different from the correct labels. We record the smallest perturbations in terms of ℓ_∞ norm that make the models predict a wrong label **(best viewed in color)**.

(a) clean example

(b) adversarial example by boundary attack with random spatial transformation

(c) clean example

(d) adversarial example by boundary attack with random spatial transformation

(e) clean example

(f) adversarial example by boundary attack with random spatial transformation

Figure 4.14: Adversarial examples by boundary attack with random spatial transformation on the ResNet-50 model trained by a variant of TRADES. The ground-truth label is 'bicycle', and our robust model recognizes the adversarial examples correctly as 'bicycle'. It shows in the second column that all of adversarial images have obvious feature of 'bird' **(best viewed in color)**.

(a) clean example

(b) adversarial example by boundary attack with random spatial transformation

(c) clean example

(d) adversarial example by boundary attack with random spatial transformation

(e) clean example

(f) adversarial example by boundary attack with random spatial transformation

Figure 4.15: Adversarial examples by boundary attack with random spatial transformation on the ResNet-50 model trained by a variant of TRADES. The ground-truth label is 'bird', and our robust model recognizes the adversarial examples correctly as 'bird'. It shows in the second column that all of adversarial images have obvious feature of 'bicycle' **(best viewed in color)**.

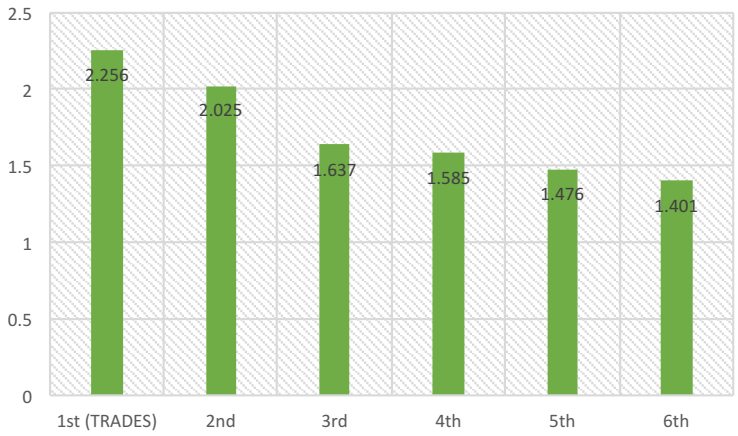


Figure 4.16: Top-6 results (out of 1,995 submissions) in the NeurIPS 2018 Adversarial Vision Challenge (Robust Model Track). The vertical axis represents the mean ℓ_2 perturbation distance that makes robust models fail to output correct labels.

Proof. For any classifier f, we have

$$
\begin{aligned}
&\Pr(\exists \mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{X}')) \neq Y | \mathbf{X} = \mathbf{x}) \\
&= \Pr(Y = 1, \exists \mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{X}')) = -1 | \mathbf{X} = \mathbf{x}) \\
&\quad + \Pr(Y = -1, \exists \mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{X}')) = 1 | \mathbf{X} = \mathbf{x}) \\
&= \mathbb{E}[\mathbf{1}\{Y = 1\}\mathbf{1}\{\exists \mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{X}')) = -1\} | \mathbf{X} = \mathbf{x}] \\
&\quad + \mathbb{E}[\mathbf{1}\{Y = -1\}\mathbf{1}\{\exists \mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{X}')) = 1\} | \mathbf{X} = \mathbf{x}] \\
&= \mathbf{1}\{\exists \mathbf{x}' \in \mathbb{B}(\mathbf{x}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{x}')) = -1\} \mathbb{E}\mathbf{1}\{Y = 1 | \mathbf{X} = \mathbf{x}\} \\
&\quad + \mathbf{1}\{\exists \mathbf{x}' \in \mathbb{B}(\mathbf{x}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{x}')) = 1\} \mathbb{E}\mathbf{1}\{Y = -1 | \mathbf{X} = \mathbf{x}\} \\
&= \mathbf{1}\{\exists \mathbf{x}' \in \mathbb{B}(\mathbf{x}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{x}')) = -1\} \eta(\mathbf{x}) + \mathbf{1}\{\exists \mathbf{x}' \in \mathbb{B}(\mathbf{x}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{x}')) = 1\}(1 - \eta(\mathbf{x})) \\
&= \begin{cases} 1, & \mathbf{x} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), \\ \mathbf{1}\{\operatorname{sign}(f(\mathbf{x})) = -1\}(2\eta(\mathbf{x}) - 1) + (1 - \eta(\mathbf{x})), & \text{otherwise.} \end{cases}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathcal{R}_{\mathrm{adv}}(f) \\
&= \int_{\mathcal{X}} \Pr[\exists \mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{X}')) \neq Y | \mathbf{X} = \mathbf{x}] d \Pr_{\mathbf{X}}(\mathbf{x}) \\
&= \int_{\mathbb{B}(\mathrm{DB}(f), \epsilon)} \Pr[\exists \mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{X}')) \neq Y | \mathbf{X} = \mathbf{x}] d \Pr_{\mathbf{X}}(\mathbf{x}) \\
&\quad + \int_{\mathbb{B}(\mathrm{DB}(f), \epsilon)^{\perp}} \Pr[\exists \mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon) \text{ s.t. } \operatorname{sign}(f(\mathbf{X}')) \neq Y | \mathbf{X} = \mathbf{x}] d \Pr_{\mathbf{X}}(\mathbf{x}) \\
&= \Pr(\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)) \\
&\quad + \int_{\mathbb{B}(\mathrm{DB}(f), \epsilon)^{\perp}} [\mathbf{1}\{\operatorname{sign}(f(\mathbf{x})) = -1\}(2\eta(\mathbf{x}) - 1) + (1 - \eta(\mathbf{x}))] d \Pr_{\mathbf{X}}(\mathbf{x}).
\end{aligned}
$$

