Algorithms for Matrix Approximation: Sketching, Sampling, and Sparse Optimization

Taisuke Yasuda

CMU-CS-24-110

May 2024

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

Thesis Committee:
David P. Woodruff, Chair
Anupam Gupta
Richard Peng
Cameron Musco (University of Massachusetts Amherst)

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

This research was sponsored by the Simons Foundation under award number 689863 and the Office of Naval Research under award number N000141812562. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Matrix approximation, sketching, sampling, sparse optimization

To my family

Abstract

The approximation of matrices by smaller, simpler, or structured matrices is a fundamental problem in various fields of mathematics and computer science including numerical linear algebra, graph algorithms, computational geometry, signal processing, statistics, machine learning, and optimization. Recently, matrix approximation has been particularly important in modern computing as a key technique for efficiently processing enormous datasets in running time and memory scaling linearly, or even sublinearly, in the size of the dataset. In this thesis, we develop new and improved algorithms for a wide variety of matrix approximation tasks, drawing particularly heavily from *sketching* and *sampling* techniques from randomized numerical linear algebra, as well as *sparse optimization* techniques. We also utilize and develop connections of these problems with the literature of *geometric functional analysis*.

We develop and improve foundational tools for matrix approximation, and find novel applications of these building blocks to solve central questions in matrix approximation. Some of the basic tools that we develop and sharpen include nearly optimal constructions of oblivious and non-oblivious subspace embeddings, improved low rank approximation algorithms, and new properties of ℓ_1 regularization. Using our improved understanding of these primitives, we obtain a suite of applications such as the first polynomial space algorithms for high-dimensional computational geometry, nearly optimal algorithms for active linear regression, and the first nearly optimal coresets for multiple regression and subspace approximation. Many of our results have implications in big data computing settings, such as streaming, online, and distributed computation.

Acknowledgements

I would first like to thank my advisor David Woodruff, who inspired me to pursue a doctorate in theoretical computer science and guided me to be the researcher I am today. I was an undergraduate at CMU when David's lectures on Algorithms for Big Data first captivated my fascination. Later, his generosity to patiently mentor me despite countless failed ideas and months of stagnancy gave me the experiences I needed to commit to a research career. It was a no-brainer to come back to work with David again at CMU for my PhD, and our collaborations since have been a constant stream of wonderful and exciting discoveries. I will always remember my time solving problems with David as some of my best memories.

I would next like to thank my thesis committee, Anupam Gupta, Richard Peng, and Cameron Musco. Every one of my committee members has been a huge influence throughout my work during my PhD, and they each have works that I have completely obsessed over for a period of time. Anupam's work with Sanjoy Dasgupta on the Johnson–Lindenstrauss lemma gives a proof that was simple enough for me to understand as an inexperienced undergraduate, and marks one of my earliest inspirations in the area of sketching. Richard's work with Michael Cohen on Lewis weight sampling was a paper that I studied many times over when I first started working on sampling, and caused my jaw to drop when it introduced me to chaining techniques. Cameron has many papers that I have studied intensely, but a particularly mind-blowing one is his work on ridge leverage score sampling with Michael Cohen and Chris Musco, which I also remember as a paper that made me feel electrified.

I thank my collaborators and friends at Google Research for an incredible two years spent on machine learning research. I am especially grateful to my hosts Kyriakos Axiotis, Thomas Fu, and Vahab Mirrokni. Thomas patiently taught me everything I currently know on conducting empirical machine learning research. Kyriakos, an expert in sparse optimization, has been a great inspiration to me in a theoretical field separate from my work at CMU. Matthew Fahrbach and Rajesh Jayaram have been great mentors at Google who gave me moral support and friendship when I faced my peak frustration at not being able to make progress. I have also had great conversations with my other Google collaborators, MohammadHossein Bateni and Lin Chen, as well as Renato Paes Leme, Jon Schneider, Balu Sivan, Manfred Warmuth, and Shuran Zheng.

Next, I would like to thank my academic collaborators Gregory Dexter, Petros Drineas, Cameron Musco, Chris Musco, and Yi Li, as well as my research friends Manuel Fernandez, Mehrdad Ghadiri, Praneeth Kacham, Yang Liu, Naren Manoj, Raphael Meyer, Swati Padmanabhan, Eliot Robson, Kshiteej Sheth, Fred Zhang, and Samson Zhou. Through our conversations, you have all given me a sense of belonging to a research community that I will always cherish.

I thank Ian Tice, who mentored me throughout my undergraduate years at CMU and raised me from a freshman just learning my way around basic mathematics all the way to a master's degree in PDEs. Although I didn't end up in the same field of research as Ian, I owe a great deal of my mathematical and research abilities and tastes to his mentorship.

Lastly, I would like to thank my family, my in-law family, and my wife Joyce. You have all been my rock throughout some of my hardest years.

Contents

List of Figures xiii

1 Introduction 1
- 1.1 Randomized numerical linear algebra . . . 1
  - 1.1.1 Sketching . . . 2
  - 1.1.2 Subspace embeddings . . . 2
- 1.2 Oblivious sketches . . . 3
  - 1.2.1 Why are oblivious subspace embeddings useful? . . . 4
  - 1.2.2 Oblivious ℓ_2 subspace embeddings . . . 4
  - 1.2.3 Overview of Part I . . . 5
- 1.3 Sampling . . . 6
  - 1.3.1 Coresets and sensitivity sampling . . . 6
  - 1.3.2 Leverage score sampling . . . 9
  - 1.3.3 Streaming and online coresets . . . 10
  - 1.3.4 Applications of sampling algorithms beyond coresets . . . 12
  - 1.3.5 Overview of Part II . . . 13
- 1.4 Sparse optimization . . . 13
  - 1.4.1 Sparse linear regression . . . 13
  - 1.4.2 Column subset selection . . . 14
  - 1.4.3 Overview of Part III . . . 15
- 1.5 Connections to geometric functional analysis . . . 15
  - 1.5.1 Lewis weights and embedding subspaces of ℓ_p . . . 15
  - 1.5.2 Well-conditioned bases and spanning sets . . . 16

2 Preliminaries 19
- 2.1 Notation . . . 19
  - 2.1.1 Linear algebra . . . 19
  - 2.1.2 Inequalities . . . 19
  - 2.1.3 Probability . . . 20
- 2.2 Streaming . . . 20
  - 2.2.1 INDEX . . . 20
- 2.3 Random processes . . . 20
  - 2.3.1 Symmetrization: reduction to a Rademacher process . . . 21
  - 2.3.2 Subgaussian processes . . . 22

2.3.3 Chaining and Dudley's inequality . . . 22

I Oblivious Sketching **27**

3 High distortion embeddings for ℓ_p [WY23a] **29**
3.1 The question of well-conditioned bases . . . 31
3.2 Relaxing linear bases to spanning sets . . . 31
3.3 Proof of Theorem 3.0.4 . . . 32

4 Low distortion embeddings for ℓ_1 [LWY21] **35**
4.1 Overview of sketch construction and analysis . . . 35
4.1.1 Sketching a single vector . . . 36
4.1.2 Extension to subspaces . . . 37
4.2 No expansion . . . 38
4.2.1 Bounding badly concentrated levels . . . 38
4.2.2 Bounding well-concentrated levels . . . 39
4.2.3 Bounding oversampled levels . . . 40
4.2.4 Bounding tiny levels . . . 41
4.2.5 Net argument . . . 41
4.3 No contraction . . . 43
4.3.1 Essential weight classes . . . 43
4.3.2 Hashing lemmas . . . 44
4.3.3 Preserving weight classes . . . 46
4.3.4 Net argument . . . 48
4.4 Endgame . . . 48

5 Future directions for oblivious ℓ_p subspace embeddings **51**

II Sampling Algorithms and Coresets **53**

6 ℓ_p Lewis weight sampling [WY23b] **55**
6.1 Sampling algorithms for ℓ_p subspace embeddings . . . 55
6.1.1 ℓ_p sensitivity sampling . . . 56
6.1.2 ℓ_p well-conditioned basis sampling . . . 56
6.1.3 ℓ_p Lewis weight sampling . . . 57
6.2 Properties of one-sided ℓ_p Lewis weights . . . 60
6.3 Analysis of ℓ_p Lewis weight sampling: reduction to a Rademacher process . . . 62
6.3.1 Regularizing the Rademacher process . . . 62
6.3.2 Flattening the Rademacher process: $p < 2$ . . . 64
6.3.3 Flattening the Rademacher process: $p > 2$ [WY23b] . . . 65
6.4 Analysis of ℓ_p Lewis weight sampling: Dudley's entropy integral . . . 67
6.4.1 Bounds on the pseudo-metric . . . 67

6.4.2 Entropy bounds . . . 68
6.4.3 Entropy integral for ℓ_p Lewis weight sampling . . . 72
6.5 Analysis of ℓ_p Lewis weight sampling: endgame . . . 76
6.6 Online ℓ_p Lewis weight sampling . . . 77
6.6.1 Lemmas from linear algebra . . . 77
6.6.2 Properties of online ℓ_p Lewis weights . . . 78
6.6.3 The sum of online ℓ_p Lewis weights . . . 80

7 ℓ_p sensitivity sampling [WY23c] 81
7.1 Beyond ℓ_p Lewis weight sampling . . . 81
7.2 Structured matrices with small total sensitivity, $p > 2$ . . . 82
7.3 Properties of ℓ_p sensitivities . . . 85
7.3.1 Monotonicity of max ℓ_p sensitivity . . . 85
7.3.2 Flattening ℓ_p sensitivities . . . 86
7.3.3 Total sensitivity . . . 86
7.4 Analysis of ℓ_p sensitivity sampling . . . 88
7.4.1 Dudley's entropy integral . . . 88
7.4.2 Sensitivity sampling, $p < 2$ . . . 89
7.4.3 Sensitivity sampling, $p > 2$ . . . 93

8 Root leverage score sampling [WY23c, WY24b] 99
8.1 Analysis of root leverage score sampling . . . 101
8.1.1 Reduction to a small number of scales . . . 102
8.1.2 Reduction to a Rademacher process with flat sensitivities . . . 103
8.1.3 Bounds on the Rademacher process . . . 106
8.1.4 Proof of main sampling theorems . . . 107

9 High-distortion ℓ_p subspace embeddings [WY22a] 109
9.1 Lewis weight switching . . . 110
9.2 Change of density . . . 111

10 Subspace embeddings for general losses [MMWY22] 115
10.1 M-estimators preliminaries . . . 117
10.2 Sensitivities upper bounds . . . 118
10.2.1 Efficient algorithm for sensitivity upper bounds . . . 120
10.2.2 Sharper sensitivity bounds . . . 123
10.3 Sensitivity lower bounds . . . 124

11 Applications: streaming ℓ_∞ subspace embeddings and computational geometry [WY22a] 127
11.1 Nearly optimal sum of online leverage scores . . . 128
11.2 Online coresets for ℓ_∞ subspace embeddings . . . 130
11.3 Near-optimal bounds for restricted instances . . . 132
11.3.1 Lower bound . . . 132

11.3.2 Upper bound . . . 133
11.4 Applications to streaming algorithms for geometric problems in high dimensions 134
11.4.1 Directional width . . . 135
11.4.2 Convex hulls . . . 137
11.4.3 Löwner–John ellipsoids . . . 138
11.4.4 Volume maximization . . . 141
11.4.5 Minimum-width spherical shell . . . 143

12 Applications: active ℓ_p linear regression [MMWY22, WY23a] 145
12.1 Active ℓ_p linear regression . . . 145
12.2 Constant factor solution . . . 147
12.2.1 Probability boosting for constant factor approximation . . . 148
12.3 $(1+\varepsilon)$ factor solution . . . 149
12.3.1 Closeness of nearly optimal solutions . . . 150
12.3.2 Iterative size reduction argument . . . 151
12.3.3 High probability . . . 154
12.4 ℓ_p Lewis weight sampling for differences . . . 155
12.5 Rademacher process bounds . . . 158
12.5.1 Estimates on the outlier term . . . 159
12.5.2 Estimates on the sensitivity term . . . 159
12.6 Lower bounds . . . 165
12.6.1 Lower bounds for $p \in (0,1)$ . . . 166
12.6.2 Lower bounds for $p \in (1,2)$ . . . 168
12.6.3 Lower bounds for $p \in (2,\infty)$ . . . 170
12.6.4 A $1/\delta^{p-1}$ lower bound for sampling-and-reweighting algorithms . . . 171

13 Applications: coresets for multiple ℓ_p regression [WY24a] 173
13.1 Multiple ℓ_p regression . . . 173
13.1.1 Coreset constructions for $p = 2$ . . . 173
13.1.2 Challenges for $p \neq 2$ . . . 174
13.1.3 Strong coresets for multiple ℓ_p regression . . . 175
13.1.4 Weak coresets for multiple ℓ_p regression . . . 176
13.1.5 Applications: sublinear algorithms for Euclidean power means . . . 177
13.1.6 Applications: spanning coresets for ℓ_p subspace approximation . . . 178
13.2 Strong coresets . . . 179
13.3 Weak coresets . . . 181
13.3.1 Closeness of nearly optimal solutions . . . 181
13.3.2 Iterative size reduction argument . . . 182
13.4 Sublinear algorithm for Euclidean power means . . . 185
13.5 Spanning coresets for ℓ_p subspace approximation . . . 186
13.6 Lower bounds . . . 187
13.6.1 Strong coresets . . . 187
13.6.2 Weak coresets . . . 189
13.6.3 Spanning coresets . . . 190

14 Applications: strong coresets for ℓ_p subspace approximation [WY23a, WY24b] 193
14.1 Coresets for ℓ_p subspace approximation . . . 194
14.1.1 Technical overview . . . 198
14.1.2 Corollaries . . . 203
14.2 Representative subspace theorem for ℓ_p subspace approximation . . . 204
14.2.1 Sharper scalar inequalities . . . 205
14.2.2 Proof of the representative subspace theorem . . . 206
14.3 Preliminaries . . . 208
14.3.1 Dvoretzky's theorem . . . 208
14.3.2 Flattening . . . 209
14.3.3 Properties of ridge leverage scores . . . 210
14.4 Reduction to additive-multiplicative ℓ_p affine embeddings . . . 211
14.5 Main sampling theorems . . . 213
14.5.1 Affine embedding . . . 213
14.5.2 Results for $p > 2$ . . . 215
14.5.3 Results for $p < 2$ . . . 218
14.6 Streaming and online coresets . . . 220
14.6.1 Online coresets . . . 221
14.6.2 Streaming coresets . . . 222

15 Future directions for sampling and coreset algorithms 223
15.1 Questions on ℓ_p subspace embeddings . . . 223
15.2 Questions on coresets . . . 224

III Sparse Optimization 227

16 Sparse convex optimization via ℓ_1 regularization [YBC+23, AY23] 229
16.1 Introduction . . . 229
16.1.1 Related work: prior guarantees for ℓ_1 regularization . . . 230
16.1.2 Our results . . . 231
16.1.3 Related work: the Forward Stagewise Regression conjecture . . . 238
16.1.4 Related work: algorithms for sparse convex optimization . . . 238
16.1.5 Open directions . . . 238
16.2 Preliminaries . . . 239
16.2.1 Fenchel duality . . . 239
16.2.2 Berge's theorem . . . 240
16.3 Equivalence of Group Sequential LASSO and Group Orthogonal Matching Pursuit 240
16.3.1 The dual problem . . . 240
16.3.2 Selection of features . . . 242
16.4 Guarantees for Group Orthogonal Matching Pursuit . . . 243
16.4.1 Group OMP with Replacement . . . 246
16.5 Equivalence of Group Sequential Attention and Group Sequential LASSO . . . 248
16.6 Experiments: feature selection via Sequential Attention . . . 249

16.6.1 Small-scale experiments . . . 249
16.6.2 Large-scale experiments . . . 250
16.6.3 Visualization of selected MNIST features . . . 250

17 Column subset selection with entrywise losses [WY23a] 253
17.1 Algorithms for general entrywise losses . . . 254
17.1.1 An improved structural result on uniform sampling . . . 256
17.1.2 Sharper guarantees for the [SWZ19] algorithm . . . 257
17.2 Huber column subset selection . . . 260
17.3 Algorithms for the entrywise ℓ_p norm . . . 262
17.3.1 Improved existential result . . . 262
17.3.2 Lower bounds . . . 264
17.4 Reduction from existential to algorithmic column subset selection . . . 265

18 Spectral low rank approximation for sparse singular vectors [WY22b] 269
18.1 Technical overview . . . 270
18.2 Proof of Theorem 18.0.3 . . . 272
18.2.1 Approximating singular components . . . 272
18.2.2 Finding the support of singular vectors with large singular value . . . 274
18.2.3 Approximating large singular values . . . 275
18.2.4 Approximating small singular values . . . 279

19 Future directions for sparse optimization 281

Bibliography 283

List of Figures

16.1 Feature selection results for small-scale neural network experiments. Here, SA = Sequential Attention, LLY = [LLY21], GL = Group LASSO, SL = Sequential LASSO, OMP = OMP, and CAE = Concrete Autoencoder [BAZ19]. 250
16.2 AUC and log loss when selecting $k \in \{10, 15, 20, 25, 30, 35\}$ features for Criteo dataset. 251
16.3 Visualizations of the $k = 50$ pixels selected by the feature selection algorithms on MNIST. 251

Chapter 1

Introduction

Matrices are one of the most fundamental forms of representing data. As data-driven technologies proliferate throughout modern computer science, large matrices that represent enormous datasets have become central objects of study, and designing approximation algorithms for efficiently handling these matrices and datasets has become one of the most important computational challenges today. A natural and highly effective idea for computing with such large matrices is to first approximate them by smaller or more structured matrices, so that downstream algorithms enjoy a more well-behaved instance. Popular approximations that have been studied include low dimensional embeddings, low rank approximations, approximations by a small subset of rows or columns (subset selection and feature selection), projections onto a collection of points or subspaces (clustering and projective clustering), sparse linear combinations (sparse dictionary learning), or in general, any other structured object with efficient representation. We generally refer to such problems as *matrix approximation*, and algorithms and lower bounds for matrix approximation tasks are the focus of this thesis.

1.1 Randomized numerical linear algebra

The rise of the field of *randomized numerical linear algebra* [Mah11, Woo14, MT20] in the past two decades has been particularly fruitful for the development of algorithmic results for matrix approximation, and many results of this thesis are best placed in the context of this literature. Traditionally, the study of computational and numerical aspects of matrices and linear algebra, or *numerical linear algebra* [TB97, GVL13], focused on designing deterministic algorithms for manipulating matrices at machine precision, with input instances that are small enough to fit in memory. Thus, in this regime, algorithms are generally assumed to have access to the entire input instance, and running time and space complexity scaling polynomially in the input dimensions are acceptable. In contrast, randomized numerical linear algebra places an emphasis on massive input instances that arise in big data settings, where the datasets are so large that only small parts of the input can be accessed at a time, and algorithms must run in at most linear or even sublinear time to be considered practical. To handle such inputs, we allow for *randomized* and *approximate* algorithms, that is, the algorithm is allowed to fail with some small probability δ (say 1% probability), and is considered successful if it outputs a solution that is correct up to some

small tolerance parameter ε (say 1% error in some appropriate sense).

1.1.1 Sketching

Linear sketching is a fundamental technique of randomized numerical linear algebra, where input matrices are compressed by multiplication with a random matrix. More concretely, sketching algorithms roughly follow the following framework. Given as input an $n \times d$ matrix $\mathbf{A}$, first apply a random $r \times n$ matrix $\mathbf{S}$ for some $r \ll n$ to obtain a compressed $r \times d$ input $\mathbf{SA}$. Then, perform some computations on the compressed instance $\mathbf{SA}$, and use the result to output a solution for the original instance $\mathbf{A}$. By drawing the random sketching matrix $\mathbf{S}$ from a carefully chosen distribution, this framework yields the fastest known algorithms for a wide range of linear algebraic tasks including linear regression, low rank approximation, matrix multiplication, and trace estimation [DMMW12, CW13, CEM+15, CNW16, MMMW21, CSWZ23, CDDR23] as well as further applications to clustering [CEM+15, MMR19], sparse dictionary learning [DDWY23], subspace approximation [FMSW10, CW15a], minimum volume enclosing ellipsoids [CCLY19, WY22a], graph algorithms [CKL22, AKY23], tensor decompositions [MWZ24], and far beyond. Sketching algorithms often (but not always) take $\mathbf{S}$ to be drawn *independently* of the input $\mathbf{A}$, which facilitates its application in settings when $\mathbf{A}$ may change, for example if $\mathbf{A}$ undergoes additive updates in a stream or is multiplied on the right by some projection. These sketches are known as *oblivious sketches*, and our results for oblivious sketching algorithms for matrix approximation is the subject of Part I of this thesis.

1.1.2 Subspace embeddings

The study of sketching algorithms and their applications has generated a number of foundational definitions for formalizing useful guarantees in matrix approximation. One such definition is that of a *subspace embedding*, which we describe here to provide a more detailed illustration of the sketching paradigm.

A subspace embedding is a notion of matrix approximation which considers an approximation $\mathbf{A}'$ to be close to a matrix $\mathbf{A}$ if the norms of vectors in $\mathrm{colspan}(\mathbf{A}') = \{\mathbf{A}'\mathbf{x} : \mathbf{x} \in \mathbb{R}^d\}$ are close to those of $\mathrm{colspan}(\mathbf{A}) = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{R}^d\}$. While $\mathbf{A}'$ could, in principle, be constructed in any way, we will always focus on constructions of the form $\mathbf{A}' = \mathbf{SA}$ for a sketching matrix $\mathbf{S}$ in this thesis.

Definition 1.1.1 (Subspace embedding)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{S} \in \mathbb{R}^{r\times n}$. Let $\kappa \geq 1$ be a distortion parameter and let $\|\cdot\|$ be a norm. Then, $\mathbf{S}$ is a *κ-approximate subspace embedding* if

$$\text{for every } \mathbf{x} \in \mathbb{R}^d, \qquad \|\mathbf{Ax}\| \leq \|\mathbf{SAx}\| \leq \kappa\|\mathbf{Ax}\|.$$

One of the most ideal settings for the application of subspace embeddings is the design of efficient approximation algorithms for the overdetermined least squares linear regression problem [DMM06a, Sar06]. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a tall ($n \gg d$) design matrix and let $\mathbf{b} \in \mathbb{R}^n$ be a label vector, and suppose that we want to efficiently approximate

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{Ax} - \mathbf{b}\|_2^2,$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm given by $\|\mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{n} \mathbf{y}(i)^2}$ for an n-dimensional vector $\mathbf{y}$. Classically, this problem requires $O(nd^2)$ time to solve exactly by using Gaussian elimination.[1] However, if n and d are large, then this running time is much larger than the size of the input which is nd, and thus may be prohibitive. We will now describe a way to design much faster algorithms by using efficient constructions of subspace embeddings. Suppose that we have an algorithm for efficiently computing a κ-approximate subspace embedding $\mathbf{S} \in \mathbb{R}^{r \times n}$ in the ℓ_2 norm for the $n \times (d+1)$ matrix $[\mathbf{A}\ \mathbf{b}]$, that is, $\mathbf{A}$ together with $\mathbf{b}$ appended as an additional column. Note then that,

$$\text{for every } \mathbf{x} \in \mathbb{R}^d, \qquad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \leq \|\mathbf{S}\mathbf{A}\mathbf{x} - \mathbf{S}\mathbf{b}\|_2^2 \leq \kappa^2 \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2, \tag{1.1}$$

since $\mathbf{A}\mathbf{x} - \mathbf{b}$ is in the column span of $[\mathbf{A}\ \mathbf{b}]$. Now suppose we set

$$\hat{\mathbf{x}} \coloneqq \arg\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{S}\mathbf{A}\mathbf{x} - \mathbf{S}\mathbf{b}\|_2^2$$
$$\mathbf{x}^* \coloneqq \arg\min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

Then, $\hat{\mathbf{x}}$ is a κ^2-approximately optimal solution since

$$\begin{aligned}
\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_2^2 &\leq \|\mathbf{S}\mathbf{A}\hat{\mathbf{x}} - \mathbf{S}\mathbf{b}\|_2^2 && \text{(1.1)} \\
&\leq \|\mathbf{S}\mathbf{A}\mathbf{x}^* - \mathbf{S}\mathbf{b}\|_2^2 && \text{optimality of } \hat{\mathbf{x}} \\
&\leq \kappa^2 \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2 = \kappa^2 \min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 && \text{(1.1)}
\end{aligned}$$

and furthermore, it can be computed in the time that it takes to compute $\mathbf{SA}$ and $\mathbf{Sb}$, plus only $O(rd^2)$ time. This can potentially be much faster than the original $O(nd^2)$ time, if $r \ll n$ and the computation of $\mathbf{SA}$ and $\mathbf{Sb}$ is fast. Indeed, this framework has been applied to develop some of the fastest known algorithms for least squares linear regression, as well as a variety of other related linear algebraic tasks [DMMW12, CW13, CCKW22, CSWZ23, CDDR23].

Throughout this thesis, we will place great emphasis on developing subspace embeddings for norms $\|\cdot\|$ beyond the ℓ_2 norm, as well as new applications of related ideas to solve problems such as streaming computational geometry, active regression, multiple regression, subspace approximation, and column subset selection.

1.2 Oblivious sketches

We begin our discussion of algorithms for subspace embeddings (Definition 1.1.1) by considering a particularly useful restricted class of subspace embeddings known as *oblivious* subspace embeddings. As the name suggests, the construction of oblivious subspace embeddings $\mathbf{S}$ are oblivious to the input matrix $\mathbf{A}$, that is, they are constructed independently of $\mathbf{A}$. More formally, we have the following definition:

[1] By using fast matrix multiplication, this running time can be improved to $O(nd^{\omega-1})$ for $\omega \approx 2.372$ [DWZ23, WXXZ23], but such algorithms are rarely used in practice.

Definition 1.2.1 (Oblivious subspace embedding)**.** Let $0 < \delta < 1$. Let $\mathcal{D}$ be a distribution over $r \times n$ matrices $\mathbf{S}$. Let $\kappa \geq 1$ be a distortion parameter and let $\|\cdot\|$ be a norm. Then, $\mathcal{D}$ is a *κ-approximate oblivious subspace embedding with probability* $1 - \delta$ if for every $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{S} \sim \mathcal{D}$ is a κ-approximate subspace embedding (Definition 1.1.1) for $\mathbf{A}$ with probability at least $1 - \delta$.

Remark 1.2.2. Note that in order to construct nontrivial oblivious subspace embeddings embeddings, i.e. constructions with $r < n$ and $\kappa < \infty$, then randomized constructions are necessary. Indeed, otherwise, for any fixed $\mathbf{S} \in \mathbb{R}^{r \times n}$ with $r < n$, there exists some nonzero $\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{S}\mathbf{y} = 0$.

1.2.1 Why are oblivious subspace embeddings useful?

Oblivious subspace embeddings are particularly useful in settings where we need to efficiently update the sketch $\mathbf{SA}$, for example in streaming algorithms or distributed computing. For instance, consider the turnstile streaming model, in which our input matrix $\mathbf{A}$ undergoes entrywise additive updates of the form $\mathbf{A}_{i,j} \leftarrow \mathbf{A}_{i,j} + \Delta$ for $\Delta \in \mathbb{R}$. Then, the sketch $\mathbf{SA}$ can be updated efficiently under this model by using the linearity of the sketch, and furthermore, $\mathbf{S}$ is still a subspace embedding for the updated matrix with high probability due to the oblivious property. This simple yet powerful observation has been extremely useful in the design of streaming algorithms for many problems in numerical linear algebra and machine learning, including ℓ_p linear regression [SW11, MM13, WZ13, WW19, LWY21, WY23a], ridge regression [KW22], low rank approximation [CW09, BWZ16], robust regression [CW15b], and logistic regression [MOW21, MOW23]. Similar benefits apply in distributed models of computation, where the input matrix $\mathbf{A}$ is represented as a sum of matrices, each of which is stored in a separate server.

1.2.2 Oblivious ℓ_2 subspace embeddings

Consider the problem of computing an oblivious subspace embedding for the ℓ_2 norm for $d = 1$, which simply corresponds to the problem of finding a norm-preserving linear map for a single vector. This natural problem is resolved by a classical result due to Johnson and Lindenstrauss [JL84], which states that given a set $S \subseteq \mathbb{R}^n$ of m vectors in n dimensions, a random linear projection $\mathbf{S} \in \mathbb{R}^{r \times n}$ from n dimensions to $r = O(\varepsilon^{-2} \log m)$ dimensions has the property that

$$\|\mathbf{S}\mathbf{y}\|_2 = (1 \pm \varepsilon)\|\mathbf{y}\|_2$$

simultaneously for every $\mathbf{y} \in S$, with probability at least $2/3$. Thus, the ℓ_2 norm of a finite number of vectors can be preserved up to $(1 \pm \varepsilon)$ factors. Furthermore, note that the matrix $\mathbf{S}$ is an independent random linear map, so it satisfies the obliviousness property of Definition 1.2.1. Thus, for $d = 1$, oblivious subspace embeddings for the exist for the ℓ_2 norm with dimension $r = O(\varepsilon^{-2})$ and distortion $\kappa = (1 + \varepsilon)$.

It may not be immediately clear that the technique of random projections also solves the problem of computing a subspace embedding for $d > 1$, since this involves preserving the ℓ_2 norm of every vector in the column space of $\mathbf{A}$, which is an *uncountably infinite* number of vectors,

rather than a finite number m. Nevertheless, the following seminal result of Sarlos [Sar06] shows that random projections in fact do yield ℓ_2 subspace embeddings with distortion $\kappa = (1+\varepsilon)$.

Theorem 1.2.3 (Sarlos [Sar06])**.** Let $\mathbf{S}$ be an $r \times n$ matrix of i.i.d. Gaussian random variables. There is an $r = O(\varepsilon^{-2} d \log d)$ such that for any $\mathbf{A} \in \mathbb{R}^{n \times d}$,

$$\mathbf{Pr}\big\{\text{for all } \mathbf{x} \in \mathbb{R}^d, \quad \|\mathbf{A}\mathbf{x}\|_2 \le \|\mathbf{S}\mathbf{A}\mathbf{x}\|_2 \le (1+\varepsilon)\|\mathbf{A}\mathbf{x}\|_2\big\} \ge \frac{99}{100}$$

that is, $\mathbf{S}$ is an ℓ_2 subspace embedding of $\mathbf{A}$ with distortion $(1+\varepsilon)$, with probability at least $99/100$.

It is known that the bound on the embedding dimension r in Theorem 1.2.3 is nearly optimal for oblivious subspace embeddings [NN14]. Since the result of Theorem 1.2.3, a long line of work has studied further improvements to oblivious ℓ_2 subspace embeddings, yielding smaller embedding dimensions, faster algorithms, and simpler proofs [Sar06, CW13, NN13, Coh16, CCKW22, CSWZ23, CDDR23]. Similar techniques have been also been applied to embedding objects beyond subspaces, including sparse vectors, collections of subspaces, and manifolds [BDN15].

1.2.3 Overview of Part I

With the construction of oblivious ℓ_2 norms established, a natural question is whether similar results can be obtained for ℓ_p *norms* rather than ℓ_2 norms, where the ℓ_p norm of a vector $\mathbf{y} \in \mathbb{R}^n$ for $1 \le p < \infty$ is defined by

$$\|\mathbf{y}\|_p := \left(\sum_{i=1}^{n} |\mathbf{y}(i)|^p\right)^{1/p}.$$

For example, for $p = 1$, the ℓ_1 norm corresponds to the use of the *absolute deviations* loss in the context of linear regression, and captures the average loss rather than the average squared loss as in the ℓ_2 norm. This gives a more *robust* loss function that is less sensitive to outliers, and is more appropriate in settings where the data may be more prone to corruption by noise. On the other hand, if $p = \infty$, the ℓ_∞ norm measures the maximum error in the context of linear regression, and is more appropriate when the worst-case error must be minimized. In general, ℓ_p norms allow us to smoothly interpolate between these two cases, with $p < 2$ giving more robust losses and $p > 2$ giving more worst-case losses.

Unfortunately, it is in fact *impossible* to construct oblivious ℓ_p subspace embeddings that match the guarantees of Theorem 1.2.3 in general for $p \neq 2$. As we discuss further in Chapter 3, there are lower bounds that prohibit the sketching dimension r from being subpolynomial in n for $p > 2$, and subexponential in d for $p < 2$, if we insist on a distortion of $\kappa = (1+\varepsilon)$. For $p < 2$, however, we can at least salvage a useful result if d is relatively small compared to n. In Chapter 3, we will show how to obtain the best possible oblivious ℓ_p subspace embedding when we allow for κ to be as large as $\kappa = \mathrm{poly}(d)$, where we achieve a trade-off of $\kappa = \tilde{O}(d^{1/p})$ distortion with $r = \tilde{O}(d)$ sketching dimension, based on work of [WY23a]. In Chapter 4, we show new results when we insist on a distortion of $\kappa = (1+\varepsilon)$, showing an upper bound of $r = \exp(\tilde{O}(d/\varepsilon))$ based on work of [LWY21], which exponentially improves upon a prior bound of $r = \exp\exp(O(d))$.

Open questions arising from work in this part are collected in Chapter 5.

1.3 Sampling

Another particularly useful special case of sketching techniques is *sampling*, in which the sketching matrix $\mathbf{S}$ has at most one nonzero entry in each row. Sampling methods are appealing since they can be interpreted as subset selection (also known as *coresets*), which is an important and well-studied problem in its own right. Another advantage of sampling is that it often preserves useful properties of the input such as sparsity or tensor product structure, which can be important for efficient algorithms when consuming the resulting sketch $\mathbf{SA}$. Furthermore, despite the restriction to a simpler class of sketching matrices $\mathbf{S}$, sampling-based sketches, in many cases, in fact achieve the best known upper bounds in a variety of sketching problems. Sampling-based matrix approximation algorithms is a major focus of this thesis, and our results in this setting are found in Part II of this thesis.

Unlike matrix approximations discussed in Part I based on oblivious sketching, sampling-based methods will generally be constructed as a function of the input matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, that is, they are *non-oblivious*. The use of additional information about the input matrix $\mathbf{A}$ has both drawbacks and advantages. On one hand, the requirement of knowing $\mathbf{A}$ beforehand prevents the use of these techniques in various information-limited settings, such as turnstile streaming (see Section 1.2.1). However, this additional information will, in many cases, lead to dramatically improved guarantees such as smaller approximation errors, greater compression of the matrix size, or both. Furthermore, when the input matrix $\mathbf{A}$ is structured, sampling methods often produce matrix approximations that preserve these structural properties, for example if the rows of the matrix are sparse or carry tensor product structure. These characteristics make sampling methods a highly attractive class of matrix approximation algorithms to study.

In this section, we introduce the notion of *coresets* and its relationship to sampling algorithms in Section 1.3.1, the technique of *leverage score sampling* for sampling ℓ_2 subspace embeddings in Section 1.3.2, and the *online coreset model* in Section 1.3.3 which extends the problem of constructing coresets to the setting of online algorithms. All of these concepts will be recurring fundamental themes throughout Part II of this thesis. Finally, we give an overview of the rest of Part II in Section 1.3.5.

1.3.1 Coresets and sensitivity sampling

Sampling-based matrix approximations are intimately related to the notion of *coresets*, which broadly refer to the paradigm of solving computational problems on large datasets by first approximating the dataset by a small reweighted subset of the dataset. The "computational problem on datasets" that have been studied in this context range from statistical inference tasks such as mean and median estimation, linear regression, and logistic regression, computational geometric tasks such as low rank approximation, clustering, convex hull estimation, and ellipsoidal rounding, machine learning tasks such as training deep neural networks, and far beyond. In all of these problems, datasets are typically represented as matrices, and thus each of these coreset problems correspond to a different matrix approximation problem. Furthermore, sampling methods naturally give approximations that are of the form of a subset of the dataset, and thus there is a rich interplay between the literature of coreset algorithms and matrix approximation. We refer the reader to [Fel20] for a survey on the literature of coreset algorithms.

Sensitivity sampling

One of the most important sampling-based approaches to constructing coresets is the *sensitivity sampling* method, and will form the starting point for many of the sampling-based algorithms that we study in Part II. The sensitivity framework was introduced by [LS10, FL11] and further optimized by [BFL16, FSS20] in order to develop a unified approach to sampling-based approximation algorithms for a wide range of problems including clustering, projective clustering, low rank approximation and subspace approximation, empirical risk minimization, and others.

In this general framework, we seek to approximate an objective function $f : X \to \mathbb{R}_{\geq 0}$ of the form of a sum

$$f(\mathbf{x}) \coloneqq \sum_{i=1}^{n} f_i(\mathbf{x})$$

by sampling a subset $S \subseteq [n]$ as well as associated weights $\mathbf{w}_i$ for $i \in S$, so that

$$f(\mathbf{x}) = (1 \pm \varepsilon) \sum_{i \in S} \mathbf{w}_i f_i(\mathbf{x}) \tag{1.2}$$

simultaneously for every $\mathbf{x} \in X$. As an example instantiation of this framework, we may consider the approximate mean estimation problem, where given an input matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, we wish to find some vector $\hat{\mathbf{x}} \in \mathbb{R}^d$ such that

$$\sum_{i=1}^{n} \|\mathbf{a}_i - \hat{\mathbf{x}}\|_2^2 \leq (1 + \varepsilon) \min_{\mathbf{x} \in \mathbb{R}^d} \sum_{i=1}^{n} \|\mathbf{a}_i - \mathbf{x}\|_2^2.$$

The functions f_i in this case correspond to $f_i(\mathbf{x}) = \|\mathbf{a}_i - \mathbf{x}\|_2^2$.

To sample our approximation, we define *sensitivity scores* $\boldsymbol{\sigma}_i$, and sample functions f_i with probabilities p_i proportional to $\boldsymbol{\sigma}_i$ with weights $\mathbf{w}_i = 1/p_i$.

Definition 1.3.1 (Sensitivity score [LS10, FL11])**.** For $i \in [n]$, let $f_i : X \to \mathbb{R}_{\geq 0}$ be functions. Then, the *ith sensitivity score* is defined as

$$\boldsymbol{\sigma}_i \coloneqq \sup_{\mathbf{x} \in X} \frac{f_i(\mathbf{x})}{\sum_{j=1}^{n} f_j(\mathbf{x})}$$

and the *total sensitivity* is defined as $\mathfrak{S} \coloneqq \sum_{i=1}^{n} \boldsymbol{\sigma}_i$.

A preview of sampling arguments

Suppose that we construct weights $\{\mathbf{w}_i\}_{i=1}^n$ via the sensitivity sampling framework by sampling the weight $\mathbf{w}_i = 1/p_i$ with probability $\min\{1, \boldsymbol{\sigma}_i/\alpha\}$ for some $\alpha > 0$ to be determined later. Then for any fixed $\mathbf{x} \in X$, we have that

$$\mathbf{E}\left[\sum_{i \in S} \mathbf{w}_i f_i(\mathbf{x})\right] = \sum_{i=1}^{n} \mathbf{E}[\mathbf{w}_i f_i(\mathbf{x})] = \sum_{i=1}^{n} p_i \frac{f_i(\mathbf{x})}{p_i} = \sum_{i=1}^{n} f_i(\mathbf{x})$$

so $\sum_{i\in S} \mathbf{w}_i f_i(\mathbf{x})$ is an unbiased estimator of $\sum_{i=1}^n f_i(\mathbf{x})$. We can also bound the variance as

$$\begin{aligned}\mathbf{Var}\left[\sum_{i\in S} \mathbf{w}_i f_i(\mathbf{x})\right] &= \sum_{i=1}^n \mathbf{Var}(\mathbf{w}_i f_i(\mathbf{x})) \leq \sum_{i=1}^n p_i \frac{f_i(\mathbf{x})^2}{p_i^2} \\ &= \sum_{i=1}^n \alpha \frac{f_i(\mathbf{x})}{\boldsymbol{\sigma}_i} f_i(\mathbf{x}) \leq \alpha \left(\sum_{i=1}^n f_i(\mathbf{x})\right)^2.\end{aligned}$$

Thus, it follows from Chebyshev's inequality that

$$\sum_{i\in S} \mathbf{w}_i f_i(\mathbf{x}) = (1 \pm O(\sqrt{\alpha})) \sum_{i=1}^n f_i(\mathbf{x})$$

with constant probability, so setting $\alpha = O(\varepsilon^2)$ gives a $(1\pm\varepsilon)$ approximation for any fixed $\mathbf{x}$ with constant probability. We can further replace the use of Chebyshev's inequality with a Bernstein bound to get improved concentration, so that if we set $\alpha = O(\varepsilon^2)/\log\frac{1}{\delta}$, then the above bound holds with probability at least $1-\delta$, for each fixed $\mathbf{x}\in X$. Note that the support size required for this guarantee is $\mathbf{E}|S| = O(\mathfrak{S}/\alpha) = O(\varepsilon^{-2}\mathfrak{S}\log\frac{1}{\delta})$ in expectation.

The sampling argument so far gives a $(1\pm\varepsilon)$ approximation of the objective function $f(\mathbf{x})$ for any fixed $\mathbf{x}\in X$, but this alone is not sufficient to guarantee that $\sum_{i\in S}\mathbf{w}_i f_i(\mathbf{x}) = (1\pm\varepsilon)f(\mathbf{x})$ *simultaneously for every* $\mathbf{x}\in X$, which is what we need if we wish to find an approximate minimizer of f. To obtain such a guarantee, we need a *net argument* where we approximate the domain X by a finite subset $\mathcal{N}\subseteq X$ known as a *net*, for which $(1\pm\varepsilon)$ approximations on $\mathcal{N}$ imply $(1\pm O(\varepsilon))$ approximations on X. We can then apply the previous result with $\delta = O(1/|\mathcal{N}|)$ and union bound over all $\mathbf{x}\in\mathcal{N}$ to conclude that the sampling weights $\mathbf{w}$ yield approximations for every $\mathbf{x}\in X$. For X in d-dimensional space, $|\mathcal{N}|$ is typically on the order of $(1/\varepsilon)^d$, and thus the above argument typically gives a coreset size of roughly $|S| = \tilde{O}(\varepsilon^{-2}\mathfrak{S}d)$. In fact, for a wide variety of applications, it can be shown that sampling $|S| = \tilde{O}(\varepsilon^{-2}\mathfrak{S}d)$ functions f_i is sufficient to achieve the guarantee of (1.2), where d is a complexity parameter known as the *VC-dimension* of a certain set system associated with the functions $\{f_i\}_{i=1}^n$ [LS10, FL11, BFL16, FSS20].

Beyond VC-dimension arguments

While the above basic sensitivity sampling framework gives a strong baseline sampling result, the resulting guarantees are often far from optimal, and more sophisticated arguments are needed to obtain nearly optimal coreset sizes $|S|$. A large fraction of Part II of this thesis will be concerned with studying and developing arguments to improve over the basic union bound and/or VC-dimension argument, which often lead to nearly optimal bounds on coreset sizes for a variety of problems. In particular, we will work extensively with a technique known as *chaining* in later chapters, which improves the basic union bound argument discussed above by carefully constructing a *sequence* of nets at different levels of granularity, rather than considering only a single nets.

1.3.2 Leverage score sampling

Suppose that we apply the sensitivity sampling framework to the objective function $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|_2^2 = \sum_{i=1}^n \langle \mathbf{a}_i, \mathbf{x}\rangle^2$, where $\mathbf{A}$ is an $n \times d$ matrix and the domain of f is $\mathbb{R}^d$. In this case, the sensitivity sampling framework samples weights $\mathbf{w}_i$ for $i \in S \subseteq [n]$ such that

$$\sum_{i \in S} \mathbf{w}_i \langle \mathbf{a}_i, \mathbf{x}\rangle^2 = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_2^2.$$

Note that if we set $\mathbf{S}$ to be a diagonal matrix with $\mathbf{S}_{i,i} = \sqrt{\mathbf{w}_i}$ whenever $i \in S$ is sampled, then we can write the above guarantee as

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_2^2$$

simultaneously for every $\mathbf{x} \in \mathbb{R}^d$. That is, $\mathbf{S}$ is an ℓ_2 subspace embedding (Definition 1.1.1) for $\mathbf{A}$! In this special case of sensitivity sampling, the sensitivity scores have been known for a long time in the statistics literature as the *leverage scores of* $\mathbf{A}$. Indeed, the use of leverage scores in the context of coreset algorithms preceded the development of the sensitivity sampling framework, and was studied in works such as [BSST13, BK15] in the context of graph sparsification as well as [DMM06a, DMM06b, DKM06a, DKM06b, RV07, DMM08, Mag10] to construct coresets for ℓ_2 linear regression and Frobenius norm low rank approximation. Furthermore, it is known that leverage scores can be approximated extremely efficiently by combining ideas from sketching and recursive sampling [SS11, DMMW12, CW13, LMP13, CLM+15], in fact in time $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^\omega)$, and thus leverage score sampling has become an important primitive for designing fast randomized algorithms for numerical linear algebra.

Definition 1.3.2 (Leverage scores)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then for each $i \in [n]$, the *ith leverage score of* $\mathbf{A}$ is defined to be

$$\boldsymbol{\tau}_i(\mathbf{A}) \coloneqq \sup_{\mathbf{A}\mathbf{x} \neq 0} \frac{[\mathbf{A}\mathbf{x}](i)^2}{\|\mathbf{A}\mathbf{x}\|_2^2} = \mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{A})^- \mathbf{a}_i,$$

where $\mathbf{a}_i = \mathbf{e}_i^\top \mathbf{A}$ is the ith row of $\mathbf{A}$.

Let us now instantiate the sensitivity sampling guarantee discussed in Section 1.3.1 to the setting of ℓ_2 subspace embeddings. We first bound the total sensitivity $\mathfrak{S} = \sum_{i=1}^n \boldsymbol{\tau}_i(\mathbf{A})$. Let $\mathbf{U} \in \mathbb{R}^{n \times \mathrm{rank}(\mathbf{A})}$ be an orthogonal basis for the column space of $\mathbf{A}$. Then, the supremum characterization of leverage scores in Definition 1.3.2 does not depend on the particular basis chosen for the column space of $\mathbf{A}$, so we have that

$$\boldsymbol{\tau}_i(\mathbf{A}) = \sup_{\mathbf{U}\mathbf{x} \neq 0} \frac{[\mathbf{U}\mathbf{x}](i)^2}{\|\mathbf{U}\mathbf{x}\|_2^2} = \sup_{\mathbf{U}\mathbf{x} \neq 0} \frac{[\mathbf{U}\mathbf{x}](i)^2}{\|\mathbf{x}\|_2^2} = \sup_{\|\mathbf{x}\|_2 = 1} [\mathbf{U}\mathbf{x}](i)^2 = \|\mathbf{e}_i^\top \mathbf{U}\|_2^2.$$

It is then easy to see that

$$\sum_{i=1}^n \boldsymbol{\tau}_i(\mathbf{A}) = \sum_{i=1}^n \|\mathbf{e}_i^\top \mathbf{U}\|_2^2 = \|\mathbf{U}\|_F^2 = \mathrm{rank}(\mathbf{A}). \tag{1.3}$$

Thus, the bound given by the sensitivity sampling framework on the number of rows r sampled by the leverage score sampling-based ℓ_2 subspace embedding is

$$r = \tilde{O}(\varepsilon^{-2}\mathfrak{S}d) = \tilde{O}(\varepsilon^{-2}d^2)$$

Recall from Theorem 1.2.3 in Part I, however, that a bound of $r = \tilde{O}(\varepsilon^{-2}d)$ is achievable via oblivious random projections, even without knowing $\mathbf{A}$. A bound of $r = \tilde{O}(\varepsilon^{-2}d)$ is thus quite pessimistic, and in fact, the analysis can be improved to a bound of $r = \tilde{O}(\varepsilon^{-2}d)$. Indeed, a series of works have culminated in the following guarantee for leverage score sampling, which achieves a nearly linear dependence on d in the row count r.

Theorem 1.3.3 (Leverage score sampling [DMM06a, RV07, Mag10])**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\alpha > 0$ and let $p_i = \min\{1, \boldsymbol{\tau}_i(\mathbf{A})/\alpha\}$ for $i \in [n]$. Let $\mathbf{S} \in \mathbb{R}^{n\times n}$ be the diagonal matrix formed by independently setting

$$\mathbf{S}_{i,i} = \begin{cases} \frac{1}{\sqrt{p}_i} & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{cases}$$

for each $i \in [n]$. Then, there is an α such that with probability at least $99/100$, $\mathbf{S}$ is an ℓ_2 subspace embedding satisfying Definition 1.1.1 with $\kappa = (1+\varepsilon)$, and furthermore, $\mathbf{S}$ has at most $r = O(\varepsilon^{-2}d\log d)$ nonzero rows.

The proofs of the above result given by the works [RV07, Mag10] crucially rely on the special structure of the ℓ_2 norm, which allows one to translate the problem of analyzing the maximum sampling error $\sup_{\|\mathbf{Ax}\|_2\leq 1}|\|\mathbf{SAx}\|_2^2 - \|\mathbf{Ax}\|_2^2|$ to a problem about analyzing the spectral norm of random matrices. However, such an argument does not generalize easily to other sampling problems such as ℓ_p subspace embeddings. In order to obtain similar improvements for these other problems, we will instead need a technique known as *chaining*. These arguments will be discussed extensively in later chapters.

1.3.3 Streaming and online coresets

In many big data settings, one does not have the luxury of accessing the entire dataset at once, and one must instead settle for accessing the dataset one row at a time. This is known as the *row arrival streaming model* or the *geometric streaming model*, and represents many realistic settings, for example when data is loaded into memory in batches of rows from disk. Coresets are a valuable tool in this setting, since they can be used to store representatives of large amounts of data in a small amount of space, for example via the use of the *merge-and-reduce* technique which repeatedly accumulates data and computes coresets to maintain a coreset without ever requiring a large amount of working memory. A further useful restriction to this setting is to require that the coreset be constructed in an *online* fashion, meaning that the a row must be selected to be included in the coreset irrevocably at the time of arrival. This setting is known as the *online coreset* model introduced by [CMP16, CMP20], and has proven to be a highly valuable tool for streaming algorithms, with applications to the design of sliding window algorithms [BDM+20] as well as smaller space row arrival algorithms [CWZ23]. Throughout our study of coresets and sampling algorithms, we will often simultaneously study how to extend our constructions to the online setting.

Online leverage score sampling

As an introduction to results and techniques in the online coreset setting, we discuss a result of [CMP16, CMP20] which shows how to construct online coresets for ℓ_2 subspace embeddings. In this setting, our input is an $n \times d$ matrix $\mathbf{A}$ whose rows $\mathbf{a}_i \in \mathbb{R}^d$ arrive one at a time. At each time step $i \in [n]$, the row $\mathbf{a}_i$ arrives, and we must irrevocably commit to some weight $\mathbf{s}_i$ for this row and discard $\mathbf{a}_i$ if $\mathbf{s}_i = 0$ or keep $\mathbf{a}_i$ if $\mathbf{s}_i \neq 0$. Finally, we must have that $\|\mathbf{S}_i\mathbf{A}_i\mathbf{x}\|_2^2 = (1 \pm \varepsilon)\|\mathbf{A}_i\mathbf{x}\|_2^2$ simultaneously for every $\mathbf{x} \in \mathbb{R}^d$ for each $i \in [n]$, where $\mathbf{A}_i \in \mathbb{R}^{n \times d}$ denotes the matrix formed by the first i rows of $\mathbf{A}$. Our goal is to take $r = \mathsf{nnz}(\mathbf{S})$ to be as small as possible.

To analyze algorithms for this problem, we first introduce a couple of definitions. The first and most important definition is the *online leverage scores of* $\mathbf{A}$, which defines a version of leverage scores that can be computed in a manner compatible with the online coreset model.

Definition 1.3.4 (Definition 2.1 of [BDM+20], Theorem 2.2 of [CMP20])**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then, for each $i \in [n]$, the *ith online leverage score of* $\mathbf{A}$ is defined as

$$\boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}) \coloneqq \begin{cases} \min\left\{\mathbf{a}_i^\top(\mathbf{A}_{i-1}^\top\mathbf{A}_{i-1})^-\mathbf{a}_i, 1\right\} & \mathbf{a}_i \in \mathrm{rowspan}(\mathbf{A}_{i-1}) \\ 1 & \text{otherwise} \end{cases}$$

where $\mathbf{A}_j \in \mathbb{R}^{j \times d}$ denotes the submatrix of $\mathbf{A}$ formed by the first j rows.

It is not hard to see that the online leverage scores upper bound the usual leverage scores of $\mathbf{A}$, since each of the rows $\mathbf{a}_i$ are being compared with a smaller quadratic form.

Lemma 1.3.5 (Online leverage scores bound leverage scores)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then, for each $i \in [n]$,

$$\boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}) \geq \boldsymbol{\tau}_i(\mathbf{A}).$$

Furthermore, the quadratic form $\mathbf{A}_{i-1}^\top\mathbf{A}_{i-1}$ needed to compute the online leverage scores can be maintained in d^2 words of space. Thus, if the sum of the leverage scores is not too large, then sampling proportionally to the online leverage score immediately yields online coresets for ℓ_2 subspace embeddings. The remaining task is to show that the sum of the online leverage scores can be bounded. We define the *online condition number of* $\mathbf{A}$, which is a quantity that will characterize this crucial quantity.

Definition 1.3.6 (Online condition number)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then, the *online condition number of* $\mathbf{A}$ is defined as

$$\kappa^{\mathsf{OL}} \coloneqq \|\mathbf{A}\|_2 \max_{i=1}^{n} \|(\mathbf{A}_i)^-\|_2.$$

One of the main results of [CMP16, CMP20] is to show that the online leverage scores sum to at most $O(d \log \kappa^{\mathsf{OL}})$. It is known that such a dependence on the condition number is necessary [CMP16, CMP20]. A crucial lemma used in the bound on the sum of online leverage scores is the *matrix determinant lemma*:

Lemma 1.3.7 (Matrix determinant lemma)**.** Let $\mathbf{M} \in \mathbb{R}^{d \times d}$ be an invertible matrix and let $\mathbf{a} \in \mathbb{R}^d$. Then,

$$\det(\mathbf{M} + \mathbf{a}\mathbf{a}^\top) = \det(\mathbf{M})(1 + \mathbf{a}^\top\mathbf{M}^{-1}\mathbf{a})$$

Lemma 1.3.8 (Sum of online leverage scores [CMP16, CMP20])**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then, $\sum_{i=1}^n \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}) = O(d \log \kappa^{\mathsf{OL}})$.

Online leverage scores for integer points [WY22a]

Although the $O(d \log \kappa^{\mathsf{OL}})$ bound is necessary in general, this bound is quite pessimistic, especially if $\mathbf{A}$ is an integer matrix with entries bounded by, say, $\mathrm{poly}(n)$. In this case, the online condition number can be as large as $\mathrm{poly}(n)^d$ [AVu97], so Lemma 1.3.8 results in a bound of $O(d^2 \log n)$ on the sum of online leverage scores. Note that this is quadratically worse than d bound that holds for offline leverage scores (1.3). In fact, we show in [WY22b] (see Theorem 11.1.1 in Chapter 11) that for integer matrices with entries bounded by $\mathrm{poly}(n)$, we can improve the bound on the sum of online leverage scores to $O(d \log n)$, matching the offline result up to just a $O(\log n)$ factor.

1.3.4 Applications of sampling algorithms beyond coresets

So far, we have motivated sampling algorithms mostly as a way of obtaining dataset compressions via small coresets. While this is certainly a compelling application, the special form of compression by sampling lends itself to many uses beyond this simple use case.

One setting in which sampling algorithms excel is the setting of *active learning*. In this setting, we wish to solve a supervised learning problem which usually involves solving an optimization problem with training examples as well as corresponding labels. Furthermore, we consider accessing labels to be an expensive process; for example, labeling a training example may involve conducting a survey, a physical experiment, or a time-intensive computer simulation. Thus, it is desirable to solve this supervised learning problem while minimizing the number of label entries that we need to read. In this setting, sampling algorithms are a natural strategy to consider, since if a small subset of the dataset is sufficient to solve the optimization problem, then we only need to access the labels corresponding to these dataset entries. This observation has indeed been utilized to solve least squares regression in work of [CP19], and we use a similar strategy to obtain nearly optimal algorithms for active ℓ_p linear regression in Chapter 12.

Another setting involving the analysis of sampling algorithms to design query-efficient algorithms is *sublinear power mean estimation*. In this problem, we have query access to a set of n points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ in d dimensions, and we wish to estimate the *power mean* of this dataset, that is, the point $\mathbf{x} \in \mathbb{R}^d$ that minimizes the ℓ_p norm of the Euclidean distances of $\mathbf{x}_i$ to $\mathbf{x}$. It is in fact possible to approximately solve this problem in *sublinear time*, i.e., in time substantially less than the time it takes to read the entire dataset. Indeed, prior results of [CSS21] show that a uniform sample of $\mathrm{poly}(\varepsilon^{-1})$ points is sufficient to optimize the objective up to a $(1+\varepsilon)$ factor. In Chapter 13, we will show that our techniques from active ℓ_p linear regression can in fact be applied this problem to obtain a tight upper bound on the query complexity of this problem.

Finally, we note that sampling-based compressions, and especially online coreset techniques, are particularly well-suited to solve problems in the geometric streaming model (see Section 1.3.3) since the dataset is partitioned into rows in this streaming model, which is naturally how sampling algorithms interact with the data. One natural question in the geometric streaming model is the problem of maintaining a small ellipsoid that covers all of the points of the stream, known as the *streaming minimum enclosing ellipsoid* problem. Perhaps surprisingly, it was not known how to solve this simple problem in a small amount of space despite consideration by various works [AHV04, AHV05, MSS10, AS15]. In Chapter 11, we use insights from our study of online coresets to solve this decade-old question in computational geometry.

1.3.5 Overview of Part II

We have alluded to the fact that sampling-based methods for matrix approximation can give improved guarantees over the oblivious methods discussed in Part I. We will see the full power of this fact in Chapter 6, where we study the method of ℓ_p Lewis weight sampling for constructing $(1+\varepsilon)$-approximate ℓ_p subspace embeddings *for every* $p > 0$, and describe some improvements and extensions to the online coreset model given by [WY23b]. In Chapters 7 and 8, we study two other sampling methods, ℓ_p sensitivity sampling and root leverage score sampling, that give similar guarantees that can at times be more useful than ℓ_p Lewis weight sampling. The results in this section are taken from [WY23c] and [WY24b]. Chapter 9 studies the question of how the row count r of ℓ_p subspace embeddings can be improved if we allow for the distortion κ to be substantially larger than $(1+\varepsilon)$. In particular, we study trade-offs shown in [WY22b] that improve the row count r by $\mathrm{poly}(d)$ factors if the distortion can be as large as $\mathrm{poly}(d)$. Chapter 10 shows that subspace embeddings can be constructed for loss functions for beyond ℓ_p norms by using the sensitivity sampling technique, based on results in [MMWY22].

While Chapters 6 through 10 have focused intensely on developing sampling techniques for subspace embeddings, the next two chapters look more towards applications. Chapter 11 continues the study of high-distortion ℓ_p subspace embeddings and presents the results of [WY22b] that obtains the first streaming, and in fact online, coreset for ℓ_∞ subspace embeddings, which has profound implications in streaming computational geometry. In Chapter 12, we present the results of [MMWY22, WY23a] that obtains nearly optimal bounds on the problem of *active* ℓ_p *linear regression*, which solves ℓ_p linear regression while reading the minimal number of labels. This is applied in Chapter 13 to obtain a wide variety of coreset results for *multiple* ℓ_p regression with important corollaries for sublinear power mean estimation and subspace approximation, based on work of [WY24a]. In Chapter 14, we construct the first nearly optimal coresets for the problem of ℓ_p subspace approximation, based on the work of [WY24b].

Open questions arising from work in this part are collected in Chapter 15.

1.4 Sparse optimization

The approximation of matrices by a simpler matrix is quite similar in spirit to the problem of *sparse optimization*. Indeed, "approximation" is often captured by some (often convex) optimization problem, while a "simple" or "structured" matrix is often captured by sparsity. It is thus natural to expect that techniques in the literature of sparse optimization would be of great utility in matrix approximation problems. In Part III, we depart from the sketching and sampling-based approaches discussed in Parts I and II to discuss our results on matrix approximation that draw from problems and techniques more closely related to sparse optimization.

1.4.1 Sparse linear regression

We start with a discussion of *sparse linear regression*, which is arguably the most fundamental question in sparse optimization. In sparse linear regression, we are given an instance of the usual linear regression problem, i.e. a design matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and a target vector $\mathbf{b}$, with an additional

parameter k which specifies a *sparsity* parameter. The goal is then to output a coefficient vector $\mathbf{x}$ *with at most* k *nonzero entries* that minimizes $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$. That is, we wish to solve linear regression, restricted to k-sparse vectors $\mathbf{x}$. One of the most popular motivations for this problem is for applications to *feature selection*, in which we wish to select a small subset of features to use for a prediction model. This often leads to improvements in efficiency, generalizability, and interpretability.

Unfortunately, the sparse linear regression problem is NP-hard to solve in the worst case, even with bicriteria sparsity and large multiplicative error [Nat95, FKT15, GV21, PSZ22]. To overcome this intractability, there has been much work on studying popular efficient algorithms used in practice that can at least successfully solve well-behaved instances. One popular approach is convex relaxation, where the sparsity constraint is relaxed to an ℓ_1 constraint, which leads to an approach known as the LASSO [Tib96]. The LASSO and its variants are known to provably solve the sparse linear regression problem when the design matrix $\mathbf{A}$ satisfies a condition known as the *restricted isometry property (RIP)* [DS89, CDS98, Tro06, CRT06, CT07, Can08, BRT09, Zho09, RWY10, BCFS14] and $\mathbf{b}$ is (almost) exactly a k-sparse linear combination of the columns of $\mathbf{A}$. The combination of efficient algorithms and provable guarantees has made the LASSO one of the most widely used methods for sparse linear regression.

Another class of popular algorithms for sparse linear regression are *greedy algorithms*, which are based on the idea of greedily adding the column of $\mathbf{A}$ with the largest "local improvement" to the current set of columns. Greedy algorithms are perhaps one of the oldest and most natural approaches, and provable guarantees for such greedy methods have been established under similar RIP-like conditions [SSZ10, DK11, LS17, EKDN18]. In fact, these works have succeeded in proving similar results beyond the sparse linear regression setting, and work in general for the more general setting of sparse convex optimization, under an appropriate generalization of the restricted isometry property to general convex functions via restricted strong convexity and restricted smoothness.

1.4.2 Column subset selection

Another well-studied linear algebraic problem in the area of sparse optimization is *column subset selection*. This problem has origins in the numerical linear algebra literature [Cha86, GE96, GVL13], and asks for a small subset of columns of a matrix $\mathbf{A}$ that "best represents" $\mathbf{A}$. This problem was also studied in the randomized numerical linear algebra literature by making use of sampling techniques that sample a subset of columns that span a nearly optimal (bicriteria) low rank approximation of $\mathbf{A}$ [FKV04, DV06, DKM06b, DMM06b, DMM08, BWZ16]. In this formulation, we seek a sparse set of columns S that minimizes $\|\mathbf{A}|^S\mathbf{X} - \mathbf{A}\|_F^2$, and thus this problem can also be viewed as a special case of a sparse linear regression problem with multiple responses. One can thus expect greedy algorithms to be applicable for the column subset selection problem as well, and indeed, such results have been established by several works [SVW15, ABF+16, LS17]. Related greedy algorithms have also been analyzed for loss functions other than the Frobenius norm [SWZ17, CGK+17, DWZ+19, SWZ19, JLL+21, MW21, WY23a, AY23].

1.4.3 Overview of Part III

In Chapter 16, we develop a new connection between the LASSO and greedy algorithms by showing that when applied sequentially, the LASSO is in fact *equivalent* to a greedy algorithm known as orthogonal matching pursuit for sparse convex optimization. In fact, this holds even in the *group* sparse convex optimization setting, which allows our results to be immediately applied to the matrix column subset selection problem. These results give theoretical explanations for a novel feature selection algorithm proposed in [YBC+23] called Sequential Attention, which achieves remarkable results in experiments. These results are based on the works [YBC+23, AY23].

As mentioned previously, there has been substantial interest in developing algorithms for column subset selection under loss functions beyond the Frobenius norm. In Chapter 17, we move away from the general setting of sparse convex optimization, and focus specifically on the matrix column subset selection problem with entrywise loss functions. Based on work developed in [WY23a], we show how the idea of well-conditioned spanning sets, which we used in Chapter 3 to obtain nearly optimal oblivious ℓ_p subspace embeddings, can be used to obtain improved (and even optimal) column subset selection algorithms for entrywise loss functions.

Finally, we change perspectives on the theme of sparse optimization and low rank approximation by studying algorithms for low rank approximation when the input is assumed to have a sparse optimal solution. In Chapter 18, we show that if a matrix $\mathbf{A}$ is promised to have sparse singular vectors, then a $(1+\varepsilon)$-approximate spectral rank k approximation of $\mathbf{A}$ can be computed in roughly $\mathsf{nnz}(\mathbf{A})/\sqrt{\varepsilon}$ time, improving the best known algorithm for spectral low rank approximation [MM15] by a factor of k.

Open questions arising from work in this part are collected in Chapter 19.

1.5 Connections to geometric functional analysis

In addition to randomized numerical algebra and sparse optimization, much of the theory developed in this thesis interacts heavily with the literature of *geometric functional analysis*. Indeed, many matrix approximation problems can be phrased as *embedding* problems, i.e. the problem of constructing "nice" mappings from one space to another, which is a fundamental theme in geometric functional analysis.

1.5.1 Lewis weights and embedding subspaces of ℓ_p

The ℓ_p spaces, the space of vectors (or functions) equipped with the ℓ_p norm, are a fundamental object in functional analysis, and the embedding of subspaces of ℓ_p into other spaces is similarly one of the most well-studied questions in functional analysis. In our thesis, we make contributions to two such questions, one on embeddings into ℓ_p^n for a small n, and one on embeddings into ℓ_q.

Embeddings from ℓ_p to ℓ_p: ℓ_p Lewis weight sampling

It is a well-known fact that any d-dimensional subspace V of ℓ_2 (regardless of the ambient dimension) is *isometric* to $\mathbb{R}^d$ equipped with the ℓ_2 norm. That is, there is an invertible linear map

$T : V \to \mathbb{R}^d$ such that

$$\|T(\mathbf{y})\|_2 = \|\mathbf{y}\|_2$$

for every $\mathbf{y} \in V$. Such a result is not possible for the ℓ_p norm if we insist on an isometry, but in general, we can ask for a trade-off between the target dimension r and a distortion parameter κ. This leads to the following natural question that was studied by many works [Sch87, BLM89, Tal90, LT91, Tal95, Zva00, SZ01, Sch11]:

Question 1.5.1. Given a distortion parameter κ, what is the smallest target dimension r such that every d-dimensional subspace V of ℓ_p admits an invertible linear map $T : V \to \mathbb{R}^r$ such that

$$\|\mathbf{y}\|_p \leq \|T(\mathbf{y})\|_p \leq \kappa\|\mathbf{y}\|_p?$$

We may recognize that this is exactly the question of constructing subspace embeddings for the ℓ_p norm (Definition 1.1.1). In the regime of $\kappa = (1 + \varepsilon)$ for a small ε, this question is nearly optimally resolved by works of [BLM89, Tal90, LT91, Tal95, SZ01] (see Theorem 6.1.4), even algorithmically [CP15], via a technique known as ℓ_p *Lewis weight sampling* (see Chapter 6). Throughout this thesis, we will develop various generalizations of this machinery to solve a wide variety of problems in theoretical computer science, including active ℓ_p regression (Chapter 12), multiple ℓ_p regression (Chapter 13), and ℓ_p subspace approximation (Chapter 14), even in information-limited settings such as online, streaming, and distributed computation.

Embeddings from ℓ_p to ℓ_q

Another natural question on embedding subspaces of ℓ_p, studied by [LT80], is the following:

Question 1.5.2. *What is the smallest distortion κ such that every d-dimensional subspace V of ℓ_p admits an invertible linear map $T : V \to V'$ for any space V' such that*

$$\|\mathbf{y}\|_p \leq \|T(\mathbf{y})\|_q \leq \kappa\|\mathbf{y}\|_p?$$

That is, what is the smallest distortion incurred when embedding a subspace of ℓ_p into any subspace of ℓ_q? This question is in fact nearly optimally resolved by [LT80], but their proof is quite sophisticated, relying on deep results from the factorization theory of operator ideals [Pie80]. On the other hand, we show the same theorem with an elementary proof based on a simple yet new property of ℓ_p Lewis weights in Chapter 9. Our observations may be useful in gaining further insight into the deeper results employed by [LT80] and simplifying some of the techniques in this field.

1.5.2 Well-conditioned bases and spanning sets

Another example of the connections between matrix approximation and geometric functional analysis appears in the construction of oblivious ℓ_p subspace embeddings (see Section 1.2.3). The work of [SW11] first showed that oblivious ℓ_1 subspace embeddings for a matrix $\mathbf{A}$ could be constructed by using the *existence* of a basis for the subspace $V = \text{colspan}(\mathbf{A})$ with certain "well-conditioning" properties that generalizes the notion of orthogonal bases for the ℓ_2 norm.

More specifically, for a d-dimensional subspace $V \subseteq \mathbb{R}^n$, an orthogonal basis $\mathbf{U} \in \mathbb{R}^{n \times d}$ is a set of d ℓ_2-unit vectors spanning V such that

$$\|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$$

for every $\mathbf{x} \in \mathbb{R}^d$. Similarly, an ℓ_1 well-conditioned basis $\mathbf{U} \in \mathbb{R}^{n \times d}$ is a set of d ℓ_1-unit vectors spanning V such that

$$\|\mathbf{U}\mathbf{x}\|_1 \geq \|\mathbf{x}\|_\infty.$$

In fact, the construction of such bases for the ℓ_1 norm implies *nearly optimal* constructions [LWW21], and can be accomplished by using a construction from geometric functional analysis known as *Auerbach bases* [Aue30]. In general, it is known that the appropriate generalization to ℓ_p norms for $1 < p < 2$ similarly implies nearly optimal oblivious ℓ_p subspace embeddings, that is, a set of d ℓ_p-unit vectors spanning V such that

$$\|\mathbf{U}\mathbf{x}\|_p \geq \|\mathbf{x}\|_q,$$

where q is the Hölder conjugate exponent of p. However, the existence of such bases is still not known, despite substantial interest in well-conditioned bases for subspaces of ℓ_p. Some partial attempts leading to suboptimal trade-offs include the use of Auerbach bases [Aue30], John ellipsoids [Joh48], and Lewis bases [Lew78].

In Chapter 3, we show that by relaxing the requirement of constructing a basis for the subspace to a *spanning set*, we can construct a set of $O(d)$ ℓ_p-unit vectors $\mathbf{U}$ spanning V such that for any $\mathbf{A}\mathbf{x} \in V$, we can write $\mathbf{A}\mathbf{x} = \mathbf{U}\mathbf{y}$ such that

$$\|\mathbf{U}\mathbf{y}\|_p \geq c\|\mathbf{y}\|_q$$

for some universal constant c. This relaxation is in fact sufficient to recover nearly optimal ℓ_p subspace embeddings, resolving an old question in randomized numerical linear algebra.

Chapter 2

Preliminaries

2.1 Notation

- We write $[n]$ to denote the set $\{1, 2, 3, \ldots, n\} = \{i \in \mathbb{N} : 1 \leq i \leq n\}$.
- For an n-dimensional vector $\mathbf{y} \in \mathbb{R}^n$, we write $\mathbf{y}(i)$ or $\mathbf{y}_i$ to denote the ith entry of $\mathbf{y}$. For an $n \times d$ matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, we write $\mathbf{A}_{i,j}$ to denote the (i, j)th entry of $\mathbf{A}$.

2.1.1 Linear algebra

- We write $\mathbf{e}_i$ to denote the i-th standard basis vector, i.e., the vector with 1 in the i-th entry and 0s everywhere else.
- We write $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ to denote the identity matrix in d dimensions, that is, the matrix with $\mathbf{I}_d(i, i) = 1$ for $i \in [d]$ and $\mathbf{I}_d(i, j) = 0$ for $i \neq j \in [d]$.
- We write $\mathsf{nnz}(\mathbf{A})$ for the number of nonzero entries of a matrix $\mathbf{A}$.
- For a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, we write $\mathrm{colspan}(\mathbf{A}) = \{\mathbf{A}\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \in \mathbb{R}^d\}$ for the column span of $\mathbf{A}$, and $\mathrm{rowspan}(\mathbf{A}) = \{\mathbf{y}^\top \mathbf{A} \in \mathbb{R}^d : \mathbf{y} \in \mathbb{R}^n\}$ for the row span of $\mathbf{A}$.
- For a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, we write $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ for the singular value decomposition (SVD) of $\mathbf{A}$.
- For a rank parameter k, we let $\mathbf{\Sigma}_k$ denote the matrix $\mathbf{\Sigma}$ with all but the top k singular values zeroed out, $\mathbf{\Sigma}_{\setminus k}$ for the matrix $\mathbf{\Sigma}$ with all but the bottom $d - k$ singular values zeroed out, and $\mathbf{A}_k = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^\top$ for the optimal rank k approximation of $\mathbf{A}$ under the Frobenius norm.
- For a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, we write $\mathbf{A}^- \in \mathbb{R}^{d \times n}$ to denote the Moore–Penrose pseudoinverse, or simply the pseudoinverse, of $\mathbf{A}$.

2.1.2 Inequalities

We repeatedly use the following inequalities.

Fact 2.1.1. For any $p \geq 1$ and any $a, b \in \mathbb{R}$, $|a + b|^p \leq 2^{p-1}(|a|^p + |b|^p) = O(|a|^p + |b|^p)$.

Fact 2.1.2 (Corollary A.2, [MMR19])**.** For any $p \geq 1$, $\varepsilon > 0$, and any $a, b \in \mathbb{R}$, $|a+b|^p \leq (1+\varepsilon)|a|^p + \frac{(1+\varepsilon)^{p-1}}{\varepsilon^{p-1}}|b|^p$.

Fact 2.1.3. For any $p \geq 1$ and any $a, b \in \mathbb{R}$, $|a|^p - |b|^p \leq p|a-b|(|a|^{p-1} + |b|^{p-1})$.

2.1.3 Probability

- We write $\boldsymbol{\varepsilon} \sim \{\pm 1\}^n$ to denote a random n-dimensional Rademacher vector $\boldsymbol{\varepsilon}$ drawn with each entry $\boldsymbol{\varepsilon}_i$ drawn independently and uniformly from the set $\{\pm 1\}$.
- For a vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive semidefinite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, we write $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2.2 Streaming

2.2.1 INDEX

In the INDEX problem, Alice has a set $A \subseteq [n]$ and Bob has an index $i \in [n]$. Alice then generates a message M using a possibly randomized algorithm $\mathcal{A}$ as a function of A, and then passes the message M to Bob. Bob must then determine whether $i \in A$ or not.

Theorem 2.2.1 (INDEX lower bound [KNR99])**.** Suppose $\mathcal{A}$ solves the INDEX problem with probability at least $2/3$. Then, $\mathcal{A}$ must use at least $\Omega(n)$ bits.

2.3 Random processes

In our study of sampling algorithms (see Part II), we often seek to approximate a separable function

$$f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$$

by a weighted subset of summands given by

$$\tilde{f}(\mathbf{x}) = \sum_{i=1}^{n} \mathbf{w}_i f_i(\mathbf{x})$$

where most of weights $\mathbf{w}_i$ are zero (i.e., $\mathsf{nnz}(\mathbf{w}) \ll n$) so that $\tilde{f}$ provides a compressed approximation. Suppose that the random weights $\mathbf{w}_i$ are generated by independently setting $\mathbf{w}_i = 1/q_i$ with probability q_i and 0 with probability $1 - q_i$ for some sampling probability $q_i \in (0, 1)$, as is the case for most sampling algorithms. Then, the sampling error at any particular point $\mathbf{x} \in X$ in the domain, given by

$$|\tilde{f}(\mathbf{x}) - f(\mathbf{x})| = \left| \sum_{i=1}^{n} (\mathbf{w}_i - 1) f_i(\mathbf{x}) \right|,$$

is a random variable that lives in the probability space defined by the randomness of sampling the weights $\mathbf{w}$. Thus, the collection of all of these random variables, indexed by $\mathbf{x} \in X$, naturally defines a *random process*.

Definition 2.3.1 (Random process)**.** A *random process* is a collection $(X_t)_{t \in T}$ of random variables X_t on the same probability space, which is indexed by some set T.

To bound the error of this approximation, or the *sampling error*, we will be interested in bounding the quantity

$$\Lambda \coloneqq \sup_{\mathbf{x} \in X} |\tilde{f}(\mathbf{x}) - f(\mathbf{x})| = \sup_{\mathbf{x} \in X} \left| \sum_{i=1}^{n} (\mathbf{w}_i - 1) f_i(\mathbf{x}) \right|$$

for some bounded domain set X, which represents the largest error in the approximation over all $\mathbf{x}$ in the set X. That is, we wish to bound the supremum of this random process. In particular, we seek moment bounds of the form $\mathbf{E}[\Lambda^l] \leq \varepsilon^l$ for $l = O(\log \frac{1}{\delta})$, which implies that $\Lambda \leq O(\varepsilon)$ with probability at least $1 - \delta$ by Markov's inequality.

2.3.1 Symmetrization: reduction to a Rademacher process

In most cases, directly analyzing the sampling error Λ is inconvenient, and we instead resort to bounding Λ by a simpler random process via a technique known as *symmetrization*, which was first introduced in the study of empirical processes [GZ84]. This exploits the fact that the random variables $(\mathbf{w}_i - 1)$ have zero mean to bound the original random process by a *Rademacher process*, where the randomness is transferred from the sampling weights $\mathbf{w}_i$ to independent random signs $\boldsymbol{\varepsilon}_i \sim \{\pm 1\}$, up to a constant factor.

Lemma 2.3.2 (Symmetrization)**.** Suppose that for each $i \in [n]$, $\mathbf{w}_i$ is independently set to $1/q_i$ with probability q_i and 0 with probability $1 - q_i$ for some sampling probability $q_i \in (0, 1)$. Then,

$$\mathop{\mathbf{E}}_{\mathbf{w}} \left[\sup_{\mathbf{x} \in X} \left| \sum_{i=1}^{n} (\mathbf{w}_i - 1) f_i(\mathbf{x}) \right|^l \right] \leq 2^l \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon} \sim \{\pm 1\}^n, \mathbf{w}} \left[\sup_{\mathbf{x} \in X} \left| \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i \mathbf{w}_i f_i(\mathbf{x}) \right|^l \right]$$

Proof. Note that $Z(\mathbf{x}) = \sum_{i=1}^{n} (\mathbf{w}_i - 1) f_i(\mathbf{x})$ is a zero mean random variable. Then if Z' is an independent copy of Z with sampling weights $\mathbf{w}_i$, then by Jensen's inequality, we have that

$$\mathop{\mathbf{E}}_{\mathbf{w}, \mathbf{w}'} \sup_{\mathbf{x} \in X} |Z(\mathbf{x})|^l \leq \mathop{\mathbf{E}}_{\mathbf{w}, \mathbf{w}'} \sup_{\mathbf{x} \in X} |Z(\mathbf{x}) - Z'(\mathbf{x})|^l = \mathop{\mathbf{E}}_{\mathbf{w}, \mathbf{w}'} \sup_{\mathbf{x} \in X} \left| \sum_{i=1}^{n} (\mathbf{w}_i - \mathbf{w}_i') f_i(\mathbf{x}) \right|^l.$$

Furthermore, by independence, the distribution does not change if we multiply each summand $i \in [n]$ with signs $\boldsymbol{\varepsilon}_i \in \{\pm 1\}$. In particular, we can take these signs to be random Rademacher signs, so the above quantity is equal to

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon} \sim \{\pm 1\}^n, \mathbf{w}, \mathbf{w}'} \sup_{\mathbf{x} \in X} \left| \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i (\mathbf{w}_i - \mathbf{w}_i') f_i(\mathbf{x}) \right|^l.$$

This is at most

$$2^{l-1}\left(\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n,\mathbf{w}}\sup_{\mathbf{x}\in X}\left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i\mathbf{w}_i f_i(\mathbf{x})\right|^l + \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n,\mathbf{w}'}\sup_{\mathbf{x}\in X}\left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i\mathbf{w}_i' f_i(\mathbf{x})\right|^l\right)$$

by Fact 2.1.1, which proves the claimed bound since $\mathbf{w}$ and $\mathbf{w}'$ have the same distribution. □

In particular, it suffices to fix a choice of weights $\mathbf{w}$ and then bound the Rademacher process

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n}\left[\sup_{\mathbf{x}\in X}\left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i\mathbf{w}_i f_i(\mathbf{x})\right|^l\right].$$

2.3.2 Subgaussian processes

One useful fact about Rademacher processes is that it is a *subgaussian process*, a property which we will fully exploit in the next section.

Definition 2.3.3 (Subgaussian process)**.** A random process $(X_t)_{t\in T}$ equipped with a distance d is *subgaussian* if $\mathbf{E}[X_t] = 0$ for every $t \in T$ and

$$\mathbf{E}[\exp(\lambda(X_s - X_t))] \leq \exp(\lambda^2 d(s,t)^2/2)$$

for every $s, t \in T$ and $\lambda \geq 0$. Equivalently,

$$\mathbf{Pr}\{|X_s - X_t| \geq \lambda d(s,t)\} \leq K\exp(-\lambda^2/K)$$

for some universal constant K and any $s, t \in T$ and $\lambda \geq 0$.

For a Rademacher process $(X_t)_{t\in T}$, the *natural pseudo-metric* can be used as the distance d that makes $(X_t)_{t\in T}$ a subgaussian process.

Definition 2.3.4 (Natural pseudo-metric)**.** Let $(X_t)_{t\in T}$ be a random process. Then, the *natural pseudo-metric* associated with the random process is

$$d_X(s,t) \coloneqq \sqrt{\mathbf{E}|X_s - X_t|^2} = \|X_s - X_t\|_2.$$

2.3.3 Chaining and Dudley's inequality

We will now introduce Dudley's inequality, which is a powerful tool for bounding the suprema of subgaussian processes. Our exposition here largely follows [vH14].

To introduce Dudley's inequality, we consider the problem of bounding the first moment

$$\mathbf{E}\sup_{s\in T}|X_s - X_t|$$

for some fixed $t \in T$. Similar ideas will also allow us to bound higher moments.

For a subgaussian process $(X_t)_{t\in T}$ with a finite indexing set T, the subgaussian tail inequality of Definition 2.3.3 allows for a bound on the supremum with only a polylogarithmic dependence on the size $|T|$. Indeed, suppose we fix some $t \in T$. Then, for each fixed $s \in T$, we have that

$$\mathbf{Pr}\{|X_s - X_t| \geq \lambda d(s,t)\} \leq K \exp(-\lambda^2/K) \leq \frac{1}{100T}$$

by setting $\lambda = O(\sqrt{\log|T|})$. Then by a union bound over all the elements $s \in T$, we have that $|X_s - X_t| \leq O(\sqrt{\log|T|})d(s,t)$ simultaneously for every $s \in T$, with probability at least $99/100$. That is, with probability at least $99/100$, we have that

$$\sup_{s\in T}|X_s - X_t| \leq O(\sqrt{\log|T|}) \sup_{s\in T} d(s,t) \leq O(\sqrt{\log|T|})\operatorname{diam}(T)$$

where

$$\operatorname{diam}(T) = \sup_{s,t\in T} d(s,t)$$

denotes the *diameter* of the set T with respect to the distance d. A similar argument shows that the same bound holds in expectation, that is, $\mathbf{E}[\sup_{s\in T}|X_s - X_t|] \leq O(\sqrt{\log|T|})\operatorname{diam}(T)$.

While the $O(\sqrt{\log|T|})$ dependence is acceptable for small finite sets T, we are often interested in infinite sets T, for example the interval $[0,1]$ or the Euclidean ball. Thus, we will need to modify our strategy to handle these cases.

To make use of our previous result for finite sets T, we may consider approximating the infinite set T by a finite set N. For this, we make use of *nets*.

Definition 2.3.5 (ε-net)**.** A finite set N is an ε-net of a set T with distance d if for every $t \in T$, there exists $\pi(t) \in N$ such that $d(t,\pi(t)) \leq \varepsilon$.

If N is an ε-net for T, then we have that

$$\sup_{s\in T}|X_s - X_t| \leq \sup_{s\in T}|X_s - X_{\pi(s)}| + \sup_{s\in T}|X_{\pi(s)} - X_t|$$

so $\sup_{s\in T}|X_s - X_t|$ can now be controlled by the error from approximating T by N and the supremum of X_t over a finite set N. The latter can be handled by our previous union bound argument, so it suffices to consider the former term. Note that for any fixed s, we may expect the "net error" $\sup_{s\in T}|X_s - X_{\pi(s)}|$ to be at most roughly ε by the net construction.

At this point, one option is to directly consider a construction of a net N which allows the net error to be directly controlled, which can be accomplished under additional assumptions on T and d such as Lipschitzness [vH14]. Another elegant option, however, is to *iterate* this argument and introduce another net, say an $(\varepsilon/2)$-net N_2, which approximates T to a finer error $\varepsilon/2$. If $\pi_2 : T \to N_2$ is the mapping for N_2, then we have that

$$\sup_{s\in T}|X_s - X_{\pi(s)}| \leq \sup_{s\in T}|X_s - X_{\pi_2(s)}| + \sup_{s\in T}|X_{\pi_2(s)} - X_{\pi(s)}|.$$

We have now achieved a smaller "net error" by introducing another finite supremum term, and this process can now be repeated indefinitely. In general, we set $T_1 = \{t\}$ to be a $\operatorname{diam}(T)$-net

and T_i to be a $2^{-i}\operatorname{diam}(T)$-net for $i \in [I]$ so that

$$\sup_{s\in T}|X_s - X_t| \le \sup_{s\in T}|X_s - X_{\pi_{I+1}(s)}| + \sum_{i=1}^{I}\sup_{s\in T}|X_{\pi_i(s)} - X_{\pi_{i+1}(s)}|$$

Note that the finite supremum bound gives

$$\begin{aligned}\mathbf{E}\sup_{s\in T}|X_{\pi_i(s)} - X_{\pi_{i+1}(s)}| &\le O(2^{-i}\operatorname{diam}(T))\sqrt{\log|N_i \times N_{i+1}|}\\ &\le O(2^{-i}\operatorname{diam}(T))\sqrt{\log|N_{i+1}|}\end{aligned}$$

summing gives us the bound

$$\mathbf{E}\sup_{s\in T}|X_s - X_t| \le \mathbf{E}\sup_{s\in T}|X_s - X_{\pi_{I+1}(s)}| + O(\operatorname{diam}(T))\sum_{i=1}^{I}2^{-i}\sqrt{\log|N_{i+1}|}.$$

Finally, note that for T of any finite size, the first term vanishes for sufficiently large I, so we obtain the bound

$$\mathbf{E}\sup_{s\in T}|X_s - X_t| \le \sum_{i=1}^{\infty}O(2^{-i}\operatorname{diam}(T))\sqrt{\log|N_{i+1}|}$$

whenever T is finite. This bound extends to the setting of infinite T under mild assumptions such as separability of T, which gives us the result known as *Dudley's inequality* [Dud67].

Dudley's inequality is often stated in terms of *metric entropy* numbers $E(T, d, u)$ which denotes the minimal number of d-balls of radius u required to cover T. Furthermore, we can also write this bound as an integral, giving the bound

$$\mathbf{E}\sup_{s\in T}|X_s - X_t| \le O(1)\int_0^{\infty}\sqrt{\log E(T, d, u)}\,du$$

known as *Dudley's entropy integral*. In fact, a very similar argument also gives tail bounds, which we state in the following theorem.

Theorem 2.3.6 (Dudley's entropy integral, Theorem 8.1.6, [Ver18])**.** Let $(X_t)_{t\in T}$ be a subgaussian process with pseudo-metric $d_X(s,t) := \|X_s - X_t\|_2$. Let $E(T, d_X, u)$ denote the minimal number of d_X-balls of radius u required to cover T. Then, for every $z \ge 0$, we have that

$$\mathbf{Pr}\left\{\sup_{s,t\in T}|X_s - X_t| \ge C\left[\int_0^{\infty}\sqrt{\log E(T, d_X, u)}\,du + z\cdot\operatorname{diam}(T)\right]\right\} \le 2\exp(-z^2)$$

The way in which the entropy numbers $E(T, d_X, u)$ and the diameter $\operatorname{diam}(T)$ are bounded is heavily problem-dependent, and will be discussed further in the coming chapters.

In order to recover moment bounds for subgaussian processes, we can integrate the tail bound given by Theorem 2.3.6. This is executed in the following lemma.

Lemma 2.3.7 (Moment bounds)**.** Let Λ be the supremum of a subgaussian process with domain T and distance d_X. Let $\mathcal{E} := \int_0^{\infty}\sqrt{\log E(T, d_X, u)}\,du$ and $\mathcal{D} = \operatorname{diam}(X)$. Then, for $l \in \mathbb{N}$,

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)}[|\Lambda|^l] \le (2\mathcal{E})^l(\mathcal{E}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l$$

Proof. By Theorem 2.3.6, we have that

$$\mathbf{Pr}\left\{\Lambda \geq C\left[\int_0^\infty \sqrt{\log E(X, d_G, u)}\, du + z \cdot \mathrm{diam}(X)\right]\right\} \leq 2\exp(-z^2)$$

for a constant $C = O(1)$. Then,

$$\begin{aligned}
\mathbf{E}[(\Lambda/\mathcal{D})^l] &= l\int_0^\infty z^l\, \mathbf{Pr}\{\Lambda \geq z\mathcal{D}\}\, dz \\
&\leq (2\mathcal{E}/\mathcal{D})^{l+1} + l\int_{2\mathcal{E}/\mathcal{D}}^\infty z^l\, \mathbf{Pr}\{\Lambda \geq z\mathcal{D}\}\, dz \\
&\leq (2\mathcal{E}/\mathcal{D})^{l+1} + l\int_{2\mathcal{E}/\mathcal{D}}^\infty z^l\, \mathbf{Pr}\{\Lambda \geq \mathcal{E} + (z/2)\mathcal{D}\}\, dz \\
&\leq (2\mathcal{E}/\mathcal{D})^{l+1} + 2l\int_0^\infty z^l \exp(-z^2/4)\, dz \\
&\leq (2\mathcal{E}/\mathcal{D})^{l+1} + O(l)^{l/2}
\end{aligned}$$

so

$$\mathbf{E}[\Lambda^l] \leq (2\mathcal{E})^l(\mathcal{E}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l.$$

□

Part I

Oblivious Sketching

Chapter 3

High distortion embeddings for ℓ_p [WY23a]

Given the oblivious ℓ_2 subspace embedding result of Theorem 1.2.3, a natural question to ask is whether similar results exist for ℓ_p norms for $p \neq 2$. For $d = 1$, the more general question of whether the ℓ_p norm of a single vector can be preserved given a small number of linear measurements of the vector is well-studied in the *streaming* literature [SS02, BJKS04, IW05, Ind06], where for $p < 2$, $\tilde{\Theta}(\varepsilon^{-2})$ measurements is necessary and sufficient for approximation up to a factor of $(1+\varepsilon)$, while for $p > 2$, the ℓ_p norm cannot be approximated to within a constant factor unless $\Omega(n^{1-2/p})$ measurements are used. The latter result already prohibits a result of the form of Theorem 1.2.3 for $p > 2$, if the number of rows r of $\mathbf{S}$ must be subpolynomial in n. Thus, the key question is whether a theorem analogous to Theorem 1.2.3 is possible for $p < 2$.

One idea is to take inspiration from the proof of Theorem 1.2.3 and a classic streaming algorithm for ℓ_p norm estimation for vectors due to Indyk [Ind06]. In Theorem 1.2.3 for the case of $p = 2$, the sketch $\mathbf{S}$ can be taken to be a matrix with i.i.d. Gaussian entries [DG03], largely owing to the fact that the Gaussian distribution is 2-*stable*, that is, if $\mathbf{g} \in \mathbb{R}^n$ is an i.i.d. Gaussian vector and $\mathbf{y} \in \mathbb{R}^n$ is an arbitrary vector, then $\langle \mathbf{g}, \mathbf{y} \rangle$ is distributed as a single Gaussian random variable, scaled by $\|\mathbf{y}\|_2$. In fact, an analogous result is known for ℓ_p norms for $p < 2$:

Theorem 3.0.1 (Standard p-stable distributions [Ind06, Nol20])**.** For $0 < p \leq 2$, there exists a probability distribution $\mathcal{D}_p$ called the *standard p-stable distribution* such that if $\mathbf{g} \in \mathbb{R}^n$ has entries drawn i.i.d. from $\mathcal{D}_p$, then for any $\mathbf{y} \in \mathbb{R}^n$ $\langle \mathbf{g}, \mathbf{y} \rangle$ is distributed as $\|\mathbf{y}\|_p g$, for $g \sim \mathcal{D}_p$.

While Theorem 3.0.1 takes a step in the right direction, several challenges remain. For $p < 2$, the p-stable distributions $\mathcal{D}_p$ are *heavy-tailed* (unlike the 2-stable Gaussian distribution which enjoys sub-Gaussian tails), and thus in order to obtain $(1 \pm \varepsilon)$-approximate estimates with high probability, we need to take a *median* of independent measurements of $|\langle \mathbf{g}, \mathbf{y} \rangle|$ to approximate $\|\mathbf{y}\|_p$. However, the approximation that we seek, of the form of Definition 1.1.1, would take a *mean* of the measurements, which in turn results in either a much higher distortion, or a much higher number of rows r for the sketch $\mathbf{S}$.

In fact, it turns out that this loss for $p < 2$ is inherent for oblivious ℓ_p subspace embeddings, as shown by [WW19, WW22] in the following impossibility result:

Theorem 3.0.2 (Lower bounds for oblivious ℓ_p subspace embeddings, [WW19, WW22])**.** Sup-

pose that a distribution $\mathcal{D}$ over $r \times n$ matrices $\mathbf{S}$ satisfies, for any $\mathbf{A} \in \mathbb{R}^{n \times d}$,

$$\Pr_{\mathbf{S} \sim \mathcal{D}} \left\{ \text{for all } \mathbf{x} \in \mathbb{R}^d, \quad \|\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p \leq \kappa \|\mathbf{A}\mathbf{x}\|_p \right\} \geq \frac{99}{100}.$$

Then, the distortion κ is at least

$$\kappa = \Omega\left(\frac{1}{\frac{1}{d^{1/p}} \log^{2/p} r + \left(\frac{r}{n}\right)^{1/p - 1/2}} \right).$$

Note that typically, we seek $r = \text{poly}(d)$, which means that the distortion κ must be at least

$$\kappa = \Omega\left(\frac{d^{1/p}}{\log^{2/p} d} \right) = \tilde{\Omega}(d^{1/p})$$

and so the distortion must be at least polynomial in d. Thus, $(1+\varepsilon)$-approximations, or even $O(1)$-approximations, are not possible in this regime. On the other hand, the question of whether we can match the lower bound of Theorem 3.0.2 and design oblivious subspace embeddings with distortion $\kappa = \tilde{O}(d^{1/p})$ is a natural and interesting one.

The first known upper bounds for ℓ_p subspace embeddings for $p < 2$ were obtained by [SW11], who gave a construction with $r = \tilde{O}(d)$ rows and distortion $\kappa = \tilde{O}(d)$ for the case of $p = 1$. In fact, their sketch $\mathbf{S}$ is constructed analogously to Theorem 1.2.3 with the 2-stable Gaussian distribution replaced by the 1-stable distribution, also known as the Cauchy distribution. That is, $\mathbf{S}$ is just an appropriate scaling of the $r \times n$ matrix where each entry is drawn independently from the standard Cauchy distribution. Note that this result achieves a nearly optimal trade-off distortion for any $r = \text{poly}(d)$ rows by the lower bound of Theorem 3.0.2. While a dense Cauchy matrix is not as ideal to apply quickly, faster variants of this construction have been developed in subsequent works [MM13, WZ13, CDM+16, WW19, WW22].

With the resolution of the trade-offs for oblivious ℓ_1 subspace embeddings, the next natural question is to settle the analogous problem for $1 < p < 2$.

Question 3.0.3 ([WW19, WW22])**.** Do there exist oblivious ℓ_p subspace embeddings that achieve the guarantee of Definition 1.1.1 for the ℓ_p norm with $\kappa = \tilde{O}(d^{1/p})$ and $r = \text{poly}(d)$?

In fact, for a long time, the Question 3.0.3 was thought to be resolved, and many papers claimed constructions of oblivious ℓ_p subspace embeddings achieving a distortion of $\kappa = \tilde{O}(d^{1/p})$ [MM13, WZ13, WW19]. Unfortunately, all of these results relied on the existence of a certain *well-conditioned basis*, whose proof contained an error, and the revised proofs only achieves constructions with a distortion of $\kappa = \tilde{O}(d)$ [WW22] for any $p \in (1, 2)$. Thus, the resolution of Question 3.0.3 became a central open question in the study of randomized matrix approximation [WW22]. In the work [WY23a], we give a positive resolution to Question 3.0.3:

Theorem 3.0.4 (Nearly optimal oblivious ℓ_p subspace embeddings [WY23a])**.** Let $\mathbf{S}$ be an $r \times n$ matrix of i.i.d. p-stable random variables. There is an $r = \tilde{O}(d)$ such that for any $\mathbf{A} \in \mathbb{R}^{n \times d}$,

$$\mathbf{Pr}\left\{ \text{for all } \mathbf{x} \in \mathbb{R}^d, \quad \|\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p \leq \tilde{O}(d^{1/p}) \|\mathbf{A}\mathbf{x}\|_p \right\} \geq \frac{99}{100}$$

that is, $\mathbf{S}$ is an ℓ_p subspace embedding of $\mathbf{A}$ with distortion $\kappa = \tilde{O}(d^{1/p})$, with probability at least $99/100$.

We discuss our approach towards proving Theorem 3.0.4 in the rest of Chapter 3.

3.1 The question of well-conditioned bases

As alluded to previously, the central question is the existence of a well-conditioned basis. We start with a discussion of these objects.

For the ℓ_2 norm, every subspace admits an *orthogonal basis*, which is a basis for the subspace which exactly preserves the ℓ_2 norm. That is, for any matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, there exists a matrix $\mathbf{U} \in \mathbb{R}^{n\times d}$ such that for any $\mathbf{x} \in \mathbb{R}^d$, there exists $\mathbf{x}' \in \mathbb{R}^d$ such that $\mathbf{Ax} = \mathbf{Ux}'$, and furthermore, $\|\mathbf{Ux}\|_2 = \|\mathbf{x}\|_2$ for every $\mathbf{x} \in \mathbb{R}^d$. In other words, $\mathbf{U}$ is a *norm-preserving map* from $\mathbb{R}^d$ to $\mathbb{R}^n$. The existence of orthogonal bases plays a key role in the analyses of oblivious ℓ_2 subspace embeddings [NN13, Woo14]. However, for $p \neq 2$, exact analogues of orthogonal bases do not exist, in the sense that there does not necessarily exist a basis such that $\|\mathbf{Ux}\|_p = \|\mathbf{x}\|_p$. Thus, we must settle for an appropriately relaxed notion of "orthogonal bases" when working with subspaces of ℓ_p. One way to meaningfully define such an analogue was introduced by [DDH+09], based on a similar definition by [Cla05]:

Definition 3.1.1 ((α, β, p)-well-conditioned basis, Definition 3, [DDH+09])**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be rank d matrix, let $p \geq 1$, and let $q = p/(p-1)$ be the Hölder dual of p. Then, $\mathbf{U} \in \mathbb{R}^{n\times d}$ is an (α, β, p)-*well-conditioned basis* if (1) $\|\mathbf{U}\|_{p,p} \leq \alpha$ and (2) for any $\mathbf{z} \in \mathbb{R}^d$, $\|\mathbf{z}\|_q \leq \beta\|\mathbf{Uz}\|_p$.

Note that for ℓ_2, an orthogonal basis $\mathbf{U}$ corresponds to an $(\alpha, \beta, 2)$-well-conditioned basis with parameters $\alpha = d^{1/2}$ and $\beta = 1$. For ℓ_1, [SW11] showed that the well-known construction of Auerbach bases [Aue30] from the geometric functional analysis literature corresponds to an $(\alpha, \beta, 1)$-well-conditioned basis with $\alpha = d$ and $\beta = 1$. For $p \in (1, 2)$, however, the works of [MM13, WZ13, WW19] mistakenly claimed that Auerbach bases also give (α, β, p)-well-conditioned bases for $\alpha = d^{1/p}$ and $\beta = 1$, while they in fact only give $\alpha = d$ and $\beta = 1$ [WW22].

3.2 Relaxing linear bases to spanning sets

In fact, our techniques in [WY23a] do not give a construction for $(d^{1/p}, 1, p)$-well-conditioned basis. Instead, we show that by relaxing the notion of well-conditioned bases to well-conditioned *spanning sets*, we can obtain a construction that is sufficient to prove Theorem 3.0.4. More specifically, we show the following:

Theorem 3.2.1 ((α, β, p)-well-conditioned spanning set, [WY23a])**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $p \geq 1$. Then, there exists $\mathbf{U} \in \mathbb{R}^{n\times s}$ for $s = O(d)$ such that (1) $\|\mathbf{U}\|_{p,p} \leq s^{1/p}$ and (2) for any $\mathbf{x} \in \mathbb{R}^d$, there exists $\mathbf{z} \in \mathbb{R}^s$ such that $\mathbf{Ax} = \mathbf{Uz}$ and $\|\mathbf{z}\|_2 \leq \|\mathbf{Uz}\|_p$.

That is, we show that by relaxing the use of a basis, which is only allowed to contain d vectors, to a spanning set consisting of a slightly larger $O(d)$ vectors, we can obtain a spanning set which has properties that are just as good as a $(O(d^{1/p}), 1, p)$-well-conditioned basis. In fact, the well-conditioning guarantee is even better than the one in Definition 3.1.1, since the Hölder

conjugate q of p is greater than 2 for $p < 2$, so we in fact have the guarantee that

$$\|\mathbf{z}\|_q \leq \|\mathbf{z}\|_2 \leq \|\mathbf{U}\mathbf{z}\|_p.$$

Towards our result, we will need the following result on the construction of *volumetric spanners* [HK16], which, given a set of vectors, is a small subset of vectors such that the minimum volume ellipsoid enclosing the subset is also the minimum volume ellipsoid enclosing the entire set of n vectors. These objects can also be stated as *coresets for Löwner–John ellipsoids*. Related and improved constructions are also given in [KY05, Tod16, BMV23].

Theorem 3.2.2 (Theorem 1.1, [HK16])**.** Let $\mathcal{K} \in \mathbb{R}^d$ be compact. There exists $S \subseteq \mathcal{K}$ with $|S| \leq 12d$ such that $\mathcal{K} \subseteq \mathcal{E}(S)$, where $\mathcal{E}(S) = \left\{\sum_{\mathbf{v}\in S} \mathbf{v} \cdot \mathbf{x}_i \in \mathbb{R}^d : \mathbf{x} \in \mathbb{R}^{|S|}, \|\mathbf{x}\|_2 \leq 1\right\}$.

From the guarantees of Theorem 3.2.2, we immediately the following lemma about spanning sets for subspaces of ℓ_p.

Lemma 3.2.3 (Well-conditioned spanning sets for subspaces of ℓ_p)**.** Let $p \in (0, \infty)$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. There exists $\mathbf{R} \in \mathbb{R}^{d\times s}$ for $s = O(d)$ such that $\|\mathbf{A}\mathbf{R}\mathbf{e}_i\|_p \leq 1$ for every $i \in [s]$, and for any $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{A}\mathbf{x}\|_p \leq 1$, there exists $\mathbf{y} \in \mathbb{R}^s$ such that $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{R}\mathbf{y}$ and $\|\mathbf{y}\|_2 \leq 1$.

Proof. We take $\mathcal{K} = \{\mathbf{A}\mathbf{x} : \|\mathbf{A}\mathbf{x}\|_p \leq 1\}$. Then by Theorem 3.2.2, there is a subset of at most $s \leq 12d$ vectors, say the columns of $\mathbf{A}\mathbf{R}$ for some $\mathbf{R} \in \mathbb{R}^{d\times s}$, such that any $\mathbf{v} \in \mathcal{K}$ can be written as $\mathbf{v} = \mathbf{A}\mathbf{R}\mathbf{y}$ with $\|\mathbf{y}\|_2 \leq 1$. □

Given the above lemma, the proof of Theorem 3.2.1 is immediate:

Proof of Theorem 3.2.1. We simply translate the guarantees of Lemma 3.2.3 into that of Theorem 3.2.1. First, we take $\mathbf{U} = \mathbf{A}\mathbf{R}$, where $\mathbf{R}$ is given by Lemma 3.2.3. Then, the entrywise ℓ_p norm of $\mathbf{U}$ is bounded since

$$\|\mathbf{U}\|_{p,p}^p = \sum_{j=1}^{s} \|\mathbf{U}\mathbf{e}_j\|_p^p = \sum_{j=1}^{s} \|\mathbf{A}\mathbf{R}\mathbf{e}_j\|_p^p \leq s.$$

Next, let $\mathbf{x} \in \mathbb{R}^d$ satisfy $\|\mathbf{A}\mathbf{x}\|_p = 1$. Then, by Lemma 3.2.3, we may identify a $\mathbf{z} \in \mathbb{R}^s$ such that $\mathbf{A}\mathbf{x} = \mathbf{U}\mathbf{z}$ and

$$\|\mathbf{z}\|_2 \leq 1 \leq \|\mathbf{A}\mathbf{x}\|_p = \|\mathbf{U}\mathbf{z}\|_p.$$

The result for general $\mathbf{x} \in \mathbb{R}^d$ follows by scaling. □

3.3 Proof of Theorem 3.0.4

Finally, with the construction of well-conditioned spanning sets (Theorem 3.2.1) in hand, we obtain a construction of nearly optimal oblivious ℓ_p subspace embeddings. We give a simple proof based on the work of [SW11] for the ℓ_1 norm, which uses a dense p-stable embedding $\mathbf{S}$. This embedding has a much slower running time to apply to the matrix $\mathbf{A}$, but gives a simple proof and still gives a nearly optimal trade-off between the embedding dimension r and the distortion κ. Constructions with faster running time can be obtained by combining our ideas with results in [WZ13, WW19, WW22], but we omit the details for sake of simplicity.