We have

$$
\begin{aligned}
&\mathcal{R}_{\mathrm{adv}}(f) - \mathcal{R}_{\mathrm{nat}}(f^*) \\
&= \Pr(\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)) + \int_{\mathbb{B}(\mathrm{DB}(f),\epsilon)^\perp} [\mathbf{1}\{\mathrm{sign}(f(\mathbf{x})) = -1\}(2\eta(\mathbf{x}) - 1) + (1 - \eta(\mathbf{x}))] d \Pr_{\mathbf{X}}(\mathbf{x}) \\
&\quad - \int_{\mathbb{B}(\mathrm{DB}(f),\epsilon)^\perp} [\mathbf{1}\{\mathrm{sign}(f^*(\mathbf{x})) = -1\}(2\eta(\mathbf{x}) - 1) + (1 - \eta(\mathbf{x}))] d \Pr_{\mathbf{X}}(\mathbf{x}) \\
&\quad - \int_{\mathbb{B}(\mathrm{DB}(f),\epsilon)} [\mathbf{1}\{\mathrm{sign}(f^*(\mathbf{x})) = -1\}(2\eta(\mathbf{x}) - 1) + (1 - \eta(\mathbf{x}))] d \Pr_{\mathbf{X}}(\mathbf{x}) \\
&= \Pr(\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)) - \int_{\mathbb{B}(\mathrm{DB}(f),\epsilon)} [\mathbf{1}\{\mathrm{sign}(f^*(\mathbf{x})) = -1\}(2\eta(\mathbf{x}) - 1) + (1 - \eta(\mathbf{x}))] d \Pr_{\mathbf{X}}(\mathbf{x}) \\
&\quad + \mathbb{E}[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(\eta(\mathbf{X}) - 1/2), \mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)^\perp\} |2\eta(\mathbf{X}) - 1|] \\
&= \Pr(\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)) - \mathbb{E}[\mathbf{1}\{\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)\} \min\{\eta(\mathbf{X}), 1 - \eta(\mathbf{X})\}] \\
&\quad + \mathbb{E}[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(\eta(\mathbf{X}) - 1/2), \mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)^\perp\} |2\eta(\mathbf{X}) - 1|] \\
&= \mathbb{E}[\mathbf{1}\{\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)\} \max\{\eta(\mathbf{X}), 1 - \eta(\mathbf{X})\}] \\
&\quad + \mathbb{E}[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(\eta(\mathbf{X}) - 1/2), \mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)^\perp\} |2\eta(\mathbf{X}) - 1|] \\
&= \Pr[\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), \mathrm{sign}(f^*(\mathbf{X})) = Y] \\
&\quad + \mathbb{E}[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(f^*(\mathbf{X})), \mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)^\perp\} |2\eta(\mathbf{X}) - 1|].
\end{aligned}
$$

□

Now we are ready to prove Theorem 51.

Theorem 51 (restated). *Under Assumption 2, for any non-negative loss function ϕ such that $\phi(0) \geq 1$, any measurable $f : \mathcal{X} \to \mathbb{R}$, any probability distribution on $\mathcal{X} \times \{\pm 1\}$, and any $\lambda > 0$, we have*

$$
\begin{aligned}
\mathcal{R}_{\mathrm{adv}}(f) - \mathcal{R}^*_{\mathrm{nat}} &\leq \psi^{-1}(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi) + \Pr[\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), c_0(\mathbf{X}) = Y] \\
&\leq \psi^{-1}(\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi) + \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')f(\mathbf{X})/\lambda),
\end{aligned}
$$

where $\mathcal{R}^*_\phi := \min_f \mathcal{R}_\phi(f)$ and $c_0(\cdot) = \mathrm{sign}(2\eta(\cdot) - 1)$ is the Bayes optimal classifier.

Proof. By Lemma 75, we note that

$$
\begin{aligned}
&\psi(\mathcal{R}_{\mathrm{adv}}(f) - \mathcal{R}_{\mathrm{nat}}(f^*) - \Pr[\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), \mathrm{sign}(f^*(\mathbf{X})) = Y]) \\
&= \psi(\mathbb{E}[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(f^*(\mathbf{X})), \mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)^\perp\} |2\eta(\mathbf{X}) - 1|]) \\
&\leq \mathbb{E}[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(f^*(\mathbf{X})), \mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)^\perp\} \psi(|2\eta(\mathbf{X}) - 1|)] \\
&\leq \mathbb{E}[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(f^*(\mathbf{X}))\} \psi(|2\eta(\mathbf{X}) - 1|)] \\
&= \mathbb{E}[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(f^*(\mathbf{X}))\} \times (H^-(\eta(\mathbf{X})) - H(\eta(\mathbf{X})))] \\
&= \mathbb{E}\left[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(f^*(\mathbf{X}))\} \times \left(\inf_{\alpha : \alpha(2\eta(\mathbf{X}) - 1) \leq 0} C_{\eta(\mathbf{X})}(\alpha) - H(\eta(\mathbf{X}))\right)\right] \\
&\leq \mathbb{E}[C_{\eta(\mathbf{X})}(f(\mathbf{X})) - H(\eta(\mathbf{X}))] \\
&= \mathcal{R}_\phi(f) - \mathcal{R}^*_\phi.
\end{aligned}
$$

Also, notice that

$$\begin{aligned}\Pr[\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), \mathrm{sign}(f^*(\mathbf{X})) = Y] &\le \Pr[\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)] \\ &= \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \mathbf{1}\{f(\mathbf{X}') \neq f(\mathbf{X})\} \\ &= \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \mathbf{1}\{f(\mathbf{X}')f(\mathbf{X})/\lambda < 0\} \\ &\le \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')f(\mathbf{X})/\lambda),\end{aligned}$$

as desired. □

Proofs of Theorem 52

Theorem 52 (restated). *Suppose that $|\mathcal{X}| \ge 2$. Under Assumption 2, for any non-negative loss function ϕ such that $\phi(x) \to 0$ as $x \to +\infty$, any $\xi > 0$, and any $\theta \in [0, 1]$, there exists a probability distribution on $\mathcal{X} \times \{\pm 1\}$, a function $f : \mathbb{R}^d \to \mathbb{R}$, and a regularization parameter $\lambda > 0$ such that $\mathcal{R}_{\mathrm{adv}}(f) - \mathcal{R}^*_{\mathrm{nat}} = \theta$ and*

$$\psi\Big(\theta - \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')f(\mathbf{X})/\lambda)\Big) \le \mathcal{R}_\phi(f) - \mathcal{R}^*_\phi \le \psi\left(\theta - \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')f(\mathbf{X})/\lambda)\right) + \xi.$$

Proof. The first inequality follows from Theorem 51. Thus it suffices to prove the second inequality.

Fix $\epsilon > 0$ and $\theta \in [0, 1]$. By the definition of ψ and its continuity, we can choose $\gamma, \alpha_1, \alpha_2 \in [0, 1]$ such that $\theta = \gamma\alpha_1 + (1 - \gamma)\alpha_2$ and $\psi(\theta) \ge \gamma\tilde{\psi}(\alpha_1) + (1 - \gamma)\tilde{\psi}(\alpha_2) - \epsilon/3$. For two distinct points $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we set $\mathcal{P}_\mathcal{X}$ such that $\Pr[\mathbf{X} = \mathbf{x}_1] = \gamma$, $\Pr[\mathbf{X} = \mathbf{x}_2] = 1 - \gamma$, $\eta(\mathbf{x}_1) = (1 + \alpha_1)/2$, and $\eta(\mathbf{x}_2) = (1 + \alpha_2)/2$. By the definition of H^-, we choose function $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(\mathbf{x}) < 0$ for all $\mathbf{x} \in \mathcal{X}$, $C_{\eta(\mathbf{x}_1)}(f(\mathbf{x}_1)) \le H^-(\eta(\mathbf{x}_1)) + \epsilon/3$, and $C_{\eta(\mathbf{x}_2)}(f(\mathbf{x}_2)) \le H^-(\eta(\mathbf{x}_2)) + \epsilon/3$. By the continuity of ψ, there is an $\epsilon' > 0$ such that $\psi(\theta) \le \psi(\theta - \epsilon_0) + \epsilon/3$ for all $0 \le \epsilon_0 < \epsilon'$. We also note that there exists an $\lambda_0 > 0$ such that for any $0 < \lambda < \lambda_0$, we have

$$0 \le \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')f(\mathbf{X})/\lambda) < \epsilon'.$$

Thus, we have

$$\begin{aligned}\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi &= \mathbb{E}\phi(Yf(\mathbf{X})) - \inf_f \mathbb{E}\phi(Yf(\mathbf{X})) \\ &= \gamma[C_{\eta(\mathbf{x}_1)}(f(\mathbf{x}_1)) - H(\eta(\mathbf{x}_1))] + (1 - \gamma)[C_{\eta(\mathbf{x}_2)}(f(\mathbf{x}_2)) - H(\eta(\mathbf{x}_2))] \\ &\le \gamma[H^-(\eta(\mathbf{x}_1)) - H(\eta(\mathbf{x}_1))] + (1 - \gamma)[H^-(\eta(\mathbf{x}_2)) - H(\eta(\mathbf{x}_2))] + \epsilon/3 \\ &= \gamma\tilde{\psi}(\alpha_1) + (1 - \gamma)\tilde{\psi}(\alpha_2) + \epsilon/3 \\ &\le \psi(\theta) + 2\epsilon/3 \\ &\le \psi\left(\theta - \mathbb{E} \max_{\mathbf{X}' \in \mathbb{B}(\mathbf{X}, \epsilon)} \phi(f(\mathbf{X}')f(\mathbf{X})/\lambda)\right) + \epsilon.\end{aligned}$$

Furthermore, by Lemma 75,

$$\begin{aligned}\mathcal{R}_{\mathrm{adv}}(f) - \mathcal{R}^*_{\mathrm{nat}} &= \mathbb{E}[\mathbf{1}\{\mathrm{sign}(f(\mathbf{X})) \neq \mathrm{sign}(f^*(\mathbf{X})), \mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon)^{\perp}\}|2\eta(\mathbf{X}) - 1|] \\ &\quad + \Pr[\mathbf{X} \in \mathbb{B}(\mathrm{DB}(f), \epsilon), \mathrm{sign}(f^*(\mathbf{X})) = Y] \\ &= \mathbb{E}|2\eta(\mathbf{X}) - 1| \\ &= \gamma(2\eta(\mathbf{x}_1) - 1) + (1 - \gamma)(2\eta(\mathbf{x}_2) - 1) \\ &= \theta,\end{aligned}$$

where f^* is the Bayes optimal classifier which outputs "positive" for all data points. □

Chapter 5

Conclusion and Discussion

Throughout this thesis, we have witnessed that sparse learning, deep networks and adversarial learning are ubiquitous in various machine learning topics, ranging from learning from sparsity, learning with low-rank approximations, to learning with deep neural networks and more. These correspond to the data assumptions that the underlying data is sparse, is of low-rank, and is drawn from low-dimensional manifold, respectively. On the other hand, although these new paradigms have been widely applied to the real-world problems such as AlphaGo, AlphaStar, autonomous vehicle, medical AI and so on, many fundamental questions remain unresolved. This requires us to better understand these learning problems from ROSE perspective: Robustness, Optimization, and Sample Efficiency.

5.1 Robustness

Robustness is always at the heart of machine learning models; without it, learning models cannot be deployed to the real applications as they will be vulnerable to adversarial attacks. In this thesis, we discuss the robustness of sparse active learning, robust PCA, as well as the classification by deep neural networks, while we use three different techniques to analyze the models. For sparse active learning, we show that the localization technique itself is robust to adversarial noise model; for robust PCA, replacing ℓ_0 norm with ℓ_1 norm is robust to sparse corruption as well; and for deep neural networks, we identify a trade-off between robustness and accuracy that serves as a guiding principle in the design of defenses against adversarial examples. Our contributions are not only from theoretical aspects, but also are from practical aspects. For example, we implement our robust training algorithms of supervised classification by deep neural networks. The methodology is the winning submission of NeurIPS 2018 Adversarial Vision Challenge, in which we won the first place out of 1,995 submissions, surpassing the runner-up approach by a large margin. Our model TRADES also create a new record in the Unrestricted Adversarial Examples Challenge hosted by Google.

5.2 Optimization

Non-convex learning is notorious for its hardness to achieve global optimality in polynomial time. For generic non-convex learning problems, it has been proved to be NP-hard to get global optimality. However, if we can make use of nice structure of some non-convex learning problems, we can hopefully solve for their global optimality in polynomial time. In this thesis, we show that the idea works for many non-convex learning problems, ranging from margin-based active learning under Massart and adversarial noise models, matrix factorization with random sampling of its entries, to provable approximate global optimality of deep neural networks and GANs with multi-branch architecture. At the core of our analysis is the duality gap which serves as a measure to characterize the non-convexity of learning problems.