Proof of Theorem 3.0.4. We take $\mathbf{S} \in \mathbb{R}^{r \times d}$ to be drawn with i.i.d. p-stable random variables [Nol20], scaled by $C/r^{1/p}$ for some large enough constant C. For every $(i,j) \in [r] \times [d]$, $\mathbf{e}_i^\top \mathbf{S}\mathbf{U}\mathbf{e}_j$ is distributed as $C\|\mathbf{U}\mathbf{e}_j\|_p / r^{1/p}$ times a p-stable variable $X_{i,j}$, by definition of p-stable variables. With probability at least $1 - 1/\operatorname{poly}(rd)$, $|X_{i,j}|$ is at most $\operatorname{poly}(rd)$, so by a union bound over all rd choices of (i,j), this is true for every $(i,j) \in [r] \times [d]$. Call this event $\mathcal{E}$. Conditioned on this event, the expectation of $|X_{i,j}|$ is $O(\log(rd))$, so by linearity of expectation, we have

$$\mathbf{E}\Big[\|\mathbf{S}\mathbf{U}\|_{p,p}^p | \mathcal{E}\Big] = \sum_{i=1}^{r}\sum_{j=1}^{d} \mathbf{E}\big[\big|\mathbf{e}_i^\top \mathbf{S}\mathbf{U}\mathbf{e}_j\big|^p | \mathcal{E}\big] \le O(1)\sum_{i=1}^{r}\sum_{j=1}^{d} \frac{\|\mathbf{U}\mathbf{e}_j\|_p^p}{r} = O(\|\mathbf{U}\|_{p,p}^p \log(rd)).$$

By Markov's inequality, this bound holds up to constant factors with probability at least $199/200$. We condition on this event. Then, for any $\mathbf{x} \in \mathbb{R}^{n \times d}$, we write $\mathbf{A}\mathbf{x} = \mathbf{U}\mathbf{z}$ for $\mathbf{z}$ promised by Theorem 3.2.1, so that

$$\begin{aligned}
\|\mathbf{S}\mathbf{U}\mathbf{z}\|_p^p &= \sum_{i=1}^{n} \big|\mathbf{e}_i^\top \mathbf{S}\mathbf{U}\mathbf{z}\big|^p \\
&\le \|\mathbf{z}\|_q^p \sum_{i=1}^{n} \big|\mathbf{e}_i^\top \mathbf{S}\mathbf{U}\big|_p^p && \text{Hölder's inequality} \\
&\le \|\mathbf{z}\|_2^p \|\mathbf{S}\mathbf{U}\|_{p,p}^p \\
&\le \|\mathbf{U}\mathbf{z}\|_p^p \|\mathbf{S}\mathbf{U}\|_{p,p}^p \\
&\le \|\mathbf{U}\mathbf{z}\|_p^p O(\|\mathbf{U}\|_{p,p}^p \log(rd)) \\
&\le O(d\log(rd))\|\mathbf{U}\mathbf{z}\|_p^p.
\end{aligned}$$

Taking pth roots gives the upper inequality.

For the lower inequality, we use the following concentration lemma [WW19, Lemma 2.12]:

Lemma 3.3.1 (Lemma 2.12 of [WW19])**.** Let $\{X_i\}_{i=1}^n$ be independent p-stable random variables. Then for sufficiently large n and T,

$$\mathbf{Pr}\left\{\sum_{i=1}^{n} |X_i|^p \ge L_p n \log\frac{n}{\log T}\right\} \ge 1 - \frac{1}{T}$$

for some constant L_p.

For every $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{A}\mathbf{x}\|_p = 1$, $\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p$ is the sum of r independent p-stable random variables, raised to the p and scaled by r. We then apply the above lemma with $n = r$ and $T = \exp(r)$ to conclude that for every $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{A}\mathbf{x}\|_p = 1$, $\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p \ge 1$ with probability at least $1 - \exp(-r)$, by choosing our constant C large enough. By a standard net argument (see, e.g., [SW11]), this is true for every $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{A}\mathbf{x}\|_p = 1$. This in turn implies the lower tail inequality for every $\mathbf{x} \in \mathbb{R}^d$ by scaling. □

Chapter 4

Low distortion embeddings for ℓ_1 [LWY21]

In Chapter 3, we studied algorithms for oblivious ℓ_p subspace embeddings with distortion κ on the order of $\mathrm{poly}(d)$, as the lower bound of Theorem 3.0.2 prohibited a construction with smaller distortion, if we insist on $r = \mathrm{poly}(d)$. However, if we are allowed to make r as large as $\exp(\mathrm{poly}(d))$, then the lower bound of Theorem 3.0.2 no longer gives a lower bound, and we can hope for a distortion of $\kappa = (1+\varepsilon)$. Indeed, [WW19, WW22] studied the question of whether $(1+\varepsilon)$ approximations are possible if we allow for superpolynomial dependencies on d, and showed that if r is doubly exponential, i.e. $r = \exp(\exp(\mathrm{poly}(d)))$, then a dense Cauchy embedding (similarly to that used in [SW11]) admits oblivious ℓ_1 subspace embeddings with $(1+\varepsilon)$ distortion. However, this leads to an *exponential* gap in the bound on r. A natural question is to resolve this gap:

Question 4.0.1 ([WW19, WW22])**.** Do there exist oblivious ℓ_p subspace embeddings that achieves the guarantee of Definition 1.1.1 for the ℓ_p norm with $\kappa = (1+\varepsilon)$ and $r = \exp(\mathrm{poly}(d, \varepsilon^{-1}))$?

In [LWY21], we study Question 4.0.1 and answer it affirmatively for $p = 1$ with the following theorem:

Theorem 4.0.2 ($(1+\varepsilon)$ oblivious ℓ_1 subspace embeddings [LWY21])**.** There exists a distribution over $r \times n$ matrices $\mathbf{S}$ for $r = \exp(\tilde{O}(d/\varepsilon))$ such that for any $\mathbf{A} \in \mathbb{R}^{n\times d}$,

$$\mathbf{Pr}\big\{\text{for all } \mathbf{x} \in \mathbb{R}^d, \quad \|\mathbf{Ax}\|_1 \le \|\mathbf{SAx}\|_1 \le (1+\varepsilon)\|\mathbf{Ax}\|_1\big\} \ge \frac{99}{100}$$

that is, $\mathbf{S}$ is an ℓ_1 subspace embedding of $\mathbf{A}$ with distortion $\kappa = (1+\varepsilon)$, with probability at least $99/100$.

Our techniques developed in this work have been further developed in [MOW23] to design streaming algorithms for logistic regression and ℓ_1 regression. The remainder of Chapter 4 will be devoted to proving Theorem 4.0.2.

4.1 Overview of sketch construction and analysis

Our sketch $\mathbf{S}$ for proving Theorem 4.0.2 requires a number of novel ideas over prior constructions. Instead of using the typical approach for oblivious ℓ_1 subspace embeddings based on Cauchy

sketches [SW11, WW19, WW22], we instead start with the M-sketch of [CW15a], which is based on classic techniques of hashing and subsampling from the streaming literature [IW05].

Let us now discuss the original construction of the M-sketch and its analysis, as well as its shortcomings that we will need to overcome when proving Theorem 4.0.2. The M-sketch matrix $\mathbf{S}$ first samples the rows of $\mathbf{A}$ at $O(\log n)$ geometrically decreasing scales of sampling probabilities ranging from $p = 1$ to $p = 1/n$. We will refer to each of these scales as *levels*, and denote the sampling matrix at level h by $\mathbf{S}^{(h)}$. At level h, $\mathbf{S}^{(h)}$ will sample each row $i \in [n]$ with probability $p_h = B^{-h}$ for some branching factor B to be chosen later. For each of these sampling levels, we apply a CountSketch matrix [CW13] $\mathbf{C}^{(h)}$, given by the following definition:

Definition 4.1.1 (CountSketch [CW13])**.** The $r \times n$ CountSketch matrix is a random matrix associated with a random hash function $H : [n] \to [r]$ and random signs $\Lambda_i \sim \{\pm 1\}$ for $i \in [n]$, so that for each $i \in [n]$, $\mathbf{C}_{H(i),i} = \Lambda_i$.

Finally, the M-sketch construction $\mathbf{S}$ takes the vertical concatenation of the matrices $\mathbf{C}^{(h)}\mathbf{S}^{(h)}$.

4.1.1 Sketching a single vector

Let us first discuss what this sketch does for a fixed vector $\mathbf{y} \in \mathbb{R}^n$. We focus on showing that our construction does not expand the ℓ_1 norm too much, i.e., $\|\mathbf{S}\mathbf{y}\|_1 \leq (1+\varepsilon)\|\mathbf{y}\|_1$, since the other inequality $\|\mathbf{S}\mathbf{y}\|_1 \geq (1-\varepsilon)\|\mathbf{y}\|_1$ will turn out to be much simpler. As we justify later, we can assume that $\mathbf{y}$ is a vector with m coordinates of ones and $n - m$ coordinates of zeros without loss of generality. We first consider three cases on the sampling probability p used in the M-sketch. Two of these are "easy", while the last is a challenge we will need to overcome. In the first case, p is much smaller than $1/m$, which means that none of the m nonzero entries of $\mathbf{y}$ are sampled. In this case, there is no contribution towards $\|\mathbf{S}\mathbf{y}\|_1$ in this sampling level, so the analysis is simple. On the other hand, if p is much larger than $1/(\varepsilon^2 m)$, then we will have good concentration by Chernoff bounds and thus the sampled mass is $\|\mathbf{S}^{(h)}\mathbf{y}\|_1 = (1 \pm \varepsilon)\|\mathbf{y}\|_1$. Then when we apply a CountSketch matrix $\mathbf{C}^{(h)}$ on the sampled coordinates $\mathbf{S}^{(h)}\mathbf{y}$, we will perfectly hash the sampled coordinates if this number is small relative to the number of hash buckets, or introduce many collisions if the number of sampled coordinates is much larger than the number of hash buckets. In the former case, we continue to preserve the ℓ_1 norm so $\left\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{y}\right\|_1 = (1 \pm \varepsilon)\left\|\mathbf{S}^{(h)}\mathbf{y}\right\|_1$, while in the latter case, the collisions will cause a substantial reduction in ℓ_1 mass due to cancellations due to the random signs of CountSketch. Both of these will be amenable to analysis.

The last case is the challenging case, where the sampling probability p is larger than $1/m$ so that we will sample some of the nonzero entries of $\mathbf{y}$ with constant probability, but smaller than $1/(\varepsilon^2 m)$ so that we cannot expect $(1 \pm \varepsilon)$ approximation with high enough probability. We will refer to this as the "badly concentrated" levels. When such badly concentrated levels exist, then we cannot hope for our sketch to achieve $(1 \pm \varepsilon)$ approximation. However, for any fixed sampling probability p, one can define a hard instance for this algorithm by taking the input vector to be the binary vector supported on $m = \Theta(1/p)$ entries. The crucial idea for getting around this problem is to *randomize the sampling probability itself*. To implement this idea, we draw a uniformly random offset $u \sim [0, 1]$ and take the sampling probability at level h to now be $p_h = B^{-(u+h)}$. Then, for any fixed support size m, the probability that any fixed p_h is between $1/m$ and $1/(\varepsilon^2 m)$

is at most

$$\mathbf{Pr}\{p_h \in [1/m, 1/(\varepsilon^2 m)]\} = \mathbf{Pr}\{-(u+h) \in [0, \log_B \varepsilon^{-2}] - \log_B m\} \leq \log_B \varepsilon^{-2}.$$

Now if we take $B = (\varepsilon^{-2})^{1/\delta}$, then this probability is at most δ, so the expected contribution from badly concentrated levels is now at most $\delta\|\mathbf{y}\|_1$. We can then set $\delta = \varepsilon$ so that the expected contribution is at most an ε fraction of the entire mass, and this idea is sufficient to carry the proof out when applying the sketch to one vector. At this point, we may formally introduce the construction we use to prove Theorem 4.0.2, which we call *random boundary M-sketch*.

Definition 4.1.2 (Random boundary M-sketch)**.** Let $\mathbf{C}^{(0)}$ be a $N_0 \times n$ CountSketch matrix (Definition 4.1.1) and for each $h \in [h_{\max}]$, let $\mathbf{C}^{(h)}$ be a $N \times n$ CountSketch matrix. Let $u \sim [0, 1]$ be uniformly random, and for each $h \in [h_{\max}]$, let $p_h = B^{-(u+h-1)}$ and let $\mathbf{S}^{(h)}$ be the diagonal sampling matrix with $\mathbf{S}^{(h)}_{i,i}$ set to $1/p_h$ with probability p_h and 0 otherwise. Then, the random boundary M-sketch matrix is defined as the vertical concatenation

$$\mathbf{S} \coloneqq \begin{pmatrix} \mathbf{C}^{(0)} \\ \mathbf{C}^{(1)}\mathbf{S}^{(1)} \\ \mathbf{C}^{(2)}\mathbf{S}^{(2)} \\ \vdots \\ \mathbf{C}^{(h_{\max})}\mathbf{S}^{(h_{\max})} \end{pmatrix}$$

4.1.2 Extension to subspaces

We now turn to extending the analysis to the case of general d. A common technique in the sketching literature is to extend an analysis for a single vector to a whole d-dimensional subspace through the use of a net argument, in which the single vector analysis is applied with failure probability $\delta = \exp(-d)$. However, this is a problem in our case since the dependence on the failure rate is exponential in the analysis of badly concentrated levels and thus this would still lead to a doubly exponential dependence in d, which is the original bound we sought to improve. On the other hand, the analysis of badly concentrated levels is the only place in which this problem occurs; in all other parts of the analysis, a union bound is sufficient.

A second key insight we need to get around this problem is that when analyzing the badly concentrated levels, we can apply the analysis just once to the vector of ℓ_1 *sensitivities*, that is, the vector $\boldsymbol{\sigma}^1(\mathbf{A}) \in \mathbb{R}^n$ given by

$$\boldsymbol{\sigma}^1_i(\mathbf{A}) \coloneqq \sup_{\mathbf{Ax} \neq 0} \frac{|[\mathbf{Ax}](i)|}{\|\mathbf{Ax}\|_1}. \tag{4.1}$$

The ith ℓ_1 sensitivity of $\mathbf{A}$ captures the largest value that $\mathbf{Ax}$ can take on the ith coordinate ranging over all ℓ_1 unit vectors $\mathbf{Ax}$, and thus bounding badly concentrated levels of the vector of sensitivities *simultaneously* bounds the badly concentrated levels of every column space vector $\mathbf{Ax}$. Furthermore, it is known that $\|\boldsymbol{\sigma}^1(\mathbf{A})\|_1 \leq d$ [WY23c] and thus we only incur an additional factor of d in the error. This can be handled by replacing ε by ε/d, which only leads to a singly exponential dependence on d, rather than doubly exponential.

4.2 No expansion

Our goal in this section is to show that $\|\mathbf{SAx}\|_1 \le (1+\varepsilon)\|\mathbf{Ax}\|_1$ simultaneously for every $\mathbf{x} \in \mathbb{R}^d$, with probability at least $1-\delta$. We first introduce some notation.

Definition 4.2.1 (Sensitivity weight classes)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Then, for each $q \in \mathbb{N}$, we let $W^q(\mathbf{A}) \subseteq [n]$ denote the set

$$W^q(\mathbf{A}) \coloneqq \left\{i \in [n] : \boldsymbol{\sigma}^1_i(\mathbf{A}) \in (2^{-q}, 2^{1-q}]\right\}$$

and let $\mathbf{A}^{(q)}$ and $\boldsymbol{\sigma}^{(q)}$ denote the restriction of $\mathbf{A}$ and $\boldsymbol{\sigma}^1$ to the rows indexed by $W^q(\mathbf{A})$, respectively.

We decompose the sketch by the sampling level h and the sensitivity weight class q as

$$\|\mathbf{SAx}\|_1 \le \|\mathbf{C}^{(0)}\mathbf{Ax}\|_1 + \sum_{q=1}^{\infty}\sum_{h=1}^{h_{\max}}\left\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\right\|_1$$

We will then bound this quantity by casework on h and q, and in particular, by casing on the expected number of sampled elements $p_h|W^q(\mathbf{A})|$.

4.2.1 Bounding badly concentrated levels

Lemma 4.2.2. Suppose that $0 < a < b$ are such that $\log_B(b/a) \le 1$. Then with probability at least $1-\delta$, we have that simultaneously for every $\mathbf{x} \in \mathbb{R}^d$,

$$\sum_{q=1}^{\infty}\sum_{h=1}^{h_{\max}}\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 \cdot \mathbb{1}\{p_h|W^q(\mathbf{A})| \in [a,b]\} \le \frac{2d}{\delta}\log_B\frac{b}{a}\cdot\|\mathbf{Ax}\|_1.$$

Proof. By the definition of sensitivities (4.1), we have that for every h and q,

$$\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 \le \|\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 \le \|\mathbf{Ax}\|_1 \cdot \|\mathbf{S}^{(h)}\boldsymbol{\sigma}^{(q)}\|_1.$$

Note that for each q, there are at most two choices of $h \in [h_{\max}]$ such that

$$\Pr_{u\sim[0,1]}\{p_h|W^q(\mathbf{A})| \in [a,b]\} = \Pr_{u\sim[0,1]}\{-(u+h-1) \in [\log_B a, \log_B b] - \log_B|W^q(\mathbf{A})|\} > 0$$

since $[\log_B a, \log_B b]$ is an interval of length at most 1. In this case, this probability is bounded by $\log_B(b/a)$. Then,

$$\begin{aligned}\mathbf{E}\left[\sum_{q=1}^{\infty}\sum_{h=1}^{h_{\max}}\|\mathbf{S}^{(h)}\boldsymbol{\sigma}^{(q)}\|_1 \cdot \mathbb{1}\{p_h|W^q(\mathbf{A})| \in [a,b]\}\right] &\le \sum_{q=1}^{\infty}\mathbf{E}\left[\|\mathbf{S}^{(h)}\boldsymbol{\sigma}^{(q)}\|_1\right]\cdot 2\log_B(b/a)\\ &\le 2d\log_B(b/a).\end{aligned}$$

Thus by Markov's inequality, this quantity is at most $2d\log_B(b/a)/\delta$ with probability at least $1-\delta$. □

The above lemma allows us to bound all levels below some threshold expectation $p_h|W^q(\mathbf{A})| < m_{\text{crowd}}$.

Lemma 4.2.3. Let $m_{\text{crowd}} \geq \delta/h_{\max}q_{\max}$ be some threshold and suppose

$$B \geq \exp\left(\frac{2d}{\delta\varepsilon}\log\frac{h_{\max}q_{\max}m_{\text{crowd}}}{\delta}\right)$$

Then with probability at least $1-2\delta$, we have for all $\mathbf{x} \in \mathbb{R}^d$ that

$$\sum_{h\in[h_{\max}]}\sum_{q\in[q_{\max}]}\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1\mathbb{1}\{p_h|W^q(\mathbf{A})| \in [0, m_{\text{crowd}})\} \leq 2\varepsilon\|\mathbf{A}\mathbf{x}\|_1.$$

Proof. We case on $p_h|W^q(\mathbf{A})|$ by intervals $[0, \delta/h_{\max}q_{\max})$ and $[\delta/h_{\max}q_{\max}, m_{\text{crowd}})$.

- **Dead levels**: First consider the h for which $p_h|W^q(\mathbf{A})| < \delta/h_{\max}q_{\max}$. In this case, the probability that we sample any row $i \in W^q(\mathbf{A})$ is at most $p_h|W^q(\mathbf{A})| < \delta/h_{\max}q_{\max}$ by a union bound over the $|W^q(\mathbf{A})|$ rows. Then by a further union bound over all pairs $(h,q) \in [h_{\max}] \times [q_{\max}]$, this category of levels contributes no mass with probability at least $1-\delta$.
- **Badly concentrated levels**: Consider the subsampling levels with $p_h|W^q(\mathbf{A})| \in [\delta/h_{\max}q_{\max}, m_{\text{crowd}})$. Then, B is chosen large enough such that by Lemma 4.2.2, the contribution from these levels is at most $\varepsilon\|\mathbf{A}\mathbf{x}\|_1$ simultaneously for every $\mathbf{x} \in \mathbb{R}^d$, with probability at least $1-\delta$.

We thus conclude by a union bound over the above three events. □

4.2.2 Bounding well-concentrated levels

For each weight class q, we will bound exactly one sampling level h by $(1+\varepsilon)\|\mathbf{A}^{(q)}\mathbf{x}\|_1$ by showing that the sampling $\mathbf{S}^{(h)}$ concentrates.

Lemma 4.2.4. Let $a > 0$. For each q, let $h_q \in [h_{\max}]$ such that $p_h|W^q(\mathbf{A})| \geq a$. Let

$$X \coloneqq \sum_{q=1}^{\infty}\|\mathbf{S}^{(h_q)}\mathbf{A}^{(q)}\mathbf{x}\|_1$$

Then, with probability at least $1-\exp(-a\varepsilon^2/8d)$, we have $X \leq \mathbf{E}[X] + \varepsilon\|\mathbf{A}\mathbf{x}\|_1$.

Proof. We will show that at this level, $\mathbf{S}^{(h_q)}$ performs sensitivity sampling. Note that

$$|[\mathbf{A}^{(q)}\mathbf{x}](i)| \leq \|\mathbf{A}\mathbf{x}\|_1 \cdot \max_{j\in W^q(\mathbf{A})}\boldsymbol{\sigma}_j^1(\mathbf{A}) \leq 2^{1-q}\|\mathbf{A}\mathbf{x}\|_1$$

for each $i \in W^q(\mathbf{A})$, by definition of sensitivities (4.1). We also have that $d \geq \|\boldsymbol{\sigma}^1(\mathbf{A})\|_1 \geq 2^{-q}|W^q(\mathbf{A})|$ so

$$p_{h_q} \geq \frac{a}{|W^q(\mathbf{A})|} \geq \frac{a}{d}\cdot 2^{-q}.$$

Then, the variance of X is

$$\begin{aligned}\mathbf{Var}(X) &= \sum_{q=1}^{\infty} \sum_{i \in W^q(\mathbf{A})} p_{h_q} \frac{|[\mathbf{A}^{(q)}\mathbf{x}](i)|^2}{p_{h_q}^2} = \sum_{q=1}^{\infty} \sum_{i \in W^q(\mathbf{A})} \frac{|[\mathbf{A}^{(q)}\mathbf{x}](i)|^2}{p_{h_q}} \\ &\leq \sum_{q=1}^{\infty} \sum_{i \in W^q(\mathbf{A})} \frac{2d}{a} \|\mathbf{A}\mathbf{x}\|_1 \cdot |[\mathbf{A}^{(q)}\mathbf{x}](i)| \leq \frac{2d}{a} \|\mathbf{A}\mathbf{x}\|_1^2\end{aligned}$$

and each term is bounded by $|[\mathbf{A}^{(q)}\mathbf{x}](i)|/p_{h_q} \leq (2d/a)\|\mathbf{A}\mathbf{x}\|_1$. Then by Bernstein's inequality,

$$\begin{aligned}\mathbf{Pr}\{X \geq \mathbf{E}[X] + \varepsilon \|\mathbf{A}\mathbf{x}\|_1\} &\leq \exp\left(-\frac{1}{2} \frac{\varepsilon^2 \|\mathbf{A}\mathbf{x}\|_1^2}{(2d/a)\|\mathbf{A}\mathbf{x}\|_1(\|\mathbf{A}\mathbf{x}\|_1 + \varepsilon\|\mathbf{A}\mathbf{x}\|_1/3)}\right) \\ &\leq \exp\left(-\frac{a\varepsilon^2}{8d}\right). \qquad \square\end{aligned}$$

4.2.3 Bounding oversampled levels

Lemma 4.2.5. Let $a > 0$. Let $p_h|W^q(\mathbf{A})| \geq a$. Then, for each $\mathbf{x} \in \mathbb{R}^d$, with probability at least $1 - 2N\exp(-a/3N) - \delta$, we have that

$$\left\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\right\|_1 \leq \frac{2\sqrt{2}d\sqrt{N}}{\sqrt{a}} \sqrt{\log(N/\delta)} \cdot \|\mathbf{A}\mathbf{x}\|_1.$$

Proof. Let $\mu = p_h|W^q(\mathbf{A})| \geq a$. By Chernoff's bound, the probability that a fixed CountSketch hash bucket in level h gets more than $X \geq 2\mu/N$ elements from $W^q(\mathbf{A})$ is at least

$$\mathbf{Pr}\left\{\left|X - \frac{\mu}{N}\right| \leq \frac{\mu}{N}\right\} \geq 1 - 2\exp\left(-\frac{\mu/N}{3}\right) = 1 - 2\exp\left(-\frac{\mu}{3N}\right)$$

By a union bound over the N buckets, this is true for every bucket with probability at least $1 - 2N\exp(-\mu/3N)$. We condition on this event. Then by Hoeffding's bound, the inner product of m elements $\{a_i\}_{i=1}^m$ bounded by 1 with random signs $s_i \sim \{\pm 1\}$ is bounded by

$$\mathbf{Pr}\left\{\left|\sum_{i=1}^m s_i a_i\right| > \sqrt{m\log(N/\delta)}\right\} \leq \frac{\delta}{N}.$$

Now let X_i denote the number of elements sampled from $W^q(\mathbf{A})$ in the ith hash bucket. Then, applying the above Hoeffding bound gives

$$\begin{aligned}\left|[\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}](i)\right| &\leq \frac{2^{1-q}}{p_h} \cdot \|\mathbf{A}\mathbf{x}\|_1 \sqrt{X_i \log(N/\delta)} \\ &\leq \frac{2^{1-q}|W^q(\mathbf{A})|}{p_h|W^q(\mathbf{A})|} \cdot \|\mathbf{A}\mathbf{x}\|_1 \sqrt{\frac{2\mu}{N}\log(N/\delta)} \\ &\leq \frac{2d}{\mu} \cdot \|\mathbf{A}\mathbf{x}\|_1 \sqrt{\frac{2\mu}{N}\log(N/\delta)}\end{aligned}$$

By a union bound over N buckets, with probability at least $1-\delta$, the above bound holds for all N buckets. Summing over the N buckets gives

$$\left\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\right\|_1 \leq \frac{2\sqrt{2}d\sqrt{N}}{\sqrt{\mu}}\sqrt{\log(N/\delta)}\cdot\|\mathbf{A}\mathbf{x}\|_1. \qquad \square$$

4.2.4 Bounding tiny levels

Lemma 4.2.6. Let $q_{\max} \geq \log(2nh_{\max}/\delta\varepsilon)$. Then with probability at least $1-\delta$, it holds for all $\mathbf{x}\in\mathbb{R}^d$ that

$$\sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\|\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 \leq \varepsilon\|\mathbf{A}\mathbf{x}\|_1.$$

Proof. For the weight classes $q > q_{\max}$, the total sensitivity mass contribution is bounded by

$$\sum_{q>q_{\max}}\|\boldsymbol{\sigma}^{(q)}\|_1 \leq \sum_{q>q_{\max}} 2^{1-q}|W^q(\mathbf{A})| \leq \frac{\delta\varepsilon}{nh_{\max}}\sum_{q>q_{\max}}|W^q(\mathbf{A})| \leq \frac{\delta\varepsilon}{h_{\max}}.$$

Then in expectation, the sum of the sampled and scaled sensitivity scores is bounded by

$$\mathbf{E}\left(\sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\|\mathbf{S}^{(h)}\boldsymbol{\sigma}^{(q)}\|_1\right) = \sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\|\boldsymbol{\sigma}^{(q)}\|_1 \leq \sum_{h\in[h_{\max}]}\frac{\delta\varepsilon}{h_{\max}} = \delta\varepsilon.$$

Then with probability at least $1-\delta$, the above sum is at most ε. We condition on this event. Then, for all $\mathbf{x}\in\mathbb{R}^d$,

$$\begin{aligned}\sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\|\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1 &\leq \|\mathbf{A}\mathbf{x}\|_1\sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\sum_{i\in W^q(\mathbf{A})}[\mathbf{S}^{(h)}\boldsymbol{\sigma}^1(\mathbf{A})](i)\\ &\leq \|\mathbf{A}\mathbf{x}\|_1\sum_{h\in[h_{\max}]}\sum_{q>q_{\max}}\|\mathbf{S}^{(h)}\boldsymbol{\sigma}^1(\mathbf{A})\|_1 \leq \varepsilon\|\mathbf{A}\mathbf{x}\|_1\end{aligned}$$

as desired. $\square$

4.2.5 Net argument

In this section, we collect the bounds obtained in previous sections and conclude with a net argument.

Lemma 4.2.7. Let the randomized boundary M-sketch $\mathbf{S}$ satisfy the hypotheses of Lemmas 4.2.3 and 4.2.6. Let $\alpha\in(0,1)$. Let

$$m_{\text{crowd}} \geq \frac{8d}{\varepsilon^2}\log\frac{1}{\alpha} + \frac{N_0}{B}\frac{8d^2q_{\max}^2}{\varepsilon^2}\log\frac{N_0q_{\max}}{\alpha} + \frac{N}{B}\frac{8d^2q_{\max}^2}{\varepsilon^2}\log\frac{N_0q_{\max}}{\alpha}$$

There is an event with probability $1-3\delta$ such that conditioned on this event, for every $\mathbf{x}\in\mathbb{R}^d$,

$$\mathbf{Pr}(\|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \leq (1+5\varepsilon)\|\mathbf{A}\mathbf{x}\|_1) \geq 1-2\alpha.$$

Proof. By Lemma 4.2.6, the contribution from weight classes $q > q_{\max}$ is at most $\varepsilon\|\mathbf{Ax}\|_1$ with probability at least $1-\delta$. We let this event be $\mathcal{E}_1$ and restrict our attention to $q \le q_{\max}$.

For each $q \in [q_{\max}]$, we bound the mass contribution of rows corresponding to $W^q(\mathbf{A})$ at each subsampling level $\{0\} \cup [h_{\max}]$. Note that by Lemma 4.2.3, there is an event $\mathcal{E}_2$ with probability at least $1-2\delta$ such that all levels h, q except for those such that $h = 0$, or $p_h|W^q(\mathbf{A})| \in [m_{\text{crowd}}, Bm_{\text{crowd}})$, or $p_h|W^q(\mathbf{A})| \in [Bm_{\text{crowd}}, \infty)$ are bounded by at most $2\varepsilon\|\mathbf{Ax}\|_1$, so it remains to bound these levels.

Note that for each $q \in [q_{\max}]$, there exists at most one level $h_q \in [h_{\max}]$ such that $p_{h_q}|W^q(\mathbf{A})| \in [m_{\text{crowd}}, Bm_{\text{crowd}})$, since p_h varies in factors of B by construction. For these levels, we have by Lemma 4.2.4 that with probability at least $1-\alpha$,

$$\sum_{q\in[q_{\max}]} \|\mathbf{S}^{(h_q)}\mathbf{A}^{(q)}\mathbf{x}\|_1 \le \sum_{q\in[q_{\max}]} \|\mathbf{A}^{(q)}\mathbf{x}\|_1 + \varepsilon\|\mathbf{Ax}\|_1.$$

If such a sampling level h_q exists, then we have $|W^q(\mathbf{A})| \ge m_{\text{crowd}}/p_{h_q} \ge Bm_{\text{crowd}}$. Then by Lemma 4.2.5, the $h = 0$ level of sampling level contributes mass at most $(\varepsilon/q_{\max})\|\mathbf{Ax}\|_1$ with probability at least $1-\alpha/q_{\max}$. Thus by a union bound over all $q \in [q_{\max}]$ with a Goldilocks level and summing over these, the 0th level contributes at most $\varepsilon\|\mathbf{Ax}\|_1$. Otherwise, if a weight class q has no level h_q, then we have by the triangle inequality that $\|\mathbf{C}^{(0)}\mathbf{A}^{(q)}\mathbf{x}\|_1 \le \|\mathbf{A}^{(q)}\mathbf{x}\|_1$ and thus we simply bound the contribution of the 0th level by $\|\mathbf{A}^{(q)}\mathbf{x}\|_1$. Finally, note that for levels with $p_h|W^q(\mathbf{A})| \in [Bm_{\text{crowd}}, \infty)$, a similar application of Lemma 4.2.5 shows that the total contribution of all of these levels is at most $\varepsilon\|\mathbf{Ax}\|_1$.

Note that $\mathcal{E}_1 \cap \mathcal{E}_2$ occurs with probability at least $1-3\delta$. Then conditioned on this event, every $\mathbf{x} \in \mathbb{R}^d$ has a $1-2\alpha$ probability that

$$\begin{aligned}
\|\mathbf{SAx}\|_1 &= \left[\sum_{q>q_{\max}} \|\mathbf{SA}^{(q)}\mathbf{x}\|_1\right] + \sum_{q\in[q_{\max}]} \left[\|\mathbf{C}^{(0}\mathbf{A}^{(q)}\mathbf{x}\|_1 + \sum_{h\in[h_{\max}]} \|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{A}^{(q)}\mathbf{x}\|_1\right] \\
&\le \varepsilon\|\mathbf{Ax}\|_1 + \underbrace{(1+\varepsilon)\|\mathbf{Ax}\|_1}_{h_q \text{ or 0th level}} + \underbrace{\varepsilon\|\mathbf{Ax}\|_1}_{\text{0th level if } h_q \text{ level exists}} + \underbrace{2\varepsilon\|\mathbf{Ax}\|_1}_{\text{badly concentrated and oversampled levels}} \\
&\le (1+5\varepsilon)\|\mathbf{Ax}\|_1
\end{aligned}$$

which is the desired bound. □

We conclude by a standard net argument.

Theorem 4.2.8 (No expansion)**.** Let the randomized boundary M-sketch $\mathbf{S}$ satisfy the hypotheses of Lemmas 4.2.3, 4.2.6, and 4.2.7. Let $\alpha = \delta\exp(-d\log(3/\varepsilon))/2$. With probability at least $1-4\delta$, we have that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{SAx}\|_1 \le (1+8\varepsilon)\|\mathbf{Ax}\|_1.$$

Proof. By Lemma 4.2.7, there is an event with probability at least $1-3\delta$ such that conditioned on this event, for each $\mathbf{x}$, there is a $1-2\alpha$ probability that

$$\|\mathbf{SAx}\|_1 \le (1+5\varepsilon)\|\mathbf{Ax}\|_1. \tag{4.2}$$

It is well-known (see e.g., [BLM89]), that there exists an ε-net $\mathcal{N}$ of size at most $(3/\varepsilon)^d = \exp(d\log(3/\varepsilon))$ over the set $\{\mathbf{Ax} : \mathbf{x} \in \mathbb{R}^d, \|\mathbf{Ax}\| = 1\}$. Then by a union bound over the net, (4.2) holds for every $\mathbf{Ax} \in \mathcal{N}$ with probability at least $1 - \delta$.

Finally, let $\mathbf{x} \in \mathbb{R}^d$ be arbitrary with $\|\mathbf{Ax}\|_1 = 1$. It is shown in [WW19, Theorem 3.5] that $\mathbf{Ax} = \sum_{i=0}^{\infty} \mathbf{y}^{(i)}$ where each nonzero $\mathbf{y}^{(i)}$ has $\mathbf{y}^{(i)}/\|\mathbf{y}^{(i)}\|_1 \in \mathcal{N}$ and $\|\mathbf{y}^{(i)}\|_1 \le \varepsilon^i$. We then have that

$$\|\mathbf{SAx}\|_1 = \|\mathbf{S}\sum_{i=0}^{\infty}\mathbf{y}^{(i)}\|_1 \le \sum_{i=0}^{\infty}\|\mathbf{Sy}^{(i)}\|_1 \le (1+5\varepsilon)\sum_{i=0}^{\infty}\|\mathbf{y}^{(i)}\|_1 \le (1+5\varepsilon)\sum_{i=0}^{\infty}\varepsilon^i \le 1+8\varepsilon.$$

We conclude by homogeneity. □

4.3 No contraction

Our goal in this section is to show that $\|\mathbf{SAx}\|_1 \le (1-\varepsilon)\|\mathbf{Ax}\|_1$ simultaneously for every $\mathbf{x} \in \mathbb{R}^d$, with probability at least $1 - \delta$.

We analyze the no contraction lemma for each unit vector $\mathbf{y} \in \mathbb{R}^n$, and conclude by a union bound over a net (Section 4.3.4). We thus define weight classes based on an individual vector $\mathbf{y}$.

Definition 4.3.1 (Weight classes)**.** Let $\mathbf{y} \in \mathbb{R}^n$ be an ℓ_1 unit vector. Then, for each $q \in \mathbb{N}$, we let $W^q(\mathbf{y}) \subseteq [n]$ denote the set

$$W^q(\mathbf{y}) \coloneqq \{i \in [n] : |\mathbf{y}(i)| \in (2^{-q}, 2^{1-q}]\}$$

and let $\mathbf{y}^{(q)}$ denote the restriction of $\mathbf{y}$ to the rows indexed by $W^q(\mathbf{y})$.

4.3.1 Essential weight classes

We first reduce the analysis of preserving $\|\mathbf{y}\|_1$ to the analysis of preserving a subset of the weight classes $\|\mathbf{y}^{(q)}\|_1$.

Lemma 4.3.2. Consider a random boundary M-sketch (Definition 4.1.2). Let $q_{\max} = \log_2(n/\varepsilon)$, $q_0 = \log_2(m_{\min}/p_1\varepsilon)$, and $m_{\min} \ge 1$. Define

$$\begin{aligned}
\hat{Q}_0 &\coloneqq \{q \in [q_{\max}] : p_1|W^q(\mathbf{y})| < m_{\min}\} \\
\hat{Q}_h &\coloneqq \{q \in [q_{\max}] : p_h|W^q(\mathbf{y})| \in [m_{\min}, Bm_{\min})\}, \qquad h \in [h_{\max}] \\
Q_0 &\coloneqq \{q \in \hat{Q}_0 : q \le q_0, \|\mathbf{y}^{(q)}\|_1 \ge \varepsilon/q_0\} \\
Q_h &\coloneqq \{q \in \hat{Q}_h : q \le \min \hat{Q}_h + \log_2(B/\varepsilon), \|\mathbf{y}^{(q)}\|_1 \ge \varepsilon/q_{\max}\}, \qquad h \in [h_{\max}]
\end{aligned}$$

and let $Q^* \coloneqq \bigcup_{h=0}^{h_{\max}} Q_h$. If $h_{\max} \ge \log_B n$, then

$$\sum_{q \in Q^*} \|\mathbf{y}^{(q)}\|_1 \ge 1 - 8\varepsilon.$$

Proof. Note that

$$\sum_{q \geq q_{\max}} \|\mathbf{y}^{(q)}\|_1 \leq 2\frac{\varepsilon}{n} \sum_{q=1}^{\infty} |W^q(\mathbf{y})| \leq 2\varepsilon$$

so we only need to consider $q \in [q_{\max}]$. Note also that the $\hat{Q}_h$ for $h \in [h_{\max}] \cup \{0\}$ partition the set $[q_{\max}]$ by our choice of $h_{\max}$.

We first show that Q_0 preserves almost all of the mass of $\hat{Q}_0$. Indeed, the $q \in \hat{Q}_0$ with $q > q_0$ has total ℓ_1 mass at most

$$\sum_{q \in \hat{Q}_0, q > q_0} \|\mathbf{y}^{(q)}\|_1 \leq \frac{m_{\min}}{p_1} \sum_{q \in \hat{Q}_0, q > q_0} 2^{1-q} \leq 2\frac{m_{\min}}{p_1} \frac{\varepsilon p_1}{m_{\min}} \leq 2\varepsilon.$$

Of the levels $q \leq q_0$, the levels q with $\|\mathbf{y}^{(q)}\|_1 \leq \varepsilon/q_0$ have total mass at most ε.

Similarly, we show that Q_h preserves almost all of the mass of $\hat{Q}_h$. Indeed, the $q \in \hat{Q}_h$ with $q > \min \hat{Q}_h + \log_2(B/\varepsilon)$ has total ℓ_1 mass at most

$$\sum_{\substack{q \in \hat{Q}_h \\ q > \min \hat{Q}_h + \log_2(B/\varepsilon)}} \|\mathbf{y}^{(q)}\|_1 \leq \sum_{\substack{q \in \hat{Q}_h \\ q > \min \hat{Q}_h + \log_2(B/\varepsilon)}} 2^{1-q} \frac{B m_{\min}}{p_h} \leq 2B\frac{\varepsilon}{B} \|\mathbf{y}^{\min \hat{Q}_h}\|_1 = 2\varepsilon \|\mathbf{y}^{(\min \hat{Q}_h)}\|_1$$

so the total mass over all $h \in [h_{\max}]$ is at most 2ε. Of the levels $q \leq q_{\max}$, the levels q with $\|\mathbf{y}^{(q)}\|_1 \leq \varepsilon/q_{\max}$ have total mass at most ε.

The total lost mass is $2\varepsilon + 2\varepsilon + \varepsilon + 2\varepsilon + \varepsilon = 8\varepsilon$. □

4.3.2 Hashing lemmas

We collect lemmas on the hashing guarantees of CountSketch.

Lemma 4.3.3. Let $p_h |W^q(\mathbf{y})| \geq m_{\min}$ for $m_{\min} \geq 12\varepsilon^{-2} \log(4/\delta)$. Then, with probability at least $1 - \delta$,

$$\|\mathbf{S}^{(h)} \mathbf{y}^{(q)}\|_0 = (1 \pm \varepsilon) p_h \|\mathbf{y}^{(q)}\|_0$$
$$\|\mathbf{S}^{(h)} \mathbf{y}^{(q)}\|_1 = (1 \pm \varepsilon) p_h \|\mathbf{y}^{(q)}\|_1$$

Proof. This follows from Chernoff bounds. □

The following lemma uses a standard balls and bins martingale argument to show that most items are hashed uniquely.

Lemma 4.3.4 (Concentration for unique hashing)**.** Let $H : [n] \to [r]$ be a random hash function. Let $S \subseteq T \subseteq [n]$, $p \in (0, 1]$, and $\varepsilon \in (0, 1)$ with $\varepsilon r \geq p|T|$. Consider the process that samples each element $i \in [n]$ with probability p and hashes it to a bucket in $[r]$ if it was sampled. Let X be the number of elements $i \in S$ that are sampled and hashed to a bucket containing no other member of T. Then,

$$\mathbf{Pr}\{X \geq (1 - \varepsilon)^2 p|S|\} \leq 2 \exp\left(-\frac{\varepsilon^2}{12} p|S|\right).$$

Proof. For each $i \in S$, sample i with probability p and place the result in a uniformly random hash bucket in $[r]$ if it was sampled. Let $\mathcal{E}_i$ denote the event where i is sampled and is hashed to a bucket with no other members of T. Let $C_1, C_2, \ldots, C_{|S|}$ denote the sequence of these independent random choices and let $f(C_1, C_2, \ldots, C_s)$ denote the number of hash buckets in $[r]$ that contains members $i \in S$ satisfying $\mathcal{E}_i$ at the end of the sampling and hashing process. Note that f is 1-Lipschitz, and that

$$\mathbf{E}\, f(C_1, C_2, \ldots, C_{|S|}) = \sum_{i \in S} \mathbf{Pr}(\mathcal{E}_i) = |S| p \Big(1 - \frac{p}{r}\Big)^{|T|} \geq p|S| \left(1 - \frac{p|T|}{r}\right) \geq (1-\varepsilon) p |S|.$$

Now consider the Doob martingale

$$Z_k \coloneqq \mathbf{E}\big[f_q(C_1, C_2, \ldots, C_{|S|}) \mid C_1, C_2, \ldots, C_k\big].$$

Note that the increments $Z_k - Z_{k-1}$ conditioned on $C_1, C_2, \ldots, C_{k-1}$ is simply the indicator variable of whether on choice C_k we sampled an entry and placed it in a new bucket or not. Then $Z_k - Z_{k-1} = 1$ with probability at most p and thus $\mathbf{E}_{k-1}(Z_k - Z_{k-1})^2 \leq p$. Then by Freedman's inequality [Fre75],

$$\mathbf{Pr}\big(|Z_{|S|} - Z_0| \geq \varepsilon Z_0\big) \leq 2\exp\left(-\frac{1}{2}\frac{(\varepsilon(1-\varepsilon)p|S|)^2}{p|S| + \varepsilon(1-\varepsilon)p|S|/3}\right) \leq 2\exp\left(-\frac{\varepsilon^2}{12} p|S|\right). \qquad \square$$

We apply Lemma 4.3.4 in the context of the M-sketch in the following lemma.

Lemma 4.3.5 (Approximately perfect hashing)**.** Let $h \in [h_{\max}]$ and $Q \subseteq \{q : p_h |W_q| \geq m_{\min}\}$ for $m_{\min} \geq 12\varepsilon^{-2}\log(4/\delta)$. Let $\hat{W} \supset W_Q$ for $W_Q \coloneqq \bigcup_{q \in Q} W^q(\mathbf{y})$ and suppose that $p_h|\hat{W}| \leq \varepsilon N$ for some $\varepsilon \in (0, 1/2)$. Then with probability at least $1 - 2|Q|\delta$, every $W^q(\mathbf{y})$ has a subset $W^q_* \subset W^q(\mathbf{y})$ that gets sampled and placed in a hash bucket with no other members of $\hat{W}$, and

$$\|\mathbf{y}^{(q)}_*\|_0 \geq (1-2\varepsilon) p_h \|\mathbf{y}^{(q)}\|_0$$
$$\|\mathbf{y}^{(q)}_*\|_1 \geq (1-7\varepsilon) p_h \|\mathbf{y}^{(q)}\|_1$$

where $\mathbf{y}^{(q)}_*$ is the restriction of $\mathbf{y}^{(q)}$ to W^q_*.

Proof. We apply Lemma 4.3.4 to see that with probability at least

$$1 - 2\exp\left(-\frac{\varepsilon^2}{12} p_h |W^q(\mathbf{y})|\right) \leq 1 - \delta,$$

there is a set $W^q_* \subseteq W^q(\mathbf{y})$ of elements that are hashed to a bucket with no other element of $\hat{W}$ in it and of size $|W^q_*| \geq (1-\varepsilon)^2 p_h |W^q(\mathbf{y})| \geq (1-2\varepsilon) p_h |W^q(\mathbf{y})|$. We condition on this event.

By Lemma 4.3.3, with probability at least $1 - \delta$, we sample $(1 \pm \varepsilon) p_h |W^q(\mathbf{y})|$ elements with mass $(1 \pm \varepsilon) p_h \|\mathbf{y}^{(q)}\|_1$. Note then that there are at most $3\varepsilon p_h |W^q(\mathbf{y})|$ sampled elements that do not belong to W^q_*. The mass of these elements is at most

$$3\varepsilon p_h |W^q(\mathbf{y})| 2^{1-q} \leq 6\varepsilon p_h \|\mathbf{y}^{(q)}\|_1.$$

Thus,

$$\|\mathbf{y}^{(q)}_*\|_1 \geq (1-\varepsilon) p_h \|\mathbf{y}^{(q)}\|_1 - 6\varepsilon p_h \|\mathbf{y}^{(q)}\|_1 = (1-7\varepsilon) p_h \|\mathbf{y}^{(q)}\|_1.$$

We conclude by a union bound over the weight classes Q. $\square$

Lemma 4.3.6 (Balls and bins)**.** Let $W \subseteq [n]$ such that $\|\mathbf{y}|_W\|_\infty \le T$ and let the number of hash buckets N be at least $N \ge \|\mathbf{y}|_W\|_1/T$. For $k \in [N]$, let $L_k \subseteq W$ denote the indices from W hashed to the hash bucket k. Then, with probability at least $1-\delta$,

$$\max_{k=1}^{N} \|\mathbf{y}|_{L_k}\|_1 \le 3T\log(N/\delta)$$

Proof. Fix a single bucket $k \in [N]$. Then, $\mathbf{E}\|\mathbf{y}|_{L_k}\|_1 = \|\mathbf{y}|_W\|_1/N \le T$. Furthermore, $\|\mathbf{y}|_{L_k}\|_1$ is the independent sum of $|W|$ random variables that are bounded by T with variance

$$\mathbf{Var}[\|\mathbf{y}|_{L_k}\|_1] = \sum_{i\in W} \frac{\mathbf{y}(i)^2}{N} = \frac{\|\mathbf{y}|_W\|_2^2}{N} \le \frac{T\|\mathbf{y}|_W\|_1}{N} \le T^2.$$

Then by Bernstein bounds, we have that

$$\begin{aligned}\mathbf{Pr}\{\|\mathbf{y}|_{L_k}\|_1 \ge \mathbf{E}\|\mathbf{y}|_{L_k}\|_1 + 2T\log(N/\delta)\} &\le \exp\left(-\frac{1}{2}\frac{4T^2(\log(N/\delta))^2}{T^2 + 2T^2(\log(N/\delta)/3}\right)\\ &\le \exp(-\log(N/\delta)) = \delta/N.\end{aligned}$$

A union bound over the N buckets yields the claim. □

4.3.3 Preserving weight classes

We now use the previous results on approximate perfect hashing to show the main result, that the random boundary M-sketch $\mathbf{S}$ does not contract ℓ_1 norms.

We first show the no contraction lemma for sampling levels $h \in [h_{\max}]$.

Lemma 4.3.7. Let the number of hash buckets N satisfy

$$N \ge \frac{6Bm_{\min}q_{\max}\log(Nh_{\max}\log_2(B/\varepsilon)/\delta)}{\varepsilon^3}$$

Then, we have with probability at least $1-2\delta$ that

$$\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{y}\|_1 \ge (1-8\varepsilon)\sum_{q\in Q_h}\|\mathbf{y}^{(q)}\|_1.$$

for every $h \in [h_{\max}]$.

Proof. Note that for any $i \in W^q(\mathbf{y})$ for $q \in Q_h$,

$$|\mathbf{y}(i)| \ge 2^{-q} = \frac{2^{1-q}p_h|W^q(\mathbf{y})|}{2p_h|W^q(\mathbf{y})|} \ge \frac{p_h\|\mathbf{y}^{(q)}\|_1}{2Bm_{\min}} \ge \frac{p_h\varepsilon}{2q_{\max}Bm_{\min}} =: \tau_h.$$

By Lemma 4.3.6, as long as we avoid hashing i with any entry that is larger than an ε fraction of this (i.e. $T \approx \varepsilon\tau_h$), then the total ℓ_1 mass of all other elements in the hash bucket will only be roughly an ε fraction of τ_h. We will now carry out this analysis.

Let $T_h := \varepsilon\tau_h/3\log(Nh_{\max}\log_2(B/\varepsilon)/\delta)$ and let $W_h^{\text{large}} := \bigcup_{q=1}^{\log_2(1/T_h)} W^q(\mathbf{y})$. Note that $i \in W_h^{\text{large}}$ satisfies $\mathbf{y}(i) \geq T_h$, so

$$|W_h^{\text{large}}| \leq \frac{1}{T_h} = \frac{1}{p_h}\frac{6Bm_{\min}q_{\max}\log(Nh_{\max}\log_2(B/\varepsilon)/\delta)}{\varepsilon^2} \leq \frac{\varepsilon N}{p_h},$$

that is, $p_h|W_h^{\text{large}}| \leq \varepsilon N$. Furthermore, $W_{Q_h} \subseteq W_h^{\text{large}}$. We may thus apply Lemma 4.3.5 with $\hat{W} = W_h^{\text{large}}$ and $Q = Q_h$. We condition on the success of Lemma 4.3.5.

Now for each $q \in Q_h$ and $i \in W_*^q$ given by Lemma 4.3.5, let $k_i \in [N]$ be the hash bucket containing i. Then, by applying Lemma 4.3.6 with $W = [n] \setminus W_h^{\text{large}}$ with $N \geq p_h/T_h$ buckets, we have that the total ℓ_1 mass in bucket k_i besides the item i is at most $3T_h\log(Nh_{\max}\log_2(B/\varepsilon)/\delta) \leq \varepsilon\tau_h$, with probability at least $1 - \delta/h_{\max}\log_2(B/\varepsilon)$. Thus, we have that

$$|[\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{y}](k_i)| \geq \frac{1}{p_h}(|\mathbf{y}(i)| - \varepsilon\tau_h) \geq (1-\varepsilon)\frac{|\mathbf{y}(i)|}{p_h}.$$

This holds simultaneously for every $q \in Q_h$ with probability at least $1 - |Q_h|\delta/h_{\max}\log_2(B/\varepsilon) \geq 1 - \delta/h_{\max}$ by a union bound. By summing over all $q \in Q_h$ and $i \in W_*^q$, we have that

$$\begin{aligned}
\|\mathbf{C}^{(h)}\mathbf{S}^{(h)}\mathbf{y}\|_1 &\geq (1-\varepsilon)\sum_{q\in Q_h}\sum_{i\in W_*^q}\frac{|\mathbf{y}(i)|}{p_h} \\
&\geq (1-\varepsilon)(1-7\varepsilon)\sum_{q\in Q_h}\|\mathbf{y}^{(q)}\|_1 && \text{Lemma 4.3.5} \\
&\geq (1-8\varepsilon)\sum_{q\in Q_h}\|\mathbf{y}^{(q)}\|_1.
\end{aligned}$$

By another union bound over $h \in [h_{\max}]$, this is true for every sampling level h with probability at least $1 - \delta$. □

We show a similar result for the $h = 0$ level. Instead of using a concentration-based argument via Lemma 4.3.5, we instead show that important elements at this level can be perfectly hashed.

Lemma 4.3.8. Let the number of hash buckets N_0 satisfy

$$N_0 \geq \frac{6\log(N_0/\delta)q_0^2 m_{\min}}{\delta\varepsilon^2 p_1}$$

Then, we have with probability at least $1 - 2\delta$ that

$$\|\mathbf{C}^{(0)}\mathbf{y}\|_1 \geq (1-\varepsilon)\sum_{q\in Q_0}\|\mathbf{y}^{(q)}\|_1$$

Proof. Note that for any $i \in W^q(\mathbf{y})$ for $q \in Q_0$,

$$|\mathbf{y}(i)| \geq 2^{-q} = \frac{2^{1-q}p_1|W^q(\mathbf{y})|}{2p_1|W^q(\mathbf{y})|} \geq \frac{p_1\|\mathbf{y}^{(q)}\|_1}{2m_{\min}} \geq \frac{p_1\varepsilon}{2q_0 m_{\min}} =: \tau_0.$$

Let $T_0 \coloneqq \varepsilon\tau_0/3\log(N_0/\delta)$ and let $W_0^{\text{large}} \coloneqq \bigcup_{q=1}^{\log_2(1/T_0)} W^q(\mathbf{y})$. With at least $N_0 \geq 1/T_0$, by Lemma 4.3.6, any hash bucket $k \in [N_0]$ has a total ℓ_1 contribution from elements outside W_0^{large} of at most $\varepsilon\tau_0$, with probability at least $1-\delta$. Furthermore, with $N_0 \geq q_0/\delta T_0$ buckets, we can perfectly hash all indices in $W^q(\mathbf{A})$ to different buckets from W_0^{large} with probability at least $1-\delta$. Thus, for each $i \in W_0^{\text{large}}$ and hash bucket $k_i \in [N_0]$ containing i, we have

$$|[\mathbf{C}^{(0)}\mathbf{y}](k_i)| \geq |\mathbf{y}(i)| - \varepsilon\tau_0 \geq (1-\varepsilon)|\mathbf{y}(i)|.$$

By summing over all $i \in W_0^{\text{large}}$, we obtain that

$$\|\mathbf{C}^{(0)}\mathbf{y}\| \geq (1-\varepsilon)\sum_{i\in W_0^{\text{large}}} |\mathbf{y}(i)| \geq (1-\varepsilon)\sum_{q\in Q_0}\|\mathbf{y}^{(q)}\|_1.$$

□

4.3.4 Net argument

Finally, we will assemble our previous lemmas to show that the randomized boundary M-sketch does not contract ℓ_1 norms. The proof is analogous to that of Theorem 4.2.8.

Theorem 4.3.9 (No contraction)**.** Let the randomized boundary M-sketch $\mathbf{S}$ satisfy the hypotheses of Lemmas 4.3.7 and 4.3.8. Let $\alpha = \delta\exp(-d\log(3/\varepsilon))/4$. With probability at least $1-\delta$, we have that for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{SAx}\|_1 \geq (1-67\varepsilon)\|\mathbf{Ax}\|_1.$$

Proof. For any fixed vector $\mathbf{y}$, Lemmas 4.3.7 and 4.3.8 hold with probability at least $1-\alpha$ by a union bound. By summing over the results of these lemmas and then applying Lemma 4.3.2, we have that

$$\|\mathbf{Sy}\|_1 \geq (1-8\varepsilon)\sum_{q\in Q^*}\|\mathbf{y}^{(q)}\|_1 \geq (1-8\varepsilon)^2\|\mathbf{y}\|_1.$$

We then conclude by a net argument simular to the proof of Theorem 4.2.8. □

4.4 Endgame

We first verify that the parameters of the randomized boundary M-sketch can be chosen to satisfy all the hypotheses necessary to satisfy Theorems 4.2.8 and 4.3.9 to obtain the following.

Theorem 4.4.1 (Singly exponential oblivious ℓ_1 subspace embeddings)**.** Let $\mathbf{S}$ be a randomized boundary M-sketch (Definition 4.1.2) with parameters $h_{\max} = \log_2 n$ and

$$B, N_0, N = \exp\left(O\left(\frac{d}{\delta\varepsilon}\log\frac{d\log n}{\delta\varepsilon}\right)\right)$$

Then, $\mathbf{S}$ has

$$r = \exp\left(O\left(\frac{d}{\delta\varepsilon}\log\frac{d\log n}{\delta\varepsilon}\right)\right)$$

rows and satisfies

$$\mathbf{Pr}\big\{\text{for all } \mathbf{x} \in \mathbb{R}^d, \quad (1-\varepsilon)\|\mathbf{A}\mathbf{x}\|_1 \le \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \le (1+\varepsilon)\|\mathbf{A}\mathbf{x}\|_1\big\} \ge 1-\delta.$$

The above theorem is almost the claimed result for Theorem 4.0.2, up to the $\log\log n$ dependence in the exponent. To remove this dependence, we will use the dense Cauchy sketch of [WW19, WW22], which removes the dependence on n at a cost of a doubly exponential dependence on d. While [WW19, WW22] only prove the result for $O(1)$ approximation, we improve their analysis to $(1+\varepsilon)$ approximations.

Theorem 4.4.2 (Doubly exponential oblivious ℓ_1 subspace embeddings [WW19, WW22])**.** There exists a distribution over $r \times n$ matrices $\mathbf{S}$ for $r = \exp(\exp(\tilde{O}(d/\varepsilon^2))/\delta)$ such that for any $\mathbf{A} \in \mathbb{R}^{n\times d}$,

$$\mathbf{Pr}\big\{\text{for all } \mathbf{x} \in \mathbb{R}^d, \quad (1-\varepsilon)\|\mathbf{A}\mathbf{x}\|_1 \le \|\mathbf{S}\mathbf{A}\mathbf{x}\|_1 \le (1+\varepsilon)\|\mathbf{A}\mathbf{x}\|_1\big\} \ge 1-\delta$$

Proof. We take $\mathbf{S}$ to be an appropriate scaling of a $r \times n$ dense Cauchy matrix.

Let $X_1, X_2, \ldots X_r$ be independent Cauchy variables. Let $R = \Theta(\frac{r\log r}{\log\log r})$ and let $\mathcal{E}$ denote the event that $\max_{i=1}^r |X_i| \le R$. Note that $\mathbf{Pr}[|X_i| \ge t] \le O(1/t)$ for Cauchy variables, so by a union bound, we have that $\mathbf{Pr}(\mathcal{E}) \ge 1 - O(r/R) \ge 1 - O((\log\log r)/\log r)$. Furthermore, we have by linearity of expectation that $\mathbf{E}[\|X\|_1 \mid \mathcal{E}] = \Omega(r\log R)$. Then by Chernoff bounds,

$$\mathbf{Pr}\{\|X\|_1 = (1\pm\varepsilon)\,\mathbf{E}[\|X\|_1 \mid \mathcal{E}] \mid \mathcal{E}\} \ge 1 - 2\exp\left(-C\varepsilon^2\frac{r\log R}{R}\right) \ge 1 - 2\exp\big(-C\varepsilon^2\log\log r\big)$$

for some constant C. Thus for $r = \exp(\exp(\tilde{O}(d/\varepsilon^2))/\delta)$, we have

$$\mathbf{Pr}\{\|X\|_1 = (1\pm\varepsilon)\,\mathbf{E}[\|X\|_1 \mid \mathcal{E}]\} \ge 1 - \delta(\varepsilon/3)^d$$

Finally, note that $X = \mathbf{S}\mathbf{y}$ is distributed as r independent Cauchy variables scaled by $\|\mathbf{y}\|_1$ by the 1-stability of Cauchy variables. Thus, by the above result, $\|\mathbf{S}\mathbf{y}\|_1$ concentrates around some scaling of $\|\mathbf{y}\|_1$ up to a $(1\pm\varepsilon)$ factor, with probability at least $1-\delta(\varepsilon/3)^d$. We may then perform a net argument just as in Theorems 4.2.8 and 4.3.9 to conclude the theorem. □

With the above theorem in hand, we finally arrive at a proof of Theorem 4.0.2.

Proof of Theorem 4.0.2. We first apply Theorem 4.4.2 to reduce n to $\exp(\exp(\tilde{O}(d/\varepsilon^2)))$. Then, applying Theorem 4.4.1 once reduces the number of rows to $\exp(\tilde{O}(d^2/\varepsilon))$, and then applying it again further reduces the number of rows to $\exp(\tilde{O}(d/\varepsilon))$, as claimed. □

Chapter 5

Future directions for oblivious ℓ_p subspace embeddings

While we have been able to resolve many of the outstanding gaps in our understanding of oblivious ℓ_p subspace embeddings, several interesting questions still remain to be explored. Perhaps one of the most notable unresolved challenges is to resolve the dependence on the accuracy parameter ε for $(1+\varepsilon)$ oblivious ℓ_1 subspace embeddings. Our upper bounds in [LWY21] have a singly exponential dependence on $1/\varepsilon$, while there is no known lower bound which rules out an upper bound of the form $r = \exp(\mathrm{poly}(d))/\mathrm{poly}(\varepsilon)$. We conjecture that our upper bound is tight, and ask whether one can show an exponential lower bound in ε, even for $d = O(1)$.

Question 5.0.1. Is there an $\exp(\mathrm{poly}(1/\varepsilon))$ lower bound on r for $(1+\varepsilon)$ oblivious ℓ_1 subspace embeddings for $d = O(1)$?

A second question is to pin down the polynomial dependence on d in the exponent. The lower bound result of [WW19, WW22] of Theorem 3.0.2 shows that for a distortion of $\kappa = O(1)$, we need $d = O((\log r)^2)$, or $r = \exp(\Omega(\sqrt{d}))$. On the other hand, our upper bound is linear in d in the exponent, i.e., $r = \exp(\tilde{O}(d))$. Thus, an interesting question is to resolve this gap.

Question 5.0.2. Is there an $\exp(\Omega(d))$ lower bound on r for $O(1)$ oblivious ℓ_1 subspace embeddings?

Part II

Sampling Algorithms and Coresets

Chapter 6

ℓ_p Lewis weight sampling [WY23b]

In Section 1.3.2, we have discussed the leverage score sampling algorithm, which gives a sampling-based approach to constructing nearly optimal ℓ_2 subspace embeddings. A highly fruitful direction of research is to explore how this result can be generalized to other loss functions, and in particular ℓ_p losses. That is, we will study randomized algorithms for constructing $(1 \pm \varepsilon)$-approximate subspace embeddings $\mathbf{S} \in \mathbb{R}^{n \times n}$ for the ℓ_p norm (Definition 1.1.1) satisfying

$$\mathbf{Pr}\big\{\text{for all } \mathbf{x} \in \mathbb{R}^d, \|\mathbf{SAx}\|_p^p = (1 \pm \varepsilon)\|\mathbf{Ax}\|_p^p\big\} \geq 1 - \delta.$$

Here, $\mathbf{S}$ will be a sampling matrix, that is, $\mathbf{S}$ is a diagonal matrix with few nonzero entries, and we will seek to minimize the *row count* $r = \mathsf{nnz}(\mathbf{S})$ of $\mathbf{S}$.

In this chapter, we will introduce the sampling methods for constructing ℓ_p subspace embeddings, with a particular focus on the technique of ℓ_p Lewis weight sampling, and discuss two improvements to this technique made in the work of [WY23b]. In particular, we give a nearly optimal "one-shot" ℓ_p Lewis weight sampling theorem and an online ℓ_p Lewis weight sampling theorem in Section 6.6.

6.1 Sampling algorithms for ℓ_p subspace embeddings

There are many possible natural generalizations of leverage scores to the setting of ℓ_p subspace embeddings, but not all are known to achieve the best trade-offs between the dimension d, the accuracy parameter ε, and the row count r. We discuss several of these approaches and their shortcomings before introducing ℓ_p *Lewis weights*, which is the main subject study for most of this chapter. We introduce the following definition to facilitate our discussion:

Definition 6.1.1 (ℓ_p sampling matrix)**.** Let $p \geq 1$. A random diagonal matrix $\mathbf{S} \in \mathbb{R}^{n \times n}$ is a *random ℓ_p sampling matrix with sampling probabilities* $\{q_i\}_{i=1}^n$ if for each $i \in [n]$, the ith diagonal entry is independently set to be

$$\mathbf{S}_{i,i} = \begin{cases} 1/q_i^{1/p} & \text{with probability } q_i \\ 0 & \text{otherwise} \end{cases}$$

6.1.1 ℓ_p sensitivity sampling

We originally motivated the definition of leverage scores in Section 1.3.2 as the sampling algorithm obtained when specializing the general technique of sensitivity sampling to the setting of ℓ_2 subspace embeddings. Doing the same for ℓ_p subspace embeddings yields a sampling algorithm known as ℓ_p *sensitivity sampling*.

Definition 6.1.2 (ℓ_p sensitivities)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $0 < p < \infty$. Then for each $i \in [n]$, the *ith* ℓ_p *sensitivity of* $\mathbf{A}$ is defined to be

$$\boldsymbol{\sigma}_i^p(\mathbf{A}) \coloneqq \sup_{\mathbf{Ax} \neq 0} \frac{|[\mathbf{Ax}](i)|^p}{\|\mathbf{Ax}\|_p^p}.$$

The *total* ℓ_p *sensitivity* is $\mathfrak{S}^p(\mathbf{A}) \coloneqq \sum_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A})$.

Recall that the sensitivity sampling framework of [LS10, FL11] (see Section 1.3.1) gives a bound of $r = \tilde{O}(\varepsilon^{-2}\mathfrak{S}d)$ on the row count required to sample an ℓ_p subspace embedding.

While aesthetically pleasing, the ℓ_p sensitivity sampling has a number of disadvantages. Perhaps the most glaring is that ℓ_p sensitivity sampling is not known to achieve nearly optimal row counts for ℓ_p subspace embeddings for worst-case matrices. Indeed, we will see that ℓ_p Lewis weight sampling achieves nearly optimal row counts, whereas ℓ_p sensitivity sampling is only known to match these results for $p = 2$. Another drawback of ℓ_p sensitivity sampling is that we do not currently know fast algorithms for estimating the ℓ_p sensitivity scores. Indeed, naïvely computing the ℓ_p sensitivity scores essentially requires solving an ℓ_p linear regression problem of the form of

$$\frac{1}{\boldsymbol{\sigma}_i^p(\mathbf{A})} = \min_{[\mathbf{Ax}](i)=1} \|\mathbf{Ax}\|_p^p$$

for each $i \in [n]$, and algorithms that can compute these scores in time as fast as solving linear regression as in the case of $p = 2$ [SS11, DMMW12, CW13, LMP13, CLM+15] are not known. A recent work of [PWZ23] shows a trade-off that the number of sensitivity calculations can be reduced to $O(n/\alpha)$ at a cost of an α factor blow-up in the total sensitivity, but this still does not match the known results for ℓ_2 leverage score sampling. For these reasons, ℓ_p sensitivity sampling has not attracted as much attention from the literature of randomized numerical linear algebra.

Nonetheless, the study of ℓ_p sensitivity sampling does have benefits over other sampling algorithms for ℓ_p subspace embeddings, in particular for constructing ℓ_p subspace embeddings for $p > 2$ for input matrices $\mathbf{A}$ with total sensitivity $\mathfrak{S}^p(\mathbf{A})$ much less than the worst case of $d^{p/2}$. We will give a much more in-depth study of ℓ_p sensitivity sampling in Chapter 7, together with another natural generalization of leverage score sampling known as root leverage score sampling.

6.1.2 ℓ_p well-conditioned basis sampling

Although ℓ_p sensitivity scores appear to be difficult to quickly estimate up to a small constant factor, if one is willing to sacrifice on the row count up to $\mathrm{poly}(d)$ factors, then fast routines do exist. Indeed, some of the earliest works on sampling-based algorithms for ℓ_p linear regression proceed in this manner. The main idea is to generalize the observation that the leverage scores

can be characterized as the row norms of any orthogonal basis of $\mathbf{A}$. That is, if $\mathbf{U} \in \mathbb{R}^{n \times d}$ is an orthogonal basis of $\mathbf{A} \in \mathbb{R}^{n \times d}$, then it is not hard to see that

$$\boldsymbol{\tau}_i(\mathbf{A}) = \|\mathbf{e}_i^\top \mathbf{U}\|_2^2$$

for every $i \in [n]$ (see (1.3)). To generalize this to the ℓ_p norm setting, we can then recall constructions of well-conditioned bases $\mathbf{U}$ for subspaces of ℓ_p (Definition 3.1.1) and define analogous scores that are proportional to $\|\mathbf{e}_i^\top \mathbf{U}\|_p^p$. Indeed, such approaches were considered and used to obtain ℓ_p subspace embeddings with $r = \text{poly}(d/\varepsilon)$ rows and $\kappa = (1+\varepsilon)$ distortion [Cla05, DDH+09]:

Theorem 6.1.3 (ℓ_p well-conditioned basis sampling [Cla05, DDH+09])**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $1 \leq p < \infty$. Let $\mathbf{U} \in \mathbb{R}^{n \times d}$ be a $(\text{poly}(d), 1, p)$-well-conditioned basis for the column space of $\mathbf{A}$ (see Definition 3.1.1). Let $\alpha > 0$ and let $q_i = \min\{1, \|\mathbf{e}_i^\top \mathbf{U}\|_p^p / \alpha\}$ for $i \in [n]$. Let $\mathbf{S} \in \mathbb{R}^{n \times n}$ be the $\ell)p$ sampling matrix with probabilities $\{q_i\}_{i=1}^n$ (Definition 6.1.1). Then, there is an α such that with probability at least $99/100$ $\mathbf{S}$ is an ℓ_p subspace embedding satisfying Definition 1.1.1 with $\kappa = (1+\varepsilon)$, and furthermore, $\mathbf{S}$ has at most $r = \text{poly}(d/\varepsilon)$ nonzero rows.

In the results of [Cla05, DDH+09], the crucial subroutine of computing the well-conditioned basis $\mathbf{U}$ was done by using algorithms for computing Löwner–John ellipsoids for ℓ_p balls. This results in running times of the form $n\,\text{poly}(d)$. While this avoids the $O(n^2)$ running time cost resulting from solving n ℓ_p linear regression instances of size $n \times d$ for sensitivity sampling, this is still far less efficient than the running time of $\tilde{O}(\text{nnz}(\mathbf{A}) + d^\omega)$ for leverage score sampling.

6.1.3 ℓ_p Lewis weight sampling

The work of [CP15] observed that the problem of constructing ℓ_p subspace embeddings of the form of Definition 1.1.1 has actually been studied decades ago in the geometric functional analysis literature, and obtains *nearly optimal* trade-offs between the number of rows r and the accuracy parameter ε. Indeed, a series of works [Lew78, BLM89, Tal90, LT91, Tal95, SZ01] culminated in the following result:

Theorem 6.1.4 (ℓ_p subspace embeddings, existential version [Lew78, BLM89, LT91, SZ01])**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $0 < p < \infty$. Then, there exists an ℓ_p subspace embedding $\mathbf{S} \in \mathbb{R}^{r \times n}$ with distortion $\kappa = (1+\varepsilon)$ with

$$r = \begin{cases} O(\varepsilon^{-2} d (\log d)^2 \log(d/\varepsilon)) & 0 < p < 2 \\ O(\varepsilon^{-2} d^{p/2} (\log d)^2 \log(d/\varepsilon)) & 2 < p < \infty \end{cases}$$

We will present an algorithmic version of the bounds of Theorem 6.1.4 in the following sections of this chapter, following the proofs presented in the works of [LT91, CP15] as well as improvements obtained in our work [WY23b]. We note that the statement of Theorem 6.1.4 is slightly suboptimal in the logarithmic factors compared to the best known results [BLM89, Tal90, Tal95, Zva00, CP15], but we present this version as it uses a simpler proof that we work extensively with, while achieving the best known dependencies on d and ε, up to polylogarithmic factors.

It has recently been shown that the upper bound of Theorem 6.1.4 is nearly optimal for $p < 2$, while for $p > 2$, the dependence on ε and d are individually optimal [LWW21, LLW23] when $d = \Omega(\log(1/\varepsilon))$.

Theorem 6.1.5 ([LWW21])**.** Let $p \in [1, \infty) \setminus 2\mathbb{Z}$. Suppose that $\mathbf{S} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $r = \mathsf{nnz}(\mathbf{S})$ such that

$$\text{for all } \mathbf{x} \in \mathbb{R}^d, \|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_p^p$$

Then, $r = \tilde{\Omega}(d/\varepsilon^2)$. Furthermore, for $p > 2$, $r = \tilde{\Omega}(\varepsilon^{-1} d^{p/2})$.

Algorithmic aspects: approximating ℓ_p Lewis weights

The proof of Theorem 6.1.4 is *almost* algorithmic, as the proof is based on the probabilistic method; the only component which is not algorithmic is the construction of a certain set of weights known as the *Lewis weights* [Lew78], which can be viewed as a certain generalization of the leverage scores (Definition 1.3.2) for ℓ_p that differs from the other scores that we have discussed so far.

Definition 6.1.6 (ℓ_p Lewis weights [Lew78, CP15])**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $0 < p < \infty$. Then, the *ℓ_p Lewis weights of* $\mathbf{A}$ are the unique set of weights $\mathbf{w} \in \mathbb{R}^n_{\geq 0}$ such that for every $i \in [n]$,

$$\mathbf{w}_i = \boldsymbol{\tau}_i(\mathbf{W}^{1/2-1/p}\mathbf{A}),$$

where $\mathbf{W} = \operatorname{diag}(\mathbf{w})$. We will denote the ℓ_p Lewis weights of $\mathbf{A}$ as $\mathbf{w}_i^p(\mathbf{A})$ for $i \in [n]$.

The work of Cohen and Peng [CP15] addresses the problem of the *algorithmic computation* of Lewis weights by showing that Lewis weights can, in fact, be approximated efficiently, and even in nearly input sparsity time for $p \in (0, 4)$. For $p \in (0, 4)$, their algorithm uses the following equivalent and more algorithmically useful characterization of ℓ_p Lewis weights in a fixed point iteration algorithm

$$\mathbf{w}_i = \left(\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A})^- \mathbf{a}_i\right)^{p/2}. \tag{6.1}$$

Follow-up works have further refined algorithms for approximating ℓ_p Lewis weights [Lee16, CCLY19, FLPS22, JLS22], and approximate ℓ_p Lewis weights that are compatible with sampling can now be computed in nearly input sparsity time for all $p > 0$ [JLS22]. In particular, a crucial relaxation for the efficient computation of ℓ_p Lewis weights is the notion of *one-sided ℓ_p Lewis weights*, which we show is sufficient for sampling:

Definition 6.1.7 (One-sided ℓ_p Lewis weights [JLS22, WY22b])**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $p \in (0, \infty)$. Let $\gamma \in (0, 1]$. Then, weights $\mathbf{w} \in \mathbb{R}^n$ are *γ-one-sided ℓ_p Lewis weights* if

$$\mathbf{w}_i \geq \gamma \cdot \boldsymbol{\tau}_i(\mathbf{W}^{1/2-1/p}\mathbf{A}),$$

where $\mathbf{W} \coloneqq \operatorname{diag}(\mathbf{w})$, or equivalently,

$$\mathbf{w}_i \geq \gamma^{p/2}\left[\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A})\mathbf{a}_i\right]^{p/2}.$$

If $\gamma = 1$, we just say that $\mathbf{w}$ are *one-sided ℓ_p Lewis weights*.

The following theorem collects the results of [CP15, JLS22] on the fastest known algorithms for approximating one-sided ℓ_p Lewis weights:

Theorem 6.1.8 ([CP15, JLS22])**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $0 < p < \infty$. Then, there is an algorithm which computes one-sided ℓ_p Lewis weights (Definition 6.1.7) $\mathbf{w}$ such that $d \leq \|\mathbf{w}\|_1 \leq 2d$ in $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^\omega)$ time.

Algorithmic aspects: sampling

While the works above address the question of approximating Lewis weights, using the Lewis weights to sample ℓ_p subspace embeddings is an orthogonal direction of investigation. By an appropriate adaptation of the earlier work in geometric functional analysis [BLM89, LT91, SZ01], as well as the construction of ℓ_p Lewis weights due to [CP15], one can obtain algorithmic constructions of ℓ_p subspace embeddings which match the guarantees of Theorem 6.1.4 [MMWY22]. However, this construction has the drawback that the sampling algorithm requires a sophisticated *recursive* structure in which the number of rows are reduced by half for $O(\log n)$ recursive rounds of sampling. This hinders the use of Lewis weight sampling in one-pass streaming settings [WY23b], and poses a gap from algorithms for ℓ_2 leverage score sampling, which admits ℓ_2 subspace embeddings just by sampling proportionally to the leverage scores in a "one-shot" sampling algorithm [DMM06a, RV07, Mag10], as well as streaming and online variants [CMP16, CMP20]. Indeed, the work of [CP15] studies the problem of obtaining ℓ_p subspace embeddings via sampling algorithms that simply sample rows proportionally to the Lewis weights in a "one-shot" manner analogous to leverage score sampling as in Theorem 1.3.3, rather than using a recursive sampling algorithm. In fact, such results are possible, and [CP15] obtain the following result:

Theorem 6.1.9 (ℓ_p Lewis weight sampling [CP15])**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $1 \leq p < \infty$. Let $\alpha > 0$ and let $q_i = \min\{1, \mathbf{w}_i^p(\mathbf{A})/\alpha\}$ for $i \in [n]$. Let $\mathbf{S} \in \mathbb{R}^{n \times n}$ be the ℓ_p sampling matrix with probabilities $\{q_i\}_{i=1}^n$ (Definition 6.1.1). Then, there is an α such that, with probability at least $99/100$, $\mathbf{S}$ is an ℓ_p subspace embedding satisfying Definition 1.1.1 with $\kappa = (1+\varepsilon)$, and furthermore, $\mathbf{S}$ has at most r nonzero rows, for

$$r = \begin{cases} O(\varepsilon^{-2} d \log(d/\varepsilon)) & p = 1 \\ O(\varepsilon^{-2} d \log(d/\varepsilon) \log\log(d/\varepsilon)) & 1 < p < 2 \\ O(\varepsilon^{-5} d^{p/2} (\log d) \log(1/\varepsilon)) & 2 < p < \infty \end{cases}$$

However, a notable gap exists between the algorithmic results of Theorem 6.1.9 based on "one-shot" sampling versus the existential results of Theorem 6.1.4 for $p > 2$ and its algorithmic version based on recursive sampling, where Theorem 6.1.4 achieves a quadratic dependence on ε, while Theorem 6.1.9 incurs a dependence of ε^5. An important question in the study of ℓ_p Lewis weight sampling is whether this gap can be closed:

Question 6.1.10. For $p > 2$, can the guarantee of one-shot ℓ_p Lewis weight sampling in Theorem 6.1.9 be improved to $\tilde{O}(\varepsilon^{-2} d^{p/2})$?

One of the main results we obtain in [WY23b] is a positive resolution to Question 6.1.10:

Theorem 6.1.11 (ℓ_p Lewis weight sampling, improved [WY23b])**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $2 < p < \infty$. Then, Theorem 6.1.9 holds with

$$r = O(\varepsilon^{-2} d^{p/2} (\log d)^2 \log(d/\varepsilon)).$$

6.2 Properties of one-sided ℓ_p Lewis weights

We collect some elementary properties of one-sided ℓ_p Lewis weights. We will extensively use the notion of Lewis bases, which is the change of basis matrix $\mathbf{R}$ such that $\mathbf{W}^{1/2-1/p}\mathbf{AR}$ is an orthonormal matrix.

The first lemma relates one-sided Lewis weights and Lewis bases.

Lemma 6.2.1. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $0 < p < \infty$. The following hold: Let $\mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights, and let $\mathbf{R}$ be the corresponding one-sided Lewis basis. Then, for each $i \in [n]$,

$$\mathbf{w}_i \geq \gamma^{p/2} \cdot \left\|\mathbf{e}_i^\top \mathbf{AR}\right\|_2^p.$$

Proof. We have that

$$\mathbf{w}_i \geq \gamma \cdot \boldsymbol{\tau}_i(\mathbf{W}^{1/2-1/p}\mathbf{A}) = \gamma \cdot \left\|\mathbf{e}_i^\top \mathbf{W}^{1/2-1/p}\mathbf{AR}\right\|_2^2 = \gamma \cdot \mathbf{w}_i^{1-2/p}\left\|\mathbf{e}_i^\top \mathbf{AR}\right\|_2^2$$

which rearranges to the desired result. □

We will also use the following two lemmas relating Lewis-reweighted ℓ_2 norms and ℓ_p norms.

Lemma 6.2.2. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $\mathbf{w}$ be γ-one-sided ℓ_p Lewis weights for $\mathbf{A}$. Then,

$$\left\|\mathbf{W}^{1/2-1/p}\mathbf{Ax}\right\|_2 \leq \begin{cases} \|\mathbf{w}\|_1^{1/2-1/p}\|\mathbf{Ax}\|_p & p > 2 \\ \gamma^{1/2-1/p}\|\mathbf{Ax}\|_p & p < 2 \end{cases}$$

Proof. Let $\mathbf{R} \in \mathbb{R}^{d\times d}$ be a change of basis matrix such that $\mathbf{W}^{1/2-1/p}\mathbf{AR}$ is orthonormal. If $p \geq 2$, then by Hölder's inequality,

$$\begin{aligned}\left\|\mathbf{W}^{1/2-1/p}\mathbf{ARx}\right\|_2^2 &= \sum_{i=1}^n \mathbf{w}_i^{1-2/p}\left[\mathbf{e}_i^\top \mathbf{ARx}\right]^2 \\ &\leq \left[\sum_{i=1}^n \mathbf{w}_i\right]^{1-2/p}\left[\sum_{i=1}^n \left|\mathbf{e}_i^\top \mathbf{ARx}\right|^p\right]^{2/p} = \|\mathbf{w}\|_1^{1-2/p}\|\mathbf{ARx}\|_p^2\end{aligned}$$

and if $p \leq 2$, then

$$\begin{aligned}\left\|\mathbf{W}^{1/2-1/p}\mathbf{ARx}\right\|_2^2 &= \sum_{i=1}^n \mathbf{w}_i^{1-2/p}\left[\mathbf{e}_i^\top \mathbf{ARx}\right]^{2-p}\left[\mathbf{e}_i^\top \mathbf{ARx}\right]^p \\ &\leq \sum_{i=1}^n \mathbf{w}_i^{1-2/p}\left\|\mathbf{e}_i^\top \mathbf{AR}\right\|_2^{2-p}\|\mathbf{x}\|_2^{2-p}\left[\mathbf{e}_i^\top \mathbf{ARx}\right]^p \qquad \text{Cauchy–Schwarz}\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \mathbf{w}_i^{1-2/p} \cdot (\mathbf{w}_i/\gamma^{p/2})^{2/p-1} \|\mathbf{x}\|_2^{2-p} \left[\mathbf{e}_i^\top \mathbf{A}\mathbf{R}\mathbf{x}\right]^p && \text{Lemma 6.2.1} \\
&= \gamma^{p/2-1} \|\mathbf{x}\|_2^{2-p} \|\mathbf{A}\mathbf{R}\mathbf{x}\|_p^p \\
&= \gamma^{p/2-1} \left\|\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}\mathbf{x}\right\|_2^{2-p} \|\mathbf{A}\mathbf{R}\mathbf{x}\|_p^p
\end{aligned}
$$

□

Lemma 6.2.3. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $\mathbf{w}$ be γ-one-sided ℓ_p Lewis weights for $\mathbf{A}$. Then,

$$
\|\mathbf{A}\mathbf{x}\|_p \leq \begin{cases} \gamma^{1/p-1/2} \left\|\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{x}\right\|_2 & p > 2 \\ \|\mathbf{w}\|_1^{1/p-1/2} \|\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{x}\|_2 & p < 2 \end{cases}
$$

Proof. Let $\mathbf{R} \in \mathbb{R}^{d\times d}$ be a change of basis matrix such that $\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{x}$ is orthonormal. If $p > 2$, then

$$
\begin{aligned}
\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p^p &= \sum_{i=1}^{n} |[\mathbf{A}\mathbf{R}\mathbf{x}](i)|^p = \sum_{i=1}^{n} |[\mathbf{A}\mathbf{R}\mathbf{x}](i)|^2 |[\mathbf{A}\mathbf{R}\mathbf{x}](i)|^{p-2} \\
&= \|\mathbf{x}\|_2^{p-2} \sum_{i=1}^{n} |[\mathbf{A}\mathbf{R}\mathbf{x}](i)|^2 \|\mathbf{e}_i^\top \mathbf{A}\mathbf{R}\|_2^{p-2} && \text{Cauchy-Schwarz} \\
&= \|\mathbf{x}\|_2^{p-2} \sum_{i=1}^{n} |[\mathbf{A}\mathbf{R}\mathbf{x}](i)|^2 (\mathbf{w}_i/\gamma^{p/2})^{1-2/p} && \text{Lemma 6.2.1} \\
&= \|\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}\mathbf{x}\|_2^{p-2} \sum_{i=1}^{n} |[\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)|^2 \gamma^{1-p/2} \\
&= \gamma^{1-p/2} \|\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}\mathbf{x}\|_2^p
\end{aligned}
$$

and

$$
\begin{aligned}
\|\mathbf{A}\mathbf{R}\mathbf{x}\|_p^p &= \sum_{i=1}^{n} |[\mathbf{A}\mathbf{R}\mathbf{x}](i)|^p = \sum_{i=1}^{n} \mathbf{w}^{1-p/2} |[\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}\mathbf{x}](i)|^p \\
&= \left[\sum_{i=1}^{n} \mathbf{w}_i\right]^{1-p/2} \|\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}\mathbf{x}\|_2^p && \text{Hölder's inequality}
\end{aligned}
$$

□

The next lemma uses the above result to bound ℓ_p sensitivities by one-sided Lewis weights.

Lemma 6.2.4 (One-sided Lewis weights bound sensitivities)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 < p < \infty$. Let $\mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights. Then,

$$
\sup_{\mathbf{x}\in \text{rowspan}(\mathbf{A})\setminus\{0\}} \frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \leq \begin{cases} \gamma^{-p/2} \|\mathbf{w}\|_1^{p/2-1} \cdot \mathbf{w}_i & p > 2 \\ \gamma^{-1} \cdot \mathbf{w}_i & p < 2 \end{cases}
$$

Proof. Let $\mathbf{R} \in \mathbb{R}^{d\times d}$ be a change of basis matrix such that $\mathbf{W}^{1/2-1/p}\mathbf{AR}$ is orthonormal. Then by Lemmas 6.2.1 and 6.2.2, we have

$$\frac{|[\mathbf{ARx}](i)|^p}{\|\mathbf{ARx}\|_p^p} \le \frac{\left\|\mathbf{e}_i^\top \mathbf{AR}\right\|_2^p \|\mathbf{x}\|_2^p}{\|\mathbf{ARx}\|_p^p} \le \frac{\|\mathbf{w}\|_1^{p/2-1}\|\mathbf{ARx}\|_p^p}{\gamma^{p/2}\|\mathbf{ARx}\|_p^p}\mathbf{w}_i = \gamma^{-p/2}\|\mathbf{w}\|_1^{p/2-1}\cdot\mathbf{w}_i$$

for $p > 2$ and

$$\frac{|[\mathbf{ARx}](i)|^p}{\|\mathbf{ARx}\|_p^p} \le \frac{\left\|\mathbf{e}_i^\top \mathbf{AR}\right\|_2^p \|\mathbf{x}\|_2^p}{\|\mathbf{ARx}\|_p^p} \le \frac{\gamma^{p/2-1}\|\mathbf{ARx}\|_p^p}{\gamma^{p/2}\|\mathbf{ARx}\|_p^p}\mathbf{w}_i = \gamma^{-1}\cdot\mathbf{w}_i$$

for $p < 2$. □

6.3 Analysis of ℓ_p Lewis weight sampling: reduction to a Rademacher process

We start off our analysis of ℓ_p Lewis weight sampling by a standard symmetrization argument (see Section 2.3 and Lemma 2.3.2). However, the Rademacher process given by Lemma 2.3.2 alone is still hard to analyze. Throughout Section 6.3, we will make a series of reductions to bound the process by successively "simpler" Rademacher processes.

6.3.1 Regularizing the Rademacher process

We first specialize our Rademacher process for sampling-based algorithms for ℓ_p subspace embeddings as well as related problems, such as ℓ_p affine embeddings and ℓ_p linear regression. Our end goal is to reduce the analysis to bounding a Rademacher process of the form of

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}'\mathbf{x}\|_p\le 1} \left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i |[\mathbf{A}'\mathbf{x}](i)|^p\right|^l \tag{6.2}$$

where $\mathbf{A}'$ is a matrix whose ℓ_p Lewis weights are uniformly bounded by $\alpha \approx \varepsilon^2$. These special properties about the Rademacher process will be necessary in the next step of the analysis when we bound the Rademacher process in Section 6.4.

Note that (6.2) differs from our original Rademacher process in Lemma 2.3.2 in two aspects. The first is that the matrix $\mathbf{A}'$ that appears in the objective function is the same matrix $\mathbf{A}'$ that defines the domain $X = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{A}'\mathbf{x}\|_p \le 1\}$. This allows us to use an ℓ_p norm bound on the objective function when analyzing this Rademacher process. Note that this does not hold a priori in the Rademacher process of Lemma 2.3.2, since the objective is reweighted by weights $\mathbf{w}_i$, whereas the domain only bounds the ℓ_p norm of $\mathbf{Ax}$ itself. The next lemma, based on [CP15], addresses this problem. The second aspect to address is the flatness of the ℓ_p Lewis weights of $\mathbf{A}'$. Intuitively, we expect this to be true since a row with ℓ_p Lewis weight $\mathbf{w}_i^p(\mathbf{A})$ is reweighted by at most $(\alpha/\mathbf{w}_i^p(\mathbf{A}))^{1/p}$, which would lead to an ℓ_p Lewis weight of at most α. This intuition will be formalized in Sections 6.3.2 and 6.3.3.

Lemma 6.3.1. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $X = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{A}\mathbf{x}\|_p^p \leq 1\}$. Furthermore, let $\mathbf{B} \in \mathbb{R}^{m\times d}$ and $C \geq 1$ be such that $\|\mathbf{B}\mathbf{x}\|_p^p \leq C\|\mathbf{A}\mathbf{x}\|_p^p$ for every $\mathbf{x} \in \mathbb{R}^d$. For every setting of the weights $\mathbf{w}$, let $\mathbf{S}_\mathbf{w}$ be the $n \times n$ sampling matrix with $(\mathbf{S}_\mathbf{w})_{i,i} = \mathbf{w}_i^{1/p}$ and let $\mathbf{B}_\mathbf{w}$ denote the $(n+m)\times d$ matrix obtained by a vertical concatenation of $\mathbf{S}_\mathbf{w}\mathbf{A}$ and $\mathbf{B}$. Suppose that

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{B}_\mathbf{w}\mathbf{x}\|_p^p \leq R} \left|\sum_{i=1}^n \mathbf{g}_i \mathbf{w}_i f_i(\mathbf{x})\right|^l \leq R^l \varepsilon^l \delta$$

for each fixed $\mathbf{w}$ and $R \geq 1$. Then, we have that

$$\mathop{\mathbf{E}}_{\mathbf{w}} \sup_{\mathbf{x}\in X} \left|\sum_{i=1}^n (\mathbf{w}_i - 1) f_i(\mathbf{x})\right|^l \leq \frac{(C+1)^l (2\sqrt{2\pi}\varepsilon)^l \delta}{1 - (2\varepsilon)^l \delta}$$

Proof. Fix a setting of the weights $\mathbf{w}$ and define

$$F_\mathbf{w} \coloneqq \sup_{\mathbf{x}\in X} \left|\sum_{i=1}^n (\mathbf{w}_i - 1)|[\mathbf{A}\mathbf{x}](i)|^p\right|$$

Then for any $\mathbf{x} \in \mathbb{R}^d$, we have that

$$\|\mathbf{S}_\mathbf{w}\mathbf{A}\mathbf{x}\|_p^p \leq (1 + F_\mathbf{w})\|\mathbf{A}\mathbf{x}\|_p^p$$

so $\|\mathbf{B}_\mathbf{w}\mathbf{x}\|_p^p \leq (C + 1 + F_\mathbf{w})\|\mathbf{A}\mathbf{x}\|_p^p$. Thus, we have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x}\|_p^p \leq 1} \left|\sum_{i=1}^n \mathbf{g}_i \mathbf{w}_i f_i(\mathbf{x})\right|^l &\leq \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{B}_\mathbf{w}\mathbf{x}\|_p^p \leq C+1+F_\mathbf{w}} \left|\sum_{i=1}^n \mathbf{g}_i \mathbf{w}_i f_i(\mathbf{x})\right|^l \\
&\leq (C + 1 + F_\mathbf{w})^l \cdot \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{B}_\mathbf{w}\mathbf{x}\|_p^p \leq 1} \left|\sum_{i=1}^n \mathbf{g}_i \mathbf{w}_i f_i(\mathbf{x})\right|^l \\
&\leq (C + 1 + F_\mathbf{w})^l \varepsilon^l \delta \leq ((C+1)^l + F_\mathbf{w}^l)(2\varepsilon)^l \delta
\end{aligned}$$

Then by Lemma 2.3.2,

$$\mathop{\mathbf{E}}_{\mathbf{w}}[F_\mathbf{w}^l] \leq \mathop{\mathbf{E}}_{\mathbf{w}}((C+1)^l + F_\mathbf{w}^l)(\sqrt{2\pi})^l (2\varepsilon)^l \delta = (C+1)^l (2\varepsilon)^l \delta + \mathop{\mathbf{E}}_{\mathbf{w}}[F_\mathbf{w}^l](2\sqrt{2\pi}\varepsilon)^l \delta$$

or

$$\mathop{\mathbf{E}}_{\mathbf{w}}[F_\mathbf{w}^l] \leq \frac{(C+1)^l (2\sqrt{2\pi}\varepsilon)^l \delta}{1 - (2\varepsilon)^l \delta}$$

□

6.3.2 Flattening the Rademacher process: $p < 2$

As discussed in Section 6.3.1, the next step in our analysis of ℓ_p Lewis weight sampling is to flatten the ℓ_p Lewis weights of the Rademacher process resulting from Lemma 2.3.2. We wish to argue that if we reweight a row $i \in [n]$ of $\mathbf{A}$ by $(\alpha/\mathbf{w}_i^p(\mathbf{A}))^{1/p}$, then the ℓ_p Lewis weight of the reweighted row increases by at most an $\alpha/\mathbf{w}_i^p(\mathbf{A})$ factor to a new ℓ_p Lewis weight of α. However, directly arguing as such as difficult, due to the recursive nature of the definition of ℓ_p Lewis weights.

The observation of [CP15] is that such an argument works if we concatenate the sampled matrix $\mathbf{SA}$ with $\mathbf{A}$. Note that this is exactly the matrix $\mathbf{B}_{\mathbf{w}}$ constructed in Lemma 2.3.2 with $\mathbf{B} = \mathbf{A}$. In this case, by (6.1), the ℓ_p Lewis weight of any row i of $\mathbf{B}_{\mathbf{w}}$ corresponding to a row $\mathbf{SA}$ satisfies

$$\begin{aligned}\mathbf{w}_i^p(\mathbf{B}_{\mathbf{w}}) &= \frac{\alpha}{\mathbf{w}_i^p(\mathbf{A})}\left(\mathbf{a}_i^\top(\mathbf{B}_{\mathbf{w}}\mathbf{W}(\mathbf{B}_{\mathbf{w}})^{1-2/p}\mathbf{B}_{\mathbf{w}})^{-}\mathbf{a}_i\right)^{p/2} \\ &\leq \frac{\alpha}{\mathbf{w}_i^p(\mathbf{A})}\left(\mathbf{a}_i^\top(\mathbf{A}\mathbf{W}(\mathbf{A})^{1-2/p}\mathbf{A})^{-}\mathbf{a}_i\right)^{p/2} \leq \alpha\end{aligned}$$

where $\mathbf{W}(\mathbf{B}_{\mathbf{w}})$ denotes the diagonal matrix of the ℓ_p Lewis weights of $\mathbf{B}_{\mathbf{w}}$, and $\mathbf{W}(\mathbf{A})$ denotes the ℓ_p Lewis weights for $\mathbf{A}$.

Two problems remain. The first is that if we concatenate $\mathbf{A}$ with $\mathbf{SA}$, then $\mathbf{A}$ may not have uniformly bounded ℓ_p Lewis weights, even if $\mathbf{SA}$ does. This can be addressed by *flattening* $\mathbf{A}$, that is, we take any row $i \in [n]$ of $\mathbf{A}$ with a large ℓ_p Lewis weight and replace it with k copies of $\mathbf{a}_i/k^{1/p}$. We will show that splitting a row into k copies reduces the ℓ_p Lewis weight of each of the rows by a factor of k, so we can take $k = 1/\alpha$ for every row to reduce the maximum ℓ_p Lewis weight of the flattened matrix to α. Furthermore, flattening does not change the Lewis quadratic $\mathbf{A}\mathbf{W}(\mathbf{A})^{1-2/p}\mathbf{A}$ and thus the argument above still holds.

Lemma 6.3.2 (Flattening ℓ_p Lewis weights)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 \leq \mathbf{w} \in \mathbb{R}^n$. Let $0 < p < \infty$. Let $\mathbf{A}' \in \mathbb{R}^{(n+k-1)\times d}$ be obtained by replacing some row $i \in [n]$ with k copies of $\mathbf{a}_i/k^{1/p}$ and let $\mathbf{w}' \in \mathbb{R}^{(n+k-1)\times d}$ be obtained by replacing $\mathbf{w}_i$ with k copies of $\mathbf{w}_i/k$. Then, $\|\mathbf{A}\mathbf{x}\|_p^p = \|\mathbf{A}'\mathbf{x}\|_p^p$ for every $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{A}^\top\mathbf{W}^{1-2/p}\mathbf{A} = \mathbf{A}'^\top\mathbf{W}'^{1-2/p}\mathbf{A}'$.

Proof. We have for every $\mathbf{x} \in \mathbb{R}^d$ that

$$|\langle \mathbf{a}_i, \mathbf{x}\rangle|^p = k\left|\langle \mathbf{a}_i/k^{1/p}, \mathbf{x}\rangle\right|^p$$

which shows that $\|\mathbf{A}\mathbf{x}\|_p^p = \|\mathbf{A}'\mathbf{x}\|_p^p$. Furthermore,

$$\mathbf{w}_i^{1-2/p}\mathbf{a}_i\mathbf{a}_i^\top = k\cdot\left(\frac{\mathbf{w}_i}{k}\right)^{1-2/p}\frac{\mathbf{a}_i}{k^{1/p}}\frac{\mathbf{a}_i^\top}{k^{1/p}}$$

which shows that $\mathbf{A}^\top\mathbf{W}^{1-2/p}\mathbf{A} = \mathbf{A}'^\top\mathbf{W}'^{1-2/p}\mathbf{A}'$. □

The second problem is that while flattening may reduce the ℓ_p Lewis weights of $\mathbf{A}$ alone, the ℓ_p Lewis weights may change when concatenated with $\mathbf{SA}$. Fortunately, for $p < 2$, it can be shown that the ℓ_p Lewis weights can in fact only decrease after concatenations (Lemma 6.3.3). Note that this property does *not* hold for $p > 2$, and thus we will need a different argument, which we show in Section 6.3.3.

Lemma 6.3.3 (Monotonicity of ℓ_p Lewis weights, Lemma 5.5, [CP15])**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 < p \leq 2$. Let $m \geq n$ and let $\mathbf{A}' \in \mathbb{R}^{m\times d}$ be a matrix such that $\mathbf{e}_i^\top \mathbf{A} = \mathbf{e}_i^\top \mathbf{A}'$ for all $i \in [n]$, that is, $\mathbf{A}'$ is obtained by adding rows to $\mathbf{A}$. Then, $\mathbf{w}_i^p(\mathbf{A}) \geq \mathbf{w}_i^p(\mathbf{A}')$ for every $i \in [n]$.

6.3.3 Flattening the Rademacher process: $p > 2$ [WY23b]

As discussed in Section 6.3.2, for $p > 2$, we need to overcome the lack of monotonicity of ℓ_p Lewis weights to flatten the Rademacher process. In the work [WY23b], we show how to circumvent the issue of non-monotonicity by directly constructing one-sided ℓ_p Lewis weights (Definition 6.1.7) for the concatenation of $\mathbf{SA}$ and the flattened version of $\mathbf{A}$ that still allows the argument from Section 6.3.2 to go through. In particular, we wish to construct one-sided ℓ_p Lewis weights such that the Lewis quadratic of the concatenated matrix is at least the Lewis quadratic $\mathbf{A}\mathbf{W}^{1-2/p}\mathbf{A}$, in order to argue that the ℓ_p Lewis weights of $\mathbf{SA}$ are at most α. The next lemma constructs such weights.

Lemma 6.3.4 (Batch online ℓ_p Lewis weights, $2 \leq p < \infty$)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $\mathbf{M} = \mathbf{L}^\top\mathbf{L} \in \mathbb{R}^{d\times d}$ be a symmetric PSD matrix, and let $2 \leq p < \infty$. There exists weights $\mathbf{w} \in \mathbb{R}^n$ such that for $i \in [n]$,

$$\mathbf{w}_i = \left(\frac{p}{2}\right)^{\frac{p/2}{1-2/p}} (\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A} + \mathbf{M})^{-1}\mathbf{a}_i)^{p/2}$$

and

$$\sum_{i=1}^n \mathbf{w}_i \leq \left(\frac{p}{2}\right)^{\frac{1}{1-2/p}} d.$$

Proof. Consider the following optimization problem over symmetric PSD matrices $\mathbf{Q}$:

$$\begin{aligned} \text{maximize} \quad & \det(\mathbf{Q}) \\ \text{subject to} \quad & \sum_{i=1}^n (\mathbf{a}_i^\top \mathbf{Q}\mathbf{a}_i)^{p/2} + \sum_{j=1}^d \mathbf{l}_j^\top \mathbf{Q}\mathbf{l}_j \leq d \\ & \mathbf{Q} \succeq 0 \end{aligned}$$

where $\mathbf{a}_i$ is the ith row of $\mathbf{A}$ and $\mathbf{l}_j$ is the jth row of $\mathbf{L}$. Let $\mathbf{Q}$ be any matrix which attains this maximum. Note then that

$$\sum_{i=1}^n (\mathbf{a}_i^\top \mathbf{Q}\mathbf{a}_i)^{p/2} + \sum_{j=1}^d \mathbf{l}_j^\top \mathbf{Q}\mathbf{l}_j = d$$

since otherwise scaling $\mathbf{Q}$ up can increase the objective function. Furthermore, by considering Lagrange multipliers, the gradient of the constraint is some scalar C times the gradient of of the objective, so

$$\sum_{i=1}^n \frac{p}{2}(\mathbf{a}_i^\top \mathbf{Q}\mathbf{a}_i)^{p/2-1}\mathbf{a}_i\mathbf{a}_i^\top + \sum_{j=1}^d \mathbf{l}_j\mathbf{l}_j^\top = C\det(\mathbf{Q})\mathbf{Q}^{-1}.$$

We now define

$$\mathbf{w}_i \coloneqq \left(\frac{p}{2}\right)^{\frac{1}{1-2/p}} (\mathbf{a}_i^\top \mathbf{Q}\mathbf{a}_i)^{p/2}.$$

Then, we have that

$$\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A} + \mathbf{M} = C\det(\mathbf{Q})\mathbf{Q}^{-1}$$

for $\mathbf{W} = \operatorname{diag}(\mathbf{w})$. Rearranging, we have that

$$\mathbf{Q} = C\det(\mathbf{Q})(\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A} + \mathbf{M})^{-1}$$

so

$$\mathbf{w}_i = \left(\frac{p}{2}\right)^{\frac{1}{1-2/p}} (\mathbf{a}_i \mathbf{Q} \mathbf{a}_i)^{p/2} = \left(\frac{p}{2}\right)^{\frac{1}{1-2/p}} (C\det(\mathbf{Q}))^{p/2} [\mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A} + \mathbf{M})^{-1}\mathbf{a}_i]^{p/2}$$

and thus

$$\begin{aligned}\mathbf{w}_i &= \left(\frac{p}{2}\right)^{\frac{2/p}{1-2/p}} (C\det(\mathbf{Q}))[(\mathbf{w}_i^{1/2-1/p}\mathbf{a}_i)^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A} + \mathbf{M})^{-1}(\mathbf{w}_i^{1/2-1/p}\mathbf{a}_i)] \\ &= \left(\frac{p}{2}\right)^{\frac{2/p}{1-2/p}} (C\det(\mathbf{Q}))\boldsymbol{\tau}_i(\mathbf{B})\end{aligned}$$

where $\mathbf{B}$ is the vertical concatenation of $\mathbf{W}^{1/2-1/p}\mathbf{A}$ and $\mathbf{L}$. Note also that for rows j corresponding to $\mathbf{L}$ in $\mathbf{B}$, we have that

$$(C\det(\mathbf{Q}))\boldsymbol{\tau}_j(\mathbf{B}) = (C\det(\mathbf{Q}))\mathbf{l}_j^\top (\mathbf{A}^\top \mathbf{W}^{1-2/p}\mathbf{A} + \mathbf{M})^{-1}\mathbf{l}_j = \mathbf{l}_j^\top \mathbf{Q}\mathbf{l}_j.$$

Now by the normalization constraint, we have that

$$\sum_{i=1}^n \left(\frac{2}{p}\right)^{\frac{1}{1-2/p}} \mathbf{w}_i + \sum_{j=1}^d \mathbf{l}_j^\top \mathbf{Q}\mathbf{l}_j = \sum_{i=1}^n (\mathbf{a}_i^\top \mathbf{Q}\mathbf{a}_i)^{p/2} + \sum_{j=1}^d \mathbf{l}_j^\top \mathbf{Q}\mathbf{l}_j = d.$$

However,

$$\left(\frac{2}{p}\right)^{\frac{1}{1-2/p}} \mathbf{w}_i = \left(\frac{p}{2}\right)^{\frac{-1}{1-2/p}} \left(\frac{p}{2}\right)^{\frac{2/p}{1-2/p}} (C\det(\mathbf{Q}))\boldsymbol{\tau}_i(\mathbf{B}) = \frac{2}{p}(C\det(\mathbf{Q}))\boldsymbol{\tau}_i(\mathbf{B})$$

so we must have that $p/2 = C\det(\mathbf{Q})$. The result follows. □

Remark 6.3.5. Note that if we set $\mathbf{M} = 0$ and redefine $\mathbf{w}_i' := \mathbf{w}_i/(p/2)^{\frac{1}{1-2/p}}$, then we will retrieve the usual definition of ℓ_p Lewis weights.

By setting $\mathbf{M} = \mathbf{A}\mathbf{W}^{1-2/p}\mathbf{A}$, we can use Lemma 6.3.4 to construct one-sided ℓ_p Lewis weights for the concatenation of $\mathbf{S}\mathbf{A}$ and the flattened version of $\mathbf{A}$ such that the Lewis quadratic is bounded below by $\mathbf{A}\mathbf{W}^{1-2/p}\mathbf{A}$, which makes the same argument as in Section 6.3.2 go through even for $p > 2$.

We make one final reduction of the Rademacher process by restricting to the set of rows $i \in [n]$ with significantly large ℓ_p Lewis weights.