5.3 Sample Efficiency

This thesis also cares about the sample/label efficiency of sparse learning algorithms, targeting on designing learning algorithms with (near-)optimal sample complexity. This requires us to understand the hardness of problem independent of any specific algorithm (by showing the lower bound), as well as providing *tight* upper bound by analyzing the proposed algorithms. For almost all the problems that were studied in this thesis (we did not try to discuss the sample complexity of learning by deep neural network), we provide tight (upper and lower) bounds for the sample complexity of sparse learning problems. These models include margin-based active learning under Massart and adversarial noise models, adaptive compressed sensing, matrix completion, robust PCA, and property testing of matrix rank.

5.4 Future Directions

Finally, we propose some interesting directions for the future study.

5.4.1 Small-data learning by self-supervised and semi-supervised learning

Supervised learning has been widely applied to various fields ranging from computer vision to natural language processing. However, the framework is data-hungry: to learn a deep neural network for image classification tasks, it typically requires more than $10,000$ labelled data, while collecting labels are labor-intensive and expensive. Therefore, small-data learning has become more and more popular in recent years. It will not only reduce the cost of data collection, but small data is also friendly to short-time training procedures. This is highly related to the "sample-efficiency" theme in this thesis.

We are particularly interested in how to use the structure of data itself to learn representation for some pretext tasks, and apply the representation to the downstream tasks. This is also known as the self-supervised learning, because the practicableness of the pretext tasks typically comes from the fake labels of data itself. For example, given a image of cat with patch "cat nose" and "cat right ear", we know that the latter patch should be at the top right conner of the former patch.

This positional relation serves as a fake label, and we hope that by taking this into account, the number of true required labels can be reduced.

Semi-supervised learning is another way to improve the label complexity. Typically semi-supervised learning includes co-training and Transductive SVM. We might combine our discovery in this thesis with these new frameworks to achieve improved sample-efficient learning.

5.4.2 Robust learning by self-supervised and semi-supervised learning

The idea of self-supervised learning or co-training might be applied to the analysis of adversarial examples as well. In the co-training and some pretext tasks in the self-supervised learning, there are two views $\mathbf{X}_1$ and $\mathbf{X}_2$ for each instance $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. For example, $\mathbf{X}$ may represent an image of a cat, X_1 represents a patch of the image corresponding to cat nose, and $\mathbf{X}_2$ represents a patch of the image corresponding to cat's right ear. Therefore, $\mathbf{X}_1$ and $\mathbf{X}_2$ are two different views of the same cat image $\mathbf{X}$. However, $\mathbf{X}_1$ and X_2 are highly related: it is known that $\mathbf{X}_2$ should be at the top right conner compared with the position of X_1 in the image. Suppose there is an adversary who generates an adversarial example $\mathbf{X}' = (\mathbf{X}'_1, \mathbf{X}'_2)$ by adding small perturbation on the top of $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, that is, $\|\mathbf{X} - \mathbf{X}'\| \leq \epsilon$. Since the standard training method of Deep Neural Networks (DNNs) implicitly processes each patch separately without taking into account the global relations among various patches in the image (see [45]), it is quite possible that DNNs fail to recognize that $\mathbf{X}'_2$ should be at the top right conner of $\mathbf{X}'_1$. With this, we can detect whether a given image is an adversarial example or not. Furthermore, if we extract two features related to two different views $\mathbf{X}_1$ and $\mathbf{X}_2$ with considerations of their positional relationship, the DNNs trained by the two features intuitively should be much more robust to the adversarial examples, because there is one extra constraint (e.g., the positional relationship) for the generation of adversarial examples, which is harder. In summary, the "consistency" and "compatibility" between the two views $\mathbf{X}_1$ and $\mathbf{X}_2$ in the language of self-supervised learning and co-training might play a key role in the defense against adversarial examples $(\mathbf{X}'_1, \mathbf{X}'_2)$, in the hope that $\mathbf{X}'_1$ might be inconsistent and incompatible with $\mathbf{X}'_2$.

Another interesting open question is that whether unlabeled data may help alleviate the vulnerability of DNNs to adversarial examples. It is known that adversarially robust generalization requires more data [205]. However, labeled data is very expensive, while unlabeled data is cheap to collect. So it is interesting to see whether unlabeled data (especially in the framework of semi-supervised learning) can help. Fortunately, model (4.35) has been ready for this purpose: note that the second term in model (4.35) measures the difference between $f(\mathbf{X})$ and its adversarial counterpart $f(\mathbf{X}')$; no label information is required here. Therefore, model (4.35) can be directly applied to the semi-supervised framework (using the unlabeled data to estimate the second term more precisely without the requirement of label information). In fact, there are some previous works which have explored this idea in the name of virtual adversarial training [171]. However, there is no theoretical support in this line of research. The existing analysis of co-training might be a good starting point for this problem.

5.4.3 Other potential directions

Perturbation resilience is a famous data assumption that the optimum clustering to the objective is preserved under small multiplicative perturbations to distances between points [19, 23]. This data assumption may have straightforward connections with the analysis of TRADES in Section 4.3: the success of the regularization term in Equ. (4.35) may have implicitly built upon perturbation resilience, as it requires the neighbour of data points should have stable output. It is interesting to see how the techniques of perturbation resilience [19, 23] can be applied to the robustness analysis of supervised learning.

Bibliography

[1] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, pages 1171–1197, 2012.

[2] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima for nonconvex optimization in linear time. *Annual ACM Symposium on Theory of Computing*, 2017.

[3] Rima Alaifari, Giovanni S Alberti, and Tandri Gauksson. ADef: an iterative algorithm to construct adversarial deformations. In *International Conference on Learning Representations*, 2019.

[4] Akram Aldroubi, Haichao Wang, and Kourosh Zarringhalam. Sequential adaptive compressed sampling via Huffman codes. *arXiv preprint arXiv:0810.4916*, 2008.

[5] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Annual ACM Symposium on Theory of Computing*, 2016.

[6] Anima Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *Annual Conference on Learning Theory*, 2016.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[8] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *ACM symposium on Theory of computing*, pages 145–162, 2012.

[9] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, pages 224–232, 2017.

[10] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, 2018.

[11] Özlem Aslan, Xinhua Zhang, and Dale Schuurmans. Convex deep learning via normalized kernels. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2014.

[12] Sepehr Assadi, Sanjeev Khanna, and Yang Li. On estimating maximum matching size in graph streams. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1723–1742, 2017.

[13] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false

sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[14] Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. In *Annual ACM Symposium on Theory of Computing*, pages 449–458. ACM, 2014.

[15] Pranjal Awasthi, Maria Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In *Annual Conference on Learning Theory*, 2015.

[16] Pranjal Awasthi, Maria Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM*, 2015.

[17] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Annual Conference on Learning Theory*, pages 152–192, 2016.

[18] Khanh Do Ba, Piotr Indyk, Eric Price, and David P. Woodruff. Lower bounds for sparse recovery. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1190–1197, 2010.

[19] Maria Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. *SIAM Journal on Computing*, 45(1):102–155, 2016.

[20] Maria Florina Balcan and Phillip M. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2013.

[21] Maria-Florina Balcan and Hongyang Zhang. Noise-tolerant life-long matrix completion via adaptive sampling. In *Advances in Neural Information Processing Systems*, pages 2955–2963, 2016.

[22] Maria Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *Annual Conference on Learning Theory*, 2007.

[23] Maria-Florina Balcan, Nika Haghtalab, and Colin White. k-center clustering under perturbation resilience. *arXiv preprint arXiv:1505.03924*, 2015.

[24] Maria-Florina Balcan, Travis Dick, Yingyu Liang, Wenlong Mou, and Hongyang Zhang. Differentially private clustering in high-dimensional euclidean spaces. In *International Conference on Machine Learning*, pages 322–331, 2017.

[25] Maria-Florina Balcan, Yi Li, David P Woodruff, and Hongyang Zhang. Testing matrix rank, optimally. In *Annual ACM-SIAM Symposium on Discrete Algorithms*, 2018.

[26] Maria-Florina Balcan, Yingyu Liang, David P. Woodruff, and Hongyang Zhang. Matrix completion and related problems via strong duality. In *Innovations in Theoretical Computer Science*, volume 94, 2018.

[27] Maria-Florina Balcan, Zhao Liang, Yingyu Song, David P. Woodruff, and Hongyang Zhang. Non-convex matrix completion and related problems via strong duality. *Journal of Machine Learning Research*, 2019.

[28] Maria-Florina F Balcan and Hongyang Zhang. Sample and computationally efficient learning algorithms under s-concave distributions. In *Advances in Neural Information*

Processing Systems, pages 4799–4808, 2017.

[29] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

[30] Ziv Bar-Yossef, TS Jayram, Ravi Kumar, and D Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68:702–732, 2004.

[31] Peter Bartlett and Shai Ben-David. Hardness results for neural network approximation problems. In *European Conference on Computational Learning Theory*, pages 50–62, 1999.

[32] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[33] Amir Beck and Yonina C Eldar. Strong duality in nonconvex quadratic optimization with two quadratic constraints. *SIAM Journal on Optimization*, 17(3):844–860, 2006.

[34] David Berthelot, Thomas Schumm, and Luke Metz. BEGAN: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[35] D Bertsekas. Min common/max crossing duality: A geometric view of conjugacy in convex optimization. *Lab. for Information and Decision Systems, MIT, Tech. Rep. Report LIDS-P-2796*, 2009.

[36] Dimitri P Bertsekas and Nils R Sandell. Estimates of the duality gap for large-scale separable nonconvex optimization problems. In *IEEE Conference on Decision and Control*, volume 21, pages 782–785, 1982.

[37] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[38] Yingjie Bi and Ao Tang. Refined Shapely-Folkman lemma and its application in duality gap estimation. *arXiv preprint arXiv:1610.05416*, 2016.

[39] Avrim Blum and Ronald L Rivest. Training a 3-node neural network is NP-complete. In *Advances in neural information processing systems*, pages 494–501, 1989.

[40] Avrim Blum, Alan M. Frieze, Ravi Kannan, and Santosh Vempala. A polynomial time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1997.

[41] Émile Borel. Sur les principes de la théorie cinétique des gaz. *Annales scientifiques de l'École Normale Supérieure, Série 3*, 23:9–32, 1904.

[42] Olivier Bousquet, Stéphane Boucheron, and Gabor Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:9:323–375, 2005.

[43] Thierry Bouwmans, Sajid Javed, Hongyang Zhang, Zhouchen Lin, and Ricardo Otazo. On the applications of robust PCA in image and video processing. *Proceedings of the IEEE*, 106(8):1427–1457, 2018.

[44] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[45] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations*, 2019.

[46] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018.

[47] Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Marcel Salathé, Sharada P Mohanty, and Matthias Bethge. Adversarial vision challenge. *arXiv preprint arXiv:1808.01976*, 2018.

[48] Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018.

[49] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a ConvNet with Gaussian inputs. In *International Conference on Machine Learning*, 2017.

[50] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations*, 2018.

[51] Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from cryptographic pseudo-random generators. *arXiv preprint arXiv:1811.06418*, 2018.

[52] Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

[53] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018.

[54] Marc Bury and Chris Schwiegelshohn. Sublinear estimation of weighted matchings in dynamic data streams. In *Algorithms-ESA 2015*, pages 263–274. 2015.

[55] Emmanuel J. Candès and Ben Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[56] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[57] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

[58] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

[59] Rui M. Castro, Jarvis Haupt, Robert Nowak, and Gil M. Raz. Finding needles in noisy haystacks. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5133–5136, 2008.

[60] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *ACM-SIAM Symposium on Discrete*

Algorithms, pages 1193–1203, 2014.

[61] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6): 805–849, 2012.

[62] Minming Chen, Zhouchen Lin, Yi Ma, and Leqin Wu. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Coordinated Science Laboratory Report no. UILU-ENG-09-2215*, 2009.

[63] Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

[64] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[65] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of ImageNet as an alternative to the CIFAR datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[66] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, 2017.

[67] Kenneth L Clarkson and David P Woodruff. Numerical linear algebra in the streaming model. In *ACM Symposium on Theory of Computing*, pages 205–214, 2009.

[68] Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *ACM Symposium on Theory of Computing*, pages 81–90, 2013.

[69] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[70] Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. PAC-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pages 228–239, 2018.

[71] Amit Daniely. Complexity theoretic limitations on learning halfspaces. *CoRR*, abs/1505.05800, 2015.

[72] Amit Daniely. A PTAS for agnostically learning halfspaces. In *Annual Conference on Learning Theory*, 2015.

[73] Bhaskar DasGupta, Hava T Siegelmann, and Eduardo Sontag. On the complexity of training neural networks with continuous activation functions. *IEEE Transactions on Neural Networks*, 6(6):1490–1504, 1995.