Lemma 6.3.6. Let $J \supseteq \{i \in [n] : \boldsymbol{\sigma}_i^p(\mathbf{A}) \geq \varepsilon/n\}$. Then,

$$\left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i |[\mathbf{A}\mathbf{x}](i)|^p\right|^l \leq (2\varepsilon)^l + 2^l \left|\sum_{i\in J} \boldsymbol{\varepsilon}_i |[\mathbf{A}\mathbf{x}](i)|^p\right|^l$$

for any $\mathbf{x}$ such that $\|\mathbf{A}\mathbf{x}\|_p \leq 1$.

Proof. We have that

$$\left|\sum_{i\notin J}\boldsymbol{\varepsilon}_i|[\mathbf{Ax}](i)|^p\right| \leq \sum_{i\notin J}|\boldsymbol{\varepsilon}_i|[\mathbf{Ax}](i)|^p| \leq \sum_{i\notin J}\frac{\varepsilon}{n}\|\mathbf{Ax}\|_p^p \leq \varepsilon\|\mathbf{Ax}\|_p^p$$

which proves the claim. □

6.4 Analysis of ℓ_p Lewis weight sampling: Dudley's entropy integral

In the previous section, we have reduced our task to bounding a Rademacher process of the form of (6.2), where $\mathbf{A}'$ is a matrix whose ℓ_p Lewis weights are uniformly bounded by $\alpha \approx \varepsilon^2$. We will finally tackle the task of bounding this Rademacher process via Dudley's entropy integral. Our task is thus to estimate the entropy numbers $E(T, d_X, u)$ appearing in Theorem 2.3.6. Our calculations will be slightly more general than required for the analysis of ℓ_p Lewis weight sampling, to facilitate further applications in this thesis.

6.4.1 Bounds on the pseudo-metric

The Rademacher process that we study is indexed by the index set $T = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{Ax}\|_p \leq 1\}$, and is given by

$$X_{\mathbf{x}} = \sum_{i\in J}\boldsymbol{\varepsilon}_i|[\mathbf{Ax}](i)|^p$$

We will now estimate the pseudo-metric.

Lemma 6.4.1. Let $1 \leq p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Define the pseudo-metric

$$d_X(\mathbf{x},\mathbf{x}') := \left(\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n}\left|\sum_{i\in J}\boldsymbol{\varepsilon}_i|[\mathbf{Ax}](i)|^p - \sum_{i\in J}\boldsymbol{\varepsilon}_i|[\mathbf{Ax}'](i)|^p\right|^2\right)^{1/2}$$

Let $\sigma \geq \max_{i\in J}\boldsymbol{\sigma}_i^p(\mathbf{A})$. Then, for any $\mathbf{x},\mathbf{x}' \in \mathbb{R}^d$ such that $\|\mathbf{Ax}\|_p, \|\mathbf{Ax}'\|_p \leq 1$,

$$d_X(\mathbf{x},\mathbf{x}') \leq \begin{cases} 2\|(\mathbf{Ax}-\mathbf{Ax}')|_J\|_\infty^{p/2} & p<2 \\ 2p\cdot\sigma^{1/2-1/p}\cdot\|(\mathbf{Ax}-\mathbf{Ax}')|_J\|_\infty & p>2 \end{cases}$$

Proof. Note first that by expanding out the square and noting that $\mathbf{E}[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j] = \mathbb{1}(i=j)$, we have

$$d_X(\mathbf{x},\mathbf{x}') = \left(\sum_{i\in J}(|\mathbf{Ax}(i)|^p - |\mathbf{Ax}'(i)|^p)^2\right)^{1/2}$$

Let $\mathbf{y} = \mathbf{Ax}$ and $\mathbf{y}' = \mathbf{Ax}'$. For $p<2$, we bound this as

$$d_X(\mathbf{x},\mathbf{x}')^2 = \sum_{i\in J}(|\mathbf{y}(i)|^p - |\mathbf{y}'(i)|^p)^2$$

$$
\begin{aligned}
&= \sum_{i\in J} (|\mathbf{y}(i)|^{p/2} - |\mathbf{y}'(i)|^{p/2})^2 (|\mathbf{y}(i)|^{p/2} + |\mathbf{y}'(i)|^{p/2})^2 \\
&\leq \sum_{i\in J} (|\mathbf{y}(i) - \mathbf{y}'(i)|^{p/2})^2 (|\mathbf{y}(i)|^{p/2} + |\mathbf{y}'(i)|^{p/2})^2 \\
&\leq 2\|(\mathbf{y}-\mathbf{y}')|_J\|_\infty^p \sum_{i\in J} |\mathbf{y}(i)|^p + |\mathbf{y}'(i)|^p \\
&\leq 4\|(\mathbf{y}-\mathbf{y}')|_J\|_\infty^p.
\end{aligned}
$$

For $p > 2$, we have by convexity that

$$
|\mathbf{y}(i)|^p - |\mathbf{y}'(i)|^p \leq p|(\mathbf{y}(i) - \mathbf{y}'(i))|_J|(|\mathbf{y}(i)|^{p-1} + |\mathbf{y}'(i)|^{p-1})
$$

and that $\|\mathbf{y}|_J\|_\infty \leq \sigma^{1/p}$, so we have

$$
\begin{aligned}
d_X(\mathbf{x},\mathbf{x}')^2 &= \sum_{i\in J} (|\mathbf{y}(i)|^p - |\mathbf{y}'(i)|^p)^2 \\
&\leq p^2 \sum_{i\in J} |\mathbf{y}(i) - \mathbf{y}'(i)|^2 (|\mathbf{y}(i)|^{p-1} + |\mathbf{y}'(i)|^{p-1})^2 \\
&\leq 2p^2 \|(\mathbf{y}-\mathbf{y}')|_J\|_\infty^2 \sum_{i=1}^n |\mathbf{y}(i)|^{2p-2} + |\mathbf{y}'(i)|^{2p-2} \\
&\leq 2p^2 \max\{\|\mathbf{y}|_J\|_\infty, \|\mathbf{y}'|_J\|_\infty\}^{p-2} \|(\mathbf{y}-\mathbf{y}')|_J\|_\infty^2 \sum_{i=1}^n |\mathbf{y}(i)|^p + |\mathbf{y}'(i)|^p \\
&\leq 4p^2 \sigma^{1-2/p} \|(\mathbf{y}-\mathbf{y}')|_J\|_\infty^2.
\end{aligned}
$$

□

From the above lemma, we also immediately obtain diameter bounds.

Lemma 6.4.2. Let $1 \leq p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\sigma \geq \max_{i\in J} \boldsymbol{\sigma}_i^p(\mathbf{A})$. Then, the diameter of $T = \{\mathbf{x} : \|\mathbf{A}\mathbf{x}\|_p \leq 1\}$ with respect to d_X is bounded by

$$
\operatorname{diam}(T) \leq \begin{cases} 4 \cdot \sigma^{1/2} & p < 2 \\ 4p \cdot \sigma^{1/2} & p > 2 \end{cases}
$$

Proof. For any $\mathbf{y} = \mathbf{A}\mathbf{x}$ with $\|\mathbf{A}\mathbf{x}\|_p \leq 1$, we have that $\|\mathbf{y}|_J\|_\infty \leq \sigma^{1/p}$, so combining the triangle inequality and Lemma 6.4.1 yields the result. □

6.4.2 Entropy bounds

With bounds on the pseudo-metric d_X in hand, we can estimate the entropy numbers $E(T, d_X, u)$ as required by Theorem 2.3.6. The bounds in this section are taken from [WY23c], which in turn follows [BLM89]. We first introduce the dual Sudakov minoration theorem, which is a general tool for bounding covering numbers of the Euclidean ball.

Definition 6.4.3 (Levy mean)**.** The Levy mean is defined as

$$
M_X = \int_{\mathbb{S}^{d-1}} \|\mathbf{x}\| \, d\sigma(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathbb{S}^{d-1}} \|\mathbf{x}\|.
$$

Remark 6.4.4. By noting that $\mathbf{x} \sim \mathbb{S}^{d-1}$ is the same as drawing a Gaussian vector and normalizing, that is,

$$M_X = \mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_d)} \left\| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \right\| = \frac{\mathbf{E}\|\mathbf{g}\|_2}{\mathbf{E}\|\mathbf{g}\|_2} \mathbf{E} \left\| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \right\| = \frac{1}{\mathbf{E}\|\mathbf{g}\|_2} \mathbf{E}\|\mathbf{g}\|$$

since the norm of the Gaussian is independent of its direction.

Lemma 6.4.5 (Dual Sudakov minoration (Proposition 4.2, [BLM89]))**.** Let $(X, \|\cdot\|)$ be Banach space on $\mathbb{R}^d$ and let be the Levy mean of $\|\cdot\|$. Then, for some constant $C > 0$, we have that

$$\log E(B_2, t \cdot B_X) \leq C \cdot d \left(\frac{M_X}{t} \right)^2$$

where $B_2 = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\}$ and $B_X = \{\mathbf{x} : \|\mathbf{x}\| \leq 1\}$.

We will compute the above Levy mean bound for reweighted ℓ_q norms, defined below.

Definition 6.4.6. Let $0 \leq \mathbf{w} \in \mathbb{R}^n$. We define the $\mathbf{w}$-weighted ℓ_q norm by

$$\|\mathbf{y}\|_{\mathbf{w},q} := \left(\sum_{i=1}^n \mathbf{w}_i |\mathbf{y}(i)|^q \right)^{1/q}.$$

For a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $B^q_{\mathbf{w}}(\mathbf{A}) = \{\mathbf{x} : \|\mathbf{A}\mathbf{x}\|_{\mathbf{w},q} \leq 1\}$ denote the corresponding unit ball in the column space of $\mathbf{A}$. If $\mathbf{w} = 1$, then we simply write $B^q(\mathbf{A})$.

Note that $\|\mathbf{y}\|_p = \|\mathbf{W}^{-1/p}\mathbf{y}\|_{\mathbf{w},p}$ for $\mathbf{W} = \mathrm{diag}(\mathbf{w})$, so we can instead prove bounds under these reweighted norms, as long as we apply $\mathbf{W}^{-1/p}$ first. We then have the following Levy mean bound.

Lemma 6.4.7. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $1 \geq \mathbf{w} \in \mathbb{R}^n$ be nonnegative weights. Let $\tau \geq \max_{i=1}^n \|\mathbf{e}_i^\top \mathbf{A}\|_2^2$. Let $1 \leq q < \infty$. Then,

$$\mathop{\mathbf{E}}_{\mathbf{g}} \|\mathbf{A}\mathbf{g}\|_{\mathbf{w},q} \leq n^{1/q}\sqrt{q\tau}$$

Proof. We have that

$$\mathop{\mathbf{E}}_{\mathbf{g}}[|[\mathbf{A}\mathbf{g}](i)|^q] = \frac{2^{q/2}\Gamma(\frac{q+1}{2})}{\sqrt{\pi}} \|\mathbf{e}_i^\top \mathbf{A}\|_2^q \leq q^{q/2} \cdot \|\mathbf{e}_i^\top \mathbf{A}\|_2^q \leq q^{q/2} \cdot \tau^{q/2}$$

Then by Jensen's inequality and linearity of expectation, we have

$$\mathop{\mathbf{E}}_{\mathbf{g}} \|\mathbf{A}\mathbf{g}\|_{\mathbf{w},q} \leq \left(\mathop{\mathbf{E}}_{\mathbf{g}} \|\mathbf{A}\mathbf{g}\|_{\mathbf{w},q}^q \right)^{1/q} \leq (n \cdot q^{q/2} \cdot \tau^{q/2})^{1/q} = n^{1/q}\sqrt{q\tau}.$$

□

By combining the above calculation with Lemma 6.4.5, we obtain the following:

Corollary 6.4.8. Let $1 \geq \mathbf{w} \in \mathbb{R}^n$ be nonnegative weights. Let $2 \leq q < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be such that $\mathbf{W}^{1/2}\mathbf{A}$ is orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$. Then,

$$\log E(B_{\mathbf{w}}^2(\mathbf{A}), B_{\mathbf{w}}^q(\mathbf{A}), t) \leq O(1)\frac{n^{2/q} q \cdot \tau}{t^2}$$

Proof. For $\mathbf{W}^{1/2}\mathbf{A}$ orthonormal, $B_{\mathbf{w}}^2(\mathbf{A}) = B^2(\mathbf{W}^{1/2}\mathbf{A})$ is isometric to the Euclidean ball in d dimensions. Thus Lemma 6.4.5 applies. □

We also get a similar result for $q = \infty$, by applying Corollary 6.4.8 with $q = O(\log n)$.

Corollary 6.4.9. Let $1 \geq \mathbf{w} \in \mathbb{R}^n$ be nonnegative weights with $\min_{i \in [n]} \mathbf{w}_i \geq \varepsilon$. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be such that $\mathbf{W}^{1/2}\mathbf{A}$ is orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$. Then,

$$\log E(B_{\mathbf{w}}^2(\mathbf{A}), B^\infty(\mathbf{A}), t) \leq O(1)\frac{\log(n/\varepsilon) \cdot \tau}{t^2}$$

Proof. This follows from the fact that for $\mathbf{y} \in \mathbb{R}^n$,

$$\Omega(1)\|\mathbf{y}\|_\infty = \varepsilon^{1/q}\|\mathbf{y}\|_\infty \leq \|\mathbf{y}\|_{\mathbf{w},q} \leq n^{1/q}\|\mathbf{y}\|_\infty = O(1)\|\mathbf{y}\|_\infty$$

for $q = O(\log(n/\varepsilon))$. □

By interpolation, we can improve the bound in Corollary 6.4.8, which is needed for our results for $p < 2$:

Lemma 6.4.10. Let $2 < r < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$. Let $1 \leq t \leq \mathrm{poly}(d)$. Then,

$$\log E(B^2(\mathbf{A}), B^r(\mathbf{A}), t) \leq O(1)\frac{1}{(t/2)^{2r/(r-2)}} \cdot \left(\frac{r}{r-2}\log d + \log n\right)\tau$$

Proof. Let $q > r$, and let $0 < \theta < 1$ satisfy

$$\frac{1}{r} = \frac{1-\theta}{2} + \frac{\theta}{q}$$

Then by Hölder's inequality, we have for any $\mathbf{y} \in \mathbb{R}^n$ that

$$\|\mathbf{y}\|_r = \left(\sum_{i=1}^n |\mathbf{y}(i)|^{r(1-\theta)}|\mathbf{y}(i)|^{r\theta}\right)^{1/r} \leq \left(\sum_{i=1}^n |\mathbf{y}(i)|^2\right)^{(1-\theta)/2}\left(\sum_{i=1}^n |\mathbf{y}(i)|^q\right)^{\theta/q} = \|\mathbf{y}\|_2^{1-\theta}\|\mathbf{y}\|_q^\theta$$

Then for any $\mathbf{y}, \mathbf{y}' \in B^2$, we have

$$\|\mathbf{y} - \mathbf{y}'\|_r \leq \|\mathbf{y} - \mathbf{y}'\|_2^{1-\theta}\|\mathbf{y} - \mathbf{y}'\|_q^\theta \leq 2\|\mathbf{y} - \mathbf{y}'\|_q^\theta$$

so

$$\log E(B^2(\mathbf{A}), B^r(\mathbf{A}), t) \leq \log E(B^2(\mathbf{A}), B^q(\mathbf{A}), (t/2)^{1/\theta}) \leq O(1)\frac{n^{2/q} q \cdot \tau}{(t/2)^{2/\theta}}$$

by Corollary 6.4.8. Now, we have

$$\frac{2}{\theta} = 2\frac{\frac{1}{2}-\frac{1}{q}}{\frac{1}{2}-\frac{1}{r}} = \frac{q-2}{q}\frac{2r}{r-2}$$

so by taking $q = O(\frac{r}{r-2}\log d + \log n)$, we have that $n^{2/q} = O(1)$ and $(t/2)^{1/\theta} = \Theta(1)(t/2)^{2r/(r-2)}$, so we conclude as claimed. □

Using Lemma 6.4.10, we obtain the following analogue of Corollary 6.4.8 for $p < 2$.

Lemma 6.4.11. Let $1 \geq \mathbf{w} \in \mathbb{R}^n$ be nonnegative weights. Let $0 < p < 2$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be such that $\mathbf{W}^{1/2}\mathbf{A}$ is orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{W}^{1/2}\mathbf{A}\right\|_2^2$. Then,

$$\log E(B_{\mathbf{w}}^p(\mathbf{A}), B^\infty(\mathbf{A}), t) \leq O(1)\frac{1}{t^p}\left(\frac{\log d}{2-p} + \log n\right)\tau.$$

Proof. In order to bound a covering of $B_{\mathbf{w}}^p(\mathbf{A})$ by $B^\infty(\mathbf{A})$, we first cover $B_{\mathbf{w}}^p(\mathbf{A})$ by $B_{\mathbf{w}}^2(\mathbf{A})$, and then use Corollary 6.4.9 to cover $B_{\mathbf{w}}^2(\mathbf{A})$ by $B^\infty(\mathbf{A})$.

We will first bound $E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), t)$ using Lemma 6.4.10. For each $k \geq 0$, let $\mathcal{E}_k \subseteq B_{\mathbf{w}}^p(\mathbf{A})$ be a maximal subset of $B_{\mathbf{w}}^p(\mathbf{A})$ such that for each distinct $\mathbf{y}, \mathbf{y}' \in \mathcal{E}_k$, $\|\mathbf{y}-\mathbf{y}'\|_{\mathbf{w},2} > 8^k t$, with $\mathcal{E}_k := \{0\}$ for $8^{k+1}t > n^{1/p-1/q}$. Note then that

$$|\mathcal{E}_k| \geq E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), 8^k t).$$

By averaging, for each k, there exists $\mathbf{y}^{(k)} \in \mathcal{E}_k$ such that if

$$\mathcal{F}_k := \left\{\mathbf{y} \in \mathcal{E}_k : \|\mathbf{y}-\mathbf{y}^{(k)}\|_{\mathbf{w},2} \leq 8^{k+1}t\right\},$$

then

$$|\mathcal{F}_k| \geq \frac{|\mathcal{E}_k|}{E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), 8^{k+1}t)} \geq \frac{E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), 8^k t)}{E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), 8^{k+1}t)}$$

We now use this observation to construct an $\ell_{p'}$-packing of $B_{\mathbf{w}}^2(\mathbf{A})$, where p' is the Hölder conjugate of p. Let

$$\mathcal{G}_k := \left\{\frac{1}{8^{k+1}t}(\mathbf{y}-\mathbf{y}^{(k)}) : \mathbf{y} \in \mathcal{F}_k\right\}.$$

Then, $\mathcal{G}_k \subseteq B_{\mathbf{w}}^2(\mathbf{A})$ and $\mathcal{G}_k \subseteq B_{\mathbf{w}}^p(\mathbf{A}) \cdot 2/8^{k+1}t$, and $\|\mathbf{y}-\mathbf{y}'\|_{\mathbf{w},2} > 1/8$ for every distinct $\mathbf{y}, \mathbf{y}' \in \mathcal{G}_k$. Then by Hölder's inequality,

$$\frac{1}{8^2} \leq \|\mathbf{y}-\mathbf{y}'\|_{\mathbf{w},2}^2 \leq \|\mathbf{y}-\mathbf{y}'\|_{\mathbf{w},p}\|\mathbf{y}-\mathbf{y}'\|_{\mathbf{w},p'} \leq \frac{4}{8^{k+1}t}\|\mathbf{y}-\mathbf{y}'\|_{\mathbf{w},p'}$$

so $\|\mathbf{y}-\mathbf{y}'\|_{\mathbf{w},p'} \geq 2\cdot 8^{k-2}t$. Thus, $\mathcal{G}_k$ is an $\ell_{p'}$-packing of $B_{\mathbf{w}}^2(\mathbf{A})$, so

$$\begin{aligned}\log E(B_{\mathbf{w}}^2(\mathbf{A}), B_{\mathbf{w}}^{p'}(\mathbf{A}), 8^{k-2}t) &\geq \log|\mathcal{G}_k| = \log|\mathcal{F}_k| \\ &\geq \log E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), 8^k t) - \log E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), 8^{k+1}t).\end{aligned} \tag{6.3}$$

Summing over k gives

$$\begin{aligned}
&\log E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), t) \\
=\ &\sum_{k\geq 0} \log E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), 8^k t) - \log E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), 8^{k+1} t) \\
\leq\ &\sum_{k\geq 0} \log E(B_{\mathbf{w}}^2(\mathbf{A}), B_{\mathbf{w}}^{p'}(\mathbf{A}), 8^{k-2} t) && (6.3) \\
\leq\ &O(1)\frac{1}{(t/2)^{2p'/(p'-2)}} \cdot \left(\frac{p'}{p'-2}\log d + \log n\right)\tau && \text{Lem. 6.4.10, Cor. 6.4.9} \\
=\ &O(1)\frac{1}{(t/2)^{2p/(2-p)}} \cdot \left(\frac{p}{2-p}\log d + \log n\right)\tau
\end{aligned}$$

where we take $p'/(p'-2) = 1$ for $p' = \infty$. Using this and Corollary 6.4.9, we now bound

$$\begin{aligned}
\log E(B_{\mathbf{w}}^p(\mathbf{A}), B^\infty(\mathbf{A}), t) &\leq \log E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^2(\mathbf{A}), \lambda) + \log E(B_{\mathbf{w}}^2(\mathbf{A}), B^\infty(\mathbf{A}), t/\lambda) \\
&\leq O(1)\frac{1}{(\lambda/2)^{2p/(2-p)}} \cdot \left(\frac{p}{2-p}\log d + \log n\right)\tau + O(1)\frac{(\log n)\cdot\tau}{(t/\lambda)^2}
\end{aligned}$$

for any $\lambda \in [1, t]$. We choose λ satisfying

$$\frac{1}{(\lambda/2)^{2p/(2-p)}} = \frac{(\lambda/2)^2}{t^2},$$

which gives

$$(\lambda/2)^{2p/(2-p)} = \left(t^2\right)^{\frac{2p/(2-p)}{2+2p/(2-p)}} = t^p$$

so we obtain a bound of

$$O(1)\frac{1}{t^p}\left(\frac{1}{2-p}\log d + \log n\right)\tau.$$

□

6.4.3 Entropy integral for ℓ_p Lewis weight sampling

We will now specialize the general results derived previous to the case of ℓ_p Lewis weight sampling. For this setting, we will make use of reweighted ℓ_p norms. We first translate our pseudo-metric bounds to this reweighted setting.

Lemma 6.4.12. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $0 \leq \mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights. Let $J = \{i \in [n] : \mathbf{w}_i \geq \varepsilon/n\} \supseteq \{i \in [n] : \boldsymbol{\sigma}_i^p(\mathbf{A}) \geq \varepsilon/n\}$. Let $w = \max_{i\in J} \mathbf{w}_i$ and $\mathbf{W} = \operatorname{diag}(\mathbf{w})$. Let d_X be the pseudo-metric as defined in Lemma 6.4.1. Then for $q = O(\log(n/\varepsilon))$,

$$d_X(\mathbf{x}, \mathbf{x}') \leq \begin{cases} 2w^{1/2}\left\|\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}'\right\|_{\mathbf{w},q}^{p/2} & p < 2 \\ 2pw^{1/2}\cdot(\|\mathbf{w}\|_1^{p/2-1}/\gamma^{p/2})^{1/2-1/p}\left\|\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}'\right\|_{\mathbf{w},q} & p > 2 \end{cases}$$

and

$$\operatorname{diam}(B^p(\mathbf{A})) \le \begin{cases} 4 \cdot (w/\gamma)^{1/2} & p < 2 \\ 4p \cdot (\gamma^{-p/2}\|\mathbf{w}\|_1^{p/2-1} w)^{1/2} & p > 2 \end{cases}$$

Proof. We have that

$$\begin{aligned} \|(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}')|_J\|_\infty &= \left\|\mathbf{W}^{1/p}(\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}')|_J\right\|_\infty \\ &\le w^{1/p}\left\|(\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}')|_J\right\|_\infty \end{aligned}$$

and $\boldsymbol{\sigma}_i^p(\mathbf{A}) \le \gamma^{-p/2}\|\mathbf{w}\|_1^{p/2-1} w$ for $p > 2$ and $\boldsymbol{\sigma}_i^p(\mathbf{A}) \le \gamma^{-1} w$ for $p < 2$ by Lemma 6.2.4. Furthermore, note that $\mathbf{w}_i \ge \varepsilon/n$ for each $i \in J$, so

$$\frac{n}{\varepsilon}\|\mathbf{y}|_J\|_{\mathbf{w},q}^q \ge \|\mathbf{y}|_J\|_q^q \ge \|\mathbf{y}|_J\|_\infty^q$$

so for $q = O(\log(n/\varepsilon))$, we have

$$\|(\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}')|_J\|_\infty \le 2\|\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}'\|_{\mathbf{w},q}$$

Plugging these results into Lemmas 6.4.1 and 6.4.2 yields the desired result. □

Next, we obtain entropy bounds using our lemmas from Section 6.4.2 for covering the index set $T = B^p(\mathbf{A}) = B_{\mathbf{w}}^p(\mathbf{W}^{-1/p}\mathbf{A})$ of the Rademacher process by $B_{\mathbf{w}}^q(\mathbf{W}^{-1/p}\mathbf{A})$ (unit balls of the $\|\cdot\|_{\mathbf{w},q}$ norm), as required by Lemma 6.4.12.

Lemma 6.4.13 (Entropy bounds, $p > 2$)**.** Let $2 < p < \infty$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $0 \le \mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights. Let $w = \max_{i\in[n]} \mathbf{w}_i$. Then,

$$\log E(B^p(\mathbf{A}), d_X, t) \le O(p^2 w)\frac{\log(n/\varepsilon)}{\gamma^{p/2}t^2}\|\mathbf{w}\|_1^{p/2-1}.$$

Proof. We first bound

$$\log E(B^p(\mathbf{A}), d_X, t) \le \log E(B^p(\mathbf{A}), K\|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}, t)$$

for $K = 2pw^{1/2}(\|\mathbf{w}\|_1^{p/2-1}/\gamma^{p/2})^{1/2-1/p}$ by Lemma 6.4.12. We have by Lemma 6.2.2 that

$$B^p(\mathbf{A}) \subseteq \|\mathbf{w}\|_1^{1/2-1/p} \cdot B_{\mathbf{w}}^2(\mathbf{W}^{-1/p}\mathbf{A})$$

so we then have

$$\begin{aligned} & \log E(B^p(\mathbf{A}), K\|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}, t) \\ \le\ & \log E(\|\mathbf{w}\|_1^{1/2-1/p} \cdot B_{\mathbf{w}}^2(\mathbf{W}^{-1/p}\mathbf{A}), K\|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}, t) \\ \le\ & \log E(B_{\mathbf{w}}^2(\mathbf{W}^{-1/p}\mathbf{A}), \|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}, t/K\|\mathbf{w}\|_1^{1/2-1/p}). \end{aligned}$$

Note that the entropy bounds do not change if we replace $\mathbf{A}$ by $\mathbf{A}\mathbf{R}$, where $\mathbf{R}$ is the change of basis matrix such that $\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}$ is orthonormal. Then by the properties of γ-one-sided ℓ_p Lewis weights (Lemma 6.2.1), we have

$$\|\mathbf{e}_i^\top \mathbf{W}^{-1/p}\mathbf{A}\mathbf{R}\|_2^2 = \mathbf{w}_i^{-2/p}\|\mathbf{e}_i^\top \mathbf{A}\mathbf{R}\|_2^2 \le \gamma^{-1}.$$

We can then apply Corollary 6.4.8 to bound

$$\begin{aligned}
&\log E(B_{\mathbf{w}}^2(\mathbf{W}^{-1/p}\mathbf{A}), \|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}, t/K\|\mathbf{w}\|_1^{1/2-1/p}) \\
\leq\ & O(1)\frac{n^{2/q}q\cdot\gamma^{-1}}{t^2}(K\|\mathbf{w}\|_1^{1/2-1/p})^2 \\
\leq\ & O(p^2)\frac{\log(n/\varepsilon)\cdot\gamma^{-1}w}{t^2}(\|\mathbf{w}\|_1/\gamma)^{p/2-1} \\
\leq\ & O(p^2w)\frac{\log(n/\varepsilon)}{\gamma^{p/2}t^2}\|\mathbf{w}\|_1^{p/2-1}
\end{aligned}$$

□

Lemma 6.4.14 (Entropy bounds, $p<2$)**.** Let $0<p<2$. Let $\mathbf{A}\in\mathbb{R}^{n\times d}$ and let $0\leq\mathbf{w}\in\mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights. Let $w=\max_{i\in[n]}\mathbf{w}_i$ and $\mathbf{W}=\operatorname{diag}(\mathbf{w})$. Then,

$$\log E(B^p(\mathbf{A}), d_X, t)\leq O(1)\frac{w}{\gamma t^2}\left(\frac{\log d}{2-p}+\log n\right).$$

Proof. We first bound

$$\begin{aligned}
\log E(B^p(\mathbf{A}), d_X, t) &\leq \log E(B^p(\mathbf{A}), K\|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}^{p/2}, t) \\
&= \log E(B^p(\mathbf{A}), \|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}, (t/K)^{2/p})
\end{aligned}$$

for $K=2w^{1/2}$ by Lemma 6.4.12. Note that the entropy bounds do not change if we replace $\mathbf{A}$ by $\mathbf{AR}$, where $\mathbf{R}$ is the change of basis matrix such that $\mathbf{W}^{1/2-1/p}\mathbf{AR}$ is orthonormal. Then by the properties of γ-one-sided ℓ_p Lewis weights (Lemma 6.2.1), we have

$$\|\mathbf{e}_i^\top\mathbf{W}^{-1/p}\mathbf{AR}\|_2^2=\mathbf{w}_i^{-2/p}\|\mathbf{e}_i^\top\mathbf{AR}\|_2^2\leq\gamma^{-1}.$$

Then by Lemma 6.4.11, we have that

$$\log E(B^p(\mathbf{A}), \|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}, (t/K)^{2/p})\leq O(1)\frac{w}{\gamma t^2}\left(\frac{\log d}{2-p}+\log n\right).$$

□

We may now evaluate the entropy integral required in Theorem 2.3.6. We use the following calculus lemma:

Lemma 6.4.15. Let $0<\lambda\leq 1$. Then,

$$\int_0^\lambda\sqrt{\log\frac{1}{t}}\,dt=\lambda\sqrt{\log(1/\lambda)}+\frac{\sqrt{\pi}}{4}\operatorname{erfc}(\sqrt{\log(1/\lambda)})\leq\lambda\left(\sqrt{\log(1/\lambda)}+\frac{\sqrt{\pi}}{2}\right)$$

Proof. We calculate

$$\int_0^\lambda\sqrt{\log\frac{1}{t}}\,dt=2\int_{\sqrt{\log(1/\lambda)}}^\infty x^2\exp(-x^2)\,dx \qquad x=\sqrt{\log(1/t)}$$

$$
\begin{aligned}
&= -\int_{\sqrt{\log(1/\lambda)}}^{\infty} x \cdot -2x \exp(-x^2)\, dx \\
&= -\left(x \exp(-x^2)\Big|_{\sqrt{\log(1/\lambda)}}^{\infty} - \int_{\sqrt{\log(1/\lambda)}}^{\infty} \exp(-x^2)\, dx \right) \quad \text{integration by parts} \\
&= \lambda\sqrt{\log\frac{1}{\lambda}} + \frac{\sqrt{\pi}}{2}\operatorname{erfc}\left(\sqrt{\log\frac{1}{\lambda}}\right)
\end{aligned}
$$

□

Lemma 6.4.16 (Entropy integral bound for $p < 2$)**.** Let $0 < p < 2$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $0 \le \mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights. Let $w = \max_{i\in[n]} \mathbf{w}_i$. Then,

$$
\int_0^\infty \sqrt{\log E(B^p(\mathbf{A}), d_X, t)}\, dt \le O((w/\gamma)^{1/2})\left(\frac{\log d}{2-p} + \log n\right)^{1/2} \log d
$$

Proof. Note that it suffices to integrate the entropy integral to $\operatorname{diam}(B^p(\mathbf{A}))$ rather than ∞, which is at most $4(w/\gamma)^{1/2}$ for $p < 2$ by Lemma 6.4.12.

For small radii less than λ for a parameter λ to be chosen, we use a standard volume argument, which shows that

$$
\log E(B^p(\mathbf{A}), d_X, t) \le O(d)\log\frac{n}{t}
$$

so

$$
\begin{aligned}
\int_0^\lambda \sqrt{\log E(B^p(\mathbf{A}), d_X, t)}\, dt &= \int_0^\lambda \sqrt{d\log\frac{n}{t}}\, dt \\
&\le \lambda\sqrt{d\log n} + \sqrt{d}\int_0^\lambda \sqrt{\log\frac{1}{t}}\, dt \\
&\le \lambda\sqrt{d\log n} + \sqrt{d}\left(\lambda\sqrt{\log\frac{1}{\lambda}} + \frac{\sqrt{\pi}}{2}\lambda\right) \quad \text{Lemma 6.4.15} \\
&\le O(\lambda)\sqrt{d\log\frac{n}{\lambda}}
\end{aligned}
$$

On the other hand, for large radii larger than λ, we use the bounds of Lemma 6.4.14, which gives

$$
\log E(B^p(\mathbf{A}), d_X, t) \le O(1)\frac{1}{t^2}\left(\frac{\log d}{2-p} + \log n\right)(w/\gamma)
$$

so the entropy integral gives a bound of

$$
\begin{aligned}
&O(1)\left[\left(\frac{\log d}{2-p} + \log n\right)(w/\gamma)\right]^{1/2} \int_\lambda^{4(w/\gamma)^{1/2}} \frac{1}{t}\, dt \\
=\ &O(1)\left[\left(\frac{\log d}{2-p} + \log n\right)(w/\gamma)\right]^{1/2} \log\frac{4p(w/\gamma)^{1/2}}{\lambda}.
\end{aligned}
$$

We choose $\lambda = \sqrt{w/\gamma d}$, which yields the claimed conclusion. □

An analogous result and proof holds for $p > 2$.

Lemma 6.4.17 (Entropy integral bound for $p > 2$)**.** Let $2 < p < \infty$. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let $0 \leq \mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights. Let $w = \max_{i \in [n]} \mathbf{w}_i$. Then,

$$\int_0^\infty \sqrt{\log E(B^p(\mathbf{A}), d_X, t)}\, dt \leq O(pw^{1/2})(\|\mathbf{w}\|_1^{p/2-1}/\gamma^{p/2})^{1/2}(\log(n/\varepsilon))^{1/2} \log d$$

Proof. The proof is similar to the case of $p < 2$. We again introduce a parameter λ. For radii below λ, the bound is the same as Lemma 6.4.16. For radii above λ, we use Lemma 6.4.13 to bound

$$\log E(B^p(\mathbf{A}), d_X, t) \leq O(p^2 w) \frac{\log(n/\varepsilon)}{\gamma^{p/2} t^2} \|\mathbf{w}\|_1^{p/2-1}.$$

so the entropy integral gives a bound of

$$\begin{aligned} & O(pw^{1/2})(\|\mathbf{w}\|_1^{p/2-1}/\gamma^{p/2})^{1/2}(\log(n/\varepsilon))^{1/2} \cdot \int_\lambda^{\operatorname{diam}(B^p(\mathbf{A}))} \frac{1}{t}\, dt \\ \leq\ & O(pw^{1/2})(\|\mathbf{w}\|_1^{p/2-1}/\gamma^{p/2})^{1/2}(\log(n/\varepsilon))^{1/2} \log \frac{4p \cdot (\gamma^{-p/2}\|\mathbf{w}\|_1^{p/2-1} w)^{1/2}}{\lambda} \quad \text{Lemma 6.4.12} \end{aligned}$$

Choosing $\lambda = pw^{1/2}(\gamma^{-p/2}\|\mathbf{w}\|_1^{p/2-1})^{1/2}/\sqrt{d}$ yields the claimed conclusion. □

6.5 Analysis of ℓ_p Lewis weight sampling: endgame

We will now assemble our previous lemmas to prove the main sampling theorem for γ-one-sided ℓ_p Lewis weight sampling.

Theorem 6.5.1. Let $0 < p < \infty$. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let $0 \leq \mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights. Let $\alpha > 0$ and let $p_i = \min\{1, \mathbf{w}_i^p(\mathbf{A})/\alpha\}$ for $i \in [n]$. Let $\mathbf{S} \in \mathbb{R}^{n \times n}$ be the diagonal matrix formed by independently setting $\mathbf{S}_{i,i} = 1/p_i^{1/p}$ with probability p_i and 0 otherwise for each $i \in [n]$. Then for

$$\alpha = \begin{cases} \dfrac{\gamma \varepsilon^2}{(\log d)^2 (\log n) \log \frac{1}{\delta}} & p < 2 \\ \dfrac{\gamma^{p/2} \varepsilon^2}{\|\mathbf{w}\|_1^{p/2-1} (\log d)^2 \log(n/\varepsilon) \log \frac{1}{\delta}} & p > 2 \end{cases}$$

$$\|\mathbf{SAx}\|_p^p = (1 \pm \varepsilon)\|\mathbf{Ax}\|_p^p$$

for every $\mathbf{x} \in \mathbb{R}^d$. With probability at least $1 - \delta$, the number of rows sampled is at most

$$\mathsf{nnz}(\mathbf{S}) = \begin{cases} O\left(\|\mathbf{w}\|_1 \dfrac{(\log d)^2 (\log n) \log \frac{1}{\delta}}{\gamma \varepsilon^2} \right) & p < 2 \\ O\left(\|\mathbf{w}\|_1^{p/2} \dfrac{(\log d)^2 \log(n/\varepsilon) \log \frac{1}{\delta}}{\gamma^{p/2} \varepsilon^2} \right) & p > 2 \end{cases}.$$

Proof. In Lemma 6.3.1 we have reduced our task of bounding the sampling error to bounding the l-th moments of a certain Rademacher process. The bound for this Rademacher process is reduced to a bound for another Rademacher process of the form of (6.2). Here, $\mathbf{A}'$ is an $m \times d$ matrix with $m = O(n/\alpha)$ rows whose ℓ_p Lewis weights are uniformly bounded by α, for any $\alpha \in (0, 1)$ of our choosing. We bound the tail of this Rademacher process via Dudley's entropy integral (Theorem 2.3.6), which then leads to moment bounds via integration (Lemma 2.3.7).

We now evaluate the moment bounds. The entropy integral $\mathcal{E}$ and diameter $\mathcal{D}$ required in Lemma 2.3.7 are given by

$$\mathcal{E} = \begin{cases} O((\alpha/\gamma)^{1/2})\left(\frac{\log d}{2-p} + \log n\right)^{1/2} \log d & p < 2 \\ O(p\alpha^{1/2})(\gamma^{-p/2}\|\mathbf{w}\|_1^{p/2-1})^{1/2}(\log(n/\varepsilon))^{1/2} \log d & p > 2 \end{cases}$$

by Lemmas 6.4.16 and 6.4.17, and

$$\mathcal{D} = \begin{cases} 4 \cdot (\alpha/\gamma)^{1/2} & p < 2 \\ 4p \cdot \alpha^{1/2}(\gamma^{-p/2}\|\mathbf{w}\|_1^{p/2-1})^{1/2} & p > 2 \end{cases}$$

by Lemma 6.4.12. Then for α as chosen in the theorem statement, the moment bound is at most $\delta\varepsilon^l$. This is the bound requested by Lemma 6.3.1, and thus this proves the theorem. □

6.6 Online ℓ_p Lewis weight sampling

In this section, we obtain the first *online* ℓ_p subspace embeddings which achieve guarantees which nearly match those of Theorems 6.1.9 and 6.1.11. This provides a generalization of the results of [CMP16, CMP20] for ℓ_2 subspace embeddings to ℓ_p subspace embeddings, and answers open questions of [BDM+20] and [CLS22].

We define online ℓ_p Lewis weights, which are defined analogously to online leverage scores (Definition 1.3.4).

Definition 6.6.1 (Online ℓ_p Lewis weights)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 < p < \infty$. Then, for each $i \in [n]$, the ith online ℓ_p Lewis weight is defined as

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) := \begin{cases} \min\left\{\left[\mathbf{a}_i^\top(\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1})^-\mathbf{a}_i\right]^{p/2}, 1\right\} & \text{if } \mathbf{a}_i \in \text{rowspan}(\mathbf{A}_{i-1}) \\ 1 & \text{otherwise} \end{cases}$$

where $\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_j$ is the $j \times j$ diagonal matrix with $\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_j(i,i) = \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})$.

Note that by maintaining the online Lewis quadratic $\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p}\mathbf{A}_{i-1}$, we can access $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})$ upon the arrival of row $\mathbf{a}_i$ by using only $O(d^2)$ words of memory.

6.6.1 Lemmas from linear algebra

Before we prove our results about online ℓ_p Lewis weights, we need a few linear algebraic lemmas.

Lemma 6.6.2. Let $\mathbf{R} = \mathbf{V}\tilde{\mathbf{R}}\mathbf{V}^\top \in \mathbb{R}^{d\times d}$ where $\tilde{\mathbf{R}} \in \mathbb{R}^{r\times r}$ is a symmetric positive definite matrix and $\mathbf{V} \in \mathbb{R}^{d\times r}$ has orthonormal columns. Then,

$$\mathbf{R}^- = \mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top.$$

Proof. Note that $\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top$ is an inverse for the column space of $\mathbf{R}$, i.e.,

$$\mathbf{R}(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top)\mathbf{R} = \mathbf{V}\tilde{\mathbf{R}}\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{R}}\mathbf{V}^\top = \mathbf{R}$$

and a weak inverse, i.e.,

$$(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top)\mathbf{R}(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top) = \mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top.$$

One can also easily check that both $\mathbf{R}(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top)$ and $(\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top)\mathbf{R}$ are Hermitian. Thus, $\mathbf{V}\tilde{\mathbf{R}}^{-1}\mathbf{V}^\top$ is uniquely determined to be the pseudoinverse of $\mathbf{R}$. □

Lemma 6.6.3. Let $0 \preceq \mathbf{R} \preceq \mathbf{S} \in \mathbb{R}^{d\times d}$ by symmetric positive semidefinite matrices. Let $\mathbf{a} \in \mathrm{rowspan}(\mathbf{R})$. Then,

$$\mathbf{a}^\top\mathbf{R}^-\mathbf{a} \geq \mathbf{a}^\top\mathbf{S}^-\mathbf{a}.$$

Proof. Let $\mathbf{V} \in \mathbb{R}^{d\times r}$ be an orthonormal basis for $V := \mathrm{rowspan}(\mathbf{R})$, where $r = \dim(V)$. Let $\mathbf{P} = \mathbf{V}\mathbf{V}^\top$ be the projection matrix onto V. Write $\mathbf{a} = \mathbf{V}\mathbf{b}$ for $\mathbf{b} \in \mathbb{R}^r$ and $\mathbf{R} = \mathbf{V}\tilde{\mathbf{R}}\mathbf{V}^\top$, $\mathbf{PSP} = \mathbf{V}\tilde{\mathbf{S}}\mathbf{V}^\top$ for $\tilde{\mathbf{R}}, \tilde{\mathbf{S}} \in \mathbb{R}^{r\times r}$. Then, we have that

$$\mathbf{a}^\top\mathbf{R}^-\mathbf{a} = \mathbf{b}^\top\mathbf{V}^\top(\mathbf{V}\tilde{\mathbf{R}}\mathbf{V}^\top)^-\mathbf{V}\mathbf{b} = \mathbf{b}^\top\tilde{\mathbf{R}}^{-1}\mathbf{b}$$

and

$$\mathbf{a}^\top\mathbf{S}^-\mathbf{a} = \mathbf{b}^\top\mathbf{V}^\top(\mathbf{V}\tilde{\mathbf{S}}\mathbf{V}^\top)^-\mathbf{V}\mathbf{b} = \mathbf{b}^\top\tilde{\mathbf{S}}^{-1}\mathbf{b}.$$

Furthermore, for all $\mathbf{x} \in \mathbb{R}^r$, we have that

$$\mathbf{x}^\top\mathbf{V}^\top\mathbf{R}\mathbf{V}\mathbf{x} \leq \mathbf{x}^\top\mathbf{V}^\top\mathbf{S}\mathbf{V}\mathbf{x} = \mathbf{x}^\top\mathbf{V}^\top\mathbf{P}\mathbf{S}\mathbf{P}\mathbf{V}\mathbf{x}$$

so $\tilde{\mathbf{R}} \preceq \tilde{\mathbf{S}}$, meaning that $\tilde{\mathbf{R}}^{-1} \succeq \tilde{\mathbf{S}}^{-1}$. Thus,

$$\mathbf{a}^\top\mathbf{R}^-\mathbf{a} = \mathbf{b}^\top\tilde{\mathbf{R}}^{-1}\mathbf{b} \geq \mathbf{b}^\top\tilde{\mathbf{S}}^{-1}\mathbf{b} = \mathbf{a}^\top\mathbf{S}^-\mathbf{a}.$$ □

6.6.2 Properties of online ℓ_p Lewis weights

We first show that for $0 < p < 2$, the online Lewis weights upper bound Lewis weights.

Lemma 6.6.4. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $0 < p < 2$. Then, for each $i \in [n]$,

$$\mathbf{w}_i^p(\mathbf{A}) \leq \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})$$

Proof. We proceed by induction. It suffices to consider the case when $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) < 1$, since $\mathbf{w}_i^p(\mathbf{A}) \le 1$ for every $i \in [n]$. In particular, $\mathbf{a}_i \in \operatorname{rowspan}(\mathbf{A}_{i-1})$ and

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = \left[\mathbf{a}_i^\top (\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1})^- \mathbf{a}_i\right]^{p/2}.$$

Then, since $1 - \frac{2}{p} < 0$, we have that

$$\begin{aligned}
\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1} \succeq \mathbf{W}^p(\mathbf{A})_{i-1} \succ 0 &\implies \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \preceq \mathbf{W}^p(\mathbf{A})_{i-1}^{1-2/p} \\
&\implies \mathbf{A}_{i-1}^\top (\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} - \mathbf{W}^p(\mathbf{A})_{i-1}^{1-2/p}) \mathbf{A}_{i-1} \preceq 0 \\
&\implies \mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1} \preceq \mathbf{A}_{i-1}^\top \mathbf{W}^p(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1}.
\end{aligned}$$

By Lemma 6.6.3, it follows that for every $\mathbf{a} \in \operatorname{rowspan}(\mathbf{A}_{i-1})$,

$$\mathbf{a}^\top (\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1})^- \mathbf{a} \ge \mathbf{a}^\top (\mathbf{A}_{i-1}^\top \mathbf{W}^p(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1})^- \mathbf{a}.$$

Similarly, we have that

$$\mathbf{a}^\top (\mathbf{A}_{i-1}^\top \mathbf{W}^p(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1})^- \mathbf{a} \ge \mathbf{a}^\top (\mathbf{A}^\top \mathbf{W}^p(\mathbf{A})^{1-2/p} \mathbf{A})^- \mathbf{a}$$

for every $\mathbf{a} \in \operatorname{rowspan}(\mathbf{A}_{i-1})$. The result follows by taking $p/2$-th roots on the chain of inequalities. □

Note that for $p > 2$, the above proof fails since $1 - \frac{2}{p} > 0$, which causes the inequalities to go the wrong way. Nevertheless, we show that these weights satisfy the *one-sided Lewis property*, which we have shown to be sufficient for sampling in Theorem 6.5.1.

Lemma 6.6.5 (One-sided Lewis property of online ℓ_p Lewis weights)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $0 < p < \infty$. Then, for each $i \in [n]$,

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) \ge \boldsymbol{\tau}_i(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p} \mathbf{A}).$$

Proof. We already have the result when $\mathbf{a}_i \notin \operatorname{rowspan}(\mathbf{A}_{i-1})$, so we assume $\mathbf{a}_i \in \operatorname{rowspan}(\mathbf{A}_{i-1})$. Similarly, we can assume that $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) < 1$. In this case,

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = \left[\mathbf{a}_i^\top (\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1})^- \mathbf{a}_i\right]^{p/2}$$

which rearranges to

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = (\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p} \mathbf{a}_i)^\top (\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1})^- (\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p} \mathbf{a}_i).$$

By Lemma 6.6.3, this is bounded below by

$$(\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p} \mathbf{a}_i)^\top (\mathbf{A}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1-2/p} \mathbf{A})^- (\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p} \mathbf{a}_i) = \boldsymbol{\tau}_i(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p} \mathbf{A}),$$

which is the claimed result. □

6.6.3 The sum of online ℓ_p Lewis weights

Finally, we bound the sum of online Lewis weights, using bounds on the sum of online leverage scores. Our proof substantially simplifies the proofs of [BDM+20, Lemma 4.7, Lemma 5.15], which relied on an elaborate argument involving recursive applications of a "whack-a-mole" lemma of [CLM+15], and also slightly improves the bound by logarithmic factors.

Lemma 6.6.6 (Sum of online ℓ_p Lewis weights)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $0 < p < \infty$. Then,

$$\sum_{i=1}^{n} \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) \leq O(d) \log(n\kappa^{\mathsf{OL}}(\mathbf{A})).$$

Proof. Our analysis is similar to those given by [CMP20] and [WY22b]. For $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) < 1$, we have that

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = \left[\mathbf{a}_i^\top (\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1})^- \mathbf{a}_i\right]^{p/2}.$$

This rearranges to

$$\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = (\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{a}_i)^\top (\mathbf{A}_{i-1}^\top \mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})_{i-1}^{1-2/p} \mathbf{A}_{i-1})^- (\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{a}_i),$$

which is exactly the ith online leverage score of $\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A}$. Similar reasoning for $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = 1$ shows that $\mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A})$. Thus,

$$\sum_{i=1}^{n} \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) = \sum_{i=1}^{n} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A}) \leq O(d \log \kappa^{\mathsf{OL}}(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A}))$$

by bounds on the sum of online leverage scores (Lemma 1.3.8). If $p < 2$, then we have for any $\mathbf{x} \in \mathbb{R}^d$ and $i \in [n]$ that

$$\begin{aligned}\|\mathbf{A}_i\mathbf{x}\|_2 \leq \left\|\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A}_i)^{1/2-1/p}\mathbf{A}_i\mathbf{x}\right\|_2 &\leq \left\|\mathbf{W}^{p}(\mathbf{A}_i)^{1/2-1/p}\mathbf{A}_i\mathbf{x}\right\|_2 \\ &\leq d^{|1/2-1/p|}\|\mathbf{A}_i\mathbf{x}\|_p \leq (nd)^{|1/2-1/p|}\|\mathbf{A}_i\mathbf{x}\|_2,\end{aligned}$$

so $\kappa^{\mathsf{OL}}(\mathbf{A}) = \mathrm{poly}(n)\kappa^{\mathsf{OL}}(\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A})^{1/2-1/p}\mathbf{A})$. If $p > 2$, then by Lemma 6.2.3,

$$\|\mathbf{A}_i\mathbf{x}\|_p \leq \left\|\mathbf{W}^{p,\mathsf{OL}}(\mathbf{A}_i)^{1/2-1/p}\mathbf{A}_i\mathbf{x}\right\|_2 \leq \|\mathbf{A}_i\mathbf{x}\|_2.$$

Thus,

$$\sum_{i=1}^{n} \mathbf{w}_i^{p,\mathsf{OL}}(\mathbf{A}) \leq O(d) \log(n\kappa^{\mathsf{OL}}(\mathbf{A})).$$ □

Now that we have established the one-sided Lewis property of the online ℓ_p Lewis weights and bounded their sum, sampling results immediately follow from our results on sampling with one-sided ℓ_p Lewis weights in Theorem 6.5.1.

Chapter 7

ℓ_p sensitivity sampling [WY23c]

7.1 Beyond ℓ_p Lewis weight sampling

In this section, we study the ℓ_p sensitivity sampling algorithm for sampling ℓ_p subspace embeddings. Recall from our discussion in Section 1.3.1 that the general sensitivity sampling framework provides an approach towards constructing coresets for an extremely wide class of shape-fitting problems, and when specialized to the case of ℓ_p subspace embeddings, sensitivity sampling achieves a $\mathrm{poly}(d)/\varepsilon^2$ bound on the row count for any fixed $0 < p < \infty$ (Section 6.1), although is not known to achieve the nearly optimal row counts that are possible with ℓ_p Lewis weight sampling. However, for $p > 2$, ℓ_p sensitivity sampling in fact has the potential to produce a *smaller* number of rows than ℓ_p Lewis weight sampling, if the total sensitivity $\mathfrak{S}^p(\mathbf{A})$ is small. Indeed, $\mathfrak{S}^p(\mathbf{A})$ can be as small as d even for $p > 2$ (while the worst-case bound is $d^{p/2}$), in which case one obtains a sample complexity of $\tilde{O}(\varepsilon^{-2}\mathfrak{S}d) = \tilde{O}(\varepsilon^{-2}d^2)$ for such matrices, while Lewis weight sampling would require $\tilde{O}(\varepsilon^{-2}d^{p/2})$, which is polynomially worse for $p > 4$. We discuss several explicit families of matrices with small total sensitivity in Section 7.2. Thus, despite the fact that ℓ_p Lewis weight sampling already achieves nearly optimal bounds in the worst case (see Chapter 6), the study of sensitivity sampling using the scores of Definition 6.1.2 is still interesting for two reasons:

1. The definition of sensitivities can be massively generalized to a wide variety of sampling-based approximation problems.
2. For $p > 2$, sensitivity sampling admits matrix-dependent bounds which can circumvent the lower bounds of Theorem 6.1.5.

For these reasons, our work in [WY23c] studies the problem of obtaining the tightest possible bounds for ℓ_p sensitivity sampling:

Question 7.1.1. What is the smallest sample complexity possible for the ℓ_p sensitivity sampling algorithm?

While we are not able to complete resolve Question 7.1.1, we make progress towards this question by giving an analysis of ℓ_p sensitivity sampling which goes beyond the general case bound of $\tilde{O}(\varepsilon^{-2}\mathfrak{S}d)$ for $p > 2$. Our analysis also gives a similar result for $p < 2$, although this result is superseded by an analysis that relates ℓ_p sensitivity scores to ℓ_p Lewis weights, which

achieves a bound of $\tilde{O}(\varepsilon^{-2}d^{1-p/2}\mathfrak{S})$ [CD21, MO23].

Theorem 7.1.2 (ℓ_p sensitivity sampling [WY23c])**.** Let $1 \le p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\alpha > 0$ and let $q_i = \min\{1, 1/n + \boldsymbol{\sigma}_i^p(\mathbf{A})/\alpha\}$ for $i \in [n]$. Let $\mathbf{S} \in \mathbb{R}^{n\times n}$ be the ℓ_p sampling matrix with probabilities $\{q_i\}_{i=1}^n$. Then, with probability at least $99/100$, there is an α such that $\mathbf{S}$ is an ℓ_p subspace embedding satisfying Definition 1.1.1 with $\kappa = (1+\varepsilon)$, and furthermore, $\mathbf{S}$ has at most r nonzero rows, for

$$r = \begin{cases} \varepsilon^{-2}\mathfrak{S}^p(\mathbf{A})^{2/p}\operatorname{poly}\log n & 1 \le p < 2 \\ \varepsilon^{-2}\mathfrak{S}^p(\mathbf{A})^{2-2/p}\operatorname{poly}\log n & 2 < p < \infty \end{cases}$$

Our improved analysis of ℓ_p sensitivity sampling is largely based off of the analysis of ℓ_p Lewis weight sampling in the works of [BLM89, LT91], and in particular makes use of similar chaining arguments (see Chapter 6 for further details). In these arguments when ℓ_p Lewis weights are used as sampling probabilities, then such a chaining argument goes through due to the fact that the resulting matrix has uniformly bounded leverage scores and ℓ_p sensitivities, which in turn is a consequence of the specific definition of Lewis weights. However, when we instead use the ℓ_p sensitivities as the sampling probabilities, we no longer have this property, and the analysis needs to be modified.

To address this problem, we observe that although ℓ_p sensitivity sampling does not directly lead to uniformly bounded leverage scores, it *does* lead to uniformly bounded ℓ_p sensitivities in the resulting matrix. We then show that this in turn implies approximately uniformly bounded leverage scores, by relating the ℓ_p sensitivities to the leverage scores.

7.2 Structured matrices with small total sensitivity, $p > 2$

We first show several examples in structured regression problems in which our new sensitivity sampling results give the best known sample complexity results for ℓ_p subspace embeddings for $p > 2$. We start by presenting a couple of lemmas which show that certain natural classes of matrices have total ℓ_p sensitivity $\ll d^{p/2}$.

The first result is a lemma extracted from a result of [MMM+22] bounding the total ℓ_p sensitivity for a sparse perturbation of low rank matrices:

Lemma 7.2.1 (Sensitivity bounds for low rank + sparse matrices [MMM+22])**.** Let $\mathbf{A} = \mathbf{K} + \mathbf{S} \in \mathbb{R}^{n\times d}$ for a rank k matrix $\mathbf{K}$ and an $\mathbf{S}$ with at most s nonzero entries per row. Let $1 \le p < \infty$. Then, $\mathfrak{S}^p(\mathbf{A}) \le d^s(k+s)^p$.

Proof. Let r be an integer such that $2^r \le p < 2^{r+1}$. Then, for each $i \in [n]$, we may write

$$\mathbf{a}_i = \mathbf{k}_i + \mathbf{s}_i = \sum_{j=1}^{k} \alpha_{i,j}\mathbf{v}_j + \sum_{j=1}^{s} \beta_{i,j}\mathbf{e}_{i_j}$$

where $\mathbf{v}_j \in \mathbb{R}^d$ for $j \in [k]$. Then, the tensor product $\mathbf{a}_i^{\otimes 2^r}$ of $\mathbf{a}_i$ with itself 2^r times can be written as a linear combination of tensor products $\mathbf{y}_1 \otimes \cdots \otimes \mathbf{y}_{2^r}$, where each $\mathbf{y}_q$ for $q \in [2^r]$ is one of $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k, \mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \ldots, \mathbf{e}_{i_s}\}$. Thus, $\mathbf{a}_i^{\otimes r}$ lies in the span of at most $(k+s)^{2^r}$ vectors, for a fixed choice of $\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \ldots, \mathbf{e}_{i_s}$. Since there are at most d^s possible choices of the sparsity pattern,

every $\mathbf{a}_i^{\otimes 2^r}$ for $i \in [n]$ lies in the span of at most $d' \coloneqq d^s(k+s)^{2^r}$ vectors. That is, if $\mathbf{A}^{\otimes 2^r}$ is the Khatri-Rao 2^rth power of $\mathbf{A}$, then $\mathbf{A}^{\otimes 2^r}$ is a rank d' matrix. Then, we have that

$$|[\mathbf{A}\mathbf{x}](i)|^p = (|[\mathbf{A}\mathbf{x}](i)|^{2^r})^{p/2^r} = (\langle \mathbf{a}_i, \mathbf{x}\rangle^{2^r})^{p/2^r} = \left|\left\langle \mathbf{a}_i^{\otimes 2^r}, \mathbf{x}^{\otimes 2^r}\right\rangle\right|^{p/2^r}$$

so

$$\sup_{\mathbf{A}\mathbf{x}\neq 0} \frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} = \sup_{\mathbf{A}\mathbf{x}\neq 0} \frac{\left|[\mathbf{A}^{\otimes 2^r}\mathbf{x}^{\otimes 2^r}](i)\right|^{p/2^r}}{\|\mathbf{A}^{\otimes 2^r}\mathbf{x}^{\otimes 2^r}\|_{p/2^r}^{p/2^r}} \leq \sup_{\mathbf{A}^{\otimes 2^r}\mathbf{x}\neq 0} \frac{\left|[\mathbf{A}^{\otimes 2^r}\mathbf{x}](i)\right|^{p/2^r}}{\|\mathbf{A}^{\otimes 2^r}\mathbf{x}\|_{p/2^r}^{p/2^r}}$$

that is, the $\ell_{p/2^r}$ sensitivities of $\mathbf{A}^{\otimes 2^r}$ upper bound the ℓ_p sensitivities of $\mathbf{A}$. Since $p/2^r \leq 2$, the total $\ell_{p/2^r}$ sensitivity of $\mathbf{A}^{\otimes 2^r}$ is bounded by its rank, which is d'. □

In a second example, we show that "concatenated Vandermonde" matrices, which were studied in, e.g., [ASW13], also have small total ℓ_p sensitivity. These matrices naturally arise as the result of applying a polynomial feature map to a matrix.

Definition 7.2.2 (Vandermonde matrix)**.** Given a vector $\mathbf{a} \in \mathbb{R}^n$, the degree q Vandermonde matrix $V^q(\mathbf{a}) \in \mathbb{R}^{n\times(q+1)}$ is defined entrywise as $V^q(\mathbf{a})_{i,j} = \mathbf{a}_i^j$ for $j = 0, 1, \ldots, q$.

Definition 7.2.3 (Polynomial feature map)**.** Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times k}$ and an integer q, we define the matrix $V^q(\mathbf{A}) \in \mathbb{R}^{n\times k(q+1)}$ to be the horizontal concatenation of the Vandermonde matrices $V^q(\mathbf{A}\mathbf{e}_1), V^q(\mathbf{A}\mathbf{e}_2), \ldots, V^q(\mathbf{A}\mathbf{e}_k)$.

We then have the following:

Lemma 7.2.4 (Sensitivity bounds for matrices under polynomial feature maps)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times k}$ and let q be an integer. Let $1 \leq p < \infty$. Then, $\mathfrak{S}^p(V^q(\mathbf{A})) \leq (pq+1)^k$.

Proof. Let r be an integer such that $2^r \leq p < 2^{r+1}$. Fix some $\mathbf{x} \in \mathbb{R}^{k(q+1)}$. Now consider the vector $\langle \mathbf{a}, \mathbf{x}\rangle$, where $\mathbf{a}$ is a $k(q+1)$-dimensional vector of monomials of degree 0 through q of the indeterminate variables $a_1, a_2, \ldots, a_k$, that is,

$$\mathbf{a} = (1, a_1, a_1^2, \ldots, a_1^q, \quad 1, a_2, a_2^2, \ldots, a_2^q, \quad \ldots, \quad 1, a_k, a_k^2, \ldots, a_k^q).$$

Then, $\langle \mathbf{a}, \mathbf{x}\rangle$ is a degree q polynomial in the indeterminates $a_1, a_2, \ldots, a_k$ with coefficients specified by $\mathbf{x}$, so $\langle \mathbf{a}, \mathbf{x}\rangle^{2^r}$ is a polynomial in the indeterminates $a_1, a_2, \ldots, a_k$, such that every monomial term is at most degree $2^r q$ in each variable. Note that there are at most k variables, so there can be at most $(2^r q+1)^k$ possible monomials, by choosing the degree of each of the monomials. Let $\mathbf{x}'$ denote the coefficients of this polynomial in the monomial basis, for a given set of original coefficients $\mathbf{x}$.

Now consider the matrix $V^q(\mathbf{A})$. Then, for a fixed $\mathbf{x} \in \mathbb{R}^{k(q+1)}$, $[V^q(\mathbf{A})\mathbf{x}](i)^{2^r}$ is the evaluation of $\langle \mathbf{a}, \mathbf{x}\rangle^{2^r}$ at the ith row $\mathbf{a}_i$ of $\mathbf{A}$ for the indeterminates $a_1, a_2, \ldots, a_k$, so it can be written as the linear combination of at most $(2^r q+1)^k$ monomials evaluated at $\mathbf{a}_i$, with coefficients $\mathbf{x}'$. Thus, $[V^q(\mathbf{A})\mathbf{x}](i)^{2^r} = \mathbf{A}'\mathbf{x}'$ for some $\mathbf{A}'$ with rank at most $(2^r q+1)^k$.

Finally, note that

$$|[V^q(\mathbf{A})\mathbf{x}](i)|^p = (|[V^q(\mathbf{A})\mathbf{x}](i)|^{2^r})^{p/2^r} = |[\mathbf{A}'\mathbf{x}']|(i)|^{p/2^r}.$$

Thus, the total ℓ_p sensitivity of $V^q(\mathbf{A})$ is bounded by the total $\ell_{p/2^r}$ sensitivity of $\mathbf{A}'$, which is at most $(2^r q+1)^k \leq (pq+1)^k$. □

This generalizes a result of [MMM+22], which bounds the ℓ_p sensitivities of a single Vandermonde matrix.

In the low-sensitivity matrices of Lemma 7.2.1 and Lemma 7.2.4, it is in fact possible to apply Lewis weight sampling to obtain sampling bounds that match these sensitivity bounds, by using the *tensoring trick* [MMM+22]. However, when a tiny amount of noise is added to these matrices, then algebraic tricks such as tensoring break down, and the sensitivity bounds derived from Lewis weights increase substantially to $d^{p/2}$ for $p > 2$. On the other hand, sensitivity sampling itself is robust with respect to the addition of noise, as it depends only on norms rather than brittle quantities such as rank.

Lemma 7.2.5. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a rank d matrix with minimum singular value $\sigma_{\min}$. Let $\mathbf{E} \in \mathbb{R}^{n\times d}$ be an arbitrary perturbation matrix with

$$\|\mathbf{E}\|_2 \leq \frac{\sigma_{\min}}{2n^{1+1/p}}.$$

Then, $\mathfrak{S}^p(\mathbf{A}+\mathbf{E}) \leq 2^p(\mathfrak{S}^p(\mathbf{A})+1)$.

Proof. For any $\mathbf{x} \in \mathbb{R}^d$, we have that

$$\begin{aligned}
\|(\mathbf{A}+\mathbf{E})\mathbf{x}\|_p &= \|\mathbf{A}\mathbf{x}\|_p \pm \|\mathbf{E}\mathbf{x}\|_p \\
&= \|\mathbf{A}\mathbf{x}\|_p \pm \sqrt{n}\|\mathbf{E}\mathbf{x}\|_2 \\
&= \|\mathbf{A}\mathbf{x}\|_p \pm \frac{\sigma_{\min}}{\sqrt{n}}\|\mathbf{x}\|_2 \\
&= \|\mathbf{A}\mathbf{x}\|_p \pm \frac{\sigma_{\min}}{\sqrt{n}}\frac{1}{\sigma_{\min}}\|\mathbf{A}\mathbf{x}\|_2 \\
&= \|\mathbf{A}\mathbf{x}\|_p \pm \frac{1}{2}\|\mathbf{A}\mathbf{x}\|_p \\
&= (1 \pm 1/2)\|\mathbf{A}\mathbf{x}\|_p
\end{aligned}$$

so

$$\frac{|[(\mathbf{A}+\mathbf{E})\mathbf{x}](i)|^p}{\|(\mathbf{A}+\mathbf{E})\mathbf{x}\|_p^p} \leq 2^{p-1}\frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|(\mathbf{A}+\mathbf{E})\mathbf{x}\|_p^p} + 2^{p-1}\frac{|[\mathbf{E}\mathbf{x}](i)|^p}{\|(\mathbf{A}+\mathbf{E})\mathbf{x}\|_p^p} \leq 2^p\frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} + 2^p\frac{|[\mathbf{E}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p}.$$

The first term is clearly bounded by $2^p\boldsymbol{\sigma}_i^p(\mathbf{A})$ for any $\mathbf{x}$. On the other hand, the second term is bounded by

$$\begin{aligned}
2^p\frac{|[\mathbf{E}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} &\leq 2^p\frac{\|\mathbf{E}\mathbf{x}\|_p^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \leq 2^p n^{p/2}\frac{\|\mathbf{E}\mathbf{x}\|_2^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \\
&\leq 2^p\frac{\sigma_{\min}^p}{n^{p/2+1}}\frac{\|\mathbf{x}\|_2^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \leq 2^p\frac{1}{n^{p/2+1}}\frac{\|\mathbf{A}\mathbf{x}\|_2^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \leq 2^p\frac{\|\mathbf{A}\mathbf{x}\|_p^p}{n\|\mathbf{A}\mathbf{x}\|_p^p} = \frac{2^p}{n}.
\end{aligned}$$

Thus, the total sensitivity is bounded by

$$2^p\sum_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A}) + \frac{1}{n} = 2^p(\mathfrak{S}^p(\mathbf{A})+1).$$

□

Thus, for small perturbations of structured matrices with small ℓ_p sensitivity as specified by Lemma 7.2.5, Theorem 7.1.2 give the tightest known bounds on the sample complexity for ℓ_p subspace embeddings. Such perturbations may arise due to roundoff error or finite precision on a computer, and no prior bounds beating Lewis weight sampling or the naïve $\mathfrak{S}d$ bound for sensitivity sampling were known for the applications above.

7.3 Properties of ℓ_p sensitivities

We will first collect several results on ℓ_p sensitivities that we will use.

7.3.1 Monotonicity of max ℓ_p sensitivity

Lemma 7.3.1 (Monotonicity of max ℓ_p sensitivity)**.** Let $q \geq p > 0$ and $\mathbf{y} \in \mathbb{R}^n$. Then,

$$\frac{\|\mathbf{y}\|_\infty^p}{\|\mathbf{y}\|_p^p} \leq \frac{\|\mathbf{y}\|_\infty^q}{\|\mathbf{y}\|_q^q}.$$

Proof. We have that

$$\|\mathbf{y}\|_q^q = \sum_{i=1}^n |\mathbf{y}(i)|^q \leq \|\mathbf{y}\|_\infty^{q-p} \sum_{i=1}^n |\mathbf{y}(i)|^p = \|\mathbf{y}\|_\infty^{q-p} \|\mathbf{y}\|_p^p,$$

so

$$\frac{\|\mathbf{y}\|_\infty^p}{\|\mathbf{y}\|_p^p} \leq \frac{\|\mathbf{y}\|_\infty^p}{\|\mathbf{y}\|_q^q / \|\mathbf{y}\|_\infty^{q-p}} = \frac{\|\mathbf{y}\|_\infty^q}{\|\mathbf{y}\|_q^q}.$$

□

We also use an "approximate converse" of the above result:

Lemma 7.3.2 (Reverse monotonicity of max ℓ_p sensitivity)**.** Let $q \geq p > 0$ and $\mathbf{y} \in \mathbb{R}^n$. Then,

$$\frac{\|\mathbf{y}\|_\infty^q}{\|\mathbf{y}\|_q^q} \leq \left(\frac{\|\mathbf{y}\|_\infty^p}{\|\mathbf{y}\|_p^p} \right)^{q/p} n^{q/p-1}.$$

Proof. Since $\|\mathbf{y}\|_p \leq \|\mathbf{y}\|_q n^{1/p-1/q}$, we have that

$$\frac{\|\mathbf{y}\|_\infty^q}{\|\mathbf{y}\|_q^q} \leq \frac{|\mathbf{y}(i)|^q}{\|\mathbf{y}\|_p^q \cdot n^{1-q/p}} \leq \frac{\|\mathbf{y}\|_\infty^q}{\|\mathbf{y}\|_p^q \cdot n^{1-q/p}} = \left(\frac{\|\mathbf{y}\|_\infty^p}{\|\mathbf{y}\|_p^p} \right)^{q/p} n^{q/p-1}.$$

□

7.3.2 Flattening ℓ_p sensitivities

We give a sensitivity flattening lemma, analogous to the ℓ_p Lewis weight flattening lemma of Lemma 6.3.2.

Lemma 7.3.3 (ℓ_p Sensitivity Flattening)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $1 \le p < \infty$. Let $C \ge 1$. Then, there exists a $\mathbf{A}' \in \mathbb{R}^{m\times d}$ for $m = (1 + 1/C)n$ such that $\|\mathbf{A}\mathbf{x}\|_p = \|\mathbf{A}'\mathbf{x}\|_p$ for every $\mathbf{x} \in \mathbb{R}^d$, $\mathfrak{S}^p(\mathbf{A}) = \mathfrak{S}^p(\mathbf{A}')$, and $\boldsymbol{\sigma}^p_{i'}(\mathbf{A}') \le C\mathfrak{S}^p(\mathbf{A})/n$ for every $i' \in [m]$.

Proof. Suppose that for any row $\mathbf{a}_i \in \mathbb{R}^d$ of $\mathbf{A}$ for $i \in [n]$ with $\boldsymbol{\sigma}^p_i(\mathbf{A}) \ge C\mathfrak{S}^p(\mathbf{A})/n$, we replace the row with $k := \lceil \boldsymbol{\sigma}^p_i(\mathbf{A})/(C\mathfrak{S}^p(\mathbf{A})/n)\rceil$ copies of $\mathbf{a}_i/k^{1/p}$ to form a new matrix $\mathbf{A}'$. Then, we add at most

$$\sum_{i:\boldsymbol{\sigma}^p_i(\mathbf{A})\ge\mathfrak{S}^p(\mathbf{A})/n} \left\lceil \frac{\boldsymbol{\sigma}^p_i(\mathbf{A})}{C\mathfrak{S}^p(\mathbf{A})/n} \right\rceil - 1 \le \sum_{i:\boldsymbol{\sigma}^p_i(\mathbf{A})\ge\mathfrak{S}^p(\mathbf{A})/n} \frac{\boldsymbol{\sigma}^p_i(\mathbf{A})}{C\mathfrak{S}^p(\mathbf{A})/n} = \frac{\mathfrak{S}^p(\mathbf{A})}{C\mathfrak{S}^p(\mathbf{A})/n} = \frac{n}{C}$$

rows. Furthermore, we clearly have that $\|\mathbf{A}\mathbf{x}\|_p = \|\mathbf{A}'\mathbf{x}\|_p$ for every $\mathbf{x} \in \mathbb{R}^d$, and also for any row $i' \in [m]$ that comes from row $i \in [n]$ in the original matrix,

$$\frac{|[\mathbf{A}'\mathbf{x}](i')|^p}{\|\mathbf{A}'\mathbf{x}\|_p^p} \le \frac{C\mathfrak{S}^p(\mathbf{A})/n}{\boldsymbol{\sigma}^p_i(\mathbf{A})} \frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \le \frac{C\mathfrak{S}^p(\mathbf{A})}{n}.$$

Finally, it is also clear that the sum of the sensitivities is also preserved, since the sum of the sensitivities of the k copies of each row $i \in [n]$ in the original matrix is $\boldsymbol{\sigma}^p_i(\mathbf{A})$. □

7.3.3 Total sensitivity

Here we collect several bounds on the total ℓ_p sensitivity.

Lemma 7.3.4 (Sampling preserves total sensitivity)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $1 \le p < \infty$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix such that with probability at least $3/4$,

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p = (1 \pm 1/2)\|\mathbf{A}\mathbf{x}\|_p$$

simultaneously for every $\mathbf{x} \in \mathbb{R}^d$. Then, with probability at least $1/2$,

$$\mathbf{Pr}\{\mathfrak{S}^p(\mathbf{S}\mathbf{A}) \le 8\mathfrak{S}^p(\mathbf{A})\} \ge \frac{1}{2}.$$

Proof. We have that

$$\begin{aligned}\mathfrak{S}^p(\mathbf{S}\mathbf{A}) &= \sum_{i=1}^n \sup_{\mathbf{S}\mathbf{A}\mathbf{x}\neq 0} \frac{|[\mathbf{S}\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p} = \sum_{i=1}^n \mathbf{S}^p_{i,i} \sup_{\mathbf{S}\mathbf{A}\mathbf{x}\neq 0} \frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \frac{\|\mathbf{A}\mathbf{x}\|_p^p}{\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p} \\ &\le \sum_{i=1}^n \mathbf{S}^p_{i,i}\boldsymbol{\sigma}^p_i(\mathbf{A}) \sup_{\mathbf{S}\mathbf{A}\mathbf{x}\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_p^p}{\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p}.\end{aligned}$$

We are guaranteed that

$$\mathbf{Pr}\left\{\sup_{\mathbf{SAx}\neq 0}\frac{\|\mathbf{Ax}\|_p^p}{\|\mathbf{SAx}\|_p^p}\leq 2\right\}\geq\frac{3}{4}.$$

On the other hand, we have that

$$\mathbf{E}\left[\sum_{i=1}^n \mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})\right]=\sum_{i=1}^n\mathbf{E}[\mathbf{S}_{i,i}^p]\boldsymbol{\sigma}_i^p(\mathbf{A})=\mathfrak{S}^p(\mathbf{A})$$

so by Markov's inequality,

$$\mathbf{Pr}\left\{\sum_{i=1}^n \mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})\leq 4\mathfrak{S}^p(\mathbf{A})\right\}\geq\frac{3}{4}.$$

By a union bound,

$$\mathbf{Pr}\{\mathfrak{S}^p(\mathbf{SA})\leq 8\mathfrak{S}^p(\mathbf{A})\}\geq\frac{1}{2}.$$

□

We also prove a high probability and high accuracy version of Lemma 7.3.4.

Lemma 7.3.5 (Sensitivity sampling preserves total sensitivity: high probability and accuracy)**.** Let $\mathbf{A}\in\mathbb{R}^{n\times d}$ and $1\leq p<\infty$. Let $0<\varepsilon,\delta<1$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix such that with probability at least $1-\delta$,

$$\|\mathbf{SAx}\|_p=(1\pm\varepsilon)\|\mathbf{Ax}\|_p$$

simultaneously for every $\mathbf{x}\in\mathbb{R}^d$. Furthermore, suppose that

$$\frac{\boldsymbol{\sigma}_i}{q_i}\leq M\coloneqq\frac{\varepsilon^2\mathfrak{S}^p(\mathbf{A})}{3\log\frac{2}{\delta}}$$

for every $i\in[n]$. Then, with probability at least $1-2\delta$,

$$\mathbf{Pr}\{\mathfrak{S}^p(\mathbf{SA})=(1\pm O(\varepsilon))\mathfrak{S}^p(\mathbf{A})\}\geq 1-2\delta.$$

Proof. The proof follows Lemma 7.3.4. Just as in Lemma 7.3.4, we have that

$$\mathfrak{S}^p(\mathbf{SA})\leq\sum_{i=1}^n\mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})\sup_{\mathbf{SAx}\neq 0}\frac{\|\mathbf{Ax}\|_p^p}{\|\mathbf{SAx}\|_p^p}.$$

Similarly,

$$\mathfrak{S}^p(\mathbf{SA})\geq\sum_{i=1}^n\mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})\inf_{\mathbf{SAx}\neq 0}\frac{\|\mathbf{Ax}\|_p^p}{\|\mathbf{SAx}\|_p^p}.$$

Furthermore, since $\boldsymbol{\sigma}_i/q_i\leq M$, $\mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})/M$ is a random variable bounded by 1, with.

$$\mathbf{E}\left[\sum_{i=1}^n\frac{\mathbf{S}_{i,i}^p\boldsymbol{\sigma}_i^p(\mathbf{A})}{M}\right]=\frac{\mathfrak{S}^p(\mathbf{A})}{M}\geq\frac{3}{\varepsilon^2}\log\frac{2}{\delta}.$$

Thus by Chernoff bounds, we have that

$$\mathbf{Pr}\left\{\sum_{i=1}^{n}\mathbf{S}_{i,i}^{p}\boldsymbol{\sigma}_{i}^{p}(\mathbf{A})=(1\pm\varepsilon)\mathfrak{S}^{p}(\mathbf{A})\right\}\geq 1-\delta.$$

We conclude by a union bound as in Lemma 7.3.4. □

7.4 Analysis of ℓ_p sensitivity sampling

We will now turn towards an analysis of the ℓ_p sensitivity algorithm. Most of the components of the chaining argument for ℓ_p Lewis weight sampling discussed in Chapter 6 will apply to our setting, such as the reduction to a Rademacher process in Section 6.3 and the use of Dudley's entropy integral in Section 6.4.

7.4.1 Dudley's entropy integral

The crucial change to the argument comes from the fact that we need to separately control the leverage scores and sensitivity scores when bounding the Rademacher process associated with sensitivity sampling. We have the following lemma which bounds Dudley's entropy integral in terms of the maximum leverage score τ and maximum sensitivity score σ.

Lemma 7.4.1 (Entropy integral bound for $p<2$)**.** Let $1\leq p<2$ and let $\mathbf{A}\in\mathbb{R}^{n\times d}$ be orthonormal. Let $\tau\geq\max_{i=1}^{n}\left\|\mathbf{e}_i^\top\mathbf{A}\right\|_2^2$ and let $\sigma\geq\max_{i=1}^{n}\boldsymbol{\sigma}_i^p(\mathbf{A})$. Then,

$$\int_0^\infty\sqrt{\log E(B^p,d_X,t)}\,dt\leq O(\tau^{1/2})\left(\frac{\log d}{2-p}+\log n\right)^{1/2}\log\frac{d\sigma}{\tau}$$

Proof. Note that it suffices to integrate the entropy integral to $\operatorname{diam}(B^p(\mathbf{A}))$ rather than ∞, which is at most $4\sigma^{1/2}$ for $p<2$ and $4p\sigma^{1/2}$ for $p>2$ by Lemma 6.4.2.

By Lemma 6.4.1, we have that

$$\log E(B^p,d_X,t)\leq\log E(B^p,2\|\cdot\|_\infty^{p/2},t)=\log E(B^p,B^\infty,(t/2)^{2/p})$$

For small radii less than λ for a parameter λ to be chosen, we use a standard volume argument, which shows that

$$\log E(B^p,B^\infty,t)\leq O(d)\log\frac{n}{t}$$

so

$$\begin{aligned}\int_0^\lambda\sqrt{\log E(B^p,B^\infty,t)}\,dt&=\int_0^\lambda\sqrt{d\log\frac{n}{t}}\,dt\\&\leq\lambda\sqrt{d\log n}+\sqrt{d}\int_0^\lambda\sqrt{\log\frac{1}{t}}\,dt\\&\leq\lambda\sqrt{d\log n}+\sqrt{d}\left(\lambda\sqrt{\log\frac{1}{\lambda}}+\frac{\sqrt{\pi}}{2}\lambda\right)\qquad\text{Lemma 6.4.15}\end{aligned}$$

$$\leq O(\lambda)\sqrt{d \log \frac{n}{\lambda}}$$

On the other hand, for large radii larger than λ, we use the bounds of Lemma 6.4.11, which gives

$$\log E(B^p, B^\infty, (t/2)^{2/p}) \leq O(1)\frac{1}{t^2}\left(\frac{\log d}{2-p} + \log n\right)\tau$$

so the entropy integral gives a bound of

$$O(1)\left[\left(\frac{\log d}{2-p} + \log n\right)\tau\right]^{1/2} \int_\lambda^{4p\sigma^{1/2}} \frac{1}{t}\, dt = O(1)\left[\left(\frac{\log d}{2-p} + \log n\right)\tau\right]^{1/2} \log \frac{4p\sigma^{1/2}}{\lambda}.$$

We choose $\lambda = \sqrt{\tau/d}$, which yields the claimed conclusion. □

An analogous result and proof holds for $p > 2$.

Lemma 7.4.2 (Entropy integral bound for $p > 2$)**.** Let $2 < p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be orthonormal. Let $\tau \geq \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$ and let $\sigma \geq \max_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A})$. Then,

$$\int_0^\infty \sqrt{\log E(B^p, d_X, t)}\, dt \leq O(p\tau^{1/2}) \cdot (\sigma n)^{1/2-1/p} (\log n)^{1/2} \cdot \log \frac{p^2 d\sigma}{\tau}$$

Proof. The proof is similar to the case of $p < 2$. We again introduce a parameter λ. For radii below λ, the bound is the same as Lemma 7.4.1. For radii above λ, we use Lemma 6.4.1 to bound

$$\log E(B^p, d_X, t) \leq \log E(B^p, 2p \cdot \sigma^{1/2-1/p} \cdot \|\cdot\|_\infty, t) \leq \log E(B^p, B^\infty, t/2p \cdot \sigma^{1/2-1/p})$$

Then by Corollary 6.4.9,

$$\begin{aligned}\log E(B^p, B^\infty, t/2p \cdot \sigma^{1/2-1/p}) &\leq \log E(B^2, B^\infty, t/2p \cdot (\sigma n)^{1/2-1/p}) \\ &\leq O(p^2)\frac{(\log n)\cdot\tau}{t^2} \cdot (\sigma n)^{1-2/p}\end{aligned}$$

so the entropy integral gives a bound of

$$O(p\tau^{1/2})\cdot(\sigma n)^{1/2-1/p}(\log n)^{1/2}\cdot\int_\lambda^{\operatorname{diam}(B^p(\mathbf{A}))} \frac{1}{t}\, dt \leq O(p\tau^{1/2})\cdot(\sigma n)^{1/2-1/p}(\log n)^{1/2}\cdot\log\frac{p\sigma^{1/2}}{\lambda}$$

Choosing $\lambda = \sqrt{\tau/d}$ yields the claimed conclusion. □

7.4.2 Sensitivity sampling, $p < 2$

Our first result is a sensitivity sampling guarantee for $p < 2$.

Theorem 7.4.3 (Sensitivity sampling for $p < 2$)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $1 \le p < 2$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix with sampling probabilities $q_i = \min\{1, 1/n + \boldsymbol{\sigma}_i^p(\mathbf{A})/\alpha\}$ for an oversampling parameter α set to

$$\begin{aligned}\frac{1}{\alpha} &= \frac{\mathfrak{S}^p(\mathbf{A})^{2/p-1}}{\varepsilon^2}\left[O(l\log n)^{2/p-1}\left(\frac{\log d}{2-p} + \log\frac{l\log n}{\varepsilon}\right)(\log d)^2 + l\right] \\ &= \frac{\mathfrak{S}^p(\mathbf{A})^{2/p-1}}{\varepsilon^2}\operatorname{poly}\left(\log n, \log\frac{1}{\delta}, \frac{1}{2-p}\right)\end{aligned}$$

for

$$l = O\left(\log\frac{1}{\delta} + \log\log n + \log\frac{1}{2-p} + \log\frac{d}{\varepsilon}\right).$$

Then, with probability at least $1-\delta$, simultaneously for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{x}\|_p^p.$$

Furthermore, with probability at least $1-\delta$, $\mathbf{S}$ samples

$$\frac{\mathfrak{S}^p(\mathbf{A})^{2/p}}{\varepsilon^2}\operatorname{poly}\left(\log n, \log\frac{1}{\delta}, \log\frac{1}{2-p}\right)$$

rows.

Proof. Our approach is to bound

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p - 1\right|^l$$

for a large even integer l. Using Lemma 2.3.2, we first bound

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p - 1\right|^l \le (2\pi)^{l/2}\mathop{\mathbf{E}}_{\mathbf{S}}\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n}\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\sum_{i\in S}\boldsymbol{\varepsilon}_i|[\mathbf{S}\mathbf{A}\mathbf{x}](i)|^p\right|^l$$

where $S = \{i \in [n] : q_i < 1\}$. For simplicity of presentation, we assume $S = [n]$, which will not affect our proof.

By Theorem 6.1.9, there exists a matrix $\mathbf{A}' \in \mathbb{R}^{m_1\times d}$ with $m_1 = O(d(\log d)^3)$ such that

$$\|\mathbf{A}'\mathbf{x}\|_p^p = (1\pm 1/2)\|\mathbf{A}\mathbf{x}\|_p^p$$

for all $\mathbf{x} \in \mathbb{R}^d$. Furthermore, because $\mathbf{A}'$ in Theorem 6.1.9 is constructed by random sampling, Lemma 7.3.4 shows that $\mathfrak{S}^p(\mathbf{A}') \le 8\mathfrak{S}^p(\mathbf{A})$ (note that we only need existence of this matrix). We then construct a matrix $\mathbf{A}'' \in \mathbb{R}^{m_2\times d}$ with $m_2 = O(\alpha^{-1}\mathfrak{S}^p(\mathbf{A}) + d(\log d)^3) = O(\alpha^{-1}\mathfrak{S}^p(\mathbf{A}))$ such that

$$\sigma \coloneqq \max_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A}'') \le \alpha,$$

$\mathfrak{S}^p(\mathbf{A}') = \mathfrak{S}^p(\mathbf{A}'')$, and $\|\mathbf{A}'\mathbf{x}\|_p = \|\mathbf{A}''\mathbf{x}\|_p$ for all $\mathbf{x} \in \mathbb{R}^d$ by viewing $\mathbf{A}'$ as an $(m_1 + \alpha^{-1}\mathfrak{S}^p(\mathbf{A}))\times d$ matrix with all zeros except for the first m_1 rows and then applying Lemma 7.3.3.

Now let

$$\mathbf{A}''' := \begin{pmatrix} \mathbf{A}'' \\ \mathbf{SA} \end{pmatrix}$$

be the $(m_2 + n_{\mathbf{S}}) \times d$ matrix formed by the vertical concatenation of $\mathbf{A}''$ with $\mathbf{SA}$, where $n_{\mathbf{S}}$ is the number of rows sampled by $\mathbf{S}$.

Sensitivity bounds for $\mathbf{A}'''$. We will first bound the ℓ_p sensitivities of $\mathbf{A}'''$. For any row i corresponding to a row of $\mathbf{A}''$, the ℓ_p sensitivities are already bounded by α, and furthermore, ℓ_p sensitivities can clearly only decrease with row additions. For any row i corresponding to a row of $\mathbf{SA}$ that is sampled with probability $q_i < 1$, we have that

$$\frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{A}'''\mathbf{x}\|_p^p} \leq 2\frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{Ax}\|_p^p} \leq 2\frac{1}{q_i}\frac{|[\mathbf{Ax}](i)|^p}{\|\mathbf{Ax}\|_p^p} \leq 2\frac{\boldsymbol{\sigma}_i^p(\mathbf{A})}{q_i} = 2\alpha.$$

Thus, we have that $\boldsymbol{\sigma}_i^p(\mathbf{A}''') \leq 2\alpha$ for every row i of $\mathbf{A}'''$.

With a bound on the ℓ_p sensitivities of $\mathbf{A}'''$ in hand, we may then convert this into a bound on the leverage scores of $\mathbf{A}'''$ using Lemma 7.3.2, which gives

$$\tau := \max_{i=1}^{n} \boldsymbol{\tau}_i(\mathbf{A}''') \leq (2\alpha)^{2/p}(m_2 + n_{\mathbf{S}})^{2/p-1}$$

where $n_{\mathbf{S}}$ is the number of nonzero entries of $\mathbf{S}$.

Moment bounds on the sampling error. We now fix a choice of $\mathbf{S}$, and define

$$F_{\mathbf{S}} := \sup_{\|\mathbf{Ax}\|_p=1} \Big| \|\mathbf{SAx}\|_p^p - 1 \Big|.$$

Note that the event that $n_{\mathbf{S}}$ is at least

$$n_{\text{thresh}} := O(l \log n)\, \mathbf{E}[n_S] = O(l \log n)\alpha^{-1}\mathfrak{S}^p(\mathbf{A}),$$

occurs with probability at most $\text{poly}(n)^{-l}$ by Chernoff bounds over the randomness of $\mathbf{S}$, and

$$F_{\mathbf{S}}^l \leq \left[1 + \sum_{i=1}^{n} \frac{1}{q_i}\right]^l \leq (n+1)^{2l},$$

and thus this event contributes at most $\text{poly}(n)^{-l}$ to the moment bound $\mathbf{E}\, F_{\mathbf{S}}^l$. Thus, we focus on bounding $\mathbf{E}\, F_{\mathbf{S}}^l$ conditioned on $n_{\mathbf{S}} \leq n_{\text{thresh}}$. Define

$$G_{\mathbf{S}} := \sup_{\|\mathbf{A}'''\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|$$

for $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{m_2+n_{\mathbf{S}}})$. Then,

$$\|\mathbf{A}'''\mathbf{x}\|_p^p \leq (1 + 2 + F_{\mathbf{S}})\|\mathbf{Ax}\|_p^p$$

so

$$\begin{aligned} F_{\mathbf{S}}^l &\le 2^l \sup_{\|\mathbf{A}\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|^l \\ &\le 2^l (1+2+F_{\mathbf{S}})^l \sup_{\|\mathbf{A}'''\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|^l \\ &\le 2^{2l-1}(3^l + F_{\mathbf{S}}^l) G_{\mathbf{S}}^l. \end{aligned} \tag{7.1}$$

We then take expectations on both sides with respect to $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{m_2+n_{\mathbf{S}}})$, and bound the right hand side using Lemma 2.3.7, which gives

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \le (2\mathcal{E})^l \frac{\mathcal{E}}{\mathcal{D}} + O(\sqrt{l}\mathcal{D})^l$$

where $\mathcal{E}$ is the entropy integral and $\mathcal{D} = 4\sigma^{1/2}$ is the diameter by Lemma 6.4.2. We have by Lemma 7.4.1 that

$$\begin{aligned} \mathcal{E} &\le O(\tau^{1/2}) \left(\frac{\log d}{2-p} + \log(m_2 + n_{\mathbf{S}}) \right)^{1/2} \log \frac{d\sigma}{\tau} \\ &\le O(\alpha^{1/p} (m_2 + n_{\mathbf{S}})^{1/p-1/2}) \left(\frac{\log d}{2-p} + \log(m_2 + n_{\mathbf{S}}) \right)^{1/2} \log \frac{d\sigma}{\tau} \end{aligned}$$

Thus, conditioned on $n_{\mathbf{S}} \le n_{\text{thresh}}$, we have that

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \le \left[O(\alpha^{1/p} n_{\text{thresh}}^{1/p-1/2}) \left(\frac{\log d}{2-p} + \log n_{\text{thresh}} \right)^{1/2} \log d \right]^l + O(\sqrt{l}\sqrt{\alpha})^l.$$

Note that

$$\alpha^{1/p} n_{\text{thresh}}^{1/p-1/2} = O(l \log n)^{1/p-1/2} \alpha^{1/p} (\alpha^{-1} \mathfrak{S}^p(\mathbf{A}))^{1/p-1/2} = O(l \log n)^{1/p-1/2} \alpha^{1/2} \mathfrak{S}^p(\mathbf{A})^{1/p-1/2},$$

which shows that

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \le \varepsilon^l \delta$$

due to our choice of α and l.

Now if we take conditional expectations on both sides of (7.1) conditioned on the event $\mathcal{F}$ that $n_{\mathbf{S}} \le n_{\text{thresh}}$, then we have

$$\mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}] \le 2^{2l-1}(3^l + \mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}])\varepsilon^l \delta \le (3^l + \mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}])(4\varepsilon)^l \delta$$

which means

$$\mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}] \le \frac{(12\varepsilon)^l \delta}{1 - (4\varepsilon)^l \delta} \le 2(12\varepsilon)^l \delta$$

for $(4\varepsilon)^l \delta \le 1/2$. We thus have

$$\mathbf{E}[F_{\mathbf{S}}^l] \le \frac{(12\varepsilon)^l \delta}{1 - (4\varepsilon)^l \delta} \le 2(12\varepsilon)^l \delta + \text{poly}(n)^{-l}$$

altogether. Finally, we have by a Markov bound that

$$F_{\mathbf{S}}^l \le 2(12\varepsilon)^l + \frac{1}{\delta}\operatorname{poly}(n)^l \le 3(12\varepsilon)^l$$

with probability at least $1-\delta$, which means that

$$F_{\mathbf{S}} \le 3 \cdot 12\varepsilon = 36\varepsilon$$

with probability at least $1-\delta$. Rescaling ε by constant factors yields the claimed result. □

7.4.3 Sensitivity sampling, $p > 2$

For $p > 2$, we first need a construction of a matrix with a small number of rows and small sensitivity. While this construction can be made to be a randomized algorithm succeeding with high probability, it uses a sophisticated recursive sampling strategy which may be undesirable. In Theorem 7.4.5, we use this result to show that a more direct one-shot sensitivity sampling can in fact achieve a similar guarantee.

Lemma 7.4.4 (Recursive sensitivity sampling)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $2 < p < \infty$. Let $0 < \varepsilon < 1$. Then, there exists a matrix $\mathbf{A}' \in \mathbb{R}^{m\times d}$ for

$$m = O(p^2)\frac{\mathfrak{S}^p(\mathbf{A})^{2-2/p}}{\varepsilon^2}\log(pd)^2\log\frac{pd}{\varepsilon}$$

such that

$$\|\mathbf{A}'\mathbf{x}\|_p^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{x}\|_p^p$$

for every $\mathbf{x} \in \mathbb{R}^d$ and $\mathfrak{S}^p(\mathbf{A}') \le (1+O(\varepsilon))\mathfrak{S}^p(\mathbf{A})$.

Proof. Let $\mathbf{A}' \in \mathbb{R}^{m\times d}$ be the flattened isometric matrix given by Lemma 7.3.3 with $C = 4$, where $m \le (5/4)n$. Then for all $i \in [m]$, we have that

$$\boldsymbol{\sigma}_i^p(\mathbf{A}') \le 4\frac{\mathfrak{S}^p(\mathbf{A})}{n} \le 5\frac{\mathfrak{S}^p(\mathbf{A}')}{m}.$$

Now consider the random sampling matrix $\mathbf{S}$ with sampling probabilities $q_i = 1/2$. Note then that sampling with probability $q_i = 1/2$ and scaling by $1/q_i = 2$ corresponds to muliplying by the random variable $\boldsymbol{\varepsilon}_i + 1$, where $\boldsymbol{\varepsilon}_i$ is a Rademacher variable. Thus,

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{S}}\sup_{\|\mathbf{A}'\mathbf{x}\|_p=1}\left|\|\mathbf{S}\mathbf{A}'\mathbf{x}\|_p^p - 1\right| &= \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}}\sup_{\|\mathbf{A}'\mathbf{x}\|_p=1}\left|\sum_{i=1}^n(\boldsymbol{\varepsilon}_i+1)|[\mathbf{A}'\mathbf{x}](i)|^p - |[\mathbf{A}'\mathbf{x}](i)|^p\right| \\ &= \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}}\sup_{\|\mathbf{A}'\mathbf{x}\|_p=1}\left|\sum_{i=1}^n\boldsymbol{\varepsilon}_i|[\mathbf{A}'\mathbf{x}](i)|^p\right|.\end{aligned}$$

By Lemma 2.3.2 and Theorem 2.3.6, this is bounded by

$$O(1)\int_0^\infty \sqrt{\log E(T, d_X, u)}\,du \le O(p\tau^{1/2})\cdot(\sigma n)^{1/2-1/p}(\log n)^{1/2}\cdot\log\frac{p^2 d\sigma}{\tau}$$

where τ is an upper bound on the leverage scores of $\mathbf{A}'$ and σ is an upper bound on the ℓ_p sensitivities of $\mathbf{A}'$. By Lemma 7.3.1, we have that $\tau \leq \sigma$, and furthermore, we can take $\sigma = 5\mathfrak{S}^p(\mathbf{A}')/m$. Thus, the resulting bound on the expected sampling error is at most

$$\varepsilon_{\mathbf{A}} \coloneqq O(p)\frac{\mathfrak{S}^p(\mathbf{A})^{1-1/p}}{\sqrt{n}}(\log n)^{1/2}\log(pd)$$

so with probability at least $99/100$, the same bound holds up to a factor of 100. Furthermore, $\mathbf{S}$ samples $m/2 \leq (5/8)n$ rows in expectation, so by Markov's inequality, it samples at most $(3/2)m/2 \leq (15/16)n$ rows with probability at least $1/3$. We also have that

$$\frac{\boldsymbol{\sigma}_i^p(\mathbf{A})}{q_i} = 2\boldsymbol{\sigma}_i^p(\mathbf{A}') \leq 10\frac{\mathfrak{S}^p(\mathbf{A}')}{m}$$

so by Lemma 7.3.5, we have that

$$\mathbf{Pr}\{\mathfrak{S}^p(\mathbf{S}\mathbf{A}') = (1 \pm O(\varepsilon_{\mathbf{A}}))\mathfrak{S}^p(\mathbf{A})\} \geq \frac{99}{100}.$$

By a union bound, $\mathbf{S}\mathbf{A}'$ samples at most $(15/16)n$ rows, has sampling error at most ε_n, and has ℓ_p total sensitivity at most $(1 + O(\varepsilon_{\mathbf{A}}))\mathfrak{S}^p(\mathbf{A})$ with probability at least $1/3 - 1/100 - 1/100 > 0$. Thus, such an instantiation of $\mathbf{S}\mathbf{A}'$ exists.

We now recursively apply our reasoning, by repeatedly applying the flattening and sampling operation. Note that each time we repeat this procedure, the number of rows goes down by a factor of $15/16$, while the total sensitivity and total sampling error accumulates. Let $\mathbf{A}_l$ denote the matrix obtained after l recursive applications of this procedure and let n_l denote the number of rows of $\mathbf{A}_l$. Then,

$$\begin{aligned}
\varepsilon_{\mathbf{A}_{l+1}} &= O(p)\frac{\mathfrak{S}^p(\mathbf{A}_{l+1})^{1-1/p}}{\sqrt{n}_{l+1}}(\log n_{l+1})^{1/2}\log(pd) \\
&\geq (1 - O(\varepsilon_{\mathbf{A}_l}))O(p)\frac{\mathfrak{S}^p(\mathbf{A}_l)^{1-1/p}}{\sqrt{n}_{l+1}}(\log n_{l+1})^{1/2}\log(pd) \\
&\geq \sqrt{\frac{16}{15}}(1 - O(\varepsilon_{\mathbf{A}_l}))O(p)\frac{\mathfrak{S}^p(\mathbf{A}_l)^{1-1/p}}{\sqrt{n}_l}(\log n_l)^{1/2}\log(pd) \\
&\geq \frac{101}{100}\cdot\varepsilon_{\mathbf{A}_l}
\end{aligned}$$

as long as $\varepsilon_{\mathbf{A}_l}$ is less than some absolute constant. Thus, the sum of the $\varepsilon_{\mathbf{A}_l}$ are dominated by the last $\varepsilon_{\mathbf{A}_l}$, up to a constant factor. Now let L be the smallest integer l such that $\varepsilon_{\mathbf{A}_l} \leq \varepsilon$. Then, we have that

$$\mathfrak{S}^p(\mathbf{A}_L) \leq (1 + O(\varepsilon))\mathfrak{S}^p(\mathbf{A})$$

and thus

$$\|\mathbf{A}_L\mathbf{x}\|_p^p = (1 \pm O(\varepsilon))\|\mathbf{A}\mathbf{x}\|_p^p$$

for every $\mathbf{x} \in \mathbb{R}^d$. Furthermore, n_L satisfies

$$\varepsilon = O(p)\frac{\mathfrak{S}^p(\mathbf{A})^{1-1/p}}{\sqrt{n}_L}(\log n_L)^{1/2}\log(pd)$$

or

$$n_L = O(p^2)\frac{\mathfrak{S}^p(\mathbf{A})^{2-2/p}}{\varepsilon^2}\log(pd)^2\log\frac{pd}{\varepsilon}.$$

□

Theorem 7.4.5 (Sensitivity Sampling for $p > 2$)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $2 < p < \infty$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix with sampling probabilities $q_i = \min\{1, 1/n + \boldsymbol{\sigma}_i^p(\mathbf{A})/\alpha\}$ for an oversampling parameter α set to

$$\frac{1}{\alpha} = O(p^2)\mathfrak{S}^p(\mathbf{A})^{1-2/p}(l\log n)^{1-2/p}\log(pd)\log\frac{l\log n}{\varepsilon} + O(p^2)l$$

for

$$l = O\left(\log\frac{1}{\delta} + \log\log n + \log p + \log\frac{\mathfrak{S}^p(\mathbf{A})}{\varepsilon}\right).$$

Then, with probability at least $1-\delta$, simultaneously for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{x}\|_p^p.$$

Furthermore, with probability at least $1-\delta$, $\mathbf{S}$ samples

$$\frac{\mathfrak{S}^p(\mathbf{A})^{2-2/p}}{\varepsilon^2}\operatorname{poly}\left(\log n, \log\frac{1}{\delta}, p\right)$$

rows.

Proof. Our approach is to bound

$$\mathop{\mathbf{E}}_{\mathbf{S}}\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p - 1\right|^l$$

for a large even integer l. Using Lemma 2.3.2, we first bound

$$\mathop{\mathbf{E}}_{\mathbf{S}}\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p - 1\right|^l \le (2\pi)^{l/2}\mathop{\mathbf{E}}_{\mathbf{S}}\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_n)}\sup_{\|\mathbf{A}\mathbf{x}\|_p=1}\left|\sum_{i\in S}\mathbf{g}_i|[\mathbf{S}\mathbf{A}\mathbf{x}](i)|^p\right|^l$$

where $S = \{i \in [n] : q_i < 1\}$. For simplicity of presentation, we assume $S = [n]$, which will not affect our proof.

By Lemma 7.4.4, there exists a matrix $\mathbf{A}' \in \mathbb{R}^{m_1\times d}$ with $m_1 = O(\mathfrak{S}^{2-2/p}\log(pd)^3)$ such that

$$\|\mathbf{A}'\mathbf{x}\|_p^p = (1\pm 1/2)\|\mathbf{A}\mathbf{x}\|_p^p$$

for all $\mathbf{x} \in \mathbb{R}^d$, and $\mathfrak{S}^p(\mathbf{A}') \le O(1)\mathfrak{S}^p(\mathbf{A})$. Then for $m_2 = O(m_1 + \mathfrak{S}^p(\mathbf{A})\alpha^{-1})$, let $\mathbf{A}'' \in \mathbb{R}^{m_2\times d}$ be the matrix given by Lemma 7.3.3 such that $\boldsymbol{\sigma}_i^p(\mathbf{A}'') \le \alpha$ for every $i \in [m_2]$ and $\|\mathbf{A}''\mathbf{x}\|_p = \|\mathbf{A}'\mathbf{x}\|_p$ for every $\mathbf{x} \in \mathbb{R}^d$. Now let

$$\mathbf{A}''' := \begin{pmatrix}\mathbf{A}''\\ \mathbf{S}\mathbf{A}\end{pmatrix}$$

be the $(m_2 + n_{\mathbf{S}}) \times d$ matrix formed by the vertical concatenation of $\mathbf{A}''$ with $\mathbf{S}\mathbf{A}$, where $n_{\mathbf{S}}$ is the number of rows sampled by $\mathbf{S}$.

Sensitivity bounds for $\mathbf{A}'''$. We will first bound the ℓ_p sensitivities of $\mathbf{A}'''$. For any row i corresponding to a row of $\mathbf{A}''$, the ℓ_p sensitivities are already bounded by α, and furthermore, ℓ_p sensitivities can only decrease with row additions. For any row i corresponding to a row of $\mathbf{SA}$ that is sampled with probability $q_i < 1$, we have that

$$\frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{A}'''\mathbf{x}\|_p^p} \leq \frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{A}''\mathbf{x}\|_p^p} \leq 2\frac{|[\mathbf{SAx}](i)|^p}{\|\mathbf{Ax}\|_p^p} \leq 2\alpha.$$

By Lemma 7.3.1, this immediately implies that the ℓ_2 sensitivities, or the leverage scores, are also bounded by 2α.