[74] Constantinos Daskalakis, Ilias Diakonikolas, Rocco A Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1833–1852, 2013.

[75] Alexandre d'Aspremont and Igor Colin. An approximate Shapley-Folkman theorem. *arXiv preprint arXiv:1712.08559*, 2017.

[76] Amit Deshpande, Madhur Tulsiani, and Nisheeth K Vishnoi. Algorithms and hardness for subspace approximation. In *ACM-SIAM symposium on Discrete Algorithms*, pages 482–496, 2011.

[77] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

[78] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *International Conference on Learning Representations*, 2017.

[79] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673*, 2016.

[80] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

[81] Logan Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.

[82] Hossein Esfandiari, Mohammad T Hajiaghayi, Vahid Liaghat, Morteza Monemizadeh, and Krzysztof Onak. Streaming algorithms for estimating the matching size in planar graphs and beyond. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1217–1233, 2014.

[83] Ethan X Fang, Han Liu, and Mengdi Wang. Blessing of massive scale: Spatial graphical model estimation with a total cardinality constraint. 2015.

[84] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[85] David Gamarnik, Quan Li, and Hongyi Zhang. Matrix completion from $O(n)$ samples in linear time. In *Annual Conference on Learning Theory*, 2017.

[86] Xavier Gastaldi. Shake-Shake regularization. *arXiv preprint arXiv:1705.07485*, 2017.

[87] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Annual Conference on Learning Theory*, pages 797–842, 2015.

[88] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[89] Rong Ge, Chi Jin, and Zheng Yi. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *International Conference on Machine Learning*, 2017.

[90] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2017.

[91] Arnab Ghosh, Viveka Kulharia, Vinay Namboodiri, Philip HS Torr, and Puneet K Dokania. Multi-agent diverse generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8513–8521, 2017.

[92] Anna C. Gilbert, Yi Li, Ely Porat, and Martin J. Strauss. Approximate sparse recovery: optimizing time and measurements. *SIAM Journal on Computing*, 41(2):436–453, 2012.

[93] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

[94] Andreas Goerdt and André Lanka. An approximation hardness result for bipartite clique. In *Electronic Colloquium on Computational Complexity, Report*, volume 48, 2004.

[95] Ian Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[96] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[97] Sivakant Gopi, Praneeth Netrapalli, Prateek Jain, and Aditya Nori. One-bit compressed sensing: Provable support and vector recovery. In *International Conference on Machine Learning*, pages 154–162, 2013.

[98] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[99] Christian Grussler, Anders Rantzer, and Pontus Giselsson. Low-rank optimization with convex constraints. *arXiv preprint arXiv:1606.01793*, 2016.

[100] Quanquan Gu, Zhaoran Wang, and Han Liu. Low-rank and sparse structure pursuit via alternating minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 600–609, 2016.

[101] Rishi Gupta, Piotr Indyk, Eric Price, and Yaron Rachlin. Compressive sensing with local geometric features. *International Journal of Computational Geometry & Applications*, 22 (04):365–390, 2012.

[102] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. In *Annual IEEE Symposium on Foundations of Computer Science*, 2006.

[103] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[104] Benjamin Haeffele, Eric Young, and Rene Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *International Conference on Machine Learning*, pages 2007–2015, 2014.

[105] Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

[106] Moritz Hardt. Understanding alternating minimization for matrix completion. In *IEEE Symposium on Foundations of Computer Science*, pages 651–660, 2014.

[107] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. *COLT*, 2013.

[108] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Annual Conference on Learning Theory*, pages 354–375, 2013.

[109] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for

differentially private data release. In *Advances in Neural Information Processing Systems*, pages 2339–2347, 2012.

[110] Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *Annual Conference on Learning Theory*, pages 703–725, 2014.

[111] Jarvis Haupt, Waheed U Bajwa, Michael Rabbat, and Robert Nowak. Compressed sensing for networked data. *IEEE Signal Processing Magazine*, 25(2):92–101, 2008.

[112] Jarvis Haupt, Robert Nowak, and Rui Castro. Adaptive sensing for sparse signal recovery. In *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop*, pages 702–707, 2009.

[113] Jarvis D. Haupt, Richard G. Baraniuk, Rui M. Castro, and Robert D. Nowak. Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements. In *Asilomar Conference on Signals, Systems and Computers*, pages 1551–1555, 2009.

[114] Bingsheng He and Xiaoming Yuan. On the $O(1/n)$ convergence rate of the douglas–rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[115] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[116] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defenses: Ensembles of weak defenses are not strong. *arXiv preprint arXiv:1706.04701*, 2017.

[117] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[118] Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. MGAN: Training generative adversarial nets with multiple generators. In *International Conference on Learning Representations*, 2018.

[119] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

[120] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*, 2016.

[121] Piotr Indyk, Eric Price, and David P. Woodruff. On the power of adaptivity in sparse recovery. In *Annual IEEE Symposium on Foundations of Computer Science*, pages 285–294, 2011.

[122] Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.

[123] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using

alternating minimization. In *ACM Symposium on Theory of Computing*, pages 665–674, 2013.

[124] Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008.

[125] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Empirical Methods in Natural Language Processing*, 2017.

[126] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *International Conference on Machine Learning*, 2017.

[127] Raghunandan M. Kainkaryam, Angela Bruex, Anna C. Gilbert, John Schiefelbein, and Peter J. Woolf. poolmc: Smart pooling of mrna samples in microarray experiments. *BMC Bioinformatics*, 11:299, 2010.

[128] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. In *Annual IEEE Symposium on Foundations of Computer Science*, 2005.

[129] Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In *Annual ACM Symposium on Theory of Computing*, pages 629–638, 2008.

[130] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[131] Kenji Kawaguchi. Deep learning without poor local minima. *arXiv preprint arXiv:1605.07110*, 2016.

[132] Michael Kearns and Ming Li. Learning in the presence of malicious errors. In *Annual ACM Symposium on Theory of Computing*, 1988.

[133] Michael J Kearns and Umesh Virkumar Vazirani. *An introduction to computational learning theory*. MIT Press, Cambridge, MA, 1994.

[134] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[135] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[136] Adam Klivans and Pravesh Kothari. Embedding hard learning problems into gaussian space. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, 28:793–809, 2014.

[137] J Zico Kolter and Eric Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, 2018.

[138] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[139] Robert Krauthgamer and Ori Sasson. Property testing of data dimensionality. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 18–27, 2003.

[140] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny

images. 2009.

[141] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

[142] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

[143] Thomas Laurent and James von Brecht. The multilinear structure of ReLU networks. *arXiv preprint arXiv:1712.10132*, 2017.

[144] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[145] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Society, 2005.

[146] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. A Series of Modern Surveys in Mathematics Series. Springer, 1991.

[147] Hao Li, Zheng Xu, Gavin Taylor, and Tom Goldstein. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.

[148] Yi Li and Vasileios Nakos. Sublinear-time algorithms for compressive phase retrieval. *arXiv preprint arXiv:1709.02917*, 2017.

[149] Yi Li and David P. Woodruff. On approximating functions of the singular values in a stream. In *ACM Symposium on Theory of Computing*, pages 726–739, 2016.

[150] Yi Li and David P. Woodruff. Tight bounds for sketching the operator norm, Schatten norms, and subspace embeddings. In *International Conference on Approximation Algorithms for Combinatorial Optimization Problems, and International Conference on Randomization and Computation*, pages 39:1–39:11, 2016.

[151] Yi Li and David P. Woodruff. Embeddings of Schatten norms with applications to data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 60:1–60:14, 2017.

[152] Yi Li, Huy L. Nguyen, and David P. Woodruff. On sketching matrix norms and the top singular vector. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 1562–1581, 2014.

[153] Yi Li, Zhengyu Wang, and David P. Woodruff. Improved testing of low rank matrices. In *International Conference on Knowledge Discovery and Data Mining*, pages 691–700, 2014.

[154] Yi Li, Huy L. Nguyên, and David P. Woodruff. On approximating matrix norms in a stream. 2017. Submitted.

[155] Yuanzhi Li, Yingyu Liang, and Andrej Risteski. Recovery guarantee of weighted low-rank approximation via alternating minimization. In *International Conference on Machine Learning*, pages 2358–2367, 2016.

[156] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Annual*

Conference on Learning Theory, 2017.

[157] Shiyu Liang, Ruoyu Sun, Yixuan Li, and R Srikant. Understanding the loss surface of neural networks for binary classification. In *International Conference on Machine Learning*, 2018.

[158] Zhouchen Lin and Hongyang Zhang. *Low Rank Models for Visual Analysis: Theories, Algorithms and Applications*. Elsevier, 2017.

[159] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

[160] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures and Algorithms*, 30(3):307–358, 2007.

[161] Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv:1702.08580*, 2017.

[162] F. Lust-Piquard. Inégalités de Khintchine dans c_p $(1 < p < \infty)$. *C. R. Math. Acad. Sci. Paris Sér I Math.*, 303:289–292, 1986.

[163] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Michael E Houle, Grant Schoenebeck, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

[164] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[165] Malik Magdon-Ismail. Row sampling for matrix algorithms via a non-commutative Bernstein bound. *arXiv preprint arXiv:1008.0587*, 2010.

[166] Thomas L Magnanti, Jeremy F Shapiro, and Michael H Wagner. Generalized linear programming solves the dual. *Management Science*, 22(11):1195–1203, 1976.

[167] Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[168] Dmitry M. Malioutov, Sujay Sanghavi, and Alan S. Willsky. Compressed sensing with sequential observations. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3357–3360, 2008.

[169] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 135–147, 2017.

[170] Marvin L Minsky and Seymour A Papert. *Perceptrons - Expanded Edition: An Introduction to Computational Geometry*. MIT press Boston, MA:, 1987.

[171] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[172] Ankur Moitra. An almost optimal algorithm for computing nonnegative rank. *SIAM Journal on Computing*, 45(1):156–173, 2016.

[173] Tom Morgan and Jelani Nelson. A note on reductions between compressed sensing guarantees. *CoRR*, abs/1606.00757, 2016.

[174] Shanmugavelayutham Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236, 2005.

[175] Taesik Na, Jong Hwan Ko, and Saibal Mukhopadhyay. Cascade adversarial machine learning regularized with a unified embedding. *arXiv preprint arXiv:1708.02582*, 2017.

[176] Vasileios Nakos, Xiaofei Shi, David P Woodruff, and Hongyang Zhang. Improved algorithms for adaptive compressed sensing. In *International Colloquium on Automata, Languages, and Programming*, pages 90:1–90:14, 2018.

[177] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[178] Behnam Neyshabur, Ruslan Salakhutdinov, and Nati Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2015.

[179] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

[180] Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung. Dual discriminator generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2670–2680, 2017.

[181] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

[182] Michael L Overton and Robert S Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992.

[183] Michal Parnas and Dana Ron. Testing metric properties. *Information and Computation*, 187(2):155–195, 2003.

[184] Bernardo Ávila Pires and Csaba Szepesvári. Multiclass classification calibration functions. *arXiv preprint arXiv:1609.06385*, 2016.

[185] 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.

[186] Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

[187] Eric Price and David P. Woodruff. (1+eps)-approximate sparse recovery. In *IEEE Symposium on Foundations of Computer Science*, pages 295–304, 2011.

[188] Eric Price and David P. Woodruff. Lower bounds for adaptive sparse recovery. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 652–663, 2013.

[189] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning

with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[190] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

[191] Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10899–10909, 2018.

[192] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox v0. 8.0: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

[193] Ilya Razenshteyn, Zhao Song, and David P. Woodruff. Weighted low rank approximations with provable guarantees. In *ACM Symposium on Theory of Computing*, pages 250–263, 2016.

[194] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[195] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[196] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *arXiv preprint arXiv:1711.09404*, 2017.

[197] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54(4):21, 2007.

[198] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[199] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *International Conference on Machine Learning*, 2017.

[200] Michael Saks and Xiaodong Sun. Space lower bounds for distance approximation in the data stream model. In *ACM Symposium on Theory of Computing*, pages 360–369, 2002.

[201] Ruslan Salakhutdinov and Hugo Larochelle. Efficient learning of deep Boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 693–700, 2010.

[202] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[203] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

[204] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv:1312.6120*, 2013.

[205] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander

Mądry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems 31*, pages 5019–5031, 2018.

[206] Rocco Anthony Servedio. *Efficient algorithms in computational learning theory*. Harvard University, 2001.

[207] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[208] Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of gradient-based deep learning. In *International Conference on Machine Learning*, 2017.