Moment bounds on sampling error. We now fix a choice of $\mathbf{S}$, and define

$$F_{\mathbf{S}} \coloneqq \sup_{\|\mathbf{Ax}\|_p=1} \left| \|\mathbf{SAx}\|_p^p - 1 \right|$$

Note that the event that $n_{\mathbf{S}}$ is at least

$$n_{\text{thresh}} \coloneqq O(l \log n)\, \mathbf{E}[n_S] = O(l \log n)\alpha^{-1}\mathfrak{S}^p(\mathbf{A}),$$

occurs with probability at most $\text{poly}(n)^{-l}$ by Chernoff bounds over the randomness of $\mathbf{S}$, and

$$F_{\mathbf{S}}^l \leq \left[1 + \sum_{i=1}^{n} \frac{1}{q_i}\right]^l \leq (n+1)^{2l},$$

and thus this event contributes at most $\text{poly}(n)^{-l}$ to the moment bound $\mathbf{E}\, F_{\mathbf{S}}^l$. Thus, we focus on bounding $\mathbf{E}\, F_{\mathbf{S}}^l$ conditioned on $n_{\mathbf{S}} \leq n_{\text{thresh}}$. Now define

$$G_{\mathbf{S}} \coloneqq \sup_{\|\mathbf{A}'''\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|$$

for $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{m_2+n_{\mathbf{S}}})$. Then,

$$\|\mathbf{A}'''\mathbf{x}\|_p^p \leq (1 + 2 + F_{\mathbf{S}})\|\mathbf{Ax}\|_p^p$$

so

$$\begin{aligned} F_{\mathbf{S}}^l &\leq 2^l \sup_{\|\mathbf{Ax}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|^l \\ &\leq 2^l (1 + 2 + F_{\mathbf{S}})^l \sup_{\|\mathbf{A}'''\mathbf{x}\|_p=1} \left| \sum_{i=1}^{m_2+n_{\mathbf{S}}} \mathbf{g}_i |[\mathbf{A}'''\mathbf{x}](i)|^p \right|^l \\ &\leq 2^{2l-1}(3^l + F_{\mathbf{S}}^l) G_{\mathbf{S}}^l. \end{aligned} \tag{7.2}$$

We then take expectations on both sides with respect to $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{m_2+n_{\mathbf{S}}})$, and bound the right hand side using Lemma 2.3.7, which gives

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \leq (2\mathcal{E})^l \frac{\mathcal{E}}{\mathcal{D}} + O(\sqrt{l}\mathcal{D})^l$$

where $\mathcal{E}$ is the entropy integral and $\mathcal{D} = 4p\sigma^{1/2}$ is the diameter by Lemma 6.4.2. We have by Lemma 7.4.2 that

$$\begin{aligned}\mathcal{E} &\le O(p\tau^{1/2}) \cdot (\sigma(m_2 + n_{\mathbf{S}}))^{1/2-1/p}(\log(m_2 + n_{\mathbf{S}}))^{1/2} \cdot \log \frac{p^2 d\sigma}{\tau} \\ &\le O(p\alpha^{1/2}) \cdot (\alpha(m_2 + n_{\mathbf{S}}))^{1/2-1/p}(\log(m_2 + n_{\mathbf{S}}))^{1/2} \cdot \log(pd).\end{aligned}$$

Thus, conditioned on $n_{\mathbf{S}} \le n_{\text{thresh}}$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l &\le \left[O(p\alpha^{1/2})(\alpha(m_2 + n_{\text{thresh}}))^{1/2-1/p}(\log(m_2 + n_{\text{thresh}}))^{1/2} \log(pd)\right]^l + O(\sqrt{l}p\sqrt{\alpha})^l \\ &\le \left[O(p\alpha^{1-1/p})n_{\text{thresh}}^{1/2-1/p}(\log n_{\text{thresh}})^{1/2} \log(pd)\right]^l + O(\sqrt{l}p\sqrt{\alpha})^l\end{aligned}$$

Note that

$$\alpha^{1-1/p}n_{\text{thresh}}^{1/2-1/p} = O(l\log n)^{1/2-1/p}\alpha^{1-1/p}(\alpha^{-1}\mathfrak{S}^p(\mathbf{A}))^{1/2-1/p} = O(l\log n)^{1/2-1/p}\alpha^{1/2}\mathfrak{S}^p(\mathbf{A})^{1/2-1/p},$$

which shows that

$$\mathop{\mathbf{E}}_{\mathbf{g}\sim\mathcal{N}(0,\mathbf{I}_{m_2+n_{\mathbf{S}}})} G_{\mathbf{S}}^l \le \varepsilon^l\delta$$

due to our choice of α and l.

Now if we take conditional expectations on both sides of (7.2) conditioned on the event $\mathcal{F}$ that $n_{\mathbf{S}} \le n_{\text{thresh}}$, then we have

$$\mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}] \le 2^{2l-1}(3^l + \mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}])\varepsilon^l\delta \le (3^l + \mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}])(4\varepsilon)^l\delta$$

which means

$$\mathbf{E}[F_{\mathbf{S}}^l \mid \mathcal{F}] \le \frac{(12\varepsilon)^l\delta}{1-(4\varepsilon)^l\delta} \le 2(12\varepsilon)^l\delta$$

for $(4\varepsilon)^l\delta \le 1/2$. We thus have

$$\mathbf{E}[F_{\mathbf{S}}^l] \le \frac{(12\varepsilon)^l\delta}{1-(4\varepsilon)^l\delta} \le 2(12\varepsilon)^l\delta + \text{poly}(n)^{-l}$$

altogether. Finally, we have by a Markov bound that

$$F_{\mathbf{S}}^l \le 2(12\varepsilon)^l + \frac{1}{\delta}\,\text{poly}(n)^l \le 3(12\varepsilon)^l$$

with probability at least $1-\delta$, which means that

$$F_{\mathbf{S}} \le 3 \cdot 12\varepsilon = 36\varepsilon$$

with probability at least $1-\delta$. Rescaling ε by constant factors yields the claimed result. □

Chapter 8

Root leverage score sampling [WY23c, WY24b]

Our techniques and observations used to obtain our improved ℓ_p sensitivity sampling theorem (Theorem 7.1.2) lead to improved guarantees for yet another generalization of ℓ_2 leverage score sampling, known as *root leverage score sampling*. In root leverage score sampling, the sampling probabilities are taken to be proportional to the $(p/2)$-*th root* of the ℓ_2 leverage scores, and have found applications as upper bounds to sensitivities for more general loss functions with less structure than the ℓ_p losses, including the Huber loss and the logistic loss [CW15a, MSSW18, GPV21]. Here, the idea is that the ℓ_p sensitivities (Definition 6.1.2) can be bounded by the $(p/2)$-th roots of the ℓ_2 leverage scores, since for $p < 2$, we have

$$\frac{|\mathbf{y}(i)|^p}{\|\mathbf{y}\|_p^p} \leq \frac{|\mathbf{y}(i)|^p}{\|\mathbf{y}\|_2^p} \leq \left(\frac{|\mathbf{y}(i)|^2}{\|\mathbf{y}\|_2^2}\right)^{p/2}$$

and for $p > 2$, we have

$$\frac{|\mathbf{y}(i)|^p}{\|\mathbf{y}\|_p^p} \leq n^{p/2-1}\frac{|\mathbf{y}(i)|^p}{\|\mathbf{y}\|_2^p} \leq n^{p/2-1}\left(\frac{|\mathbf{y}(i)|^2}{\|\mathbf{y}\|_2^2}\right)^{p/2}$$

(see also Lemma 7.3.2) and thus an ℓ_p sensitivity sampling argument immediately applies for the $(p/2)$-th roots of the ℓ_2 leverage scores. Furthermore, the unlike the ℓ_p sensitivities themselves, ℓ_2 leverage scores can be computed quickly [SS11, DMMW12, CW13, LMP13, CLM+15] and thus root leverage score sampling has been a popular choice for fast algorithms.

Note that the number of rows sampled is actually $\mathrm{poly}(n)$ for any $p \neq 2$, which can be far larger than the usual bound of $\mathrm{poly}(d)$. Indeed, we have the following tight bounds on the sum of these scores:

Lemma 8.0.1 (Sum of root leverage scores[WY24b])**.** Let $0 < p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Then for $p < 2$,

$$\sum_{i=1}^{n} \boldsymbol{\tau}_i(\mathbf{A})^{p/2} \leq n^{1-p/2}d^{p/2}$$

and for $p > 2$,

$$\sum_{i=1}^{n} \min\{1, n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2}\} \leq 2n^{1-2/p}d.$$

Proof. The bound for $p < 2$ follows from relating ℓ_q norms, that is, if $\mathbf{y}(i) = \boldsymbol{\tau}_i(\mathbf{A})^{p/2}$, then

$$\sum_{i=1}^{n} \boldsymbol{\tau}_i(\mathbf{A})^{p/2} = \|\mathbf{y}\|_1 \leq n^{1-p/2}\|\mathbf{y}\|_{2/p} = n^{1-p/2}\left(\sum_{i=1}^{n} \boldsymbol{\tau}_i(\mathbf{A})\right)^{p/2} = n^{1-p/2}d^{p/2}.$$

For $p > 2$, we bound the sum of the scores for rows $i \in [n]$ with $n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2} \leq 1 \iff \boldsymbol{\tau}_i(\mathbf{A}) \leq n^{2/p-1}$ by

$$\begin{aligned}
\sum_{i\in[n]:n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2}\leq 1} n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2} &= n^{p/2-1} \sum_{i\in[n]:n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2}\leq 1} \boldsymbol{\tau}_i(\mathbf{A}) \cdot \boldsymbol{\tau}_i(\mathbf{A})^{p/2-1} \\
&\leq n^{p/2-1} \sum_{i\in[n]:n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2}\leq 1} \boldsymbol{\tau}_i(\mathbf{A}) \cdot (n^{2/p-1})^{p/2-1} \\
&\leq (n^{2/p})^{p/2-1} \sum_{i\in[n]:n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2}\leq 1} \boldsymbol{\tau}_i(\mathbf{A}) \leq n^{1-2/p}d
\end{aligned}$$

On the other hand, there are at most $n^{1-2/p}d$ rows with $\boldsymbol{\tau}_i(\mathbf{A}) \geq n^{2/p-1}$, so the contribution of the rest of the rows is also at most $n^{1-2/p}d$. □

However, because the exponent on n is less than 1 in both cases, we repeatedly apply a subsampling procedure to reduce the number of rows to $\text{poly}(d)$. In fact, in the work of [WY23c], we show that $(p/2)$-th root leverage score sampling allows us to simultaneously control both the ℓ_p sensitivity scores and ℓ_2 leverage scores just as with ℓ_p Lewis weight sampling, and thus similar chaining arguments allow us to show that the sum of the root leverage scores in Lemma 8.0.1 is the resulting row count of this sampling algorithm, up to roughly an ε^2 factor. If we recursively apply this sampling algorithm multiply times until the row reduction gives no more improvements, then the row count that we converge to is roughly

$$n = \varepsilon^{-2}n^{1-p/2}d^{p/2} \iff n = \varepsilon^{-4/p}d$$

for $p < 2$ and

$$n = \varepsilon^{-2}n^{1-2/p}d \iff n = \varepsilon^{-p}d^{p/2}$$

for $p > 2$. Furthermore, it is not hard to see that we converge to this value up to a constant factor in roughly $O(\log\log n)$ rounds.

Theorem 8.0.2 (Root leverage score sampling, $p < 2$ [WY23c])**.** Let $1 \leq p < 2$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\alpha > 0$ and let $q_i = \min\{1, \boldsymbol{\tau}_i^p(\mathbf{A})^{p/2}/\alpha\}$ for $i \in [n]$. Let $\mathbf{S} \in \mathbb{R}^{n\times n}$ be the ℓ_p sampling matrix with probabilities $\{q_i\}_{i=1}^n$. Then, with probability at least $1 - 1/\text{poly}(n)$, there is an α such that $\mathbf{S}$ is an ℓ_p subspace embedding satisfying Definition 1.1.1 with $\kappa = (1+\varepsilon)$, and furthermore, $\mathbf{S}$ has at most r nonzero rows, for

$$r = \varepsilon^{-2}n^{1-p/2}d^{p/2}\,\text{poly}\log n.$$

Recursively applying this result gives a matrix $\mathbf{S}$ with

$$r = \varepsilon^{-4/p} d \operatorname{poly}\log n.$$

We obtain a similar result for $p > 2$ in the work [WY24b]:

Theorem 8.0.3 (Root leverage score sampling, $p > 2$ [WY24b])**.** Let $2 < p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Let $\alpha > 0$ and let $q_i = \min\{1, n^{p/2-1}\boldsymbol{\tau}_i^p(\mathbf{A})^{p/2}/\alpha\}$ for $i \in [n]$. Let $\mathbf{S} \in \mathbb{R}^{n \times n}$ be the ℓ_p sampling matrix with probabilities $\{q_i\}_{i=1}^n$. Then, with probability at least $1 - 1/\operatorname{poly}(n)$, there is an α such that $\mathbf{S}$ is an ℓ_p subspace embedding satisfying Definition 1.1.1 with $\kappa = (1+\varepsilon)$, and furthermore, $\mathbf{S}$ has at most r nonzero rows, for

$$r = \varepsilon^{-2} n^{1-2/p} d \operatorname{poly}\log n.$$

Recursively applying this result gives a matrix $\mathbf{S}$ with

$$r = \varepsilon^{-p} d^{p/2} \operatorname{poly}\log n.$$

Remark 8.0.4. Note that Theorems 8.0.2 and 8.0.3 achieve a nearly optimal dependence on d, while it is suboptimal in the ε dependence (see Theorem 6.1.5 in Chapter 6 for a discussion on the lower bound results of [LWW21]).

8.1 Analysis of root leverage score sampling

The idea of bounding Dudley's entropy integral by separately parameterizing by the maximum leverage score and maximum sensitivity score in Section 7.4.1 is also useful in the analysis of root leverage score sampling. We will show the following theorem, which shows that root leverage score sampling yields ℓ_p affine embeddings, which is a slight extension of our results for ℓ_p subspace embeddings that we will later need in a later chapter (Chapter 14). Note that our ℓ_p subspace embedding theorems, Theorems 8.0.2 and 8.0.3, are easily recovered by setting $\mathbf{b} = 0$ and $R = 0$.

Theorem 8.1.1 (Root leverage score sampling)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Let $1 \le p < \infty$. Let $R \ge \|\mathbf{b}\|_p$. Suppose that

$$\frac{|\mathbf{b}(i)|^p}{R^p} \le \begin{cases} \min\{1, n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2}\} & p > 2 \\ \boldsymbol{\tau}_i(\mathbf{A})^{p/2} & p < 2 \end{cases}$$

for every $i \in [n]$. Let $\alpha = \Theta(\varepsilon^2)/((\log n)^3 + \log(1/\delta))$ and let

$$q_i \ge \begin{cases} \min\{1, n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2}/\alpha\} & p > 2 \\ \min\{1, \boldsymbol{\tau}_i(\mathbf{A})^{p/2}/\alpha\} & p < 2 \end{cases}$$

Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 6.1.1) with sampling probabilities $\{q_i\}_{i=1}^n$. Then, with probability at least $1 - \delta$, for every $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{S}(\mathbf{A}\mathbf{x} + \mathbf{b})\|_p^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_p^p \pm \varepsilon R^p$$

and

$$\mathsf{nnz}(\mathbf{S}) = \begin{cases} O(\varepsilon^{-2} n^{1-2/p} d\big((\log n)^3 + \log \frac{1}{\delta}\big) & p > 2 \\ O(\varepsilon^{-2} n^{1-p/2} d^{p/2}\big((\log n)^3 + \log \frac{1}{\delta}\big) & p < 2 \end{cases}.$$

Recursively applying this result for $O(\log \log n)$ reduces the number of rows to

$$\begin{cases} O(\varepsilon^{-p} d^{p/2}\big((\log n)^3 + \log \frac{1}{\delta}\big)^{p/2} & p > 2 \\ O(\varepsilon^{-4/p} d\big((\log n)^3 + \log \frac{1}{\delta}\big)^{2/p} & p < 2 \end{cases}.$$

8.1.1 Reduction to a small number of scales

Our first task is to reduce the proof of Theorem 8.1.1 to showing a similar theorem when $\mathbf{Ax}$ is restricted to a certain scale, for a small number of scales. This is shown in the following lemma:

Lemma 8.1.2. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Let $1 \leq p < \infty$. Let $R_0 \geq \|\mathbf{b}\|_p$ and $0 < \varepsilon < 1/2$. Suppose that

$$\sup_{\|\mathbf{Ax}\|_p = 1} \Big| \|\mathbf{SAx}\|_p^p - \|\mathbf{Ax}\|_p^p \Big| \leq \varepsilon$$

and that

$$\sup_{\|\mathbf{Ax}\|_p \leq R_i} \Big| \|\mathbf{S}(\mathbf{Ax} + \mathbf{b})\|_p^p - \|\mathbf{Ax} + \mathbf{b}\|_p^p \Big| \leq \varepsilon R_i^p$$

holds for each $R_i = 2^i \cdot R_0$, $i \in [I]$, where $I = O(\log \varepsilon^{-1})$. Then,

$$\|\mathbf{S}(\mathbf{Ax} + \mathbf{b})\|_p^p = (1 \pm 2 \cdot 4^p \varepsilon)\|\mathbf{Ax} + \mathbf{b}\|_p^p \pm 2^p \varepsilon R_0^p$$

for every $\mathbf{x} \in \mathbb{R}^d$.

Proof. First note that if $\|\mathbf{Ax}\|_p \leq R_1$, then we immediately have

$$\|\mathbf{S}(\mathbf{Ax} + \mathbf{b})\|_p^p = \|\mathbf{Ax} + \mathbf{b}\|_p^p \pm \varepsilon R_1^p = \|\mathbf{Ax} + \mathbf{b}\|_p^p \pm 2^p \varepsilon R_0^p.$$

Next, suppose that $\|\mathbf{Ax}\|_p \geq R_0/\varepsilon$. Note that

$$\|\mathbf{Ax} + \mathbf{b}\|_p = \|\mathbf{Ax}\|_p \pm \|\mathbf{b}\|_p = (1 \pm \varepsilon)\|\mathbf{Ax}\|_p$$

and similarly,

$$\|\mathbf{S}(\mathbf{Ax} + \mathbf{b})\|_p = \|\mathbf{SAx}\|_p \pm \|\mathbf{Sb}\|_p = (1 \pm 4\varepsilon)\|\mathbf{SAx}\|_p$$

Thus,

$$\|\mathbf{S}(\mathbf{Ax} + \mathbf{b})\|_p = (1 \pm 4\varepsilon)(1 \pm \varepsilon)\|\mathbf{Ax} + \mathbf{b}\|_p = (1 \pm 7\varepsilon)\|\mathbf{Ax} + \mathbf{b}\|_p.$$

Finally, we handle the intermediate scales between R_0 and R_0/ε. Consider $\mathbf{x}$ such that $R_i \leq \|\mathbf{Ax}\|_p < 2 \cdot R_i$. Note then that

$$\|\mathbf{Ax} + \mathbf{b}\|_p \geq \|\mathbf{Ax}\|_p - \|\mathbf{b}\|_p \geq \|\mathbf{Ax}\|_p/2 \geq R_i/2$$

so

$$\|\mathbf{S}(\mathbf{Ax} + \mathbf{b})\|_p^p = \|\mathbf{Ax} + \mathbf{b}\|_p^p \pm \varepsilon \cdot (2R_i)^p = \|\mathbf{Ax} + \mathbf{b}\|_p^p \pm \varepsilon \cdot (4\|\mathbf{Ax} + \mathbf{b}\|_p)^p.$$

This covers all cases. □

8.1.2 Reduction to a Rademacher process with flat sensitivities

We now work towards bounding a quantity as the one used in Lemma 8.1.2. The following lemma follows from a standard symmetrization (Lemma 2.3.2) argument.

Lemma 8.1.3 (Reduction to Rademacher processes)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Let $1 \le p < \infty$. Let $R \ge \|\mathbf{b}\|_p$. Let $\mathbf{S}$ be a random ℓ_p sampling matrix (Definition 6.1.1). Then,

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{Ax}\|_p \le R} \left| \|\mathbf{S}(\mathbf{Ax}+\mathbf{b})\|_p^p - \|\mathbf{Ax}+\mathbf{b}\|_p^p \right|^l \le \mathop{\mathbf{E}}_{\mathbf{S}} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm 1\}^n} \sup_{\|\mathbf{Ax}\|_p \le R} \left| \sum_{i\in T} \boldsymbol{\varepsilon}_i |[\mathbf{S}(\mathbf{Ax}+\mathbf{b})](i)|^p \right|^l,$$

where $T \subseteq [n]$ is the set of rows with sampling probability $q_i < 1$.

We will further reduce the problem to a similar problem for an instance with "flat sensitivities". For this, we show the following flattening lemma, which shows how to obtain an ℓ_p isometry that simultaneously flatten all ℓ_q sensitivities.

Lemma 8.1.4 (Flattening all sensitivities)**.** Let $1 \le p < \infty$ and $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Let $0 < \alpha < 1$. Then, there exists $\mathbf{A}' \in \mathbb{R}^{m\times d}$ and $\mathbf{b}' \in \mathbb{R}^m$ for $m = O(n\alpha^{-1})$ such that

$$\boldsymbol{\sigma}_i^q(\mathbf{A}') \le \alpha, \qquad \boldsymbol{\sigma}_i^q(\mathbf{b}') = \frac{|\mathbf{b}(i)|^q}{\|\mathbf{b}\|_q^q} \le \alpha$$

for every $i \in [m]$ and $1 \le q < \infty$. Furthermore, for any $1 \le q < \infty$ and $\mathbf{x} \in \mathbb{R}^d$, we have that $\|\mathbf{A}'\mathbf{x}+\mathbf{b}'\|_q = \Theta(\alpha^{1/p-1/q})\|\mathbf{Ax}+\mathbf{b}\|_q$.

Proof. Let $k \coloneqq \lceil 1/\alpha \rceil$. Then, we construct $\mathbf{A}' \in \mathbb{R}^{m\times d}$ for $m = nk$ by replacing the ith row $\mathbf{a}_i$ of $\mathbf{A}$ for every $i \in [n]$ with k copies of $\mathbf{a}/k^{1/p}$, and similarly for $\mathbf{b}$. Then, for every row $j \in [m]$ that is a copy of row $i \in [n]$, we have that

$$\boldsymbol{\sigma}_j^q(\mathbf{A}) = \sup_{\mathbf{Ax}\neq 0} \frac{|[\mathbf{A}'\mathbf{x}](i)|^q}{\|\mathbf{A}'\mathbf{x}\|_q^q} \le \sup_{\mathbf{Ax}\neq 0} \frac{\left|[k^{-1/p}\mathbf{Ax}](i)\right|^q}{k\cdot|[k^{-1/p}\mathbf{Ax}](i)|^q} \le \frac{1}{k} \le \alpha$$

as desired, and similarly for $\mathbf{b}$. The second conclusion holds since

$$\|\mathbf{A}'\mathbf{x}+\mathbf{b}'\|_q^q = k\cdot k^{-q/p}\|\mathbf{Ax}+\mathbf{b}\|_q^q = k^{1-q/p}\|\mathbf{Ax}+\mathbf{b}\|_q^q.$$

□

Using Lemma 8.1.4, we construct the following new instance with bounded ℓ_2 and ℓ_p sensitivities:

Lemma 8.1.5 (Flattened instance)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^{n\times d}$. Let $R \ge \|\mathbf{b}\|_p$. Suppose that

$$\frac{|\mathbf{b}(i)|^p}{R^p} \le \begin{cases} \min\{1, n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2}\} & p > 2 \\ \boldsymbol{\tau}_i(\mathbf{A})^{p/2} & p < 2 \end{cases}$$

for every $i \in [n]$. Let $0 < \alpha < 1$ and let

$$q_i \geq \begin{cases} \min\{1, n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})^{p/2}/\alpha\} & p > 2 \\ \min\{1, \boldsymbol{\tau}_i(\mathbf{A})^{p/2}/\alpha\} & p < 2 \end{cases}$$

Let $T \subseteq [n]$ be the set of rows $i \in [n]$ with $q_i < 1$. Let $\mathbf{S}$ be a diagonal matrix with $\mathbf{S}_{i,i} \leq 1/q_i^{1/p}$. Then, there is $\mathbf{A}'' \in \mathbb{R}^{m \times d}$ and $\mathbf{b}'' \in \mathbb{R}^m$ for $m = O(n/\alpha)$ such that

- $\boldsymbol{\tau}_i(\mathbf{A}'') \leq O(\alpha)$ for $p < 2$ and $\boldsymbol{\tau}_i(\mathbf{A}'') \leq O(\alpha)/n^{1-2/p}$ for $p > 2$, for every $i \in [m]$
- $\boldsymbol{\sigma}_i^p(\mathbf{A}'') \leq O(\alpha)$ and $|\mathbf{b}''(i)|^p/R^p \leq O(\alpha)$ for every $i \in [m]$
- $\|\mathbf{A}''\mathbf{x} + \mathbf{b}''\|_p^p = \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_p^p + \|\mathbf{S}|_T(\mathbf{A}\mathbf{x} + \mathbf{b})\|_p^p$ for every $\mathbf{x} \in \mathbb{R}^d$
- $\|\mathbf{A}''\mathbf{x}\|_p^p = \|\mathbf{A}\mathbf{x}\|_p^p + \|\mathbf{S}|_T\mathbf{A}\mathbf{x}\|_p^p$ for every $\mathbf{x} \in \mathbb{R}^d$

Proof. Let $\mathbf{A}' \in \mathbb{R}^{m \times d}$ and $\mathbf{b}' \in \mathbb{R}^m$ be the flattened instances given by Lemma 8.1.4, where $m = O(n/\alpha)$. Now let

$$\mathbf{A}'' \coloneqq \begin{pmatrix} \mathbf{A}' \\ \mathbf{S}|_T\mathbf{A} \end{pmatrix}, \qquad \mathbf{b}'' \coloneqq \begin{pmatrix} \mathbf{b}' \\ \mathbf{S}|_T\mathbf{b} \end{pmatrix}$$

be the $(m + n_{\mathbf{S}}) \times d$ matrix and $(m + n_{\mathbf{S}})$-dimensional vector formed by the vertical concatenation of $\mathbf{A}'$ and $\mathbf{b}'$ with $\mathbf{SA}$ and $\mathbf{Sb}$, where $n_{\mathbf{S}}$ is the number of rows sampled by $\mathbf{S}$.

We now show how to bound the sensitivities of $\mathbf{A}''$ and $\mathbf{b}''$.

For any row i corresponding to a row of $\mathbf{A}'$, the ℓ_2 sensitivities are already bounded by α, and furthermore, ℓ_2 sensitivities can clearly only decrease with row additions. For any row i corresponding to a row of $\mathbf{SA}$ that is sampled with probability $q_i < 1$, we have that

$$\frac{|[\mathbf{SA}\mathbf{x}](i)|^2}{\|\mathbf{A}''\mathbf{x}\|_2^2} \leq \frac{|[\mathbf{SA}\mathbf{x}](i)|^2}{\|\mathbf{A}'\mathbf{x}\|_2^2} = \frac{|[\mathbf{SA}\mathbf{x}](i)|^2}{\Theta(\alpha^{2/p-1})\|\mathbf{A}\mathbf{x}\|_2^2} \leq \frac{1}{q_i^{2/p}} \frac{|[\mathbf{A}\mathbf{x}](i)|^2}{\Theta(\alpha^{2/p-1})\|\mathbf{A}\mathbf{x}\|_2^2} \leq \frac{\boldsymbol{\tau}_i(\mathbf{A})}{\Theta(\alpha^{2/p-1})q_i^{2/p}} = O(\alpha).$$

In fact, for $p > 2$, we have the stronger bound of

$$\frac{\boldsymbol{\tau}_i(\mathbf{A})}{\Theta(\alpha^{2/p-1})q_i^{2/p}} \leq \frac{O(\alpha)}{n^{1-2/p}}.$$

Thus, we have that $\boldsymbol{\tau}_i(\mathbf{A}'') = \boldsymbol{\sigma}_i^2(\mathbf{A}'') \leq O(\alpha)$ for every row i of $\mathbf{A}''$.

For $p < 2$, the max sensitivity is bounded by $O(\alpha)$ by the monotonicity of max sensitivities. For $p > 2$, we have by reverse monotonicity of max sensitivities that

$$\boldsymbol{\sigma}_i^p(\mathbf{A}) \leq n^{p/2-1}\boldsymbol{\tau}_i(\mathbf{A})$$

so for any row i corresponding to a row of $\mathbf{SA}$ sampled with probability $q_i < 1$, we have that

$$\frac{|[\mathbf{SA}\mathbf{x}](i)|^p}{\|\mathbf{A}''\mathbf{x}\|_p^p} \leq \frac{|[\mathbf{SA}\mathbf{x}](i)|^p}{\|\mathbf{A}'\mathbf{x}\|_p^p} = \frac{|[\mathbf{SA}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \leq \frac{1}{q_i} \frac{|[\mathbf{A}\mathbf{x}](i)|^p}{\|\mathbf{A}\mathbf{x}\|_p^p} \leq \frac{\boldsymbol{\sigma}_i^p(\mathbf{A})}{q_i} \leq O(\alpha).$$

By similar reasoning, we have that

$$\frac{|[\mathbf{Sb}](i)|^p}{R^p} \leq \frac{1}{q_i} \frac{|\mathbf{b}(i)|^p}{R^p} \leq O(\alpha).$$

□

The next lemma shows that in order to bound the Rademacher process in Lemma 8.1.3, it suffices to bound a similar Rademacher process for $\mathbf{A}''$ and $\mathbf{b}''$.

Lemma 8.1.6 (Reduction to flattened instance)**.** Let $\mathbf{S}$, $\mathbf{A}, \mathbf{b}$, and $\mathbf{A}'', \mathbf{b}''$ be as given in Lemma 8.1.5. Let δ, ε, l be such that $\delta\varepsilon^l \leq 1/2$. Furthermore, suppose that

$$\mathop{\mathbf{E}}_{\mathbf{S}} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm 1\}^n} \sup_{\|\mathbf{A}''\mathbf{x}\|_p \leq R} \left| \sum_{i\in T} \boldsymbol{\varepsilon}_i |[\mathbf{A}''\mathbf{x} + \mathbf{b}](i)|^p \right|^l \leq \delta\varepsilon^l (R^p)^l$$

for every $R \geq \|\mathbf{b}\|_p$. Then,

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{A}\mathbf{x}\|_p \leq R} \left| \|\mathbf{S}(\mathbf{A}\mathbf{x} + \mathbf{b})\|_p^p - \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_p^p \right|^l \leq 2\delta(2^{3p}\varepsilon R^p)^l$$

for every $R \geq \|\mathbf{b}\|_p$.

Proof. Fix an outcome of $\mathbf{S}$ and let

$$F_{\mathbf{S},R} = \sup_{\|\mathbf{A}\mathbf{x}\|_p \leq R} \left| \|\mathbf{S}(\mathbf{A}\mathbf{x} + \mathbf{b})\|_p^p - \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_p^p \right|.$$

Note then that for any $\|\mathbf{A}\mathbf{x}\|_p \leq R$, we have

$$\begin{aligned}
\|\mathbf{A}''\mathbf{x}\|_p &\leq \|\mathbf{A}\mathbf{x}\|_p + \|\mathbf{S}(\mathbf{A}\mathbf{x} + \mathbf{b})\|_p + \|\mathbf{S}\mathbf{b}\|_p \\
&\leq R + \left(\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_p^p + \left| \|\mathbf{S}(\mathbf{A}\mathbf{x} + \mathbf{b})\|_p^p - \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_p^p \right| \right)^{1/p} + \left(\|\mathbf{b}\|_p^p + \left| \|\mathbf{S}\mathbf{b}\|_p^p - \|\mathbf{b}\|_p^p \right| \right)^{1/p} \\
&\leq R + \left(\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_p^p + F_{\mathbf{S},R} \right)^{1/p} + \left(\|\mathbf{b}\|_p^p + F_{\mathbf{S},R} \right)^{1/p} \\
&\leq R + \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_p + F_{\mathbf{S},R}^{1/p} + \|\mathbf{b}\|_p + F_{\mathbf{S},R}^{1/p} \\
&\leq 4R + 2F_{\mathbf{S},R}^{1/p}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{S}} F_{\mathbf{S},R}^l &\leq \mathop{\mathbf{E}}_{\mathbf{S}} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm 1\}^n} \sup_{\|\mathbf{A}\mathbf{x}\|_p \leq R} \left| \sum_{i\in T} \boldsymbol{\varepsilon}_i |[\mathbf{S}(\mathbf{A}\mathbf{x} + \mathbf{b})](i)|^p \right|^l && \text{Lemma 8.1.3} \\
&\leq \mathop{\mathbf{E}}_{\mathbf{S}} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm 1\}^n} \sup_{\|\mathbf{A}\mathbf{x}\|_p \leq R} \left| \sum_{i} \boldsymbol{\varepsilon}_i |[\mathbf{A}''\mathbf{x} + \mathbf{b}''](i)|^p \right|^l \\
&\leq \mathop{\mathbf{E}}_{\mathbf{S}} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm 1\}^n} \sup_{\|\mathbf{A}''\mathbf{x}\|_p \leq 4R + 2F_{\mathbf{S},R}^{1/p}} \left| \sum_{i} \boldsymbol{\varepsilon}_i |[\mathbf{A}''\mathbf{x} + \mathbf{b}''](i)|^p \right|^l \\
&\leq \mathop{\mathbf{E}}_{\mathbf{S}} \delta\varepsilon^l ((4R + 2F_{\mathbf{S},R}^{1/p})^p)^l && \text{by hypothesis} \\
&\leq \mathop{\mathbf{E}}_{\mathbf{S}} \delta(2^{2p}\varepsilon)^l (((2R)^p)^l + F_{\mathbf{S},R}^l)
\end{aligned}$$

$$= \delta(2^{2p}\varepsilon)^l \left[((2R)^p)^l + \mathop{\mathbf{E}}_{\mathbf{S}} F_{\mathbf{S},R}^l \right]$$

so rearranging gives

$$\frac{\mathbf{E}_{\mathbf{S}} F_{\mathbf{S},R}^l}{(2^p R^p)^l + \mathbf{E}_{\mathbf{S}} F_{\mathbf{S},R}^l} \leq \delta(2^{2p}\varepsilon)^l.$$

In turn, this implies that

$$\mathop{\mathbf{E}}_{\mathbf{S}} F_{\mathbf{S},R}^l \leq \frac{\delta(2^{2p}\varepsilon)^l (2^p R^p)^l}{1 - \delta(2^{2p}\varepsilon)^l} \leq 2\delta(2^{3p}\varepsilon R^p)^l. \qquad \square$$

8.1.3 Bounds on the Rademacher process

In this section, we present results from [WY23c] (which in turn are based on [BLM89, LT91]) which will allow us to bound a Rademacher process of the form of Lemma 8.1.6.

The following is a straightforward generalization of Lemma 6.4.1.

Lemma 8.1.7. Let $1 \leq p < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Let $R \geq \|\mathbf{b}\|_p$. Define the pseudo-metric

$$d_X(\mathbf{y}, \mathbf{y}') := \left(\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon} \sim \{\pm 1\}^n} \left| \sum_{i=1}^n \boldsymbol{\varepsilon}_i |\mathbf{y}(i)|^p - \sum_{i=1}^n \boldsymbol{\varepsilon}_i |\mathbf{y}'(i)|^p \right|^2 \right)^{1/2}$$

Let $\sigma \geq \max_{i \in S}^n \boldsymbol{\sigma}_i^p(\mathbf{A}) + |\mathbf{b}(i)|^p / R^p$. Then, for any $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ and $\mathbf{y}' = \mathbf{A}\mathbf{x}' + \mathbf{b}$ with $\|\mathbf{A}\mathbf{x}\|_p, \|\mathbf{A}\mathbf{x}'\|_p \leq R$,

$$d_X(\mathbf{y}, \mathbf{y}') \leq \begin{cases} O(1) \|\mathbf{A}(\mathbf{x} - \mathbf{x}')\|_\infty^{p/2} R^{p/2} & p < 2 \\ O(1) \sigma^{1/2 - 1/p} \cdot \|\mathbf{A}(\mathbf{x} - \mathbf{x}')\|_\infty R^{p-1} & p > 2 \end{cases}$$

With Lemma 8.1.7 in hand, we show the following.

Theorem 8.1.8. Let $1 \leq p < \infty$ be fixed and let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Let $R \geq \|\mathbf{b}\|_p$. Let $\tau \geq \max_{i=1}^n \boldsymbol{\tau}_i(\mathbf{A})$ and let $\sigma \geq \max_{i=1}^n \boldsymbol{\sigma}_i^p(\mathbf{A})$. Define

$$\mathcal{E} := \begin{cases} \tau^{1/2} \cdot (\log n)^{3/2} & p < 2 \\ \tau^{1/2} (\sigma n)^{1/2 - 1/p} \cdot (\log n)^{3/2} & p > 2 \end{cases}.$$

Then,

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon} \sim \{\pm 1\}^n} \sup_{\|\mathbf{A}\mathbf{x}\|_p \leq R} \left| \sum_{i=1}^n \boldsymbol{\varepsilon}_i |[\mathbf{A}\mathbf{x} + \mathbf{b}](i)|^p \right|^l \leq \left[(2\mathcal{E})^l (\mathcal{E}/\sigma^{1/2}) + O(\sqrt{l}\sigma^{1/2})^l \right] (R^p)^l$$

Proof. Let $T = \{\mathbf{A}\mathbf{x} : \|\mathbf{A}\mathbf{x}\|_p \leq 1\}$. We have that

$$\int_0^\infty \sqrt{\log E(T, d_X, u)}\, du = O(\mathcal{E}) R^p$$

by trivial modifications of Lemmas 7.4.1 and 7.4.2 in Chapter 7. We also have the diameter bound

$$\operatorname{diam}(T) \le O(\sigma^{1/2})R^p$$

by trivial modifications of Lemma 6.4.2. We may then conclude by the moment bounds of Lemma 2.3.7 that are obtained by integrating Dudley's tail bound (Theorem 2.3.6). □

8.1.4 Proof of main sampling theorems

We now prove Theorem 8.1.1 by combining the previous results of this section.

Proof of Theorem 8.1.1. By Lemma 8.1.2, it suffices to show that

$$\sup_{\|\mathbf{Ax}\|_p \le R_i} \left| \|\mathbf{S}(\mathbf{Ax}+\mathbf{b})\|_p^p - \|\mathbf{Ax}+\mathbf{b}\|_p^p \right| \le \varepsilon R_i^p \tag{8.1}$$

for $R_i = 2^i R$ for $i \in [I]$, $I = O(\log \varepsilon^{-1})$. The corresponding statement for bounding

$$\sup_{\|\mathbf{Ax}\|_p = 1} \left| \|\mathbf{SAx}\|_p^p - \|\mathbf{Ax}\|_p^p \right| \le \varepsilon$$

will follow from the exact same analysis by setting $\mathbf{b} = 0$ and $R = 0$.

In order to obtain (8.1) for a single scale with high probability, we will bound the lth moment for a large even power l. We will bound this quantity by passing to a Rademacher process bound in Lemma 8.1.6, which we can in turn bound using Theorem 8.1.8. Note that by our construction of the flattened instance $\mathbf{A}''$ and $\mathbf{b}''$ in Lemmas 8.1.5 and 8.1.6, we have $\tau = O(\alpha)$ and $\sigma = O(\alpha)$ for $p < 2$ and $\tau = O(\alpha)/n^{1-2/p}$ and $\sigma = O(\alpha)$ for $p > 2$, so we can bound the $\mathcal{E}$ parameter in Theorem 8.1.8 by

$$\mathcal{E} \le \begin{cases} O(\alpha^{1-1/p})(\log n)^{3/2} & p > 2 \\ O(\alpha^{1/2})(\log n)^{3/2} & p < 2 \end{cases}$$

In turn, the bound on the Rademacher process in Theorem 8.1.8 is $\delta\varepsilon^l (R_i^p)^l/(I+1)$ by our choice of α, for

$$l = O\left(\log\log n + \log\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right).$$

That is, we have shown that

$$\mathop{\mathbf{E}}_{\mathbf{S}} \sup_{\|\mathbf{Ax}\|_p \le R_i} \left| \|\mathbf{S}(\mathbf{Ax}+\mathbf{b})\|_p^p - \|\mathbf{Ax}+\mathbf{b}\|_p^p \right|^l \le \frac{\delta}{I+1}\varepsilon^l (R_i^p)^l.$$

Then by Markov's inequality, we have that

$$\Pr_{\mathbf{S}}\left\{ \sup_{\|\mathbf{Ax}\|_p \le R_i} \left| \|\mathbf{S}(\mathbf{Ax}+\mathbf{b})\|_p^p - \|\mathbf{Ax}+\mathbf{b}\|_p^p \right| \le \varepsilon R_i^p \right\} \ge 1 - \frac{\delta}{I+1}.$$

Now by a union bound, this is simultaneously true for all $i \in [I]$ as well as for $\mathbf{b} = 0$ and $R = 0$ by a union bound, all with probability at least $1 - \delta$. In turn, we have that

$$\|\mathbf{S}(\mathbf{A}\mathbf{x} + \mathbf{b})\|_p^p = (1 \pm 2 \cdot 4^p \varepsilon)\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_p^p \pm 2^p \varepsilon R_0^p$$

by Lemma 8.1.2. We have the desired conclusion by rescaling ε and δ up to constant factors.

Finally, to analyze the recursive application of this theorem, we use the following elementary recurrence:

Lemma 8.1.9. Suppose $(a_i)_{i=0}^{\infty}$ satisfies the recurrence $a_{i+1} = \lambda a_i + b$ for some $b > 0$ and $\lambda \in (0, 1)$. Then,

$$a_i = \frac{1}{1-\lambda}\big(b - \lambda^i(b - (1-\lambda)a_0)\big).$$

Proof. Let x satisfy $x = \lambda x + b$, that is, $x = b/(1-\lambda)$. Then, the sequence $a'_{i+1} = a_i - x$ satisfies $a'_{i+1} = \lambda a'_i$ so $a'_i = \lambda^i a'_0$. Thus, $a_i = a'_i + x = \lambda^i(a_0 - x) + x$. □

We apply the above result with failure probability δ/R and accuracy $\varepsilon/$ recursively for at most $R = \Theta(\log \log n)$ rounds, until the number of rows is at most the claimed bound. By a union bound, we succeed at achieving ε/R sampling error and row count bound on all R rounds, that is, for any number of rows m_i on the ith round, we reduce the number of rows to at most

$$\begin{cases} O(\varepsilon^{-2} m_i^{1-2/p} d\big((\log n)^3 + \log \frac{1}{\delta}\big) & p > 2 \\ O(\varepsilon^{-2} m_i^{1-p/2} d^{p/2}\big((\log n)^3 + \log \frac{1}{\delta}\big) & p < 2 \end{cases}$$

rows. We then apply the recurrence lemma (Lemma 8.1.9) on the logarithm of the above bound with $a_i = m_i$, $\lambda = (1 - p/2)$ for $p < 2$ and $\lambda = (1 - 2/p)$ for $p > 2$ and

$$b = \begin{cases} \log O(\varepsilon^{-2} d\big((\log n)^3 + \log \frac{1}{\delta}\big) & p > 2 \\ \log O(\varepsilon^{-2} d^{p/2}\big((\log n)^3 + \log \frac{1}{\delta}\big) & p < 2 \end{cases}.$$

This gives a bound of

$$a_i = \log m_i \le \frac{b+1}{1-\lambda},$$

that is, a final row count of at most

$$\begin{cases} O(\varepsilon^{-p} d^{p/2}\big((\log n)^3 + \log \frac{1}{\delta}\big)^{p/2} & p > 2 \\ O(\varepsilon^{-4/p} d\big((\log n)^3 + \log \frac{1}{\delta}\big)^{2/p} & p < 2 \end{cases}.$$

□

Chapter 9

High-distortion ℓ_p subspace embeddings [WY22a]

Until now, we have focused on subspace embeddings which achieve a distortion of $(1+\varepsilon)$. However, in certain applications, such a high accuracy may not be necessary, and a natural question is whether the number of rows r of the sketch $\mathbf{S}$ can be improved or not if larger errors are allowed. Note that $(1+\varepsilon)$ distortion is essentially the end of the story of $0 < p \leq 2$, as the upper bounds obtained by ℓ_p Lewis weight sampling (Theorem 6.1.9) already achieve a bound of $\tilde{O}(\varepsilon^{-2}d)$, and it is easy to see that at least d rows is needed for any subspace embedding, even just to maintain the rank. On the other hand, for $p > 2$, one could still ask for more, since if we require $\Theta(1)$ distortion, then the number of rows necessary is $r = \Omega(d^{p/2})$ [LWW21], whose exponential dependence on p may be prohibitive for large p. In the work of [WY22a], we study the following question:

Question 9.0.1. For $p > 2$, what trade-offs between the number of rows r and the distortion κ are possible in the regime where $\kappa \gg 1$?

In [WY22a], we provide a nearly optimal trade-off between r and κ as a solution to Question 9.0.1.

Theorem 9.0.2 (High-distortion ℓ_p Lewis weight sampling [WY22a])**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $2 < p < \infty$. Then, for any $2 < q < p$, there is a randomized algorithm for constructing a diagonal map $\mathbf{S} \in \mathbb{R}^{n\times n}$ such that

$$\mathbf{Pr}\Big\{\text{for all } \mathbf{x}\in\mathbb{R}^d, \quad \|\mathbf{Ax}\|_p \leq \|\mathbf{SAx}\|_q \leq O(d^{\frac{1}{2}(1-\frac{q}{p})})\|\mathbf{Ax}\|_p\Big\} \geq \frac{99}{100}$$

and furthermore, $\mathbf{S}$ has at most r nonzero rows, for $r = O(d^{q/2}(\log d)^3)$. Furthermore, any randomized algorithm which constructs a data structure $\mathcal{Q}$ such that

$$\mathbf{Pr}\Big\{\text{for all } \mathbf{x}\in\mathbb{R}^d, \quad \|\mathbf{Ax}\|_p \leq \mathcal{Q}(\mathbf{x}) \leq o(d^{\frac{1}{2}(1-\frac{q}{p})})\|\mathbf{Ax}\|_p\Big\} \geq \frac{99}{100}$$

requires $\Omega(d^{q/2+1})$ bits of space.

We note that the lower bound is shown in [LWW21]. Our proof of the upper bound in Theorem 9.0.2 proceeds in two steps: (1) we first show that we can approximate $\|\mathbf{Ax}\|_p$ by $\|\mathbf{W}^{\frac{1}{q}-\frac{1}{p}}\mathbf{Ax}\|_q$

for some diagonal reweighting map $\mathbf{W}$ up to a factor of $d^{\frac{1}{2}\left(1-\frac{q}{p}\right)}$, and (2) we use ℓ_q Lewis weight sampling to reduce the number of rows to $\tilde{O}(d^{q/2})$ while preserving the distortion up to $\Theta(1)$ factors. Step (2) is simply using Theorem 6.1.11, so the key ingredient here is step (1).

Perhaps surprisingly, we show that step (1) can in fact also be implemented using ℓ_p Lewis weights, and the reweighting map $\mathbf{W}$ can be simply be taken to be the ℓ_p Lewis weights. More specifically, we show the following theorem:

Theorem 9.0.3 (ℓ_p Lewis weight change of density [WY22a])**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $0 < q < p < \infty$. Let $\mathbf{W} = \operatorname{diag}(\mathbf{w}^p(\mathbf{A}))$ be the diagonal map given by the ℓ_p Lewis weights of $\mathbf{A}$. Then, there is a scaling factor c such that for every $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{A}\mathbf{x}\|_p \le c\|\mathbf{W}^{\frac{1}{q}-\frac{1}{p}}\mathbf{A}\mathbf{x}\|_q \le \kappa\|\mathbf{A}\mathbf{x}\|_p$$

for

$$\kappa = \begin{cases} d^{\frac{1}{q}-\frac{1}{p}} & \min(p,q) \le 2 \\ d^{\frac{1}{2}\left(1-\frac{q}{p}\right)} & \min(p,q) \ge 2 \end{cases}$$

In fact, the result of Theorem 9.0.3 provides an elementary proof of a result of [LT80] from the geometric functional analysis literature, who proved the existence of a diagonal map satisfying the guarantees of Theorem 9.0.3 by using sophisticated results from the theory of factorization of operators, p-summing norms, and operator ideals. On the other hand, our proof of Theorem 9.0.3 only requires elementary inequalities and ℓ_p Lewis weights. One of the key insights we use is that if $\mathbf{W}$ are the ℓ_p Lewis weights, then the $\mathbf{W}$ is also the ℓ_q Lewis weights of the matrix $\mathbf{W}^{\frac{1}{q}-\frac{1}{p}}\mathbf{A}$. We will now develop this idea in the next sections.

9.1 Lewis weight switching

We first show the following crucial identity, which shows that one can reweight a matrix by ℓ_p Lewis weights, so that the ℓ_q Lewis weights of the resulting matrix coincides with the ℓ_p Lewis weights of the original matrix. Note that this identity is true by definition for $q = 2$ (Definition 6.1.6), since ℓ_2 Lewis weights are just leverage scores. Thus, this result shows that although Lewis weights are defined by normalizing a change of density with respect to ℓ_2 (see [CP15]), they actually simultaneously satisfy the analogous property for all ℓ_q as well.

Lemma 9.1.1 (Lewis weight switching)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $p, q > 0$. Let $\mathbf{B} := \mathbf{W}^p(\mathbf{A})^{1/q-1/p}\mathbf{A}$. Then, for each $i \in [n]$,

$$\mathbf{w}_i^q(\mathbf{B}) = \mathbf{w}_i^p(\mathbf{A}).$$

Furthermore, the two Lewis bases coincide.

Proof. We have that

$$\begin{aligned}\boldsymbol{\tau}_i(\mathbf{W}^p(\mathbf{A})^{1/2-1/q}\mathbf{B}) &= \boldsymbol{\tau}_i(\mathbf{W}^p(\mathbf{A})^{1/2-1/q} \cdot \mathbf{W}^p(\mathbf{A})^{1/q-1/p}\mathbf{A}) \\ &= \boldsymbol{\tau}_i(\mathbf{W}^p(\mathbf{A})^{1/2-1/p}\mathbf{A}) = \mathbf{w}_i^p(\mathbf{A})\end{aligned}$$

so

$$\mathbf{w}_i^q(\mathbf{W}^p(\mathbf{A})^{1/q-1/p}\mathbf{A}) = \mathbf{w}_i^p(\mathbf{A})$$

by uniqueness of Lewis weights [CP15]. The Lewis bases coincide since

$$\mathbf{W}^p(\mathbf{A})^{1/2-1/p}\mathbf{A}\mathbf{R} = \mathbf{W}^p(\mathbf{A})^{1/2-1/q}\mathbf{W}^p(\mathbf{A})^{1/q-1/p}\mathbf{A}\mathbf{R} = \mathbf{W}^p(\mathbf{A})^{1/2-1/q}\mathbf{B}\mathbf{R}.$$

□

In fact, given only one-sided ℓ_p Lewis weights (Definition 6.1.7), we can prove a similar inequality:

Lemma 9.1.2 (One-sided Lewis weight switching)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $p, q > 0$. Let $\mathbf{w} \in \mathbb{R}^n$ be one-sided ℓ_p Lewis weights for $\mathbf{A}$ and let $\mathbf{R}$ be the corresponding one-sided ℓ_p Lewis basis. Let $\mathbf{W} = \operatorname{diag}(\mathbf{w})$ and $\mathbf{B} \coloneqq \mathbf{W}^{1/q-1/p}\mathbf{A}$. Then, $\mathbf{w}$ are one-sided ℓ_q Lewis weights and $\mathbf{R}$ is a one-sided ℓ_q Lewis basis for $\mathbf{B}$, i.e.,

$$\boldsymbol{\tau}_i(\mathbf{W}^{1/2-1/q}\mathbf{B}) \le \mathbf{w}_i.$$

Proof. We have that $\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}$ is orthonormal, which means

$$\mathbf{W}^{1/2-1/q}\mathbf{B}\mathbf{R} = \mathbf{W}^{1/2-1/q}\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{R} = \mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}$$

is as well. Then,

$$\boldsymbol{\tau}_i(\mathbf{W}^{1/2-1/q}\mathbf{B}) = \left\|\mathbf{e}_i^\top\mathbf{W}^{1/2-1/q}\mathbf{B}\mathbf{R}\right\|_2^2 = \left\|\mathbf{e}_i^\top\mathbf{W}^{1/2-1/p}\mathbf{A}\mathbf{R}\right\|_2^2 = \boldsymbol{\tau}_i(\mathbf{W}^{1/2-1/p}\mathbf{A}) \le \mathbf{w}_i$$

as desired. □

9.2 Change of density

Using the above, we show that reweighting the rows of $\mathbf{A}$ by a scalar multiple of the ℓ_p Lewis weights provide optimal approximations of ℓ_p by ℓ_q. The following lemmas show the upper bounds and lower bounds. The proofs roughly follow, but are still slightly different from, the estimates in Lemma 2.6 of [JLS22] and Lemma 8 in Chapter III.B of [Woj91], which show the analogous results for $q = 2$. The estimates are an elementary combination of Lewis weight switching (Lemma 9.1.1/Lemma 9.1.2), sensitivity bounds (Lemma 6.2.4), and Hölder's inequality.

Lemma 9.2.1 (Upper bound, $p \ge q$)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \ge q > 0$. Let $\mathbf{w} \in \mathbb{R}^n$ be one-sided ℓ_p Lewis weights for $\mathbf{A}$. Let $\mathbf{W} = \operatorname{diag}(\mathbf{w})$. For all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{A}\mathbf{x}\|_p \le \|\mathbf{w}\|_1^{[0\vee(1/2-1/q)](p-q)/p}\left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q$$

Proof. For $i \in [n]$, we have that

$$\begin{aligned}|[\mathbf{A}\mathbf{x}](i)| &= \mathbf{w}_i^{1/p-1/q}\cdot[\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}](i)\\ &\le \mathbf{w}_i^{1/p-1/q}\cdot\left[\|\mathbf{w}\|_1^{0\vee(q/2-1)}\cdot\mathbf{w}_i\cdot\left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q^q\right]^{1/q} \qquad \text{Lemmas 6.2.4, 9.1.2}\end{aligned}$$

$$= \|\mathbf{w}\|_1^{0\vee(1/2-1/q)} \cdot \mathbf{w}_i^{1/p} \left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q.$$

Then,

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}\|_p^p &= \sum_{i=1}^n |[\mathbf{A}\mathbf{x}](i)|^p = \sum_{i=1}^n |[\mathbf{A}\mathbf{x}](i)|^{p-q} \cdot |[\mathbf{A}\mathbf{x}](i)|^q \\
&\le \sum_{i=1}^n \|\mathbf{w}\|_1^{[0\vee(1/2-1/q)](p-q)} \cdot \mathbf{w}_i^{(p-q)/p} \left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q^{p-q} \cdot |[\mathbf{A}\mathbf{x}](i)|^q \\
&= \|\mathbf{w}\|_1^{[0\vee(1/2-1/q)](p-q)} \left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q^{p-q} \sum_{i=1}^n \mathbf{w}_i^{q(1/q-1/p)} \cdot |[\mathbf{A}\mathbf{x}](i)|^q \\
&= \|\mathbf{w}\|_1^{[0\vee(1/2-1/q)](p-q)} \left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q^{p-q} \cdot \left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q^q \\
&= \|\mathbf{w}\|_1^{[0\vee(1/2-1/q)](p-q)} \left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q^p.
\end{aligned}$$

Taking pth roots on both sides gives the desired result. □

Lemma 9.2.2 (Upper bound, $q \ge p$)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $q \ge p > 0$. Let $\kappa \ge 1$ and let $\mathbf{w} \in \mathbb{R}^n$ be α-approximate ℓ_p Lewis weights for $\mathbf{A}$. Let $\mathbf{W} = \operatorname{diag}(\mathbf{w})$. For all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{A}\mathbf{x}\|_p \le \|\mathbf{w}\|_1^{1/p-1/q} \left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q$$

Proof. We have

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}\|_p^p &= \sum_{i=1}^n \mathbf{w}_i^{1-p/q} \cdot [\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}](i)^p \\
&= \left[\sum_{i=1}^n \mathbf{w}_i^{(1-p/q)/(1-p/q)}\right]^{1-p/q} \left[\sum_{i=1}^n [\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}](i)^q\right]^{p/q} \qquad \text{Hölder's inequality} \\
&= \|\mathbf{w}\|_1^{1-p/q} \left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q^p.
\end{aligned}$$

Taking pth roots on both sides gives the desired result. □

Lemma 9.2.3 (Lower bound, $p \ge q$)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $p \ge q > 0$. Let $\kappa \ge 1$ and let $\mathbf{w} \in \mathbb{R}^n$ be α-approximate ℓ_p Lewis weights for $\mathbf{A}$. Let $\mathbf{W} = \operatorname{diag}(\mathbf{w})$. For all $\mathbf{x} \in \mathbb{R}^d$,

$$\left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q \le \|\mathbf{w}\|_1^{1/q-1/p} \|\mathbf{A}\mathbf{x}\|_p$$

Proof. We have

$$\begin{aligned}
\left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q^q &= \sum_{i=1}^n \mathbf{w}_i^{1-q/p} [\mathbf{A}\mathbf{x}](i)^q \\
&\le \left[\sum_{i=1}^n \mathbf{w}_i^{(1-q/p)/(1-q/p)}\right]^{1-q/p} \left[\sum_{i=1}^n [\mathbf{A}\mathbf{x}](i)^p\right]^{q/p} \qquad \text{Hölder's inequality} \\
&\le \|\mathbf{w}\|_1^{1-q/p} \|\mathbf{A}\mathbf{x}\|_p^q.
\end{aligned}$$

Taking qth roots on both sides gives the desired result. □

Lemma 9.2.4 (Lower bound, $q \geq p$)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $q \geq p > 0$. Let $\mathbf{w} \in \mathbb{R}^n$ be one-sided ℓ_p Lewis weights for $\mathbf{A}$. Let $\mathbf{W} = \operatorname{diag}(\mathbf{w})$. For all $\mathbf{x} \in \mathbb{R}^d$,

$$\left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q \leq \|\mathbf{w}\|_1^{[0\vee(1/2-1/p)](q-p)/q}\|\mathbf{A}\mathbf{x}\|_p$$

Proof. We have

$$\begin{aligned}
\left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q^q &= \sum_{i=1}^n \mathbf{w}_i^{1-q/p}[\mathbf{A}\mathbf{x}](i)^q = \sum_{i=1}^n \mathbf{w}_i^{1-q/p}[\mathbf{A}\mathbf{x}](i)^{q-p}[\mathbf{A}\mathbf{x}](i)^p \\
&\leq \sum_{i=1}^n \mathbf{w}_i^{1-q/p}\left[\|\mathbf{w}\|_1^{0\vee(p/2-1)}\mathbf{v}_i\|\mathbf{A}\mathbf{x}\|_p^p\right]^{(q-p)/p}[\mathbf{A}\mathbf{x}](i)^p \qquad \text{Lemma 6.2.4} \\
&= \|\mathbf{A}\mathbf{x}\|_p^{q-p}\|\mathbf{w}\|_1^{[0\vee(p/2-1)](q-p)/p}\sum_{i=1}^n [\mathbf{A}\mathbf{x}](i)^p \\
&= \|\mathbf{w}\|_1^{[0\vee(p/2-1)](q-p)/p}\|\mathbf{A}\mathbf{x}\|_p^q.
\end{aligned}$$

Taking qth roots on both sides gives the desired result. □

Combining the above lemmas yields the following conclusion.

Theorem 9.2.5 (Change of density via approximate Lewis weights)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $0 < p, q < \infty$. Let $\mathbf{w} \in \mathbb{R}^n$ be one-sided ℓ_p Lewis weight (Definition 6.1.7) and $\mathbf{W} = \operatorname{diag}(\mathbf{w})$. For $p \geq q$, let

$$\begin{aligned}
\kappa_{d,p,q} &\coloneqq \|\mathbf{w}\|_1^{1/q-1/p} \\
\lambda_{d,p,q} &\coloneqq \|\mathbf{w}\|_1^{[0\vee(1/2-1/q)](p-q)/p}
\end{aligned}$$

and let $\kappa_{d,p,q} \coloneqq \kappa_{d,q,p}^{-1}$, $\lambda_{d,p,q} \coloneqq \lambda_{d,q,p}^{-1}$ if $q \leq p$. Then for all $\mathbf{x} \in \mathbb{R}^d$ we have the following:

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}\|_p \leq \left\|\lambda_{d,p,q} \cdot \mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q \leq \kappa_{d,p,q}\lambda_{d,p,q}\|\mathbf{A}\mathbf{x}\|_p \qquad \text{if } p \geq q \\
\|\mathbf{A}\mathbf{x}\|_p \leq \left\|\kappa_{d,p,q} \cdot \mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q \leq \kappa_{d,p,q}\lambda_{d,p,q}\|\mathbf{A}\mathbf{x}\|_p \qquad \text{if } q \geq p
\end{aligned}$$

Note that

$$\kappa_{d,p,q}\lambda_{d,p,q} = \begin{cases} \|\mathbf{w}\|_1^{\left|\frac{1}{q}-\frac{1}{p}\right|} & \text{if } \min(p,q) \leq 2 \\ \|\mathbf{w}\|_1^{\frac{1}{2}\left(1-\frac{p\wedge q}{p\vee q}\right)} & \text{if } \min(p,q) \geq 2 \end{cases}$$

Proof. Let $\mathbf{v}$ be the one-sided ℓ_p Lewis weights such that $\mathbf{w} \geq \mathbf{v}$. First consider the case of $p \geq q > 0$. Note that in this case, $1/q - 1/p \geq 0$ so Lemmas 9.2.1 and 9.2.3 yield that

$$\|\mathbf{A}\mathbf{x}\|_p \leq \lambda_{d,p,q}\left\|\mathbf{V}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q \leq \lambda_{d,p,q}\left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q \leq \kappa_{d,p,q}\lambda_{d,p,q}\|\mathbf{A}\mathbf{x}\|_p.$$

On the other hand, if $q \geq p > 0$, then $1/q - 1/p \leq 0$ so Lemmas 9.2.2 and 9.2.4 yield that

$$\|\mathbf{A}\mathbf{x}\|_p \leq \kappa_{d,p,q}\left\|\mathbf{W}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q \leq \kappa_{d,p,q}\left\|\mathbf{V}^{1/q-1/p}\mathbf{A}\mathbf{x}\right\|_q \leq \kappa_{d,p,q}\lambda_{d,p,q}\|\mathbf{A}\mathbf{x}\|_p.$$

For $p \geq q$, we have that the total distortion is $\|\mathbf{w}\|_1^\beta$, for

$$\beta = \left[\frac{1}{2} - \frac{1}{q}\right]\frac{p-q}{p} + \left[\frac{1}{q} - \frac{1}{p}\right] = \left[\frac{1}{2} - \frac{1}{q}\right]\frac{p-q}{p} + \frac{1}{q}\frac{p-q}{p} = \frac{p-q}{2p} = \frac{1}{2}\left(1 - \frac{q}{p}\right)$$

if $q \geq 2$, and $\|\mathbf{w}\|_1^{\frac{1}{q}-\frac{1}{p}}$ if $q \leq 2$. Next, when $q \geq p > 0$, then we have that the total distortion is $\|\mathbf{w}\|_1^\beta$ for

$$\beta = \left[\frac{1}{2} - \frac{1}{p}\right]\frac{q-p}{q} + \left[\frac{1}{p} - \frac{1}{q}\right] = \left[\frac{1}{2} - \frac{1}{p}\right]\frac{q-p}{q} + \frac{1}{p}\frac{q-p}{q} = \frac{q-p}{2q} = \frac{1}{2}\left(1 - \frac{p}{q}\right)$$

if $p \geq 2$, and $\|\mathbf{w}\|_1^{\frac{1}{p}-\frac{1}{q}}$ if $p \leq 2$. These yield the claimed bounds.

□

Chapter 10

Subspace embeddings for general losses [MMWY22]

Up until now, we have studied subspace embeddings for the ℓ_p loss, with applications to ℓ_p regression in mind. In fact, the problem of computing subspace embeddings makes sense in a far more generalized setting, where we wish to approximate loss functions of the form

$$\|\mathbf{A}\mathbf{x}\|_{g,\mathbf{w}} \coloneqq \sum_{i=1}^{n} \mathbf{w}_i \cdot g([\mathbf{A}\mathbf{x}](i)), \tag{10.1}$$

where we denote the loss function as a norm in an abuse of notation, despite the fact that $\|\cdot\|_{g,\mathbf{w}}$ may not be a norm. For example, taking the weights $\mathbf{w}_i$ to be all ones and g to be the so-called *Huber loss* H defined as

$$H(x) \coloneqq \begin{cases} x^2/2 & |x| \leq 1 \\ |x| - 1/2 & |x| \geq 1 \end{cases}$$

is useful in solving linear regression with the Huber loss, which is a popular loss function in the literature of robust statistics [CW15a]. Similarly, taking g to be the *Tukey loss* T defined as

$$T(x) \coloneqq \begin{cases} 1 - (1 - x^2)^3 & |x| \leq 1 \\ 1 & |x| \geq 1 \end{cases}$$

is another popular choice for robust regression [CWW19]. Yet another example is to take g to be the *logistic loss*, given by

$$g(x) \coloneqq \log(1 + e^x)$$

which corresponds to logistic regression [MSSW18, MMR21].

Improved sensitivity bounds for general loss functions. In fact, we have already discussed a generalized approach to estimating functions of the form of (10.1) in Chapter 7, via *sensitivity sampling*. Recall that in this framework, we wish to compute upper bounds on the sensitivity scores $\boldsymbol{\sigma}_i$, which in this case are given by

$$\boldsymbol{\sigma}_i(\mathbf{A}) \coloneqq \sup_{\mathbf{A}\mathbf{x} \neq 0} \frac{\mathbf{w}_i \cdot g([\mathbf{A}\mathbf{x}](i))}{\sum_{j=1}^{n} \mathbf{w}_j \cdot g([\mathbf{A}\mathbf{x}](j))}.$$

Given upper bounds $\tilde{\boldsymbol{\sigma}}_i \geq \boldsymbol{\sigma}_i(\mathbf{A})$ on the sensitivity scores, we almost immediately obtain a sampling algorithm which samples at most $\tilde{O}(\varepsilon^{-2}\tilde{\mathfrak{S}}d)$ rows of $\mathbf{A}$, where $\tilde{\mathfrak{S}} = \sum_{i=1}^{n} \tilde{\boldsymbol{\sigma}}_i$. The primary difficult in this approach is efficiently obtaining the sensitivity upper bounds $\tilde{\boldsymbol{\sigma}}_i$. Previously, an approach based on ellipsoidal rounding of the balls induced by the norm $\|\mathbf{A}\mathbf{x}\|_{g,\mathbf{w}}$ has been proposed by [TMF20]. However, computing Löwner–John ellipsoids for general convex bodies is computational expensive, and furthermore, leads to $\mathrm{poly}(d)$ factor losses in the total sensitivity upper bound $\tilde{\mathfrak{S}}$ and thus in the sample complexity.

In the work of [MMWY22], we obtain a significantly improved algorithm for estimating sensitivity scores, which is nearly optimal for a wide class of loss functions.

Theorem 10.0.1 (Sensitivity upper bounds for general loss functions, Theorem 4.9, [MMWY22])**.** Let $M : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be increasing, has $M(0) = 0$, and has at most quadratic growth, that is,

$$\frac{M(y)}{M(x)} \leq c\left(\frac{y}{x}\right)^2$$

for all $y > x$. Let $g(x) \coloneqq M(|x|)$. Then, there is an algorithm that computes upper bounds $\tilde{\boldsymbol{\sigma}}_i$ to the sensitivities with respect to g such that $\tilde{\mathfrak{S}} = \sum_{i=1}^{n} \boldsymbol{\sigma}_i \leq O(d \log^2 n + \tau)$ in time

$$O\left(\mathsf{nnz}(\mathbf{A}) \log^3 n + \frac{nd^\omega}{\tau} \log n\right).$$

The class of functions handled by Theorem 10.0.1 include the Huber loss, any ℓ_p loss for $p \leq 2$, as well as a wide variety of loss functions considered in the robust statistics literature that behave similarly to the Huber loss, that is, quadratic growth near the origin and linear growth away from the origin.

The idea behind Theorem 10.0.1 starts from an observation from the streaming literature [BO10] that for functions g of at most quadratic growth, entries $i \in [n]$ of a vector $\mathbf{y}$ which are "heavy" in the g loss, that is, $g(\mathbf{y}_i)/\|\mathbf{y}\|_g = \Omega(1)$, must also be "heavy" in the ℓ_2 loss, that is, $|\mathbf{y}_i|^2/\|\mathbf{y}\|_2^2 = \Omega(1)$. Thus, a superset of heavy elements in the g loss can be identified by identifying the heavy elements in the ℓ_2 loss, and furthermore, this superset is not too large by the definition of heaviness. This can then be generalized to identifying ε-heavy elements, that is, $g(\mathbf{y}_i)/\|\mathbf{y}\|_g \geq \varepsilon$, based on a standard argument. This argument is based on random hashing, and if we randomly hash the entries of $\mathbf{y}$ into $O(1/\varepsilon)$ buckets, then within this bucket, an ε-heavy entry is likely to be $\Omega(1)$-heavy.

Finally, we can now draw an analogy between "heavy" entries under the g loss with rows of $\mathbf{A}$ with large sensitivity $\boldsymbol{\sigma}_i$, as well as "heavy" entries under the ℓ_2 loss with rows of $\mathbf{A}$ with large ℓ_2 leverage score. Thus, by combining leverage score estimation with a hashing trick, we arrive at our Theorem 10.0.1.

In Section 10.1, we collect basic definitions and lemmas concerning M-estimators. Section 10.2 develops basic notions for sensitivity sampling for M-estimators. In Section 10.2.1, we describe our efficient algorithm for computing sensitivities for a broad class of M-estimators. In Section 10.2.2, we show that a variation on our efficient algorithm can be used to show an existential bound of $O(d^{\max\{1,p_M/2\}} \log n)$ total sensitivity for the same class of M-estimators. Finally, in Section 10.3, we show that the Tukey loss has a total sensitivity of $\Omega(d \log n)$, and that the Huber loss has a total sensitivity of $\Omega(d \log \log n)$.

10.1 M-estimators preliminaries

In this section, we define M-norms and collect some of their geometric properties. This is a slight generalization of Section 4.1 of [CW15a] which allows for a broader class of M-norms (namely with a relaxed polynomial lower bound condition). With applications to active regression in mind, we also slightly generalize the results to handle translations by a single vector $\mathbf{b}$, which can be taken to be 0 to retrieve the original results.

Definition 10.1.1. Let $M : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be increasing. If there exist constants $p > 0$ and $c_U \geq 1$ such that for all $y > x$,

$$\frac{M(y)}{M(x)} \leq c_U \left(\frac{y}{x}\right)^p,$$

then we say that M is *polynomially bounded above with degree p and constant c_U*. Similarly, if there exists constants $q > 0$ and $c_L \geq 1$ such that for all $y > x$,

$$\frac{M(y)}{M(x)} \geq c_L \left(\frac{y}{x}\right)^q,$$

then we say that M is *polynomially bounded below with degree q and constant c_L*.

Remark 10.1.2. As noted in [CW15a], it can be shown that convex functions are polynomially bounded below with degree 1.

Remark 10.1.3. Throughout this work, we will consider the constants p, q, c_U, c_L in Definition 10.1.1 to be absolute constants that don't depend on other parameters under consideration.

We define the M-norm as follows. Note that despite our abuse of notation and terminology, the M-norm need not be an actual norm.

Definition 10.1.4 (M-norm)**.** Let $M : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be such that

- $M(0) = 0$
- M is nondecreasing
- M is polynomially bounded above with degree p_M and constant c_U (see Definition 10.1.1)

Let $\mathbf{w} \in \mathbb{R}^n$ be a set of weights such that

$$\mathbf{w}_i \geq 1$$

for all $i \in [n]$. Then, we define the M-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ as

$$\|\mathbf{x}\|_{M,\mathbf{w}} := \left[\sum_{i=1}^{n} \mathbf{w}_i M(|\mathbf{x}_i|)\right]^{1/p_M}.$$

If $\mathbf{w}$ is the vector of all ones, we simple write $\|\mathbf{x}\|_M$ for $\|\mathbf{x}\|_{M,\mathbf{w}}$. If $M(x) = |x|^p$ for some $p > 0$, then we write $\|\mathbf{x}\|_{p,\mathbf{w}}$ for $\|\mathbf{x}\|_{M,\mathbf{w}}$.

Definition 10.1.5 (M balls and spheres)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let $\mathcal{V} = \text{span}(\mathbf{A})$. Let $M : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfy the conditions of Definition 10.1.4, and let $\mathbf{w} \geq \mathbf{1}_n$ be a set of weights. Define the ball $\mathcal{B}_\rho^{M,\mathbf{w}}$ of radius $\rho > 0$ as

$$\mathcal{B}_\rho^{M,\mathbf{w}} := \left\{\mathbf{y} \in \mathcal{V} : \|\mathbf{y}\|_{M,\mathbf{w}} \leq \rho\right\}.$$

Similarly define the sphere $\mathcal{S}_\rho^{M,\mathbf{w}}$ of radius $\rho > 0$ as

$$\mathcal{S}_\rho^{M,\mathbf{w}} \coloneqq \Big\{\mathbf{y} \in \mathcal{V} : \|\mathbf{y}\|_{M,\mathbf{w}} = \rho\Big\}.$$

If $\mathbf{w} = \mathbf{1}_n$, then we simply write $\mathcal{B}_\rho^M$ and $\mathcal{S}_\rho^M$, respectively.

The next lemma compares important entries using the polynomial boundedness condition.

Lemma 10.1.6. Let M be polynomially bounded above with degree p and constant $c_U \geq 1$. Let $\mathbf{x} \in \mathbb{R}^n$ be a vector with entries arranged in order, i.e., $|\mathbf{x}_1| \geq |\mathbf{x}_2| \geq \cdots \geq |\mathbf{x}_n|$. Then,

$$\frac{M(|\mathbf{x}_1|)}{\|\mathbf{x}\|_M^p} \leq c_U \frac{|\mathbf{x}_1|^p}{\|\mathbf{x}\|_p^p}.$$

Proof. Note that for all $i \geq 2$, we have by the polynomially boundedness condition that

$$\frac{M(|\mathbf{x}_1|)}{M(|\mathbf{x}_i|)} \leq c_u \left(\frac{|\mathbf{x}_1|}{|\mathbf{x}_i|}\right)^p = c_U \frac{|\mathbf{x}_1|^p}{|\mathbf{x}_i|^p}.$$

Then,

$$\begin{aligned}
\frac{M(|\mathbf{x}_1|)}{\|\mathbf{x}\|_M^p} &= \frac{M(|\mathbf{x}_1|)}{M(|\mathbf{x}_1|) + \sum_{i=2}^n M(|\mathbf{x}_i|)} \\
&= \left[1 + \sum_{i=2}^n \frac{M(|\mathbf{x}_i|)}{M(|\mathbf{x}_1|)}\right]^{-1} \leq \left[1 + \sum_{i=2}^n \frac{1}{c_U} \frac{|\mathbf{x}_i|^p}{|\mathbf{x}_1|^p}\right]^{-1} \\
&= \frac{|\mathbf{x}_1|^p}{|\mathbf{x}_1|^p + c_U^{-1} \sum_{i=2}^n |\mathbf{x}_i|^p} \leq c_U \frac{|\mathbf{x}_1|^p}{\|\mathbf{x}\|_p^p}.
\end{aligned}$$

□

10.2 Sensitivities upper bounds

Because M-estimators are defined as coordinate-wise sums, one can naturally define analogues of sensitivities, just as was done for ℓ_p norms.

Definition 10.2.1 (M-sensitivity)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let $\|\cdot\|_M$ be an M-norm. Then, the *ith M-sensitivity* is defined as

$$\mathbf{s}_i^M(\mathbf{A}) \coloneqq \sup_{\mathbf{x} \in \mathbb{R}^d, \mathbf{A}\mathbf{x} \neq 0} \frac{M(|[\mathbf{A}\mathbf{x}](i)|)}{\|\mathbf{A}\mathbf{x}\|_M^{p_M}}$$

and the *total M-sensitivity* is defined as

$$\mathcal{T}^M(\mathbf{A}) \coloneqq \sum_{i=1}^n \mathbf{s}_i^M(\mathbf{A}).$$

Let $\mathbf{w} \geq \mathbf{1}_n$ be a set of weights. Then, the *ith weighted M-sensitivity* is defined as

$$\mathbf{s}_i^{M,\mathbf{w}}(\mathbf{A}) \coloneqq \sup_{\mathbf{x} \in \mathbb{R}^d, \mathbf{A}\mathbf{x} \neq 0} \frac{\mathbf{w}_i M(|[\mathbf{A}\mathbf{x}](i)|)}{\|\mathbf{A}\mathbf{x}\|_{M,\mathbf{w}}^{p_M}}$$

and the *total weighted M-sensitivity* is defined as

$$\mathcal{T}^{M,\mathbf{w}}(\mathbf{A}) \coloneqq \sum_{i=1}^{n} \mathbf{s}_i^{M,\mathbf{w}}(\mathbf{A}).$$

When $M(x) = |x|^p$, i.e. for the case of ℓ_p norms, it is known that sampling with probabilities proportional to upper bounds on sensitivities yields subspace embeddings [BLM89, DDH+09, CP15]. Analogous results are known as well for M-estimators [CW15b, CW15a, CWW19] and Orlicz norms [SWY+19].

Definition 10.2.2 (Sensitivity Sampling for M-Estimators)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$, let $\|\cdot\|_M$ be an M-norm, and let $\mathbf{w} \geq \mathbf{1}_n$ be a set of weights. Let m be an oversampling parameter. Then, a random set of weights $\mathbf{w}'$ is sampled according to sensitivity upper bounds $\tilde{\mathbf{s}}_i^{M,\mathbf{w}}(\mathbf{A}) \geq \mathbf{s}_i^{M,\mathbf{w}}(\mathbf{A})$ (see Definition 10.2.1) if

$$\mathbf{w}'_i \coloneqq \begin{cases} \mathbf{w}_i / \mathbf{p}_i & \text{w.p. } \mathbf{p}_i \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{p}_i \coloneqq \min\{1, m \cdot \tilde{\mathbf{s}}_i^{M,\mathbf{w}}(\mathbf{A})\}$.

Note that in the case of M-estimators, the lack of scale invariance means that we get norm preservation guarantees for spheres rather than for entire subspaces. That is, we can get the following lemma, similar Lemma 43 of [CW15a]:

Lemma 10.2.3. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Let $\varepsilon \in (0, 1)$, $\delta > 0$, and let $\rho \geq 1$. Let $M : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfy the conditions of Definition 10.1.4, and furthermore that

- M^{1/p_M} is subadditive
- M is polynomially bounded below with degree q_M and constant c_L (see Definition 10.1.1)

Let $\mathbf{w} \geq \mathbf{1}_n$ be a set of weights. Let $\tilde{\mathbf{s}}_i^{M,\mathbf{w}}(\mathbf{A}) \geq \mathbf{s}_i^{M,\mathbf{w}}(\mathbf{A})$ be sensitivity upper bounds. Let $m \geq m_0$ be an oversampling parameter larger than some

$$m_0 = O\left(\frac{d}{\varepsilon^2}\left(\log\frac{1}{\varepsilon}\right)\left(\log\frac{1}{\delta}\right)\right).$$

Let $\mathbf{w}' \geq \mathbf{1}_n$ be sampled according to Definition 10.2.2. Then with probability at least $1 - \delta$,

$$\|\mathbf{y}\|_{M,\mathbf{w}'}^{p_M} = (1 \pm \varepsilon)\|\mathbf{y}\|_{M,\mathbf{w}}^{p_M}$$

for all $\mathbf{y} \in \mathcal{S}_\rho^M$ (see Definition 10.1.5). Furthermore,

$$\mathbf{E}\,\mathsf{nnz}(\mathbf{w}') \leq m \sum_{i=1}^{n} \tilde{\mathbf{s}}_i^{M,\mathbf{w}} = m\tilde{\mathcal{T}}^{M,\mathbf{w}}(\mathbf{A})$$

Note that the sample complexity of Lemma 10.2.3 is weaker than the corresponding bound for ℓ_p Lewis weights by a factor of d. This extra factor of d can in fact be removed, as shown by a remarkable recent result of [JLLS23].

10.2.1 Efficient algorithm for sensitivity upper bounds

We first show that algorithmically, one can compute upper bounds to the M-estimator sensitivities that sum to at most $O(d^{\max\{1,p_M/2\}}\log^2 n + \tau)$ in time

$$O\left(\mathsf{nnz}(\mathbf{A})\log^3 n + \frac{nT}{\tau}\log n\right),$$

where $T = T(n,d)$ is such that constant factor ℓ_{p_M} Lewis weight approximation for an $n \times d$ matrix $\mathbf{B}$ takes $O(\mathsf{nnz}(\mathbf{B})\log n + T)$ time. For example, it is known that ℓ_p Lewis weights for $0 < p < 4$ can be approximated up to constant factors in $O(\mathsf{nnz}(\mathbf{A})\log n + d^\omega)$ time, so for $\tau = T = d^\omega$, we obtain a nearly input sparsity time algorithm that computes upper bounds to M-estimator sensitivities that sum to at most $O(d^{\max\{1,p_M/2\}}\log^2 n + d^\omega)$. In applications, this is enough to compute a set of $\mathrm{poly}(d)\log^2 n$ rows that approximates the original matrix well, at which point we can compute sensitivities that sum to only $O(d\log^2 n)$ in an additional $\mathrm{poly}(d\log n)$ time.

The algorithm draws ideas from a theorem of [CWW19, Theorem 3.4], which shows an input sparsity time algorithm for locating "heavy entries" for the Tukey loss, which is equivalent to finding coordinates with high Tukey sensitivity.

Algorithm 1 Sensitivity upper bounds

input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, M-norm M, parameter $1 \le \tau \le n$.
output: Upper bounds $\tilde{\mathbf{s}}_i^M(\mathbf{A})$ on $\mathbf{s}_i^M(\mathbf{A})$.

1: Initialize $\tilde{\mathbf{s}}_i^M(\mathbf{A}) \leftarrow 2\tau/n$.
2: **for** $r \in [\lceil \log_2(n/\tau)\rceil]$ **do**
3: **for** $t \in [O(\log n)]$ **do**
4: Hash the rows of $\mathbf{A}$ into $B = 10\cdot 2^r$ buckets $S_1, S_2, \ldots, S_B$.
5: Compute $O(1)$-approximate ℓ_p Lewis weights of each of the B buckets $\mathbf{A}\mid_{S_1}, \mathbf{A}\mid_{S_2}, \ldots, \mathbf{A}\mid_{S_B}$.
6: For any row i with an ℓ_{p_M} Lewis weight of at least $\Omega(1)$, set $\tilde{\mathbf{s}}_i^M(\mathbf{A}) \leftarrow \max\{2/2^r, \tilde{\mathbf{s}}_i^M(\mathbf{A})\}$.

Theorem 10.2.4 (Sensitivity upper bounds)**.** Let $\|\cdot\|_M$ be an M-norm. Let $1 \le \tau \le n$ be a parameter. Then, with probability at least $99/100$, Algorithm 1 computes sensitivity upper bounds $\tilde{\mathbf{s}}_i^M(\mathbf{A}) \ge \mathbf{s}_i^M(\mathbf{A})$ that sum to at most

$$\tilde{\mathcal{T}}^M(\mathbf{A}) \coloneqq \sum_{i=1}^n \tilde{\mathbf{s}}_i^M(\mathbf{A}) = O(d^{\max\{1,p_M/2\}}\log^2 n + \tau).$$

If constant factor Lewis weight approximation takes for an $n\times d$ matrix $\mathbf{B}$ takes $O(\mathsf{nnz}(\mathbf{B})\log n + T)$ time (see Theorem 6.1.8), then the total running time is

$$O\left(\mathsf{nnz}(\mathbf{A})\log^3 n + \frac{nT}{\tau}\log n\right).$$

Proof. We first show correctness of the algorithm, then show the sensitivity bound, and finally the running time guarantee.

Correctness. Let $r \in [[\log_2(n/\tau)]]$. Consider a coordinate $i \in [n]$ that has M-sensitivity between $1/2^r$ and $2/2^r$ and let $\mathbf{y} = \mathbf{A}\mathbf{x}$ be a corresponding vector which satisfies

$$\frac{M(|\mathbf{y}_i|)}{\|\mathbf{y}\|_M^{p_M}} \in \left[\frac{1}{2^r}, \frac{2}{2^r}\right].$$

Note then that there are at most $2^r - 1$ entries j of $\mathbf{y}$ such that $|\mathbf{y}_j| > |\mathbf{y}_i|$. In our algorithm (line 4), we randomly hash the n rows of $\mathbf{A}$ into $B = 10 \cdot 2^r$ buckets. Then, the probability that any one of these $2^r - 1$ entries is hashed to the same bucket as i is $1/B$, so by a union bound, the probability that $\mathbf{y}_i$ has the largest absolute value in its hash bucket is at least $1 - 2^r/B \geq 9/10$. Call this event $\mathcal{E}$.

Now let S be the set of indices which hash to the same bucket as row i and let $S' = S \setminus \{i\}$. By Markov's inequality, with probability at least $9/10$, i is hashed to a bucket such that the M-norm of all other entries is most $\|\mathbf{y}\,|_{S'}\|_M^{p_M} \leq \|\mathbf{y}\|_M^{p_M}/2^r$, where $\mathbf{y}\,|_{S'}$ is the restriction of $\mathbf{y}$ to the indices in S'. Call this event $\mathcal{F}$.

Condition on $\mathcal{E}$ and $\mathcal{F}$. We have by Lemma 10.1.6 that

$$\frac{|\mathbf{y}_i|^{p_M}}{\|\mathbf{y}\,|_S\|_{p_M}^{p_M}} \geq \frac{M(|\mathbf{y}_i|^{p_M})}{\|\mathbf{y}\,|_S\|_M^{p_M}} \geq \frac{M(|\mathbf{y}_i|^{p_M})}{M(|\mathbf{y}_i|^{p_M}) + \|\mathbf{y}\,|_{S'}\|_M^{p_M}} \geq \frac{1/2^r}{2/2^r + 1/2^r} = \frac{1}{3}.$$

The above holds with probability at least $4/5$. Thus, by repeating the hashing process $O(\log n)$ times, with probability at least $1 - 1/(100n\log_2 n) = 1 - 1/\operatorname{poly}(n)$, there exists some trial where the ℓ_{p_M} sensitivity of the ith row *in the matrix* $\mathbf{A}\,|_S$ is at least $1/3$. In this trial, our algorithm will correctly set $\tilde{\mathbf{s}}_i^M(\mathbf{A}) \geq 2/2^r$ (line 6). By a union bound over $O(\log(n/\tau))$ levels r and the n rows, our algorithm succeeds with probability at least $99/100$.

Sensitivity bound. By Lemma 6.2.4, the ℓ_{p_M} sensitivities sum to at most $d^{\max\{1,p_M/2\}}$. Thus, each time we compute $O(1)$-approximate ℓ_{p_M} Lewis weights (line 5), we find at most $O(d^{\max\{1,p_M/2\}})$ entries with M-sensitivity at least $2/2^r$. Thus, for each r and each iteration, we increase the sum of our upper bounds on M-sensitivities by a total of at most

$$B \cdot O(d^{\max\{1,p_M/2\}}) \cdot \frac{2}{2^r} = O(d^{\max\{1,p_M/2\}}).$$

This occurs at most $O((\log n)(\log(n/\tau)))$ times, and we start at a sensitivity bound of

$$\frac{2\tau}{n} \cdot n = O(\tau)$$

so our upper bounds on the sensitivities sum to at most

$$O(d^{\max\{1,p_M/2\}}(\log n)(\log(n/\tau)) + \tau) \leq O(d^{\max\{1,p_M/2\}}\log^2 n + \tau).$$

Running time. For a given r, the dominating running time cost of the inner-most loop of Algorithm 1 is the computation of ℓ_{p_M} Lewis weights for $O(2^r)$ matrices whose sparsities sum to

$\mathsf{nnz}(\mathbf{A})$. Thus, if Lewis weight computation for an $n \times d$ matrix $\mathbf{B}$ takes $O(\mathsf{nnz}(\mathbf{B}) \log n + T)$ time, then the total running time is

$$\sum_{r=1}^{\lceil \log_2(n/\tau) \rceil} O(\log n) \cdot O(\mathsf{nnz}(\mathbf{A}) \log n + 2^r T) = O\left(\mathsf{nnz}(\mathbf{A}) \log^3 n + \frac{nT}{\tau} \log n\right). \qquad \square$$

Remark 10.2.5. As noted by [TMF20], if we can control the sensitivities of functions M_1 and M_2, then it is straightforward to control the sensitivities of the sum of these two functions, i.e., $M = M_1 + M_2$. This applies to our algorithm as well. Suppose that $\mathbf{s}_i^M(\mathbf{A}) \in [1/2^r, 2/2^r]$ and let $\mathbf{y} = \mathbf{A}\mathbf{x}$ be such that

$$\frac{M_1(|\mathbf{y}_i|) + M_2(|\mathbf{y}_i|)}{\|\mathbf{y}\|_{M_1}^{p_{M_1}} + \|\mathbf{y}\|_{M_2}^{p_{M_2}}} \in \left[\frac{1}{2^r}, \frac{2}{2^r}\right].$$

Then,

$$\frac{M_1(|\mathbf{y}_i|)}{\|\mathbf{y}\|_{M_1}^{p_{M_1}}} + \frac{M_2(|\mathbf{y}_i|)}{\|\mathbf{y}\|_{M_2}^{p_{M_2}}} \geq \frac{1}{2^r}$$

so then there is some $j \in \{1, 2\}$ such that

$$\frac{M_j(|\mathbf{y}_i|)}{\|\mathbf{y}\|_{M_j}^{p_{M_j}}} \geq \frac{1}{2} \cdot \frac{1}{2^r},$$

so the ℓ_{p_j} Lewis weight of the ith coordinate must be large by using a similar proof as Theorem 10.2.4. Thus, we can obtain similar sensitivity upper bounds up to a constant factor loss. Similarly, if M_2 is a "flat" sensitivity function in the sense of [TMF20], that is, if

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \frac{M_2(|[\mathbf{A}\mathbf{x}](i)|)}{\|\mathbf{A}\mathbf{x}\|_{M_2}^{p_{M_2}}} = O\left(\frac{1}{n}\right)$$

for all $i \in [n]$, then this just means that either the M_1 sensitivity is large, or the M sensitivity is at most $O(1/n)$, in which case we still get the same bounds.

Although Theorem 10.2.4 only handles unweighted M-estimators, this result can be generalized to weighted M-estimators by splitting into level sets, similarly to Lemma 39 of [CW15a].

Lemma 10.2.6. Let $\|\cdot\|_M$ be an M-norm. Let $\mathbf{w} \geq \mathbf{1}_n$ be a set of weights. Let $N \coloneqq \lceil \log_2(1 + \|\mathbf{w}\|_\infty) \rceil$. For $j \in [N]$, let

$$T_j \coloneqq \left\{i \in [n] : 2^{j-1} \leq \mathbf{w}_i < 2^j\right\},$$

and let $\mathbf{A}\,|_{T_j}$ denote the restriction of $\mathbf{A}$ to the rows of T_j. Then,

$$\mathbf{s}_i^{M,\mathbf{w}}(\mathbf{A}) \leq 2 \cdot \mathbf{s}_i^M(\mathbf{A}\,|_{T_j})$$

for $i \in T_j$.

Proof. Let $i \in T_j$ for some $j \in [N]$. We have that

$$\begin{aligned}
\mathbf{s}_i^{M,\mathbf{w}}(\mathbf{A}) &= \sup_{\mathbf{x}\in\mathbb{R}^n,\mathbf{A}\mathbf{x}\neq 0} \frac{\mathbf{w}_i M([\mathbf{A}\mathbf{x}](i))}{\|\mathbf{A}\mathbf{x}\|_{M,\mathbf{w}}^{p_M}} \\
&\leq \sup_{\mathbf{x}\in\mathbb{R}^n,\mathbf{A}\mathbf{x}\neq 0} \frac{\mathbf{w}_i M([\mathbf{A}\mid_{T_j}\mathbf{x}](i))}{\left\|\mathbf{A}\mid_{T_j}\mathbf{x}\right\|_{M,\mathbf{w}}^{p_M}} \\
&\leq \sup_{\mathbf{x}\in\mathbb{R}^n,\mathbf{A}\mathbf{x}\neq 0} \frac{2^j M([\mathbf{A}\mid_{T_j}\mathbf{x}](i))}{2^{j-1}\left\|\mathbf{A}\mid_{T_j}\mathbf{x}\right\|_{M}^{p_M}} \\
&= 2\cdot\mathbf{s}_i^M(\mathbf{A}\mid_{T_j})
\end{aligned}$$

as desired. □

This leads to an algorithm that achieves guarantees similar to Theorem 10.2.4 for weighted M-sensitivities, up to a loss of a factor of N in the running time and sensitivity bound.

Corollary 10.2.7. Let $\|\cdot\|_M$ be an M-norm. Let $\mathbf{w} \geq \mathbf{1}_n$ be a set of weights. Define N as in Lemma 10.2.6. There is an algorithm that computes weighted M-estimator sensitivities that sum to at most

$$O(Nd^{\max\{1,p_M/2\}}\log^2 n + N\tau)$$

in time

$$O\left(\mathsf{nnz}(\mathbf{A})\log^3 n + N\frac{nT}{\tau}\log n\right),$$

where T is such that constant factor Lewis weight approximation for an $n\times d$ matrix $\mathbf{B}$ takes $O(\mathsf{nnz}(\mathbf{B})+T)$ time (see Theorem 6.1.8).

Proof. This is simply the result of applying Theorem 10.2.4 on the N matrices $\mathbf{A}\mid_{T_j}$ as defined in Lemma 10.2.6. Note that the $\mathsf{nnz}(\mathbf{A}\mid_{T_j})$ terms add up to $\mathsf{nnz}(\mathbf{A})$ in the running time. □

10.2.2 Sharper sensitivity bounds

We show that we may modify the proof of our input sparsity time algorithm to show that the sum of sensitivities is at most $O(d^{\max\{1,p_M/2\}}\log n)$, if we do not need to efficient algorithms for constructing these sensitivities.

Theorem 10.2.8. Let $\|\cdot\|_M$ be an M-norm. Then, the total M-sensitivity of $\mathbf{A}$ is at most

$$O(d^{\max\{1,p_M/2\}}\log n).$$

Proof. Our idea is essentially to run Algorithm 1 with $\tau = d$ without the $O(\log n)$ repetitions of the hashing process.

Let $r \in [\lceil\log_2 n\rceil]$ and let I_r be the set of coordinates with M-sensitivity in $[1/2^r, 2/2^r]$. Suppose we hash the rows of $\mathbf{A}$ into $B = 10\cdot 2^r$ buckets. Then, as in the proof of Theorem 10.2.4, for each $i\in I_r$, there is at least a $9/10$ probability that i has ℓ_{p_M} Lewis weight at least $1/3$ in its hash bucket. Thus, the number of such i is $(9/10)|I_r|$ in expectation, so there exists some hashing

such that at least $(9/10)|I_r|$ of the indices $i \in I_r$ have ℓ_p Lewis weight at least $1/3$ in its hash bucket. However, there can be at most $B \cdot d^{\max\{1,p/2\}}$ such indices, so we must have that

$$\frac{9}{10}|I_r| \leq B \cdot d^{\max\{1,p/2\}}$$

so

$$|I_r| = O(B \cdot d^{\max\{1,p/2\}}) = O(2^r \cdot d^{\max\{1,p/2\}}).$$

By summing over the r, we obtain a bound of

$$\sum_{r=1}^{\lceil \log_2 n \rceil} \frac{2}{2^r}|I_r| \leq \sum_{r=1}^{\lceil \log_2 n \rceil} \frac{2}{2^r} O(2^r \cdot d^{\max\{1,p/2\}}) = O(d^{\max\{1,p/2\}} \log n) = O(d^{\max\{1,p/2\}} \log n)$$

on the total M-sensitivity, as claimed. □

10.3 Sensitivity lower bounds

Finally, we show that our sensitivity upper bounds are tight by showing that the Tukey loss can have a total sensitivity as large as $\Omega(d \log(n/d))$. We also show a weaker lower bound of $\Omega(d \log \log(n/d))$ for the Huber loss. This is in contrast to sensitivities for the ℓ_p loss for $0 < p < \infty$, which is always at most $d^{\max\{1,p/2\}}$ due to the existence of Lewis bases [Lew78, SZ01], and thus has no dependence on n. The necessity for a dependence on n can be attributed to the lack of scale invariance for these M-estimator losses. A similar observation has been made previously in [SWZ19, Theorem 1.3], which shows that the column subset selection problem with the entrywise Huber loss exhibits a lower bound of $\Omega(\sqrt{\log n})$ columns, also attributed to the lack of scale invariance.

We simultaneously handle the Tukey and Huber losses by analyzing the ℓ_2-ℓ_p loss for $p \in [0, 1]$, which grows quadratically near the origin and as ℓ_p away from the origin, and is polynomially bounded above with degree 2.

Lemma 10.3.1 (Sensitivity lower bound for the ℓ_2-ℓ_p loss)**.** Define the ℓ_2-ℓ_p loss of width τ to be

$$M(x) = \begin{cases} x^2 & |x| \leq \tau \\ (\tau^2/\tau^p) \cdot x^p & |x| > \tau \end{cases}.$$

For $d \geq 1$ and $n \geq d$, there exists an $n \times d$ matrix $\mathbf{A}$ with total M-sensitivity that is at least

$$\mathcal{T}^T(\mathbf{A}) \geq \begin{cases} \Omega\big(d \log \frac{n}{d}\big) & \text{if } p \in [0, 1) \\ \Omega\big(d \log \log \frac{n}{d}\big) & \text{if } p = 1. \end{cases}$$

Proof. Let $\ell = \lfloor \log_2 n \rfloor$ and let $\mathbf{x} \in \mathbb{R}^n$ be a vector with 2^i coordinates of value $\tau/2^i$ for $i \in [\ell]$. We will show a sensitivity lower bound of $\Omega(\ell) = \Omega(\log n)$ for the $n \times 1$ matrix formed by the vector $\mathbf{x}$. By considering d disjoint copies of this vector, each on n/d coordinates, this implies a lower bound of $\Omega(d \log(n/d))$.

Let $j \in [\ell]$. Then,

$$\begin{aligned}
\left\|2^j \cdot \mathbf{x}\right\|_M^2 &= \sum_{i=1}^{\ell} 2^i \cdot M\left(\tau \frac{2^j}{2^i}\right) \\
&\leq \frac{\tau^2}{\tau^p} \sum_{i=1}^{j} 2^i \cdot \left(\tau \frac{2^j}{2^i}\right)^p + \sum_{i=j+1}^{\ell} 2^i \cdot \left(\tau \frac{2^j}{2^i}\right)^2 \\
&= \tau^2 2^{pj} \sum_{i=1}^{j} 2^{(1-p)i} + \tau^2 2^{2j} \sum_{i=j+1}^{\ell} \frac{1}{2^i} \\
&= \begin{cases} O(\tau^2 \cdot 2^j) & \text{if } p \in [0,1) \\ O(\tau^2 \cdot j2^j) & \text{if } p = 1 \end{cases}
\end{aligned}$$

so for each $j \in [\ell]$, there are 2^j coordinates i such that

$$\begin{aligned}
\frac{M(2^j \cdot \mathbf{x}_i)}{\left\|2^j \cdot \mathbf{x}\right\|_M^2} &= \begin{cases} \Omega\left(\frac{\tau^2}{\tau^2 \cdot 2^j}\right) & \text{if } p \in [0,1) \\ O\left(\frac{\tau^2}{\tau^2 \cdot j2^j}\right) & \text{if } p = 1 \end{cases} \\
&= \begin{cases} \Omega\left(\frac{1}{2^j}\right) & \text{if } p \in [0,1) \\ O\left(\frac{1}{j2^j}\right) & \text{if } p = 1 \end{cases}.
\end{aligned}$$

Thus, the sum of sensitivities for the Tukey loss for this matrix is at least

$$\sum_{j=1}^{\ell} 2^j \cdot \Omega\left(\frac{1}{2^j}\right) = \Omega(\ell) = \Omega(\log n)$$

for $p \in [0,1)$ and

$$\sum_{j=1}^{\ell} 2^j \cdot \Omega\left(\frac{1}{j2^j}\right) = \Omega(\log \ell) = \Omega(\log \log n)$$

for $p = 1$. □

Chapter 11

Applications: streaming ℓ_∞ subspace embeddings and computational geometry [WY22a]

An investigation of high-distortion ℓ_p subspace embeddings for $p > 2$ in the previous Chapter 9 prompts a closely related study in the *streaming setting*, in which we must compute an ℓ_p subspace embedding of the matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, when $\mathbf{A}$ is presented as n rows $\mathbf{a}_i \in \mathbb{R}^d$ which arrive one by one in one pass over a stream (see Section 1.3.3).

Note that when we have algorithms for subspace embeddings with $(1+\varepsilon)$ distortion, then we can easily obtain a streaming algorithm by a technique known as *merge-and-reduce*, in which we iteratively perform the operations of concatenating new rows and reducing the size of the stored subspace embedding by re-computing a subspace embedding. These operations can be performed in a way such that the subspace embedding is re-computed at a "depth" of only $O(\log n)$ if the input matrix $\mathbf{A}$ has n rows, meaning that if we compute subspace embeddings with distortion $(1+\varepsilon/\log n)$ at each step, then the total distortion is only $(1+\varepsilon/\log n)^{\log n} = (1+O(\varepsilon))$. However, this trick does not work for when our distortions are $\kappa = (1+\Omega(1))$, and leads to $\mathrm{poly}(n)$ factor total distortions when applied in this case.

Perhaps the most important case of this problem is that of computing ℓ_∞ subspace embeddings in the streaming model. In this case, Theorem 9.0.3, both in the upper bound and lower bound, can be generalized to show that ℓ_∞ subspace embeddings with $\kappa = \sqrt{d}$ distortion and $r = d$ rows can be obtained, and that the upper bound comes from ℓ_∞ Lewis weights, which corresponds to the well-studied problem of *Löwner–John ellipsoids* [Joh48, Tod16], also known as minimum volume enclosing ellipsoids. However, the question of computing Löwner–John ellipsoids in the streaming setting using only $\mathrm{poly}(d)$ bits of space is a central unresolved problem in the literature of computational geometry [MSS10, AS15]. Indeed, the only known prior results for computing Löwner–John ellipsoids in a stream uses $\exp(\mathrm{poly}(d))$ bits of space in order to estimate the extent of every direction in $\mathbb{R}^d$ using a net [AHV04, AHV05], rather than polynomial in d. Thus the question of efficiently maintaining ℓ_∞ subspace embeddings in a stream is an important problem.

In our work of [WY22a], we resolve both the problem of maintaining ℓ_∞ subspace embeddings and Löwner–John ellipsoids in the streaming setting, and in fact, a multitude of other problems in the streaming computational geometry literature which previously only admitted upper bounds

with exponential dependencies in the dimension. Our central theorem is the following:

Theorem 11.0.1 (Streaming ℓ_∞ subspace embedding [WY22a])**.** There is a deterministic streaming algorithm such that, for any $\mathbf{A} \in \mathbb{Z}^{n\times d}$ presented in a geometric stream, the algorithm maintains $\mathbf{SA}$ for a matrix $\mathbf{S} \in \mathbb{Z}^{r\times n}$ such that for every $\mathbf{x} \in \mathbb{R}^d$,

$$\|\mathbf{Ax}\|_\infty \le \|\mathbf{SAx}\|_\infty \le O(\sqrt{d \log n})\|\mathbf{Ax}\|_\infty.$$

Furthermore, the algorithm uses at most $O(d^2(\log n)^2)$ bits of space.

Our main technique is the use of *online leverage scores* (see Section 1.3.3) as a tool both to discover directions $\mathbf{x} \in \mathbb{R}^d$ in which the ℓ_∞ norm $\|\mathbf{Ax}\|_\infty$ is updated significantly in a stream, and to bound the total number of such updates which can occur.

A related result on maintaining Löwner–John ellipsoids in the streaming setting has been obtained in concurrent work of [MMO22], which achieve results that depend on a certain condition number of the ellipsoid. The case of asymmetric polytopes was later handled by [MMO23].

11.1 Nearly optimal sum of online leverage scores

We begin with a theorem establishing a tight bound on the sum of online leverage scores when $\mathbf{A}$ has integer entries bounded by $\mathrm{poly}(n)$.

Theorem 11.1.1. Let $\mathbf{A} \in \mathbb{Z}^{n\times d}$ have entries bounded in absolute value by $\mathrm{poly}(n)$. Then,

$$\sum_{i=1}^{n} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}) = O(d \log n).$$

Our argument will need the notion of a pseudodeterminant.

Definition 11.1.2 (Pseudodeterminant)**.** Let $\mathbf{M} \in \mathbb{R}^{d\times d}$ be a symmetric matrix of rank r. Then, the *pseudodeterminant* $\mathrm{pdet}(\mathbf{M})$ of $\mathbf{M}$ is the product of the nonzero eigenvalues of $\mathbf{M}$.

We need the following simple lemmas which dictate the evolution of pseudodeterminants under row additions. The first shows how to handle the additional of orthogonal rows.

Lemma 11.1.3. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $\mathbf{a} \in \mathbb{R}^d$ be a vector that is orthogonal to the row span of $\mathbf{A}$. Then,

$$\mathrm{pdet}(\mathbf{A}^\top\mathbf{A} + \mathbf{aa}^\top) = \|\mathbf{a}\|_2^2 \cdot \mathrm{pdet}(\mathbf{A}^\top\mathbf{A}).$$

Proof. Let $\mathbf{A} = \mathbf{U\Sigma V}^\top$. Note that the SVD of the concatenation $\mathbf{A}' \in \mathbb{R}^{(n+1)\times d}$ of $\mathbf{A}$ and $\mathbf{a}$ is

$$\mathbf{A}' = \begin{pmatrix}\mathbf{A}\\ \mathbf{a}\end{pmatrix} = \begin{pmatrix}\mathbf{U} & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma} & 0\\ 0 & \|\mathbf{a}\|_2\end{pmatrix}\begin{pmatrix}\mathbf{V}^\top\\ \mathbf{a}^\top/\|\mathbf{a}\|_2\end{pmatrix}.$$

Thus,

$$\mathrm{pdet}(\mathbf{A}^\top\mathbf{A} + \mathbf{aa}^\top) = \mathrm{pdet}(\mathbf{A}'^\top\mathbf{A}') = \|\mathbf{a}\|_2^2 \prod_{j=1}^{d} \sigma_j^2 = \|\mathbf{a}\|_2^2 \cdot \mathrm{pdet}(\mathbf{A}^\top\mathbf{A}),$$

as claimed. □

Our second lemma is a generalization of the matrix determinant lemma to pseudodeterminants.

Lemma 11.1.4 (Matrix pseudodeterminant lemma)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let $\mathbf{a} \in \mathbb{R}^d$ be a vector that is in the row span of $\mathbf{A}$. Then,

$$\operatorname{pdet}(\mathbf{A}^\top \mathbf{A} + \mathbf{a}\mathbf{a}^\top) = \operatorname{pdet}(\mathbf{A}^\top \mathbf{A})(1 + \mathbf{a}(\mathbf{A}^\top \mathbf{A})^- \mathbf{a}).$$

Proof. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the truncated SVD of $\mathbf{A}$. Since $\mathbf{a}$ is in the row span of $\mathbf{V}$, we may write $\mathbf{a} = \mathbf{V}\mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^r$, where $r = \operatorname{rank}(\mathbf{A})$. Then,

$$\begin{aligned}
\operatorname{pdet}(\mathbf{A}^\top \mathbf{A} + \mathbf{a}\mathbf{a}^\top) &= \operatorname{pdet}(\mathbf{\Sigma}^2 + \mathbf{b}\mathbf{b}^\top) \\
&= \det(\mathbf{\Sigma}^2 + \mathbf{b}\mathbf{b}^\top) \\
&= \det(\mathbf{\Sigma}^2)(1 + \mathbf{b}^\top \mathbf{\Sigma}^{-2} \mathbf{b}) && \text{matrix determinant lemma} \\
&= \operatorname{pdet}(\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top)(1 + \mathbf{b}^\top \mathbf{V}^\top (\mathbf{V}\mathbf{\Sigma}^{-2}\mathbf{V}^\top)\mathbf{V}\mathbf{b}) \\
&= \operatorname{pdet}(\mathbf{A}^\top \mathbf{A})(1 + \mathbf{a}(\mathbf{A}^\top \mathbf{A})^- \mathbf{a})
\end{aligned}$$

as desired. □

We also need the following identity found in [GK10], which states that the volume of a parallelotope is given by both the determinant of the Gram matrix as well as the product of the heights of the dimensions of the parallelotope.