[209] Yuan Shen, Zaiwen Wen, and Yin Zhang. Augmented lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods and Software*, 29(2):239–263, 2014.

[210] Noam Shental, Amnon Amir, and Or Zuk. Rare-allele detection using compressed se(que)nsing. *CoRR*, abs/0909.0400, 2009.

[211] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *arXiv preprint arXiv:1707.04926*, 2017.

[212] Tasuku Soma and Yuichi Yoshida. Non-convex compressed sensing with the sum-of-squares method. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 570–579, 2016.

[213] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018.

[214] Zhao Song, David P Woodruff, and Peilin Zhong. Low rank approximation with entrywise ℓ_1-norm error. In *Proceedings of the 49th Annual Symposium on the Theory of Computing*, 2017.

[215] Zhao Song, David P Woodruff, and Peilin Zhong. Relative error tensor low rank approximation. In *ArXiv preprints*, 2017.

[216] Ross M Starr. Quasi-equilibria in markets with non-convex preferences. *Econometrica: Journal of the Econometric Society*, pages 25–38, 1969.

[217] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy? — a comprehensive study on the robustness of 18 deep image classification models. In *European Conference on Computer Vision*, 2018.

[218] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. In *IEEE International Symposium on Information Theory*, pages 2379–2383, 2016.

[219] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2): 853–884, 2017.

[220] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via nonconvex factorization.

In *IEEE Symposium on Foundations of Computer Science*, pages 270–289, 2015.

[221] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[222] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017.

[223] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[224] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[225] Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *ICML*, 2016.

[226] Madeleine Udell and Stephen Boyd. Bounding duality gap for separable problems with linear constraints. *Computational Optimization and Applications*, 64(2):355–378, 2016.

[227] Madeleine Udell, Corinne Horn, Reza Zadeh, Stephen Boyd, et al. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.

[228] Jonathan Uesato, Brendan O'Donoghue, Pushmeet Kohli, and Aaron van den Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pages 5025–5034, 2018.

[229] Leslie Valiant. Graph-theoretic arguments in low-level complexity. *Mathematical Foundations of Computer Science*, pages 162–176, 1977.

[230] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[231] Vladimir Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[232] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, pages 550–558, 2016.

[233] Elad Verbin and Wei Yu. The streaming complexity of cycle counting, sorting by reversals, and other problems. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 11–25, 2011.

[234] Roman Vershynin. Lectures in geometric functional analysis. pages 1–76, 2009.

[235] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint: 1011.3027*, 2010.

[236] Roman Vershynin. Estimation in high dimensions: A geometric perspective. In *Sampling theory, a renaissance*, pages 3–66. Springer, 2015.

[237] Ruohan Wang, Antoine Cully, Hyung Jin Chang, and Yiannis Demiris. MAGAN: Margin adaptation for generative adversarial networks. *arXiv preprint arXiv:1704.03817*, 2017.

[238] Yu-Xiang Wang and Huan Xu. Stability of matrix factorization for collaborative filtering.

In *International Conference on Machine Learning*, pages 417–424, 2012.

[239] Zaiwen Wen, Wotao Yin, and Yin Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.

[240] E Wong, F Schmidt, JH Metzen, and JZ Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, 2018.

[241] David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[242] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

[243] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018.

[244] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *International Conference on Computer Vision*, 2017.

[245] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5987–5995, 2017.

[246] Yichong Xu, Hongyang Zhang, Kyle Miller, Aarti Singh, and Artur Dubrawski. Noise-tolerant interactive learning using pairwise comparisons. In *Advances in Neural Information Processing Systems*, pages 2431–2440, 2017.

[247] 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.

[248] Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. *arXiv preprint arXiv:1810.11914*, 2018.

[249] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, pages 87.1–87.12, 2016.

[250] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2016.

[251] Hongyang Zhang, Zhouchen Lin, and Chao Zhang. A counterexample for the validity of using nuclear norm as a convex surrogate of rank. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 226–241, 2013.

[252] Hongyang Zhang, Zhouchen Lin, Chao Zhang, and Junbin Gao. Robust latent low rank representation for subspace clustering. *Neurocomputing*, 145:369–373, 2014.

[253] Hongyang Zhang, Zhouchen Lin, Chao Zhang, and Edward Y Chang. Exact recoverability of robust PCA via outlier pursuit with tight recovery bounds. In *AAAI Conference on*

Artificial Intelligence, pages 3143–3149, 2015.

[254] Hongyang Zhang, Zhouchen Lin, Chao Zhang, and Junbin Gao. Relations among some low rank subspace recovery models. *Neural Computation*, 27:1915–1950, 2015.

[255] Hongyang Zhang, Zhouchen Lin, and Chao Zhang. Completing low-rank matrices with corrupted samples from few coefficients in general basis. *IEEE Transactions on Information Theory*, 62(8):4748–4768, 2016.

[256] Hongyang Zhang, Shan You, Zhouchen Lin, and Chao Xu. Fast compressive phase retrieval under bounded noise. In *AAAI Conference on Artificial Intelligence*, 2017.

[257] Hongyang Zhang, Susu Xu, Jiantao Jiao, Pengtao Xie, Ruslan Salakhutdinov, and Eric P Xing. Stackelberg gan: Towards provable minimax equilibrium via multi-generator architectures. *arXiv preprint arXiv:1811.08010*, 2018.

[258] Hongyang Zhang, Junru Shao, and Ruslan Salakhutdinov. Deep neural networks with multi-branch architectures are intrinsically less non-convex. In *International Conference on Artificial Intelligence and Statistics*, pages 1099–1109, 2019.

[259] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

[260] Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh. The limitations of adversarial training and the blind-spot attack. In *International Conference on Learning Representations*, 2019.

[261] Tong Zhang. Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research*, 2:527–550, 2002.

[262] Xiao Zhang, Lingxiao Wang, and Quanquan Gu. A nonconvex free lunch for low-rank plus sparse matrix recovery. *arXiv preprint arXiv:1702.06525*, 2017.

[263] Yuchen Zhang, Jason Lee, Martin Wainwright, and Michael Jordan. On the learnability of fully-connected neural networks. In *Artificial Intelligence and Statistics*, pages 83–91, 2017.

[264] Yuchen Zhang, Percy Liang, and Martin J Wainwright. Convexified convolutional neural networks. In *International Conference on Machine Learning*, 2017.

[265] Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

[266] Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[267] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4480–4488, 2016.