Lemma 11.1.5 (Determinant volume identity [GK10])**.** Let $\mathbf{A} \in \mathbb{R}^{r \times d}$ have linearly independent rows. Then,

$$\sqrt{\det(\mathbf{A}\mathbf{A}^\top)} = \prod_{i=1}^{r} \left\| \mathbf{a}_i^\perp \right\|_2$$

where $\mathbf{a}_1^\perp = \mathbf{a}_1$ and $\mathbf{a}_i^\perp$ is the projection of $\mathbf{a}_i$ onto the orthogonal complement of the row span of $\mathbf{A}_{i-1}$ for $i \geq 2$.

We now prove the following main theorem of this section.

Proof of Theorem 11.1.1. Our proof is a careful improvement of the original proof by [CMP20] under our bit complexity assumption. Let $i \in [n]$. If $\mathbf{a}_{i+1}$ is in the row span of $\mathbf{A}_i$, then by Lemma 11.1.4, we have that

$$\begin{aligned}
\operatorname{pdet}(\mathbf{A}_{i+1}^\top \mathbf{A}_{i+1}) &= \operatorname{pdet}(\mathbf{A}_i^\top \mathbf{A}_i)(1 + \mathbf{a}_{i+1}(\mathbf{A}_i^\top \mathbf{A}_i)^- \mathbf{a}_{i+1}) \\
&\geq \operatorname{pdet}(\mathbf{A}_i^\top \mathbf{A}_i)(1 + \boldsymbol{\tau}_{i+1}^{\mathsf{OL}}(\mathbf{A})) \\
&\geq \operatorname{pdet}(\mathbf{A}_i^\top \mathbf{A}_i) \exp(\boldsymbol{\tau}_{i+1}^{\mathsf{OL}}(\mathbf{A})/2)
\end{aligned}$$

and otherwise, let $\mathbf{a}_{i+1} = \mathbf{a}_{i+1}^{\|} + \mathbf{a}_{i+1}^{\perp}$, where $\mathbf{a}^{\|}$ is the projection of $\mathbf{a}_{i+1}$ onto the row span of $\mathbf{A}_i$ and $\mathbf{a}^{\perp}$ is the residual. We have that

$$\begin{aligned}
\operatorname{pdet}(\mathbf{A}_{i+1}^\top \mathbf{A}_{i+1}) &= \operatorname{pdet}(\mathbf{A}_i^\top \mathbf{A}_i + \mathbf{a}_{i+1}\mathbf{a}_{i+1}^\top) \\
&= \operatorname{pdet}(\mathbf{A}_i^\top \mathbf{A}_i + \mathbf{a}_{i+1}^{\|}(\mathbf{a}_{i+1}^{\|})^\top + \mathbf{a}_{i+1}^{\perp}(\mathbf{a}_{i+1}^{\perp})^\top)
\end{aligned}$$

$$
\begin{aligned}
&= \left\|\mathbf{a}_{i+1}^{\perp}\right\|_2^2 \cdot \operatorname{pdet}(\mathbf{A}_i^\top \mathbf{A}_i + \mathbf{a}_{i+1}^{\parallel}(\mathbf{a}_{i+1}^{\parallel})^\top) && \text{Lemma 11.1.3}\\
&= \left\|\mathbf{a}_{i+1}^{\perp}\right\|_2^2 \cdot \operatorname{pdet}(\mathbf{A}_i^\top \mathbf{A}_i)(1 + \mathbf{a}_{i+1}^{\parallel}(\mathbf{A}_i^\top \mathbf{A})^{-}\mathbf{a}_{i+1}^{\parallel}) && \text{Lemma 11.1.4}\\
&\geq \left\|\mathbf{a}_{i+1}^{\perp}\right\|_2^2 \cdot \operatorname{pdet}(\mathbf{A}_i^\top \mathbf{A}_i)
\end{aligned}
$$

Now let $S \subseteq [n]$ denote the at most d indices such that $\mathbf{a}_i$ is not in the row span of $\mathbf{A}_{i-1}$. Note that we take $1 \in S$ so that $\mathbf{a}_1^{\perp} = \mathbf{a}_1$. We then have by induction that

$$
\begin{aligned}
\operatorname{pdet}(\mathbf{A}^\top \mathbf{A}) = \operatorname{pdet}(\mathbf{A}_n^\top \mathbf{A}_n) &\geq \prod_{i \in [n]\setminus S} \exp(\boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A})/2) \prod_{j \in S} \left\|\mathbf{a}_j^{\perp}\right\|_2^2 \\
&= \exp\left(\frac{1}{2}\sum_{i \in [n]\setminus S} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A})\right) \prod_{j \in S} \left\|\mathbf{a}_j^{\perp}\right\|_2^2 \\
&= \exp\left(\frac{1}{2}\sum_{i \in [n]\setminus S} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A})\right) \det(\mathbf{A}|_S \mathbf{A}|_S^\top) \qquad \text{Lemma 11.1.5}
\end{aligned}
$$

where $\mathbf{A}|_S$ is the restriction of $\mathbf{A}$ to the rows indexed by S. By bounding each eigenvalue by the operator norm, we have that $\operatorname{pdet}(\mathbf{A}^\top \mathbf{A}) \leq \left\|\mathbf{A}^\top \mathbf{A}\right\|_2^d \leq \operatorname{poly}(n)^d$. Furthermore, since $\mathbf{A}|_S \mathbf{A}|_S^\top$ is a nonsingular integer Gram matrix, it has positive integer determinant, which is in particular at least 1. We thus have that

$$
\exp\left(\frac{1}{2}\sum_{i \in [n]\setminus S} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A})\right) \leq \operatorname{poly}(n)^d \implies \sum_{i \in [n]\setminus S} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}) \leq O(d \log n).
$$

Finally, $|S| \leq d$, which implies that

$$
\sum_{i=1}^{n} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}) = \sum_{i \in [n]\setminus S} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}) + \sum_{i \in S} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}) \leq O(d \log n) + d = O(d \log n),
$$

as claimed. □

11.2 Online coresets for ℓ_∞ subspace embeddings

We will analyze the following algorithm for constructing ℓ_∞ subspace embeddings in the online coreset model, which keeps a new row $\mathbf{a}_i$ if and only if it exceeds the "ℓ_2 width" of the previously kept rows.

We show the following guarantee for Algorithm 2:

Theorem 11.2.1. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ such that for any subset $S' \subseteq [n]$, the sum of online leverage scores is bounded by

$$
\sum_{i \in S'} \boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}|_{S'}) \leq T
$$

and let S be the output of Algorithm 2. Then:

Algorithm 2 Online ℓ_∞ subspace sketch coreset

input: $\mathbf{A} \in \mathbb{R}^{n \times d}$.
output: Coreset $S \subseteq [n]$.

1: $S \leftarrow \varnothing$
2: **for** $i \in [n]$ **do**
3: **if** $\exists \mathbf{x} \in \mathbb{R}^n : \langle \mathbf{a}_i, \mathbf{x} \rangle^2 \geq \|\mathbf{A}|_S \mathbf{x}\|_2^2$ **then**
4: $S \leftarrow S \cup \{i\}$
5: **return** S

- $|S| \leq O(T)$
- $\frac{1}{\Delta}\|\mathbf{A}\mathbf{x}\|_\infty \leq \|\mathbf{A}|_S \mathbf{x}\|_\infty \leq \|\mathbf{A}\mathbf{x}\|_\infty$ for all $\mathbf{x} \in \mathbb{R}^d$, for $\Delta = O(\sqrt{T})$.

In particular, if $\mathbf{A} \in \mathbb{Z}^{n \times d}$ is an integer matrix with entries bounded by $\mathrm{poly}(d)$, then by storing the rows of S, we obtain an algorithm for the streaming ℓ_∞ subspace sketch problem using $O(d^2 \log^2 n)$ bits of space, and if $\mathbf{A} \in \mathbb{R}^{n \times d}$ has online pseudo condition number κ^{OL}, then we obtain an online coreset algorithm storing at most $O(d \log(n\kappa^{\mathsf{OL}}))$ rows and achieves distortion at most $O(\sqrt{d \log(n\kappa^{\mathsf{OL}})})$.

Proof. We first bound $|S|$. Note that for every $i \in S$,

$$\boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}|_S) = \Omega(1)$$

since if $\mathbf{a}_i \in \mathrm{rowspan}((\mathbf{A}|_S)_{i-1})$, then by Line 3 and Lemma 1.3.5,

$$\boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}|_S) = \sup_{(\mathbf{A}|_S)_{i-1}\mathbf{x} \neq 0} \frac{\langle \mathbf{a}_i, \mathbf{x} \rangle^2}{\|(\mathbf{A}|_S)_{i-1}\mathbf{x}\|_2^2} \geq 1$$

while if $\mathbf{a}_i \notin \mathrm{rowspan}((\mathbf{A}|_S)_{i-1})$, then $\boldsymbol{\tau}_i^{\mathsf{OL}}(\mathbf{A}|_S) = 1$. Since the online leverage scores of $\mathbf{A}_S$ sum to at most T, it follows that $|S| \leq O(T)$.

Next, we bound the distortion Δ. Note that $\|\mathbf{A}|_S \mathbf{x}\|_\infty \leq \|\mathbf{A}\mathbf{x}\|_\infty$ is trivial, so it suffices to show the lower bound. Let $\mathbf{x} \in \mathbb{R}^d$ and let $i_* \in [n]$ satisfy $\|\mathbf{A}\mathbf{x}\|_\infty = |\langle \mathbf{a}_{i_*}, \mathbf{x} \rangle|$, i.e., the row that witnesses the max. If $i_* \in S$, then we already have that

$$\|\mathbf{A}|_S \mathbf{x}\|_\infty \geq \|\mathbf{A}\mathbf{x}\|_\infty$$

so assume that $i_* \notin S$. Then,

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}\|_\infty^2 &= \langle \mathbf{a}_{i_*}, \mathbf{x} \rangle^2 \\
&\leq \|\mathbf{A}|_S \mathbf{x}\|_2^2 && \text{Line 3} \\
&\leq |S| \cdot \|\mathbf{A}|_S \mathbf{x}\|_\infty^2 \\
&\leq O(T)\|\mathbf{A}|_S \mathbf{x}\|_\infty^2
\end{aligned}$$

which yields the claimed bound on Δ. The guarantee for streaming algorithms for $\mathbf{A}$ with bounded bit complexity follow from online leverage score bound from Theorem 11.1.1. The guarantee for online coreset algorithms follows from Lemma 1.3.8 and by noting that

$$\kappa^{\mathsf{OL}}(\mathbf{A}_S) \leq n \cdot \kappa^{\mathsf{OL}}(\mathbf{A})$$

since for any $i \in [n]$,

$$\begin{aligned}\left\|(\mathbf{A}_S)_i^-\right\|_2^{-1} &= \min_{\|\mathbf{x}\|_2=1, \mathbf{x}\in\text{rowspan}((\mathbf{A}|_S)_i)} \|(\mathbf{A}|_S)_i\mathbf{x}\|_2 \\ &\geq \min_{\|\mathbf{x}\|_2=1, \mathbf{x}\in\text{rowspan}((\mathbf{A}|_S)_i)} \|(\mathbf{A}|_S)_i\mathbf{x}\|_\infty \\ &\geq \frac{1}{\Delta} \min_{\|\mathbf{x}\|_2=1, \mathbf{x}\in\text{rowspan}((\mathbf{A}|_S)_i)} \|\mathbf{A}_i\mathbf{x}\|_\infty \\ &\geq \frac{1}{\Delta\sqrt{n}} \min_{\|\mathbf{x}\|_2=1, \mathbf{x}\in\text{rowspan}((\mathbf{A}|_S)_i)} \|\mathbf{A}_i\mathbf{x}\|_2 \geq \frac{1}{n}\left\|\mathbf{A}_i^-\right\|_2^{-1}.\end{aligned}$$

Note that we use that S has the ℓ_∞ subspace embedding guarantee here, rather than using an upper bound on T. □

Remark 11.2.2. It is not hard to see that $\|\mathbf{A}|_S\mathbf{x}\|_2$ can be used as the subspace sketch estimator in Theorem 11.2.1 instead of $\|\mathbf{A}|_S\mathbf{x}\|_\infty$. In this case, one can obtain a space complexity of $O(d^2 \log n)$ bits of space instead of $O(d^2 \log^2 n)$, by storing the quadratic form $\mathbf{A}|_S^\top \mathbf{A}|_S \in \mathbb{Z}^{d\times d}$. In particular, we obtain an ellipsoid $\left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{A}|_S\mathbf{x}\|_2 \leq 1\right\}$ approximating the polytope $\left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{A}\mathbf{x}\|_\infty \leq 1\right\}$ up to a factor of $O(\sqrt{d \log n})$. If we instead just store the rows themselves in the online coreset model, we store $O(d \log(n\kappa^{\mathsf{OL}}))$ rows for a distortion of $O(\sqrt{d \log(n\kappa^{\mathsf{OL}})})$ between the polytope and ellipsoid. See also Theorem 11.4.7.

11.3 Near-optimal bounds for restricted instances

In this section, we study a restricted variant of the ℓ_∞ subspace sketch problem, and give near optimal algorithms and lower bounds, i.e., without extra $\log n$ factors.

Definition 11.3.1 (Restricted ℓ_∞ subspace sketch)**.** We define the restricted ℓ_∞ subspace sketch problem as follows. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a matrix with row norms all $\Theta(1)$. Then, we must design a data structure Q that receives a row arrival stream of $\mathbf{A}$ and answers queries $\mathbf{x} \in \mathbb{R}^d$ at the end of the stream. Furthermore, we must output

$$Q(\mathbf{x}) \leq \kappa\|\mathbf{A}\mathbf{x}\|_\infty$$

for all $\mathbf{x} \in \mathbb{R}^d$, while we must output

$$Q(\mathbf{x}) \geq \|\mathbf{A}\mathbf{x}\|_\infty$$

when $\mathbf{x}$ is an input point, i.e., $\mathbf{x}$ is one of the rows $\mathbf{a}_i$ of $\mathbf{A}$.

We show that the lower bound instance of [LWW21] is captured by this restriction, and show an algorithm that matches the lower bound up to logarithmic factors.

11.3.1 Lower bound

We will use the following lemma from coding theory.

Theorem 11.3.2 ([PTB13])**.** For any $p \geq 1$ and $d = 2^k - 1$ for some integer k, there exists a set $S \subseteq \{-1, 1\}^d$ and a constant C_p depending only on p which satisfy

- $|S| = d^p$
- For any $s, t \in S$ such that $s \neq t$, $|\langle s, t \rangle| \leq C_p \sqrt{d}$

We then have the following:

Theorem 11.3.3. Let $n = d^q$ for some integer q. Suppose that a streaming algorithm $\mathcal{A}$ solves the restricted ℓ_∞ subspace sketch problem (Definition 11.3.1) with $\kappa = c\sqrt{d}$ for some sufficiently small constant $c > 0$. Then, $\mathcal{A}$ must use $\Omega(n)$ bits of space.

Proof. We show the result by reduction from the INDEX problem (Theorem 2.2.1).

Let $S \subseteq \{-1, 1\}^d$ be the set of vectors given by Theorem 11.3.2 with $n = d^q$. Suppose that Alice has a subset $A \subseteq [n]$. Then, Alice can feed the vectors of S corresponding to her subset A, normalized to have norm $\Theta(1)$, and then pass the memory state of $\mathcal{A}$ to Bob. Now suppose that Bob has the index $b \in [n]$. Then, Bob queries the subspace sketch data structure the vector $\mathbf{x}_b \in S$ corresponding to the index b.

If $b \in A$, then we have that $Q(\mathbf{x}_b) \geq \|\mathbf{A}\mathbf{x}_b\|_\infty = \|\mathbf{x}_b\|_2^2 = \Theta(1)$. On the other hand, if $b \notin A$, then we have that

$$Q(\mathbf{x}_b) \leq \kappa \|\mathbf{A}\mathbf{x}\|_\infty \leq c\sqrt{d} \cdot \frac{1}{\Theta(\sqrt{d})} = \Theta(c).$$

Thus for c sufficiently small, Bob can distinguish whether $b \in A$ or not and thus $\mathcal{A}$ must use at least $\Omega(n)$ bits of space. □

Remark 11.3.4. By replacing our use of Theorem 11.3.2 with n random unit vectors in d dimensions, we can instead get a collection of vectors with inner product $\Theta(\sqrt{(\log n)/d})$, which leads to an $\Omega(n)$ bit lower bound for distortions better than $O(\sqrt{d/\log n})$, even for n larger than $\mathrm{poly}(d)$.

11.3.2 Upper bound

In this section, we design an algorithm solving the restricted ℓ_∞ subspace sketch problem (Definition 11.3.1). Our algorithm is given in Algorithm 3.

Algorithm 3 Restricted ℓ_∞ subspace sketch

input: $\mathbf{A} \in \mathbb{R}^{n \times d}$ in a row arrival stream of $\Theta(1)$ norm rows.
output: Coreset $S \subseteq [n]$.

```
S ← ∅
for i ∈ [n] do
    if there is no j ∈ S s.t. |⟨a_j/‖a_j‖_2, a_i/‖a_i‖_2⟩| ≥ 1/√(2d − 1) then
        S ← S ∪ {i}
```

Our analysis will use the well-known Welch bound from coding theory.

Theorem 11.3.5 (Inner product lower bound [Wel74])**.** Let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M \in \mathbb{R}^d$ be a set of M unit vectors. Let $k \geq 1$ be an integer. Then,

$$\max_{i \neq j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^{2k} \geq \frac{1}{M-1} \left[\frac{M}{\binom{d+k-1}{k}} - 1 \right]$$

Using Theorem 11.3.5, we show the following.

Theorem 11.3.6. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a matrix with rows with norm $\Theta(1)$. Then, Algorithm 3 outputs a coreset $S \subseteq [n]$ such that

$$C\sqrt{d} \|\mathbf{A}_S \mathbf{a}_i\|_\infty \geq \|\mathbf{A}\mathbf{a}_i\|_\infty$$

for all $i \in [n]$, for some $C > 0$ a sufficiently large constant. Furthermore, Algorithm 3 uses $O(d^2 \log n)$ bits of space.

Before proving Theorem 11.3.6, note that the result implies that Algorithm 3 solves the restricted ℓ_∞ subspace sketch problem, since trivially, we have that

$$C\sqrt{d} \|\mathbf{A}_S \mathbf{x}\|_\infty \leq C\sqrt{d} \|\mathbf{A}\mathbf{x}\|_\infty$$

for all $\mathbf{x} \in \mathbb{R}^d$.

Proof of Theorem 11.3.6. First note that by assuming that $\mathbf{a}_i$ are unit vectors, we only lose $\Theta(1)$ factors in the distortion parameter κ, so we make this assumption without loss of generality.

Note that the correctness guarantee is trivial from the construction of the algorithm, since every input point that doesn't satisfy line 3 is kept by the coreset. It suffices to argue the space complexity of the algorithm.

We will argue that the algorithm keeps at most $O(d)$ points in S. If we apply Theorem 11.3.5 with $k = 1$ and $M = 2d$, we get that for any set of $M = 2d$ unit vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M \in \mathbb{R}^d$,

$$\max_{i \neq j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 \geq \frac{1}{2d-1} \left[\frac{2d}{d} - 1 \right] = \frac{1}{2d-1}$$

and thus the algorithm cannot keep more than $2d - 1$ points. Storing these points only requires $O(d^2 \log n)$ bits of space. □

11.4 Applications to streaming algorithms for geometric problems in high dimensions

We now show that our ℓ_∞ subspace sketch algorithm gives the first polynomial space algorithms for many important problems in streaming computational geometry, including fundamental problems such as symmetric width, convex hull, and Löwner–John ellipsoids. Previous algorithms for these problems had an exponential dependence on d, due to reliance on ε-kernels [AHV04, AHV05]. In particular, in the high-dimensional regime of $d \geq C \log n$ for a large enough constant C, the memory bound for known results becomes larger than $\tilde{\Theta}(nd)$, and thus *there were no previously*

known nontrivial algorithms in this regime, despite the fact that algorithms that work in the high-dimensional regime have been sought after for over a decade since they were suggested by [AHV04, AHV05, Cha06, ZC06] and others.

In the following discussion, we generally assume a centrally symmetric input instance, that is, if $\mathbf{a} \in \mathbb{R}^d$ is a point in the input point set, then so is $-\mathbf{a}$. Note that for most geometric problems falling under the class of *extent measure* problems [AHV04, AS15], considering only centrally symmetric instances is without loss of generality, up to constant factor losses in the distortion. For illustration, consider the directional width problem, in which we wish to estimate $\max_{i=1}^n \langle \mathbf{a}_i, \mathbf{x} \rangle - \min_{j=1}^n \langle \mathbf{a}_j, \mathbf{x} \rangle$ for any query direction $\mathbf{x} \in \mathbb{R}^d$. One can translate the entire point set by one of the input points, say $\mathbf{a}_1$, so that $0 \in \mathbb{R}^d$ is one of the elements of the point set. This preserves the directional width. Note then that $\max_{i=1}^n \langle \mathbf{a}_i - \mathbf{a}_1, \mathbf{x} \rangle \geq \langle 0, \mathbf{x} \rangle = 0$ and $\min_{j=1}^n \langle \mathbf{a}_j - \mathbf{a}_1, \mathbf{x} \rangle \leq \langle 0, \mathbf{x} \rangle = 0$, so

$$\begin{aligned} \max_{i=1}^n \langle \mathbf{a}_i - \mathbf{a}_1, \mathbf{x} \rangle - \min_{j=1}^n \langle \mathbf{a}_j - \mathbf{a}_1, \mathbf{x} \rangle &= \left| \max_{i=1}^n \langle \mathbf{a}_i - \mathbf{a}_1, \mathbf{x} \rangle \right| + \left| \min_{j=1}^n \langle \mathbf{a}_j - \mathbf{a}_1, \mathbf{x} \rangle \right| \\ &= 2 \max_{i=1}^n |\langle \mathbf{a}_i - \mathbf{a}_1, \mathbf{x} \rangle|. \end{aligned}$$

Then for each translated point $\mathbf{a}_i - \mathbf{a}_1$, we add its negation $-(\mathbf{a}_i - \mathbf{a}_1)$, which preserves the latter value. Similar arguments apply to other problems, such as convex hull, Löwner–John ellipsoids, etc.

We show that our techniques for the streaming subspace sketch problem yield the first one pass $\mathrm{poly}(d)$ space algorithms for a wide variety of geometric approximation problems that are symmetric with respect to the origin. For these problems, the previously known techniques typically only yielded space bounds of the form $\varepsilon^{-\Theta(d)}$ for a $(1+\varepsilon)$ approximation. In contrast, we show how to obtain $\mathrm{poly}(d)$ approximations using $\mathrm{poly}(d)$ bits of space. Because our ℓ_∞ subspace sketch algorithm is online, many of our algorithms for streaming geometry are online as well, and we present results in both the row arrival streaming and online coreset models.

11.4.1 Directional width

The most direct application of our results is that of approximating the *directional height* oh a point set, which is a symmetric version of the more well-known *directional width*:

Definition 11.4.1 (Directional width and height)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. The *directional width* [AHV05] of $\mathbf{A}$ with respect to a unit vector $\mathbf{x}$ is defined to be

$$\omega(\mathbf{x}, \mathbf{A}) = \max_{i \in [n]} \langle \mathbf{x}, \mathbf{a}_i \rangle - \min_{i \in [n]} \langle \mathbf{x}, \mathbf{a}_i \rangle$$

and the *directional height* [IMGR20, MRWZ20] of $\mathbf{A}$ with respect to a unit vector $\mathbf{x}$ is defined to be

$$h(\mathbf{x}, \mathbf{A}) = \max_{i \in [n]} |\langle \mathbf{x}, \mathbf{a}_i \rangle|.$$

The definition of directional height is equivalent to an ℓ_∞ subspace sketch data structure, which means that Theorem 11.2.1 directly yields the result by providing a coreset result for the

problem in the high-dimensional regime. Furthermore, Theorem 11.3.3 improves the lower bound of [AS15] for directional width from $\Omega(d^{1/3})$ to $\Omega(d^{1/2})$. This in turn shows a lower bound of a $\Omega(d^{1/2})$ factor distortion for the convex hull estimation problem as well, which we discuss in Section 11.4.2.

By using the "peeling" technique of [AHY08], we extend this to *k-robust directional width*. We define this for centrally symmetric instances as follows:

Definition 11.4.2 (Centrally symmetric k-robust directional width [AHY08])**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a set of n points in d dimensions. We consider each row $\mathbf{a}_i \in \mathbb{R}^d$ as representing both $\mathbf{a}_i$ and $-\mathbf{a}_i$, so that the input instance is centrally symmetric. Define the *level* of $\mathbf{a} \in \mathbb{R}^d$ in the direction $\mathbf{x} \in \mathbb{R}^d$ to be

$$|\{i \in [n] : |\langle \mathbf{a}_i, \mathbf{x} \rangle| > \langle \langle \mathbf{a}, \mathbf{x} \rangle \rangle\}|$$

and let $\mathbf{A}^\ell[\mathbf{x}]$ denote the point (or row) of $\mathbf{A}$ at level ℓ[1]. Then, the k-robust directional width is defined to be

$$\mathcal{E}_k(\mathbf{x}, \mathbf{A}) \coloneqq \left|\left\langle \mathbf{A}^k[\mathbf{x}], \mathbf{x} \right\rangle\right|.$$

We now turn to showing Theorem 11.4.3, which uses the reduction of [AHY08] to turn coresets for directional width for coresets for k-robust directional width, even in one-pass streams.

Theorem 11.4.3 (k-robust directional width in polynomial space)**.** Let $\mathbf{A}$ be an $n \times d$ matrix presented in one pass over a row arrival stream. There is an algorithm $\mathcal{A}$ which maintains a coreset $S \subseteq [n]$ such that

$$\frac{1}{\Delta}\mathcal{E}_k(\mathbf{x}, \mathbf{A}) \leq \mathcal{E}_k(\mathbf{x}, \mathbf{A}|_S) \leq \mathcal{E}_k(\mathbf{x}, \mathbf{A})$$

where

- in the streaming model, $\Delta = O(\sqrt{d \log n})$, $|S| = O(kd \log n)$, and $\mathcal{A}$ uses $O(kd^2 \log^2 n)$ bits of space.
- in the online coreset model, $\Delta = O(\sqrt{d \log(n\kappa^{\mathsf{OL}})})$ and $|S| = O(kd \log(n\kappa^{\mathsf{OL}}))$.

Proof. We follow the reduction described in [AHY08]. We first discuss an algorithm running in $k+1$ passes, and then describe how this can be implemented in one pass. In $k+1$ iterations, we consider a decreasing sequence of sets of rows $[n] = S_0 \supseteq S_1 \supseteq \ldots \supseteq S_k$, where $S_{i+1} = S_i \setminus \mathcal{T}_i$, where $\mathcal{T}_i \subseteq S_i$ is a coreset for directional width as constructed by our Theorem 11.2.1. The coreset we output is then $\mathcal{T} \coloneqq \bigcup_{i=0}^{k} \mathcal{T}_i$.

We first argue correctness. Consider an arbitrary direction $\mathbf{x} \in \mathbb{R}^d$. Say that the ith iteration is *successful* if $\mathbf{A}^j[\mathbf{x}] \in \mathcal{T}_i$ for some $j \in \{0, 1, \ldots, k\}$, and *unsuccessful* otherwise. Now if $\mathbf{A}^j[\mathbf{x}] \in \mathcal{T}$ for every j, then we already have that $\mathcal{E}_k(\mathbf{x}, \mathbf{A}|_{\mathcal{T}_i}) \geq \mathcal{E}_k(\mathbf{x}, \mathbf{A})$, so we assume that there exists some j such that $\mathbf{A}^j[\mathbf{x}] \notin \mathcal{T}$. It then follows that $\mathbf{A}^j[\mathbf{x}] \notin \mathcal{T}_i$ for every iteration i. Then, let i be any iteration in which the algorithm is unsuccessful in the direction $\mathbf{x}$. Then,

$$\begin{aligned}
\mathcal{E}_0(\mathbf{x}, \mathbf{A}|_{\mathcal{T}_i}) &\geq \frac{1}{\Delta}\mathcal{E}_0(\mathbf{x}, \mathbf{A}|_{S_i}) && \text{since } \mathcal{T}_i \text{ is a coreset for } S_i \\
&\geq \frac{1}{\Delta}\mathcal{E}_j(\mathbf{x}, \mathbf{A}) && \text{since } \mathbf{A}^j[\mathbf{x}] \in S_i
\end{aligned}$$

[1] For simplicity, we assume that there is at most one vector at a given level, as done in [AHY08].

$$\geq \frac{1}{\Delta}\mathcal{E}_k(\mathbf{x}, \mathbf{A})$$

Furthermore, the $\mathbf{A}|_{\mathcal{T}_i}^{0}[\mathbf{x}]$ witnessing the above inequality is not one of the $\mathbf{A}^j[\mathbf{x}]$ of the entire dataset $\mathbf{A}$, since this iteration was unsuccessful. Thus, no matter whether the iteration is successful or unsuccessful, the final coreset $\mathcal{T}$ gains a vector $\mathbf{a} \in \mathbb{R}^d$ with $|\langle \mathbf{a}, \mathbf{x}\rangle| \geq \mathcal{E}_k(\mathbf{x}, \mathbf{A})/\Delta$ in each iteration.

To turn this into a one-algorithm, we can follow Section 2.4 of [AHY08] and maintain $k+1$ copies of our coreset data structure in parallel, where the ith data structure gets inserted with a row $\mathbf{a}$ if either the $(i-1)$th copy of the algorithm does not add $\mathbf{a}$ to $\mathcal{T}_{i-1}$. Note that our base coreset algorithm does not delete points, so we do not need to handle this as [AHY08] does. □

11.4.2 Convex hulls

A fundamental problem in computational geometry is the approximation of the convex hull of n points $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^d$. For $(1+\varepsilon)$-approximation, ε-kernels [AHV04, AHV05] give coresets of near-optimal size of $\varepsilon^{-\Theta(d)}$, even in the streaming model [Cha06, Cha16]. More recently, [BBK+18] removed the exponential dependence on d for certain beyond-worst-case instances. However, a general streaming algorithm for convex hull in $\mathrm{poly}(d, \log n)$ bits of space, even with $\mathrm{poly}(d, \log n)$ distortion, remained elusive. In the offline setting, this is possible via coresets for Löwner–John ellipsoids (see Section 3.6 of [Tod16]).

By using our coreset for ℓ_∞ subspace sketch, we obtain coresets for approximating symmetric convex hulls, with $\mathrm{poly}(d, \log n)$ bits of space and distortion. This is done by noticing that our ℓ_∞ subspace sketch result yields an online coreset for approximating a polytope defined by the intersection of the linear inequalities specified by each of the rows, and then using the fact that this linear inequality polytope is the *polar body* of the symmetric convex hulls of the corresponding rows [HW20].

The following are standard elementary facts about polars that we will need:

Lemma 11.4.4 (Polars and their properties, Exercises 1.1.14, 2.3.2 of [HW20], Section 3.5 of [Tod16])**.** Let $K \subset \mathbb{R}^d$ be a convex body and define the polar K° as

$$K^\circ := \left\{\mathbf{x} \in \mathbb{R}^d : \forall \mathbf{x}' \in \mathbb{R}^d, \langle \mathbf{x}, \mathbf{x}'\rangle \leq 1\right\}.$$

Then, the following hold:

- if $K \subset L$, then $K^\circ \supset L^\circ$
- for $r > 0$, $(r \cdot K)^\circ = r^{-1} \cdot K^\circ$
- if $0 \in \operatorname{int} K$, then $(K^\circ)^\circ = K$
- for $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\operatorname{conv}(\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\})^\circ = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{A}\mathbf{x}\|_\infty \leq 1\right\}$
- for an ellipsoid $E = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}^\top \mathbf{H}\mathbf{x} \leq 1\right\}$, E° is the ellipsoid $E^\circ = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}^\top \mathbf{H}^{-1}\mathbf{x} \leq 1\right\}$

This observation, combined with Theorem 11.2.1, yields the first polynomial space algorithm for approximating convex hulls in the worst case:

Theorem 11.4.5 (Streaming convex hulls in polynomial space)**.** Let $\mathbf{A}$ be an $n \times d$ matrix presented in one pass over a row arrival stream. There is an algorithm $\mathcal{A}$ which maintains a coreset

$S \subseteq [n]$ such that

$$\operatorname{conv}(\{\pm \mathbf{a}_i\}_{i \in S}) \subseteq \operatorname{conv}(\{\pm \mathbf{a}_i\}_{i=1}^n) \subseteq \Delta \operatorname{conv}(\{\pm \mathbf{a}_i\}_{i \in S}).$$

where

- in the streaming model, $\Delta = O(\sqrt{d \log n})$, $|S| = O(d \log n)$, and $\mathcal{A}$ uses $O(d^2 \log^2 n)$ bits of space.
- in the online coreset model, $\Delta = O(\sqrt{d \log(n\kappa^{\mathsf{OL}})})$ and $|S| = O(d \log(n\kappa^{\mathsf{OL}}))$.

Proof. Let $S \subseteq [n]$ be the coreset computed by Algorithm 2. Let $K = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{A}\mathbf{x}\|_\infty \leq 1\right\}$ and let $K_S = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{A}|_S \mathbf{x}\|_\infty \leq 1\right\}$. By Theorem 11.2.1, we are guaranteed that

$$K \subseteq K_S \subseteq \Delta K$$

for $\Delta = O(\sqrt{d \log n})$ in the row arrival streaming model and $\Delta = O(\sqrt{d \log(n\kappa^{\mathsf{OL}})})$ in the online coreset model. Then by Lemma 11.4.4, we may take polars on this chain of inclusions to conclude that

$$K^\circ \supseteq K_S^\circ \supseteq \frac{1}{\Delta} K^\circ.$$

Since $K^\circ = \operatorname{conv}(\{\pm \mathbf{a}_i\}_{i=1}^n)$ and $K_S^\circ = \operatorname{conv}(\{\pm \mathbf{a}_i\}_{i \in S})$, we conclude. □

Note that this also gives us a $O(\sqrt{d \log n})^d$-factor approximation to the volume of convex hull.

11.4.3 Löwner–John ellipsoids

We consider the problem of computing an approximate Löwner–John ellipsoid of a convex symmetric polytope, also known as the problem of minimum volume enclosing ellipsoid (MVEE). We define our notion of approximation of Löwner–John ellipsoids as follows:

Definition 11.4.6. Let $K \subseteq \mathbb{R}^d$ be a convex body and let E be the Löwner–John ellipsoid of K. We say that an ellipsoid E' is an *α-approximate Löwner–John ellipsoid* for K if

$$E \subseteq E' \subseteq \alpha E.$$

Upper bound

In the literature, there are two closely related variations to this problem (see Equations (1.1.1) and (1.1.2) of [Tod16]). In one, more common in the computational geometry community, the input data set $\mathbf{A} \in \mathbb{R}^{n \times d}$ is interpreted as the convex hull of the n rows, i.e., $K = \operatorname{conv}(\{\pm \mathbf{a}_1, \pm \mathbf{a}_2, \ldots, \pm \mathbf{a}_n\})$. In the other, more common in the optimization community, the $\mathbf{A}$ is interpreted as a set of n linear constraints, and the input polytope is $K = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{A}\mathbf{x}\|_\infty \leq 1\right\}$. As noted in Section 11.4.2, these two interpretations are polars of each other.

Our results for the streaming ℓ_∞ subspace sketch problem in Theorem 11.2.1 apply most readily to the latter interpretation, i.e. the linear inequalities interpretation, and we immediately obtain the following:

Theorem 11.4.7 (Löwner–John ellipsoids in polynomial space)**.** Let $\mathbf{A}$ be an $n \times d$ matrix presented in one pass over a row arrival stream. Define the polytope $K = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{A}\mathbf{x}\|_\infty \leq 1\right\}$. There is an algorithm $\mathcal{A}$ which maintains a coreset $S \subseteq [n]$ from which we can compute an ellipsoid E' such that

$$E' \subseteq K \subseteq \Delta E'$$

where

- in the streaming model, $\Delta = O(\sqrt{d \log n})$, $|S| = O(d \log n)$, and $\mathcal{A}$ uses $O(d^2 \log^2 n)$ bits of space.
- in the online coreset model, $\Delta = O(\sqrt{d \log(n\kappa^{\mathsf{OL}})})$ and $|S| = O(d \log(n\kappa^{\mathsf{OL}}))$.

Since $K \subseteq E \subseteq \sqrt{d}K$, E' is an $O(\Delta\sqrt{d})$-approximate Löwner–John ellipsoid.

The proof is sketched in Remark 11.2.2.

We also show that we can also get results in the convex hull interpretation, by using the fact that these two interpretations of the input matrix $\mathbf{A}$ are polars of each other. See Section 3.5 of [Tod16] for a discussion on polars and Löwner–John ellipsoids.

Using basic facts about polars (Lemma 11.4.4), we obtain the following:

Corollary 11.4.8. Let $\mathbf{A}$ be an $n \times d$ matrix presented in one pass over a row arrival stream. Define the polytope $K = \operatorname{conv}(\{\pm\mathbf{a}_1, \pm\mathbf{a}_2, \ldots, \pm\mathbf{a}_n\})$. There is an algorithm $\mathcal{A}$ which maintains a coreset $S \subseteq [n]$ from which we can compute an ellipsoid E' such that

$$E' \subseteq K \subseteq \Delta E'.$$

where

- in the streaming model, $\Delta = O(\sqrt{d \log n})$, $|S| = O(d \log n)$, and $\mathcal{A}$ uses $O(d^2 \log^2 n)$ bits of space.
- in the online coreset model, $\Delta = O(\sqrt{d \log(n\kappa^{\mathsf{OL}})})$ and $|S| = O(d \log(n\kappa^{\mathsf{OL}}))$.

Since $K \subseteq E \subseteq \sqrt{d}K$, E' is an $O(\Delta\sqrt{d})$-approximate Löwner–John ellipsoid.

Proof. We claim that we can just interpret the row arrival stream as in Theorem 11.4.7, and then simply invert the quadratic form of the ellipsoid. Using Theorem 11.4.7 and Lemma 11.4.4, we obtain some ellipsoid E such that

$$E \subseteq K^\circ \subseteq \lambda E$$

for some λ. Then, the ellipsoid with the inverse quadratic form of E is E° and satisfies

$$E^\circ \supseteq K \supseteq \frac{1}{\lambda} E^\circ$$

by Lemma 11.4.4. Scaling by λ gives the desired conclusion. □

Lower bound

In this section, we show the negative result that approximate Löwner–John ellipsoids cannot be maintained in the row arrival model with small space, if the desired approximation is much smaller than $\sqrt{d}$.

Our main result of the section is the following.

Theorem 11.4.9. Let $n = d^c$, where $c \geq 1$ is any constant integer. Suppose an algorithm $\mathcal{A}$ computes an α-approximate Löwner–John ellipsoid of any $n \times d$ matrix $\mathbf{A}$ with probability at least $2/3$, for $\alpha = c'\sqrt{d}$ for a sufficiently small constant c', in one pass over a row arrival stream. Then, $\mathcal{A}$ must use $\Omega(n)$ bits of space.

Proof. We show the result by reduction from the INDEX problem (Theorem 2.2.1).

Let S be the set constructed in Theorem 11.3.2 with p in the lemma set to c, so that $|S| = d^c = n$. Then, Alice constructs an $|A| \times d$ matrix $\mathbf{A}$ by choosing the vectors of S corresponding to the indices $i \in A$. Alice then runs the algorithm $\mathcal{A}$ on the rows of $\mathbf{A}$, then passes the working memory of the algorithm to Bob.

Let $i_* \in [n]$ be the index given to Bob. We claim that Bob can then figure out whether $i_* \in A$ or not using this working memory. Let $\mathbf{b} \in S$ be the vector in S indexed by i_*. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{d-1} \in \mathbb{R}^d$ be an orthonormal basis to the orthogonal complement $\left\{\mathbf{x} \in \mathbb{R}^d : \langle \mathbf{b}, \mathbf{x} \rangle = 0\right\}$ of $\mathbf{b}$. Then, Bob inserts the following rows into the working memory of $\mathcal{A}$:

- $4(d-1)$ rows $\pm d \cdot \mathbf{u}_i \pm \mathbf{b}/\sqrt{d}$ for $i \in [d-1]$
- $2(d-1)$ rows $\pm R \cdot \mathbf{u}_i$ for $i \in [d-1]$, for a large $R = \text{poly}(d)$ to be determined

Bob will then report that $i_* \in A$ if and only if $\mathbf{b}$ belongs to the α-approximate Löwner–John ellipsoid that is output by $\mathcal{A}$.

If $i_* \in A$, then it is obvious that $\mathbf{b}$ must be in the Löwner–John ellipsoid, so suppose that $i_* \notin A$. By rotating, we assume without loss of generality that $\mathbf{b} = \sqrt{d} \cdot \mathbf{e}_1$ and $\mathbf{u}_i = \mathbf{e}_{i+1}$ for $i \in [d-1]$. Now consider the exact Löwner–John ellipsoid E of the input dataset including all rows added by both Alice and Bob, and let g be the largest magnitude achieved by a point $\mathbf{g} \in E$, in the direction of $\mathbf{b}$. Suppose for contradiction that $g \geq 2$.

Replacing Alice's points by a box. Let $V = \{\pm d \cdot \mathbf{u}_i \pm \mathbf{b}/\sqrt{d} : i \in [d-1]\}$ be the rows added by Bob. We first show that the Löwner–John ellipsoid does not change if we remove all of Alice's points, by showing that the convex hull of V must contain Alice's points. Note that Alice's points all have ℓ_2 norm at most $\sqrt{d}$ and $\mathbf{e}_1$ component at most 1. If $\mathbf{x}$ is any point with $\mathbf{x}_1 = 0$, then $d \cdot \mathbf{x}/\|\mathbf{x}\|_1$ is a convex combination of $\pm d \cdot \mathbf{e}_i$. The ℓ_2 norm of this point is at least

$$\left\| d \cdot \frac{\mathbf{x}}{\|\mathbf{x}\|_1} \right\|_2 = d\frac{\|\mathbf{x}\|_2}{\|\mathbf{x}\|_1} \geq \sqrt{d}.$$

Thus, applying this to any of Alice's points $\mathbf{x}$ with the first coordinate removed, these points must lie in $\text{conv}(V)$, since $d \cdot \mathbf{x}/\|\mathbf{x}\|_1 \in \text{conv}(V)$ is a vector in the same direction with a greater magnitude as $\mathbf{x}$ that is also in $\text{conv}(V)$. It follows that $\mathbf{x}$ must lie in $\text{conv}(V)$ as well.

Reduction to a two-dimensional ellipse. Note that V is symmetric with respect to flipping signs on coordinates, and so is $\text{conv}(V)$, and thus so is the Löwner–John ellipsoid of $\text{conv}(V)$. Now let $\mathbf{v} \in V$ be any vertex of $\text{conv}(V)$, and consider the two-dimensional ellipse E' obtained by intersecting E with the plane spanned by $\mathbf{v}$, and $\mathbf{b}$. Write this ellipse as $E' = \{(x, y) : ax^2 + by^2 \leq 1\}$, where the cross term disappears due to symmetry of the ellipse. We will think of the x direction as the $\mathbf{b}$ direction, and refer to this coordinate system as the E' coordinate system.

Bounds on the ellipse. If the V vertices do not contact the ellipsoid, then they can be removed from the Löwner–John ellipsoid, which means that the Löwner–John ellipsoid would be degenerate since it would lie on a $(d-1)$-dimensional space. Thus, the vertices of V must contact the ellipsoid. Similarly, for large enough R, the points $\pm R \cdot \mathbf{u}_i$ must also contact the ellipsoid, since otherwise removing them would lead to a John ellipse of bounded radius. Note that the ℓ_2 diameter of $\mathrm{conv}(V)$ is at most $O(d)$, so a Löwner–John ellipsoid of $\mathrm{conv}(V)$ would have radius at most $O(d^{3/2})$, which means the above holds when R is chosen to be larger than some $O(d^{3/2})$. Also, we have a point $(g, 0) \in E'$ for $g > 2$. Then, we have that $a = 1/g^2$ and $b = 1/R^2$ so that

$$E' = \left\{(x, y) : \frac{1}{g^2}x^2 + \frac{1}{R^2}y^2 \leq 1\right\}.$$

Note that the V vertices have the form $(\pm d, \pm 1)$ in the E' coordinate system. However, we then have that

$$\frac{1}{g^2} + \frac{d^2}{R^2} \ll 1$$

so they in fact cannot contact the ellipse. We conclude that $g > 2$ is impossible.

Finally, even if we have an α-approximate Löwner–John ellipsoid, $\mathbf{b}$ will still not be contained in the ellipsoid, so the Bob will still output the correct answer. □

Remark 11.4.10. By replacing our use of Theorem 11.3.2 with n random unit vectors in d dimensions, we can instead get a collection of vectors with inner product $\Theta(\sqrt{(\log n)/d})$, which leads to an $\Omega(n)$ bit lower bound for distortions better than $O(\sqrt{d/\log n})$, even for n larger than $\mathrm{poly}(d)$.

Remark 11.4.11. Note that the above lower bound holds even if Alice and Bob compute a general convex body K such that

$$E \subseteq K \subseteq \alpha E,$$

since such a K can still detect whether Bob's point is in Alice's point set or not.

11.4.4 Volume maximization

We next consider the problem of selecting k rows that approximately maximizes the volume of the parallelepiped spanned by the rows, known as *volume maximization*, or maximum a posteriori (MAP) inference of determinantal point processes (DPPs) [BKLZ20]. Relative error guarantees for this problem have been studied by [IMGR19, IMGR20, MRWZ20], culminating in the following:

Theorem 11.4.12 (Streaming volume maximization, Theorem 1.9 of [MRWZ20])**.** Let $\mathbf{A} \in \mathbb{Z}^{n \times d}$ have entries bounded by $\mathrm{poly}(n)$ and $k \geq 1$. Let $C \in [1, (\log n)/k]$. There is a one-pass streaming algorithm that computes a subset $S \subseteq [n]$ of k points such that

$$\mathbf{Pr}\left\{O(Ck)^{k/2}\,\mathrm{Vol}(\mathbf{A}|_S) \geq \mathrm{Vol}(\mathbf{A}|_{S_*})\right\} \geq \frac{2}{3}$$

where $\mathrm{Vol}(\mathbf{A}|_S)$ is the volume of the parallelepiped spanned by the rows $\mathbf{A}|_S$ indexed by S and $\mathbf{A}|_{S_*}$ is a set of k rows that maximizes the volume. The algorithm uses $O(n^{O(1/C)}d)$ bits of space.

This result is obtained by combining coresets for volume maximization [IMGR19] with streaming ε-kernels for directional width [Cha06]. Note that even when $C = (\log n)/k$, the space complexity is $\exp(O(k))d$ and thus still exponential in k. By replacing ε-kernels for directional width with our ℓ_∞ subspace sketch result, we obtain the first relative error polynomial space algorithms for volume maximization[2].

Theorem 11.4.13 (Streaming volume maximization in polynomial space)**.** Let $\mathbf{A} \in \mathbb{Z}^{n\times d}$ with entries bounded by $\mathrm{poly}(n)$ and $k \geq 1$. Let $1 < C < (\log n)/k$ and $r = (\log n)/C$. There is a one-pass streaming algorithm that computes a subset $S \subseteq [n]$ of k points such that

$$\mathbf{Pr}\big\{O(r^2Ck\log^2 n)^{k/2}\,\mathrm{Vol}(\mathbf{A}|_S) \geq \mathrm{Vol}(\mathbf{A}|_{S_*})\big\} \geq \frac{2}{3}$$

where $\mathrm{Vol}(\mathbf{A}|_S)$ is the volume of the parallelepiped spanned by the rows $\mathbf{A}|_S$ indexed by S and $\mathbf{A}|_{S_*}$ is a set of k rows that maximizes the volume. The algorithm uses $O(rd\log^2 n)$ bits of space.

If only the indices (rather than the d-dimensional rows) are required, there is an algorithm using $O(k^2\log^3 n)$ bits of space with $O(k\log n)^k$ distortion.

Proof. The following is shown in [MRWZ20]:

Lemma 11.4.14 (Lemmas 5.13, 5.14 of [MRWZ20])**.** Let $r = \Theta((\log n)/C)$. Let $\mathbf{G} \in \mathbb{R}^{d\times r}$ have each entry drawn i.i.d. from the Gaussian distribution $\mathcal{N}(0, 1/r)$. Then:

- The volume of the optimal k-subset $\mathbf{A}|_{S_*}$ satisfies

$$\mathbf{Pr}\big\{2^k\,\mathrm{Vol}(\mathbf{AG}|_{S_*}) \geq \mathrm{Vol}(\mathbf{A}|_{S_*})\big\} \geq \frac{9}{10}$$

- $$\mathbf{Pr}\left\{\forall S \in \binom{[n]}{k}, \mathrm{Vol}(\mathbf{AG}|_S) \leq O(Ck)^{k/2}\,\mathrm{Vol}(\mathbf{A}|_S)\right\} \geq \frac{9}{10}$$

Thus, up to a $O(Ck)^{k/2}$ factor loss in the approximation factor, we may replace $\mathbf{A}$ by the $n \times r$ matrix $\mathbf{AG}$. Now applying the observation of Section 11.4.1, we can obtain a directional height coreset for $\mathbf{AG}$ from our Theorem 11.2.1, which produces a set $T \subseteq [n]$ of size $|T| \leq O(r\log n)$ with distortion $\kappa = O(\sqrt{r\log n})$. Observation 5.9 of [MRWZ20] and Lemma 3.3 of [IMGR19] then shows that the maximum volume subset of the directional height coreset approximates the maximum volume subset of $\mathbf{A}$ up to a factor of $\kappa^{2k} = (\kappa^4)^{k/2}$. The total approximation factor is thus $O(\kappa^4Ck)^{k/2}$, while the space complexity is $O(dr\log n + |T|d\log n) = O(|T|d\log n)$ for storing $\mathbf{G}$ and the coreset.

If we only need to output the indices of the coreset, then we can first replace the Gaussian matrix $\mathbf{G}$ with a subspace embedding with a small seed as in, e.g., [KMN11]. Then by setting $\delta = \exp(-\Theta(k^2\log n))$, we have an $O(\log\frac{1}{\delta}) \times d$ matrix $\mathbf{S}$ such that, with probability at least $1 - n^{-k}$, for any fixed $d \times k$ matrix $\mathbf{R}$, $\|\mathbf{SRx}\|_2 = \Theta(1)\|\mathbf{Rx}\|_2$ for all $\mathbf{x} \in \mathbb{R}^k$, where $\mathbf{S}$ can be generated from a seed of length $\tilde{O}(\log k + \log\frac{1}{\delta})$. In particular, $\|\mathbf{SR}\|_2 = \Theta(1)\|\mathbf{R}\|_2$ under this

[2] The algorithm of [BKLZ20] has polynomial space as well, but has an additive error guarantee

event. We now consider any subset $S \in \binom{[n]}{k}$. Then by the same reasoning as in [MRWZ20], we have that the volume spanned by $\mathbf{A}|_S = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ written in the SVD is

$$\sqrt{\det(\mathbf{A}|_S\mathbf{A}|_S^\top)} = \sqrt{\det(\mathbf{\Sigma}^2)}$$

while the volume of the embedded matrix $\mathbf{A}|_S\mathbf{S}^\top$ is at most

$$\sqrt{\det(\mathbf{A}|_S\mathbf{S}^\top\mathbf{S}\mathbf{A}|_S^\top)} = \sqrt{\det(\mathbf{\Sigma}\mathbf{V}^\top\mathbf{S}^\top\mathbf{S}\mathbf{V}\mathbf{\Sigma})} \leq \|\mathbf{S}\mathbf{V}\|_2^k\sqrt{\det(\mathbf{\Sigma}^2)}.$$

Conditioned on the operator norm preservation of $\mathbf{V}$ by $\mathbf{S}$ for all $\binom{n}{k}$ subsets S, this is at most

$$O(\|\mathbf{V}\|_2)^k\sqrt{\det(\mathbf{\Sigma}^2)} \leq \exp(O(k))\sqrt{\det(\mathbf{\Sigma}^2)}.$$

The fact that the volume of the maximal volume subset S_* does not shrink by more than $\exp(O(k))$ follows similarly as in [MRWZ20]. Then, we repeat the reasoning as before with $r = O(k^2 \log n)$ on using the directional height coresets, so that our total space usage is just the seed length for the subspace embedding and the storage of the directional height coreset, which is $O(|T| \log n) = O(r \log^2 n) = O(k^2 \log^3 n)$. The total distortion is $O(\sqrt{r \log n})^{2k} = O(k \log n)^k$. □

11.4.5 Minimum-width spherical shell

Our next application is the problem of approximating the *spherical shell* of minimum width which encloses a set of points. Formally, a spherical shell centered at $\mathbf{c} \in \mathbb{R}^d$ with inner radius r and outer radius R is $\sigma(\mathbf{c}, r, R) \coloneqq \left\{\mathbf{x} \in \mathbb{R}^d : r \leq \|\mathbf{x} - \mathbf{c}\|_2 \leq R\right\}$, and we seek relative error approximations to $R - r$. This problem has received much attention in the computational geometry literature [AS98, AAHS99, Cha02, Cha06]. We give a high-dimensional streaming algorithm for this problem in Theorem 11.4.15. Our proof of Theorem 11.4.15 for minimum-width spherical shells requires additional care to handle general instances, rather than just centrally symmetric instances.

Theorem 11.4.15 (Minimum width spherical shell in polynomial space)**.** Let $\mathbf{A}$ be an $n \times d$ matrix presented in one pass over a row arrival stream. There is an algorithm $\mathcal{A}$ which maintains a coreset $S \subseteq [n]$ from which we can compute find a center $\hat{\mathbf{c}}$, inner radius $\hat{r}$ and outer radius $\hat{R}$ such that $\sigma(\hat{\mathbf{c}}, \hat{r}, \hat{R}) \supseteq \{\mathbf{a}_i\}_{i=1}^n$ and

$$\hat{R} - \hat{r} \leq \Delta^{3/2} \min_{\sigma(\mathbf{c},r,R) \supseteq \{\mathbf{a}_i\}_{i=1}^n} R - r$$

where

- in the streaming model, $\Delta = O(\sqrt{d \log n})$, $|S| = O(d \log n)$, and $\mathcal{A}$ uses $O(d^2 \log^2 n)$ bits of space.
- in the online coreset model, $\Delta = O(\sqrt{d \log(n\kappa^{\mathsf{OL}})})$ and $|S| = O(d \log(n\kappa^{\mathsf{OL}}))$.

Proof. We will always store the first point $\mathbf{a}_1$ in order to translate our input instance to the origin. Now for each $i \in [n]$, define the vector $\mathbf{b}_i \in \mathbb{R}^{d+1}$ by setting the first d coordinates to

be $-2(\mathbf{a}_i - \mathbf{a}_1)$ and the last coordinate to be $\|\mathbf{a}_i - \mathbf{a}_1\|_2^2$. Given $\mathbf{a}_1$, we can always compute $\mathbf{b}_i$ if we have stored $\mathbf{a}_i$. Similarly, define $\mathbf{b}_i'' \in \mathbb{R}^{d+2}$ to be $\mathbf{b}_i$ with an additional 1 appended as the $(d+2)$th coordinate.

We now proceed by a variation on the standard linearization trick [AHV04]. Suppose that we wish to compute the width $R - r$ of the minimum-width spherical shell $\sigma(\mathbf{c}, r, R)$ containing $\{\mathbf{a}_i\}_{i=1}^n$, centered at some arbitrary $\mathbf{c} \in \mathbb{R}^d$. Note that the inner radius is given by $r = \min_{j=1}^n \|\mathbf{c} - \mathbf{a}_j\|_2$ while the outer radius is given by $R = \max_{i=1}^n \|\mathbf{c} - \mathbf{a}_i\|_2$. Now note that

$$\begin{aligned} R^2 - r^2 &= \max_{i=1}^n \|\mathbf{c} - \mathbf{a}_i\|_2^2 - \min_{j=1}^n \|\mathbf{c} - \mathbf{a}_j\|_2^2 \\ &= \max_{i=1}^n \|\mathbf{c}\|_2^2 - 2\langle \mathbf{c}, \mathbf{a}_i \rangle + \|\mathbf{a}_i\|_2^2 - \min_{j=1}^n \|\mathbf{c}\|_2^2 - 2\langle \mathbf{c}, \mathbf{a}_j \rangle + \|\mathbf{a}_j\|_2^2 \\ &= \max_{i=1}^n -2\langle \mathbf{c}, \mathbf{a}_i \rangle + \|\mathbf{a}_i\|_2^2 - \min_{j=1}^n -2\langle \mathbf{c}, \mathbf{a}_j \rangle + \|\mathbf{a}_j\|_2^2 \\ &= \max_{i=1}^n \langle \mathbf{b}_i, \mathbf{c}' \rangle - \min_{j=1}^n \langle \mathbf{b}_j, \mathbf{c}' \rangle \end{aligned}$$

where $\mathbf{c}' = [\mathbf{c}, 1]$. Then by the discussion in Section 11.4, our ℓ_∞ subspace sketch coreset result of Theorem 11.2.1 can estimate this up to a factor of Δ both in the row arrival streaming model and the online coreset model. Similarly, note that

$$\begin{aligned} R^2 &= \max_{i=1}^n \|\mathbf{c} - \mathbf{a}_i\|_2^2 \\ &= \max_{i=1}^n \|\mathbf{c}\|_2^2 - 2\langle \mathbf{c}, \mathbf{a}_i \rangle + \|\mathbf{a}_i\|_2^2 \\ &= \max_{i=1}^n \langle \mathbf{b}_i'', \mathbf{c}'' \rangle \end{aligned}$$

where $\mathbf{c}'' = [\mathbf{c}, 1, \|\mathbf{c}\|_2^2]$. We estimate this quantity up to a Δ factor using Theorem 11.2.1 as well. Note then that

$$R - r = \frac{R^2 - r^2}{R + r} = \Theta\left(\frac{R^2 - r^2}{R}\right)$$

and we obtain a Δ factor approximation to the numerator, while we obtain a $\sqrt{\Delta}$ factor approximation to the denominator. Thus, overall, we obtain a $\Delta^{3/2}$-approximation to the entire quantity. $\square$

Chapter 12

Applications: active ℓ_p linear regression [MMWY22, WY23a]

12.1 Active ℓ_p linear regression

One of the motivating problems for the study of subspace embeddings is the least squares linear regression problem [DMM06a, Sar06] and, more generally, the ℓ_p linear regression problem, in which we wish to approximately solve

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p.$$

When one takes a sampling-based approach to constructing the subspace embedding for the matrix $[\mathbf{A}\ \mathbf{b}]$, including many of the algorithms previously, then the final solution only depends on very few coordinates of the target vector $\mathbf{b}$, namely the r rows sampled by the subspace embedding matrix $\mathbf{S}$. Thus, this gives hope for an algorithm which minimizes the number of entries of the target vector $\mathbf{b}$ it has to read, which is a problem known as *active learning* or *active regression*.

Definition 12.1.1 (Active ℓ_p linear regression)**.** An active ℓ_p linear regression algorithm has query complexity r if, given $\mathbf{A}\in\mathbb{R}^{n\times d}$ and query access to the entries of $\mathbf{b}\in\mathbb{R}^n$, it reads r entries of the vectors and outputs $\hat{\mathbf{x}}\in\mathbb{R}^d$ such that

$$\|\mathbf{A}\hat{\mathbf{x}}-\mathbf{b}\|_p^p \le (1+\varepsilon)\min_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p.$$

Our goal is to minimize the query complexity r.

Such an algorithm has significant value in practice, since label acquisition can oftentimes require significantly more resources than the training features. Indeed, viewing a single entry of $\mathbf{b}$ might require running a survey, physical experiment, or time-intensive computer simulation [SWMW89, Puk06].

Unfortunately, the previous approach of constructing sampling-based subspace embeddings for $[\mathbf{A}\ \mathbf{b}]$ does not immediately yield active regression algorithms, since the sampling probabilities will depend on $\mathbf{b}$, and thus the algorithm needs to read all entries of $\mathbf{b}$. A natural idea to overcome this problem is to take the sampling probabilities to only depend on $\mathbf{A}$ but not $\mathbf{b}$, by, for example,

using the ℓ_p Lewis weights of the matrix $\mathbf{A}$ *without* including $\mathbf{b}$. However, the correctness of this algorithm is then no longer clear, as we no longer have the subspace embedding guarantee which includes $\mathbf{b}$. Nonetheless, prior work has shown that this approach in fact *does* yield efficient active regression algorithms in several cases.

For the most important case of $p = 2$, the work of [CP19] obtained an optimal bound of $\Theta(\varepsilon^{-1}d)$, which notably removes a $\log d$ factor that is inherent in sampling-bashed approaches, by using spectral sparsifiers developed in [LS15]. For perhaps the next most important case of $p = 1$, which corresponds to *least absolute deviations regression*, two simultaneous works [CD21, PPP21] showed that a sampling-based approach which takes the sampling probabilities to be the ℓ_1 Lewis weights of $\mathbf{A}$ (without appending $\mathbf{b}$) yields an upper bound of $O(\varepsilon^{-2}d\log(d/\varepsilon))$, with a nearly matching lower bound of $\Omega(\varepsilon^{-2}d)$. However, besides these two special cases, the true sample complexity of active ℓ_p linear regression is far from settled. The only other known bound is an upper bound of $\tilde{O}(\varepsilon^{-2}d^2\log(d/\varepsilon))$ due to [CD21] for $1 < p < 2$. This leads to the following question:

Question 12.1.2. What is the query complexity of active ℓ_p linear regression for $p \neq 1, 2$?

In two works [MMWY22, WY23a], we obtain nearly optimal solutions to Question 12.1.2 for the entire range of $0 < p < \infty$.

Theorem 12.1.3 (Nearly optimal active ℓ_p linear regression, [MMWY22, WY23a])**.** There is an active ℓ_p linear regression algorithm (see Definition 12.1.1) with query complexity at most r with probability at least $99/100$, where

$$r = \begin{cases} \tilde{O}(\varepsilon^{-2}d) & 0 < p < 1 \\ \tilde{O}(\varepsilon^{-1}d) & 1 < p < 2 \\ \tilde{O}(\varepsilon^{1-p}d^{p/2}) & 2 < p < \infty \end{cases}$$

Furthermore, for any active ℓ_p linear regression algorithm which succeeds with probability at least $99/100$, its query complexity r must be at least

$$r = \begin{cases} \Omega(\varepsilon^{-2}d) & 0 < p < 1 \\ \Omega(\varepsilon^{-1}d) & 1 < p < 2 \\ \Omega(\varepsilon^{1-p}d^{p/2}) & 2 < p < \infty \end{cases}$$

Notably, we show that there is a sharp phase transition in the behavior of the query complexity at $p = 1$, where $p > 1$ admits an upper bound of $\tilde{O}(\varepsilon^{-1}d)$ queries while $p \leq 1$ requires $\Omega(\varepsilon^{-2}d)$ queries. We also note that while we have stated Theorem 12.1.3 for constant probability, we will in general obtain an algorithm with failure probability $1 - \delta$ where the number of samples scales as $(\log \frac{1}{\delta})^2$. It is an interesting open question to reduce this dependence to linear in $\log \frac{1}{\delta}$.

The algorithm used in the proof of Theorem 12.1.3 is similar to prior ideas, and we simply take the approach of sampling rows of $\mathbf{A}$ and entries of $\mathbf{b}$ proportionally to the ℓ_p Lewis weights of $\mathbf{A}$. However, a tight analysis of this algorithm requires significantly new ideas, and in particular, we introduce two key ingredients. The first is the observation that, while the ℓ_p Lewis weights do not upper bound the sensitivity of the entries of $\mathbf{b}$, any entry $\mathbf{b}_i$ of $\mathbf{b}$ can be classified as either "too big" or "not too big" by comparing $\mathbf{b}_i$ to the ith sensitivity (see Definition 6.1.2) $\boldsymbol{\sigma}_i(\mathbf{A})$. For

entries which are "too big", we show that the loss contribution $|[\mathbf{Ax}-\mathbf{b}](i)|^p = |\langle \mathbf{a}_i, \mathbf{x}\rangle - \mathbf{b}|^p$ on the ith coordinate is dominated by $\mathbf{b}_i$ for any nearly optimal solution $\mathbf{x}$, and thus this entry can be effectively ignored. On the other hand, for entries $\mathbf{b}_i$ which are "not too big", the sensitivity of $\mathbf{b}_i$ is bounded by $\boldsymbol{\sigma}_i(\mathbf{A})$, which allows an appropriate modification of the chaining arguments for Lewis weight sampling [BLM89, LT91, SZ01] to go through. The idea above is sufficient for nearly optimal bounds for $p < 1$, but for $p > 1$, this still leads to a result that is off by a single ε factor. In order to further optimize our bounds, we additionally introduce a second novel technique which allows us to reduce the ε dependence by using the strict convexity of the ℓ_p loss for $p > 1$. This is done by noting that for $p > 1$, nearly optimal solutions must necessarily be close to the optimal solution, and this fact can be used to improve the sampling error analysis.

In Chapter 13, we will discuss various applications of these ideas developed in [WY24a]. Our work has also been used to obtain online active regression algorithms in follow-up work of [CLS22].

12.2 Constant factor solution

Our first task is to establish that the "sample-and-solve" algorithm (Algorithm 4) gives a constant factor solution to the active ℓ_p linear regression problem. Such a result has already been shown in prior work such as [DDH+09] and is based on a simple analysis that only needs the property that the ℓ_p sampling matrix $\mathbf{S}$ is an ℓ_p subspace embedding.

Algorithm 4 Constant factor ℓ_p regression

input: Matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, measurement vector $\mathbf{b} \in \mathbb{R}^n$.
output: Approximate solution $\tilde{\mathbf{x}} \in \mathbb{R}^d$ to $\min_{\mathbf{x}} \|\mathbf{Ax}-\mathbf{b}\|_p$.

1: Let $\mathbf{S} \in \mathbb{R}^{m\times n}$ be an $1/2$-approximate ℓ_p subspace embedding for $\mathbf{A}$ (Theorem 6.5.1).
2: **return** $\tilde{\mathbf{x}}$ with $\|\mathbf{SA}\tilde{\mathbf{x}}-\mathbf{Sb}\|_p \le (1+\eta)\cdot \min_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{SAx}-\mathbf{Sb}\|_p$ for $\eta \ge 0$.

Remark 12.2.1. Running Algorithm 4 only requires querying m entries of $\mathbf{b}$ in order to construct the vector $\mathbf{Sb}$. Also note that in Line 2 of the algorithm, we would have $\eta = 0$ if an exact minimizer of the subsampled regression problem $\min_{\mathbf{x}}\|\mathbf{SAx}-\mathbf{Sb}\|_p^p$ was obtained. To allow for the use of approximation algorithms in implementing Line 2, we state the method for a general $\eta \ge 0$.

We first give an algorithm which works with constant probability, and then show in Section 12.2.1 how to boost the probability to $1-\delta$ for any $\delta \in (0,1)$, while incurring an $O(\log(1/\delta))$ factor overhead in our sample complexity.

Theorem 12.2.2 (Constant factor approximation)**.** For $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^n$, and $0 < p < \infty$, let $\mathsf{OPT} = \min_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{Ax}-\mathbf{b}\|_p$. For any $\delta \in (0,1]$, if $\tilde{\mathbf{x}}$ is the output of Algorithm 4, then with probability at least $1-\delta$,

$$\|\mathbf{A}\tilde{\mathbf{x}}-\mathbf{b}\|_p \le 2^{2\max\{0,1/p-1\}+1+1/p}(3+\eta)/\delta^{1/p}\cdot \mathsf{OPT}\,.$$

When δ is constant (e.g., $\delta = 1/100$) and $(1+\eta)$ is constant (e.g., $\eta = 0$) then $\|\mathbf{A}\tilde{\mathbf{x}}-\mathbf{b}\|_p \le C\cdot\mathsf{OPT}$ for constant C.

Proof. Let $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p$. By triangle inequality for $p \geq 1$ or subadditivity and approximate triangle inequality (Fact 2.1.1) for $p \in (0,1)$,

$$\|\mathbf{A}\tilde{\mathbf{x}}-\mathbf{b}\|_p \leq 2^{\max\{0,1/p-1\}}(\|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p + \|\mathbf{A}\tilde{\mathbf{x}}-\mathbf{A}\mathbf{x}^*\|_p) = 2^{\max\{0,1/p-1\}}(\mathsf{OPT}+\|\mathbf{A}\tilde{\mathbf{x}}-\mathbf{A}\mathbf{x}^*\|_p).$$

Applying the subspace embedding property of Theorem 6.5.1 with $\varepsilon = 1/2$ and failure probability $\delta/2$, we conclude that, with probability at least $1-\delta/2$,

$$\|\mathbf{A}\tilde{\mathbf{x}}-\mathbf{b}\|_p \leq 2^{\max\{0,1/p-1\}}(\mathsf{OPT}+2\|\mathbf{S}\mathbf{A}\tilde{\mathbf{x}}-\mathbf{S}\mathbf{A}\mathbf{x}^*\|_p).$$

By similar reasoning, we have $(\|\mathbf{S}\mathbf{A}\tilde{\mathbf{x}}-\mathbf{S}\mathbf{A}\mathbf{x}^*\|_p) \leq 2^{\max\{0,1/p-1\}}(\|\mathbf{S}\mathbf{A}\tilde{\mathbf{x}}-\mathbf{S}\mathbf{b}\|_p + \|\mathbf{S}\mathbf{A}\mathbf{x}^*-\mathbf{S}\mathbf{b}\|_p)$. We know that $\|\mathbf{S}\mathbf{A}\tilde{\mathbf{x}}-\mathbf{S}\mathbf{b}\|_p \leq (1+\eta)\cdot\min_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{S}\mathbf{A}\mathbf{x}-\mathbf{S}\mathbf{b}\|_p \leq (1+\eta)\cdot\|\mathbf{S}\mathbf{A}\mathbf{x}^*-\mathbf{S}\mathbf{b}\|_p$, so we conclude that

$$\|\mathbf{S}\mathbf{A}\tilde{\mathbf{x}}-\mathbf{S}\mathbf{A}\mathbf{x}^*\|_p \leq 2^{\max\{0,1/p-1\}}(2+\eta)\|\mathbf{S}\mathbf{A}\mathbf{x}^*-\mathbf{S}\mathbf{b}\|_p.$$

Finally, note that $\mathbf{E}[\|\mathbf{S}\mathbf{A}\mathbf{x}^*-\mathbf{S}\mathbf{b}\|_p^p] = \mathsf{OPT}^p$ for ℓ_p sampling matrices $\mathbf{S}$. Then by Markov's inequality, with probability $\geq 1-\delta/2$, $\|\mathbf{S}\mathbf{A}\mathbf{x}^*-\mathbf{S}\mathbf{b}\|_p^p \leq \mathsf{OPT}^p/(\delta/2)$ and so $\|\mathbf{S}\mathbf{A}\mathbf{x}^*-\mathbf{S}\mathbf{b}\|_p \leq \mathsf{OPT}/(\delta/2)^{1/p}$. Combining all these bounds we have that with probability $1-\delta$,

$$\begin{aligned}\|\mathbf{A}\tilde{\mathbf{x}}-\mathbf{b}\|_p &\leq 2^{\max\{0,1/p-1\}}\left(\mathsf{OPT}+2\cdot 2^{\max\{0,1/p-1\}}(2+\eta)\cdot 2^{1/p}\,\mathsf{OPT}/\delta^{1/p}\right)\\ &\leq 2^{2\max\{0,1/p-1\}+1+1/p}(3+\eta)\,\mathsf{OPT}/\delta^{1/p}.\end{aligned}$$

□

12.2.1 Probability boosting for constant factor approximation

We now show a boosting step for our constant factor approximation algorithm (Algorithm 4), described in Algorithm 5. If we repeat the constant factor approximation algorithm with success probability $99/100$ for a total of $O(\log(1/\delta))$ times, then via a standard Chernoff bound, with probability at least $1-\delta$, at least $9/10$ of the computed $\mathbf{x}_c$ will satisfy the guarantee of Theorem 12.2.2 – i.e., that $\|\mathbf{A}\mathbf{x}_c-\mathbf{b}\|_p = O(\mathsf{OPT})$. Thus, we just need to identify one of these good solutions, which Algorithm 5 does, deterministically, and without reading any entries of $\mathbf{b}$. The approach simply computes pairwise distances between solutions and returns any solution with a relatively low distance to at least $1/2$ of the other solutions. For later use, we state the result in terms of a general error measure $\|\cdot\|$ which satisfies an approximate triangle inequality (for example, $\|\cdot\|_p$ for $p \in (0,1)$ satisfies an approximate triangle inequality with constant $2^{1/p-1}$ by Fact 2.1.1).

Theorem 12.2.3 (Constant factor $\|\cdot\|$ regression – success boosting)**.** Consider $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^n$, and an error measure $\|\cdot\|$ which satisfies an approximate triangle inequality, that is, there exists a constant $\kappa \geq 1$ such that $\|\mathbf{y}_1+\mathbf{y}_2\| \leq \kappa(\|\mathbf{y}_1\|+\|\mathbf{y}_2\|)$ for any two vectors $\mathbf{y}_1,\mathbf{y}_2 \in \mathbb{R}^n$. Let $\mathsf{OPT} = \min_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|$. Given a set of solution vectors $\mathbf{x}_1,\ldots,\mathbf{x}_\ell \in \mathbb{R}^d$ where $\|\mathbf{A}\mathbf{x}_i-\mathbf{b}\| \leq \alpha\cdot\mathsf{OPT}$ for at least $9/10\cdot\ell$ of the vectors, Algorithm 5 identifies $\mathbf{x}_i$ with $\|\mathbf{A}\mathbf{x}_i-\mathbf{b}\| \leq (\kappa\alpha+2\kappa^3(\alpha+1))\cdot\mathsf{OPT}$, without querying any entries of $\mathbf{b}$.

Proof. Let $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|$. Call $\mathbf{x}_i$ *good* if $\|\mathbf{A}\mathbf{x}_i-\mathbf{b}\| \leq \alpha\cdot\mathsf{OPT}$. By approximate triangle inequality , for any good $\mathbf{x}_i$,

$$\|\mathbf{A}\mathbf{x}_i-\mathbf{A}\mathbf{x}^*\| \leq \kappa(\|\mathbf{A}\mathbf{x}_i-\mathbf{b}\|_p + \|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p) = \kappa(\alpha+1)\cdot\mathsf{OPT}\,.$$

Algorithm 5 Probability Boosting for Constant Factor Active ℓ_p Regression

input: ℓ candidate solutions $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ with at least $9/10 \cdot \ell$ satisfying $\|\mathbf{A}\mathbf{x}_i - \mathbf{b}\|_p \leq \alpha \min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p$.

output: Approximate solution $\tilde{\mathbf{x}} \in \mathbb{R}^d$ to $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p$.

1: Let $\mathbf{d} \in \mathbb{R}^{\ell^2}$ contain all pairwise distances $\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\|_p$ (over ordered pairs (i, j)) sorted in increasing order. Let $\tau = \mathbf{d}(\lfloor \ell^2 \cdot 8/10 \rfloor)$ be the 80^{th} percentile distance.

2: Return any $\mathbf{x}_i$ such that $\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\|_p \leq \tau$ for at least $1/2 \cdot \ell$ vectors $\mathbf{x}_j$.

Thus, again via approximate triangle inequality, for any good $\mathbf{x}_i, \mathbf{x}_j$,

$$\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\| \leq \kappa(\kappa(\alpha+1) \cdot \mathsf{OPT} + \kappa(\alpha+1) \cdot \mathsf{OPT}) = 2\kappa^2(\alpha+1) \cdot \mathsf{OPT}\,.$$

Thus, for the pairwise distance vector $\mathbf{d} \in \mathbb{R}^{\ell^2}$ computed in line 1 of Algorithm 5, at least $(9/10)^2 \cdot \ell^2 \geq 8/10 \cdot \ell^2$ of the distances will be upper bounded by $2\kappa^2(\alpha+1) \cdot \mathsf{OPT}$. Thus, the threshold τ computed in Line 1, which is the 80^{th} percentile of the distances, gives a lower bound $\tau \leq 2\kappa^2(\alpha+1) \cdot \mathsf{OPT}$. In Line 2, we return any $\mathbf{x}_i$ with $\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\| \leq \tau$ for at least $1/2 \cdot \ell$ vectors $\mathbf{x}_j$. First observe that at least one such $\mathbf{x}_i$ must exist. Otherwise, at most $1/2 \cdot \ell^2$ of the pairwise distances would lie below τ.

Additionally, observe that since at least $9/10 \cdot \ell$ of the $\mathbf{x}_i$ are good, if $\mathbf{x}_i$ is returned, it must have $\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\| \leq \tau \leq 2\kappa^2(\alpha+1) \cdot \mathsf{OPT}$ for at least one good $\mathbf{x}_j$. Since this good $\mathbf{x}_j$ has $\|\mathbf{A}\mathbf{x}_j - \mathbf{b}\| \leq \alpha \cdot \mathsf{OPT}$, by approximate triangle inequality, the returned $\mathbf{x}_i$ must then satisfy

$$\|\mathbf{A}\mathbf{x}_i - \mathbf{b}\| \leq \kappa(\alpha + 2\kappa^2(\alpha+1)) = (\kappa\alpha + 2\kappa^3(\alpha+1)) \cdot \mathsf{OPT}\,. \qquad \square$$

12.3 $(1+\varepsilon)$ factor solution

Next, we will show the following result, which shows that the "sample-and-solve" algorithm with one-sided ℓ_p Lewis weights can achieve a nearly optimal dependence on ε, if we allow for a polynomial dependence on the failure probability $\delta \in (0, 1)$. We will separately handle the probability boosting in Section 12.3.3 to show how to achieve a $(\log \frac{1}{\delta})^2$ dependence.

Theorem 12.3.1. Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 6.1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and

$$\alpha = \begin{cases} O(\gamma)\varepsilon\delta^2 \left[(\log d)^2 \log n + \log \frac{1}{\delta}\right]^{-1} \left[\log\log \frac{1}{\varepsilon}\right]^{-2} & p < 2 \\ \frac{O(\gamma^{p/2})\varepsilon^{p-1}\delta^p}{\|\mathbf{w}\|_1^{p/2-1}} \left[(\log d)^2 \log n + \log \frac{1}{\delta}\right]^{-1} \left[\log\log \frac{1}{\varepsilon}\right]^{-p} & p > 2 \end{cases}.$$

Then, for any $\hat{\mathbf{x}} \in \mathbb{R}^d$ such that

$$\|\mathbf{S}(\mathbf{A}\hat{\mathbf{x}} - \mathbf{b})\|_p^p \leq (1+\varepsilon) \min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p,$$

we have

$$\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_p^p \leq (1 + O(\varepsilon)) \min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p.$$

Two main ingredients are necessary to prove Theorem 12.3.1. The first is a theorem which establishes that ℓ_p Lewis weight sampling preserves the cost difference $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p - \|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p^p$, which is the following theorem that we will prove in Section 12.4.

Theorem 12.3.2. Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 6.1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and

$$\alpha = \begin{cases} \dfrac{O(\gamma)\varepsilon^2}{\eta^{2/p}}\left[(\log d)^2 \log n + \log\dfrac{1}{\delta}\right]^{-1} & p < 2 \\ \dfrac{O(\gamma^{p/2})\varepsilon^p}{\eta\|\mathbf{w}\|_1^{p/2-1}}\left[(\log d)^2 \log n + \log\dfrac{1}{\delta}\right]^{-1} & p > 2 \end{cases}.$$

For each $\mathbf{x}^* \in \mathbb{R}^d$ and $\mathbf{b}^* = \mathbf{A}\mathbf{x}^* - \mathbf{b}$, with probability at least $1-\delta$,

$$\left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{b})\|_p^p - \|\mathbf{S}\mathbf{b}^*\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p - \|\mathbf{b}^*\|_p^p\right)\right| \leq \varepsilon\left(\|\mathbf{b}^*\|_p^p + \|\mathbf{S}\mathbf{b}^*\|_p^p + \frac{1}{\eta}\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p\right)$$

simultaneously for every $\mathbf{x} \in \mathbb{R}^d$.

Notably, Theorem 12.3.2 incorporates a parameter η which gives a trade-off between the closeness of $\mathbf{x}$ to the optimal solution $\mathbf{x}^*$ and the sample complexity, which will be crucial for achieving the nearly optimal dependence on ε.

The second ingredient that we need is a result which takes a near-optimality guarantee and converts it into a closeness guarantee. These are established using various measures of the strict convexity of ℓ_p norms, and are established in Section 12.3.1.

12.3.1 Closeness of nearly optimal solutions

The following lemma uses strong convexity for $p < 2$ and a Bregman divergence bound for $p > 2$ to quantify the difference between the ℓ_p norms of two vectors.

Lemma 12.3.3. For any $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^n$, we have

$$\|\mathbf{y}'\|_p^2 \geq \|\mathbf{y}\|_p^2 - 2\|\mathbf{y}\|_p^{2-p}\langle \mathbf{y}^{\circ(p-1)}, \mathbf{y}-\mathbf{y}'\rangle + \frac{p-1}{2}\|\mathbf{y}-\mathbf{y}'\|_p^2$$

if $1 < p < 2$ [BMN01, Lemma 8.1] and

$$\|\mathbf{y}'\|_p^p \geq \|\mathbf{y}\|_p^p - p\langle \mathbf{y}^{\circ(p-1)}, \mathbf{y}-\mathbf{y}'\rangle + \frac{p-1}{p2^p}\|\mathbf{y}-\mathbf{y}'\|_p^p$$

if $2 \leq p < \infty$ [AKPS19, Lemmas 3.2 and 4.6].

We need the following elementary computation.

Lemma 12.3.4 (Gradients of multiple ℓ_p regression)**.** The gradient $\nabla_{\mathbf{x}}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p$ is given by the formula

$$\sum_{i=1}^n p[\mathbf{A}\mathbf{x}-\mathbf{b}](i)^{\circ(p-1)}(\mathbf{A}^\top \mathbf{e}_i)$$

The following lemma uses Lemmas 12.3.3 and 12.3.4 to show that if $\mathbf{x}$ achieves a nearly optimal value, then $\mathbf{x}$ must be close to the optimal solution $\mathbf{x}^*$.

Lemma 12.3.5 (Closeness of nearly optimal solutions)**.** Let $1 < p < \infty$. For any $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p \leq (1+\eta)\,\mathsf{OPT}$ with $\eta \in (0,1)$, we have that

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p \leq \begin{cases} O(\eta^{1/2})\,\mathsf{OPT} & p < 2 \\ O(\eta^{1/p})\,\mathsf{OPT} & p > 2 \end{cases}$$

where $\mathbf{x}^* \coloneqq \arg\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p$.

Proof. First note that for any $\mathbf{x} \in \mathbb{R}^d$, we have

$$\left\langle (\mathbf{A}\mathbf{x}^* - \mathbf{b})^{\circ(p-1)}, \mathbf{A}\mathbf{x} \right\rangle = \sum_{i=1}^{n} [\mathbf{A}\mathbf{x}^* - \mathbf{b}](i)^{\circ(p-1)} [\mathbf{A}\mathbf{x}](i) = \left\langle \sum_{i=1}^{n} [\mathbf{A}\mathbf{x}^* - \mathbf{b}](i)^{\circ(p-1)} (\mathbf{A}^\top \mathbf{e}_i), \mathbf{x} \right\rangle.$$

The left term in the product is the gradient of the objective at the optimum by Lemma 12.3.4, so this is just 0 for any $\mathbf{x}$. Then for $p < 2$, we have by Lemma 12.3.3 that

$$\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^2 + \frac{p-1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^2 \leq \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^2 \leq (1+\eta)^2 \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^2$$

which rearranges to

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p \leq O(\eta^{1/2})\,\mathsf{OPT}\,.$$

and for $p > 2$, we have by Lemma 12.3.3 that

$$\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p + \frac{p-1}{p2^p}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \leq \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p \leq (1+\eta)^p \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$$

which rearranges to

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p \leq O(\eta^{1/p})\,\mathsf{OPT}\,.$$

□

12.3.2 Iterative size reduction argument

We now give the proof of Theorem 12.3.1.

We will need the following initial result to seed our iterative argument. Note that the dependence on ε is suboptimal by an ε factor for every $1 < p < \infty$.

Lemma 12.3.6. Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 6.1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and

$$\alpha = \begin{cases} O(\gamma)(\varepsilon\delta)^2 \left[(\log d)^2 \log n + \log\frac{1}{\delta}\right]^{-1} & 1 \leq p < 2 \\ \dfrac{O(\gamma^{p/2})(\varepsilon\delta)^p}{\|\mathbf{w}\|_1^{p/2-1}} \left[(\log d)^2 \log n + \log\frac{1}{\delta}\right]^{-1} & 2 < p < \infty \end{cases}.$$

Then, for any $\hat{\mathbf{x}} \in \mathbb{R}^d$ such that

$$\|\mathbf{S}(\mathbf{A}\hat{\mathbf{x}} - \mathbf{b})\|_p^p \leq (1+\varepsilon) \min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p,$$

we have

$$\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_p^p \leq (1+O(\varepsilon)) \min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p.$$

Proof of Lemma 12.3.6. We first show that

$$\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{A}\mathbf{x}^*\|_p^p \leq O\left(\frac{1}{\delta}\right) \mathsf{OPT}^p$$

with probability at least $1-\delta$. By using the fact that $\mathbf{S}$ is an $O(1)$-approximate ℓ_p subspace embedding, we have that

$$\begin{aligned}
\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{A}\mathbf{x}^*\|_p^p &\leq \|\mathbf{S}(\mathbf{A}\hat{\mathbf{x}} - \mathbf{A}\mathbf{x}^*)\|_p^p \\
&\leq 2^{p-1}\left(\|\mathbf{S}(\mathbf{A}\hat{\mathbf{x}} - \mathbf{b})\|_p^p + \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p\right) && \text{Fact 2.1.1} \\
&\leq 2^{p+1}\|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p && \text{Approximate optimality of } \hat{\mathbf{X}}
\end{aligned}$$

The latter quantity is at most $O(\frac{1}{\delta})\,\mathsf{OPT}^p$ with probability at least $1-\delta$ by Markov's inequality. Thus, we may replace the optimization of $\hat{\mathbf{x}}$ over all $\mathbf{x} \in \mathbb{R}^d$ with optimization over the ball $\{\mathbf{x} : \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p = O(\frac{1}{\delta})\,\mathsf{OPT}^p\}$.

We apply Theorem 12.3.2 with accuracy parameter ε set to $\varepsilon\delta$ and proximity parameter η set to 1. It follows that

$$\begin{aligned}
&\left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p - \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p - \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p\right)\right| \\
\leq\ &\varepsilon\delta\left(\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p + \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p + \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p\right) \leq O(\varepsilon)\,\mathsf{OPT}^p
\end{aligned}$$

Thus, in the ball $\{\mathbf{x} : \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p = O(\frac{1}{\delta})\,\mathsf{OPT}^p\}$, we have that

$$\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p + (\|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p - \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p) \pm O(\varepsilon)\,\mathsf{OPT}^p\,.$$

It follows that $\hat{\mathbf{x}}$ must minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p$ up to an additive $O(\varepsilon)\,\mathsf{OPT}^p$. □

Starting from this initial solution bound of Lemma 12.3.6, we can proceed via an iterative argument which alternates between using a bound on the closeness of the solution to the optimal solution to improve the approximation (Theorem 12.3.2), and using a bound on the approximation to improve the closeness to the optimum (Lemma 12.3.5). More specifically, we can show that for $1 < p < 2$, a bound of C/ε^β on the sample complexity implies that a bound of $C/\varepsilon^{2\beta/(1+\beta)}$ is sufficient as well. Iterating this argument starting from $\beta = 2$ due to Lemma 12.3.6 for $O(\log\log\frac{1}{\varepsilon})$ iterations yields the desired bound of C/ε, as claimed. Similarly, for $p > 2$, a bound of C/ε^β implies a bound of $C/\varepsilon^{p\beta/(1+\beta)}$, which results in a final bound of C/ε^{p-1}, as claimed.

Proof of Theorem 12.3.1. Let

$$C = \begin{cases} O(\gamma^{-1})\delta^{-2}\|\mathbf{w}\|_1\left[(\log d)^2\log n + \log\frac{1}{\delta}\right] & p < 2 \\ O(\gamma^{-p/2})\delta^{-p}\|\mathbf{w}\|_1^{p/2}\left[(\log d)^2\log n + \log\frac{1}{\delta}\right] & p > 2 \end{cases}$$

We will make use of the fact that $\|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p = O(\frac{1}{\delta})\|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p$ with probability at least $1 - \delta$ by Markov's inequality.

We will first give the argument for $p < 2$. Suppose that C/ε^β rows are needed for a $(1+\varepsilon)$-approximate weak coreset. Now choose a such that $a - 2 = -a\beta$, that is, $a = 2/(1+\beta)$. Then for $\eta^{2/p} = \varepsilon^a$, $C\eta^{2/p}/(\varepsilon\delta)^2 = C/\eta^{(2/p)\beta}$ rows yields a $(1+\eta^{2/p})$-approximate weak coreset. Then, a $(1+\eta^{2/p})$-approximate minimizer $\mathbf{x}$ satisfies

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \leq O(\eta)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$$

by Lemma 12.3.5. For all such $\mathbf{X}$, Theorem 12.3.2 shows that $\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p - \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p$ and $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p - \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$ are close up to an additive error of

$$\varepsilon\delta\left(\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p + \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p + \frac{1}{\eta}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p\right) = O(\varepsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$$

Thus, $C/\eta^{(2/p)\beta}$ rows in fact gives a $(1+O(\varepsilon))$-approximate minimizer. That is, if C/ε^β rows is sufficient for $(1+\varepsilon)$-approximation, then $C/\eta^{(2/p)\beta} = C/\varepsilon^{a\beta} = C/\varepsilon^{2\beta/(1+\beta)}$ rows is sufficient for $(1+\varepsilon)$-approximation as well. We may now iterate this argument. Consider the sequence β_i given by

$$\beta_0 = 2, \qquad \beta_{i+1} = \frac{2\beta_i}{1+\beta_i}.$$

The solution to this recurrence is given by the following lemma, with $p = 2$:

Lemma 12.3.7. Let $p > 1$ and let $\{\beta_i\}_{i=0}^\infty$ be defined by the recurrence relation $\beta_0 = p$ and $\beta_{i+1} = p\beta_i/(1+\beta_i)$. Then,

$$\beta_i = \frac{1}{p^{-i}(p^{-1} - (p-1)^{-1}) + (p-1)^{-1}}$$

Proof. Note that $\frac{1}{\beta_{i+1}} = \frac{1}{p}\frac{1}{\beta_i} + \frac{1}{p}$ so the sequence $\{a_i\}_{i=0}^\infty$ given by $a_i = 1/\beta_i$ satisfies the linear recurrence $a_{i+1} = \frac{1}{p}a_i + \frac{1}{p}$. Note that this recurrence has the fixed point $a = 1/(p-1)$, so the sequence $a_i' = a_i - a$ satisfies $a_{i+1}' = \frac{1}{p}a_i'$, which gives, $a_i' = p^{-i}a_0'$. Thus, $a_i - a = p^{-i}(a_0 - a)$ so

$$\begin{aligned}\beta_i = \frac{1}{a_i} &= \frac{1}{p^{-i}(a_0 - a) + a} \\ &= \frac{1}{p^{-i}(p^{-1} - (p-1)^{-1}) + (p-1)^{-1}}.\end{aligned}$$

□

Thus, applying this argument $O(\log\log\frac{1}{\varepsilon})$ times yields that $\beta_i \le 1 + O(1/\log(\frac{1}{\varepsilon}))$ which means that reading only $O(1)C/\varepsilon$ entries suffices. Union bounding over the success of the $O(\log\log\frac{1}{\varepsilon})$ rounds completes the argument.

Next, let $p > 2$. Suppose that C/ε^β rows are needed for a $(1+\varepsilon)$-approximate weak coreset. Now choose a such that $a - p = -a\beta$, that is, $a = p/(1+\beta)$. Then for $\eta = \varepsilon^a$, $C\eta/\varepsilon^p = C/\eta^\beta$ rows yields a $(1+\eta)$-approximate weak coreset. Then, a $(1+\eta)$-approximate minimizer $\mathbf{X}$ satisfies

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \le O(\eta)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$$

by Lemma 12.3.5. For all such $\mathbf{x}$, Theorem 12.3.2 shows that $\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p - \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p$ and $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p - \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$ are close up to an additive error of

$$\varepsilon\left(\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p + \frac{1}{\eta}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p\right) = O(\varepsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$$

Thus, C/η^β rows in fact gives a $(1+O(\varepsilon))$-approximate minimizer. That is, if C/ε^β rows is sufficient for $(1+\varepsilon)$-approximation, then $C/\eta^\beta = C/\varepsilon^{a\beta} = C/\varepsilon^{p\beta/(1+\beta)}$ rows is sufficient for $(1+\varepsilon)$-approximation as well. We may now iterate this argument. Consider the sequence β_i given by

$$\beta_1 = p, \qquad \beta_{i+1} = \frac{p\beta_i}{1+\beta_i}.$$

Then by Lemma 12.3.7, applying this argument $O(\log\log\frac{1}{\varepsilon})$ times yields that $\beta_i \le (p-1) + O(1/\log(\frac{1}{\varepsilon}))$ which means that reading only $O(1)C/\varepsilon^{p-1}$ entries suffices. Union bounding over the success of the $O(\log\log\frac{1}{\varepsilon})$ rounds completes the argument. □

12.3.3 High probability

Note that in the statement of Theorem 12.3.1, the dependence on the failure rate δ is polynomial. This is in fact necessary if we restrict our algorithm to be of the form of "sample-and-solve" algorithms whose sampling matrices $\mathbf{S}$ don't depend on $\mathbf{b}$ (see Theorem 12.6.7). The only reason why this dependence becomes necessary in the analysis of the upper bound is that $\|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p$ may be as large as $O(\frac{1}{\delta})\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$ with probability at least δ, and this is the source of the hardness result of Theorem 12.6.7 as well. This is a mild problem, and we show how to overcome this problem via the following two-stage procedure. First, we can obtain a constant factor solution $\hat{\mathbf{x}}$ with a polylogarithmic dependence on δ via the boosting procedure described in Section 12.2.1. Then, we can run $\log\frac{1}{\delta}$ copies of the algorithm, each which succeeds with probability $1-\delta$. Then, we can sort the runs by their estimates $\|\mathbf{S}(\mathbf{A}\hat{\mathbf{x}} - \mathbf{b})\|_p^p$ and discard half of the runs with the highest values of $\|\mathbf{S}(\mathbf{A}\hat{\mathbf{x}} - \mathbf{b})\|_p^p$. This guarantees that the remaining runs have $\|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p = O(1)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$ with probability at least $1-\delta$, which is enough for the rest of the argument to go through with only a $(\log\frac{1}{\delta})^2$ dependence on δ. This proves the following result:

Theorem 12.3.8 (Nearly optimal active ℓ_p linear regression, high probability)**.** There is an active ℓ_p linear regression algorithm (see Definition 12.1.1) with query complexity at most r with

probability at least $1-\delta$, where

$$r = \begin{cases} \tilde{O}(\varepsilon^{-2}d)\left(\log \delta^{-1}\right)^2 & 0 < p < 1 \\ \tilde{O}(\varepsilon^{-1}d)\left(\log \delta^{-1}\right)^2 & 1 < p < 2 \\ \tilde{O}(\varepsilon^{1-p}d^{p/2})\left(\log \delta^{-1}\right)^2 & 2 < p < \infty \end{cases}$$

12.4 ℓ_p Lewis weight sampling for differences

Throughout this section, we fix the following notation:

Definition 12.4.1.

- Let $1 \le p < \infty$.
- Let $\varepsilon \in (0,1)$ be an accuracy parameter and let $\delta \in (0,1)$ be a failure probability parameter.
- Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$.
- Let $\mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights for $\mathbf{A}$ such that $\max_{i=1}^n \mathbf{w}_i \le w$.
- Let $\mathbf{x}^* \in \mathbb{R}^d$ any center, let $\eta \in (0,1)$ be a proximity parameter, and let $R \ge \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$ be a scale parameter.
- For each $i \in [n]$ and $\mathbf{x} \in \mathbb{R}^d$, let

$$\Delta_i(\mathbf{x}) \coloneqq |[\mathbf{A}\mathbf{x} - \mathbf{b}](i)|^p - |[\mathbf{A}\mathbf{x}^* - \mathbf{b}](i)|^p$$

Our main result of the section is the following:

Theorem 12.4.2. Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 6.1.1) with sampling probabilities $q_i \ge \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and

$$\alpha = \begin{cases} O(\gamma)\dfrac{\varepsilon^2}{\eta^{2/p}}\left[(\log d)^2 \log n + \log\dfrac{1}{\delta}\right]^{-1} & p < 2 \\ O(\gamma^{p/2})\dfrac{\varepsilon^p}{\eta\|\mathbf{w}\|_1^{p/2-1}}\left[(\log d)^2 \log n + \log\dfrac{1}{\delta}\right]^{-1} & p > 2 \end{cases}.$$

Then for each $\mathbf{x}^* \in \mathbb{R}^d$ and $R \ge \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$, with probability at least $1-\delta$,

$$\begin{aligned} &\sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R} \left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x}-\mathbf{b})\|_p^p - \|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_p^p - \|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p^p\right)\right| \\ \le\ & \varepsilon(R + \|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p) \end{aligned}$$

We will prove Theorem 12.4.2 throughout this section. Before doing so, we state the following more convenient form of the result:

Theorem 12.3.2. Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 6.1.1) with sampling probabilities $q_i \ge \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and

$$\alpha = \begin{cases} \dfrac{O(\gamma)\varepsilon^2}{\eta^{2/p}}\left[(\log d)^2 \log n + \log\dfrac{1}{\delta}\right]^{-1} & p < 2 \\ \dfrac{O(\gamma^{p/2})\varepsilon^p}{\eta\|\mathbf{w}\|_1^{p/2-1}}\left[(\log d)^2 \log n + \log\dfrac{1}{\delta}\right]^{-1} & p > 2 \end{cases}.$$

For each $\mathbf{x}^* \in \mathbb{R}^d$ and $\mathbf{b}^* = \mathbf{A}\mathbf{x}^* - \mathbf{b}$, with probability at least $1 - \delta$,

$$\left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p - \|\mathbf{S}\mathbf{b}^*\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p - \|\mathbf{b}^*\|_p^p\right)\right| \leq \varepsilon\left(\|\mathbf{b}^*\|_p^p + \|\mathbf{S}\mathbf{b}^*\|_p^p + \frac{1}{\eta}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p\right)$$

simultaneously for every $\mathbf{x} \in \mathbb{R}^d$.

Proof. We apply Theorem 12.4.2 with δ set to δ/L for $L = O(\log(1/\delta\varepsilon))$ and R set to $2^l\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$ for $l \in [L]$. By a union bound, the conclusion holds simultaneously for every $l \in [L]$ with probability at least $1-\delta$. Furthermore, by Markov's inequality, $\|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p = O(1/\delta)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$ with probability at least $1 - \delta$.

If $\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \leq 2^L\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p = \mathrm{poly}(1/\delta\varepsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$, then the result follows immediately from applying the conclusion of Theorem 12.4.2 at the appropriate scale $l \in [L]$. Otherwise, we have that $\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \geq \mathrm{poly}(1/\delta\varepsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$, in which case

$$\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*)\|_p^p \geq \Omega(1)\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \geq \mathrm{poly}(1/\delta\varepsilon)\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p$$

so

$$\begin{aligned}\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p - \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p &= (1 \pm \varepsilon)\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*)\|_p^p \pm \frac{(1+\varepsilon)^{p-1}}{\varepsilon^{p-1}}\|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p \\ &= (1 \pm \varepsilon)\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*)\|_p^p \pm \frac{(1+\varepsilon)^{p-1}}{\delta\varepsilon^{p-1}}\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p \\ &= (1 \pm O(\varepsilon))\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*)\|_p^p\end{aligned}$$

and similarly,

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p - \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p = (1 \pm O(\varepsilon))\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p.$$

Thus it suffices to have that

$$\left|\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*)\|_p^p - \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p\right| \leq \frac{\varepsilon}{\eta}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p.$$

In fact, standard ℓ_p Lewis weight sampling guarantees give

$$\left|\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*)\|_p^p - \|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p\right| \leq \begin{cases} \frac{\varepsilon}{\eta^{1/p}}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p & p < 2 \\ \frac{\varepsilon^{p/2}}{\eta^{1/2}}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p & p > 2 \end{cases}$$

which is stronger. □

Throughout our proof of Theorem 12.4.2, we will assume without loss of generality that $\mathbf{S}_{i,i}^p > 1$, that is we only consider rows that are sampled with probability $q_i < 1$, since rows that are kept with probability $q_i = 1$ do not contribute towards the sampling error. Note first that we can write

$$\left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_p^p - \|\mathbf{S}(\mathbf{A}\mathbf{x}^* - \mathbf{b})\|_p^p\right) - \left(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p - \|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_p^p\right)\right| = \left|\sum_{i=1}^n (\mathbf{S}_{i,i}^p - 1)\Delta_i(\mathbf{x})\right|.$$

The supremum of this quantity, normalized by $(R+\|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p)^l$, over $\left\{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R\right\}$ is a random variable. We will bound the l-th moment of this random variable for $l = O(\log\frac{1}{\delta} + \log n)$.

We start with a standard symmetrization procedure (see Lemma 2.3.2). Next, we replace the Rademacher process on the right hand side of Lemma 2.3.2 by one which "removes" $\mathbf{S}_{i,i}^p$, that is, one of the form

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n}\left[\sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p\le\eta R}\left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i\Delta_i(\mathbf{x})\right|^l\right]. \tag{12.1}$$

This is roughly done by noting that if we take $\mathbf{SA}$ to be a "part of" $\mathbf{A}$, then the domain $\left\{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R\right\}$ only dilates by a constant factor as $\mathbf{S}$ preserves ℓ_p norms in the column space of $\mathbf{A}$. More formally, we have the following lemma:

Lemma 12.4.3. Let $\mathbf{B}\in\mathbb{R}^{m\times d}$ satisfy $\|\mathbf{B}\mathbf{x}\|_p^p \le C\|\mathbf{A}\mathbf{x}\|_p^p$ for every $\mathbf{x}\in\mathbb{R}^d$. For every fixing of $\mathbf{S}$, let

$$\mathbf{B}_\mathbf{S} \coloneqq \begin{pmatrix}\mathbf{S}\mathbf{A}\\ \mathbf{B}\end{pmatrix}$$

be the concatenation of $\mathbf{SA}$ and $\mathbf{B}$, and let

$$F_\mathbf{S} = \sup_{\|\mathbf{A}\mathbf{x}\|_p^p\le 1}\left|\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p - \|\mathbf{A}\mathbf{x}\|_p^p\right|.$$

Suppose that for every fixing of $\mathbf{S}$ and $R' \ge R + \|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p$, we have that

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n}\sup_{\|\mathbf{B}_\mathbf{S}\mathbf{x}-\mathbf{B}_\mathbf{S}\mathbf{x}^*\|_p^p\le\eta R'}\left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i\mathbf{S}_{i,i}^p\Delta_i(\mathbf{x})\right| \le \varepsilon^l\delta R'^l$$

Then,

$$\mathop{\mathbf{E}}_{\mathbf{S}}\frac{1}{(R+\|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p)^l}\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n}\sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p\le\eta R}\left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i\mathbf{S}_{i,i}^p\Delta_i(\mathbf{x})\right|^l \le (2\varepsilon)^l\delta\Big((1+C)^l + \mathop{\mathbf{E}}_{\mathbf{S}}[F_\mathbf{S}^l]\Big)$$

Proof. Note that

$$\|\mathbf{B}_\mathbf{S}(\mathbf{x}-\mathbf{x}^*)\|_p^p = \|\mathbf{S}\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_p^p + \|\mathbf{B}(\mathbf{x}-\mathbf{x}^*)\|_p^p \le (1+F_\mathbf{S}+C)\|\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_p^p$$

so

$$\begin{aligned}\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n}\sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p\le\eta R}\left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i\mathbf{S}_{i,i}^p\Delta_i(\mathbf{x})\right|^l &\le \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n}\sup_{\|\mathbf{B}_\mathbf{S}\mathbf{x}-\mathbf{B}_\mathbf{S}\mathbf{x}^*\|_p^p\le(1+F_\mathbf{S}+C)\eta R}\left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i\mathbf{S}_{i,i}^p\Delta_i(\mathbf{x})\right|^l\\ &\le \varepsilon^l\delta(1+F_\mathbf{S}+C)^l(R+\|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p)^l\\ &\le \varepsilon^l\delta 2^{l-1}((1+C)^l+F_\mathbf{S}^l)(R+\|\mathbf{S}(\mathbf{A}\mathbf{x}^*-\mathbf{b})\|_p^p)^l \qquad \text{Fact 2.1.1}\end{aligned}$$

Taking expectations on both sides proves the lemma. □

Note that if $\mathbf{S}$ is the ℓ_p Lewis weight sampling matrix, then $\mathbf{E}[|F_{\mathbf{S}}|^l]$ in Lemma 12.4.3 is known be bounded as $O(1)^l$ (that is $\mathbf{S}$, is an $O(1)$-approximate ℓ_p subspace embedding) by standard results on ℓ_p Lewis weight sampling [CP15, WY23b].

Furthermore, we can design $\mathbf{B}$ such that the ℓ_p Lewis weights of $\mathbf{B}_{\mathbf{S}}$ are uniformly bounded by α, where α is the oversampling parameter such that $\mathbf{S}$ samples the ith row with probability $\min\{1, \mathbf{w}_i/\alpha\}$. For $p < 2$, this simply follows by taking $\mathbf{B}$ to be a flattening of $\mathbf{A}$ where every row is duplicated $1/\alpha$ times due to the monotonicity of ℓ_p Lewis weights [CP15]. For $p > 2$, monotonicity of ℓ_p Lewis weights does not hold, but Theorem 5.2 of [WY23b] nonetheless shows that γ-one-sided ℓ_p Lewis weights can be constructed for $\mathbf{B}_{\mathbf{S}}$ with $\gamma = \Omega(1)$ that makes a similar argument go through.

Finally, it remains to bound the Rademacher process of the form of (12.1), where $\mathbf{A}$ has γ-one-sided ℓ_p Lewis weights uniformly bounded by $w = \alpha$. We will prove the following in Section 12.5. Assuming this theorem, Theorem 12.4.2 follows by setting $w = \alpha$ as stated.

Theorem 12.4.4. For all $l \in \mathbb{N}$, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R} \left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x})\right|^l \le (\varepsilon R)^l \tag{12.2}$$

where

$$\varepsilon = \begin{cases} O(w\eta^{2/p})^{1/2}\gamma^{-1/2}\left[\left((\log d)^2 \log n\right)^{1+1/l} + l\right]^{1/2} & p < 2 \\ O(w\eta\|\mathbf{w}\|_1^{p/2-1})^{1/p}\gamma^{-1/2}\left[\left((\log d)^2 \log n\right)^{1+1/l} + l\right]^{1/p} & p > 2 \end{cases}.$$

12.5 Rademacher process bounds

We continue to fix our notation from Definition 12.4.1. We will prove Theorem 12.4.4 in this section.

We split the sum in (12.2) into two parts: the part that is bounded by the γ-one-sided Lewis weights of $\mathbf{A}$, and the part that is not. To this end, define a threshold

$$\tau \coloneqq \begin{cases} \dfrac{\eta}{\gamma^{p/2}\varepsilon^p} & p < 2 \\ \dfrac{\eta\|\mathbf{w}\|_1^{p/2-1}}{\gamma^{p/2}\varepsilon^p} & p > 2 \end{cases}$$

where ε will be determined later, and define the set of "good" entries $G \subseteq [n]$ as

$$G \coloneqq \{i \in [n] : |[\mathbf{A}\mathbf{x}^* - \mathbf{b}](i)| \le \tau \mathbf{w}_i R\} \tag{12.3}$$

We then bound

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R} \left|\sum_{i=1}^n \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x})\right|^l \le 2^{l-1} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R} \left|\sum_{i\in G} \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x})\right|^l$$

$$+ 2^{l-1} \mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm 1\}^n} \sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R} \left| \sum_{i\in[n]\setminus G} \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x}) \right|^l$$

using the Fact 2.1.1, and separately estimate each term. We can think of the first term as the "sensitivity" term, where each term in the sum is bounded by the Lewis weights of $\mathbf{A}$, and the latter term as the "outlier" term, where each term in the sum is much larger than the corresponding Lewis weights.

12.5.1 Estimates on the outlier term

We first bound the outlier terms ($i \notin G$), which is much easier.

Lemma 12.5.1. With probability 1, we have that

$$\sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R} \sum_{i\in[n]\setminus G} |\Delta_i(\mathbf{x})| \le O(\varepsilon) R.$$

Proof. For each $i \in [n] \setminus G$, we have that

$$\begin{aligned}
|[\mathbf{A}\mathbf{x}-\mathbf{b}](i)| &\in |[\mathbf{A}\mathbf{x}^*-\mathbf{b}](i)| \pm |[\mathbf{A}\mathbf{x}^*-\mathbf{A}\mathbf{x}](i)| \\
&\in |[\mathbf{A}\mathbf{x}^*-\mathbf{b}](i)| \pm \gamma^{-1/2} \|\mathbf{w}\|_1^{1/2-1/p} \mathbf{w}_i^{1/p} \|\mathbf{A}\mathbf{x}^*-\mathbf{A}\mathbf{x}\|_p && \text{Lemma 6.2.4} \\
&\in |[\mathbf{A}\mathbf{x}^*-\mathbf{b}](i)| \pm \gamma^{-1/2} \eta^{1/p} \|\mathbf{w}\|_1^{1/2-1/p} \mathbf{w}_i^{1/p} R^{1/p} \\
&\in |[\mathbf{A}\mathbf{x}^*-\mathbf{b}](i)| \pm \varepsilon |[\mathbf{A}\mathbf{x}^*-\mathbf{b}](i)| && i \in [n] \setminus G
\end{aligned}$$

Thus,

$$|\Delta_i(\mathbf{x})| \le O(\varepsilon) |[\mathbf{A}\mathbf{x}^*-\mathbf{b}](i)|^p$$

so

$$\sum_{i\in[n]\setminus G} |\Delta_i(\mathbf{x})| \le \sum_{i=1}^n O(\varepsilon) |[\mathbf{A}\mathbf{x}^*-\mathbf{b}](i)|^p = O(\varepsilon) \|\mathbf{A}\mathbf{x}^*-\mathbf{b}\|_p^p \le O(\varepsilon) R. \qquad \square$$

12.5.2 Estimates on the sensitivity term

Next, we estimate the sensitivity term ($i \in G$),

$$\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm 1\}^n} \sup_{\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R} \left| \sum_{i\in G} \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x}) \right|^l.$$

To estimate this moment, we obtain a subgaussian tail bound via the tail form of Dudley's entropy integral, and then integrate it. We will crucially use that $|\Delta_i(\mathbf{x})|$ for $i \in G$ is bounded over all $\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R$, which gives the following sensitivity bound:

Lemma 12.5.2. For all $i \in G$, and $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^*\|_p^p \le \eta R$, we have $|[\mathbf{A}\mathbf{x}-\mathbf{b}](i)|^p \le O(\tau \mathbf{w}_i R)$ and $|\Delta_i(\mathbf{x})| \le O(\tau \mathbf{w}_i R)$.

Proof. We have

$$\begin{aligned}|[\mathbf{Ax}-\mathbf{b}](i)|^p &\le 2^{p-1}(|[\mathbf{Ax}^*-\mathbf{b}](i)|^p + |[\mathbf{Ax}-\mathbf{Ax}^*](i)|^p) && \text{Fact 2.1.1}\\ &\le 2^{p-1}\tau\mathbf{w}_i R + 2^{p-1}\gamma^{-p/2}\eta\|\mathbf{w}\|_1^{0\vee(p/2-1)}\mathbf{w}_i R && i\in G \text{ (see (12.3)) and Lemma 6.2.4}\\ &\le O(\tau\mathbf{w}_i R)\end{aligned}$$

The bound on $\Delta_i(\mathbf{x})$ follows easily from the above calculation. □

Bounding low-sensitivity entries

We now separately handle entries $i \in G$ with small Lewis weight. To do this end, define

$$J \coloneqq \left\{i \in G : \mathbf{w}_i \ge \frac{\varepsilon}{\tau n}\right\}.$$

We then bound the mass on the complement of J:

Lemma 12.5.3. For all $\|\mathbf{Ax}-\mathbf{Ax}^*\|_p^p \le \eta R$, we have that

$$\sum_{i\in[n]\setminus J} |\Delta_i(\mathbf{x})| \le O(\varepsilon R)$$

Proof. We have that for each $i \in [n]\setminus J$, $\mathbf{w}_i \le \varepsilon/\tau n$ so by Lemma 12.5.2,

$$\sum_{i\in[n]\setminus J} |\Delta_i(\mathbf{x})| \le \sum_{i\in[n]\setminus J} O(\tau\mathbf{w}_i R) \le \sum_{i\in[n]\setminus J} \frac{O(\varepsilon)}{n} R \le O(\varepsilon R)$$

□

Bounding high-sensitivity entries: Dudley's inequality

Finally, it remains to bound the Rademacher process only on the entries indexed by $i \in J$. Define a Rademacher process by

$$X_{\mathbf{x}} \coloneqq \sum_{i\in J} \boldsymbol{\varepsilon}_i \Delta_i(\mathbf{x})$$

with pseudo-metric

$$d_X(\mathbf{x},\mathbf{x}') \coloneqq \left(\mathop{\mathbf{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} |X_{\mathbf{x}} - X'_{\mathbf{x}}|^2\right)^{1/2} = \left(\sum_{i\in J} (\Delta_i(\mathbf{x}) - \Delta_i(\mathbf{x}'))^2\right)^{1/2}$$

We will use Dudley's entropy integral (Theorem 2.3.6) to bound the tail of this quantity, and then integrate to obtain moment bounds.

Using the sensitivity bound of Lemma 12.5.2, we obtain a bound on the pseudo-metric d_X.

Lemma 12.5.4. Let $q = O(\log(\tau n/\varepsilon))$. For $\mathbf{x}, \mathbf{x}' \in T$ for $T = \{\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^*\|_p^p \leq \eta R\}$, we have that

$$d_X(\mathbf{x}, \mathbf{x}') \leq \begin{cases} O(w^{1/2})\eta^{1/p-1/2}\|\mathbf{W}^{-1/p}\mathbf{A}(\mathbf{x} - \mathbf{x}')\|_{\mathbf{w},q}^{p/2} R^{1/2} & p < 2 \\ O(w^{1/2})\tau^{1/2-1/p}\|\mathbf{W}^{-1/p}\mathbf{A}(\mathbf{x} - \mathbf{x}')\|_{\mathbf{w},q} R^{1-1/p} & p > 2 \end{cases}$$

and

$$\operatorname{diam}(T) = \sup_{\mathbf{x},\mathbf{x}' \in T} d_X(\mathbf{x}, \mathbf{x}') \leq \begin{cases} O(w^{1/2}\eta^{1/p}\gamma^{-1/2} R) & p < 2 \\ O(\varepsilon w^{1/2}\tau^{1/2} R) & p > 2 \end{cases}$$

Proof. Let $\mathbf{y} = \mathbf{A}\mathbf{x} - \mathbf{b}$ and $\mathbf{y}' = \mathbf{A}\mathbf{x}' - \mathbf{b}$. Note then that

$$\begin{aligned} d_X(\mathbf{x}, \mathbf{x}')^2 &= \sum_{i \in J} (\Delta_i(\mathbf{x}) - \Delta_i(\mathbf{x}'))^2 = \sum_{i \in J} (|\mathbf{y}(i)|^p - |\mathbf{y}'(i)|^p)^2 \\ &\leq p^2 \sum_{i \in J} |\mathbf{y}(i) - \mathbf{y}'(i)|^2 (|\mathbf{y}(i)|^{p-1} + |\mathbf{y}'(i)|^{p-1})^2 && \text{Fact 2.1.3} \end{aligned}$$

For $p < 2$, we have that

$$\begin{aligned} d_X(\mathbf{x}, \mathbf{x}')^2 &\leq p^2 \|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^p \sum_{i \in J} (|\mathbf{y}(i) - \mathbf{y}'(i)|)^{2-p} (|\mathbf{y}(i)|^{p-1} + |\mathbf{y}'(i)|^{p-1})^2 \\ &\leq 2p^2 \|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^p \sum_{i \in J} (|\mathbf{y}(i) - \mathbf{y}'(i)|)^{2-p} (|\mathbf{y}(i)|^{2p-2} + |\mathbf{y}'(i)|^{2p-2}) \\ &\leq 2p^2 \|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^p \|\mathbf{y} - \mathbf{y}'\|_p^{2-p} (\|\mathbf{y}\|_p^{2p-2} + \|\mathbf{y}'\|_p^{2p-2}) && \text{Hölder's inequality} \\ &\leq O(\eta^{2/p-1}) \|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^p R. \end{aligned}$$

where the Hölder's inequality is applied with exponents $\frac{p}{2-p} > 1$ and $\frac{p}{2p-2} > 1$. For $p > 2$, we have that

$$\begin{aligned} d_X(\mathbf{x}, \mathbf{x}')^2 &\leq 2p^2 \|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^2 \sum_{i=1}^n |\mathbf{y}(i)|^{2p-2} + |\mathbf{y}'(i)|^{2p-2} \\ &\leq 2p^2 \max\{\|\mathbf{y}|_J\|_\infty, \|\mathbf{y}'|_J\|_\infty\}^{p-2} \|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^2 \sum_{i=1}^n |\mathbf{y}(i)|^p + |\mathbf{y}'(i)|^p \\ &\leq O(1)(\tau w R)^{1-2/p} \|(\mathbf{y} - \mathbf{y}')|_J\|_\infty^2 R && \text{Lemma 12.5.2} \end{aligned}$$

Furthermore, we have that

$$\begin{aligned} \|(\mathbf{y} - \mathbf{y}')|_J\|_\infty &= \|(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}')|_J\|_\infty \\ &= \|\mathbf{W}^{1/p}(\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}')|_J\|_\infty \\ &\leq w^{1/p} \|(\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}')|_J\|_\infty \\ &\leq 2w^{1/p} \|\mathbf{W}^{-1/p}\mathbf{A}\mathbf{x} - \mathbf{W}^{-1/p}\mathbf{A}\mathbf{x}'\|_{\mathbf{w},q} \end{aligned}$$

where the last step follows from the fact that $\mathbf{w}_i \geq \varepsilon/\tau n$ for $i \in J$ and $q = O(\log(\tau n/\varepsilon))$. Combining these bounds gives the claimed bound on $d_X(\mathbf{x}, \mathbf{x}')$.

Finally, we have by Lemma 6.2.4 that

$$\|\mathbf{W}^{-1/p}\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_\infty = \max_{i=1}^n \frac{|[\mathbf{A}(\mathbf{x}-\mathbf{x}^*)](i)|}{\mathbf{w}_i} \le \begin{cases} \gamma^{-1/p}\|\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_p & p<2 \\ \gamma^{-1/2}\|\mathbf{w}\|_1^{1/2-1/p}\|\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_p & p>2 \end{cases}$$

so we have the claimed diameter bound for the set $\{\|\mathbf{A}(\mathbf{x}-\mathbf{x}^*)\|_p^p \le \eta R\}$. □

The following entropy bounds are obtained from [WY23c], which in turn largely follow [BLM89].

Lemma 12.5.5. Let $1 \ge \mathbf{w} \in \mathbb{R}^n$ be nonnegative weights. Let $2 \le q < \infty$ and let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be such that $\mathbf{W}^{1/2}\mathbf{A}$ is orthonormal. Let $\tau \ge \max_{i=1}^n \left\|\mathbf{e}_i^\top \mathbf{A}\right\|_2^2$. Let $B_{\mathbf{w}}^p(\mathbf{A}) \coloneqq \{\mathbf{x} : \|\mathbf{A}\mathbf{x}\|_{\mathbf{w},p} \le 1\}$. Then,

$$\log E(B_{\mathbf{w}}^2(\mathbf{A}), B_{\mathbf{w}}^q(\mathbf{A}), t) \le O(1)\frac{n^{2/q}q\cdot\tau}{t^2}$$

and

$$\log E(B_{\mathbf{w}}^p(\mathbf{A}), B_{\mathbf{w}}^q(\mathbf{A}), t) \le O(1)\frac{1}{t^p}\left(\frac{\log d}{2-p} + \log n + n^{2/q}q\right)\tau.$$

for $p < 2$.

We may now evaluate Dudley's entropy integral.

Lemma 12.5.6 (Entropy integral bound for $p<2$)**.** We have that

$$\int_0^\infty \sqrt{\log E(B^p(\mathbf{A}), d_X, t)}\, dt \le O(w^{1/2}\gamma^{-1/2}\eta^{1/2}R)\left(\log\frac{\tau n}{\varepsilon}\right)^{1/2}\log d$$

Proof. Note that it suffices to integrate the entropy integral to $\operatorname{diam}(T)$, which is bounded in Lemma 12.5.4. Note also that T is just a translation of $(\eta R)^{1/p}\cdot B^p(\mathbf{A})$, so we have

$$\begin{aligned}\log E(T, d_X, t) &= \log E((\eta R)^{1/p}\cdot B^p(\mathbf{A}), d_X, t) \\ &= \log E((\eta R)^{1/p}\cdot B^p(\mathbf{A}), K\|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}^{p/2}, t) && \text{Lemma 12.5.4} \\ &= \log E(B_{\mathbf{w}}^p(\mathbf{W}^{-1/p}\mathbf{A}), B_{\mathbf{w}}^q(\mathbf{W}^{-1/p}\mathbf{A}), t^{2/p}/K^{2/p}(\eta R)^{1/p})\end{aligned}$$

where $K = O(w^{1/2}\eta^{1/p-1/2}R^{1/2})$.

For small radii less than λ for a parameter λ to be chosen, we use a standard volume argument, which shows that

$$\log E(B_{\mathbf{w}}^p(\mathbf{W}^{-1/p}\mathbf{A}), B_{\mathbf{w}}^q(\mathbf{W}^{-1/p}\mathbf{A}), t) \le O(d)\log\frac{n}{t}$$

so

$$\begin{aligned}\int_0^\lambda \sqrt{\log E(T, d_X, t)}\, dt &\le \int_0^\lambda \sqrt{d\log\frac{nK^{2/p}(\eta R)^{1/p}}{t^{2/p}}}\, dt \\ &\le \lambda\sqrt{d\log(n(\eta^{2/p}w)^{1/p})} + \sqrt{d}\int_0^\lambda \sqrt{\log\frac{R^{2/p}}{t^{2/p}}}\, dt\end{aligned}$$

$$\leq \lambda\sqrt{d\log(n(\eta^{2/p}w)^{1/p})} + \sqrt{d}\cdot O(\lambda)\sqrt{\log\frac{R}{\lambda}}$$
$$\leq O(\lambda)\sqrt{d\log\frac{n(\eta^{2/p}w)^{1/p}R}{\lambda}}$$

On the other hand, for large radii larger than λ, we use the bounds of Lemma 12.5.5. Note that the entropy bounds do not change if we replace $\mathbf{A}$ by $\mathbf{AR}$, where $\mathbf{R}$ is the change of basis matrix such that $\mathbf{W}^{1/2-1/p}\mathbf{AR}$ is orthonormal. Then by the properties of γ-one-sided ℓ_p Lewis weights (Lemma 6.2.1), we have

$$\|\mathbf{e}_i^\top\mathbf{W}^{-1/p}\mathbf{AR}\|_2^2 = \mathbf{w}_i^{-2/p}\|\mathbf{e}_i^\top\mathbf{AR}\|_2^2 \leq \gamma^{-1}.$$

Then, Lemma 12.5.5 gives

$$\log E(B_{\mathbf{w}}^p(\mathbf{W}^{-1/p}\mathbf{A}), B_{\mathbf{w}}^q(\mathbf{W}^{-1/p}\mathbf{A}), t^{2/p}/K^{2/p}(\eta R)^{1/p}) = \frac{O(w\eta^{2/p}R^2)}{\gamma t^2}\log\frac{\tau n}{\varepsilon}$$

so the entropy integral gives a bound of

$$\frac{O(w^{1/2}\eta^{1/p}R)}{\gamma^{1/2}}\left(\log\frac{\tau n}{\varepsilon}\right)^{1/2}\int_\lambda^{\operatorname{diam}(T)}\frac{1}{t}\,dt = \frac{O(w^{1/2}\eta^{1/p}R)}{\gamma^{1/2}}\left(\log\frac{\tau n}{\varepsilon}\right)^{1/2}\log\frac{\operatorname{diam}(T)}{\lambda}.$$

We choose $\lambda = \operatorname{diam}(T)/\sqrt{d}$, which yields the claimed conclusion. □

An analogous result and proof holds for $p > 2$.

Lemma 12.5.7 (Entropy integral bound for $p > 2$)**.** Let $2 < p < \infty$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $0 \leq \mathbf{w} \in \mathbb{R}^n$ be γ-one-sided ℓ_p Lewis weights. Let $w = \max_{i\in[n]}\mathbf{w}_i$. Then,

$$\int_0^\infty \sqrt{\log E(B^p(\mathbf{A}), d_X, t)}\,dt \leq O(w^{1/2}\varepsilon\tau^{1/2}R)\left(\log\frac{\tau n}{\varepsilon}\right)^{1/2}\log d$$

Proof. Note that it suffices to integrate the entropy integral to $\operatorname{diam}(T)$, which is bounded in Lemma 12.5.4. Note also that T is just a translation of $(\eta R)^{1/p}\cdot B^p(\mathbf{A})$, so we have

$$\begin{aligned}\log E(T, d_X, t) &= \log E((\eta R)^{1/p}\cdot B^p(\mathbf{A}), d_X, t)\\ &= \log E((\eta R)^{1/p}\cdot B^p(\mathbf{A}), K\|\mathbf{W}^{-1/p}\mathbf{A}(\cdot)\|_{\mathbf{w},q}, t) \qquad \text{Lemma 12.5.4}\\ &= \log E(B_{\mathbf{w}}^p(\mathbf{W}^{-1/p}\mathbf{A}), B_{\mathbf{w}}^q(\mathbf{W}^{-1/p}\mathbf{A}), t/K(\eta R)^{1/p})\end{aligned}$$

where $K = O(w^{1/2}\tau^{1/2-1/p}R^{1-1/p})$.

For small radii less than λ for a parameter λ to be chosen, we use a standard volume argument, which shows that

$$\log E(B_{\mathbf{w}}^p(\mathbf{W}^{-1/p}\mathbf{A}), B_{\mathbf{w}}^q(\mathbf{W}^{-1/p}\mathbf{A}), t) \leq O(d)\log\frac{n}{t}$$

so

$$\int_0^\lambda \sqrt{\log E(T, d_X, t)}\,dt \leq \int_0^\lambda \sqrt{d\log\frac{nK(\eta R)^{1/p}}{t}}\,dt$$

$$
\begin{aligned}
&\le \lambda\sqrt{d\log(nw^{1/2}\eta^{1/p}\tau^{1/2-1/p})} + \sqrt{d}\int_0^\lambda \sqrt{\log\frac{R}{t}}\,dt \\
&\le \lambda\sqrt{d\log(nw^{1/2}\eta^{1/p}\tau^{1/2-1/p})} + \sqrt{d}\cdot O(\lambda)\sqrt{\log\frac{R}{\lambda}} \\
&\le O(\lambda)\sqrt{d\log\frac{nw^{1/2}\eta^{1/p}\tau^{1/2-1/p}R}{\lambda}}
\end{aligned}
$$

On the other hand, for large radii larger than λ, we use the bounds of Lemma 12.5.5. Note that the entropy bounds do not change if we replace $\mathbf{A}$ by $\mathbf{AR}$, where $\mathbf{R}$ is the change of basis matrix such that $\mathbf{W}^{1/2-1/p}\mathbf{AR}$ is orthonormal. Then by the properties of γ-one-sided ℓ_p Lewis weights (Lemma 6.2.1), we have

$$
\|\mathbf{e}_i^\top \mathbf{W}^{-1/p}\mathbf{AR}\|_2^2 = \mathbf{w}_i^{-2/p}\|\mathbf{e}_i^\top \mathbf{AR}\|_2^2 \le \gamma^{-1}.
$$

Then, Lemma 6.2.2 and Lemma 12.5.5 give

$$
\begin{aligned}
&\log E(B_{\mathbf{w}}^p(\mathbf{W}^{-1/p}\mathbf{A}), B_{\mathbf{w}}^q(\mathbf{W}^{-1/p}\mathbf{A}), t/K(\eta R)^{1/p}) \\
\le\ &\log E(B_{\mathbf{w}}^2(\mathbf{W}^{-1/p}\mathbf{A}), B_{\mathbf{w}}^q(\mathbf{W}^{-1/p}\mathbf{A}), t/K(\eta R)^{1/p}\|\mathbf{w}\|_1^{1/2-1/p}) \\
\le\ &\frac{K^2(\eta R)^{2/p}\|\mathbf{w}\|_1^{1-2/p}}{\gamma t^2}\log\frac{\tau n}{\varepsilon} \\
\le\ &\frac{O(w)\varepsilon^2\tau R^2}{t^2}\log\frac{\tau n}{\varepsilon}
\end{aligned}
$$

so the entropy integral gives a bound of

$$
O(w^{1/2}\varepsilon\tau^{1/2}R)\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\int_\lambda^{\operatorname{diam}(T)}\frac{1}{t}\,dt = O(w^{1/2}\varepsilon\tau^{1/2}R)\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log\frac{\operatorname{diam}(T)}{\lambda}.
$$

We choose $\lambda = \operatorname{diam}(T)/\sqrt{d}$, which yields the claimed conclusion. □

We are now ready to prove Theorem 12.4.4.

Proof of Theorem 12.4.4. We have by Lemma 2.3.7 that the Rademacher process is bounded by

$$
(2\mathcal{E})^l(\mathcal{E}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l
$$

where

$$
\mathcal{E} \le \begin{cases} O(w^{1/2}\gamma^{-1/2}\eta^{1/p}R)\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log d & p<2 \\ O(\varepsilon w^{1/2}\tau^{1/2}R)\Big(\log\frac{\tau n}{\varepsilon}\Big)^{1/2}\log d & p>2 \end{cases}
$$

by Lemmas 12.5.6 and 12.5.7 and

$$
\mathcal{D} \le \begin{cases} O(w^{1/2}\eta^{1/p}\gamma^{-1/2}R) & p<2 \\ O(\varepsilon w^{1/2}\tau^{1/2}R) & p>2 \end{cases}
$$

by Lemma 12.5.4. This gives a bound of $(\alpha R)^l$ on the Rademacher process, where

$$\alpha = \begin{cases} O(w^{1/2}\eta^{1/p}\gamma^{-1/2})\left[\left(\left(\log\frac{\tau n}{\varepsilon}\right)^{1/2}\log d\right)^{1+1/l} + \sqrt{l}\right] & p < 2 \\ O(\varepsilon w^{1/2}\tau^{1/2}R)\left[\left(\left(\log\frac{\tau n}{\varepsilon}\right)^{1/2}\log d\right)^{1+1/l} + \sqrt{l}\right] & p > 2 \end{cases}$$

We now set $\alpha = \varepsilon$ and solve for the ε that we can obtain. From this, we see that we can set

$$\varepsilon = \begin{cases} O(w\eta^{2/p})^{1/2}\gamma^{-1/2}\left[\left((\log d)^2\log n\right)^{1+1/l} + l\right]^{1/2} & p < 2 \\ O(w\eta\|\mathbf{w}\|_1^{p/2-1})^{1/p}\gamma^{-1/2}\left[\left((\log d)^2\log n\right)^{1+1/l} + l\right]^{1/p} & p > 2 \end{cases}.$$

□

12.6 Lower bounds

We now present our lower bounds on active sampling for ℓ_p regression.

Our main results show that the number of entries of $\mathbf{b}$ read by our upper bounds are nearly optimal up to polylogarithmic factors. Our first result is a lower bound of $\Omega(d/\varepsilon^2)$ for $p \in (0,1)$. Our lower bound is similar to Theorem 5.1 of [CD21], and is based on distinguishing biased coin flips. We also show that the same instance gives a lower bound $\Omega(d/\varepsilon)$ in the range of $p \in (1,2)$. For $d = 1$, we note that the active ℓ_p regression problem is equivalent to the ℓ_p *power means* problem. Thus, these results also improve upon a query complexity lower bound for this problem by [CSS21], which shows a lower bound of $\Omega(\varepsilon^{1-p})$ in one dimension. For $p > 2$, we generalize the $\Omega(\varepsilon^{1-p})$ lower bound argument of [CSS21] and show that $\Omega(d^{p/2}/\varepsilon^{p-1})$ entries must be read when $p > 2$, which is optimal in this regime.

Finally, our last lower bound concerns algorithms which solve the ℓ_p-regression problem $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p$ up to a constant factor in a specific way. Namely, say an algorithm is a *sampling-and-reweighting* algorithm if, given an $n \times d$ input matrix $\mathbf{A}$, the algorithm first reads $\mathbf{A}$ and then decides on a subset S of s entries of $\mathbf{b}$ to read in an arbitrary way. The algorithm also decides on a diagonal rescaling matrix $\mathbf{S} \in \mathbb{R}^{s\times s}$ – $\mathbf{S}$ may be arbitrary, except that we require that if two rows of $\mathbf{A}$ are identical and are both sampled in S, they are given the same weight in $\mathbf{S}$. We also assume that the number s of samples is a function of d, ε, and the failure probability δ, and is independent of n. These assumptions hold for all importance-based sampling methods for subspace preservation.

After deciding on S and $\mathbf{S}$, the algorithm then reads the entries in $\mathbf{b}$ indexed by the set S, denoted $\mathbf{b}_S$, and sets $\mathbf{x}' = \text{argmin}_x \|\mathbf{S}\mathbf{A}_S\mathbf{x} - \mathbf{S}\mathbf{b}_S\|_p$, where $\mathbf{A}_S$ is the subset of rows of $\mathbf{A}$ corresponding to the entries in $\mathbf{b}_S$. We show that any sampling-and-reweighting algorithm which fails with probability at most δ, necessarily takes $|S| = \Omega(1/\delta^{p-1})$ samples. Moreover, this remains true even if $\|\mathbf{b}\|_p = O(1)\cdot\min_{\mathbf{x}}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p$.

We stress that our main algorithms *are not* sampling-and-reweighting algorithms due to the success probability boosting steps of Section 12.2.1 which ensure that $\|\mathbf{b}\|_p = O(1)\cdot\min_{\mathbf{x}}\|\mathbf{A}\mathbf{x} -$

$\mathbf{b}\|_p$ and $\|\mathbf{S}\mathbf{b}\|_p = O(\|\mathbf{b}\|_p)$ with probability at least $1-\delta$. With these steps, our approach achieves a $O(\log 1/\delta)$ dependence overall, an exponential improvement over what is possible by simple sampling-and-reweighting algorithms. We also remark that our lower bound becomes vacuous when $p = 1$, which is required – Chen and Dereziński [CD21] as well as Parulekar, Parulekar, and Price [PPP21] achieve $O(\log(1/\delta))$ dependence with simple sampling-and-reweighting for ℓ_1 regression.

12.6.1 Lower bounds for $p \in (0, 1)$

We first show an $\Omega(d/\varepsilon^2)$ lower bound for $p \in (0, 1)$, which is tight up to logarithmic factors. The idea is essentially the same as the lower bound of [CD21]. We use Yao's minimax principle to restrict our attention to deterministic algorithms which must succeed with high probability over a random distribution over input instances.

We first recall the result of [CD21], which provides a generic reduction from d-dimensional lower bounds to 1-dimensional lower bounds via a padding argument. Although [CD21] prove a theorem online in the case of ℓ_1, the following result is an easy generalization that is implicit from their proof:

Theorem 12.6.1 (Theorem 5.1, [CD21])**.** Let $\mathcal{D}_0$ and $\mathcal{D}_1$ be two distributions over label vectors $\mathbf{b} \in \mathbb{R}^m$ such that distinguishing between $\mathbf{b} \sim \mathcal{D}_0$ and $\mathbf{b} \sim \mathcal{D}_1$ with probability at least $2/3$ requires at least q queries to $\mathbf{b}$ in expectation, for any deterministic algorithm. Furthermore, suppose that there exists $\mathbf{a} \in \mathbb{R}^m$ such that, with probability at least $99/100$, $\mathcal{D}_0$ and $\mathcal{D}_1$ can be distinguished by $\tilde{x} \in \mathbb{R}$ such that

$$\|\mathbf{a}\tilde{x} - \mathbf{b}\|_p^p \leq (1+\varepsilon) \min_{x \in \mathbb{R}} \|\mathbf{a}x - \mathbf{b}\|_p^p.$$

Finally, suppose that there exist $R > 0$ and $c \geq 1$ such that $\min_x \|\mathbf{a}x - \mathbf{b}\|_p^p \in [R, cR]$ with probability at least $99/100$ for $\mathbf{b} \sim \frac{1}{2}(\mathcal{D}_0 + \mathcal{D}_1)$. Then, there exists an $md \times d$ matrix $\mathbf{A}$ and a distribution $\mathcal{D}$ over label vectors $\mathbf{b} \in \mathbb{R}^{md}$ such that any deterministic algorithm which outputs $\tilde{\mathbf{x}} \in \mathbb{R}^d$ such that

$$\mathbf{Pr}\left\{ \|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|_p^p \leq \left(1 + \frac{\varepsilon}{200c}\right) \min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p \right\} \geq \frac{99}{100}$$

must make at least $\Omega(dq)$ queries to $\mathbf{b}$ in expectation.

Thus, it suffices to show a 1-dimensional lower bound which suits the hypotheses of Theorem 12.6.1. Our hard input distribution will be the same as that of [CD21, Theorem 5.1].

Theorem 12.6.2. Let $0 < p < 1$ be a constant. Let $\varepsilon > 0$ be sufficiently small and let $n = 100\lceil \varepsilon^{-2} \rceil$. Let $\mathbf{a} \in \mathbb{R}^n$ be the all ones vector. Let $\mathcal{D}_0$ be the distribution over binary vectors $\mathbf{b} \in \{0, 1\}^n$ which independently draws each coordinate as a Bernoulli with bias $1/2 + \varepsilon$ and let $\mathcal{D}_1$ be the distribution which independently draws each coordinate as a Bernoulli with bias $1/2 - \varepsilon$. Then, any $\tilde{x}$ such that

$$\|\mathbf{a}\tilde{x} - \mathbf{b}\|_p^p \leq (1+\varepsilon) \min_{x \in \mathbb{R}} \|\mathbf{a}x - \mathbf{b}\|_p^p$$

distinguishes whether $\mathbf{b} \sim \mathcal{D}_0$ or $\mathbf{b} \sim \mathcal{D}_1$ with probability at least $99/100$.

Proof. Note that the optimal x^* minimizing $\|\mathbf{a}x - \mathbf{b}\|_p^p$ over $x \in \mathbb{R}$ must lie in $[0, 1]$. Indeed, if $x < 0$, then $-x$ has a strictly lower cost than x, and if $x > 1$, then $x = 1$ has a strictly lower cost. Thus, the objective function can be written as

$$\|\mathbf{a}x - \mathbf{b}\|_p^p = (n - r) \cdot x^p + r \cdot (1 - x)^p \tag{12.4}$$

where r is the number of ones in $\mathbf{b}$. As noted by [CD21], note that $r \in [(\frac{1}{2} + \frac{\varepsilon}{2})n, (\frac{1}{2} + \frac{3\varepsilon}{2})n]$ with probability at least $99/100$ if $\mathbf{b} \sim \mathcal{D}_0$, and similarly, $r \in [(\frac{1}{2} - \frac{\varepsilon}{2})n, (\frac{1}{2} - \frac{3\varepsilon}{2})n]$ with probability at least $99/100$ if $\mathbf{b} \sim \mathcal{D}_1$. Let this event be denoted as $\mathcal{E}$, and condition on this event. Write this as $r = n/2 + a\varepsilon n$ for some $a \in [1/2, 3/2]$ if $\mathbf{b} \sim \mathcal{D}_0$ and $a \in [-3/2, -1/2]$ if $\mathbf{b} \sim \mathcal{D}_1$.

Optimal Solutions. We will first compute the optimal cost. Since $x \mapsto x^p$ for $p \in (0, 1)$ is nonconvex, we have three candidate solutions for the optimum: the endpoints $x = 0$, $x = 1$, and the unique stationary point

$$x = \frac{1}{1 + (n/r - 1)^{1/(p-1)}}$$

of Equation (12.4). It can easily be seen that the costs for $x = 0$ and $x = 1$ are r and $n - r$, respectively. Now assume that $\mathbf{b} \sim \mathcal{D}_0$, since the other case follows symmetrically. Then,

$$(n/r - 1)^{1/(p-1)} = \left[\frac{1}{1/2 + a\varepsilon} - 1\right]^{1/(p-1)} = \left[\frac{1/2 + a\varepsilon}{1/2 - a\varepsilon}\right]^{1/(1-p)} = 1 + O(\varepsilon)$$

so

$$x = \frac{1}{2 + O(\varepsilon)} = \frac{1}{2} - O(\varepsilon).$$

The objective cost is thus

$$\begin{aligned}
(n - r) \cdot x^p + r \cdot (1 - x)^p &= \left(\frac{1}{2} - a\varepsilon\right)n \cdot \left(\frac{1}{2} - O(\varepsilon)\right)^p + \left(\frac{1}{2} + a\varepsilon\right)n \cdot \left(\frac{1}{2} + O(\varepsilon)\right)^p \\
&= \frac{n}{2^p}\left[\left(\frac{1}{2} - a\varepsilon\right)(1 - O(\varepsilon))^p + \left(\frac{1}{2} + a\varepsilon\right)(1 + O(\varepsilon))^p\right] \\
&= \frac{n}{2^p}(1 \pm O(\varepsilon)).
\end{aligned}$$

Since $p < 1$, this has cost worse than r or $n - r$, so the optimal solution is $n - r$, conditioned on $\mathcal{E}$. Likewise, if $\mathbf{b} \sim \mathcal{D}_1$, then the optimal solution is $x = 0$ with cost r.

Suboptimal Solutions. We now show that, given a nearly optimal solution $\tilde{x}$, we can determine whether $\mathbf{b}$ is drawn from $\mathcal{D}_0$ or $\mathcal{D}_1$ with high probability by testing whether $\tilde{x} \in [1/2, 1]$ or $\tilde{x} \in [0, 1/2]$. Again, assume by symmetry that $\mathbf{b} \sim \mathcal{D}_0$. Suppose that $x \in [0, 1/2]$. If $x > 1/2 - O(\varepsilon)$, then as calculated above, x is not even an α-factor solution for some constant α. Otherwise, we have that

$$x \le \frac{1}{2} - O(\varepsilon) = \frac{1}{1 + (n/r - 1)^{1/(p-1)}}$$

which rearranges to

$$p \cdot (n-r)x^{p-1} - pr \cdot (1-x)^{p-1} > 0$$

which means that the objective is increasing on this interval. Thus, for $x \in [0, 1/2]$, the smallest that the cost can be is r, which is a factor of $(1+\varepsilon)$ larger than $n-r$. Thus, a $(1+\varepsilon)$-factor approximation must distinguish between $\mathbf{b}$ drawn from $\mathcal{D}_0$ and $\mathcal{D}_1$. □

As discussed in [CD21], the distributions of Theorem 12.6.2 require at least $\Omega(\varepsilon^{-2})$ queries to distinguish, by standard arguments. Then, by combining Theorems 12.6.1 and 12.6.2, we arrive at the following:

Theorem 12.6.3. Let $p \in (0,1)$ be a constant. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let $\mathbf{b} \in \mathbb{R}^n$. Suppose that with probability at least $99/100$, an algorithm $\mathcal{A}$ returns $\tilde{\mathbf{x}} \in \mathbb{R}^d$ such that

$$\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|_p^p \leq (1+\varepsilon) \min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p.$$

Then, $\mathcal{A}$ queries $\Omega(d/\varepsilon^2)$ entries of $\mathbf{b}$ in expectation.

12.6.2 Lower bounds for $p \in (1,2)$

In the range of $p \in (1,2)$, we analyze the same lower bound instance as in Theorem 12.6.2. However, the nature of the objective function changes in this parameter regime, and our lower bound weakens to $\Omega(d/\varepsilon)$. In particular, the value of the endpoints $x = 0$ and $x = 1$ stay at r and $n-r$, but the value of the stationary point, which is near $x = 1/2$ and has a value of around $n/2^p$, becomes significantly better than the endpoint solutions. This causes a phase transition in the lower bound that we are able to achieve with this method.

We now present our 1-dimensional lower bound of $\Omega(\varepsilon^{-1})$ for $p \in (1,2)$. For easy reuse of our calculations from Theorem 12.6.2, we state this result as a lower bound of $\Omega(\varepsilon^{-2})$ for any algorithm achieving an $O(\varepsilon^2)$-approximation. This can be reparameterized to an $\Omega(\varepsilon^{-1})$ lower bound for $O(\varepsilon)$-approximations.

Theorem 12.6.4. Let $1 < p < 2$ be a constant. Let $\varepsilon > 0$ be sufficiently small and let $n = 100\lceil \varepsilon^{-2} \rceil$. Let $\mathbf{a} \in \mathbb{R}^n$ be the all ones vector. Let $\mathcal{D}_0$ be the distribution over binary vectors $\mathbf{b} \in \{0,1\}^n$ which independently draws each coordinate as a Bernoulli with bias $1/2 + \varepsilon$ and let $\mathcal{D}_1$ be the distribution which independently draws each coordinate as a Bernoulli with bias $1/2 - \varepsilon$. Then, there exists a constant c such that any $\tilde{x}$ such that

$$\|\mathbf{a}\tilde{x} - \mathbf{b}\|_p^p \leq (1 + c \cdot \varepsilon^2) \min_{x \in \mathbb{R}} \|\mathbf{a}x - \mathbf{b}\|_p^p$$

distinguishes whether $\mathbf{b} \sim \mathcal{D}_0$ or $\mathbf{b} \sim \mathcal{D}_1$ with probability at least $99/100$.

Proof. Many of our calculations from Theorem 12.6.2 directly carry over. Recall our notation of setting r to be the number of ones in $\mathbf{b}$, which is $r = n/2 + a\varepsilon n$ for $a \in [1/2, 3/2]$ if $\mathbf{b} \sim \mathcal{D}_0$ and $a \in [-3/2, -1/2]$ if $\mathbf{b} \sim \mathcal{D}_1$. Recall also that the unique stationary point, which is the optimum now by convexity, is

$$x = \frac{1}{1 + (n/r - 1)^{1/(p-1)}}.$$

Optimal Solutions. We now calculate the value of the optimum. We carry out calculations for $\mathbf{b} \sim \mathcal{D}_0$ since $\mathbf{b} \sim \mathcal{D}_1$ gives symmetric results. Note first that

$$\begin{aligned}(n/r-1)^{1/(p-1)} &= \left[\frac{1}{1/2+a\varepsilon}-1\right]^{1/(p-1)} = \left[\frac{1-2a\varepsilon}{1+2a\varepsilon}\right]^{1/(p-1)} = \left[(1-2a\varepsilon)\cdot\sum_{i=0}^{\infty}(-2a\varepsilon)^i\right]^{1/(p-1)} \\ &= \left[(1-2a\varepsilon)-(2a\varepsilon)(1-2a\varepsilon)+(2a\varepsilon)^2(1-2a\varepsilon)+O(\varepsilon^3)\right]^{1/(p-1)} \\ &= \left[1-4a\varepsilon+2(2a\varepsilon)^2+O(\varepsilon^3)\right]^{1/(p-1)} \\ &= 1-\frac{4a}{p-1}\varepsilon+\frac{8a^2}{p-1}\varepsilon^2+O(\varepsilon^3)\end{aligned}$$

so

$$x = \frac{1}{2}\left(1+\frac{2a}{p-1}\varepsilon-\frac{4a^2}{p-1}\varepsilon^2+O(\varepsilon^3)\right).$$

Then, the objective value at this x is

$$\begin{aligned}&\frac{n}{2^p}\left[\left(\frac{1}{2}-a\varepsilon\right)\left(1+\frac{2a}{p-1}\varepsilon-\frac{4a^2}{p-1}\varepsilon^2+O(\varepsilon^3)\right)^p+\left(\frac{1}{2}+a\varepsilon\right)\left(1-\frac{2a}{p-1}\varepsilon+\frac{4a^2}{p-1}\varepsilon^2+O(\varepsilon^3)\right)^p\right] \\ =\ &\frac{n}{2^p}\left[\left(\frac{1}{2}-a\varepsilon\right)\left(1+\frac{2pa}{p-1}\varepsilon-\frac{4pa^2}{p-1}\varepsilon^2+O(\varepsilon^3)\right)+\left(\frac{1}{2}+a\varepsilon\right)\left(1-\frac{2pa}{p-1}\varepsilon+\frac{4pa^2}{p-1}\varepsilon^2+O(\varepsilon^3)\right)\right] \\ =\ &\frac{n}{2^p}\left[1+\left(\frac{1}{2}-a\varepsilon\right)\left(\frac{2pa}{p-1}\varepsilon-\frac{4pa^2}{p-1}\varepsilon^2+O(\varepsilon^3)\right)+\left(\frac{1}{2}+a\varepsilon\right)\left(-\frac{2pa}{p-1}\varepsilon+\frac{4pa^2}{p-1}\varepsilon^2+O(\varepsilon^3)\right)\right] \\ =\ &\frac{n}{2^p}\left[1-\frac{4pa^2}{p-1}\varepsilon^2+O(\varepsilon^3)\right].\end{aligned}$$

If $\mathbf{b} \sim \mathcal{D}_1$, then we have that

$$x = \frac{1}{2}\left(1-\frac{2a}{p-1}\varepsilon+\frac{4a^2}{p-1}\varepsilon^2+O(\varepsilon^3)\right)$$

with the same objective value.

Suboptimal Solutions. We now show that, given a nearly optimal solution $\tilde{x}$, we can determine whether $\mathbf{b}$ is drawn from $\mathcal{D}_0$ or $\mathcal{D}_1$ with high probability by testing whether $\tilde{x} \in [1/2, 1]$ or $\tilde{x} \in [0, 1/2]$. Again, assume by symmetry that $\mathbf{b} \sim \mathcal{D}_0$. Suppose that $x \in [0, 1/2]$. Then,

$$x \leq \frac{1}{2} < \frac{1}{1+(n/r-1)^{1/(p-1)}}$$

which rearranges to

$$p(n-r)\cdot x^{p-1} - pr\cdot(1-x)^{p-1} < 0$$

which means that the objective is decreasing on this interval. Thus, for $x \in [0, 1/2]$, the smallest that the cost can be is $x = 1/2$, which gives a value of

$$(n-r)\cdot(1/2)^p + r\cdot(1/2)^p = \frac{n}{2^p},$$

which is a factor of $1 - \Theta(\varepsilon^2)$ larger than the optimal solution. Thus, a $(1 + \Theta(\varepsilon^2))$-approximate solution can distinguish between $\mathcal{D}_0$ and $\mathcal{D}_1$. □

Then, by combining Theorems 12.6.1 and 12.6.4, we arrive at the following:

Theorem 12.6.5. Let $p \in (1, 2)$ be a constant. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let $\mathbf{b} \in \mathbb{R}^n$. Suppose that with probability at least $99/100$, an algorithm $\mathcal{A}$ returns $\tilde{\mathbf{x}} \in \mathbb{R}^d$ such that

$$\|\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}\|_p^p \leq (1 + \varepsilon) \min_{\mathbf{x} \in \mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p.$$

Then, $\mathcal{A}$ queries $\Omega(d/\varepsilon)$ entries of $\mathbf{b}$ in expectation.

12.6.3 Lower bounds for $p \in (2, \infty)$

We give our lower bound for $p > 2$ in this section.

Theorem 12.6.6. Let $p > 2$. Suppose that a randomized algorithm solves the ℓ_p regression up to a relative error of $(1 + \varepsilon/3)$ and queries m entries in expectation and is correct with probability at least 0.99. Then, $m = \Omega(d^{p/2}/\varepsilon^{p-1})$.

Proof. By Yao's minimax principle, we may assume that the algorithm is deterministic with correctness probability at least 0.99 over a distributional hard instance. We will need the construction from Theorem 11.3.2. Let S be the set given by Theorem 11.3.2 with $q = p/2$. Set $n = s \cdot d^{p/2}$ for $s = c/\varepsilon^{p-1}$ with c a sufficiently small constant to be determined. Then, we take our matrix to be the $n \times d$ matrix formed by taking s copies of each of the $d^{p/2}$ vectors in S. Furthermore, we take our target vector $\mathbf{b}$ to be the zero vector with probability $1/2$ and $d \cdot \mathbf{e}_I$ with probability $1/2$, where $I \sim [n]$ is a uniformly random index and $\mathbf{e}_i$ is the ith standard basis vector for $i \in [n]$.

Call the deterministic algorithm $\mathcal{A}$. Suppose for contradiction that $m \leq n/100$. Consider the sequence of entries of $\mathbf{b}$ read by $\mathcal{A}$ when $\mathbf{b} = 0$. Note that this sequence is of length at most $2m$, since otherwise $\mathcal{A}$ already reads more than m entries in expectation. Furthermore, $\mathcal{A}$ must output $\mathbf{x} = 0$ as the solution if it reads a sequence of $2m$ entries of zeros, since otherwise $\mathcal{A}$ cannot achieve any relative error. Then since $\mathcal{A}$ is deterministic, $\mathcal{A}$ will always output $\mathbf{x} = 0$ if it reads $2m$ entries of zeros.

On the other hand, suppose that $\mathbf{b} = d \cdot \mathbf{e}_I$ for $I \sim [n]$. We first upper bound the optimal cost. If we choose $\mathbf{x} = \varepsilon \cdot \mathbf{a}_I$, then for the nonzero row of $\mathbf{b}$, we pay a cost of

$$(d - \varepsilon \cdot \langle \mathbf{a}_I, \mathbf{a}_I \rangle)^p = (1 - \varepsilon)^p d^p \leq (1 - \varepsilon) d^p.$$

For the other rows of $\mathbf{A}$ corresponding to copies of $\mathbf{a}_I$, we pay a cost of

$$s \cdot (\varepsilon \cdot \langle \mathbf{a}_I, \mathbf{a}_I \rangle)^p = \frac{c}{\varepsilon^{p-1}} \cdot \varepsilon^p \cdot d^p = c\varepsilon d^p.$$

For all other rows of $\mathbf{A}$ for $\mathbf{a}_j \neq \mathbf{a}_I$, we pay a cost of

$$s \cdot d^{p/2} \cdot (\varepsilon \cdot \langle \mathbf{a}_I, \mathbf{a}_j \rangle)^p = \frac{c}{\varepsilon^{p-1}} \cdot \varepsilon^p \cdot d^{p/2} \cdot C_q^p d^{p/2} = cC_q^p \varepsilon d^p.$$

Thus, if we choose $c \leq \min\{C_q^p, 1\}/3$, then the total cost is at most

$$(1-\varepsilon)d^p + c\varepsilon d^p + cC_q^p \varepsilon d^p \leq (1-\varepsilon/3)d^p.$$

Now note that if $\mathbf{b} = d \cdot \mathbf{e}_I$, then the probability that I lands on one of the $2m$ entries read by $\mathcal{A}$ when $\mathbf{b} = 0$ is at most $2m/n \leq 1/50$. Thus, with probability at least $1 - 1/50$, $\mathcal{A}$ outputs $\mathbf{x} = 0$ on this instance, which has a cost of d^p. By the above calculation, this fails to be a $(1+\varepsilon/3)$-approximate solution, which contradicts the guarantee of $\mathcal{A}$. We thus conclude that $m \geq n/100 = \Omega(d^{p/2}/\varepsilon^{p-1})$. □

12.6.4 A $1/\delta^{p-1}$ lower bound for sampling-and-reweighting algorithms

We next show that sampling-and-reweighting algorithms for ℓ_p regression must pay a polynomial dependence in the failure probability δ, contrasting with the logarithmic dependence achieved by our approach.

Theorem 12.6.7. Let $p > 1$. Any sampling-and-reweighting algorithm which, with probability at least $1 - \delta$, outputs a $(1+\epsilon)$-approximate solution $\mathbf{x}$ to the ℓ_p-regression problem, for $\epsilon > 0$ less than a sufficiently small constant, requires reading $\Omega(1/\delta^{p-1})$ entries of $\mathbf{b}$.

Proof. In our hard instance we will have $d = 1$ and require a sufficiently fine constant factor approximation with failure probability δ. Suppose, with these parameters, that there is an algorithm reading s entries. We set $n = s/\delta$, and will show that the algorithm cannot output a constant factor approximation to the ℓ_p-regression problem with probability at least $1 - \delta$.

Let $\mathbf{A}$ be a single column of n 1s. Since the entries of $\mathbf{A}$ are indistinguishable from each other, we can assume without loss of generality that the sampling-and-reweighting algorithm samples entries uniformly at random. By assumption, since the rows of $\mathbf{A}$ are all identical, the algorithm reweights the sampled rows uniformly (equivalently assigns weight 1 to each sampled entry). We choose $\mathbf{b} = \mathbf{e}_I$ for a random standard basis vector $\mathbf{e}_I$. For the optimal x, necessarily $0 \leq x \leq 1$ since if $x < 0$, replacing x with $-x$ would give lower cost. Similarly, if $x > 1$, then replacing x with 1 would give lower cost. Then the cost is $(1-x)^p + (n-1)x^p$. This is convex and differentiable for $p > 1$, and is minimized when the derivative is 0. Differentiating, the optimal x satisfies $-p(1-x)^{p-1} + p(n-1)x^{p-1} = 0$, or $(1-x)^{p-1} = (n-1)x^{p-1}$. Taking $(p-1)$-th roots, $1 - x = (n-1)^{1/(p-1)}x$, or $x = 1/(1+(n-1)^{1/(p-1)})$. The optimal cost is therefore

$$\left(1 - \frac{1}{1+(n-1)^{1/(p-1)}}\right)^p + \frac{n-1}{(1+(n-1)^{1/(p-1)})^p}.$$

For $n = \omega(1)$, this is $\Theta(1 + n^{1-p/(p-1)}) = O(1)$ for any constant $p > 1$.

On the other hand, given that $n = s/\delta$, with probability at least δ, the algorithm's sample includes $\mathbf{b}_I$. If this is the case, for the sampled problem, the cost is $(1-x)^{p-1} + (s-1)x^p$ for a given x. Setting the derivative to 0, we now have that the optimal x' for the sampled problem is: $x' = \frac{1}{1+(s-1)^{1/(p-1)}}$. Computing the cost of using x' for the original problem, our cost of using x' is

$$\left(1 - \frac{1}{(1+(s-1)^{1/(p-1)})^p}\right) + \frac{(n-1)}{(1+(s-1)^{1/(p-1)})^p}.$$

The cost is at least the second term, which for $n = \omega(1)$ and $s = \delta n = \omega(1)$ is $\Theta((s/\delta)/s^{p/(p-1)})$. This term must be $O(1)$ to be an $O(1)$-approximation, by our above calculation of the optimal cost. Hence, $s^{p/(p-1)-1} = \Omega(1/\delta)$, or $s^{1/(p-1)} = \Omega(1/\delta)$, or $s = \Omega(1/\delta^{p-1})$. □

Chapter 13

Applications: coresets for multiple ℓ_p regression [WY24a]

13.1 Multiple ℓ_p regression

Up until now, we have focused mostly on least squares and ℓ_p linear regression problems with a *single* response, i.e., there is just a single $\mathbf{b}$ vector of responses. However, it is often the case that we are interested in more than just one target $\mathbf{b}$ to predict, and in general, we may wish to simultaneously fit m target vectors that are given by a matrix $\mathbf{B} \in \mathbb{R}^{n \times m}$ and solve the minimization problem

$$\min_{\mathbf{X} \in \mathbb{R}^{d \times m}} \|\mathbf{AX} - \mathbf{B}\|_{p,p}^p = \min_{\mathbf{X} \in \mathbb{R}^{d \times m}} \sum_{j=1}^{m} \|\mathbf{AX}\mathbf{e}_j - \mathbf{B}\mathbf{e}_j\|_p^p$$

This is known as the *multiple response ℓ_p regression* problem, or simply the *multiple ℓ_p regression* problem, and is the focus of the present chapter.

13.1.1 Coreset constructions for $p = 2$

For $p = 2$, the construction of strong coresets for the multiple response problem follows almost immediately from strong coresets for the single response problem due to orthogonality and the Pythagorean theorem, and we can construct $\mathbf{S}$ such that

$$\|\mathbf{S}(\mathbf{AX} - \mathbf{B})\|_F^2 = (1 \pm \varepsilon)\|\mathbf{AX} - \mathbf{B}\|_F^2$$

with $\mathsf{nnz}(\mathbf{S}) = \tilde{O}(\varepsilon^{-2}d)$ samples. Indeed, assume without loss of generality that $\mathbf{A}$ has orthogonal columns, and suppose that $\mathbf{S}$ satisfies

- $\|\mathbf{SAx}\|_2^2 = (1 \pm \varepsilon)\|\mathbf{Ax}\|_2^2$ for every $\mathbf{x} \in \mathbb{R}^d$ (i.e. $\mathbf{S}$ is a subspace embedding)
- $\|\mathbf{S}(\mathbf{AX}^* - \mathbf{B})\|_F^2 = (1 \pm \varepsilon)\|\mathbf{AX}^* - \mathbf{B}\|_F^2$ where $\mathbf{X}^*$ is the optimal minimizer
- $\|\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}(\mathbf{AX}^* - \mathbf{B})\|_F^2 \leq (\varepsilon^2/d)\|\mathbf{A}\|_F^2 \|\mathbf{AX}^* - \mathbf{B}\|_F^2 = \varepsilon^2 \|\mathbf{AX}^* - \mathbf{B}\|_F^2$

Then, the following argument of Section 7.5 of [CW13] shows that $\mathbf{S}$ is a strong coreset. Indeed,

$$\begin{aligned}\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\|_F^2 &= \|\mathbf{S}\mathbf{A}(\mathbf{X}-\mathbf{X}^*)\|_F^2 + \|\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\|_F^2 \\ &\quad + 2\operatorname{tr}\big((\mathbf{X}-\mathbf{X}^*)^\top\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\big)\end{aligned}$$

by expanding the square, and the inner product term is bounded by

$$\begin{aligned}\big|\operatorname{tr}\big((\mathbf{X}-\mathbf{X}^*)^\top\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\big)\big| &\le \|\mathbf{X}-\mathbf{X}^*\|_F\|\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\|_F \\ &\le \varepsilon\|\mathbf{A}(\mathbf{X}-\mathbf{X}^*)\|_F\|\mathbf{A}\mathbf{X}^*-\mathbf{B}\|_F \\ &\le \varepsilon\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_F^2\end{aligned}$$

and $\mathbf{S}$ also preserves the quantities $\|\mathbf{S}\mathbf{A}(\mathbf{X}-\mathbf{X}^*)\|_F^2$ and $\|\mathbf{S}(\mathbf{A}\mathbf{X}^*-\mathbf{B})\|_F^2$ up to $(1\pm\varepsilon)$ relative error. A similar trick is available in the weak coreset setting (see, e.g., Section 3.1 of [CNW16]), which gives a bound of $\mathsf{nnz}(\mathbf{S}) = \tilde{O}(\varepsilon^{-1}d)$ for this guarantee. Unfortunately, almost every step in the above argument uses special properties of the ℓ_2 norm that are not available for the ℓ_p norm, and thus we will need completely different arguments to handle $p \neq 2$.

13.1.2 Challenges for $p \neq 2$

If we desire only weak coresets, then prior results on active ℓ_p regression (see Chapter 12) in fact almost immediately provide a solution. These results show that a weak coreset $\mathbf{S}$ for the single response ℓ_p regression problem can be constructed independently of $\mathbf{b}$, and with the dependence of $\mathsf{nnz}(\mathbf{S})$ on the failure probability δ being polylogarithmic. Thus by setting the failure rate to $\delta = 1/10m$, we can simultaneously solve every column of $\mathbf{B}$ independently with overall probability at least $9/10$.

For strong coresets, however, such a column-wise strategy must be implemented carefully. If we consider constructing a strong coreset for a single column $j \in [m]$, then the sampling probabilities now depend on the target vector $\mathbf{B}\mathbf{e}_j$, so the sampling complexity would need to scale as m rather than $\operatorname{poly}\log(m)$ as in the previous upper bound weak coresets. On the other hand, another natural strategy is to mimic the strategy for the $p = 2$ case and take the sampling probabilities to only guarantee a ℓ_p subspace embedding for the column space of $\mathbf{A}$ and that $q_i \ge \|\mathbf{e}_i^\top\mathbf{B}^*\|_p^p/\|\mathbf{B}^*\|_{p,p}^p$ for $\mathbf{B}^* \coloneqq \mathbf{A}\mathbf{X}^* - \mathbf{B}$. This is a reasonable choice of sampling probabilities, and indeed it is not hard to see that

$$\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\|_{p,p}^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_{p,p}^p$$

for any fixed $\mathbf{X} \in \mathbb{R}^{d\times m}$ with only $\mathsf{nnz}(\mathbf{S}) = \tilde{O}(\varepsilon^{-2}d)$ samples for $p < 2$ and $\mathsf{nnz}(\mathbf{S}) = \tilde{O}(\varepsilon^{-2}d^{p/2})$ samples for $p > 2$ via a Bernstein tail bound. However, it is unclear how to extend a guarantee for any single $\mathbf{X} \in \mathbb{R}^{d\times m}$ to a guarantee simultaneously for *all* $\mathbf{X} \in \mathbb{R}^{d\times m}$. Although the dependence on the failure rate δ is logarithmic, a net argument, or even more sophisticated chaining arguments, over the possible choices of $\mathbf{X} \in \mathbb{R}^{d\times m}$ seem to require a union bound over sets of size $\exp(dm)$, thus again introducing a linear dependence on m in the sample complexity $\mathsf{nnz}(\mathbf{S})$. As we show, a careful blend of these two ideas will be necessary to obtain our strong coreset result.

13.1.3 Strong coresets for multiple ℓ_p regression

Our first main result is the first construction of strong coresets for multiple ℓ_p regression that is independent of m.

Theorem 13.1.1 (Strong coresets for multiple ℓ_p regression)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, and $p \geq 1$. There is an algorithm which constructs $\mathbf{S}$ with

$$\mathsf{nnz}(\mathbf{S}) = \begin{cases} \frac{O(d)}{\varepsilon^2}\left[(\log d)^2 \log\frac{d}{\varepsilon} + \log\frac{1}{\delta}\right] & 1 \leq p < 2 \\ \frac{O(d^{p/2})}{\varepsilon^p}\left[(\log d)^2 \log\frac{d}{\varepsilon} + \log\frac{1}{\delta}\right] & p > 2 \end{cases}$$

such that with probability at least $1-\delta$,

$$\|\mathbf{S}(\mathbf{AX}-\mathbf{B})\|_{p,p}^p = (1\pm\varepsilon)\|\mathbf{AX}-\mathbf{B}\|_{p,p}^p$$

simultaneously for every $\mathbf{X} \in \mathbb{R}^{d\times m}$.

We achieve a nearly optimal dependence on d and ε, as we show that $\Omega(d^{p/2}/\varepsilon^p)$ rows are necessary for strong coresets in Theorem 13.6.1 for $p > 2$, while it is known that $\tilde{\Omega}(d/\varepsilon^2)$ rows are necessary even for $m = 1$ for $p < 2$ [LWW21]. We note that our upper bound shows that multiple ℓ_p regression is as easy as single response ℓ_p regression for $p < 2$, while our lower bound demonstrates an interesting separation between the two for $p > 2$.

Initial $\log m$ bound

We first recall Theorem 12.3.2 from Chapter 12 which shows that

$$\left|\left(\|\mathbf{S}(\mathbf{Ax}-\mathbf{b})\|_p^p - \|\mathbf{Sb}\|_p^p\right) - \left(\|\mathbf{Ax}-\mathbf{b}\|_p^p - \|\mathbf{b}\|_p^p\right)\right| \leq \varepsilon\left(\|\mathbf{b}\|_p^p + \|\mathbf{Sb}\|_p^p + \|\mathbf{Ax}\|_p^p\right) \quad (13.1)$$

This guarantee is in a form that can be summed over the m columns of $\mathbf{B}$. Thus, if a $\log m$ dependence is admissible, then we can apply the above result with failure probability $1/10m$, union bound over the m columns, and sum the results to obtain

$$\left|(\|\mathbf{S}(\mathbf{AX}-\mathbf{B})\|_{p,p}^p - \|\mathbf{SB}\|_{p,p}^p) - (\|\mathbf{AX}-\mathbf{B}\|_{p,p}^p - \|\mathbf{B}\|_{p,p}^p)\right| \leq \varepsilon\left(\|\mathbf{B}\|_{p,p}^p + \|\mathbf{SB}\|_{p,p}^p + \|\mathbf{AX}\|_{p,p}^p\right).$$

Now suppose that we additionally have

- $\|\mathbf{SB}\|_{p,p}^p = (1\pm\varepsilon)\|\mathbf{B}\|_{p,p}^p$
- $\|\mathbf{B}\|_{p,p}^p = O(\mathsf{OPT}^p)$ (which is without loss of generality by subtracting an $O(1)$-optimal solution)

Then, we have

$$\begin{aligned}\|\mathbf{S}(\mathbf{AX}-\mathbf{B})\|_{p,p}^p &= \|\mathbf{AX}-\mathbf{B}\|_{p,p}^p - \|\mathbf{B}\|_{p,p}^p + \|\mathbf{SB}\|_{p,p}^p \pm O(\varepsilon)\left(\|\mathbf{B}\|_{p,p}^p + \|\mathbf{AX}\|_{p,p}^p\right) \\ &= \|\mathbf{AX}-\mathbf{B}\|_{p,p}^p \pm \varepsilon\|\mathbf{B}\|_{p,p}^p \pm O(\varepsilon)\left(\|\mathbf{B}\|_{p,p}^p + \|\mathbf{AX}\|_{p,p}^p\right) \\ &= \|\mathbf{AX}-\mathbf{B}\|_{p,p}^p \pm O(\varepsilon)\|\mathbf{AX}-\mathbf{B}\|_{p,p}^p\end{aligned}$$

so we indeed have a strong coreset as desired.

Removing the m dependence

Next, we show how to completely remove the m dependence, which requires additional ideas. When applying (13.1) to each of the m columns, suppose that we set the failure probability to $\mathrm{poly}(\varepsilon\delta)$ instead of $O(1/m)$. Then, this guarantee will hold for a $1-\mathrm{poly}(\varepsilon\delta)$ fraction of "good" columns, for which we can obtain $(1\pm\varepsilon)$ approximations. On the remaining $\mathrm{poly}(\varepsilon\delta)$ fraction of "bad" columns, note that the mass of $\mathbf{B}$ on these columns is at most $\mathrm{poly}(\varepsilon\delta)\|\mathbf{B}\|_{p,p}^p$ with probability $1-\delta$ by Markov's inequality. Then on these columns, $\|\mathbf{S}(\mathbf{AX}-\mathbf{B})\mathbf{e}_j\|_p$ is just $\|\mathbf{SAXe}_j\|_p$ up to a small total additive error of $\mathrm{poly}(\varepsilon\delta)\|\mathbf{B}\|_{p,p}^p$. In turn, we have that $\|\mathbf{SAXe}_j\|_p = (1\pm\varepsilon)\|\mathbf{AXe}_j\|_p$ by using that $\mathbf{S}$ is an ℓ_p subspace embedding. Thus, by combining with the $(1\pm\varepsilon)$ approximation on the rest of the "good" columns, we can still ensure that $\|\mathbf{S}(\mathbf{AX}-\mathbf{B})\|_{p,p} = (1\pm\varepsilon)\|\mathbf{AX}-\mathbf{B}\|_{p,p}$.

13.1.4 Weak coresets for multiple ℓ_p regression

In the weak coreset setting, we consider a generalized multiple ℓ_p regression problem, where we are given a design matrix $\mathbf{A}\in\mathbb{R}^{n\times d}$, an "embedding" $\mathbf{G}\in\mathbb{R}^{t\times m}$, and a target matrix $\mathbf{B}\in\mathbb{R}^{n\times m}$, and we wish to approximately minimize the objective function $\|\mathbf{AXG}-\mathbf{B}\|_{p,p}$.

As noted previously, for multiple ℓ_p regression without an embedding (i.e. $\mathbf{G}=\mathbf{I}_t$) the construction of weak coresets follows relatively straightforwardly by applying active ℓ_p regression results along each column. However, this strategy fails when we must additionally handle the embedding matrix $\mathbf{G}$, as this constraint couples the columns of $\mathbf{AX}$ together. Furthermore, we argue that handling the embedding $\mathbf{G}$ is substantially more interesting that the unconstrained case. Indeed, as we see later in Sections 13.1.5 and 13.1.6, the incorporation of the embedding $\mathbf{G}$ will allow us to handle interesting extensions of our results to settings beyond the entrywise ℓ_p norm via the use of a linear embedding into this norm. We will denote the optimal value as

$$\mathsf{OPT} \coloneqq \min_{\mathbf{X}\in\mathbb{R}^{d\times t}} \|\mathbf{AXG}-\mathbf{B}\|_{p,p}$$

and let $\mathbf{X}^*$ denote the matrix achieving this optimum unless otherwise noted. We will prove the following result:

Theorem 13.1.2 (Weak coresets for multiple ℓ_p regression)**.** Let $\mathbf{A}\in\mathbb{R}^{n\times d}$, $\mathbf{G}\in\mathbb{R}^{t\times m}$, $\mathbf{B}\in\mathbb{R}^{n\times m}$, and $1\le p<\infty$. There is an algorithm which constructs $\mathbf{S}$ independently of $\mathbf{B}$ with

$$\mathsf{nnz}(\mathbf{S}) = \begin{cases} \dfrac{O(d)}{\varepsilon^2\delta^2}\left[(\log d)^2\log\dfrac{d}{\varepsilon}+\log\dfrac{1}{\delta}\right]\left(\log\log\dfrac{1}{\varepsilon}\right)^2 & p=1 \\ \dfrac{O(d)}{\varepsilon\delta^2}\left[(\log d)^2\log\dfrac{d}{\varepsilon}+\log\dfrac{1}{\delta}\right]\left(\log\log\dfrac{1}{\varepsilon}\right)^2 & 1<p<2 \\ \dfrac{O(d^{p/2})}{\varepsilon^{p-1}\delta^p}\left[(\log d)^2\log\dfrac{d}{\varepsilon}+\log\dfrac{1}{\delta}\right]\left(\log\log\dfrac{1}{\varepsilon}\right)^p & 2<p<\infty \end{cases}$$

such that with probability at least $1-\delta$, for any $\hat{\mathbf{X}}\in\mathbb{R}^{d\times t}$ such that

$$\|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B})\|_{p,p}^p \le (1+\varepsilon)\min_{\mathbf{X}\in\mathbb{R}^{d\times t}}\|\mathbf{S}(\mathbf{AXG}-\mathbf{B})\|_{p,p}^p,$$

we have

$$\|\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B}\|_{p,p}^p \leq (1+O(\varepsilon)) \min_{\mathbf{X}\in\mathbb{R}^{d\times t}} \|\mathbf{A}\mathbf{X}\mathbf{G}-\mathbf{B}\|_{p,p}^p.$$

Furthermore, conditioned on the event that $\|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p = O(\|\mathbf{A}\mathbf{X}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p)$ for the global optimizer $\mathbf{X}^*$, the dependence on δ can be replaced by a single $\log\frac{1}{\delta}$ factor and the $\mathrm{poly}(\log\log\frac{1}{\varepsilon})$ factor can be removed.

We achieve a nearly optimal dependence on d and ε, as we show that $\Omega(d^{p/2}/\varepsilon^{p-1})$ rows are necessary for weak coresets in Theorem 13.6.2 for $p>2$. Our weak coreset upper bound result together with our strong coreset lower bound of Theorem 13.6.1 shows a tight ε factor separation between the two coreset guarantees. Note that in the statement of Theorem 13.1.2, the dependence on the failure rate δ is polynomial. This occurs for the same reason as discussed in Theorem 12.6.7 for the active ℓ_p regression setting, and can be handled in the same way (see Section 12.3.3).

13.1.5 Applications: sublinear algorithms for Euclidean power means

Our first application of our results on coresets for multiple ℓ_p regression is on designing coresets for the Euclidean power means problem. In this problem, we are given as input a set of n points $\{\mathbf{b}_i\}_{i=1}^n \subseteq \mathbb{R}^t$, and we wish to find a center $\hat{\mathbf{x}}\in\mathbb{R}^t$ that minimizes the sum of the Euclidean distances to $\hat{\mathbf{x}}$, raised to the power p. That is, we seek to minimize the objective function given by

$$\sum_{i=1}^n \|\mathbf{x}-\mathbf{b}_i\|_2^p = \|\mathbf{1}\mathbf{x}^\top-\mathbf{B}\|_{p,2}^p$$

where $\mathbf{1}$ is the $n\times 1$ matrix of all ones, $\mathbf{B}\in\mathbb{R}^{n\times t}$ is the matrix with $\mathbf{b}_i$ in its n rows, and $\|\cdot\|_{p,2}$ is the $(p,2)$-norm of a matrix given by the ℓ_p norm of the Euclidean norm of the rows. This is a fundamental problem which generalizes the well-studied problems of the mean ($p=2$), geometric median ($p=1$), and minimum enclosing balls ($p=\infty$). Coresets and sampling algorithms for this problem were recently studied by [CSS21], who showed that a uniform sample of $\tilde{O}(\varepsilon^{-(p+3)})$ points suffices to output a center $\hat{\mathbf{x}}\in\mathbb{R}^t$ such that

$$\|\mathbf{1}\hat{\mathbf{x}}^\top-\mathbf{B}\|_{p,2}^p \leq (1+\varepsilon)\min_{\mathbf{x}\in\mathbb{R}^t}\|\mathbf{1}\mathbf{x}^\top-\mathbf{B}\|_{p,2}^p = (1+\varepsilon)\,\mathsf{OPT}^p\,.$$

In comparison to the upper bounds, the lower bounds given by [CSS21] was $\Omega(\varepsilon^{-(p-1)})$ which is off by a ε^4 factor compared to the upper bound, which was improved to $\Omega(\varepsilon^{-1})$ for $1<p<2$ by [MMWY22] and $\Omega(\varepsilon^{-2})$ for $p=1$ by [CD21, PPP21].

One of the main open questions highlighted by the work of [CSS21] is to obtain tight bounds for this problem: how many uniform samples are necessary and sufficient to output a $(1+\varepsilon)$-approximate solution to the Euclidean power means problem. Our main contribution is a nearly optimal algorithm which matches the lower bounds of [CD21, PPP21, CSS21, MMWY22].

Theorem 13.1.3. Let $\{\mathbf{b}_i\}_{i=1}^n \subseteq \mathbb{R}^d$. Then, there is a sublinear algorithm which uniformly samples at most

$$s=\begin{cases} O(\varepsilon^{-2})\big(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\big)\log\frac{1}{\delta} & p=1\\ O(\varepsilon^{-1})\big(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\big)\log\frac{1}{\delta} & 1<p\leq 2\\ O(\varepsilon^{1-p})\big(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\big)\log\frac{1}{\delta} & 2<p<\infty \end{cases}$$

rows $\mathbf{b}_i$ and outputs a center $\hat{\mathbf{x}}$ such that

$$\sum_{i=1}^{n} \|\hat{\mathbf{x}} - \mathbf{b}_i\|_2^p \leq (1+\varepsilon) \min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^{n} \|\mathbf{x} - \mathbf{b}_i\|_2^p$$

with probability at least $1 - \delta$.

To apply the techniques developed in this work to the Euclidean power means problem, we need to embed the $(p, 2)$-norm into the entrywise ℓ_p norm. To make this reduction, we use a classic result of Dvoretzky and Milman [Dvo61, Mil71], which shows that a random subspace of a normed space is approximately Euclidean (see Theorem 14.3.1).

Note then that if $\mathbf{G}$ is an appropriately scaled random Gaussian matrix, then we have that

$$\|\mathbf{1}\mathbf{x}^\top - \mathbf{B}\|_{p,2}^p = (1 \pm \varepsilon)\|\mathbf{1}\mathbf{x}^\top\mathbf{G} - \mathbf{B}\mathbf{G}\|_{p,p}^p$$

by the above result. We may now note that the latter optimization problem is exactly of the form of an embedded ℓ_p regression problem, and thus our weak coreset results immediately apply to this problem. In fact, handling this Dvoretzky embedding is our main motivation for studying the ℓ_p regression problem with the embedding. We also note that similar reductions are possible by making use of other linear embeddings between ℓ_p norms [WW19, LWY21, LLW23]. The full argument is given in Appendix 13.4.

In addition to sharpening the bound of [CSS21] to optimality, we note that our techniques, both algorithmically and in the analysis, are far simpler than the prior work of [CSS21]. The previous algorithm required partitioning the dataset into "rings" of points with similar costs and preprocessing these rings. Furthermore, the analysis uses a specially designed chaining argument with custom net constructions that require terminal Johnson–Lindenstrauss embeddings. On the other hand, our algorithm simply runs multiple instances of a "sample-and-solve" algorithm, where the run with lowest sampled mass is kept. Furthermore, the analysis largely builds on existing net constructions for ℓ_p regression, and does not need terminal embeddings.

13.1.6 Applications: spanning coresets for ℓ_p subspace approximation

As a second application of our results, we give the first construction of *spanning coresets for ℓ_p subspace approximation* with nearly optimal size. The ℓ_p subspace approximation is a popular generalization of the classic Frobenius norm low rank approximation problem, where the input is a set of n points $\{\mathbf{a}_i\}_{i=1}^n$ in d dimensions, and we wish to compute a rank k subspace $F \subseteq \mathbb{R}^d$ that minimizes

$$\sum_{i=1}^{n} \|\mathbf{a}_i^\top(\mathbf{I}_d - \mathbf{P}_F)\|_2^p$$

where $\mathbf{P}_F$ denotes the orthogonal projection matrix onto F. Equivalently, we can write this as

$$\min_{\mathrm{rank}(F)\leq k} \|\mathbf{A}(\mathbf{I}_d - \mathbf{P}_F)\|_{p,2}^p.$$

We also refer to Chapter 14 for further discussion of ℓ_p subspace approximation.

While strong and weak coresets for this problem have attracted much attention [FL11, SV12, SW18, HV20, FKW21, WY23a, WY24b], our main contribution to this line of research is on a different coreset guarantee, which we call *spanning coresets*. Spanning coresets are subsets of the points $\mathbf{a}_i$ which span a $(1+\varepsilon)$-optimal rank k subspace, and is another popular guarantee in this literature [DV07, SV12, CW15a]. In addition to being an interesting object in its own right [SV12], the existence of small spanning coresets have found applications to constructions for strong and weak coresets for ℓ_p subspace approximation [HV20].

Definition 13.1.4 (Spanning coreset)**.** Let $\{\mathbf{a}_i\}_{i=1}^n \subseteq \mathbb{R}^d$. A subset $S \subseteq [n]$ is a $(1+\varepsilon)$-*spanning coreset* if the points $\{\mathbf{a}_i\}_{i\in S}$ span a k-dimensional subspace $\hat{F}$ such that

$$\|\mathbf{A}(\mathbf{I}_d - \mathbf{P}_{\hat{F}})\|_{p,2}^p \leq (1+\varepsilon) \min_{\text{rank}(F)\leq k} \|\mathbf{A}(\mathbf{I}_d - \mathbf{P}_F)\|_{p,2}^p.$$

Our main result is the following upper bound on the size of spanning coresets.

Theorem 13.1.5. Let $\{\mathbf{a}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, $1 \leq p < \infty$, $k \in \mathbb{N}$, and $0 < \varepsilon < 1$. Then, there exists a $(1+\varepsilon)$-spanning coreset S of size at most

$$|S| = \begin{cases} O(\varepsilon^{-2}k)(\log(k/\varepsilon))^3 & p = 1 \\ O(\varepsilon^{-1}k)(\log(k/\varepsilon))^3 & 1 < p \leq 2 \\ O(\varepsilon^{1-p}k^{p/2})(\log(k/\varepsilon))^3 & 2 < p < \infty \end{cases}$$

In particular, we improve the previous best result of $O(\varepsilon^{-1}k^2\log(k/\varepsilon))$ due to Theorem 3.1 of [SV12] in the k dependence for all $1 \leq p < 4$. The proof of this result is given in Section 13.5. Furthermore, we give the first lower bounds on the size of spanning coresets by generalizing an argument of [DV06] for $p = 2$, showing that spanning coresets must have size at least $\Omega(\varepsilon^{-1}k)$ in Theorem 13.6.3. Together, our results settles the size of spanning coresets up to polylogarithmic factors for $1 < p < 2$. To obtain this result, we again use Dvoretzky's theorem to embed the problem to an embedded entrywise ℓ_p norm problem, and then apply our weak coreset results.

Finally, we note that our spanning coreset lower bound implies other interesting lower bounds for coresets. First, we note that weak coresets for ℓ_p subspace approximation are automatically spanning coresets, so our lower bound for spanning coresets also gives the first nontrivial lower bound on the size of weak coresets for ℓ_p subspace approximation. Secondly, we note that our proof of Theorem 13.1.5 in fact shows that any upper bound on weak coresets for ℓ_p regression with an embedding implies upper bounds for spanning coresets of the same size. Thus, our spanning coreset lower bound in fact implies an $\Omega(d/\varepsilon)$ lower bound on the size of weak coresets for ℓ_p regression with an embedding, which establishes that our weak coreset upper bound for ℓ_p regression (Theorem 13.1.2) is also nearly optimal for $1 < p < 2$ up to polylogarithmic factors.

On the other hand, for $p > 2$, our weak coreset lower bound of Theorem 13.6.2 shows that our technique of reducing spanning coresets to weak coresets cannot prove a better upper bound than the result of Theorem 13.1.5, and thus new ideas are required to improve upon the $\tilde{O}(\varepsilon^{-1}k^2)$ spanning coreset upper bound of Theorem 3.1 of [SV12]. This is an interesting open problem.

13.2 Strong coresets

We give the formal statement and proof of our strong coreset result for multiple ℓ_p regression.

Theorem 13.2.1 (Strong coresets for multiple ℓ_p regression)**.** Let $\hat{\mathbf{X}} \in \mathbb{R}^{d\times m}$ satisfy

$$\|\mathbf{A}\hat{\mathbf{X}} - \mathbf{B}\|_{p,p}^p \leq O(1) \min_{\mathbf{X}\in\mathbb{R}^{d\times m}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{p,p}^p$$

and let $\hat{\mathbf{B}} \coloneqq \mathbf{A}\hat{\mathbf{X}} - \mathbf{B}$. Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 6.1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha + \mathbf{v}_i/\beta\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$, $\mathbf{v}_i = \|\mathbf{e}_i^\top \hat{\mathbf{B}}\|_p^p / \|\hat{\mathbf{B}}\|_{p,p}^p$,

$$\alpha = \begin{cases} O(\gamma)\varepsilon^2 \left[(\log d)^2 \log n + \log\frac{1}{\delta}\right]^{-1} & p < 2 \\ \dfrac{O(\gamma^{p/2})\varepsilon^p}{\|\mathbf{w}\|_1^{p/2-1}} \left[(\log d)^2 \log n + \log\frac{1}{\delta}\right]^{-1} & p > 2 \end{cases}$$

and $\beta = O(\varepsilon^{-2}\log\frac{1}{\delta})$. Then with probability at least $1-\delta$,

$$\|\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\|_{p,p}^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_{p,p}^p$$

simultaneously for every $\mathbf{X} \in \mathbb{R}^{d\times m}$.

Proof. By replacing $\mathbf{B}$ by $\hat{\mathbf{B}} - \mathbf{A}\hat{\mathbf{X}}$, we assume that $\|\mathbf{B}\|_p = O(\mathsf{OPT})$. We apply Theorem 12.3.2 with failure probability at $\varepsilon^p\delta^2$. Now let $S \subseteq [m]$ be the set of columns for which the guarantee of Theorem 12.3.2 fails. Note then that by Markov's inequality,

$$\sum_{j\in S} \|\mathbf{B}\mathbf{e}_j\|_p^p = O(\varepsilon^p\delta)\|\mathbf{B}\|_{p,p}^p$$

with probability at least $1-\delta$. We also have that

$$\sum_{j\in S} \|\mathbf{S}\mathbf{B}\mathbf{e}_j\|_p^p \leq \frac{1}{\delta}\sum_{j\in S} \|\mathbf{B}\mathbf{e}_j\|_p^p = O(\varepsilon^p)\|\mathbf{B}\|_{p,p}^p$$

with probability at least $1-\delta$, again by Markov's inequality. Then,

$$\begin{aligned} \|\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\mathbf{e}_j\|_p^p &= (1\pm\varepsilon)\|\mathbf{S}\mathbf{A}\mathbf{X}\mathbf{e}_j\|_p^p \pm \frac{O(1)}{\varepsilon^{p-1}}\|\mathbf{S}\mathbf{B}\mathbf{e}_j\|_p^p \\ &= (1\pm\varepsilon)^2\|\mathbf{A}\mathbf{X}\mathbf{e}_j\|_p^p \pm \frac{O(1)}{\varepsilon^{p-1}}\|\mathbf{S}\mathbf{B}\mathbf{e}_j\|_p^p \end{aligned}$$

by using that $\mathbf{S}$ is a subspace embedding. Similarly, we have that

$$\|(\mathbf{A}\mathbf{X} - \mathbf{B})\mathbf{e}_j\|_p^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{X}\mathbf{e}_j\|_p^p \pm \frac{O(1)}{\varepsilon^{p-1}}\|\mathbf{B}\mathbf{e}_j\|_p^p.$$

Then summing over $j \in S$ gives that

$$\sum_{j\in S} \|\mathbf{S}(\mathbf{A}\mathbf{X} - \mathbf{B})\mathbf{e}_j\|_p^p = \sum_{j\in S} \|(\mathbf{A}\mathbf{X} - \mathbf{B})\mathbf{e}_j\|_p^p \pm O(\varepsilon)\|\mathbf{B}\|_{p,p}^p.$$

On the other hand, for $j \notin S$, Theorem 12.3.2 succeeds so we have

$$\|\mathbf{S}(\mathbf{AX}-\mathbf{B})\mathbf{e}_j\|_p^p = \|(\mathbf{AX}-\mathbf{B})\mathbf{e}_j\|_p^p - \|\mathbf{Be}_j\|_p^p + \|\mathbf{SBe}_j\|_p^p \pm \varepsilon\big(\|\mathbf{Be}_j\|_p^p + \|\mathbf{SBe}_j\|_p^p + \|\mathbf{AXe}_j\|_p^p\big)$$

Summing the guarantee of over the m columns j gives

$$\begin{aligned}\|\mathbf{S}(\mathbf{AX}-\mathbf{B})\|_{p,p}^p &= \|\mathbf{AX}-\mathbf{B}\|_{p,p}^p - \|\mathbf{B}\|_{p,p}^p + \|\mathbf{SB}\|_{p,p}^p \pm O(\varepsilon)\big(\|\mathbf{B}\|_{p,p}^p + \|\mathbf{AX}\|_{p,p}^p\big)\\ &= \|\mathbf{AX}-\mathbf{B}\|_{p,p}^p \pm \varepsilon\|\mathbf{B}\|_{p,p}^p \pm O(\varepsilon)\big(\|\mathbf{B}\|_{p,p}^p + \|\mathbf{AX}\|_{p,p}^p\big)\\ &= \|\mathbf{AX}-\mathbf{B}\|_{p,p}^p \pm O(\varepsilon)\|\mathbf{AX}-\mathbf{B}\|_{p,p}^p. \qquad \square\end{aligned}$$

13.3 Weak coresets

We give the formal statement and proof of our weak coreset result for multiple ℓ_p regression.

Theorem 13.3.1 (Weak coresets for multiple ℓ_p regression)**.** Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 6.1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and

$$\alpha = \begin{cases} O(\gamma)\varepsilon\delta^2\left[(\log d)^2\log n + \log\frac{m}{\delta}\right]^{-1}\left[\log\log\frac{1}{\varepsilon}\right]^{-2} & p < 2\\ \frac{O(\gamma^{p/2})\varepsilon^{p-1}\delta^p}{\|\mathbf{w}\|_1^{p/2-1}}\left[(\log d)^2\log n + \log\frac{m}{\delta}\right]^{-1}\left[\log\log\frac{1}{\varepsilon}\right]^{-p} & p > 2\end{cases}.$$

Then, for any $\hat{\mathbf{X}} \in \mathbb{R}^{d\times t}$ such that

$$\|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B})\|_{p,p}^p \leq (1+\varepsilon)\min_{\mathbf{X}\in\mathbb{R}^{d\times t}}\|\mathbf{S}(\mathbf{AXG}-\mathbf{B})\|_{p,p}^p,$$

we have

$$\|\mathbf{A}\hat{\mathbf{X}}\mathbf{G}-\mathbf{B}\|_{p,p}^p \leq (1+O(\varepsilon))\min_{\mathbf{X}\in\mathbb{R}^{d\times t}}\|\mathbf{AXG}-\mathbf{B}\|_{p,p}^p.$$

The argument closely follows the active ℓ_p regression argument from Chapter 12.

13.3.1 Closeness of nearly optimal solutions

We need Lemma 12.3.3 from Chapter 12 as well as the following elementary computation.

Lemma 13.3.2 (Gradients of multiple ℓ_p regression)**.** The gradient $\nabla_{\mathbf{X}}\|\mathbf{AXG}-\mathbf{B}\|_{p,p}^p$ is given by the formula

$$\sum_{i=1}^n\sum_{j=1}^m p[\mathbf{AXG}-\mathbf{B}](i,j)^{\circ(p-1)}(\mathbf{A}^\top\mathbf{e}_i)(\mathbf{e}_j^\top\mathbf{G}^\top)$$

The following lemma uses Lemmas 12.3.3 and 13.3.2 to show that if $\mathbf{X}$ achieves a nearly optimal value, then $\mathbf{X}$ must be close to the optimal solution $\mathbf{X}^*$.

Lemma 13.3.3 (Closeness of nearly optimal solutions)**.** Let $p > 1$. For any $\mathbf{X} \in \mathbb{R}^{d\times t}$ such that $\|\mathbf{AXG} - \mathbf{B}\|_{p,p} \leq (1+\eta)\,\mathsf{OPT}$ with $\eta \in (0,1)$, we have that

$$\|\mathbf{AXG} - \mathbf{AX}^*\mathbf{G}\|_{p,p} \leq \begin{cases} O(\eta^{1/2})\,\mathsf{OPT} & p < 2 \\ O(\eta^{1/p})\,\mathsf{OPT} & p > 2 \end{cases}$$

where $\mathbf{X}^* := \arg\min_{\mathbf{X}\in\mathbb{R}^{d\times t}} \|\mathbf{AXG} - \mathbf{B}\|_{p,p}$.

Proof. First note that

$$\begin{aligned}
&\left\langle (\mathbf{AX}^*\mathbf{G} - \mathbf{B})^{\circ(p-1)}, \mathbf{AX}^*\mathbf{G} - \mathbf{AXG} \right\rangle \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} [\mathbf{AX}^*\mathbf{G} - \mathbf{B}](i,j)^{\circ(p-1)} [\mathbf{A}(\mathbf{X}^* - \mathbf{X})\mathbf{G}](i,j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} [\mathbf{AX}^*\mathbf{G} - \mathbf{B}](i,j)^{\circ(p-1)} \left\langle (\mathbf{A}^\top\mathbf{e}_i)(\mathbf{e}_j^\top\mathbf{G}^\top), \mathbf{X}^* - \mathbf{X} \right\rangle \\
&= \left\langle \sum_{i=1}^{n}\sum_{j=1}^{m} [\mathbf{AX}^*\mathbf{G} - \mathbf{B}](i,j)^{\circ(p-1)} (\mathbf{A}^\top\mathbf{e}_i)(\mathbf{e}_j^\top\mathbf{G}^\top), \mathbf{X}^* - \mathbf{X} \right\rangle.
\end{aligned}$$

The left term in the product is the gradient of the objective at the optimum by Lemma 13.3.2, so this is just 0 for any $\mathbf{X}$. Then for $p < 2$, we have by Lemma 12.3.3 that

$$\|\mathbf{AX}^*\mathbf{G} - \mathbf{B}\|_{p,p}^2 + \frac{p-1}{2}\|\mathbf{AXG} - \mathbf{AX}^*\mathbf{G}\|_{p,p}^2 \leq \|\mathbf{AXG} - \mathbf{B}\|_{p,p}^2 \leq (1+\eta)^2\|\mathbf{AX}^*\mathbf{G} - \mathbf{B}\|_{p,p}^2$$

which rearranges to

$$\|\mathbf{AXG} - \mathbf{AX}^*\mathbf{G}\|_{p,p} \leq O(\eta^{1/2})\,\mathsf{OPT}\,.$$

and for $p > 2$, we have by Lemma 12.3.3 that

$$\|\mathbf{AX}^*\mathbf{G} - \mathbf{B}\|_{p,p}^p + \frac{p-1}{p2^p}\|\mathbf{AXG} - \mathbf{AX}^*\mathbf{G}\|_{p,p}^p \leq \|\mathbf{AXG} - \mathbf{B}\|_{p,p}^p \leq (1+\eta)^p\|\mathbf{AX}^*\mathbf{G} - \mathbf{B}\|_{p,p}^p$$

which rearranges to

$$\|\mathbf{AXG} - \mathbf{AX}^*\mathbf{G}\|_{p,p} \leq O(\eta^{1/p})\,\mathsf{OPT}\,.$$

□

13.3.2 Iterative size reduction argument

We will need the following initial result to seed our iterative argument. Note that the dependence on ε is suboptimal by an ε factor for every $1 < p < \infty$.

Lemma 13.3.4. Let $\mathbf{S}$ be the ℓ_p sampling matrix (Definition 6.1.1) with sampling probabilities $q_i \geq \min\{1, \mathbf{w}_i/\alpha\}$ for γ-one-sided ℓ_p Lewis weights $\mathbf{w} \in \mathbb{R}^n$ and

$$\alpha = O(\gamma)(\varepsilon\delta)^2\left[(\log d)^2 \log n + \log\frac{1}{\delta}\right]^{-1}$$

for $1 \le p < 2$ and

$$\alpha = \frac{O(\gamma^{p/2})(\varepsilon\delta)^p}{\|\mathbf{w}\|_1^{p/2-1}}\left[(\log d)^2 \log n + \log\frac{1}{\delta}\right]^{-1}$$

for $2 < p < \infty$. Then, for any $\hat{\mathbf{X}} \in \mathbb{R}^{d\times t}$ such that

$$\|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G} - \mathbf{B})\|_{p,p}^p \le (1+\varepsilon)\min_{\mathbf{X}\in\mathbb{R}^{d\times t}}\|\mathbf{S}(\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{B})\|_{p,p}^p,$$

we have

$$\|\mathbf{A}\hat{\mathbf{X}}\mathbf{G} - \mathbf{B}\|_{p,p}^p \le (1+O(\varepsilon))\min_{\mathbf{X}\in\mathbb{R}^{d\times t}}\|\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{B}\|_{p,p}^p.$$

Proof. We first show that

$$\|\mathbf{A}\hat{\mathbf{X}}\mathbf{G} - \mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p \le O\left(\frac{1}{\delta}\right)\mathsf{OPT}^p$$

with probability at least $1-\delta$. By using the fact that $\mathbf{S}$ is an $O(1)$-approximate ℓ_p subspace embedding, we have that

$$\begin{aligned}
\|\mathbf{A}\hat{\mathbf{X}}\mathbf{G} - \mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p &\le \|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G} - \mathbf{A}\mathbf{X}^*\mathbf{G})\|_{p,p}^p \\
&\le 2^{p-1}\left(\|\mathbf{S}(\mathbf{A}\hat{\mathbf{X}}\mathbf{G} - \mathbf{B})\|_{p,p}^p + \|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B})\|_{p,p}^p\right) && \text{Fact 2.1.1} \\
&\le 2^{p+1}\|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B})\|_{p,p}^p && \text{Approximate optimality of } \hat{\mathbf{X}}
\end{aligned}$$

The latter quantity is at most $O(\frac{1}{\delta})\,\mathsf{OPT}^p$ with probability at least $1-\delta$ by Markov's inequality. Thus, we may replace the optimization of $\hat{\mathbf{X}}$ over all $\mathbf{X} \in \mathbb{R}^{d\times t}$ with optimization over the ball $\{\mathbf{X} : \|\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{A}\mathbf{X}^*\mathbf{G}\|_{p,p}^p = O(\frac{1}{\delta})\,\mathsf{OPT}^p\}$.

The rest of the proof now mimics the proof of Theorem 13.2.1. We apply Theorem 12.3.2 with accuracy parameter ε set to $\varepsilon\delta$, failure parameter set to $(\varepsilon\delta)^p\delta^2$, and proximity parameter η set to 1. Let $S \subseteq [m]$ be the set of columns for which Theorem 12.3.2 fails. Then by applying Markov's inequality twice as in the proof of Theorem 13.2.1, we have that

$$\sum_{j\in S}\|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B})\mathbf{e}_j\|_p^p = O((\varepsilon\delta)^p)\,\mathsf{OPT}^p$$

and

$$\sum_{j\in S}\|(\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B})\mathbf{e}_j\|_p^p = O((\varepsilon\delta)^p)\,\mathsf{OPT}^p$$

and thus it follows that

$$\sum_{j\in S}\|\mathbf{S}(\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{B})\mathbf{e}_j\|_p^p = \sum_{j\in S}\|(\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{B})\mathbf{e}_j\|_p^p \pm O(\varepsilon\delta)\left(\|\mathbf{A}(\mathbf{X} - \mathbf{X}^*)\mathbf{G}\|_p^p + \mathsf{OPT}^p\right).$$

Summing this result with the rest of the columns $j \notin S$ gives that

$$\left|\left(\|\mathbf{S}(\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{B})\|_{p,p}^p - \|\mathbf{S}(\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B})\|_{p,p}^p\right) - \left(\|\mathbf{A}\mathbf{X}\mathbf{G} - \mathbf{B}\|_{p,p}^p - \|\mathbf{A}\mathbf{X}^*\mathbf{G} - \mathbf{B}\|_{p,p}^p\right)\right|$$

$$\leq \varepsilon\delta\big(\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p + \|\mathbf{S}(\mathbf{AX}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p + \|\mathbf{AXG}-\mathbf{AX}^*\mathbf{G}\|_{p,p}^p\big) \leq O(\varepsilon)\,\mathsf{OPT}^p$$

Thus, in the ball $\{\mathbf{X} : \|\mathbf{AXG}-\mathbf{AX}^*\mathbf{G}\|_{p,p}^p = O(\frac{1}{\delta})\,\mathsf{OPT}^p\}$, we have that

$$\|\mathbf{S}(\mathbf{AXG}-\mathbf{B})\|_{p,p}^p = \|\mathbf{AXG}-\mathbf{B}\|_{p,p}^p + (\|\mathbf{S}(\mathbf{AX}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p - \|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p) \pm O(\varepsilon)\,\mathsf{OPT}^p\,.$$

It follows that $\hat{\mathbf{X}}$ must minimize $\|\mathbf{AXG}-\mathbf{B}\|_{p,p}^p$ up to an additive $O(\varepsilon)\,\mathsf{OPT}^p$. □

Starting from this initial solution bound of Lemma 13.3.4, we proceed via an iteration argument as in Chapter 12.

Proof of Theorem 13.3.1. Let

$$C = \begin{cases} O(\gamma^{-1})\delta^{-2}\|\mathbf{w}\|_1\left[(\log d)^2\log n + \log\frac{1}{\delta}\right] & p<2 \\ O(\gamma^{-p/2})\delta^{-p}\|\mathbf{w}\|_1^{p/2}\left[(\log d)^2\log n + \log\frac{1}{\delta}\right] & p>2 \end{cases}$$

We will make use of the fact that $\|\mathbf{S}(\mathbf{AX}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p = O(\frac{1}{\delta})\|\mathbf{S}(\mathbf{AX}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p$ with probability at least $1-\delta$ by Markov's inequality.

We will first give the argument for $p<2$. Suppose that C/ε^β rows are needed for a $(1+\varepsilon)$-approximate weak coreset. Now choose a such that $a-2=-a\beta$, that is, $a=2/(1+\beta)$. Then for $\eta^{2/p}=\varepsilon^a$, $C\eta^{2/p}/(\varepsilon\delta)^2 = C/\eta^{(2/p)\beta}$ rows yields a $(1+\eta^{2/p})$-approximate weak coreset. Then, a $(1+\eta^{2/p})$-approximate minimizer $\mathbf{X}$ satisfies

$$\|\mathbf{AXG}-\mathbf{AX}^*\mathbf{G}\|_{p,p}^p \leq O(\eta)\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$$

by Lemma 13.3.3. For all such $\mathbf{X}$, an argument as done in Theorem 13.2.1 and Lemma 13.3.4 shows that $\|\mathbf{S}(\mathbf{AXG}-\mathbf{B})\|_{p,p}^p - \|\mathbf{S}(\mathbf{AX}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p$ and $\|\mathbf{AXG}-\mathbf{B}\|_{p,p}^p - \|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$ are close up to an additive error of

$$\varepsilon\delta\left(\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p + \|\mathbf{S}(\mathbf{AX}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p + \frac{1}{\eta}\|\mathbf{AXG}-\mathbf{AX}^*\mathbf{G}\|_{p,p}^p\right) = O(\varepsilon)\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$$

Thus, $C/\eta^{(2/p)\beta}$ rows in fact gives a $(1+O(\varepsilon))$-approximate minimizer. That is, if C/ε^β rows is sufficient for $(1+\varepsilon)$-approximation, then $C/\eta^{(2/p)\beta} = C/\varepsilon^{a\beta} = C/\varepsilon^{2\beta/(1+\beta)}$ rows is sufficient for $(1+\varepsilon)$-approximation as well. We may now iterate this argument. Consider the sequence β_i given by

$$\beta_0 = 2, \qquad \beta_{i+1} = \frac{2\beta_i}{1+\beta_i}.$$

The solution to this recurrence is given by the Lemma 12.3.7.

Thus, applying this argument $O(\log\log\frac{1}{\varepsilon})$ times yields that $\beta_i \leq 1 + O(1/\log(\frac{1}{\varepsilon}))$ which means that reading only $O(1)C/\varepsilon$ entries suffices. Union bounding over the success of the $O(\log\log\frac{1}{\varepsilon})$ rounds completes the argument.

Next, let $p>2$. Suppose that C/ε^β rows are needed for a $(1+\varepsilon)$-approximate weak coreset. Now choose a such that $a-p=-a\beta$, that is, $a=p/(1+\beta)$. Then for $\eta=\varepsilon^a$, $C\eta/\varepsilon^p = C/\eta^\beta$

rows yields a $(1+\eta)$-approximate weak coreset. Then, a $(1+\eta)$-approximate minimizer $\mathbf{X}$ satisfies

$$\|\mathbf{AXG}-\mathbf{AX}^*\mathbf{G}\|_{p,p}^p \le O(\eta)\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$$

by Lemma 13.3.3. For all such $\mathbf{X}$, an argument as done in Theorem 13.2.1 and Lemma 13.3.4 shows that $\|\mathbf{S}(\mathbf{AXG}-\mathbf{B})\|_{p,p}^p - \|\mathbf{S}(\mathbf{AX}^*\mathbf{G}-\mathbf{B})\|_{p,p}^p$ and $\|\mathbf{AXG}-\mathbf{B}\|_{p,p}^p - \|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$ are close up to an additive error of

$$\varepsilon\left(\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p + \frac{1}{\eta}\|\mathbf{AXG}-\mathbf{AX}^*\mathbf{G}\|_{p,p}^p\right) = O(\varepsilon)\|\mathbf{AX}^*\mathbf{G}-\mathbf{B}\|_{p,p}^p$$

Thus, C/η^β rows in fact gives a $(1+O(\varepsilon))$-approximate minimizer. That is, if C/ε^β rows is sufficient for $(1+\varepsilon)$-approximation, then $C/\eta^\beta = C/\varepsilon^{a\beta} = C/\varepsilon^{p\beta/(1+\beta)}$ rows is sufficient for $(1+\varepsilon)$-approximation as well. We may now iterate this argument. Consider the sequence β_i given by

$$\beta_1 = p, \qquad \beta_{i+1} = \frac{p\beta_i}{1+\beta_i}.$$

Then by Lemma 12.3.7, applying this argument $O(\log\log\frac{1}{\varepsilon})$ times yields that $\beta_i \le (p-1) + O(1/\log(\frac{1}{\varepsilon}))$ which means that reading only $O(1)C/\varepsilon^{p-1}$ entries suffices. Union bounding over the success of the $O(\log\log\frac{1}{\varepsilon})$ rounds completes the argument. □

13.4 Sublinear algorithm for Euclidean power means

Theorem 13.1.3. Let $\{\mathbf{b}_i\}_{i=1}^n \subseteq \mathbb{R}^d$. Then, there is a sublinear algorithm which uniformly samples at most

$$s = \begin{cases} O(\varepsilon^{-2})\left(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\right)\log\frac{1}{\delta} & p=1 \\ O(\varepsilon^{-1})\left(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\right)\log\frac{1}{\delta} & 1<p\le 2 \\ O(\varepsilon^{1-p})\left(\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\right)\log\frac{1}{\delta} & 2<p<\infty \end{cases}$$

rows $\mathbf{b}_i$ and outputs a center $\hat{\mathbf{x}}$ such that

$$\sum_{i=1}^n \|\hat{\mathbf{x}}-\mathbf{b}_i\|_2^p \le (1+\varepsilon)\min_{\mathbf{x}\in\mathbb{R}^d}\sum_{i=1}^n \|\mathbf{x}-\mathbf{b}_i\|_2^p$$

with probability at least $1-\delta$.

Proof. We will assume without loss of generality that by reading $O(\log\frac{1}{\delta})$ rows of $\mathbf{B}$, we can identify an $O(1)$-approximate solution $\hat{\mathbf{x}}$ (see, e.g., Section 3.1 of [MMWY22]). Thus by subtracting off this solution, we may assume that $\|\mathbf{B}\|_{p,2}^p = O(\mathsf{OPT}^p)$.

We then use Dvoretzky's thoerem to embed this problem into the entrywise ℓ_p norm, so that

$$\|\mathbf{1x}^\top-\mathbf{B}\|_{p,2}^p = (1\pm\varepsilon)\|\mathbf{1x}^\top\mathbf{G}-\mathbf{BG}\|_{p,p}^p$$

for every center $\mathbf{x}\in\mathbb{R}^d$. This is now in a form where we may apply our weak coreset results for multiple ℓ_p regression of Theorem 13.1.2. Note that in this particular setting, the $\mathbf{A}$ matrix

corresponds to the $n \times d$ all ones matrix with $d = 1$, and the ℓ_p Lewis weights can be taken to be uniform.

Now consider running $L = O(\log \frac{1}{\delta})$ independent instances of the weak coreset algorithm, each which has the property that the algorithm makes at most

$$O(\varepsilon^{-\rho})\left(\log \frac{1}{\varepsilon} + \log \frac{1}{\delta}\right) \tag{13.2}$$

queries for $\rho = 2$ for $p = 1$, $\rho = 1$ for $1 < p < 2$, and $\rho = p - 1$ for $2 < p < \infty$, and that if $\|\mathbf{S}(\mathbf{1}(\mathbf{x}^*)^\top \mathbf{G} - \mathbf{BG})\|_{p,p}^p = O(\|\mathbf{1}(\mathbf{x}^*)^\top \mathbf{G} - \mathbf{BG}\|_{p,p}^p)$ for the optimal solution $\mathbf{x}^*$, then it succeeds with probability at least $1 - \delta/L$. By a union bound, this holds for all L instances.

By Markov's inequality, each instance satisfies $\|\mathbf{SBG}\|_{p,p}^p = O(\|\mathbf{BG}\|_{p,p}^p)$ with probability at least $9/10$, so at least $2/3$ of the L instances must satisfy this bound with probability at least $1 - \delta$. By Dvoretzky's theorem, this means that $\|\mathbf{SB}\|_{p,2}^p = O(\|\mathbf{B}\|_{p,2}^p)$. Then, if we restrict our attention to the $(2/3)L$ instances with the smallest values of $\|\mathbf{SB}\|_{p,2}^p$, then all of these instances must output a correct $(1+\varepsilon)$-approximately optimal solution, simultaneously with probability $1 - \delta$. This gives a query bound of L times (13.2). □

13.5 Spanning coresets for ℓ_p subspace approximation

We show that weak coreset construction imply spanning sets for ℓ_p subspace approximation.

Theorem 13.1.5. Let $\{\mathbf{a}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, $1 \le p < \infty$, $k \in \mathbb{N}$, and $0 < \varepsilon < 1$. Then, there exists a $(1+\varepsilon)$-spanning coreset S of size at most

$$|S| = \begin{cases} O(\varepsilon^{-2}k)(\log(k/\varepsilon))^3 & p = 1 \\ O(\varepsilon^{-1}k)(\log(k/\varepsilon))^3 & 1 < p \le 2 \\ O(\varepsilon^{1-p}k^{p/2})(\log(k/\varepsilon))^3 & 2 < p < \infty \end{cases}$$

Proof. By first computing a strong coreset of size $\mathrm{poly}(k/\varepsilon)$ [HV20], we can assume that $n, d = \mathrm{poly}(k/\varepsilon)$.

Let $\mathbf{P} = \mathbf{VV}^\top$ be the rank k projection that minimizes $\|\mathbf{AP} - \mathbf{A}\|_{p,2}^p$. Note then that

$$\min_{\mathbf{X} \in \mathbb{R}^{k \times d}} \|\mathbf{AVX} - \mathbf{A}\|_{p,2}^p = \|\mathbf{AP} - \mathbf{A}\|_{p,2}^p.$$

We then use Dvoretzky's theorem to embed this problem into the entrywise ℓ_p norm, so that

$$\|\mathbf{AVX} - \mathbf{A}\|_{p,2}^p = (1 \pm \varepsilon)\|\mathbf{AVXG} - \mathbf{AG}\|_{p,p}^p$$

for every $\mathbf{X} \in \mathbb{R}^{k \times d}$, for some fixed $\mathbf{G} \in \mathbb{R}^{d \times m}$ with $m = \mathrm{poly}(d/\varepsilon)$. Then by our weak coreset result for multiple ℓ_p regression (Theorem 13.3.1), there is a diagonal matrix $\mathbf{S}$ with

$$\mathsf{nnz}(\mathbf{S}) \le \begin{cases} O(\varepsilon^{-2}k)(\log(k/\varepsilon))^3 & p = 1 \\ O(\varepsilon^{-1}k)(\log(k/\varepsilon))^3 & 1 < p \le 2 \\ O(\varepsilon^{1-p}k^{p/2})(\log(k/\varepsilon))^3 & 2 < p < \infty \end{cases}$$

such that any $(1+\varepsilon)$-approximate minimizer $\hat{\mathbf{X}}$ of $\|\mathbf{S}(\mathbf{AVXG}-\mathbf{AG})\|_{p,p}^p$ satisfies

$$\|\mathbf{AV}\hat{\mathbf{X}}\mathbf{G}-\mathbf{AG}\|_{p,p}^p \leq (1+\varepsilon)\min_{\mathbf{X}\in\mathbb{R}^{k\times d}}\|\mathbf{AVXG}-\mathbf{AG}\|_{p,p}^p.$$

We will take $\hat{\mathbf{X}}$ to be

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}\in\mathbb{R}^{k\times d}}\|\mathbf{S}(\mathbf{AVX}-\mathbf{A})\|_{p,2}^p$$

which is indeed a $(1+\varepsilon)$-approximate minimizer of $\|\mathbf{S}(\mathbf{AVXG}-\mathbf{AG})\|_{p,p}^p$ by Dvoretzky's theorem. Then, again by Dvoretzky's theorem, we then have for this $\hat{\mathbf{X}}$ that

$$\begin{aligned}\|\mathbf{AV}\hat{\mathbf{X}}-\mathbf{A}\|_{p,2}^p &\leq (1+O(\varepsilon))\min_{\mathbf{X}\in\mathbb{R}^{k\times d}}\|\mathbf{AVX}-\mathbf{A}\|_{p,2}^p\\ &= (1+O(\varepsilon))\|\mathbf{AP}-\mathbf{A}\|_{p,2}^p.\end{aligned}$$

Finally, note that $\hat{\mathbf{X}}$ has row span contained in the row span of $\mathbf{SA}$, since otherwise $\|\mathbf{S}(\mathbf{AVX}-\mathbf{A})\|_{p,2}^p$ can be reduced by projecting the rows of $\mathbf{X}$ onto $\mathrm{rowspan}(\mathbf{SA})$. Then, if $\mathbf{P}_F$ is the projection matrix onto $F=\mathrm{rowspan}(\hat{\mathbf{X}})$, then for each row $i\in[n]$ of $\mathbf{A}$,

$$\|\mathbf{P}_F\mathbf{a}_i-\mathbf{a}_i\|_2 = \min_{\mathbf{x}\in F}\|\mathbf{x}-\mathbf{a}_i\|_2 \leq \|\hat{\mathbf{X}}^\top\mathbf{V}^\top\mathbf{a}_i-\mathbf{a}_i\|_2$$

so

$$\|\mathbf{AP}_F-\mathbf{A}\|_{p,2}^p \leq \|\mathbf{AV}\hat{\mathbf{X}}-\mathbf{A}\|_{p,2}^p.$$

We thus conclude that there is a rank k subspace in the row span of $\mathbf{SA}$ that is $(1+\varepsilon)$-approximately optimal. □

13.6 Lower bounds

In this section, we complement our various upper bounds with matching lower bounds. Section 13.6.1 gives a nearly optimal lower bound for strong coresets, Section 13.6.2 for weak coresets, and Section 13.6.3 for spanning coresets.

13.6.1 Strong coresets

Theorem 13.6.1. Let $2<p<\infty$ be fixed. Let $\varepsilon\in(0,1)$ be less than some sufficiently small constant. Then, a strong coreset $\mathbf{S}$ for multiple ℓ_p regression requires $\mathsf{nnz}(\mathbf{S}) = \Omega(\varepsilon^{-p}d^{p/2})$ nonzero rows.

Proof of Theorem 13.6.1. Let $s=d^{p/2}$ and let $S\subseteq\{\pm1\}^d$ be a set of $|S|=s$ points given by Theorem 11.3.2 such that $\langle\mathbf{a},\mathbf{a}'\rangle \leq C_{p/2}\sqrt{d} = O(\sqrt{d})$ for some $C_{p/2}^p\geq 1$, for every distinct $\mathbf{a},\mathbf{a}'\in S$. Let $m=s\varepsilon^{-p}$, let $\mathbf{A}\in\{\pm1\}^{m\times d}$ be the matrix with ε^{-p} copies of $\mathbf{a}$ in its rows for each $\mathbf{a}\in S$, and let $\mathbf{B}=d\cdot\mathbf{I}_m$ be the $m\times m$ identity matrix scaled by d. For each row $i\in[m]$, we say that $i'\in[s]$ is its *group number* if $\mathbf{e}_i^\top\mathbf{A}$ is the i'-th point in S.

Suppose for contradiction that $\mathbf{S}$ is a strong coreset with $\mathsf{nnz}(\mathbf{S}) \leq m/16$ such that

$$\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\|_{p,p}^p = \left(1 \pm \frac{\varepsilon}{12C_{p/2}^p}\right)\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_{p,p}^p$$

for every $\mathbf{X} \in \mathbb{R}^{d\times m}$. Then, there is a subset $T \subseteq [m]$ with $|T| = m/16$ such that $\mathbf{S}$ is supported on T. For each $i' \in [s]$, let $T_{i'} \subseteq T$ denote the rows of T whose rows in $\mathbf{A}$ with group number $i' \in [s]$, so $\sum_{i'=1}^s |T_{i'}| = |T|$. Then by averaging, there are at least $(3/4)s$ groups $i' \in [s]$ such that $|T_{i'}| \leq \varepsilon^{-p}/2$. Thus, we may assume without loss of generality that $|T_{i'}| = \varepsilon^{-p}$ for the first $(1/4)s$ groups, $|T_{i'}| = \varepsilon^{-p}/2$ for the last $(3/4)s$ groups, and $|T| = (5/8)m$.

Let $W \coloneqq \sum_{i=1}^m |\mathbf{S}_{i,i}|^p$ denote the total weight mass of $\mathbf{S}$. Note then that by querying $\mathbf{X} = 0$, we must have that

$$\|\mathbf{S}\mathbf{B}\|_{p,p}^p = W = (1\pm\varepsilon)\|\mathbf{B}\|_{p,p}^p = \left(1 \pm \frac{\varepsilon}{12C_{p/2}^p}\right)m.$$

Let W_1 denote the sum of $|\mathbf{S}_{i,i}|^p$ on the first $(1/4)s$ groups, and let W_2 denote the sum of $|\mathbf{S}_{i,i}|^p$ on the last $(3/4)s$ groups. We will assume that $W_1 \leq m/4$, since the case of $W_1 \geq m/4$ is symmetric.

We now construct a query $\mathbf{X} \in \mathbb{R}^{d\times m}$ with the j-th column given by

$$\mathbf{X}\mathbf{e}_j = \begin{cases} \varepsilon \cdot \mathbf{e}_j^\top \mathbf{A} & j \in T \\ 0 & j \notin T \end{cases}$$

Note then that for each $i, j \in [m]$,

$$\mathbf{e}_i^\top \mathbf{A}\mathbf{X}\mathbf{e}_j = \begin{cases} \varepsilon d & \mathbf{e}_i^\top \mathbf{A} = \mathbf{e}_j^\top \mathbf{A}, j \in T \\ \varepsilon C_{p/2}\sqrt{d} & \mathbf{e}_i^\top \mathbf{A} \neq \mathbf{e}_j^\top \mathbf{A}, j \in T \\ 0 & j \notin T \end{cases}$$

Let $i \in [m]$ and let $i' \in [s]$ be its group number. Then the cost of row i if $i \in T$ is

$$\begin{aligned}\|\mathbf{e}_i^\top \mathbf{A}\mathbf{X} - \mathbf{e}_i^\top \mathbf{B}\|_p^p &= \sum_{j=1}^m \left|\mathbf{e}_i^\top \mathbf{A}\mathbf{X}\mathbf{e}_j - \mathbf{B}(i,j)\right|^p \\ &= \underbrace{(1-\varepsilon)^p d^p}_{i=j} + (|T_{i'}|-1)\cdot \underbrace{\varepsilon^p d^p}_{\mathbf{e}_i^\top\mathbf{A}=\mathbf{e}_j^\top\mathbf{A}} + (|T|-|T_{i'}|)\cdot \underbrace{\varepsilon^p C_{p/2}^p d^{p/2}}_{\mathbf{e}_i^\top\mathbf{A}\neq\mathbf{e}_j^\top\mathbf{A}} \\ &= (1 - p\varepsilon + |T_{i'}|\varepsilon^p + (5/8)C_{p/2}^p + o(\varepsilon))d^p\end{aligned}$$

while the cost of row $i \in [m]$ if $i \notin T$ is

$$\|\mathbf{e}_i^\top \mathbf{A}\mathbf{X} - \mathbf{e}_i^\top \mathbf{B}\|_p^p = \sum_{j=1}^m \left|\mathbf{e}_i^\top \mathbf{A}\mathbf{X}\mathbf{e}_j - \mathbf{B}(i,j)\right|^p$$

$$= \underbrace{d^p}_{i=j} + |T_{i'}| \cdot \underbrace{\varepsilon^p d^p}_{\mathbf{e}_i^\top \mathbf{A} = \mathbf{e}_j^\top \mathbf{A}} + (|T| - |T_{i'}|) \cdot \underbrace{\varepsilon^p C_{p/2}^p d^{p/2}}_{\mathbf{e}_i^\top \mathbf{A} \neq \mathbf{e}_j^\top \mathbf{A}}$$
$$= (1 + |T_i'|\varepsilon^p + (5/8)C_{p/2}^p + o(\varepsilon))d^p.$$

Let

$$c_1 = (1 - p\varepsilon + 1 + (5/8)C_{p/2}^p + o(\varepsilon))d^p$$
$$c_2 = (1 - p\varepsilon + (1/2) + (5/8)C_{p/2}^p + o(\varepsilon))d^p$$
$$c_3 = (1 + (1/2) + (5/8)C_{p/2}^p + o(\varepsilon))d^p$$

Then, the total true cost is at least

$$\begin{aligned}\|\mathbf{AX} - \mathbf{B}\|_{p,p}^p &= \frac{m}{4}c_1 + \frac{3m}{8}c_2 + \frac{3m}{8}c_3 \\ &= \frac{m}{4}c_1 + \frac{3m}{4}c_2 + \frac{3m}{8}(c_3 - c_2) \\ &\geq \frac{m}{4}c_1 + \frac{3m}{4}c_2 + \frac{3m}{4} \cdot (\varepsilon - o(\varepsilon))d^p\end{aligned}$$

while the strong coreset estimate is at most

$$\begin{aligned}\|\mathbf{S}(\mathbf{AX} - \mathbf{B})\|_{p,p}^p &= W_1 c_1 + W_2 c_2 \\ &= W_1(c_1 - c_2) + (W_1 + W_2)c_2 \\ &\leq \frac{m}{4}(c_1 - c_2) + \left(1 + \frac{\varepsilon}{12C_{p/2}^p}\right)mc_2 \\ &\leq \frac{m}{4}c_1 + \frac{3m}{4}c_2 + \frac{\varepsilon}{4}md^p.\end{aligned}$$

Furthermore,

$$\frac{\varepsilon}{12C_{p/2}^p}\left(\frac{m}{4}c_1 + \frac{3m}{4}c_2 + \frac{\varepsilon}{4}md^p\right) \leq \frac{\varepsilon}{4}md^p$$

so $(1 + \frac{\varepsilon}{12C_{p/2}^p})\|\mathbf{S}(\mathbf{AX} - \mathbf{B})\|_{p,p}^p < \|\mathbf{AX} - \mathbf{B}\|_{p,p}^p$ and thus $\mathbf{S}$ fails to be a strong coreset. Rescaling ε by constant factors gives the desired result. □

13.6.2 Weak coresets

Theorem 13.6.2. Let $2 < p < \infty$ be fixed. Let $\varepsilon \in (0, 1)$ be less than some sufficiently small constant. Then, a weak coreset $\mathbf{S}$ for multiple ℓ_p regression requires $\mathsf{nnz}(\mathbf{S}) = \Omega(\varepsilon^{1-p}d^{p/2})$ nonzero rows.

Proof of Theorem 13.6.2. Our hard instance is identical to the one of Theorem 13.6.1, except that each group has $\varepsilon^{1-p}/2C_{p/2}^p$ copies rather than ε^{-p} copies.

Note that if $\mathbf{S}$ does not sample some row $i \in [m]$, then the i-th column of $\mathbf{SB}$ is all zeros, so the solution obtained by the weak coreset is $\mathbf{Xe}_i = 0$, which has objective function value

$\|\mathbf{B}\mathbf{e}_i\|_p^p = d^p$. On the other hand, the optimal value is at most $(1-\varepsilon)^p d^p$ since we can set $\mathbf{X}\mathbf{e}_i = \varepsilon \mathbf{A}^\top \mathbf{e}_i$ so that

$$\begin{aligned}\|(\mathbf{A}\mathbf{X}-\mathbf{B})\mathbf{e}_i\|_p^p &\leq (1-\varepsilon)^p d^p + \frac{\varepsilon^{1-p}}{2C_{p/2}^p}\cdot \varepsilon^p d^p + d^{p/2}\frac{\varepsilon^{1-p}}{2C_{p/2}^p}\cdot C_{p/2}^p \varepsilon^p d^{p/2} \\ &\leq (1-\varepsilon)^p d^p + \frac{\varepsilon}{2}\cdot d^p + \frac{\varepsilon}{2}\cdot d^p \\ &\leq ((1-\varepsilon)^p + \varepsilon) d^p\end{aligned}$$

which is a $(1+\varepsilon)$ factor smaller for all ε sufficiently small. Thus, if $\mathsf{nnz}(\mathbf{S}) \leq m/2$, then the solution $\mathbf{X}$ that minimizes $\|\mathbf{S}(\mathbf{A}\mathbf{X}-\mathbf{B})\|_{p,p}^p$ must be at least an additive $\varepsilon d^p \cdot m/2$ more expensive than the optimal solution, and thus it fails to be a $(1+\varepsilon/2)$-optimal solution. □

13.6.3 Spanning coresets

Theorem 13.6.3. Let $1 \leq p < \infty$ and

$$c_p = \begin{cases} 1/6 & p \leq 2 \\ 1/(6\cdot 5^{p/2-1}) & p > 2 \end{cases}$$

Let $k \in \mathbb{N}$. Then, there is a matrix $\mathbf{B} \in \mathbb{R}^{n\times(n+1)}$ such that for every $\varepsilon \geq k/n$ and any subset of $s \leq (c_p/4)\varepsilon^{-1}k$ rows, any rank k subspace F' spanned by the s rows must have

$$\|\mathbf{B}\mathbf{P}_{F'} - \mathbf{B}\|_{p,2}^p > (1+\varepsilon)\min_{\mathrm{rank}(F)\leq k}\|\mathbf{B}\mathbf{P}_F - \mathbf{B}\|_{p,2}^p.$$

We generalize an argument of Section 4 of [DV06].

Lemma 13.6.4. Let $1 \leq p < \infty$ and

$$c_p = \begin{cases} 1/6 & p \leq 2 \\ 1/(6\cdot 5^{p/2-1}) & p > 2 \end{cases}$$

Then, there is a matrix $\mathbf{A} \in \mathbb{R}^{n\times(n+1)}$ such that for every $\varepsilon \geq 1/n$ and any subset of $s \leq c_p\varepsilon^{-1}$ rows, any rank 1 subspace F' spanned by the s rows must have

$$\|\mathbf{A}\mathbf{P}_{F'} - \mathbf{A}\|_{p,2}^p > (1+\varepsilon)\min_{\mathrm{rank}(F)\leq 1}\|\mathbf{A}\mathbf{P}_F - \mathbf{A}\|_{p,2}^p.$$

Proof. Let $n \leq \varepsilon^{-1}$ and let $\mathbf{A}$ be the $n\times(n+1)$ matrix given by $[R\cdot\mathbf{1}_n, \mathbf{I}_n]$ for some large enough $R > 0$. That is, $\mathbf{A}$ is R along the first column and the $n\times n$ identity for the last n columns. Note that the optimal value is upper bounded by

$$n((1-\varepsilon)^2 + \varepsilon^2\cdot(n-1))^{p/2} = n(1-2\varepsilon+\varepsilon^2 n)^{p/2} = n(1-\varepsilon)^{p/2}.$$

Let $\mathbf{x} \in \mathbb{R}^s$ be the coefficients of a linear combination of s rows of $\mathbf{A}$. We may assume the coefficients are nonnegative, since making the coefficients negative can only increase the cost. Note first that $1/2 \leq \|\mathbf{x}\|_1 \leq 3/2$ since otherwise

$$n\cdot|R - R\|\mathbf{x}\|_1|^p \geq n\cdot R/2$$

which cannot be $(1+\varepsilon)$-approximately optimal for $R \geq 2$.

The cost of the i-th row is $((1-\mathbf{x}_i)^2 + \|\mathbf{x}\|_2^2 - \mathbf{x}_i^2)^{p/2} = (1 - 2\mathbf{x}_i + \|\mathbf{x}\|_2^2)^{p/2}$. If $\|\mathbf{x}\|_2 \geq 2$, then

$$\left(1 - 2\mathbf{x}_i + \|\mathbf{x}\|_2^2\right)^{p/2} \geq (1 - 2\|\mathbf{x}\|_2 + \|\mathbf{x}\|_2^2)^{p/2} = (\|\mathbf{x}\|_2 - 1)^p \geq 1$$

so this cannot produce a $(1+\varepsilon)$-approximately optimal solution. Thus, assume $\|\mathbf{x}\|_2 \leq 2$. Then,

$$\left(1 - 2\mathbf{x}_i + \|\mathbf{x}\|_2^2\right)^{p/2} = \left(1 + \|\mathbf{x}\|_2^2\right)^{p/2}\left(1 - \frac{2}{1+\|\mathbf{x}\|_2^2}\mathbf{x}_i\right)^{p/2} \geq \left(1 + \|\mathbf{x}\|_2^2\right)^{p/2}\left(1 - \frac{p}{1+\|\mathbf{x}\|_2^2}\mathbf{x}_i\right)$$

so summing over the rows gives a cost of

$$\begin{aligned}
&\left(1 + \|\mathbf{x}\|_2^2\right)^{p/2}\left(n - \frac{p}{1+\|\mathbf{x}\|_2^2}\|\mathbf{x}\|_1\right) \\
&= \left(1 + \|\mathbf{x}\|_2^2\right)^{p/2} n - p(1+\|\mathbf{x}\|_2^2)^{p/2-1}\|\mathbf{x}\|_1 \\
&\geq \left(1 + \|\mathbf{x}\|_1^2/s\right)^{p/2} n - p(1+\|\mathbf{x}\|_2^2)^{p/2-1}\|\mathbf{x}\|_1 \qquad \text{since } 1/2 \leq \|\mathbf{x}\|_1 \leq 3/2 \\
&\geq (1+1/2s)^{p/2} n - (3/2)p(1+\|\mathbf{x}\|_2^2)^{p/2-1} \\
&\geq (1+p/4s)n - (3/2)p(1+\|\mathbf{x}\|_2^2)^{p/2-1} \\
&\geq \begin{cases} (1+p/4s)n - (3/2)p & p \leq 2 \\ (1+p/4s)n - (3/2)p \cdot 5^{p/2-1} & p > 2 \end{cases}
\end{aligned}$$

Thus, this fails to be a $(1+\varepsilon)$-approximately optimal solution for

$$(p/4s)n \geq \begin{cases} (3/2)p & p \leq 2 \\ (3/2)p \cdot 5^{p/2-1} & p > 2 \end{cases}$$

that is,

$$s \leq \begin{cases} n/6 & p \leq 2 \\ n/(6 \cdot 5^{p/2-1}) & p > 2 \end{cases}.$$

□

We now extend Lemma 13.6.4 to a general rank k lower bound.

Proof of Theorem 13.6.3. Let $n = \varepsilon^{-1}$ and let $\mathbf{B}$ be a $kn \times k(n+1)$ block diagonal matrix with the $n \times (n+1)$ matrix construction $\mathbf{A} \in \mathbb{R}^{n \times (n+1)}$ of Lemma 13.6.4 on the block diagonal. Consider any set S of s rows of $\mathbf{B}$, and let S_i denote the set of $|S_i| = s_i$ rows supported on the i-th block for each $i \in [k]$. Let F_i denote the optimal subspace spanned by the rows S_i on the ith block.

Let $T \subseteq [k]$ denote the set of $i \in [k]$ such that $s_i \leq c_p n$. If $i \in T$, then we by Lemma 13.6.4 that

$$\|\mathbf{A}\mathbf{P}_{F_i} - \mathbf{A}\|_{p,2}^p > \left(1 + \frac{c_p}{s_i}\right) \min_{\text{rank}(F) \leq k} \|\mathbf{A}\mathbf{P}_F - \mathbf{A}\|_{p,2}^p$$

Then, the additive error from these rows is bounded below by

$$\begin{aligned}\sum_{i\in T}\frac{c_p}{s_i}\min_{\mathrm{rank}(F)\le k}\|\mathbf{A}\mathbf{P}_F-\mathbf{A}\|_{p,2}^p &\ge |T|\cdot\frac{c_p|T|}{\sum_{i\in[k]:s_i\le c_p n}s_i}\min_{\mathrm{rank}(F)\le k}\|\mathbf{A}\mathbf{P}_F-\mathbf{A}\|_{p,2}^p \quad \text{AM-HM}\\ &\ge |T|\cdot\frac{c_p|T|}{s}\min_{\mathrm{rank}(F)\le k}\|\mathbf{A}\mathbf{P}_F-\mathbf{A}\|_{p,2}^p\\ &\ge \frac{c_p|T|^2}{ks}\min_{\mathrm{rank}(F)\le k}\|\mathbf{B}\mathbf{P}_F-\mathbf{B}\|_{p,2}^p\end{aligned}$$

Note that $|T|\ge k/2$ by averaging, so

$$\frac{c_p|T|^2}{ks}\ge\frac{c_pk}{4s}\ge\varepsilon$$

which proves the theorem. □

Chapter 14

Applications: strong coresets for ℓ_p subspace approximation [WY23a, WY24b]

In this chapter, we give a second important application of our study of sampling-based algorithms for ℓ_p subspace embeddings and construct the first *strong coresets* of nearly optimal size for a problem from computational geometry known as ℓ_p *subspace approximation*, which generalizes the well-known Frobenius norm low rank approximation problem.

Definition 14.0.1 (Rank k subspaces)**.** Let $k \in \mathbb{N}$ be a rank parameter. Then $\mathcal{F}_k$ denotes the set of all subspaces $F \subseteq \mathbb{R}^d$ with at most k dimensions, $\mathbf{V}_F \in \mathbb{R}^{d \times k}$ denotes an orthonormal basis for F, and $\mathbf{P}_F = \mathbf{V}_F \mathbf{V}_F^\top$ denotes the orthogonal projection matrix onto F.

Definition 14.0.2 ($(p,2)$-norm)**.** Let $1 \leq p < \infty$. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ have the n rows $\{\mathbf{a}_i\}_{i=1}^n \subseteq \mathbb{R}^d$. Then, we define the $(p,2)$-norm of $\mathbf{A}$ as

$$\|\mathbf{A}\|_{p,2} \coloneqq \left[\sum_{i=1}^n \|\mathbf{a}_i\|_2^p\right]^{1/p}$$

Definition 14.0.3 (ℓ_p Subspace approximation)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let k be a rank parameter. Let $1 \leq p < \infty$. Then, the ℓ_p *subspace approximation* problem is the problem of minimizing the objective function

$$\|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p = \sum_{i=1}^n \|\mathbf{a}_i^\top (\mathbf{I} - \mathbf{P}_F)\|_2^p$$

among all k dimensional subspaces $F \in \mathcal{F}_k$. We let

$$\mathsf{OPT} \coloneqq \min_{F \in \mathcal{F}_k} \|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p$$

denote the optimal value of this optimization problem, and we let $\mathbf{P}^*$ denote the projection matrix onto a rank k subspace achieving this optimum.

14.1 Coresets for ℓ_p subspace approximation

The ℓ_p subspace approximation problem, like clustering, is known to be NP-hard for any $p \neq 2$ [DTV11, GRSW12, CW15a], and thus coresets are especially important for obtaining tractable algorithms and have long been studied in the coreset literature [DRVW06, DMM06b, DV07, DMM08, FMSW10, FL11, SV12, VX12, CEM+15, CW15a, CMM17, SW18, LSW18, BLVZ19, FSS20, MRWZ20, HV20, BDM+20, FKW21, DP22, MMWY22, CW22, WY23a].

While there are many natural notions of coresets that could be defined for the ℓ_p subspace approximation problem, we will work with the requirement that the coreset approximate the objective function *for every* rank k subspace $F \in \mathcal{F}_k$. Such a coreset is known as a *strong coreset*, which we formally define as follows:

Definition 14.1.1 (Strong coresets for ℓ_p subspace approximation)**.** Let $1 \leq p < \infty$ and $0 < \varepsilon < 1$. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then, a diagonal map $\mathbf{S} \in \mathbb{R}^{n \times n}$ is a $(1 \pm \varepsilon)$ *strong coreset* for ℓ_p subspace approximation if

$$\|\mathbf{SA}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p = (1 \pm \varepsilon)\|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p$$

for every $F \in \mathcal{F}_k$. We refer to the number of nonzero entries $\mathsf{nnz}(\mathbf{S})$ of $\mathbf{S}$ as the size of the coreset.

The guarantee of Definition 14.1.1 can also be viewed as the natural generalization of *projection cost-preserving sketches* [CEM+15, CMM17, MM20] to the ℓ_p subspace approximation setting. Strong coresets are extremely powerful, and can be used to reduce the size of the input instance to at most $\mathsf{nnz}(\mathbf{S})$ points in $\mathsf{nnz}(\mathbf{S})$ dimensions. In particular, strong coresets of size $\mathsf{nnz}(\mathbf{S}) = \mathrm{poly}(k/\varepsilon)$ immediately remove the dependence of this problem on n and d, making this a powerful tool in the design of fast algorithms. Note that this guarantee is much stronger than many other possible guarantees for row subset selection that have been studied in the literature. One weaker guarantee is that of a *weak coreset*, which requires that if $\tilde{F}$ is the optimal solution to the subspace approximation problem for $\mathbf{SA}$, then it is also a $(1+\varepsilon)$-optimal solution for $\mathbf{A}$ [FL11, HV20]. Another further weaker guarantee is that the rows sampled by $\mathbf{S}$ span a $(1+\varepsilon)$-optimal solution, as studied by [DV07, SV12, CW15a]. The guarantee of Definition 14.1.1 immediately achieves both of these guarantees, and offers further benefits that cannot be realized by these other guarantees, for example applications to *constrained* versions of ℓ_p subspace approximation, or solving ℓ_p subspace approximation in distributed and streaming models via the merge-and-reduce technique [BDM+20, CWZ23].

While it has long been known that strong coresets of size $\mathrm{poly}(k, d, \varepsilon^{-1})$ independent of n exist [FL11], the first dimension-independent result, i.e. a strong coreset of size only $\mathrm{poly}(k, \varepsilon^{-1})$ independent of d, was achieved by the work of [SW18]. In fact, the result of [SW18] achieves a coreset with a nearly optimal size of $\mathsf{nnz}(\mathbf{S}) = \tilde{O}(k)\,\mathrm{poly}(\varepsilon^{-1})$ for $p < 2$ and $\mathsf{nnz}(\mathbf{S}) = \tilde{O}(k^{p/2})\,\mathrm{poly}(\varepsilon^{-1})$ for $p > 2$, which matches a lower bound of $\mathsf{nnz}(\mathbf{S}) = \tilde{\Omega}(k)$ for $p < 2$ and $\mathsf{nnz}(\mathbf{S}) = \tilde{\Omega}(k^{p/2})$ for $p > 2$ by a reduction to coreset lower bounds for ℓ_p subspace embeddings [LWW21, WY23a]. However, this result has a couple of drawbacks: (1) the construction requires time exponential in $\mathrm{poly}(k, \varepsilon^{-1})$, and (2) the result does not quite satisfy Definition 14.1.1, due to the fact that the coreset constructed by [SW18] is a weighted subset of points *with an appended coordinate*, rather than a weighted subset of the original data points themselves.

Drawback (1) was addressed in two follow up works of [HV20, FKW21], which both gave polynomial time algorithms for constructing strong coresets for ℓ_p subspace approximation. Furthermore, [HV20] also solves drawback (2) and gives the first polynomial time algorithm for constructing dimension-independent strong coresets as defined in Definition 14.1.1, using a technique known as sensitivity sampling. However, the coreset size in both of these works, while dimension-independent, is not optimal, and loses $\mathrm{poly}(k)$ factors in the coreset size. Thus, the following is one of the most central questions in the study of coresets:

Question 14.1.2. Do strong coresets for ℓ_p subspace approximation of size $\tilde{O}(k^{\max\{1,p/2\}})\,\mathrm{poly}(\varepsilon^{-1})$ exist? Is there a polynomial time algorithm for constructing such strong coresets?

The main result of this chapter is a positive resolution to Question 14.1.2 for all $1 \leq p < \infty$.

Theorem 14.1.3. Let $2 < p < \infty$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Then, there is an algorithm running in $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^{\omega})$ time which, with probability at least $1-\delta$, constructs a diagonal matrix $\mathbf{S}$ of size

$$\mathsf{nnz}(\mathbf{S}) = \frac{k^{p/2}}{\varepsilon^{p^2/2+p}}(\log(k/\varepsilon\delta))^{O(p)}$$

satisfying Definition 14.1.1, that is,

$$\|\mathbf{SA}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p = (1\pm\varepsilon)\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p \qquad \text{for every subspace } F \subseteq \mathbb{R}^d \text{ of rank at most } k.$$

Theorem 14.1.4. Let $1 \leq p < 2$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Then, there is an algorithm running in $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^{\omega})$ time which, with probability at least $1-\delta$, constructs a diagonal matrix $\mathbf{S}$ of size

$$\mathsf{nnz}(\mathbf{S}) = \frac{k}{\varepsilon^{4/p+2}}(\log(k/\varepsilon\delta))^{O(1)}$$

satisfying Definition 14.1.1, that is,

$$\|\mathbf{SA}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p = (1\pm\varepsilon)\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p \qquad \text{for every subspace } F \subseteq \mathbb{R}^d \text{ of rank at most } k.$$

We give several remarks concerning our results. First, as discussed earlier, we are the first to establish even the *existence* of a weighted subset of points with the property of Theorems 14.1.3 and 14.1.4. Furthermore, we are able to construct such a subset in nearly input-sparsity time. The running time nearly matches the time it takes to approximately solve a least squares linear regression problem in the current matrix multiplication time [CSWZ23], which is a natural barrier for the ℓ_p subspace approximation problem. Finally, we note that for $p < 2$, the fact that we achieve nearly linear size for the strong coreset guarantee implies that we simultaneously achieve the first nearly optimal size guarantees for other weaker guarantees that have been studied intensely in the past, including weak coresets [FL11, HV20] and subsets of rows spanning a $(1+\varepsilon)$-optimal solution [DV07, SV12, CW15a]. That is, for $p < 2$, we are in fact the first to resolve Question 14.1.2 even for weaker notions of coresets, both for existence and efficient constructions. For $p > 2$, we obtain the best known construction for weak coresets and the best known efficient construction for row subsets spanning a $(1+\varepsilon)$-optimal solution.

Pitfalls in prior work

The central technique of [SW18] is a structural result which shows the existence of a *representative subspace* $S \subseteq \mathbb{R}^d$ with $s = O(k)\,\mathrm{poly}(\varepsilon^{-1})$ dimensions[1] such that for any k-dimensional subspace $F \subseteq \mathbb{R}^d$,

$$\|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p = (1 \pm \varepsilon)\|[\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F), \mathbf{b}_S]\|_{p,2}^p$$

where $\mathbf{b}_S \in \mathbb{R}^n$ is the vector with ith entry given by $\mathbf{b}_S(i) = \left\|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_S)\right\|_2$, and $[\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F), \mathbf{b}_S]$ is the $n \times (d+1)$ matrix formed by the concatenation of $\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F)$ and $\mathbf{b}_S$. That is, the ℓ_p subspace approximation cost of F can be approximated by the projection cost onto the subspace S, plus the additional projection cost of the lower dimensional points $\mathbf{A}\mathbf{P}_S$ to the query subspace F. This reduces the subspace approximation problem in d dimensions to a similar problem in $s+1$ dimensions. In turn, this lower dimensional problem can be solved using dimension-dependent techniques, since the dimension is now only $s + 1 = O(k)\,\mathrm{poly}(\varepsilon^{-1})$. Then to analyze sampling algorithms, [SW18] show that Dvoretzky's theorem [Dvo61, FLM77, PVZ17] can be applied to convert the problem of approximating the $(p, 2)$-norm to a problem of approximating the (p, p)-norm, i.e. the entrywise ℓ_p norm, which can then be handled by sampling techniques for approximating ℓ_p norms of vectors in a subspace [CP15, WY23b], which admit tight sampling bounds. While this algorithm achieves a nearly optimally-sized data structure for approximating the ℓ_p subspace approximation cost, this algorithm requires exponential time, due to the fact that finding the representative subspace S requires solving the original ℓ_p subspace approximation problem to $(1 + \varepsilon)$ accuracy, which is not known to be solvable in polynomial time. The work of [FKW21] addressed this problem by introducing a polynomial time algorithm for finding such a subspace S, but the dimension of S found by this algorithm loses $\mathrm{poly}(k)$ factors, leading to suboptimal size in the coreset.

On the other hand, the result of [HV20] takes a different approach based on the classic *sensitivity sampling* technique [LS10, FL11, VX12], and uses the representative subspace constructed [SW18] in an *existential* manner rather than algorithmic. In the sensitivity sampling approach, one first defines *sensitivity scores*

$$\boldsymbol{\sigma}_i(\mathbf{A}) \coloneqq \sup_{F \in \mathcal{F}_k} \frac{\left\|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_F)\right\|_2^p}{\|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p} \tag{14.1}$$

for each row $i \in [n]$ which represent the largest fraction of the cost occupied by the ith coordinate, ranging over all queries $F \in \mathcal{F}_k$. Then, by Bernstein bounds, it follows that for any fixed $F \in \mathcal{F}_k$, sampling the rows $i \in [n]$ proportionally to the sensitivity scores preserves $\|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p$ up to $(1 \pm \varepsilon)$ factors. Naïvely, one can apply this result to every F in a net over the space of rank k subspaces F, which has size roughly $\exp(dk)$, and apply a net argument to construct coresets of size $\mathrm{poly}(d, k, \varepsilon^{-1})$. The work of [HV20] improves this argument by showing that the existence of the representative subspace S constructed by [SW18] gives an improved analysis of sensitivity sampling which converts a guarantee for coresets that only preserve the cost of an optimal subspace (known as a *weak coreset*) to a strong coreset guarantee. The fact that weak coresets admit dimension-independent bounds is an older result of [FL11], and thus [HV20]

[1] We improve the analysis of this result by a $1/\varepsilon^3$ factor in Appendix 14.2.

show that sensitivity sampling admits dimension-independent strong coresets as well. However, the key problem in this analysis is in the use of sensitivity sampling [FL11] to obtain the weak coreset, since this result uses a VC-dimension argument which loses $\mathrm{poly}(k)$ factors. In summary, the problem is that finding a representative subspace of optimal size is computationally difficult [SW18, FKW21], but we do not know how to apply tight sampling bounds if we do not have access to an explicit representative subspace and instead must settle for VC-dimension arguments which lose $\mathrm{poly}(k)$ factors in the coreset size [HV20].

Ridge leverage scores

Our algorithmic technique takes a drastically different approach compared to the prior works of [SW18, HV20, FKW21]. Our starting point is a result of [CMM17], which resolves Question 14.1.2 for the much simpler case of $p = 2$. For $p = 2$, finding an explicit rank $O(k)\,\mathrm{poly}(\varepsilon^{-1})$ with properties similar to the representative subspace S is not difficult due to the singular value decomposition (SVD) [DMM06b, DMM08, CEM+15, CMM17]. However, as noted by [CMM17], while this gives a polynomial time algorithm for low rank approximation for $p = 2$, finding these scores is already as hard as low rank approximation itself. Thus, this defeats the purpose of finding the coreset if the goal is to design faster algorithms. To address this problem, [CMM17] make use of the following alternative scores for a sampling-based algorithm, known as the *ridge leverage scores*.

Definition 14.1.5 (Ridge leverage scores [AM15, CMM17])**.** Let $\lambda > 0$ and $\mathbf{A} \in \mathbb{R}^{n\times d}$. Then, for each $i \in [n]$, the ith ridge leverage score is defined as

$$\boldsymbol{\tau}_i^{\lambda}(\mathbf{A}) \coloneqq \mathbf{a}_i^{\top}(\mathbf{A}^{\top}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{a}_i = \sup_{\mathbf{x}\in\mathbb{R}^d} \frac{[\mathbf{A}\mathbf{x}](i)^2}{\|\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2}.$$

Ridge leverage scores can be approximated very quickly [SS11, DMMW12, CW13, CLM+15], and can be approximated up to $O(1)$ factors in just $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^{\omega})$ time, where ω is the exponent of matrix multiplication.

The main result of [CMM17] establishes that if we set $\lambda = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$, then sampling $\tilde{O}(k/\varepsilon^2)$ rows $\mathbf{a}_i$ of $\mathbf{A}$ proportionally to their ridge leverage scores (see Definition 6.1.1) yields a strong coreset $\mathbf{S}$ of nearly optimal size satisfying Definition 14.1.1 for $p = 2$. Furthermore, the scores $\boldsymbol{\tau}_i^{\lambda}(\mathbf{A})$ only depend on a constant factor approximation to the *value* of the optimal low rank approximation, which can be obtained more readily than a subspace which (approximately) witnesses this value. However, the analysis of [CMM17] is highly specific to the ℓ_2 norm, for instance making heavy use of the structural properties of the SVD and the fact that the $(p, 2)$-norm is an entrywise norm for $p = 2$, and thus does not apply to $p \neq 2$. Nonetheless, several key ideas still do carry over to the setting of $p \neq 2$, which will be crucial to our analysis.

- First of all, we show that the ridge leverage scores are useful as sampling probabilities for ℓ_p subspace approximation if we take their $(p/2)$-th roots. By doing so, we are able to tap into the remarkable fact that the ridge leverage scores sum to at most $O(k)$ (Lemma 14.3.5). Note that this fact crucially relies on the special structure of the SVD, which is a factorization that is generally only useful for the Frobenius norm rather than the $(p, 2)$-norm, so this may be somewhat surprising.

- A second idea is that ridge leverage score sampling provides a *subspace embedding* guarantee with much fewer row samples than a standard relative error subspace embedding, by trading off the sample complexity for an additive error. That is, when $\lambda \to 0$, then it is known that sampling the rows of $\mathbf{A}$ proportionally to the ridge leverage scores gives the guarantee that with constant probability,

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_2^2 \qquad \text{for every } \mathbf{x} \in \mathbb{R}^d$$

  when $\tilde{O}(d/\varepsilon^2)$ rows are sampled [DMM06a, CLM+15]. Although sample sizes scaling as d are too expensive in our setting, [CMM17] show that if $\lambda = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$ and we sample rows of $\mathbf{A}$ proportionally to $\boldsymbol{\tau}_i^{\lambda}(\mathbf{A})$, then with only $\tilde{O}(k/\varepsilon^2)$ rows, we can get the guarantee that

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_2^2 \pm \varepsilon\lambda\|\mathbf{x}\|_2^2.$$

  In the context of low rank approximation, this additive error is small enough that it only distorts the approximate cost by an additive $\varepsilon \cdot \mathsf{OPT} = \varepsilon \cdot \|\mathbf{A} - \mathbf{A}_k\|_F^2$, and we will make a similar argument for ℓ_p subspace approximation as well.
- A third key idea is the observation that the ridge leverage scores provide sampling scores which are agnostic to any specific subspace (as opposed to, e.g., leverage scores of a fixed low-dimensional subspace), which allows us to reason about a subspace S that is hard to find algorithmically. This is one of the key reasons why we are able to obtain a polynomial time algorithm for constructing our coreset, despite the heavy use of the properties of the representative subspace S in the analysis.
- Finally, we follow the rough analysis plan of splitting the quantity $\|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}$ into a "head" term that lives in the top $\tilde{O}(k)$ important dimensions, a "tail" term involving the projection off of this top subspace, and a "cross" term that involves the remaining error after considering the former two terms, which also appears in many prior works [VX12, CEM+15]. However, the concrete way in which we define these quantities and preserve them via sampling is quite different from prior work.

14.1.1 Technical overview

We now give an overview of the ideas we introduce for our sampling results.

Reduction to embedding low rank matrices

Our starting point is still based on the structural result of [SW18]: there exists an s-dimensional subspace S for $s = \dim(S) = O(k)\operatorname{poly}(\varepsilon^{-1})$ and a vector $\mathbf{b}_S \in \mathbb{R}^n$ such that for any k-dimensional subspace F,

$$\|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p = (1 \pm \varepsilon)\|[\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F), \mathbf{b}_S]\|_{p,2}^p.$$

Our analysis will roughly take two steps. First, we will show that $\mathbf{S}$ preserves the right hand side, i.e.,

$$\|\mathbf{S}[\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F), \mathbf{b}_S]\|_{p,2}^p = (1 \pm \varepsilon)\|[\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F), \mathbf{b}_S]\|_{p,2}^p$$

and then in the second step, we will show that for the same subspace S, we have

$$\|\mathbf{SA}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p = (1\pm\varepsilon)\|\mathbf{S}[\mathbf{AP}_S(\mathbf{I}-\mathbf{P}_F),\mathbf{b}_S]\|_{p,2}^p. \tag{14.2}$$

This chain of bounds will show that

$$\|\mathbf{SA}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p = (1\pm 3\varepsilon)\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p$$

which is the desired result.

In fact, the sampling algorithm analysis for both of these steps will be quite similar. Ignoring the offset vector $\mathbf{b}_S$ for now for simplicity, the first step essentially asks for the guarantee that

$$\|\mathbf{SAX}\|_{p,2}^p = (1\pm\varepsilon)\|\mathbf{AX}\|_{p,2}^p \tag{14.3}$$

for every $\mathbf{X}$ with columns that lie in the subspace S. This guarantee essentially reduces to sampling an ℓ_p subspace embedding for the subspace S, but there is an additional challenge that we cannot afford to explicitly compute S if we want polynomial time algorithms.

The second step will in fact follow from a generalization of this guarantee. The representative subspace theorem of [SW18] shows that (14.2) will follow if

$$\|\mathbf{SA}(\mathbf{P}_{S\cup F}-\mathbf{P}_S)\|_{p,2}^p \le \varepsilon^p\,\mathsf{OPT} \qquad \text{for every } F\in\mathcal{F}_k, \tag{14.4}$$

where $\mathbf{P}_{S\cup F}$ denotes the projection matrix onto $\mathrm{span}(S\cup F)$. Furthermore, by the construction of S, S already satisfies $\|\mathbf{A}(\mathbf{P}_{S\cup F}-\mathbf{P}_S)\|_{p,2}^p \le \varepsilon^p\,\mathsf{OPT}$. Thus, it suffices to show that

$$\|\mathbf{SAX}\|_{p,2}^p = O(1)\|\mathbf{AX}\|_{p,2}^p \qquad \text{for all } \mathbf{X}\in\mathbb{R}^{d\times d} \text{ with } \mathrm{rank}(\mathbf{X})\le k \text{ and } \|\mathbf{X}\|_2\le 1.$$

Note that this differs from (14.3) since it asks for $\mathbf{S}$ to preserve *all* low rank matrices, rather than $\mathbf{X}$ with columns restricted in a low dimensional subspace. Thus, this guarantee is substantially more interesting than the first guarantee, and complicates our analysis. We note that for $p=2$, (14.3) is actually sufficient to show (14.4), since if S is chosen as the top $O(k/\varepsilon^2)$ singular directions of $\mathbf{A}$, then $\mathbf{A}(\mathbf{I}-\mathbf{P}_S)$ has operator norm at most $O(\varepsilon^2/k)\|\mathbf{A}-\mathbf{A}_k\|_F^2$. This operator norm is then sufficient for (14.4). However, such operator norm-based arguments are not available for $p\neq 2$ due to the lack of an SVD.

A crucial relaxation is that it in fact suffices to show that

$$\|\mathbf{SAX}\|_{p,2}^p = (1\pm\varepsilon)\|\mathbf{AX}\|_{p,2}^p \pm \varepsilon\,\mathsf{OPT} \tag{14.5}$$

whenever we apply this sampling theorem. Thus for the rest of this technical overview, we will focus on showing (14.5).

Idea 1: additive-multiplicative ℓ_p subspace embeddings via root ridge leverage scores

We begin by using Dvoretzky's theorem to embed the ℓ_2 norm into the ℓ_p norm, so that we have

$$\|\mathbf{AX}\|_{p,2}^p = (1\pm\varepsilon)\frac{1}{m}\|\mathbf{AXH}\|_{p,p}^p$$

where $\mathbf{H} \in \mathbb{R}^{d \times m}$ is an i.i.d. standard Gaussian matrix. By embedding the $(p, 2)$-norm into an entrywise ℓ_p norm, we decouple the norm of the columns, reducing our problem to preserving the ℓ_p norm of vectors of the form $\mathbf{Ax}$. That is, we seek guarantees of the form $\|\mathbf{SAx}\|_p^p \approx \|\mathbf{Ax}\|_p^p$. Such guarantees are known as ℓ_p *subspace embeddings*, and are well-studied in the literature.

The first new ingredient in our analysis is to adapt the additive-multiplicative ℓ_2 subspace embedding idea of [CMM17]. In this result, [CMM17] show that if $\mathbf{S}$ is taken to be a sampling matrix with probabilities proportional to the ridge leverage scores $\boldsymbol{\tau}_i^\lambda(\mathbf{A}) = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$, then one obtains the additive-multiplicative guarantee

$$\|\mathbf{SAx}\|_2^2 = (1 \pm \varepsilon)\|\mathbf{Ax}\|_2^2 \pm \varepsilon\lambda\|\mathbf{x}\|_2^2$$

with only $\tilde{O}(k/\varepsilon^2)$ samples. This fact immediately follows from applying the more standard guarantee for *leverage score sampling* on a concatenated matrix $[\mathbf{A}; \sqrt{\lambda}\mathbf{I}] \in \mathbb{R}^{(n+d)\times d}$, where $\mathbf{I}$ is the $d \times d$ identity. For an ℓ_p version of this result, we use the *root leverage scores* discussed in Chapter 8.

With the ℓ_p subspace embedding theorem in hand, we can now apply a similar trick as [CMM17]: we set $\lambda = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$ so that we only sample $\tilde{O}(k^{p/2})\,\mathrm{poly}(\varepsilon^{-1})$ rows for $p > 2$ and $\tilde{O}(k)\,\mathrm{poly}(\varepsilon^{-1})$ rows for $p < 2$, and then obtain an additive-multiplicative subspace embedding guarantee by viewing it as a subspace embedding for the matrix formed by concatenating $\mathbf{A}$ with $\sqrt{\lambda}\mathbf{I}$, so that the leverage scores of the concatenated matrix correspond to the ridge leverage scores of $\mathbf{A}$. The resulting guarantee is that

$$\|\mathbf{SAx}\|_p^p = (1 \pm \varepsilon)\|\mathbf{Ax}\|_p^p \pm \varepsilon\lambda^{p/2}\|\mathbf{x}\|_p^p.$$

Now, we can apply the above ℓ_p affine embedding guarantees for the sampling matrix $\mathbf{S}$ on each column of $\frac{1}{m}\|\mathbf{AXH}\|_{p,p}^p$ to obtain the approximation guarantee

$$\frac{1}{m}\|\mathbf{SAXH}\|_{p,p}^p = (1 \pm \varepsilon)\frac{1}{m}\|\mathbf{AXH}\|_{p,p}^p \pm \varepsilon\frac{\lambda^{p/2}}{m}\|\mathbf{XH}\|_{p,p}^p.$$

Now by applying Dvoretzky's theorem to revert the (p, p)-norm back to the $(p, 2)$-norm, we obtain

$$\|\mathbf{SAX}\|_{p,2}^p = (1 \pm \varepsilon)\|\mathbf{AX}\|_{p,2}^p \pm \varepsilon\lambda^{p/2}\|\mathbf{X}\|_{p,2}^p.$$

Finally, it remains to bound $\lambda^{p/2}\|\mathbf{X}\|_{p,2}^p$, but here we will encounter some problems.

Problems when bounding the additive error

To bound the additive error $\lambda^{p/2}\|\mathbf{X}\|_{p,2}^p$, we will case on $p < 2$ and $p > 2$. We may assume without loss of generality that $\mathbf{X}$ has at most n rows, by restricting the analysis to the row span of $\mathbf{A}$ throughout. Then for $p < 2$, $\lambda^{p/2}$ is at most

$$\lambda^{p/2} = \frac{\|\mathbf{A} - \mathbf{A}_k\|_F^p}{k^{p/2}} \leq \frac{\|\mathbf{A}(\mathbf{I} - \mathbf{P}^*)\|_F^p}{k^{p/2}} \leq \frac{\|\mathbf{A}(\mathbf{I} - \mathbf{P}^*)\|_{p,2}^p}{k^{p/2}} = \frac{\mathsf{OPT}}{k^{p/2}} \tag{14.6}$$

by the monotonicity of ℓ_p norms, while for $p > 2$, $\lambda^{p/2}$ is at most

$$\lambda^{p/2} = \frac{\|\mathbf{A} - \mathbf{A}_k\|_F^p}{k^{p/2}} \leq \frac{\|\mathbf{A}(\mathbf{I} - \mathbf{P}^*)\|_F^p}{k^{p/2}} \leq \frac{n^{p/2-1}\|\mathbf{A}(\mathbf{I} - \mathbf{P}^*)\|_{p,2}^p}{k^{p/2}} = \frac{n^{p/2-1}\,\mathsf{OPT}}{k^{p/2}}. \tag{14.7}$$

Furthermore,

$$\|\mathbf{X}\|_{p,2}^p \leq \begin{cases} n^{1-p/2}\|\mathbf{X}\|_{2,2}^p & \text{if } p < 2 \\ \|\mathbf{X}\|_{2,2}^p & \text{if } p > 2 \end{cases} \leq \begin{cases} n^{1-p/2}s^{p/2} & \text{if } p < 2 \\ s^{p/2} & \text{if } p > 2 \end{cases}$$

by relating ℓ_p and ℓ_2 norms in n dimensions and using that $\mathrm{rank}(\mathbf{X}) \leq s$ and $\|\mathbf{X}\|_2 \leq 1$. Then overall, we obtain a bound of

$$\lambda^{p/2}\|\mathbf{X}\|_{p,2}^p \leq \begin{cases} n^{1-p/2}s^{p/2}\dfrac{\mathsf{OPT}}{k^{p/2}} & \text{if } p < 2 \\ n^{p/2-1}s^{p/2}\dfrac{\mathsf{OPT}}{k^{p/2}} & \text{if } p > 2 \end{cases}$$

Note that if we use rank s root ridge leverage scores instead of rank k root ridge leverage scores, we would be able to replace the $k^{p/2}$ on the denominator by $s^{p/2}$ to cancel out the $s^{p/2}$ in the numerator, with only a $\mathrm{poly}(\varepsilon^{-1})$ cost to the sample complexity. However, even still, our bound is $n^{1-p/2}\,\mathsf{OPT}$ for $p < 2$ and $n^{p/2-1}\,\mathsf{OPT}$ for $p > 2$, which is off by $\mathrm{poly}(n)$ factors from our goal of OPT in either case.

In order to fix this problem and improve our analysis by $\mathrm{poly}(n)$ factors, we will use two different types of "flattening" tricks, one for $p < 2$ and one for $p > 2$, which we discuss in the next two sections.

Idea 2: Dvoretzky's theorem for sharper additive error bounds for $p > 2$

To overcome the previous issue for $p > 2$, we note that we have an additional degree of freedom when choosing to concatenate $\mathbf{A}$ with $\sqrt{\lambda}\mathbf{I}$ when analyzing the ridge leverage score sampling algorithm. Indeed, as long as we concatenate $\mathbf{A}$ with $\sqrt{\lambda}\mathbf{U}$ for any orthonormal matrix $\mathbf{U}$, then the leverage scores of $\mathbf{A}$ concatenated with $\sqrt{\lambda}\mathbf{U}$ will have leverage scores which coincide with the ridge leverage scores of $\mathbf{A}$, since

$$\mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{A} + \lambda\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{a}_i = \mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{a}_i.$$

The resulting guarantee is that

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_p^p = (1 \pm \varepsilon)\|\mathbf{A}\mathbf{x}\|_p^p \pm \varepsilon\lambda^{p/2}\|\mathbf{U}\mathbf{x}\|_p^p, \tag{14.8}$$

so we may select $\mathbf{U}$ to be an orthonormal matrix which makes this additive error as small as possible. We will choose $\mathbf{U}$ to be a random $n \times d$ orthonormal matrix $\mathbf{G}$, which has the advantage of flattening the mass of $\mathbf{x}$ and thus minimizing the ℓ_p norm.

By Dvoretzky's theorem [Dvo61, FLM77, PVZ17], it follows that as long as n is at least $\tilde{O}(s^{p/2})\,\mathrm{poly}(\varepsilon^{-1}) = \tilde{O}(k^{p/2})\,\mathrm{poly}(\varepsilon^{-1})$, then for any $\mathbf{x}$ in a fixed s-dimensional subspace, we will have that

$$\|\mathbf{G}\mathbf{x}\|_p^p = (1 \pm \varepsilon)n^{1-p/2}\|\mathbf{x}\|_2^p. \tag{14.9}$$

This cancels out with the factor of $n^{p/2-1}$ that we lost in (14.7), giving us a sharp enough additive error. It may be tempting to reduce the additive error even further by choosing $\mathbf{G}$ to have $m \gg n$

rows rather than just n. However, this would affect the total number of rows sampled, since we would then need to oversample the root ridge leverage scores by a factor of $m^{p/2-1}$, which would increase the sample complexity. Note also that once the additive error is sufficiently small, (14.8) would give the purely multiplicative subspace embedding guarantee $\|\mathbf{SAx}\|_p^p = (1 \pm \varepsilon)\|\mathbf{Ax}\|_p^p$, for which there is a sample complexity lower bound of $\Omega(d^{p/2}/\varepsilon)$ [LWW21].

Although we have fixed the $n^{p/2-1}$ factor, we must now address a subtle issue. The above analysis works for a fixed rank s subspace specified by the low rank matrix $\mathbf{X}$. However, if we want this guarantee for every rank s matrix $\mathbf{X}$ with operator norm 1 as we need, then we run into problems, since for any fixed embedding $\mathbf{G}$ of dimension only $\tilde{O}(k^{p/2})\,\mathrm{poly}(\varepsilon^{-1})$, there exists a choice of $\mathbf{X}$ which causes (14.9) to fail. To fix our final problem, we crucially exploit independence in our analysis. We first fix the sampling matrix $\mathbf{S}$ and let $\mathbf{X} \subseteq \mathbb{R}^{d\times d}$ be the rank s matrix with $\|\mathbf{AX}\|_{p,2}^p + \mathsf{OPT}\|\mathbf{X}\|_2^p \le 1$ that maximizes the sampling error $|\|\mathbf{SAX}\|_{p,2}^p - \|\mathbf{AX}\|_{p,2}^p|$. Note then that $\mathbf{X}$ depends only on $\mathbf{S}$ but not on $\mathbf{G}$, so we may bound $\|\mathbf{GX}\|_{p,2}^p$ as we did before. This completes our proof sketch for $p > 2$.

Idea 3: splitting rows for sharper additive error bounds for $p < 2$

To improve our argument for $p < 2$, we will sharpen the bound of (14.6). The loose bound that we will tighten is bounding the Frobenius norm loss $\|\mathbf{A}(\mathbf{I} - \mathbf{P}^*)\|_F^p$ by the $(p, 2)$-norm loss $\|\mathbf{A}(\mathbf{I} - \mathbf{P}^*)\|_{p,2}^p$. For general matrices, this bound is indeed tight since the rows of $\mathbf{A}(\mathbf{I} - \mathbf{P}^*)$ could be imbalanced so that most of the mass is concentrated on a few rows. However, this bound is loose when the rows are flat, in which case there can be a $\mathrm{poly}(n)$ factor separation in the two quantities. We will show how to recover this separation.

A classic result of [VX12] shows that the sensitivity scores (14.1) for ℓ_p subspace approximation sum to at most $O(k)$ for $p < 2$. Then, a standard flattening argument shows that by replacing rows $\mathbf{a}_i$ with large sensitivity with l copies of the scaled row $\mathbf{a}_i/l^{1/p}$, we obtain a new matrix $\mathbf{A}'$ with $n' \le 2n$ rows that are each just scaled copies of rows of $\mathbf{A}$, such that $\boldsymbol{\sigma}_{i'}(\mathbf{A}') = O(k/n)$ for every row $i' \in [n']$ and $\|\mathbf{A}'(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p = \|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p$ for every $F \in \mathcal{F}_k$. Because this matrix is now flat, it can be shown that

$$\|\mathbf{A}'(\mathbf{I} - \mathbf{P}^*)\|_F^2 \lesssim (k/n)^{2/p-1}\,\mathsf{OPT}^{2/p}.$$

Thus by replacing $\mathbf{A}$ with $\mathbf{A}'$, we obtain a matrix formed by the rows of $\mathbf{A}$ that gives the same objective function, yet has a much smaller additive error when bounding λ, giving

$$\lambda^{p/2} = \frac{\|\mathbf{A}' - \mathbf{A}'_k\|_F^p}{k^{p/2}} \le \frac{\|\mathbf{A}'(\mathbf{I} - \mathbf{P}^*)\|_F^p}{k^{p/2}} \le (k/n)^{1-p/2}\frac{\|\mathbf{A}'(\mathbf{I} - \mathbf{P}^*)\|_{p,2}^p}{k^{p/2}} = (k/n)^{1-p/2}\frac{\mathsf{OPT}}{k^{p/2}}.$$

rather than the original bound in (14.6). We note, however, that this argument is still lossy, since the standard sensitivity-based flattening argument would flatten *any* matrix, whereas we only need this result for a single constant factor approximate subspace $\tilde{F}$. Thus, we instead explicitly compute a constant factor bicriteria solution, which can be done very quickly [DTV11, FKW21, WY23a] (see Lemma 14.3.2), and flatten this particular solution nearly optimally, so that we instead get the bound

$$\lambda^{p/2} = \frac{\|\mathbf{A}' - \mathbf{A}'_k\|_F^p}{k^{p/2}} \le \frac{\|\mathbf{A}'(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_F^p}{k^{p/2}} \le (1/n)^{1-p/2}\frac{\|\mathbf{A}'(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_{p,2}^p}{k^{p/2}} = n^{p/2-1}\frac{O(\mathsf{OPT})}{k^{p/2}}.$$

Thus, we recover the extra factor of $n^{1-p/2}$ lost when converting from the ℓ_p norm to the ℓ_2 norm. This completes our proof sketch for $p < 2$.

14.1.2 Corollaries

Streaming and distributed models

A simple corollary of our nearly optimal constructions for strong coresets is that we immediately obtain similar results in *streaming* and *distributed* models of computation (see Section 1.3.3). In the streaming model, the rows $\mathbf{a}_i$ of the input matrix $\mathbf{A}$ arrive one at a time, and we wish to maintain a strong coreset for $\mathbf{A}$. In this setting, the classic *merge-and-reduce* technique (see, e.g., [BDM+20] for a discussion) shows that a construction for a coreset of size $\tilde{O}(k^c)\operatorname{poly}(\varepsilon^{-1})$ can be converted into a streaming implementation of size $\tilde{O}(k^c)\operatorname{poly}(\varepsilon^{-1}\log n)$ by setting the accuracy parameter to $\varepsilon' = \varepsilon/\log n$ and composing the coreset construction in a binary tree fashion. Recent work of [CWZ23] shows that this argument can in fact be sharpened to a $\operatorname{poly}(\log\log n)$ factor overhead rather than $\operatorname{poly}(\log n)$, by first computing an online coreset. Similarly, in the distributed model, the rows of $\mathbf{A}$ are partitioned among t servers, and we wish to communicate a strong coreset to a central coordinator. This task can be solved nearly optimally if each server computes a coreset, sends their coreset to the central coordinator, and the central coordinator computes a coreset for the collection of coresets.

Online coresets

Next, we note an application of our result to designing algorithms for *online* coresets for ℓ_p subspace approximation. For ℓ_p subspace approximation, the works of [BLVZ19, BDM+20] studied the case of $p = 2$ based on the result of [CMM17], while [WY23a] studied the case of $p \neq 2$, achieving a coreset size of roughly $\tilde{O}(k^{p+O(1)})\operatorname{poly}(\varepsilon^{-1})$ by analyzing an algorithm based on sensitivity sampling [HV20].[2] One of the main open questions left in [WY23a] is whether there exists an online coreset algorithm which samples only $\tilde{O}(k^{p/2+O(1)})\operatorname{poly}(\varepsilon^{-1})$ rows for $p > 2$. Our ℓ_p subspace approximation coreset result resolves this question nearly optimally. Our results here are given in Section 14.6.1.

Entrywise ℓ_p low rank approximation

Finally, we note that for $p < 2$, our nearly optimal coresets for ℓ_p subspace approximation imply new algorithms for the related problem of *entrywise ℓ_p low rank approximation*.

Definition 14.1.6. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let k be a rank parameter. Let $1 \leq p < \infty$. Then, the *entrywise ℓ_p low rank approximation* problem is the problem of minimizing the objective function

$$\|\mathbf{A} - \mathbf{X}\|_{p,p}^p = \sum_{i=1}^{n}\sum_{j=1}^{d}|(\mathbf{A}-\mathbf{X})_{i,j}|^p$$

[2] In our discussion of online coresets, we allow for the $\tilde{O}(\cdot)$ notation to suppress polylogarithmic factors in n and an "online condition number" quantity κ^{OL} which appears in all prior works on online coresets and is known to be necessary.

among all rank k matrices $\mathbf{X} \in \mathbb{R}^{n\times d}$.

This problem is another computationally difficult variant of the low rank approximation problem, and approximation algorithms and hardness have been studied in a long line of work [SWZ17, CGK+17, DWZ+19, MW21, JLL+21, WY23a]. The works of [JLL+21, WY23a] show that for $p < 2$, if we multiply $\mathbf{A}$ on the right by a dense matrix $\mathbf{G}$ of p-stable random variables [Nol20] and then compute an ℓ_p subspace approximation coreset $\mathbf{S}$ of $\mathbf{AG}$ of size $\tilde{O}(k)$, then there exists a rank k matrix $\mathbf{V}$ such that

$$\|\mathbf{A} - \mathbf{VSA}\|_{p,p}^p \leq \tilde{O}(k^{1/p-1/2}) \min_{\mathrm{rank}(\mathbf{X})\leq k} \|\mathbf{A} - \mathbf{X}\|_{p,p}^p.$$

Among subset selection algorithms, this approximation guarantee is nearly optimal [MW21]. Furthermore, because $\mathbf{S}$ is constructed based on sketching and coresets for ℓ_p subspace approximation, this algorithm can be implemented in streaming and distributed settings, and previously discussed. However, these prior results had drawbacks. The result of [JLL+21] relied on the coreset construction of [SW18], and thus required exponential time to run. In the work of [WY23a], this idea was applied in the setting of online coresets, but their online coreset required a size of at least k^4, which resulted in a suboptimal approximation factor of at least $k^{4(1/p-1/2)}$. Our result fixes both of these problems, by substantially speeding up the algorithm of [JLL+21] to achieve the first polynomial time subset selection algorithm selecting $\tilde{O}(k)$ columns with $\tilde{O}(k^{1/p-1/2})$ distortion, as well as the first online coreset algorithm which selects $\tilde{O}(k)$ rows with a $\tilde{O}(k^{1/p-1/2})$ distortion. Note that the prior best efficient subset selection algorithm of [MW21] selects $O(k \log d)$ rows for $\tilde{O}(k^{1/p-1/2})$ distortion.

14.2 Representative subspace theorem for ℓ_p subspace approximation

One of the main technical ingredients for our strong coreset is the *representative subspace theorem* of [SW18, Theorem 10], which shows that the ℓ_p subspace approximation cost can approximately be decomposed into a cost onto a low dimensional subspace plus the cost of projecting onto this subspace. We provide sharper bounds for this result in this section.

Theorem 14.2.1 (Representative subspace theorem)**.** Let $1 \leq p < \infty$. Suppose that an s-dimensional subspace S satisfies

$$\|\mathbf{A}(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_{p,2}^p \leq \varepsilon^p \cdot \mathsf{OPT}$$

for every $F \in \mathcal{F}_k$. Then if $\mathbf{P}_S$ is the projection matrix onto S and $\mathbf{b}_S \in \mathbb{R}^n$ is the vector defined by

$$\mathbf{b}_S(i) \coloneqq \left\|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_S)\right\|_2,$$

then

$$\text{for all } F \in \mathcal{F}_k, \qquad \|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p = (1 \pm \varepsilon)\|[\mathbf{AP}_S(\mathbf{I} - \mathbf{P}_F),\ \mathbf{b}_S]\|_{p,2}^p, \tag{14.10}$$

where $[\mathbf{A}\mathbf{P}_S(\mathbf{I}-\mathbf{P}_F),\ \mathbf{b}_S]$ denotes the $n \times (d+1)$ concatenation of $\mathbf{A}\mathbf{P}_S(\mathbf{I}-\mathbf{P}_F) \in \mathbb{R}^{n\times d}$ and $\mathbf{b}_S \in \mathbb{R}$. Furthermore, such a subspace S exists for

$$s = \frac{O(k)}{\varepsilon^{\max\{2,p\}}}$$

such that $\|\mathbf{b}_S\|_p^p \leq \mathsf{OPT}$.

We note that any subspace S' that contains a subspace S satisfying the properties of Theorem 14.2.1 will continue to have the same properties.

Lemma 14.2.2. Let $S' \supseteq S$ be two subspaces such that S satisfies the guarantees of Theorem 14.2.1. Then, S' does as well.

Proof. Note that for any $\mathbf{a} \in \mathbb{R}^d$ and k-dimensional subspace F,

$$\|\mathbf{a}^\top(\mathbf{P}_{S'} - \mathbf{P}_{S'\cup F})\|_2^2 \leq \|\mathbf{a}^\top(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_2^2$$

since $(\mathbf{P}_{S\cup F} - \mathbf{P}_S)$ maps its input to the component of F orthogonal to S and similarly for $(\mathbf{P}_{S'\cup F} - \mathbf{P}_{S'})$. Thus,

$$\|\mathbf{A}(\mathbf{P}_{S'} - \mathbf{P}_{S'\cup F})\|_{p,2}^p \leq \|\mathbf{A}(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_{p,2}^p \leq \varepsilon^p \cdot \mathsf{OPT}$$

It follows that (14.10) holds from arguments in [SW18, Theorem 10]. We also have that

$$\|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_{S'})\|_2^2 \leq \|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_S)\|_2^2$$

so $\|\mathbf{b}_{S'}\|_p^p \leq \|\mathbf{b}_S\|_p^p \leq \mathsf{OPT}$ holds as well. □

14.2.1 Sharper scalar inequalities

The following result simplifies and sharpens [SW18, Claim 2].

Lemma 14.2.3. Let $u, v, w \geq 0$ satisfy $u^2 = v^2 - w^2$. Then,

$$u^p \leq \begin{cases} \min\{\varepsilon v^p, 2^{p-1}\varepsilon^{1-2/p}(v^p - w^p)\} & 1 \leq p \leq 2 \\ v^p - w^p & 2 \leq p < \infty \end{cases}$$

Proof. The second inequality follows from the subadditivity of $(\cdot)^{p/2}$ [SW18] so it remains to show the first. We may assume that $v = 1$ by scaling. We also reparameterize $w = 1 - x$ for some $0 \leq x \leq 1$. Then,

$$u^p = (1 - (1-x)^2)^{p/2} \leq (2x)^{p/2}$$

and

$$\frac{u^p}{v^p - w^p} = \frac{u^p}{1 - (1-x)^p} \leq \frac{(2x)^{p/2}}{x} = 2^{p/2}x^{p/2-1}$$

Thus, if $x \leq \varepsilon^{2/p}/2$, then $u^p \leq \varepsilon$, and if $x \geq \varepsilon^{2/p}/2$, then $u^p \leq 2^{p-1}\varepsilon^{1-2/p}(v^p - w^p)$. □

The following result sharpens [SW18, Claim 5].

Lemma 14.2.4. Let $u, v \geq 0$ and $1 \leq p < \infty$. Then,

$$(u+v)^p \leq (1+\varepsilon)u^p + \frac{(2p)^p}{\varepsilon^{p-1}}v^p$$

Proof. We may assume that $u = 1$ by scaling. If $v \geq 1$, then $(1+v)^p \leq 2^p v^p$ so assume that $v \leq 1$. Then, $(1+v)^p \leq 1 + 2pv$ so if $2pv \leq \varepsilon$, then $(1+v)^p \leq \varepsilon$, while if $2pv \geq \varepsilon$, then

$$(1+v)^p \leq 1 + 2pv = 1 + \frac{2p}{v^{p-1}}v^p \leq 1 + \frac{(2p)^p}{\varepsilon^{p-1}}v^p.$$

□

The following result generalizes [SW18, Lemma 4] to $p > 1$.

Lemma 14.2.5. Let $a, b, f, g \geq 0$. Then,

$$|(a^2+b^2)^{p/2} - (f^2+g^2)^{p/2}| \leq \frac{(4p)^p}{2\varepsilon^{p-1}}(|a-f|^p + |b-g|^p) + \varepsilon\big((a^2+b^2)^{p/2} + (f^2+g^2)^{p/2}\big).$$

Proof. By Lemma 14.2.4, we have that

$$\|(a,b)\|_2^p \leq (\|(a-f,b-g)\|_2 + \|(f,g)\|_2)^p \leq (1+\varepsilon)\|(f,g)\|_2^p + \frac{(2p)^p}{\varepsilon^{p-1}}\|(a-f,b-g)\|_2^p$$

and similarly

$$\|(f,g)\|_2^p \leq (\|(a-f,b-g)\|_2 + \|(a,b)\|_2)^p \leq (1+\varepsilon)\|(a,b)\|_2^p + \frac{(2p)^p}{\varepsilon^{p-1}}\|(a-f,b-g)\|_2^p.$$

Thus,

$$|\|(a,b)\|_2^p - \|(f,g)\|_2^p| \leq \frac{(2p)^p}{\varepsilon^{p-1}}\|(a-f,b-g)\|_2^p + \varepsilon(\|(a,b)\|_2^p + \|(f,g)\|_2^p).$$

Finally, we bound

$$\|(a-f,b-g)\|_2^p \leq \|(a-f,b-g)\|_1^p \leq 2^{p-1}(|a-f|^p + |b-g|^p).$$

□

14.2.2 Proof of the representative subspace theorem

The first lemma shows that if $\|\mathbf{a}^\top(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_2$ is small, then the projection of a vector $\mathbf{a}$ onto $S \cup F$ is close to its projection onto S, and the projection of $\mathbf{a}^\top\mathbf{P}_{S\cup F}$ onto F is close to the projection of $\mathbf{a}^\top\mathbf{P}_S$ onto F.

Lemma 14.2.6. Let $S, F \subseteq \mathbb{R}^d$ be subspaces and let $\mathbf{a} \in \mathbb{R}^d$ be a vector. Then,

- $\|\mathbf{a}^\top(\mathbf{I} - \mathbf{P}_{S\cup F})\|_2 = \|\mathbf{a}^\top(\mathbf{I} - \mathbf{P}_S)\|_2 \pm \|\mathbf{a}^\top(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_2$
- $\|\mathbf{a}^\top(\mathbf{P}_{S\cup F} - \mathbf{P}_F)\|_2 = \|\mathbf{a}^\top\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F)\|_2 \pm \|\mathbf{a}^\top(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_2$

Proof. These are proven in [SW18]. We reproduce a proof for the reader's convenience. The first inequality is just the triangle inequality, so it remains to show the latter. One direction of the inequality follows by

$$\begin{aligned}\|\mathbf{a}^\top(\mathbf{P}_{S\cup F}-\mathbf{P}_F)\|_2 &= \min_{\mathbf{x}\in F}\|\mathbf{a}^\top\mathbf{P}_{S\cup F}-\mathbf{x}\|_2\\ &\le \|\mathbf{a}^\top\mathbf{P}_{S\cup F}-\mathbf{a}^\top\mathbf{P}_S\mathbf{P}_F\|_2 \le \|\mathbf{a}^\top(\mathbf{P}_{S\cup F}-\mathbf{P}_S)\|_2+\|\mathbf{a}^\top\mathbf{P}_S(\mathbf{I}-\mathbf{P}_F)\|_2\end{aligned}$$

and the other by

$$\begin{aligned}\|\mathbf{a}^\top\mathbf{P}_S(\mathbf{I}-\mathbf{P}_F)\|_2 &= \min_{\mathbf{x}\in F}\|\mathbf{a}^\top\mathbf{P}_S-\mathbf{x}\|_2\\ &\le \|\mathbf{a}^\top\mathbf{P}_S-\mathbf{a}^\top\mathbf{P}_F\|_2 \le \|\mathbf{a}^\top(\mathbf{P}_S-\mathbf{P}_{S\cup F})\|_2+\|\mathbf{a}^\top(\mathbf{P}_{S\cup F}-\mathbf{P}_F)\|_2\end{aligned}$$

□

We may combine Lemma 14.2.6 with Lemma 14.2.5 to show the following, which states that the projection cost of $\mathbf{a}$ onto F is approximately the sum of the cost of projecting onto S, and then projecting onto F.

Lemma 14.2.7. Let $S, F \subseteq \mathbb{R}^d$ be subspaces and let $\mathbf{a} \in \mathbb{R}^d$ be a vector. Then,

$$\begin{aligned}&\left|\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_F)\|_2^p - (\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_S)\|_2^2+\|\mathbf{a}^\top\mathbf{P}_S(\mathbf{I}-\mathbf{P}_F)\|_2^2)^{p/2}\right|\\ \le\ &\left(\frac{(4p)^p}{\varepsilon^{p-1}}+2^{p-1}\varepsilon\right)\|\mathbf{a}^\top(\mathbf{P}_S-\mathbf{P}_{S\cup F})\|_2^p + (2^{p-1}+1)\varepsilon\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_F)\|_2^p\end{aligned}$$

Proof. Note that by orthogonality,

$$\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_F)\|_2^2 = \|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_{S\cup F})\|_2^2+\|\mathbf{a}^\top(\mathbf{P}_{S\cup F}-\mathbf{P}_F)\|_2^2.$$

Then, we apply Lemma 14.2.5 with $a = \|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_{S\cup F})\|_2$, $b = \|\mathbf{a}^\top(\mathbf{P}_{S\cup F}-\mathbf{P}_F)\|_2$, $f = \|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_S)\|_2$, and $g = \|\mathbf{a}^\top\mathbf{P}_S(\mathbf{I}-\mathbf{P}_F)\|_2$ as well as the bound

$$|a-b|, |f-g| \le \|\mathbf{a}^\top(\mathbf{P}_S-\mathbf{P}_{S\cup F})\|_2$$

from Lemma 14.2.6 to see that

$$\begin{aligned}|\|(a,b)\|_2^2-\|(f,g)\|_2^2|^{p/2} &= \left|\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_F)\|_2^p - (\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_S)\|_2^2+\|\mathbf{a}^\top\mathbf{P}_S(\mathbf{I}-\mathbf{P}_F)\|_2^2)^{p/2}\right|\\ &\le \frac{(4p)^p}{2\varepsilon^{p-1}}(|a-f|^p+|b-g|^p)+\varepsilon(\|(a,b)\|_2^p+\|(f,g)\|_2^p)\\ &\le \frac{(4p)^p}{\varepsilon^{p-1}}\|\mathbf{a}^\top(\mathbf{P}_S-\mathbf{P}_{S\cup F})\|_2^p\\ &\qquad +\varepsilon(\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_F)\|_2^p+(\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_S)\|_2^2+\|\mathbf{a}^\top\mathbf{P}_S(\mathbf{I}-\mathbf{P}_F)\|_2^2)^{p/2}).\end{aligned}$$

Note that

$$\begin{aligned}(\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_S)\|_2^2+\|\mathbf{a}^\top\mathbf{P}_S(\mathbf{I}-\mathbf{P}_F)\|_2^2)^{p/2} &\le 2^{p-1}(\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_{S\cup F})\|_2^2+\|\mathbf{a}^\top(\mathbf{P}_{S\cup F}-\mathbf{P}_F)\|_2^2)^{p/2}\\ &\qquad +2^{p-1}\|\mathbf{a}^\top(\mathbf{P}_S-\mathbf{P}_{S\cup F})\|_2^p\\ &= 2^{p-1}\|\mathbf{a}^\top(\mathbf{I}-\mathbf{P}_F)\|_2^p+2^{p-1}\|\mathbf{a}^\top(\mathbf{P}_S-\mathbf{P}_{S\cup F})\|_2^p\end{aligned}$$

so combining the bounds gives the claimed result. □

It remains to construct a subspace S such that $\|\mathbf{A}(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_{p,2}^p$ is small for every k-dimensional subspace F.

Lemma 14.2.8. Let $1 \le p < \infty$ and $k \in \mathbb{N}$. There is an s-dimensional subspace S where $s = O(k/\varepsilon^{\max\{2,p\}})$ such that for every k-dimensional subspace F,

$$\|\mathbf{A}(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_{p,2}^p \le \varepsilon^p \,\mathsf{OPT}$$

where $\mathsf{OPT} = \min_{F\in\mathcal{F}_k} \|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p$.

Proof. The proof largely follows [SW18] combined with our improved inequalities proved earlier. We reproduce a proof for the reader's convenience.

Lemma 6 in [SW18] shows that there is an s-dimensional subspace S such that

$$\|\mathbf{A}(\mathbf{I} - \mathbf{P}_S)\|_{p,2}^p - \|\mathbf{A}(\mathbf{I} - \mathbf{P}_{S\cup F})\|_{p,2}^p \le \varepsilon^{\max\{2,p\}} \,\mathsf{OPT} \tag{14.11}$$

for every k-dimensional subspace $F \in \mathcal{F}_k$. We now use the fact that for any vector $\mathbf{a} \in \mathbb{R}^d$,

$$\|\mathbf{a}^\top(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_2^2 = \|\mathbf{a}^\top(\mathbf{I} - \mathbf{P}_S)\|_2^2 - \|\mathbf{a}^\top(\mathbf{I} - \mathbf{P}_{S\cup F})\|_2^2$$

by orthogonality and Lemma 14.2.3 (with $\varepsilon' = \varepsilon^p$) to show that

$$\|\mathbf{A}(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_{p,2}^p \le \begin{cases} \varepsilon^p \|\mathbf{A}(\mathbf{I} - \mathbf{P}_S)\|_{p,2}^p + 2\varepsilon^{p-2}(\|\mathbf{A}(\mathbf{I} - \mathbf{P}_S)\|_{p,2}^p - \|\mathbf{A}(\mathbf{I} - \mathbf{P}_{S\cup F})\|_{p,2}^p) & 1 \le p < 2 \\ \|\mathbf{A}(\mathbf{I} - \mathbf{P}_S)\|_{p,2}^p - \|\mathbf{A}(\mathbf{I} - \mathbf{P}_{S\cup F})\|_{p,2}^p & 2 \le p < \infty \end{cases}$$

by summing up the inequalities over vectors $\mathbf{a}_i$ for $i \in [n]$. By (14.11), we have that

$$\|\mathbf{A}(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_{p,2}^p \le 3\varepsilon^p \,\mathsf{OPT}$$

in any case. Rescaling ε by constant factors yields the statement of the theorem. □

Finally, we combine this bound with Lemma 14.2.7 to conclude Theorem 14.2.1.

14.3 Preliminaries

14.3.1 Dvoretzky's theorem

A classic result of Dvoretzky and Milman [Dvo61, Mil71] shows that a random subspace of a normed space is approximately Euclidean. We will need the following version of this result for ℓ_p norms:

Theorem 14.3.1 (Dvoretzky's theorem for ℓ_p norms [FLM77, PVZ17])**.** Let $1 \le p < \infty$ and $0 < \varepsilon < 1/p$. Let $n \ge O(\max\{\varepsilon^{-2}k, \varepsilon^{-1}k^{p/2}\})$, and let $\mathbf{G} \in \mathbb{R}^{n\times k}$ be an i.i.d. random Gaussian matrix. Then,

$$\mathbf{Pr}\Big\{\text{for all } \mathbf{x} \in \mathbb{R}^k,\ \|\mathbf{G}\mathbf{x}\|_p^p = (1 \pm \varepsilon)n\|\mathbf{x}\|_2^p\Big\} \ge \frac{2}{3}$$

14.3.2 Flattening

It is known that constant factor bicriteria solutions for ℓ_p subspace approximation can be computed quickly via convex relaxations [DTV11] or by combining sketching techniques with ℓ_p Lewis weight sampling [FKW21, WY23a]. The following lemma gives a version of [WY23a, Algorithm 3] that is optimized for running time.

Lemma 14.3.2 (Fast constant factor approximation)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, $1 \le p \le 2$, and $k \in \mathbb{N}$. Let $\mathbf{G} \in \mathbb{R}^{t\times d}$ be a sparse embedding matrix [NN13, Coh16] with $t = O(k\log(n/\delta))$ and sparsity $s = O(\log(n/\delta))$. Let $\tilde{F}$ denote the span of $O(t\log(t/\delta))$ rows sampled according to the ℓ_p Lewis weights of $\mathbf{A}\mathbf{G}^\top$ [CP15]. Then, with probability at least $1-\delta$, the following hold:

- $\|\mathbf{A}(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_{p,2}^p \le O(\mathsf{OPT})$.
- The subspace $\tilde{F}$ can be computed in $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + t^\omega)$ time

Proof. The correctness is shown in [WY23a], so it remains to argue the running time. The sparse embedding matrix $\mathbf{G}$ only requires time $\tilde{O}(\mathsf{nnz}(\mathbf{A})\log(1/\delta))$ to apply due to its sparsity. The ℓ_p Lewis weights of $\mathbf{A}\mathbf{G}^\top$ can then be computed in time $\tilde{O}(\mathsf{nnz}(\mathbf{A}\mathbf{G}^\top) + t^\omega) = \tilde{O}(\mathsf{nnz}(\mathbf{A}) + t^\omega)$ [CP15]. □

By using Lemma 14.3.2, we will obtain a fast algorithm for quickly *flattening* a matrix by splitting rows, which will be a crucial component of our sampling algorithm for $p < 2$. Similar techniques have long been used in the literature of ℓ_p subspace embeddings [BLM89, CP15, MMWY22, WY23b].

Lemma 14.3.3 (Flattening)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, $1 \le p < \infty$, and $k \in \mathbb{N}$. Let $F \subseteq \mathbb{R}^d$ be a subspace. Then, there is an $n' \times d$ matrix $\mathbf{A}'$ with $n \le n' \le (3/2)n$ such that $\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p = \|\mathbf{A}'(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p$ for every $F \in \mathcal{F}_k$ and

$$\|\mathbf{a}_i'^\top(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_2^p \le \frac{2}{n}\|\mathbf{A}'(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_{p,2}^p$$

for every $i \in [n']$. Furthermore, the rows of $\mathbf{A}'$ are reweighted rows of $\mathbf{A}$.

Proof. The proof follows, e.g., [MMWY22, Lemma 2.10]. Note that if we replace a row $\mathbf{a}_i$ by l copies of the scaled row $\mathbf{a}_i/l^{1/p}$, then $\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p = \|\mathbf{A}'(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p$ and for every row i' in $\mathbf{A}'$ that is a copy of $\mathbf{A}$, $\|\mathbf{a}_{i'}'^\top(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_2^p = \|\mathbf{a}_i^\top(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_2^p/l$. Now for every row i in $\mathbf{A}$ such that $\|\mathbf{a}_i^\top(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_2^p \ge 2\|\mathbf{A}^\top(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_{p,2}^p/n$, replace the row $\mathbf{a}_i$ with

$$l_i := \left\lceil \frac{\|\mathbf{a}_i^\top(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_2^p/\|\mathbf{A}^\top(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_{p,2}^p}{2/n} \right\rceil$$

copies of $\mathbf{a}_i/l_i^{1/p}$. Note then that the number of rows we add is at most

$$\sum_{i=1}^n (l_i - 1) \le \sum_{i=1}^n \frac{\|\mathbf{a}_i^\top(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_2^p/\|\mathbf{A}^\top(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_{p,2}^p}{2/n} \le \frac{n}{2}.$$

Furthermore, by construction, every row in the new matrix $\mathbf{A}'$ has sensitivity at most $C_2 k/n$. □

The advantage of flattening is that for $p < 2$, it makes the ℓ_2 subspace approximation cost much smaller than the ℓ_p subspace approximation cost. We will exploit the following result later in our results for $p < 2$.

Lemma 14.3.4. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, $1 \le p \le 2$, and $k \in \mathbb{N}$. Suppose that $\boldsymbol{\sigma}_i(\mathbf{A}) \le C\|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_2^p\|\mathbf{A}^\top(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_{p,2}^p/n$ for every $i \in [n]$. Then, we have

$$\|\mathbf{A}(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_F \le (C/n)^{1/p-1/2}\|\mathbf{A}(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_{p,2}.$$

Proof. We have

$$\begin{aligned}\|\mathbf{A}(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_F^2 &= \sum_{i=1}^n \|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_2^2 = \sum_{i=1}^n \|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_2^p(\|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_2^p)^{2/p-1} \\ &\le \sum_{i=1}^n \|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_2^p\left(\frac{Ck}{n}\|\mathbf{A}(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_{p,2}^p\right)^{2/p-1} \\ &= (Ck/n)^{2/p-1}\|\mathbf{A}(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_{p,2}^{2/p}.\end{aligned}$$

□

14.3.3 Properties of ridge leverage scores

It is known that for $\lambda = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$, the ridge leverage scores have a small sum.

Lemma 14.3.5 (Sum of ridge leverage scores [CMM17])**.** Let $\lambda = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$. Then,

$$\sum_{i=1}^n \boldsymbol{\tau}_i^\lambda(\mathbf{A}) \le 2k$$

Next, we show that ridge leverage scores upper bound the ℓ_p subspace approximation ℓ_2 sensitivities (14.1).

Lemma 14.3.6 (Ridge leverage scores bound sensitivities)**.** Let $\lambda = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$. Then,

$$\boldsymbol{\tau}_i^\lambda(\mathbf{A}) \ge \frac{1}{48}\sup_{F\in\mathcal{F}_k}\frac{\left\|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_F)\right\|_2^2}{\|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_F^2}$$

for every $i \in [n]$.

Proof. Note that

$$\|\mathbf{A} - \mathbf{A}_{2k}\|_2^2 = \boldsymbol{\sigma}_{k+1}^2(\mathbf{A} - \mathbf{A}_k) \le \frac{1}{k}\sum_{j=1}^k \boldsymbol{\sigma}_j^2(\mathbf{A} - \mathbf{A}_k) \le \frac{\|\mathbf{A} - \mathbf{A}_k\|_F^2}{k} = \lambda$$

so

$$\boldsymbol{\tau}_i^\lambda(\mathbf{A}) = \sup_{\mathbf{x}\in\mathbb{R}^d}\frac{[\mathbf{A}\mathbf{x}](i)^2}{\|\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2}$$

$$= \sup_{\mathbf{x}\in\mathbb{R}^d} \frac{[\mathbf{A}\mathbf{x}](i)^2}{\|\mathbf{A}_{2k}\mathbf{x}\|_2^2 + \|(\mathbf{A}-\mathbf{A}_{2k})\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2}$$
$$\geq \sup_{\mathbf{x}\in\mathbb{R}^d} \frac{[\mathbf{A}\mathbf{x}](i)^2}{\|\mathbf{A}_{2k}\mathbf{x}\|_2^2 + 2\lambda\|\mathbf{x}\|_2^2}.$$

Now let $F \in \mathcal{F}_k$ be any rank k subspace. Let G denote the span of the rows of $\mathbf{A}_{2k}$, F, and $\mathbf{a}_i$, which is a subspace of dimension at most $3k+1$. We then set $\mathbf{x} = \mathbf{P}_G(\mathbf{I}-\mathbf{P}_F)\mathbf{g}$ for a standard normal Gaussian vector $\mathbf{g}$. Note then that

$$[\mathbf{A}\mathbf{x}](i) = \mathbf{a}_i^\top \mathbf{P}_G(\mathbf{I}-\mathbf{P}_F)\mathbf{g} = \mathbf{a}_i^\top(\mathbf{I}-\mathbf{P}_F)\mathbf{g}$$

is distributed as a Gaussian with variance $\|\mathbf{a}_i^\top(\mathbf{I}-\mathbf{P}_F)\|_2^2$, so

$$\mathbf{Pr}\big\{[\mathbf{A}\mathbf{x}](i)^2 \geq \|\mathbf{a}_i^\top(\mathbf{I}-\mathbf{P}_F)\|_2^2/3\big\} > \frac{1}{2}.$$

Note also that

$$\mathbf{E}\big[\|\mathbf{A}_{2k}\mathbf{x}\|_2^2\big] = \mathbf{E}\big[\|\mathbf{A}_{2k}\mathbf{P}_G(\mathbf{I}-\mathbf{P}_F)\mathbf{g}\|_2^2\big] = \mathbf{E}\big[\|\mathbf{A}_{2k}(\mathbf{I}-\mathbf{P}_F)\mathbf{g}\|_2^2\big] \leq \|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_F^2$$

and

$$\mathbf{E}[\lambda\|\mathbf{x}\|_2^2] = \mathbf{E}[\lambda\|\mathbf{P}_G(\mathbf{I}-\mathbf{P}_F)\mathbf{g}\|_2^2] \leq \lambda(3k+1) \leq 4\|\mathbf{A}-\mathbf{A}_k\|_F^2.$$

Then by Markov's inequality, we have

$$\mathbf{Pr}\big\{\|\mathbf{A}_{2k}\mathbf{x}\|_2^2 + 2\lambda\|\mathbf{x}\|_2^2 \leq 16\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_F^2\big\} \geq \frac{1}{2}.$$

Thus with positive probability, there exists a vector $\mathbf{x}$ such that

$$\boldsymbol{\tau}_i^\lambda(\mathbf{A}) \geq \frac{[\mathbf{A}\mathbf{x}](i)^2}{\|\mathbf{A}_{2k}\mathbf{x}\|_2^2 + 2\lambda\|\mathbf{x}\|_2^2} \geq \frac{1}{48}\frac{\|\mathbf{a}_i^\top(\mathbf{I}-\mathbf{P}_F)\|_2^2}{\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_F^2}.$$

Since F was arbitrary, we conclude as desired. □

14.4 Reduction to additive-multiplicative ℓ_p affine embeddings

In this section, we show that in order to obtain sampling theorems that preserve $(p,2)$-norms of a matrix, it suffices to prove additive-multiplicative ℓ_p affine embedding guarantees for the sampling matrix $\mathbf{S}$. We consider the following notion of additive-multiplicative ℓ_p affine embeddings:

Definition 14.4.1 (Additive-multiplicative ℓ_p affine embedding)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$. Then, $\mathbf{S} \in \mathbb{R}^{r\times n}$ is a $(\lambda, \varepsilon, R)$-additive-multiplicative ℓ_p affine embedding if for every $\mathbf{x} \in \mathbb{R}^d$, we have

$$\|\mathbf{S}(\mathbf{A}\mathbf{x}+\mathbf{b})\|_p^p = \|\mathbf{A}\mathbf{x}+\mathbf{b}\|_p^p \pm \varepsilon(\|\mathbf{A}\mathbf{x}\|_p^p + R^p + \lambda^{p/2}\|\mathbf{x}\|_2^p)$$

Then, our main result of this section is the following lemma:

Lemma 14.4.2. Suppose that $\mathbf{S} \in \mathbb{R}^{r\times n}$ satisfies the $(\lambda, \varepsilon, R)$-additive-multiplicative ℓ_p affine embedding property for the matrix $\mathbf{A} \in \mathbb{R}^{n\times s}$ and vector $\mathbf{b} \in \mathbb{R}^n$. Then, for any $\mathbf{X} \in \mathbb{R}^{s\times d}$,

$$\|\mathbf{S}[\mathbf{AX}, \mathbf{b}]\|_{p,2}^p = \|[\mathbf{AX}, \mathbf{b}]\|_{p,2}^p \pm O(\varepsilon)\Big[\|\mathbf{AX}\|_{p,2}^p + R^p + \lambda^{p/2}s^{p/2}\|\mathbf{X}\|_2^p\Big]$$

To prove Lemma 14.4.2, we will need the following lemma on the matrix operator norm of a Gaussian matrix.

Lemma 14.4.3. Let $p > 0$ and let $m \geq O(d^{p/2})$. Let $\mathbf{G} \in \mathbb{R}^{m\times d}$ be an i.i.d. standard Gaussian matrix. Then, with probability at least $2/3$, we have that

$$\sup_{\mathbf{X}\in\mathbb{R}^{s\times d}, \|\mathbf{X}\|_2\leq 1} \left\|\mathbf{GX}^\top\right\|_{p,2}^p \leq O(s^{p/2}m).$$

Proof. Let $\mathbf{X} \in \mathbb{R}^{s\times d}$ with $\|\mathbf{X}\|_2 \leq 1$ maximize $\left\|\mathbf{GX}^\top\right\|_{p,2}^p$. Now let $\mathbf{g} \in \mathbb{R}^s$ be a random Gaussian vector and consider the vector $\mathbf{GX}^\top\mathbf{g}$. Then for each $i \in [m]$, $\mathbf{e}_i^\top\mathbf{GX}^\top\mathbf{g}$ is distributed as a Gaussian random variable with variance $\|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2$ and thus $|\mathbf{e}_i^\top\mathbf{GX}^\top\mathbf{g}| \geq \|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2/10$ with probability at least $9/10$. Then,

$$\mathbf{E}\left[\sum_{i=1}^m \|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2^p \cdot \mathbb{1}\left\{|\mathbf{e}_i^\top\mathbf{GX}^\top\mathbf{g}| \leq \|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2/10\right\}\right] \leq \frac{1}{10}\left\|\mathbf{GX}^\top\right\|_{p,2}^p$$

so by Markov's inequality, this at most $\left\|\mathbf{GX}^\top\right\|_{p,2}^p/2$ with probability at least $4/5$. Then, under this event,

$$\begin{aligned}
\left\|\mathbf{GX}^\top\mathbf{g}\right\|_p^p &\geq \sum_{i=1}^m |\mathbf{e}_i^\top\mathbf{GX}^\top\mathbf{g}|^p \cdot \mathbb{1}\left\{|\mathbf{e}_i^\top\mathbf{GX}^\top\mathbf{g}| > \|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2/10\right\} \\
&\geq \sum_{i=1}^m \frac{\|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2^p}{10^p} \cdot \mathbb{1}\left\{|\mathbf{e}_i^\top\mathbf{GX}^\top\mathbf{g}| > \|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2/10\right\} \\
&\geq \sum_{i=1}^m \frac{\|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2^p}{10^p} \cdot (1 - \mathbb{1}\left\{|\mathbf{e}_i^\top\mathbf{GX}^\top\mathbf{g}| \leq \|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2/10\right\})
\end{aligned}$$

so

$$\begin{aligned}
10^p\left\|\mathbf{GX}^\top\mathbf{g}\right\|_p^p &\geq \left\|\mathbf{GX}^\top\right\|_{p,2}^p - \sum_{i=1}^m \|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2^p \cdot \mathbb{1}\left\{|\mathbf{e}_i^\top\mathbf{GX}^\top\mathbf{g}| \leq \|\mathbf{e}_i^\top\mathbf{GX}^\top\|_2/10\right\} \\
&\geq \left\|\mathbf{GX}^\top\right\|_{p,2}^p - \left\|\mathbf{GX}^\top\right\|_{p,2}^p/2 = \left\|\mathbf{GX}^\top\right\|_{p,2}^p/2.
\end{aligned}$$

Thus, with probability at least $4/5$, we have that

$$\begin{aligned}
\sup_{\mathbf{X}\in\mathbb{R}^{s\times d}, \|\mathbf{X}\|_2\leq 1} \left\|\mathbf{GX}^\top\right\|_{p,2}^p &\leq O(1) \sup_{\mathbf{X}\in\mathbb{R}^{s\times d}, \|\mathbf{X}\|_2\leq 1} \sup_{\mathbf{v}\in\mathbb{R}^s, \|\mathbf{v}\|_2\leq\sqrt{s}} \left\|\mathbf{GX}^\top\mathbf{v}\right\|_p^p \\
&\leq O(s^{p/2}) \sup_{\mathbf{v}\in\mathbb{R}^d, \|\mathbf{v}\|_2\leq 1} \|\mathbf{Gv}\|_p^p.
\end{aligned}$$

The latter quantity is at most $O(s^{p/2}m)$ by Dvoretzky's theorem (Theorem 14.3.1) with probability at least $99/100$, so we conclude. □

We may then return to the proof of Lemma 14.4.2:

Proof of Lemma 14.4.2. Let m be a large enough number to be chosen, and let $\mathbf{G} \in \mathbb{R}^{m \times d}$ and $\mathbf{g} \in \mathbb{R}^m$ be drawn with standard Gaussian entries. If $m \geq O(\max\{\varepsilon^{-2}s, \varepsilon^{-1}s^{p/2}\})$, then

$$\begin{aligned}
\|[\mathbf{AX},\ \mathbf{b}]\|_{p,2}^p &= (1 \pm \varepsilon)\frac{1}{m}\left\|[\mathbf{AX},\ \mathbf{b}]\begin{pmatrix}\mathbf{G}^\top \\ \mathbf{g}^\top\end{pmatrix}\right\|_{p,p}^p && \text{Theorem 14.3.1} \\
&= (1 \pm \varepsilon)\frac{1}{m}\left\|\mathbf{AXG}^\top + \mathbf{bg}^\top\right\|_{p,p}^p \\
&= (1 \pm \varepsilon)\frac{1}{m}\sum_{j=1}^m \left\|\mathbf{AXG}^\top\mathbf{e}_j + \mathbf{bg}^\top\mathbf{e}_j\right\|_p^p
\end{aligned}$$

Now using the additive-multiplicative ℓ_p affine embedding property, we have that

$$\begin{aligned}
\left\|\mathbf{S}(\mathbf{AXG}^\top\mathbf{e}_j + \mathbf{bg}^\top\mathbf{e}_j)\right\|_p^p = (1 \pm \varepsilon)\left\|\mathbf{AXG}^\top\mathbf{e}_j + \mathbf{bg}^\top\mathbf{e}_j\right\|_p^p \pm \\
\varepsilon(\left\|\mathbf{AXG}^\top\mathbf{e}_j\right\|_p^p + R^p + \lambda^{p/2}\left\|\mathbf{XG}^\top\mathbf{e}_j\right\|_2^p)
\end{aligned}$$

Note then that the total error is

$$\begin{aligned}
\frac{\varepsilon}{m}\sum_{j=1}^m \left\|\mathbf{AXG}^\top\mathbf{e}_j\right\|_p^p + R^p + \lambda^{p/2}\left\|\mathbf{XG}^\top\mathbf{e}_j\right\|_2^p &= \frac{\varepsilon}{m}\left[\left\|\mathbf{AXG}^\top\right\|_{p,p}^p + mR^p + \lambda^{p/2}\left\|\mathbf{GX}^\top\right\|_{p,2}^p\right] \\
&= O(\varepsilon)\left[\|\mathbf{AX}\|_{p,2}^p + R^p + \frac{\lambda^{p/2}}{m}\left\|\mathbf{GX}^\top\right\|_{p,2}^p\right]
\end{aligned}$$

where we have again used Dvoretzky's theorem (Theorem 14.3.1). Now if $m \geq O(d^{p/2})$, then by Lemma 14.4.3, we have with constant probability that

$$\frac{\lambda^{p/2}}{m}\left\|\mathbf{GX}^\top\right\|_{p,2}^p \leq \frac{\lambda^{p/2}}{m}O(s^{p/2}m)\|\mathbf{X}\|_2^p = O(\lambda^{p/2}s^{p/2})\|\mathbf{X}\|_2^p. \qquad \square$$

14.5 Main sampling theorems

14.5.1 Affine embedding

We first show an affine embedding guarantee for root ridge leverage score sampling, which will be used to apply Lemma 14.4.2. The main workhorse behind this lemma is Theorem 8.1.1, which establishes a general ℓ_p affine embedding theorem for root ridge leverage score sampling, and generalizes recent work of [WY23c] by handling the case of $p > 2$ as well as allowing for an affine translation rather than just subspaces.

Lemma 14.5.1. Let $1 \leq p < \infty$. Let $\alpha = \Theta(\varepsilon^2)/((\log n)^3 + \log(1/\delta))$. Let $\mathbf{S}$ be the ℓ_p sampling matrix with probabilities $\{q_i\}_{i=1}^n$ for

$$q_i \geq \begin{cases} \min\{1, n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}/\alpha\} & \text{if } p > 2 \\ \min\{1, \boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}/\alpha\} & \text{if } p < 2 \end{cases}$$

with $\lambda = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$. Let S be an s-dimensional subspace for some $s \leq n$ such that $\|\mathbf{b}_S\|_p^p \leq \mathsf{OPT}$ where $\mathbf{b}_S(i) = \|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}_S)\|_2$, and let $\mathbf{P}_S = \mathbf{V}_S\mathbf{V}_S^\top$ be the orthogonal projection matrix onto S. Let $\mathbf{U} \in \mathbb{R}^{2n\times s}$ satisfy $\mathbf{U}^\top\mathbf{U} \succeq \frac{1}{C}\mathbf{I}$ for some constant $C = O(1)$. Then, with probability at least $1-\delta$, we have simultaneously for every $\mathbf{x} \in \mathbb{R}^s$ that

$$\|\mathbf{S}[\mathbf{A}\mathbf{V}_S\mathbf{x} + \mathbf{b}_S]\|_p^p = \|\mathbf{A}\mathbf{V}_S\mathbf{x} + \mathbf{b}_S\|_p^p \pm \varepsilon\Big(\|\mathbf{A}\mathbf{V}_S\mathbf{x}\|_p^p + \mathsf{OPT} + \lambda^{p/2}\|\mathbf{U}\mathbf{x}\|_p^p\Big).$$

Proof. We have that

$$\begin{aligned}
\boldsymbol{\tau}_i^\lambda(\mathbf{A}) &= \sup_{\mathbf{x}\in\mathbb{R}^d} \frac{[\mathbf{A}\mathbf{x}](i)^2}{\|\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2} \\
&\geq \sup_{\mathbf{x}=\mathbf{V}_S\mathbf{z},\mathbf{z}\in\mathbb{R}^s} \frac{[\mathbf{A}\mathbf{x}](i)^2}{\|\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2} = \sup_{\mathbf{z}\in\mathbb{R}^s} \frac{[\mathbf{A}\mathbf{V}_S\mathbf{z}](i)^2}{\|\mathbf{A}\mathbf{V}_S\mathbf{z}\|_2^2 + \lambda\|\mathbf{V}_S\mathbf{z}\|_2^2} = \sup_{\mathbf{z}\in\mathbb{R}^s} \frac{[\mathbf{A}\mathbf{V}_S\mathbf{z}](i)^2}{\|\mathbf{A}\mathbf{V}_S\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_2^2} \\
&\geq \sup_{\mathbf{z}\in\mathbb{R}^s} \frac{[\mathbf{A}\mathbf{V}_S\mathbf{z}](i)^2}{\|\mathbf{A}\mathbf{V}_S\mathbf{z}\|_2^2 + C\lambda\|\mathbf{U}\mathbf{z}\|_2^2}
\end{aligned}$$

so $\boldsymbol{\tau}_i^\lambda(\mathbf{A})$ upper bounds the i-th *leverage score* of the $3n \times s$ matrix given by

$$\mathbf{A}' \coloneqq \begin{pmatrix} \mathbf{A}\mathbf{V}_S \\ \sqrt{C\lambda}\mathbf{U} \end{pmatrix}$$

Now note that

$$\frac{|\mathbf{y}(i)|^p}{\|\mathbf{y}\|_p^p} \leq n^{p/2-1}\frac{|\mathbf{y}(i)|^p}{\|\mathbf{y}\|_2^p} = n^{p/2-1}\left(\frac{|\mathbf{y}(i)|^2}{\|\mathbf{y}\|_2^2}\right)^{p/2}. \tag{14.12}$$

Then by Lemma 14.3.6 and (14.12), we have that $\min\{1, n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}\}$ upper bounds the rank k ℓ_p subspace approximation sensitivities for $p > 2$. Similarly, $\min\{1, \boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}\}$ upper bounds the rank k ℓ_p subspace approximation sensitivities for $p < 2$. Thus,

$$\frac{|\mathbf{b}_S(i)|^p}{\mathsf{OPT}} \leq \frac{\left\|\mathbf{a}_i^\top(\mathbf{I} - \mathbf{P}^*)\right\|_2^p}{\|\mathbf{A}(\mathbf{I} - \mathbf{P}^*)\|_{p,2}^p} \leq \boldsymbol{\tau}_i^\lambda(\mathbf{A}).$$

We then define $\mathbf{b}' = [\mathbf{b}_S; 0] \in \mathbb{R}^{3n}$ to be the vector $\mathbf{b}_S$ with $2n$ zeros appended to it. Finally, let $\mathbf{S}' \in \mathbb{R}^{3n\times 3n}$ be the ℓ_p sampling matrix which samples the first n rows according to $\mathbf{S}$ and the last n rows with probability 1. Then by Theorem 8.1.1, we have the ℓ_p affine embedding guarantee for $\mathbf{A}'$ and thus with probability at least $1-\delta$, simultaneously for every $\mathbf{x} \in \mathbb{R}^s$, we have

$$\begin{aligned}
\|\mathbf{S}'[\mathbf{A}'\mathbf{x} + \mathbf{b}']\|_p^p &= \|\mathbf{S}[\mathbf{A}\mathbf{V}_S\mathbf{x} + \mathbf{b}_S]\|_p^p + (C\lambda)^{p/2}\|\mathbf{U}\mathbf{x}\|_p^p \\
&= (1\pm\varepsilon)\|\mathbf{A}'\mathbf{x} + \mathbf{b}'\|_p^p \pm \varepsilon\,\mathsf{OPT} \\
&= (1\pm\varepsilon)\left[\|\mathbf{A}\mathbf{V}_S\mathbf{x} + \mathbf{b}_S\|_p^p + (C\lambda)^{p/2}\|\mathbf{U}\mathbf{x}\|_p^p\right] \pm \varepsilon\,\mathsf{OPT}\,.
\end{aligned}$$

Now by subtracting $(C\lambda)^{p/2}\|\mathbf{U}\mathbf{x}\|_p^p$ from both sides of the inequality, we conclude that

$$\|\mathbf{S}[\mathbf{A}\mathbf{V}_S\mathbf{x} + \mathbf{b}_S]\|_p^p = (1\pm\varepsilon)\|\mathbf{A}\mathbf{V}_S\mathbf{x} + \mathbf{b}_S\|_p^p \pm \varepsilon(C\lambda)^{p/2}\|\mathbf{U}\mathbf{x}\|_p^p \pm \varepsilon\,\mathsf{OPT}\,.$$

Scaling ε by constant factors yields the claimed result. □

Next, we convert the affine embedding guarantee in Lemma 14.5.1 into a guarantee about preserving the norms of matrices under $\mathbf{S}$ via Dvoretzky's theorem.

Lemma 14.5.2. Let $1 \le p < \infty$. Let $\alpha = \Theta(\varepsilon^2)/((\log n)^3 + \log(1/\delta))$. Let $\mathbf{S}$ be the ℓ_p sampling matrix with probabilities $\{q_i\}_{i=1}^n$ for

$$q_i \ge \begin{cases} \min\{1, n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}/\alpha\} & \text{if } p > 2 \\ \min\{1, \boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}/\alpha\} & \text{if } p < 2 \end{cases}$$

with $\lambda = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$. Let $\mathbf{U} \in \mathbb{R}^{2n \times d}$ satisfy $\mathbf{U}^\top\mathbf{U} \succeq \mathbf{I}/C$ for some $C = O(1)$. Then for all matrices $\mathbf{X} \in \mathbb{R}^{d \times d}$,

$$\|\mathbf{SAX}\|_{p,2}^p = (1 \pm \varepsilon)\|\mathbf{AX}\|_{p,2}^p \pm \varepsilon\lambda^{p/2}\left\|\mathbf{UV}_R^\top\mathbf{X}\right\|_{p,2}^p$$

where $R = \mathrm{rowspan}(\mathbf{A})$.

Proof. Note that for an i.i.d. standard Gaussian matrix $\mathbf{H} \in \mathbb{R}^{d \times m}$ for m sufficiently large, we have by Dvoretzky's theorem (Theorem 14.3.1) that

$$\|\mathbf{SAX}\|_{p,2}^p = (1 \pm \varepsilon)\frac{1}{m}\|\mathbf{SAXH}\|_{p,p}^p = (1 \pm \varepsilon)\frac{1}{m}\sum_{j=1}^m \|\mathbf{SAXHe}_j\|_p^p \tag{14.13}$$

We now apply Lemma 14.5.1 with the subspace S set to be the row span R of $\mathbf{A}$ which has dimension at most n, so that we have the following additive-multiplicative subspace embedding guarantee for every $\mathbf{x} \in \mathbb{R}^d$, with probability at least $1 - \delta$,

$$\|\mathbf{Sax}\|_p^p = \|\mathbf{SAP}_R\mathbf{x}\|_p^p = \|\mathbf{AP}_R\mathbf{x}\|_p^p \pm \varepsilon\left(\|\mathbf{AP}_R\mathbf{x}\|_p^p + (C\lambda)^{p/2}\left\|\mathbf{UV}_R^\top\mathbf{x}\right\|_p^p\right) \tag{14.14}$$

Applying this guarantee to each summand in (14.13) shows that

$$\begin{aligned} \frac{1}{m}\sum_{j=1}^m \|\mathbf{SAXHe}_j\|_p^p &= \frac{1}{m}\sum_{j=1}^m (1 \pm \varepsilon)\|\mathbf{AXHe}_j\|_p^p \pm \varepsilon(C\lambda)^{p/2}\left\|\mathbf{UV}_R^\top\mathbf{XHe}_j\right\|_p^p \\ &= (1 \pm \varepsilon)\frac{1}{m}\|\mathbf{AXH}\|_{p,p}^p \pm \varepsilon\frac{1}{m}(C\lambda)^{p/2}\left\|\mathbf{UV}_R^\top\mathbf{XH}\right\|_{p,p}^p \\ &= (1 \pm \varepsilon)\|\mathbf{AX}\|_{p,2}^p \pm \varepsilon(C\lambda)^{p/2}\left\|\mathbf{UV}_R^\top\mathbf{X}\right\|_{p,2}^p. \end{aligned}$$

Rescaling ε by constant factors yields the desired result. □

14.5.2 Results for $p > 2$

For $p > 2$, we will first give the following refinement of Lemma 14.5.2 for low rank matrices.

Lemma 14.5.3. Let $2 \le p < \infty$. Let $\alpha = \Theta(\varepsilon^2)/((\log n)^3 + \log(1/\delta))$. Let $\mathbf{S}$ be the ℓ_p sampling matrix with probabilities $\{q_i\}_{i=1}^n$ for

$$q_i \ge \min\{1, n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}/\alpha\}$$

with $\lambda = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$. Then for all rank s matrices $\mathbf{X} \in \mathbb{R}^{d \times d}$,

$$\|\mathbf{SAX}\|_{p,2}^p = (1 \pm \varepsilon)\|\mathbf{AX}\|_{p,2}^p \pm \varepsilon\lambda^{p/2}n^{1-p/2}s^{p/2}\|\mathbf{X}\|_2^p.$$

Proof. Let $\mathbf{X} \in \mathbb{R}^{d\times d}$ be a fixed rank s matrix (depending on $\mathbf{SA}$ and $\mathbf{A}$) that maximizes

$$\left|\|\mathbf{SAX}\|_{p,2}^p - \|\mathbf{AX}\|_{p,2}^p\right|$$

over all rank s matrices $\mathbf{X} \in \mathbb{R}^{d\times d}$ such that $\|\mathbf{AX}\|_{p,2}^p + \lambda^{p/2}n^{1-p/2}s^{p/2}\|\mathbf{X}\|_2^p \le 1$. Note that we may WLOG assume that $\mathbf{X} = \mathbf{P}_R\mathbf{X} = \mathbf{V}_R\mathbf{V}_R^\top\mathbf{X}$ where R is the row span of $\mathbf{A}$, since any component outside of the row span of $\mathbf{X}$ will vanish after multiplying by $\mathbf{A}$.

Let $\mathbf{G} \in \mathbb{R}^{2n\times\dim(R)}$ be a random standard Gaussian matrix. It is well-known that $\mathbf{I}/C \preceq \frac{1}{2n}\mathbf{G}^\top\mathbf{G}$ with probability at least $2/3$ [RV09]. Furthermore, by Lemma 14.4.3, we have that

$$\Pr_{\mathbf{G}}\left\{\left\|\mathbf{GV}_R^\top\mathbf{X}\right\|_{p,2}^p = O(ns^{p/2})\|\mathbf{X}\|_2^p\right\} \ge \frac{2}{3}. \tag{14.15}$$

Thus by a union bound, $\mathbf{G}$ satisfies both of these events with probability at least $1/3$.

Now consider $t = O(\log(1/\delta))$ independent drawings of $\mathbf{G}$, say $\mathbf{G}^{(i)}$ for $i \in [t]$. For each $\mathbf{G}^{(i)}$, $\mathbf{S}$ has a $1-\delta/2t$ probability of succeeding in the guarantee of Lemma 14.5.2, if $\mathbf{G}^{(i)}$ satisfies the condition that $\frac{1}{2n}(\mathbf{G}^{(i)})^\top\mathbf{G}^{(i)} \succeq \mathbf{I}/C$. By a union bound, this holds for all $i \in [t]$ simultaneously with probability at least $1-\delta/2$. Furthermore, with probability at least $1-\delta/2$, there is at least one $i \in [t]$ such that (14.15) and $\frac{1}{2n}(\mathbf{G}^{(i)})^\top\mathbf{G}^{(i)} \succeq \mathbf{I}/C$ holds. Thus over all, with probability at least $1-\delta$, there is a matrix $\mathbf{G}$ such that for $\mathbf{U} = \mathbf{G}/\sqrt{2n}$, we have

$$\begin{aligned}
\|\mathbf{SAX}\|_{p,2}^p &= (1\pm\varepsilon)\|\mathbf{AX}\|_{p,2}^p \pm \varepsilon\lambda^{p/2}\left\|\mathbf{UV}_R^\top\mathbf{X}\right\|_{p,2}^p \\
&= (1\pm\varepsilon)\|\mathbf{AX}\|_{p,2}^p \pm \varepsilon\lambda^{p/2}n^{-p/2}\left\|\mathbf{GV}_R^\top\mathbf{X}\right\|_{p,2}^p \\
&= (1\pm\varepsilon)\|\mathbf{AX}\|_{p,2}^p \pm O(\varepsilon)\lambda^{p/2}n^{1-p/2}s^{p/2}\|\mathbf{X}\|_2^p.
\end{aligned}$$

Rescaling ε by constant factors yields the desired result. □

The next theorem gives an error bound after one round of root ridge leverage score sampling.

Theorem 14.5.4. Let $p > 2$ and let $s = O(k/\varepsilon^p)$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ with $n \ge n'$ for some $n' = O(s^{p/2}/\varepsilon)$. Let $\alpha = \Theta(\varepsilon^2)/((\log n)^3 + \log(1/\delta))$. Let $\mathbf{S}$ be the ℓ_p sampling matrix with probabilities $\{q_i\}_{i=1}^n$ for

$$q_i \ge \min\left\{1, n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}/\alpha\right\}$$

with $\lambda = \|\mathbf{A}-\mathbf{A}_s\|_F^2/s$. Then, with probability at least $1-\delta$, for every $F \in \mathcal{F}_k$,

$$\|\mathbf{SA}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p = (1\pm\varepsilon)\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p.$$

Proof. First note that

$$\lambda^{p/2} = \frac{\|\mathbf{A}-\mathbf{A}_s\|_F^p}{s^{p/2}} \le \frac{\|\mathbf{A}(\mathbf{I}-\mathbf{P}^*)\|_F^p}{s^{p/2}} \le n^{p/2-1}\frac{\|\mathbf{A}(\mathbf{I}-\mathbf{P}^*)\|_{p,2}^p}{s^{p/2}} = \frac{n^{p/2-1}}{s^{p/2}}\,\mathsf{OPT} \tag{14.16}$$

By Theorem 14.2.1, we have that

$$\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p = (1\pm\varepsilon)\|[\mathbf{AP}_S(\mathbf{I}-\mathbf{P}_F),\ \mathbf{b}_S]\|_{p,2}^p$$

$$= (1 \pm \varepsilon)\left\|[\mathbf{A}\mathbf{V}_S\mathbf{V}_S^\top(\mathbf{I} - \mathbf{P}_F),\ \mathbf{b}_S]\right\|_{p,2}^p$$

Let $\mathbf{G} \in \mathbb{R}^{2n\times s}$ be a random Gaussian matrix. It is well-known that $\mathbf{I}/C \preceq \frac{1}{2n}\mathbf{G}^\top\mathbf{G}$ with probability at least $2/3$ [RV09]. Fix such a matrix $\mathbf{G}$. By Lemma 14.5.1 with $\mathbf{U} = \mathbf{G}/\sqrt{2n}$, we then have

$$\begin{aligned}
\|\mathbf{S}[\mathbf{A}\mathbf{V}_S\mathbf{x} + \mathbf{b}_S]\|_p^p &= \|\mathbf{A}\mathbf{V}_S\mathbf{x} + \mathbf{b}_S\|_p^p \pm \varepsilon\Big(\|\mathbf{A}\mathbf{V}_S\mathbf{x}\|_p^p + \mathsf{OPT} + \lambda^{p/2}\|\mathbf{U}\mathbf{x}\|_p^p\Big) \\
&= \|\mathbf{A}\mathbf{V}_S\mathbf{x} + \mathbf{b}_S\|_p^p \pm \varepsilon\Big(\|\mathbf{A}\mathbf{V}_S\mathbf{x}\|_p^p + \mathsf{OPT} + (2n)^{1-p/2}\lambda^{p/2}\|\mathbf{x}\|_2^p\Big) && \text{Dvoretzky's theorem} \\
&= \|\mathbf{A}\mathbf{V}_S\mathbf{x} + \mathbf{b}_S\|_p^p \pm O(\varepsilon)\Big(\|\mathbf{A}\mathbf{V}_S\mathbf{x}\|_p^p + \mathsf{OPT} + s^{-p/2}\,\mathsf{OPT}\|\mathbf{x}\|_2^p\Big) && (14.16)
\end{aligned}$$

where we have used that n is large enough to apply Dvoretzky's theorem. Then by Lemma 14.4.2, we have that

$$\|\mathbf{S}[\mathbf{A}\mathbf{V}_S\mathbf{X}, \mathbf{b}_S]\|_{p,2}^p = \|[\mathbf{A}\mathbf{V}_S\mathbf{X}, \mathbf{b}_S]\|_{p,2}^p \pm O(\varepsilon)\Big[\|\mathbf{A}\mathbf{V}_S\mathbf{X}\|_{p,2}^p + \mathsf{OPT} + \mathsf{OPT}\|\mathbf{X}\|_2^p\Big]$$

for any $\mathbf{X} \in \mathbb{R}^{s\times d}$. Then, applying this result with $\mathbf{X} = \mathbf{V}_S^\top(\mathbf{I} - \mathbf{P}_F)$, which has operator norm 1, gives

$$\begin{aligned}
\|\mathbf{S}[\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F), \mathbf{b}_S]\|_{p,2}^p &= \|[\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F), \mathbf{b}_S]\|_{p,2}^p \pm O(\varepsilon)\Big[\|\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p + \mathsf{OPT}\Big] \\
&= (1 \pm O(\varepsilon))\|[\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F), \mathbf{b}_S]\|_{p,2}^p
\end{aligned}$$

Finally, by (14.16) and the fact that $\mathbf{P}_S - \mathbf{P}_{S\cup F}$ is a matrix with rank at most k, we have

$$\begin{aligned}
\|\mathbf{S}\mathbf{A}(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_{p,2}^p &\lesssim \|\mathbf{A}(\mathbf{P}_S - \mathbf{P}_{S\cup F})\|_{p,2}^p + \lambda^{p/2}n^{1-p/2}k^{p/2} && \text{Lemma 14.5.3} \\
&\le \varepsilon^p \cdot \mathsf{OPT} + \lambda^{p/2}n^{1-p/2}k^{p/2} && \text{Theorem 14.2.1} \\
&\le \varepsilon^p \cdot \mathsf{OPT} + \varepsilon^{p^2/2} \cdot \mathsf{OPT} \le 2\varepsilon^p\,\mathsf{OPT} && (14.16)
\end{aligned}$$

Then by Theorem 14.2.1, it follows that

$$\|\mathbf{S}[\mathbf{A}\mathbf{P}_S(\mathbf{I} - \mathbf{P}_F), \mathbf{b}_S]\|_{p,2}^p = (1 \pm O(\varepsilon))\|\mathbf{S}\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_{p,2}^p$$

as claimed. Chaining together the previous bounds and rescaling ε by constant factors shows the claimed result. □

Finally, we show that by applying Theorem 14.5.4 recursively for $O(\log\log n)$ rounds, we obtain our desired sampling theorem.

Theorem 14.5.5. Let $p > 2$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. There is an algorithm that runs in time $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^\omega)$ time to construct a diagonal matrix $\mathbf{S}$ with

$$\mathsf{nnz}(\mathbf{S}) = \frac{O(k^{p/2})}{\varepsilon^{p^2/2+p}}\left[(\log n)^{3p/2} + (\log(1/\delta))^{p/2}\right](\log\log n)^{p^2/2+p} = \frac{k^{p/2}}{\varepsilon^{O(p^2)}}(\log(n/\delta))^{O(p)}$$

that satisfies Definition 14.1.1 with probability at least $1 - \delta$.

Proof. We will apply Theorem 14.5.4 for $r = O(\log \log n)$ rounds, with ε set to ε/r and δ set to δ/r. Let α and s be the values given by Theorem 14.5.4 with this setting of parameters. We first analyze the number of rows sampled at each round in expectation. Note first that if $n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A}) \geq 1$, then $\boldsymbol{\tau}_i^\lambda(\mathbf{A}) \geq n^{2/p-1}$ so there are at most $O(s)n^{1-2/p}$ such rows $i \in [n]$, since $\boldsymbol{\tau}_i^\lambda(\mathbf{A})$ sum to at most $2s$ by Lemma 14.3.5. On the other hand, if $n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}$, then $\boldsymbol{\tau}_i^\lambda(\mathbf{A}) \leq n^{2/p-1}$ so we have that

$$\sum_{i:n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}\leq 1} n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2} \leq n^{p/2-1}(n^{2/p-1})^{p/2-1} \sum_{i:n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}\leq 1} \boldsymbol{\tau}_i^\lambda(\mathbf{A}) \leq n^{1-2/p}\cdot 2s.$$

Thus, in either case, we have

$$\sum_{i=1}^{n} \min\{1, n^{p/2-1}\boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}\} \leq O(s)n^{1-2/p}.$$

Thus, the expected number of sampled rows is at most $O(s)n^{1-2/p}/\alpha$. By Chernoff bounds, if the expected number of sampled rows is at least $O(\log(r/\delta))$, then with probability at least $1-\delta/r$, the number of sampled rows is within a constant factor of the expectation. Then by a union bound, for the first r rounds of the recursive calls, we succeed in obtaining a $(1 \pm \varepsilon/r)$ approximation and reduce the number of rows from m to $O(s)m^{1-2/p}/\alpha$. We now define a_i to be the logarithm of the number of rows after the ith recursive call. Then,

$$a_{i+1} = (1-2/p)a_i + \log(O(s)/\alpha)$$

so by Lemma 8.1.9, we have that

$$a_r = \frac{p}{2}\big(\log(O(s)/\alpha) - (1-2/p)^i(\log(O(s)/\alpha) - (2/p))(\log n)\big)$$

so the number of rows is at most

$$\exp(a_r) = \left(\frac{O(s)}{\alpha}\right)^{p/2} = \frac{O(s^{p/2})}{\alpha^{p/2}} = \frac{O(k^{p/2})}{\varepsilon^{p^2/2+p}}\left[(\log n)^{3p/2} + (\log(1/\delta))^{p/2}\right](\log\log n)^{p^2/2+p}.$$

□

14.5.3 Results for $p < 2$

The next theorem gives an error bound after one round of root ridge leverage score sampling.

Theorem 14.5.6. Let $p < 2$ and let $s = O(k/\varepsilon^2)$. Let $\alpha = \Theta(\varepsilon^2)/((\log n)^3 + \log(1/\delta))$. Let $\mathbf{S}$ be the ℓ_p sampling matrix with probabilities $\{q_i\}_{i=1}^n$ for

$$q_i \geq \min\big\{1, \boldsymbol{\tau}_i^\lambda(\mathbf{A})^{p/2}/\alpha\big\}$$

with $\lambda = \|\mathbf{A}-\mathbf{A}_s\|_F^2/s$ for s at least $O(k/\varepsilon^2)$ as required by Theorem 14.2.1. Furthermore, suppose that there is a rank s subspace $\tilde{F}$ such that

$$\|\mathbf{a}_i^\top(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_2^p \leq O(1/n)\|\mathbf{A}(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_{p,2}^p$$

and

$$\|\mathbf{A}(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_{p,2}^p \le O(1)\min_{F\in\mathcal{F}_k}\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p = O(\mathsf{OPT}).$$

Then, with probability at least $1-\delta$, for every $F\in\mathcal{F}_k$,

$$\|\mathbf{SA}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p = (1\pm\varepsilon)\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p.$$

Proof. First note that

$$\lambda^{p/2} \le \frac{\|\mathbf{A}-\mathbf{A}_s\|_F^p}{s^{p/2}} \le \frac{\|\mathbf{A}(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_F^p}{s^{p/2}} \le O(1/n)^{1-p/2}\frac{\|\mathbf{A}(\mathbf{I}-\mathbf{P}_{\tilde{F}})\|_{p,2}^p}{s^{p/2}} = \frac{O(\mathsf{OPT})}{s^{p/2}n^{1-p/2}} \tag{14.17}$$

By Theorem 14.2.1, we have that

$$\begin{aligned}\|\mathbf{A}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p &= (1\pm\varepsilon)\|[\mathbf{AP}_S(\mathbf{I}-\mathbf{P}_F),\ \mathbf{b}_S]\|_{p,2}^p \\ &= (1\pm\varepsilon)\left\|[\mathbf{AV}_S\mathbf{V}_S^\top(\mathbf{I}-\mathbf{P}_F),\ \mathbf{b}_S]\right\|_{p,2}^p\end{aligned}$$

By Lemma 14.5.1 with $\mathbf{U}$ set to the identity padded with zeros,

$$\begin{aligned}\|\mathbf{S}[\mathbf{AV}_S\mathbf{x}+\mathbf{b}_S]\|_p^p &= \|\mathbf{AV}_S\mathbf{x}+\mathbf{b}_S\|_p^p \pm \varepsilon\Big(\|\mathbf{AV}_S\mathbf{x}\|_p^p + \mathsf{OPT} + \lambda^{p/2}\|\mathbf{x}\|_p^p\Big) \\ &= \|\mathbf{AV}_S\mathbf{x}+\mathbf{b}_S\|_p^p \pm \varepsilon\Big(\|\mathbf{AV}_S\mathbf{x}\|_p^p + \mathsf{OPT} + n^{1-p/2}\lambda^{p/2}\|\mathbf{x}\|_2^p\Big) \\ &= \|\mathbf{AV}_S\mathbf{x}+\mathbf{b}_S\|_p^p \pm \varepsilon\Big(\|\mathbf{AV}_S\mathbf{x}\|_p^p + \mathsf{OPT} + O(s^{-p/2})\,\mathsf{OPT}\|\mathbf{x}\|_2^p\Big) \quad (14.17)\end{aligned}$$

Then by Lemma 14.4.2, we have that

$$\|\mathbf{S}[\mathbf{AV}_S\mathbf{X},\mathbf{b}_S]\|_{p,2}^p = \|[\mathbf{AV}_S\mathbf{X},\mathbf{b}_S]\|_{p,2}^p \pm O(\varepsilon)\Big[\|\mathbf{AV}_S\mathbf{X}\|_{p,2}^p + \mathsf{OPT} + \mathsf{OPT}\|\mathbf{X}\|_2^p\Big]$$

for any $\mathbf{X}\in\mathbb{R}^{s\times d}$. Then, applying this result with $\mathbf{X}=\mathbf{V}_S^\top(\mathbf{I}-\mathbf{P}_F)$, which has operator norm 1, gives

$$\begin{aligned}\|\mathbf{S}[\mathbf{AP}_S(\mathbf{I}-\mathbf{P}_F),\mathbf{b}_S]\|_{p,2}^p &= \|[\mathbf{AP}_S(\mathbf{I}-\mathbf{P}_F),\mathbf{b}_S]\|_{p,2}^p \pm O(\varepsilon)\Big[\|\mathbf{AP}_S(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p + \mathsf{OPT}\Big] \\ &= (1\pm O(\varepsilon))\|[\mathbf{AP}_S(\mathbf{I}-\mathbf{P}_F),\mathbf{b}_S]\|_{p,2}^p\end{aligned}$$

Finally, by using Lemma 14.5.2 with $\mathbf{U}=\mathbf{I}$, (14.17), and the fact that $\mathbf{P}_S-\mathbf{P}_{S\cup F}$ is a matrix with rank at most k, we have

$$\begin{aligned}\|\mathbf{SA}(\mathbf{P}_S-\mathbf{P}_{S\cup F})\|_{p,2}^p &\lesssim \|\mathbf{A}(\mathbf{P}_S-\mathbf{P}_{S\cup F})\|_{p,2}^p + \lambda^{p/2}\left\|\mathbf{V}_R^\top(\mathbf{P}_S-\mathbf{P}_{S\cup F})\right\|_{p,2}^p && \text{Lemma 14.5.2} \\ &\le \|\mathbf{A}(\mathbf{P}_S-\mathbf{P}_{S\cup F})\|_{p,2}^p + \lambda^{p/2}n^{1-p/2}\left\|\mathbf{V}_R^\top(\mathbf{P}_S-\mathbf{P}_{S\cup F})\right\|_{2,2}^p && \\ &\le \varepsilon^p\cdot\mathsf{OPT} + \lambda^{p/2}n^{1-p/2}k^{p/2} && \text{Theorem 14.2.1} \\ &\le \varepsilon^p\cdot\mathsf{OPT} + O(\varepsilon^p)\cdot\mathsf{OPT} = O(\varepsilon^p)\,\mathsf{OPT} && (14.17)\end{aligned}$$

Then by Theorem 14.2.1, it follows that

$$\|\mathbf{S}[\mathbf{AP}_S(\mathbf{I}-\mathbf{P}_F),\mathbf{b}_S]\|_{p,2}^p = (1\pm O(\varepsilon))\|\mathbf{SA}(\mathbf{I}-\mathbf{P}_F)\|_{p,2}^p$$

as claimed. Chaining together the previous bounds and rescaling ε by constant factors shows the claimed result. □

Finally, we show that by applying Theorem 14.5.6 recursively for $O(\log\log n)$ rounds, we arrive at our main theorem for $p < 2$.

Theorem 14.5.7. Let $p < 2$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. There is an algorithm that runs in time $\tilde{O}(\mathsf{nnz}(\mathbf{A}) + d^{\omega})$ time to construct a diagonal matrix $\mathbf{S}$ with

$$\mathsf{nnz}(\mathbf{S}) = \frac{O(k)}{\varepsilon^{4/p+2}}\left[(\log n)^{6/p+1} + (\log(1/\delta))^{2/p+1}\right](\log\log n)^{4/p+2} = \frac{k}{\varepsilon^{O(1)}}(\log(n/\delta))^{O(1)}$$

that satisfies Definition 14.1.1 with probability at least $1 - \delta$.

Proof. We will apply Theorem 14.5.6 for $r = O(\log\log n)$ rounds, with ε set to ε/r and δ set to δ/r. In order to satisfy the precondition of the existence of a "flat" solution, we quickly compute a constant factor solution via Lemma 14.3.2 and flatten via Lemma 14.3.3, so that our condition is satisfied as long as we take s to be at least some $O(k\log(n/\delta))$.

Let α and $s = O(k/\varepsilon^2 + k\log(n/\delta))$ be the values given by Theorem 14.5.6 with this setting of parameters. We first analyze the number of rows sampled at each round in expectation. Since the ridge leverage scores sum to at most $2s'$ by Lemma 14.3.5, we have that

$$\sum_{i=1}^{n} \boldsymbol{\tau}_i^{\lambda}(\mathbf{A})^{p/2} \le n^{1-p/2}\left(\sum_{i=1}^{n} \boldsymbol{\tau}_i^{\lambda}(\mathbf{A})\right)^{p/2} = O(s^{p/2}n^{1-p/2})$$

by Hölder's inequality. Thus, the expected number of sampled rows is at most $O(s^{p/2}n^{1-p/2})/\alpha$. By Chernoff bounds, if the expected number of sampled rows is at least $O(\log(r/\delta))$, then with probability at least $1 - \delta/r$, the number of sampled rows is within a constant factor of the expectation. Then by a union bound, for the first r rounds of the recursive calls, we succeed in obtaining a $(1 \pm \varepsilon/r)$ approximation and reduce the number of rows from m to $O(s^{p/2}m^{1-p/2})/\alpha$. We now define a_i to be the logarithm of the number of rows after the ith recursive call. Then,

$$a_{i+1} = (1 - p/2)a_i + \log(O(s^{p/2})/\alpha)$$

so by Lemma 8.1.9, we have that

$$a_r = \frac{2}{p}\big(\log(O(s^{p/2})/\alpha) - (1 - p/2)^i(\log(O(s^{p/2})/\alpha) - (p/2))(\log n)\big)$$

so the number of rows is at most

$$\exp(a_r) = \left(\frac{O(s^{p/2})}{\alpha}\right)^{2/p} = \frac{O(s)}{\alpha^{2/p}} = \frac{O(k)}{\varepsilon^{4/p+2}}\left[(\log n)^{6/p+1} + (\log(1/\delta))^{2/p+1}\right](\log\log n)^{4/p+2}.$$

□

14.6 Streaming and online coresets

We present our results on streaming and online coresets for ℓ_p subspace approximation.

14.6.1 Online coresets

In this section, we note that our Theorems 14.5.5 and 14.5.7 give the first nearly optimal online coresets (see Section 1.3.3) for ℓ_p subspace approximation.

Then, the following is an immediate corollary of Theorems 14.5.5 and 14.5.7.

Corollary 14.6.1 (Online coresets)**.** Let $1 \leq p < \infty$. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ have online condition number κ^{OL}. Then, there is an online coreset algorithm which constructs a diagonal map $\mathbf{S} \in \mathbb{R}^{n \times n}$ satisfying Definition 14.1.1 with probability at least $1 - \delta$, such that

$$\mathsf{nnz}(\mathbf{S}) = \begin{cases} \frac{\tilde{O}(k^{p/2})}{\varepsilon^{p^2/2+p}} (\log(n\kappa^{\mathsf{OL}}/\delta))^{O(p)} \\ \frac{\tilde{O}(k)}{\varepsilon^{4/p+2}} (\log(n\kappa^{\mathsf{OL}}/\delta))^{O(1)} \end{cases}$$

while storing at most $O(k(\log k)(\log \kappa^{\mathsf{OL}})^2)$ additional rows in an online fashion.

Proof. The result of [BDM+20, Theorem 3.1] gives an online coreset algorithm for maintaining a $(1 \pm \varepsilon)$ strong coreset for ℓ_2 subspace approximation which stores $O(\varepsilon^{-2}k(\log k)(\log \kappa^{\mathsf{OL}})^2)$ rows. Furthermore, given such a strong coreset with $\varepsilon = O(1)$, it is shown in [BDM+20, Lemma 2.11] that one can obtain scores $\tilde{\boldsymbol{\tau}}_i$ such that

$$\tilde{\boldsymbol{\tau}}_i \geq \boldsymbol{\tau}_i^{\lambda}(\mathbf{A})$$

for $\lambda = \|\mathbf{A} - \mathbf{A}_k\|_F^2/k$, and also satisfies

$$\sum_{i=1}^{n} \tilde{\boldsymbol{\tau}}_i \leq O(k \log \kappa^{\mathsf{OL}}).$$

The result for $p > 2$ then follows as an immediate corollary of Theorem 14.5.5. For $p < 2$, we additionally need an online constant factor approximation to flatten the matrix, which is constructed in [WY23a] by obtaining online Lewis sample due to [WY23b]. The result for $p < 2$ then follows as an immediate corollary of Theorem 14.5.7. □

For integer matrices with entries bounded by Δ, we may replace the dependence on the online condition number with Δ, by using an analogous result of [BDM+20] for integer matrices.

Corollary 14.6.2 (Online coresets – integer matrices)**.** Let $1 \leq p < \infty$. Let $\mathbf{A} \in \mathbb{Z}^{n \times d}$ have entries bounded by $|\mathbf{A}_{i,j}| \leq \Delta$. Then, there is an online coreset algorithm which constructs a diagonal map $\mathbf{S} \in \mathbb{R}^{n \times n}$ satisfying Definition 14.1.1 with probability at least $1 - \delta$, such that

$$\mathsf{nnz}(\mathbf{S}) = \begin{cases} \frac{\tilde{O}(k^{p/2})}{\varepsilon^{p^2/2+p}} (\log(n\Delta/\delta))^{O(p)} \\ \frac{\tilde{O}(k)}{\varepsilon^{4/p+2}} (\log(n\Delta/\delta))^{O(1)} \end{cases}$$

while storing at most $O(k(\log \Delta)^2)$ additional rows in an online fashion.

14.6.2 Streaming coresets

Next, we state our corollaries for constructing streaming coresets in the row arrival model of streaming, which is slightly different from the online coreset model since we are allowed to remove rows from our coreset. In the streaming model, the resource measure is typically the space complexity, and thus the input is usually assumed to be an integer matrix as a bit complexity assumption. In this setting, we combine the classic merge-and-reduce technique [BDM+20] with the technique of [CWZ23] of first applying an online coreset to obtain a result with only $\mathrm{poly}(\log\log(n\Delta)$ factor overhead in the coreset size.

Corollary 14.6.3 (Streaming coresets – integer matrices)**.** Let $1 \leq p < \infty$. Let $\mathbf{A} \in \mathbb{Z}^{n\times d}$ have entries bounded by $|\mathbf{A}_{i,j}| \leq \Delta$. Then, there is an row arrival streaming algorithm which constructs a diagonal map $\mathbf{S} \in \mathbb{R}^{n\times n}$ satisfying Definition 14.1.1 with probability at least $1-\delta$, such that

$$\mathsf{nnz}(\mathbf{S}) = \begin{cases} \frac{\tilde{O}(k^{p/2})}{\varepsilon^{p^2/2+p}}(\log(k/\varepsilon\delta) + \log\log(n\Delta/\delta))^{O(p^2)} \\ \frac{\tilde{O}(k)}{\varepsilon^{4/p+2}}(\log(k/\varepsilon\delta) + \log(n\Delta/\delta))^{O(1)} \end{cases}$$

while storing at most $O(k(\log\Delta)^2)$ additional rows in an online fashion.

Proof. We may assume without loss of generality that the stream length is at most $m = \mathrm{poly}(k, \varepsilon^{-1}, \log(n\Delta/\delta))$ by first applying Corollary 14.6.2. We then apply the merge-and-reduce technique, which results in a coreset with size where the ε dependence is replaced by $\varepsilon' = \varepsilon/\log m$. This results in the claimed bounds. □

Chapter 15

Future directions for sampling and coreset algorithms

In this chapter, we present several interesting open questions on sampling and coreset algorithms arising from this thesis that remain open.

15.1 Questions on ℓ_p subspace embeddings

Nearly optimal bounds for ℓ_p subspace embeddings for $p > 2$. One of the outstanding gaps in bounds for ℓ_p subspace embeddings is the optimality of the upper bound given in Theorem 6.1.4 and Theorem 6.1.11 in terms of the dependence on d and ε for $p > 2$. So far, the upper bound is $r = \tilde{O}(\varepsilon^{-2}d^{p/2})$ for a subspace embedding $\mathbf{S}$ with r rows, while the best known lower bound is still Theorem 6.1.5 due to [LWW21], which gives a lower bound of $r = \tilde{\Omega}(\varepsilon^{-1}d^{p/2} + \varepsilon^{-2}d)$. Thus, resolving this last gap from obtaining nearly optimal trade-offs between number of rows r, d, and the accuracy parameter ε is our first open question about ℓ_p subspace embeddings.

Question 15.1.1. For $p \in (2, \infty) \setminus 2\mathbb{Z}$, what is the smallest possible number of rows r that is possible for ℓ_p subspace embeddings with $(1 + \varepsilon)$ distortion? Is there a lower bound showing that $r = \Omega(\varepsilon^{-2}d^{p/2})$ rows is necessary?

Deterministic algorithms. For $p = 2$, the seminal work of [BSS12] showed that it is possible to deterministically obtain ℓ_2 subspace embeddings with $r = O(\varepsilon^{-2}d)$ rows in polynomial time, and has spurred multiple works further improving the running time of this algorithm [Zou12, ALO15]. This algorithm, however, makes heavy use of the special structure of the ℓ_2 norm, and does not yield results for ℓ_p subspace embeddings for $p \neq 2$. Thus, an interesting question is whether polynomial time algorithms for constructing ℓ_p subspace embeddings exist or not.

Question 15.1.2. Is there a deterministic polynomial time algorithm for constructing $(1 + \varepsilon)$-approximate ℓ_p subspace embeddings with $\tilde{O}(\varepsilon^{-2}d)$ rows for $p < 2$ or $\tilde{O}(\varepsilon^{-2}d^{p/2})$ rows for $p > 2$?

In fact, even a Las Vegas algorithm for computing ℓ_p subspace embeddings may be interesting, as there are currently no known efficient algorithms for checking whether two matrices are close

in the sense of ℓ_p subspace embeddings, for any $p \neq 2$:

Question 15.1.3. Is there a polynomial time Las Vegas algorithm for constructing $(1+\varepsilon)$-approximate ℓ_p subspace embeddings with $\tilde{O}(\varepsilon^{-2}d)$ rows for $p < 2$ or $\tilde{O}(\varepsilon^{-2}d^{p/2})$ rows for $p > 2$?

Removing logarithmic factors. A closely related problem to Question 15.1.2 is the question of removing logarithmic factors in the number of rows r. In particular, the work of [BSS12] as well as its various follow-ups [Zou12, ALO15, LS15] obtain $r = O(\varepsilon^{-2}d)$, without any logarithmic factor losses. On the other hand, for independent sampling-based approaches such as Lewis weight sampling, an extra logarithmic factor is inherent due to the coupon-collector problem. However, for most values of $p \neq 2$[1], no other approaches towards obtaining $(1+\varepsilon)$-approximate ℓ_p subspace embeddings are known. Thus, an important question is the following:

Question 15.1.4. Is there an algorithm for constructing $(1+\varepsilon)$-approximate ℓ_p subspace embeddings with $r = O(\varepsilon^{-2}d)$ rows for $p < 2$ and $r = O(\varepsilon^{-2}d^{p/2})$ rows for $p > 2$?

For $p = 1$, this problem has been raised in [Sch07, HRR22].

Nearly optimal guarantees for sensitivity sampling. We re-iterate our main open question, Question 7.1.1, from the work of [WY23c] from Chapter 7: what is the smallest sample complexity possible for the ℓ_p sensitivity sampling algorithm? While we have achieved the bounds of $\tilde{O}(\varepsilon^{-2}\mathfrak{S}^{2/p})$ for $p < 2$ and $\tilde{O}(\varepsilon^{-2}\mathfrak{S}^{2-2/p})$ for $p > 2$, we conjecture that a bound of $\tilde{O}(\varepsilon^{-2}(\mathfrak{S}+d))$ is possible.

Faster algorithms for approximating sensitivities. While ℓ_p sensitivity sampling can yield the lowest known row counts for matrices with low total sensitivity, the running time for computing sensitivity scores is still much slower than leverage score or ℓ_p Lewis weight computation for $p \neq 2$ [PWZ23]. Can sensitivity approximation algorithms be sped up to compete with the running time of approximating leverage scores?

15.2 Questions on coresets

Our main questions concerning coresets are those left open by our work of [WY24b], which established the first nearly optimal strong coresets for ℓ_p subspace approximation (see Chapter 14). The main natural direction left open is to tighten the dependence on ε in the coreset size both in the upper bounds and lower bounds. Currently, the best known lower bound on the number of rows required is $\tilde{\Omega}(k/\varepsilon^2)$ for $p < 2$ and $\tilde{\Omega}(k/\varepsilon^2 + k^{p/2}/\varepsilon)$ for $p > 2$ via a reduction to lower bounds for ℓ_p subspace embeddings [LWW21, WY23a], while we have a dependence of $\varepsilon^{-O(p^2)}$ in our upper bounds.

Question 15.2.1. How many rows are necessary and sufficient for strong coresets for ℓ_p subspace approximation as a function of both k and ε?

[1] An important exception is $p \in 2\mathbb{Z}$, which admit exact isometries via other methods due to its special structure [Sch11].

In particular, we believe that the following special case is already an interesting question:

Question 15.2.2. Is there a $\varepsilon^{-\Omega(p)}$ lower bound on the size of a strong coreset for ℓ_p subspace approximation for large p? Is there a $\tilde{O}(k^{p/2})\varepsilon^{-O(p)}$ upper bound?

We note that sensitivity sampling achieves a $\tilde{O}(k^{p/2+O(1)})\varepsilon^{-O(p)}$ upper bound [HV20, WY23a], while our upper bound is $\tilde{O}(k^{p/2})\varepsilon^{-O(p^2)}$.

Similar questions can also be asked for other guarantees for row subset selection for ℓ_p subspace approximation, all of which have been intensely studied for the case of $p = 2$ but remain to be answered for $p \neq 2$.

Question 15.2.3. How many rows are necessary and sufficient for a weak coreset $\mathbf{S}$ such that $\tilde{F} := \arg\min_{F\in\mathcal{F}_k} \|\mathbf{SA}(\mathbf{I} - \mathbf{P}_F)\|_p^p$ satisfies

$$\|\mathbf{A}(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_p^p \leq (1+\varepsilon) \min_{F\in\mathcal{F}_k} \|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_p^p,$$

as a function of both k and ε?

Question 15.2.4. How many rows are necessary and sufficient for a spanning coreset $S \subseteq [n]$ such that the span of the rows in S contains a k-dimensional subspace $\tilde{F}$ such that

$$\|\mathbf{A}(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_p^p \leq (1+\varepsilon) \min_{F\in\mathcal{F}_k} \|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_p^p,$$

as a function of both k and ε?

Question 15.2.5. How many rows are necessary and sufficient for a subset $S \subseteq [n]$ such that the span $\tilde{F}$ of the rows in S of dimension $\dim(S)$ satisfies

$$\|\mathbf{A}(\mathbf{I} - \mathbf{P}_{\tilde{F}})\|_p^p \leq (1+\varepsilon) \min_{F\in\mathcal{F}_k} \|\mathbf{A}(\mathbf{I} - \mathbf{P}_F)\|_p^p,$$

as a function of both k and ε?

Our strong coreset of [WY24b] immediately implies upper bounds to all three questions above, but it is possible to improve further in some of these cases. The guarantee in Question 15.2.4 was studied by [DV07, SV12, WY24a] for $p \neq 2$ with an efficient construction achieving an upper bound of $\tilde{O}(k^2 \cdot (k/\varepsilon)^{p+1})$ due to [DV07] and an inefficient construction achieving an upper bound of $\tilde{O}(k^2/\varepsilon)$ due to [SV12, Theorem 3.1] and $\tilde{O}(k/\varepsilon)$ for $1 < p < 2$ due to Theorem 13.1.5. The result [DV07] has a better dependence on ε than our strong coresets as well as those of [HV20, WY23a], and [SV12] has a better dependence on ε for all p and a better dependence on k for $p > 4$. In particular, it is an interesting question to achieve a nearly linear dependence in k for all p, and to construct such a subset of rows in polynomial time.

Part III

Sparse Optimization

Chapter 16

Sparse convex optimization via ℓ_1 regularization [YBC+23, AY23]

16.1 Introduction

A common task in modern machine learning is to sparsify a large model by selecting a subset of its inputs. This often leads to a number of improvements to the model such as interpretability and computational efficiency due to the smaller size of the model, as well as improved generalizability due to removal of noisy or redundant features. For these reasons, feature selection and sparse optimization is a heavily studied subject in signal processing, statistics, machine learning, and theoretical computer science. We continue this line of investigation by studying the following sparse optimization problem [SSZ10, LS17, EKDN18]: design an efficient algorithm such that, given $l : \mathbb{R}^n \to \mathbb{R}$ and a sparsity parameter k, outputs a sparse solution $\tilde{\boldsymbol{\beta}}$ such that

$$l(\mathbf{0}) - l(\tilde{\boldsymbol{\beta}}) \geq \gamma\left(l(\mathbf{0}) - \min_{\boldsymbol{\beta}\in\mathbb{R}^n:\|\boldsymbol{\beta}\|_0\leq k} l(\boldsymbol{\beta})\right), \qquad \|\boldsymbol{\beta}\|_0 \coloneqq |\{i \in [n] : \boldsymbol{\beta}_i \neq 0\}| \tag{16.1}$$

for some approximation factor $\gamma > 0$. In practice, there is also much interest in feature selection for vector-valued features, due to a widespread usage of vector representations of discrete features via embeddings [SWY75, WDL+09, PSM14, PRPG22], as well as for applications to block sparsification for hardware efficiency [NUD17, RPYU18], structured sparsification when pruning neurons in neural nets [AS16, SCHU17] or channels and filters in convolutional nets [LL16, WWW+16, LKD+17, MMK20, MK21]. In such vector-valued or group settings, the n inputs $\boldsymbol{\beta} \in \mathbb{R}^n$ are partitioned into t disjoint groups of features $T_1, T_2, \ldots, T_t \subseteq [n]$, and we would like to select whole groups of features at a time. We thus also study the question of solving

$$l(\mathbf{0}) - l(\tilde{\boldsymbol{\beta}}) \geq \gamma\left(l(\mathbf{0}) - \min_{\boldsymbol{\beta}\in\mathbb{R}^n:\|\boldsymbol{\beta}\|_{\text{group}}\leq k} l(\boldsymbol{\beta})\right), \qquad \|\boldsymbol{\beta}\|_{\text{group}} \coloneqq |\{i \in [t] : \boldsymbol{\beta}|_{T_i} \neq 0\}| \tag{16.2}$$

where $\boldsymbol{\beta}|_{T_i}$ denotes the $|T_i|$-dimensional vector obtained by restricting $\boldsymbol{\beta}$ to the coordinates $j \in T_i$.[1]

[1] We also allow for $\boldsymbol{\beta}|_{T_i}$ to denote the corresponding n-dimensional vector padded with zeros outside of T_i whenever this makes sense.

Although problems (16.1) and (16.2) are computationally challenging problems in general [Nat95, FKT15, GV21, PSZ22], a multitude of highly efficient algorithms have been proposed for solving these problems in practice. Perhaps one of the most popular algorithms in practice is the use of ℓ_1 regularization. That is, if we wish to optimize a function $l : \mathbb{R}^n \to \mathbb{R}$ over k-sparse inputs $\{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}\|_0 \leq k\}$, then we instead optimize the ℓ_1-regularized objective

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} l(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1. \tag{16.3}$$

The resulting optimal solution $\boldsymbol{\beta}^*$ often has few nonzero entries and thus helps identify a sparse solution. This idea was first introduced for the linear regression problem by Tibshirani [Tib96], known as the *LASSO* in this case, and has subsequently enjoyed wide adoption in practice in applications far beyond the original scope of linear regression. For the group sparsification setting, one can consider a generalization of the LASSO known as the *Group LASSO* [Bak99, YL06], which involves minimizing the following objective:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} l(\boldsymbol{\beta}) + \lambda \sum_{i=1}^{t} \|\boldsymbol{\beta}|_{T_i}\|_2 \tag{16.4}$$

That is, the regularizer is now the sum of the ℓ_2 norms of each group of variables T_i for $i \in [t]$. In practice, this encourages groups of variables to be selected at a time, which facilitates feature selection in the group setting. We refer the reader to the monograph [HTW15] on the LASSO and its generalizations for further references and discussion.

16.1.1 Related work: prior guarantees for ℓ_1 regularization

Due to the practical importance of solving (16.3) and (16.4), there has been an intense focus on theoretical work surrounding these optimization problems, especially for the *sparse linear regression* problem, i.e., when $l(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$ is the least squares objective for a design matrix $\mathbf{X}$ and target vector $\mathbf{y}$. However, as remarked in a number of works [FKT15, GV21, PSZ22], recovery guarantees for the LASSO and the Group LASSO are strikingly lacking in settings beyond statistical problems with average-case inputs or strong assumptions on the input, and is usually considered to be a heuristic in the context of sparse convex optimization for deterministic inputs. For example, one line of work focuses on the linear regression problem in the setting where the target vector $\mathbf{y}$ is exactly a k-sparse linear combination $\mathbf{X}\boldsymbol{\beta}$ for some $\|\boldsymbol{\beta}\|_0 \leq k$ plus i.i.d. Gaussian noise, and we seek guarantees on the solution to (16.3) [DS89, CDS98, Tro06, CRT06, CT07, Can08, BRT09, Zho09, RWY10, BCFS14] when $\mathbf{X}$ satisfies the *restricted isometry property (RIP)* or its various relaxations such as the *restricted eigenvalue condition (RE)*. This can be viewed as an instantiation of (16.1) for $l(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$, under the assumption that there exists an approximate global optimum of l that is exactly k-sparse. Statistical consistency results have also been established, which also assume a "true" k-sparse target solution [ZY06, MvdGB08]. A more recent line of work has studied algorithms for sparse linear regression problem under a correlated Gaussian design matrix with other general structural assumptions on the covariance matrix [KKMR21, KKMR22, KKMR23]. All of these works exclude the consideration of worst-case error on a desired k-sparse target solution, which is an undesirable restrictive assumption

when solving (16.1) in general. Indeed, one of the most remarkable aspects about the LASSO is its empirical success on a wide variety of real input distributions that can be far from Gaussian or even general i.i.d. designs. Thus, gaining a theoretical explanation of the success of the LASSO in more general settings is a critical question in this literature.

Question 16.1.1. Why are the LASSO and Group LASSO successful on general input distributions, beyond statistical settings?

An important exception is the work of [YBC+23], which establishes that in the setting of sparse linear regression, a sequential variation on the LASSO known as the Sequential LASSO [LC14], in which the LASSO is applied sequentially k times to select k inputs one at a time, is in fact equivalent to the Orthogonal Matching Pursuit algorithm (OMP) [PRK93, Tro04].[2] The work of [DK11] showed that OMP achieves bounds of the form of (16.1) whenever $\mathbf{X}$ satisfies a restricted isometry property in the absence of additional distributional assumptions on the input instance. From [YBC+23], it follows that the Sequential LASSO does as well. Thus, the works of [DK11, YBC+23] provide a form of an answer to Question 16.1.1 for the sparse linear regression problem, for general inputs with RIP.

Given the previous success of analyzing the LASSO for general inputs under RIP, one may ask for generalizations of this result to other objective functions, such as generalized linear models, logistic regression, or even general sparse convex optimization. Indeed, as mentioned previously, the LASSO and Group LASSO are used in practice in settings far beyond linear regression, and fast algorithms for solving the optimization problems of (16.3) and (16.4) are plentiful in the literature [LLAN06, KKB07, SFR07, MvdGB08, HTF09, FHT10, BPC+11, BJM+11, HMR23]. However, none of these works provide satisfactory answers on *why* the LASSO and Group LASSO are successful at selecting a good sparse set of inputs.

Question 16.1.2. Why are the LASSO and Group LASSO successful on general convex objectives, beyond ℓ_2 linear regression? Why do they select a sparse set of inputs? Which inputs are chosen?

While the work of [YBC+23] provides answers for the sparse linear regression problem by showing that the selected inputs are precisely the inputs selected by OMP, their analysis relies on specific geometric properties of the linear regression loss such as the Pythagorean theorem and the fact that the dual of the LASSO objective is a Euclidean norm projection onto a polytope [OPT00, TT11], and thus the techniques there do not immediately generalize even to specific problems such as ℓ_p regression or regularized logistic regression. Such a generalization is left as a central open question in their work. Similarly, the work of [TBF+12] asks the question of why sequentially discarding variables using the LASSO performs so well.

16.1.2 Our results

The main result of this work is a resolution of Question 16.1.2 for both the LASSO (16.3) and the Group LASSO (16.4) setting for any strictly convex objective function l. To state our results, we first recall the (Group) Sequential LASSO and (Group) OMP algorithms in Algorithms 6 and 7,

[2] We also note a work of [TBF+12], which proposes a similar procedure called the Strong Sequential Rule of sequentially zeroing out variables using the LASSO, but does not obtain provable guarantees for the resulting selected features.

which are both iterative algorithms that maintain a set of selected features $S \subseteq [t]$ by adding one feature at a time starting with $S = \varnothing$.

Algorithm 6 Group Sequential LASSO.

function GROUPSEQUENTIALLASSO(objective l, sparsity k, iterations k')
  Initialize $S \leftarrow \varnothing$
  for $r = 1$ to k' **do**
    Let $\tau \coloneqq \sup\{\lambda > 0 : \exists i \in \overline{S}, \boldsymbol{\beta}^\lambda|_{T_i} \neq 0\}$ for

$$\boldsymbol{\beta}^\lambda \coloneqq \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^n} l(\boldsymbol{\beta}) + \lambda \sum_{i \in \overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2$$

    For $\varepsilon > 0$ sufficiently small, let $i^* \in \overline{S}$ be such that $\boldsymbol{\beta}^{\tau-\varepsilon}|_{T_{i^*}} \neq 0$
    Update $S \leftarrow S \cup \{i^*\}$
  return S

Algorithm 7 Group Orthogonal Matching Pursuit.

function GROUPOMP(objective l, sparsity k, iterations k')
  Initialize $S \leftarrow \varnothing$
  for $r = 1$ to k' **do**
    Let

$$\boldsymbol{\beta}^\infty \coloneqq \arg \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^n \\ \forall i \in \overline{S}, \boldsymbol{\beta}|_{T_i} = 0}} l(\boldsymbol{\beta})$$

    Let $i^* \in \overline{S}$ be such that $\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 = \max_{i \in \overline{S}} \|\nabla l(\boldsymbol{\beta}^\infty)|_{T_i}\|_2^2$
    Update $S \leftarrow S \cup \{i^*\}$
  return S

We show that the result of [YBC+23] generalizes to the setting of group-sparse convex optimization: the Group Sequential LASSO update rule selects a group of features $T_i \subseteq [n]$ that maximizes the ℓ_2 gradient mass $\|\nabla l(\boldsymbol{\beta})|_{T_i}\|_2^2$, i.e., the same update rule as Group OMP. Our analysis simultaneously gives a substantial simplification as well as a generalization of the analysis of [YBC+23], which gives us the flexibility to handle both group settings as well as general convex functions.

Theorem 16.1.3. Let $l : \mathbb{R}^n \to \mathbb{R}$ be strictly convex. Let $S \subseteq [t]$ be a set of currently selected features. For each $\lambda > 0$, define

$$\boldsymbol{\beta}^\lambda \coloneqq \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^n} l(\boldsymbol{\beta}) + \lambda \sum_{i \in \overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2$$

and let $\tau \coloneqq \sup\{\lambda > 0 : \exists i \in \overline{S}, \boldsymbol{\beta}^\lambda|_{T_i} \neq 0\}$ and let $\boldsymbol{\beta}^\infty \coloneqq \boldsymbol{\beta}^\tau = \lim_{\lambda \to \infty} \boldsymbol{\beta}^\lambda$. Then for $\lambda = \tau - \varepsilon$ for all $\varepsilon > 0$ sufficiently small, $\boldsymbol{\beta}^\lambda|_{T_i} \neq 0$ only if $\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_i}\|_2^2 = \max_{j \in \overline{S}} \|\nabla l(\boldsymbol{\beta}^\infty)|_{T_j}\|_2^2$.

Proof. We give our discussion of this result in Section 16.3. □

In other words, if we add the Group LASSO regularization only on unselected features $i \in \overline{S}$ and take λ as large as possible without causing the solution $\boldsymbol{\beta}^\lambda$ to be zero, then $\boldsymbol{\beta}^\lambda$ must be supported on a group of features maximizing the ℓ_2 gradient mass at $\boldsymbol{\beta}^\infty$ among the unselected features $i \in \overline{S}$. Furthermore, note that $\boldsymbol{\beta}^\infty$ is exactly the minimizer of $l(\boldsymbol{\beta})$ subject to the constraint that $\boldsymbol{\beta}|_{T_i} = 0$ for every $i \in \overline{S}$. Thus, in the non-group setting, this algorithm sequentially selects a feature $i \in [n]$ that maximizes $|\nabla l(\boldsymbol{\beta}^\infty)_i|$, which is exactly the OMP update rule analyzed in [SSZ10, LS17, EKDN18]. The works of [SSZ10, LS17, EKDN18] show that this OMP update rule gives a guarantee of the form of (16.1) with an approximation factor γ depending on the *restricted strong convexity (RSC)* of l, which is a generalization of the RIP parameter for matrices to general functions. Thus, as reasoned in [YBC+23], the Sequential LASSO for general functions l inherits this guarantee of OMP. We also show in Section 16.4 that the group version of the OMP update rule obtained here based on selecting the group with the largest ℓ_2 gradient mass $\|\nabla l(\boldsymbol{\beta})|_{T_i}\|_2^2$ in fact also gives an analogous guarantee. In particular, we give guarantees for Group OMP both in the setting of outputting exactly k-group-sparse solutions (Corollary 16.4.5) as well as bicriteria solutions that use a slightly larger sparsity to get within an additive ε of the function value of the optimal k-sparse solution (Corollary 16.4.6), restated below.

Corollary 16.4.5 (Exactly k-group-sparse solutions)**.** After k iterations of Algorithm 7, $\boldsymbol{\beta}^\infty$ (Line 4) has group sparsity $\|\boldsymbol{\beta}^\infty\|_{\text{group}} \le k$ and satisfies (16.2) with

$$\gamma = 1 - \exp\left(-\frac{\mu_{2k}}{L_1}\right),$$

where μ_{2k} is a lower bound on the restricted strong convexity constant of l at group sparsity $2k$ and L_1 is an upper bound on the restricted smoothness constant of l at group sparsity 1 (see Definition 16.4.1).

Corollary 16.4.6 (Bicriteria sparsity with ε additive error)**.** After k' iterations of Algorithm 7, for

$$k' \ge k \cdot \frac{L_1}{\mu_{k+k'}} \log \frac{l(\boldsymbol{\beta}^{(0)}) - l(\boldsymbol{\beta}^*)}{\varepsilon},$$

then $\boldsymbol{\beta}^\infty$ (Line 4) has group sparsity $\|\boldsymbol{\beta}^\infty\|_{\text{group}} \le k'$ and satisfies

$$l(\boldsymbol{\beta}^\infty) \le l(\boldsymbol{\beta}^*) + \varepsilon\,,$$

where $\mu_{k+k'}$ is a lower bound on the restricted strong convexity constant of l at group sparsity $k + k'$ and L_1 is an upper bound on the restricted smoothness constant of l at group sparsity 1 (see Definition 16.4.1).

We additionally note that our analysis also immediately extends to an analysis of a local search version of OMP, known as OMP with Replacement (Algorithm 9) [JTD11, AS20], which gives a bicriteria sparsity bound which does not depend on ε (Corollary 16.4.10).

Corollary 16.4.10 (Bicriteria sparsity with ε additive error)**.** After R iterations of Algorithm 9 with $k' \geq k\left(\frac{L_2^2}{\mu_{k+k'}^2} + 1\right)$, for

$$R \geq k \cdot \frac{L_2}{\mu_{k+k'}} \log \frac{l(\boldsymbol{\beta}^{(0)}) - l(\boldsymbol{\beta}^*)}{\varepsilon},$$

then $\boldsymbol{\beta}^\infty$ (Line 4) has group sparsity $\|\boldsymbol{\beta}^\infty\|_{\text{group}} \leq k'$ and satisfies

$$l(\boldsymbol{\beta}^\infty) \leq l(\boldsymbol{\beta}^*) + \varepsilon\,,$$

where $\mu_{k+k'}$ is a lower bound on the restricted strong convexity constant of l at group sparsity $k + k'$ and L_2 is an upper bound on the restricted smoothness constant of l at group sparsity 2 (see Definition 16.4.1).

This variant of OMP can be analogously simulated by the LASSO as well, leading to a new LASSO-based feature selection algorithm which we call (Group) Sequential LASSO with Replacement.

Techniques

Our main technique involves exploiting the correspondence between variables of a primal optimization problem with the gradient of the dual optimization problem, via the *Fenchel–Young inequality* (Theorem 16.2.2).

We start with an observation given by [GVR10]. When we take the dual of the LASSO objective, then the resulting problem involves minimizing the *Fenchel dual* l^* of l (Definition 16.2.1), subject to a hypercube constraint set. When the regularization λ is sufficiently large (say larger than some threshold τ), then this increases the size of the constraint set large enough to contain the global minimizer of the Fenchel dual l^*, and thus the gradient of l^* vanishes at this minimizer. Then by the equality case of the Fenchel–Young inequality, this implies that the corresponding primal variable $\boldsymbol{\beta}$ is zero as well. On the other hand, if λ is smaller than this threshold point τ, only some coordinates will be unconstrained (i.e. strictly feasible), while others coordinates will become constrained by the smaller λ. In this case, the strictly feasible coordinates will have zero gradient, which leads to zeroes in the corresponding primal variable $\boldsymbol{\beta}$ and thus a sparse solution. The argument until this point is known in prior work, and [GVR10] used this observation to give an algorithm which tunes the value of λ such that at least k variables are selected in a single application, while [TBF+12] proposed a sequential procedure with better empirical performance.

Our central observation, inspired by the work of [YBC+23], is that if we regularize strongly enough such that only one feature is selected at a time via the LASSO, then this feature is the one maximizing the absolute value of the gradient. Indeed, note that if λ is just slightly smaller than the threshold point τ, then the global minimizer $\mathbf{u}^* \in \mathbb{R}^n$ of l^* just slightly violates exactly a single constraint in the dual problem, which corresponds to the feature $i^* \in [n]$ with the largest absolute coordinate value $|\mathbf{u}_i^*|$ in the dual variable. We show that for such λ, all other coordinates $j \in [n] \setminus \{i^*\}$ are unconstrained optimizers and thus the gradient is $\mathbf{0}$ (Lemma 16.3.2). Thus, by the equality case of the Fenchel–Young inequality, this corresponds to a primal variable $\boldsymbol{\beta}$

that is supported only on this coordinate $i^* \in [n]$. The crucial next step then is to *apply the Fenchel–Young inequality again in the dual direction*: via the Fenchel–Young inequality, this coordinate $i^* \in [n]$ maximizes the absolute coordinate value of the dual variable $\mathbf{u}$, and thus is the coordinate that maximizes the absolute coordinate value of the gradient of the primal variable $\boldsymbol{\beta}$. Thus, this selects a coordinate which follows the first step of the OMP update rule. While we have sketched the proof only for this first step in the non-group setting, the analysis also carries through for all steps of the OMP algorithm, as well as for the group setting. Thus, this establishes the equivalence between (Group) Sequential LASSO and (Group) OMP for general convex functions.

Connections to analysis of attention mechanisms

As noted in [YBC+23], we make a connection of our work to the analysis of recently popularized techniques for discrete optimization via continuous and differentiable relaxations inspired by the *attention mechanism* [VSP+17]. The attention mechanism can be viewed as a particular algorithm for the sparse optimization problem (16.1), in which an additional set of variables $\mathbf{w} \in \mathbb{R}^n$ are introduced, and we solve a new optimization problem

$$\min_{\mathbf{w},\boldsymbol{\beta} \in \mathbb{R}^n} l(\mathrm{softmax}(\mathbf{w}) \odot \boldsymbol{\beta}), \tag{16.5}$$

where $\odot$ denotes the Hadamard (entrywise) product and $\mathrm{softmax}(\mathbf{w}) \in \mathbb{R}^n$ is defined as

$$\mathrm{softmax}(\mathbf{w})_i := \frac{\exp(\mathbf{w}_i)}{\sum_{j=1}^n \exp(\mathbf{w}_j)}.$$

The idea is that $\mathbf{w}$ serves as a measure of "importance" of each feature $i \in [n]$, and the softmax allows for a differentiable relaxation for the operation of selecting the most "important" feature when minimizing the loss l. Alternatively, $\mathbf{w}$ can be viewed as the amount of "attention" placed on feature $i \in [n]$ by the algorithm. Such ideas have been applied extremely widely in machine learning, with applications to feature selection [LLY21, YBC+23], feature attribution [AP21], permutation learning [MBLS18], neural architecture search [LSY19], and differentiable programming [NLS16]. Thus, it is a critical problem to obtain a theoretical understanding of subset selection algorithms of the form of (16.5).

The work of [YBC+23] showed that a slight variation on (16.5) is in fact amenable to analysis when l is the problem of least squares linear regression. In this case, [YBC+23] show (using a result of [Hof17]) that if we instead consider

$$\min_{\mathbf{w},\boldsymbol{\beta} \in \mathbb{R}^n} l(\mathbf{w} \odot \boldsymbol{\beta}) + \frac{\lambda}{2}\left(\|\mathbf{w}\|_2^2 + \|\boldsymbol{\beta}\|_2^2\right) \tag{16.6}$$

i.e., remove the softmax and add ℓ_2 regularization, then this is in fact equivalent to the ℓ_1-regularized problem considered in (16.3). In Lemma 16.5.1, we show a generalization of this fact to the group setting, by showing that if we have t features corresponding to disjoint subsets of coordinates $T_1, T_2, \ldots, T_t \subseteq [n]$, then multiplying each of the features $\boldsymbol{\beta}|_{T_i}$ by a single "attention weight" $\mathbf{w}_i$ for $\mathbf{w} \in \mathbb{R}^t$ gives a similar correspondence to the Group LASSO algorithm (16.4). Thus, the attention-inspired feature selection algorithm given in Algorithm 8 also enjoys the

same guarantees as the Group Sequential LASSO algorithm. We note that this generalization to the group setting is particularly important for the various applications in attention-based subset selection algorithms, due to the fact that the objects $\boldsymbol{\beta}|_{T_i}$ being selected are often large vectors in these applications.

Algorithm 8 Group Sequential Attention.

function GROUPSEQUENTIALATTENTION(objective l, sparsity k, iterations k')
  Initialize $S \leftarrow \varnothing$
  for $r = 1$ to k' **do**
    Let $\tau \coloneqq \sup\{\lambda > 0 : \exists i \in \overline{S}, \boldsymbol{\beta}^\lambda|_{T_i} \neq 0\}$ for

$$\boldsymbol{\beta}^\lambda \coloneqq \arg\min_{\mathbf{w}\in\mathbb{R}^t, \boldsymbol{\beta}\in\mathbb{R}^n} l(\boldsymbol{\beta}_{\mathbf{w}}) + \frac{\lambda}{2}\sum_{i\in\overline{S}} \mathbf{w}_i^2 + \|\boldsymbol{\beta}|_{T_i}\|_2^2, \qquad \boldsymbol{\beta}_{\mathbf{w}}|_{T_i} \coloneqq \mathbf{w}_i \cdot \boldsymbol{\beta}|_{T_i}$$

    For $\varepsilon > 0$ sufficiently small, let $i^* \in \overline{S}$ be such that $\boldsymbol{\beta}^{\tau-\varepsilon}|_{T_{i^*}} \neq 0$
    Update $S \leftarrow S \cup \{i^*\}$
  return S

Finally, we also note that our analysis of Hadamard product-type of algorithms of the form of (16.6) may prove to be useful in the analysis of similar algorithms in the literature of online convex optimization that have been developed to solve sparse optimization problems [AW20b, AW20a, Chi22].

Applications to column subset selection

As a corollary of our analyses of group feature selection algorithms, we obtain the first algorithms for the *column subset selection (CSS)* problem for general loss functions with restricted strong convexity and smoothness.

In the CSS problem, we are given an input matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, and the goal is to select a small subset of k columns $S \subseteq [d]$ of $\mathbf{X}$ that minimizes the reconstruction error

$$\min_{\mathbf{V}\in\mathbb{R}^{k\times d}} \left\|\mathbf{X} - \mathbf{X}|^S\mathbf{V}\right\|_F^2, \tag{16.7}$$

where $\mathbf{X}|^S \in \mathbb{R}^{n\times k}$ is the matrix $\mathbf{X}$ restricted to the columns indexed by S. As with sparse linear regression, this problem is known to be computationally difficult [Çiv14], and thus most works focus on approximation algorithms and bicriteria guarantees to obtain tractable results.

The CSS problem can be viewed as an unsupervised analogue of sparse convex optimization, and has been studied extensively in prior work. In particular, the works of [FGK11, ÇM12, FGK13, FEGK15, SVW15, ABF+16, LS17] gave analyses of greedy algorithms for this problem, showing that iteratively selecting columns that maximizes the improvement in reconstruction error (16.7) leads to bicriteria sparsity algorithms that depend on the sparse condition number of $\mathbf{X}$. In a separate line of work, randomized methods have been employed in the randomized numerical linear algebra literature to sample columns of $\mathbf{X}$ that span a good low rank approximation [DV06, DMM08, BMD09, DR10, BDM11, CEM+15, BW17, CMM17]. Furthermore, there has recently

been a large body of work aimed at generalizing CSS results to more general loss functions beyond the Frobenius norm, including ℓ_p norms [SWZ17, CGK+17, DWZ+19, SWZ19, JLL+21, MW21] and other entrywise losses [SWZ19, WY23b]. All of these works use complicated arguments and rely heavily on the entrywise structure of the loss function.

We show that by a surprisingly simple argument, we can immediately obtain the first results on column subset selection for general convex loss functions with restricted strong convexity and smoothness. Our key insight is to view this problem not as a column subset selection problem for $\mathbf{X}$, but rather a *row subset selection problem for* $\mathbf{V}$. That is, note that

$$\min_{|S|\leq k}\min_{\mathbf{V}\in\mathbb{R}^{k\times d}} l\big(\mathbf{X}-\mathbf{X}|^S\mathbf{V}\big) = \min_{|S|\leq k}\min_{\mathbf{V}\in\mathbb{R}^{d\times d}} l(\mathbf{X}-\mathbf{X}\mathbf{V}|_S)$$

where $\mathbf{V}|_S$ zeros out all rows of $\mathbf{V}$ not indexed by S. Then, this is just a group variable selection problem, where we have d groups given by each of the rows of $\mathbf{V}$, and thus we may write this problem as computing

$$\mathsf{OPT} = \min_{\mathbf{V}\in\mathbb{R}^{d\times d},\|\mathbf{V}\|_{\text{group}}\leq k} l(\mathbf{X}-\mathbf{X}\mathbf{V})$$

Thus, by using our guarantees for Group OMP in Corollaries 16.4.5 and 16.4.6 (which also hold for Group Sequential LASSO and Group Sequential Attention by Theorem 16.1.3 and Lemma 16.5.1), we obtain the first algorithm and analysis of the column subset selection problem under general loss functions with restricted strong convexity and smoothness. This gives a substantial generalization of results known in prior work.

Theorem 16.1.4 (Column subset selection via Group OMP)**.** Let $\mathbf{X}\in\mathbb{R}^{n\times d}$ and let $l:\mathbb{R}^{n\times d}\to\mathbb{R}$ be a strictly convex and differentiable loss function. Let $\mathbf{V}\mapsto l(\mathbf{X}-\mathbf{X}\mathbf{V})$ satisfy L_1-group-sparse smoothness and $\mu_{k+k'}$-group-sparse convexity (Definition 16.4.1), where the groups are the rows of $\mathbf{V}$. The following hold:

- Let $\kappa = L_1/\mu_{2k}$. After $k' = k$ iterations, Algorithm 7 outputs a subset $S\subseteq[n]$ of size $|S|\leq k$ such that
$$l(\mathbf{X}) - l(\mathbf{X}-\mathbf{X}|^S\mathbf{V}) \geq \big(1-e^{-\kappa}\big)(l(\mathbf{X})-\mathsf{OPT}).$$
- Let $\kappa = L_1/\mu_{k+k'}$. After $k'\geq k\cdot\kappa\log\frac{l(\mathbf{X})-\mathsf{OPT}}{\varepsilon}$ iterations, Algorithm 7 outputs a subset $S\subseteq[n]$ of size $|S|\leq k'$ such that
$$l(\mathbf{X}-\mathbf{X}|^S\mathbf{V}) \leq \mathsf{OPT}+\varepsilon.$$

Proof. This follows from applying Corollaries 16.4.5 and 16.4.6 to the group-sparse convex optimization formulation of column subset selection. □

Our proof is arguably simpler than prior work even for the Frobenius norm. Indeed, the prior works require arguments that use the special structure of Euclidean projections, whereas we simply observe that CSS is a group-sparse convex optimization problem and use a generalization of techniques for sparse regression. We also immediately obtain analyses for natural algorithms which were previously not considered in the context of column subset selection, such as Group OMP (with Replacement), Group LASSO, and attention-based algorithms. In particular, by applying guarantees for Group OMP with Replacement (Corollary 16.4.10), we obtain the first column subset selection algorithm with no dependence on ε in the sparsity, even for the Frobenius norm problem.

Theorem 16.1.5 (Column subset selection with Group OMPR)**.** Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and let $l : \mathbb{R}^{n\times d} \to \mathbb{R}$ be a strictly convex and differentiable loss function. Let $\mathbf{V} \mapsto l(\mathbf{X} - \mathbf{XV})$ satisfy L_2-group-sparse smoothness and $\mu_{k+k'}$-group-sparse convexity (Definition 16.4.1), where the groups are the rows of $\mathbf{V}$. Let $\kappa = L_2/\mu_{k+k'}$ and $k' \geq k(\kappa^2 + 1)$. After $R \geq k \cdot \kappa \log \frac{l(\mathbf{X})-\mathsf{OPT}}{\varepsilon}$ iterations, Algorithm 9 outputs a subset $S \subseteq [n]$ of size $|S| \leq k'$ such that

$$l(\mathbf{X} - \mathbf{X}|^S\mathbf{V}) \leq \mathsf{OPT} + \varepsilon.$$

Proof. This follows from applying Corollary 16.4.10 to the group-sparse convex optimization formulation of column subset selection. □

16.1.3 Related work: the Forward Stagewise Regression conjecture

A separate line of work has investigated a closely related connection between the LASSO and OMP-like algorithms. In particular, the "continuous" OMP (or coordinate descent) algorithm which updates $\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)} - \eta \cdot \mathrm{sign}(\nabla_i l(\boldsymbol{\beta}^{(t)}))\mathbf{e}_i$ for $i = \arg\max_{i=1}^n \nabla l(\boldsymbol{\beta}^{(t)})$ known as Forward Stagewise Regression is conjectured [RZH04, Conjecture 2]) to have the same solution path as the LASSO path (i.e. the set of solutions as λ ranges from 0 to ∞) when $\eta \to 0$ [EHJT04, RZH04, Tib15, FGM17]. While a full proof of this conjecture may be useful towards proving our main result, to the best of our knowledge, the only known result towards this conjecture establishes an "instantaneous" result which shows the convergence of the difference between the two paths to the gradient [RZH04, Theorem 1] under technical assumptions under the underlying loss function such as the monotonicity of the coordinates of the LASSO solution. Our result can be viewed as a full proof of this conjecture in an open ball near $\mathbf{0}$ for general strictly convex differentiable functions, and our techniques may be useful for a full resolution of this conjecture.

16.1.4 Related work: algorithms for sparse convex optimization

While we have argued so far that guarantees for ℓ_1 regularization in solving (16.1) in prior work are limited, other efficient algorithms have in fact been shown to solve (16.1), both for sparse linear regression as well as general sparse convex optimization. Via a connection between convexity and weakly submodular optimization, the works of [SSZ10, LS17, EKDN18] showed that the greedy forward algorithm and Orthogonal Matching Pursuit both give guarantees of the form of (16.1). Efficiency guarantees have also been given for OMP with Replacement (OMPR) [SSZ10, JTD11, AS20] and Iterative Hard Thresholding (IHT) [JTK14, AS22], using the restricted smoothness and strong convexity properties. Ultimately, these results show that an ϵ-approximate sparse solution can be recovered if we allow an $O(\kappa)$ blowup to the sparsity, where κ is the restricted condition number of the problem.

16.1.5 Open directions

We suggest several directions for future study. Our first question is on showing analogous results for the one-shot version of LASSO, which is used much more frequently in practice than the Sequential LASSO. That is, if λ is chosen in (16.3) such that only k nonzero entries are selected,

then can we obtain a guarantee of the form of (16.1) for this solution? It is known that one-shot variants of OMP or greedy have this type of guarantee [DK11, EKDN18] (also called "oblivious" algorithms in these works). However, our proof techniques do not immediately apply, since we crucially use the fact that for large enough regularizations λ, the resulting solution is close to the $\lambda = \infty$ solution, while this is not true when λ can be much smaller.

A second question is whether our results generalize beyond convex functions or not. For example, the analysis of OMP carries through to smooth functions that satisfy the Polyak-Łojasiewicz condition [KNS16]. Can a similar generalization be shown for our results? There are several parts of our proofs that crucially use convexity, but the LASSO is known to give good results even for nonconvex functions in practice and thus there is still a gap in our understanding of this phenomenon.

Finally, we ask if our analyses for ℓ_1 regularization can be extended to an analogous result for nuclear norm regularization for rank-constrained convex optimization. In the setting of rank-constrained convex optimization, it has been shown in special cases, such as affine rank minimization, that nuclear norm regularization can be used to efficiently recover low rank solutions [RFP10]. This suggests that our results may have a natural generalization in this setting as well. In particular, an extension of OMP to the rank-sparse setting was shown by [AS21], and thus it is possible that nuclear norm regularization can be used to simulate this algorithm as well.

16.2 Preliminaries

Let $l : \mathbb{R}^n \to \mathbb{R}$ be strictly convex and differentiable. For each $i \in [t]$, let $T_i \subseteq [n]$ denote the group of variables that belong to the i-th feature.

16.2.1 Fenchel duality

We will use the following standard facts about Fenchel duality [BV04].

Definition 16.2.1 (Fenchel dual)**.** Let $l : \mathbb{R}^n \to \mathbb{R}$. Then, the Fenchel dual l^* of l is

$$l^*(\mathbf{u}) \coloneqq \sup_{\mathbf{z} \in \mathbb{R}^n} \mathbf{u}^\top \mathbf{z} - l(\mathbf{z}).$$

Theorem 16.2.2 (Fenchel–Young inequality)**.** Let $l : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Then,

$$l(\mathbf{z}) + l^*(\mathbf{u}) \geq \mathbf{u}^\top \mathbf{z}$$

with equality if and only if $\mathbf{u} = \nabla l(\mathbf{z})$.

Theorem 16.2.3 (Conjugacy theorem)**.** Let $l : \mathbb{R}^n \to \mathbb{R}$ be convex. Then, $(l^*)^* = l$.

The following is known about the convexity and differentiability of the Fenchel dual.

Theorem 16.2.4 (Differentiability of dual, Theorem 26.3, [Roc70])**.** Let $l : \mathbb{R}^n \to \mathbb{R}$ be strictly convex and differentiable. Then, l^* is strictly convex and differentiable.

16.2.2 Berge's theorem

We will use a well-known theorem of Berge on the continuity of the argmin for constrained optimization problems with parameterized constraint sets.

Recall that a correspondence $h : \mathbb{R} \rightrightarrows \mathbb{R}^n$ is a set-valued function which maps real numbers λ to subsets $h(\lambda) \subseteq \mathbb{R}^n$. A correspondence h is upper hemicontinuous if for every $\lambda \in \mathbb{R}$ and every open set $G \subseteq \mathbb{R}^n$ such that $h(\lambda) \subset G$, there is an open set $U \subseteq \mathbb{R}$ such that $\tau \in U \implies h(\tau) \subset G$.

Theorem 16.2.5 (Berge's theorem [Ber63])**.** Let $g : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and let $\varphi : \mathbb{R} \rightrightarrows \mathbb{R}^n$ be a continuous correspondence that map into compact sets. Consider the correspondence $h : \mathbb{R} \rightrightarrows \mathbb{R}^n$ given by

$$h(\lambda) = \left\{ \mathbf{u} \in \mathbb{R}^n : g(\mathbf{u}) = \min_{\mathbf{u}' \in \varphi(\lambda)} g(\mathbf{u}') \right\}$$

Then, h is upper hemicontinuous.

The following corollary of Theorem 16.2.5 for strictly convex functions is more useful for our purposes.

Corollary 16.2.6 (Berge's theorem for convex functions)**.** Let $g : \mathbb{R}^n \to \mathbb{R}$ be a strictly convex function and let $\varphi : \mathbb{R} \rightrightarrows \mathbb{R}^n$ be a continuous correspondence that map into compact sets. Consider the function $h : \mathbb{R} \to \mathbb{R}^n$ given by

$$h(\lambda) = \arg \min_{\mathbf{u}' \in \varphi(\lambda)} g(\mathbf{u}')$$

Then, h is continuous.

Proof. Because g is strictly convex, there is a unique minimizer $\mathbf{u}^\lambda$ of g for each $\lambda \in \mathbb{R}$, so h is well-defined. Furthermore, h is upper hemicontinuous as a correspondence that maps real numbers λ to singleton sets $\{h(\lambda)\}$ by Theorem 16.2.5, and any function h that is upper hemicontinuous as a correspondence is continuous as a function. □

16.3 Equivalence of Group Sequential LASSO and Group Orthogonal Matching Pursuit

We will give our proof of Theorem 16.1.3 in this section.

16.3.1 The dual problem

Consider the Group Sequential LASSO objective:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} l(\boldsymbol{\beta}) + \lambda \sum_{i \in \overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2 \tag{16.8}$$

We will show that the dual of this problem is

$$\begin{aligned} \max_{\mathbf{u}\in\mathbb{R}^n} -l^*(-\mathbf{u}) &= -\min_{\mathbf{u}\in\mathbb{R}^n} l^*(-\mathbf{u}) \\ \text{s.t.} \qquad \|\mathbf{u}|_{T_i}\|_2 &\le \lambda \text{ for each } i \in \overline{S} \\ \|\mathbf{u}|_{T_i}\|_2 &= 0 \text{ for each } i \in S \end{aligned} \tag{16.9}$$

We write the objective of (16.8) as a constrained optimization problem in the form of

$$\begin{aligned} \min_{\mathbf{z}\in\mathbb{R}^n, \boldsymbol{\beta}\in\mathbb{R}^d} l(\mathbf{z}) + \lambda \sum_{i\in\overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2 \\ \text{s.t.} \qquad \mathbf{z} = \boldsymbol{\beta} \end{aligned}$$

Then, the Lagrangian dual of this problem is

$$\min_{\mathbf{z}\in\mathbb{R}^n, \boldsymbol{\beta}\in\mathbb{R}^n} \max_{\mathbf{u}\in\mathbb{R}^n} l(\mathbf{z}) + \lambda \sum_{i\in\overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2 + \mathbf{u}^\top(\mathbf{z}-\boldsymbol{\beta})$$

Furthermore, the objective of (16.8) is convex and strictly feasible, so strong duality holds (see, e.g., Section 5.2.3 of [BV04]) and thus we may interchange the min and the max to obtain

$$\begin{aligned} &\max_{\mathbf{u}\in\mathbb{R}^n} \min_{\mathbf{z}\in\mathbb{R}^n, \boldsymbol{\beta}\in\mathbb{R}^n} l(\mathbf{z}) + \lambda \sum_{i\in\overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2 + \mathbf{u}^\top(\mathbf{z}-\boldsymbol{\beta}) \\ = \; &\max_{\mathbf{u}\in\mathbb{R}^n} \min_{\mathbf{z}\in\mathbb{R}^n} l(\mathbf{z}) + \mathbf{u}^\top\mathbf{z} + \min_{\boldsymbol{\beta}\in\mathbb{R}^n} \lambda \sum_{i\in\overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2 - \mathbf{u}^\top\boldsymbol{\beta} \end{aligned}$$

Now note that the first minimization over $\mathbf{z} \in \mathbb{R}^n$ gives exactly the Fenchel dual objective

$$\min_{\mathbf{z}\in\mathbb{R}^n} l(\mathbf{z}) + \mathbf{u}^\top\mathbf{z} = -\max_{\mathbf{z}\in\mathbb{R}^n} (-\mathbf{u})^\top\mathbf{z} - l(\mathbf{z}) = -l^*(-\mathbf{u}).$$

On the other hand, we show in the next lemma that the second minimization over $\boldsymbol{\beta} \in \mathbb{R}^n$ gives the constraints on the variables $\mathbf{u}$ given in (16.9).

Lemma 16.3.1. We have that

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^d} \lambda \sum_{i\in\overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2 - \mathbf{u}^\top\boldsymbol{\beta} = \begin{cases} 0 & \text{if } \|\mathbf{u}|_{T_i}\|_2 \le \lambda \text{ for } i \in \overline{S} \text{ and } \|\mathbf{u}|_{T_i}\|_2 = 0 \text{ for } i \in S \\ -\infty & \text{otherwise} \end{cases}$$

Proof. If $\|\mathbf{u}|_{T_i}\|_2 > \lambda$ for some coordinate $i \in \overline{S}$, then we may choose $\boldsymbol{\beta} = \mathbf{u}|_{T_i}$ so that

$$\lambda\|\mathbf{u}|_{T_i}\|_2 - \|\mathbf{u}|_{T_i}\|_2^2 = \|\mathbf{u}|_{T_i}\|_2(\lambda - \|\mathbf{u}|_{T_i}\|_2) < 0$$

so the objective can be made arbitrarily small by scaling. If $\|\mathbf{u}|_{T_i}\|_2 > 0$ for some $i \in S$, then we may choose $\boldsymbol{\beta} = \mathbf{u}|_{T_i}$ so that

$$\lambda \sum_{i\in\overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2 - \|\mathbf{u}|_{T_i}\|_2^2 = 0 - \|\mathbf{u}|_{T_i}\|_2^2 < 0$$

so the objective can be made arbitrarily small by scaling. Otherwise, we have that

$$\begin{aligned}
\mathbf{u}^\top \boldsymbol{\beta} &= \sum_{i \in \overline{S}} \mathbf{u}|_{T_i}^\top \boldsymbol{\beta}|_{T_i} && \text{since } \mathbf{u}|_{T_i} = 0 \text{ for every } i \in S \\
&\leq \sum_{i \in \overline{S}} \|\mathbf{u}|_{T_i}\|_2 \|\boldsymbol{\beta}|_{T_i}\|_2 && \text{Cauchy–Schwarz} \\
&\leq \lambda \sum_{i \in \overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2 && \text{since } \|\mathbf{u}|_{T_i}\|_2 \leq \lambda \text{ for every } i \in \overline{S}.
\end{aligned}$$

Thus,

$$\lambda \sum_{i \in \overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2 - \mathbf{u}^\top \boldsymbol{\beta} \geq 0$$

and furthermore, this value can be achieved by $\boldsymbol{\beta} = 0$. □

16.3.2 Selection of features

We will use Berge's theorem (Theorem 16.2.5) to prove the following lemma, which characterizes the gradient of the optimal solution to the dual optimization problem given by (16.9).

Lemma 16.3.2. Let $\lambda > 0$ and let $\mathbf{u}^\lambda$ be the minimizer of (16.9). Let $\mathbf{u}^\infty$ be the minimizer of (16.9) without the constraint that $\|\mathbf{u}|_{T_i}\|_2 \leq \lambda$ for every $i \in \overline{S}$. Define the threshold $\tau := \max_{i \in \overline{S}} \|\mathbf{u}^\infty|_{T_i}\|_2$ and let $M^\tau \subseteq \overline{S}$ denote the corresponding set of indices $i \in \overline{S}$ that witnesses the max, that is,

$$M^\tau := \left\{ i \in \overline{S} : \|\mathbf{u}^\infty|_{T_i}\|_2 = \tau \right\}.$$

The following hold:

- If $\lambda \geq \tau$, then $\nabla l^*(-\mathbf{u}^\lambda)|_{T_i} = 0$ for all $i \in \overline{S}$.
- If $\lambda = \tau - \varepsilon$ for sufficiently small $\varepsilon > 0$, then $\nabla l^*(-\mathbf{u}^\lambda)|_{T_i} = 0$ for all $i \in \overline{S} \setminus M^\tau$ and $\nabla l^*(-\mathbf{u}^\lambda)|_{T_i} \neq 0$ for some $i \in M^\tau$.

Proof. If $\lambda \geq \tau$, then the constraint $\max_{i \in \overline{S}} \left\| \mathbf{u}^\lambda|_{T_i} \right\|_2 \leq \lambda$ can be removed without affecting the optimal solution, so $\mathbf{u}^\lambda = \mathbf{u}^\infty$. Then for the coordinates in T_i for $i \in \overline{S}$, $\mathbf{u}^\infty$ is a minimizer for an unconstrained optimization problem, so the gradient is 0 on these coordinates. This shows the first bullet point.

On the other hand, suppose that $\lambda = \tau - \varepsilon$ for some small $\varepsilon > 0$. Then, $\mathbf{u}^\infty$ is outside the set $\{\mathbf{u} \in \mathbb{R}^n : \max_{i \in \overline{S}} \|\mathbf{u}|_{T_i}\|_2 \leq \lambda\}$. Now consider the function

$$h(\lambda) = \max_{i \in \overline{S} \setminus M^\tau} \|\mathbf{u}^\lambda|_{T_i}\|_2,$$

i.e., the second largest value of $\|\mathbf{u}^\lambda|_{T_i}\|_2$ after excluding the maximizers $i \in M^\tau$. Note that this function is continuous since $\lambda \mapsto \mathbf{u}^\lambda$ is continuous by Corollary 16.2.6. Furthermore, we have that $h(\tau) < \tau$, since the maximum in the definition of h excludes the indices M^τ. Let τ' satisfy $h(\tau) < \tau' < \tau$. Then, for all sufficiently small ε, we have that $h(\tau - \varepsilon) < \tau'$ by the continuity of h. For these ε, we can remove the constraints of $\|\mathbf{u}|_{T_i}\|_2 \leq \lambda = \tau - \varepsilon$ for $i \in \overline{S} \setminus M^\tau$ without

affecting the optimal solution $\mathbf{u}^\lambda$ in the optimization problem of (16.9), so on the coordinates T_i for $i \in \overline{S} \setminus M^\tau$, $\mathbf{u}^\lambda$ is an unconstrained minimizer and thus has zero gradient. On the other hand, for the coordinates T_i for $i \in M^\tau$, $\mathbf{u}^\lambda$ cannot be the unconstrained minimizer and thus there must be some nonzero coordinate in the gradient due to the convexity of l^*. □

We can then show that Lemma 16.3.2 in fact characterizes the support of the optimal solution $\boldsymbol{\beta}^*$ by relating the primal and dual variables via the Fenchel–Young inequality (Theorem 16.2.2).

Lemma 16.3.3 (Primal vs dual variables)**.** We have that $-\mathbf{u} = \nabla l(\boldsymbol{\beta})$ and $\boldsymbol{\beta} = \nabla l^*(-\mathbf{u})$.

Proof. The primal variable $\mathbf{z}$ is related to the dual variable $\mathbf{u}$ via Fenchel dual, that is,

$$l^*(-\mathbf{u}) = (-\mathbf{u})^\top \mathbf{z} - l(\mathbf{z})$$

Then by the tightness of the Fenchel–Young inequality (Theorem 16.2.2) for l, we have that $-\mathbf{u} = \nabla l(\mathbf{z})$. Furthemore, by the conjugacy theorem (Theorem 16.2.3), we have that $(l^*)^* = l$, so $l^*(-\mathbf{u}) + (l^*)^*(\mathbf{z}) = (-\mathbf{u})^\top \mathbf{z}$. Then by tightness of the Fenchel–Young inequality (Theorem 16.2.2) for l^*, we have that $\boldsymbol{\beta} = \mathbf{z} = \nabla l^*(-\mathbf{u})$. □

Thus, by Lemma 16.3.2, $\boldsymbol{\beta}^\lambda$ has a nonzero support on some group T_i if and only if the group T_i maximizes $\|\mathbf{u}^\infty|_{T_i}\|_2 = \|\nabla l(\boldsymbol{\beta}^\infty)|_{T_i}\|_2$. This is precisely the Group Orthogonal Matching Pursuit selection rule (see Line 5 of Algorithm 7).

16.4 Guarantees for Group Orthogonal Matching Pursuit

In this section, we give guarantees for the Group OMP algorithm (Algorithm 7). Our analysis is similar to [SSZ10, LS17, EKDN18]. We first introduce the notion of restricted strong convexity and smoothness, generalized to the group setting.

Definition 16.4.1 (Restricted strong convexity and smoothness)**.** Let $l : \mathbb{R}^n \to \mathbb{R}$. Let $T_i \subseteq [n]$ for $i \in [t]$ form a partition of $[n]$. Then, l is *μ_s-restricted strongly convex at group sparsity s* if for any $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\boldsymbol{\Delta} \in \mathbb{R}^n$ with $\|\boldsymbol{\Delta}\|_{\text{group}} \le s$,

$$l(\boldsymbol{\beta} + \boldsymbol{\Delta}) - l(\boldsymbol{\beta}) - \langle \nabla l(\boldsymbol{\beta}), \boldsymbol{\Delta} \rangle \ge \frac{\mu_s}{2} \|\boldsymbol{\Delta}\|_2^2$$

and *L_s-restricted smooth at group sparsity s* if for any $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\boldsymbol{\Delta} \in \mathbb{R}^n$ with $\|\boldsymbol{\Delta}\|_{\text{group}} \le s$,

$$l(\boldsymbol{\beta} + \boldsymbol{\Delta}) - l(\boldsymbol{\beta}) - \langle \nabla l(\boldsymbol{\beta}), \boldsymbol{\Delta} \rangle \le \frac{L_s}{2} \|\boldsymbol{\Delta}\|_2^2.$$

Lemma 16.4.2 (Smoothness)**.** Let l be L_1-restricted smooth at group sparsity 1. Let $r \in [k']$ and let $\boldsymbol{\beta}^\infty$ and i^* be defined as in Lines 4 and 5 of Algorithm 7 on the r-th iteration. Let $\boldsymbol{\beta}' := \boldsymbol{\beta}^\infty + \boldsymbol{\Delta}$ for $\boldsymbol{\Delta} = -L_1^{-1} \nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}$. Then,

$$(2L_1)^{-1} \|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 \le l(\boldsymbol{\beta}^\infty) - l(\boldsymbol{\beta}')$$

Proof. Note that $\boldsymbol{\Delta}$ has group sparsity 1. We then have that

$$\begin{aligned} l(\boldsymbol{\beta}') - l(\boldsymbol{\beta}^\infty) &\le \langle \nabla l(\boldsymbol{\beta}^\infty), \boldsymbol{\Delta} \rangle + \frac{L_1}{2}\|\boldsymbol{\Delta}\|_2^2 && L_1\text{-restricted smoothness} \\ &= -L_1^{-1}\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 + \frac{1}{2}L_1^{-1}\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 \\ &= -\frac{1}{2}L_1^{-1}\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2. \end{aligned}$$

Rearranging gives the desired result. □

Lemma 16.4.3 (Convexity)**.** Let l be $\mu_{k+k'}$-restricted strongly convex at group sparsity $k+k'$. Let $r \in [k']$ and let $\boldsymbol{\beta}^\infty$ and i^* be defined as in Lines 4 and 5 of Algorithm 7 on the r-th iteration. Let

$$\boldsymbol{\beta}^* := \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}\|_{\text{group}} \le k} l(\boldsymbol{\beta})$$

Then,

$$\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 \ge \frac{2\mu_{k+k'}}{k}(l(\boldsymbol{\beta}^\infty) - l(\boldsymbol{\beta}^*)).$$

Proof. Let $U^* \subseteq [n]$ be the support of $\boldsymbol{\beta}^*$ and let $U \subseteq [n]$ be the support of $\boldsymbol{\beta}^\infty$. Note that $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty\|_{\text{group}} \le k + k'$. Then,

$$\begin{aligned} l(\boldsymbol{\beta}^*) - l(\boldsymbol{\beta}^\infty) &\ge \langle \nabla l(\boldsymbol{\beta}^\infty), \boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty \rangle + \frac{\mu_{k+k'}}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty\|_2^2 \\ &= \left\langle \nabla l(\boldsymbol{\beta}^\infty), (\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty)|_{U^* \setminus U} \right\rangle + \frac{\mu_{k+k'}}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty\|_2^2 && \nabla l(\boldsymbol{\beta}^\infty)|_U = \mathbf{0} \\ &\ge -\|\nabla l(\boldsymbol{\beta}^\infty)|_{U^* \setminus U}\|_2 \|(\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty)|_{U^* \setminus U}\|_2 + \frac{\mu_{k+k'}}{2}\left\|(\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty)|_{U^* \setminus U}\right\|_2^2 \\ &\ge \min_x -\|\nabla l(\boldsymbol{\beta}^\infty)|_{U^* \setminus U}\|_2 x + \frac{\mu_{k+k'}}{2}x^2 \\ &= -\frac{\|\nabla l(\boldsymbol{\beta}^\infty)|_{U^* \setminus U}\|_2^2}{2\mu_{k+k'}} \end{aligned}$$

so

$$\|\nabla l(\boldsymbol{\beta}^\infty)|_{U^* \setminus U}\|_2^2 \ge 2\mu_{k+k'}(l(\boldsymbol{\beta}^\infty) - l(\boldsymbol{\beta}^*)).$$

Now note that $U^* \setminus U$ is supported on at most k groups, so by averaging, there exists some group T_i outside of U such that

$$\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_i}\|_2^2 \ge \frac{2\mu_{k+k'}}{k}(l(\boldsymbol{\beta}^\infty) - l(\boldsymbol{\beta}^*)).$$ □

Combining Lemmas 16.4.2 and 16.4.3 leads to the following stepwise guarantee for Algorithm 7.

Lemma 16.4.4. Let $\boldsymbol{\beta}^{(r)}$ denote the value of $\boldsymbol{\beta}^\infty$ (Line 4) after r iterations of Algorithm 7 with $\boldsymbol{\beta}^{(0)} = \mathbf{0}$. Let

$$\boldsymbol{\beta}^* := \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}\|_{\text{group}} \le k} l(\boldsymbol{\beta})$$

Then,

$$l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*) \le \exp\left(-\frac{r}{k}\frac{\mu_{k+k'}}{L_1}\right)\left(l(\boldsymbol{\beta}^{(0)}) - l(\boldsymbol{\beta}^*)\right)$$

Proof. By Lemmas 16.4.2 and 16.4.3, we have that

$$l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^{(r+1)}) \geq (2L_1)^{-1} \|\nabla l(\boldsymbol{\beta}^{(r)})|_{T_{i^*}}\|_2^2 \geq \frac{1}{k}\frac{\mu_{k+k'}}{L_1}\Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*)\Big)$$

so

$$\begin{aligned} l(\boldsymbol{\beta}^{(r+1)}) - l(\boldsymbol{\beta}^*) &= l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*) - \Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^{(r+1)})\Big) \\ &\leq l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*) - \frac{1}{k}\frac{\mu_{k+k'}}{L_1}\Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*)\Big) \\ &= \left(1 - \frac{1}{k}\frac{\mu_{k+k'}}{L_1}\right)\Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*)\Big) \\ &\leq \exp\left(-\frac{1}{k}\frac{\mu_{k+k'}}{L_1}\right)\Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*)\Big) \end{aligned}$$

Applying the above inductively proves the claim. □

As a result of Lemma 16.4.4, we obtain two guarantees for Algorithm 7, one for exact k-group-sparse solutions with large approximation and one for bicriteria sparsity with ε additive error.

Corollary 16.4.5 (Exactly k-group-sparse solutions)**.** After k iterations of Algorithm 7, $\boldsymbol{\beta}^\infty$ (Line 4) has group sparsity $\|\boldsymbol{\beta}^\infty\|_{\text{group}} \leq k$ and satisfies (16.2) with

$$\gamma = 1 - \exp\left(-\frac{\mu_{2k}}{L_1}\right),$$

where μ_{2k} is a lower bound on the restricted strong convexity constant of l at group sparsity $2k$ and L_1 is an upper bound on the restricted smoothness constant of l at group sparsity 1 (see Definition 16.4.1).

Proof. After k iterations, we have by Lemma 16.4.4 applied for $k' = k$ that

$$l(\boldsymbol{\beta}^{(k)}) - l(\boldsymbol{\beta}^*) = l(\boldsymbol{\beta}^{(k)}) - l(\boldsymbol{\beta}^{(0)}) + l(\boldsymbol{\beta}^{(0)}) - l(\boldsymbol{\beta}^*) \leq \exp\left(-\frac{\mu_{2k}}{L_1}\right)\Big(l(\boldsymbol{\beta}^{(0)}) - l(\boldsymbol{\beta}^*)\Big)$$

which rearranges to

$$l(\boldsymbol{\beta}^{(0)}) - l(\boldsymbol{\beta}^{(k)}) \geq \left(1 - \exp\left(-\frac{\mu_{2k}}{L_1}\right)\right)\Big(l(\boldsymbol{\beta}^{(0)}) - l(\boldsymbol{\beta}^*)\Big)$$

□

Corollary 16.4.6 (Bicriteria sparsity with ε additive error)**.** After k' iterations of Algorithm 7, for

$$k' \geq k \cdot \frac{L_1}{\mu_{k+k'}} \log \frac{l(\boldsymbol{\beta}^{(0)}) - l(\boldsymbol{\beta}^*)}{\varepsilon},$$

then $\boldsymbol{\beta}^\infty$ (Line 4) has group sparsity $\|\boldsymbol{\beta}^\infty\|_{\text{group}} \leq k'$ and satisfies

$$l(\boldsymbol{\beta}^\infty) \leq l(\boldsymbol{\beta}^*) + \varepsilon\,,$$

where $\mu_{k+k'}$ is a lower bound on the restricted strong convexity constant of l at group sparsity $k+k'$ and L_1 is an upper bound on the restricted smoothness constant of l at group sparsity 1 (see Definition 16.4.1).

Proof. This follows immediately from the bound of Lemma 16.4.4 and rearranging. □

16.4.1 Group OMP with Replacement

In this section, we give guarantees for the Group OMP with Replacement algorithm (Algorithm 9), which is an improvement to Group OMP that can achieve a sparsity bound that is independent of the accuracy parameter ε [AS20].

Algorithm 9 Group Orthogonal Matching Pursuit with Replacement.

1: **function** GROUPOMPR(objective l, sparsity k, initial sparsity k', iterations R)
2: Initialize $S^0 \subseteq [n]$ with $|S^0| = k'$, e.g. using Algorithm 7.
3: **for** $r = 0$ to $R-1$ **do**
4: Let

$$\boldsymbol{\beta}^\infty \coloneqq \arg \min_{\substack{\boldsymbol{\beta}\in\mathbb{R}^n \\ \forall i\in\overline{S}^r, \boldsymbol{\beta}|_{T_i}=0}} l(\boldsymbol{\beta})$$

5: Let $i^* \in \overline{S}^r$ be such that $\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 = \max_{i\in\overline{S}^r}\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_i}\|_2^2$
6: Let $j^* \in S^r$ be such that $\|\boldsymbol{\beta}^\infty|_{T_{j^*}}\|_2^2 = \min_{j\in S}\|\boldsymbol{\beta}^\infty|_{T_j}\|_2^2$
7: Update $S^{r+1} \leftarrow S^r \cup \{i^*\} \setminus \{j^*\}$
8: **return** S^r, $r \in [R]$, that minimizes

$$\min_{\substack{\boldsymbol{\beta}\in\mathbb{R}^n \\ \forall i\in\overline{S}^r, \boldsymbol{\beta}|_{T_i}=0}} l(\boldsymbol{\beta})$$

Lemma 16.4.7 (Smoothness)**.** Let l be L_2-restricted smooth at group sparsity 2. Let $r \in [k']$ and let $\boldsymbol{\beta}^\infty$, i^*, j^* be defined as in Lines 4, 5 and 6 of Algorithm 9 on the r-th iteration. Let $\boldsymbol{\beta}' \coloneqq \boldsymbol{\beta}^\infty + \boldsymbol{\Delta}$ for $\boldsymbol{\Delta} = -L_2^{-1}\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}} - \boldsymbol{\beta}^\infty|_{T_{j^*}}$. Then,

$$(2L_2)^{-1}\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 - (1/2)L_2\|\boldsymbol{\beta}^\infty|_{T_{j^*}}\|_2^2 \leq l(\boldsymbol{\beta}^\infty) - l(\boldsymbol{\beta}')$$

Proof. Note that $\boldsymbol{\Delta}$ has group sparsity 2. We then have that

$$\begin{aligned} l(\boldsymbol{\beta}') - l(\boldsymbol{\beta}^\infty) &\leq \langle \nabla l(\boldsymbol{\beta}^\infty), \boldsymbol{\Delta}\rangle + \frac{L_2}{2}\|\boldsymbol{\Delta}\|_2^2 && L_2\text{-restricted smoothness} \\ &= -L_2^{-1}\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 + \frac{1}{2}L_2^{-1}\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 + \frac{1}{2}L_2\|\boldsymbol{\beta}^\infty|_{T_{j^*}}\|_2^2 && (\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{j^*}}\|_2^2 = 0) \end{aligned}$$

$$= -\frac{1}{2}L_2^{-1}\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 + \frac{1}{2}L_2\|\boldsymbol{\beta}^\infty|_{T_{j^*}}\|_2^2.$$

Rearranging gives the desired result. □

Lemma 16.4.8 (Convexity)**.** Let l be $\mu_{k+k'}$-restricted strongly convex at group sparsity $k+k'$. Let $r \in [k']$ and let $\boldsymbol{\beta}^\infty$, i^*, j^* be defined as in Lines 4, 5 and 6 of Algorithm 9 on the r-th iteration. Let

$$\boldsymbol{\beta}^* := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n:\|\boldsymbol{\beta}\|_{\text{group}}\le k} l(\boldsymbol{\beta})$$

Then,

$$\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_{i^*}}\|_2^2 \ge \frac{2\mu_{k+k'}}{k}(l(\boldsymbol{\beta}^\infty) - l(\boldsymbol{\beta}^*)) + \frac{(k'-k)\mu_{k+k'}^2}{k}\|\boldsymbol{\beta}^\infty|_{T_{j^*}}\|_2^2.$$

Proof. Let $U^* \subseteq [n]$ be the support of $\boldsymbol{\beta}^*$ and let $U \subseteq [n]$ be the support of $\boldsymbol{\beta}^\infty$. Note that $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty\|_{\text{group}} \le k + k'$. Then,

$$\begin{aligned}
&l(\boldsymbol{\beta}^*) - l(\boldsymbol{\beta}^\infty)\\
&\ge \langle \nabla l(\boldsymbol{\beta}^\infty), \boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty\rangle + \frac{\mu_{k+k'}}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty\|_2^2\\
&= \left\langle \nabla l(\boldsymbol{\beta}^\infty), (\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty)|_{U^*\setminus U}\right\rangle + \frac{\mu_{k+k'}}{2}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty\|_2^2\\
&\ge -\|\nabla l(\boldsymbol{\beta}^\infty)|_{U^*\setminus U}\|_2\|(\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty)|_{U^*\setminus U}\|_2 + \frac{\mu_{k+k'}}{2}\left\|(\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty)|_{U^*\setminus U}\right\|_2^2 + \frac{\mu_{k+k'}}{2}\left\|(\boldsymbol{\beta}^* - \boldsymbol{\beta}^\infty)|_{U\setminus U^*}\right\|_2^2\\
&\ge \min_x -\|\nabla l(\boldsymbol{\beta}^\infty)|_{U^*\setminus U}\|_2 x + \frac{\mu_{k+k'}}{2}x^2 + \frac{\mu_{k+k'}}{2}\left\|\boldsymbol{\beta}^\infty|_{U\setminus U^*}\right\|_2^2\\
&= -\frac{\|\nabla l(\boldsymbol{\beta}^\infty)|_{U^*\setminus U}\|_2^2}{2\mu_{k+k'}} + \frac{\mu_{k+k'}}{2}\left\|\boldsymbol{\beta}^\infty|_{U\setminus U^*}\right\|_2^2
\end{aligned}$$

so

$$\|\nabla l(\boldsymbol{\beta}^\infty)|_{U^*\setminus U}\|_2^2 \ge 2\mu_{k+k'}(l(\boldsymbol{\beta}^\infty) - l(\boldsymbol{\beta}^*)) + \mu_{k+k'}^2\left\|\boldsymbol{\beta}^\infty|_{U\setminus U^*}\right\|_2^2.$$

Now note that $U^* \setminus U$ is supported on at most k groups, so by averaging, there exists some group T_i outside of U such that

$$\begin{aligned}
\|\nabla l(\boldsymbol{\beta}^\infty)|_{T_i}\|_2^2 &\ge \frac{2\mu_{k+k'}}{k}(l(\boldsymbol{\beta}^\infty) - l(\boldsymbol{\beta}^*)) + \frac{\mu_{k+k'}^2}{k}\left\|\boldsymbol{\beta}^\infty|_{U\setminus U^*}\right\|_2^2\\
&\ge \frac{2\mu_{k+k'}}{k}(l(\boldsymbol{\beta}^\infty) - l(\boldsymbol{\beta}^*)) + \frac{(k'-k)\mu_{k+k'}^2}{k}\|\boldsymbol{\beta}^\infty|_{T_{j^*}}\|_2^2.
\end{aligned}$$

□

Lemma 16.4.9. Let $\boldsymbol{\beta}^{(r)}$ denote the value of $\boldsymbol{\beta}^\infty$ (Line 4) after r iterations of Algorithm 9 with $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ and $|S^0| = k' \ge k\left(\frac{L_2^2}{\mu_{k+k'}^2} + 1\right)$. Let

$$\boldsymbol{\beta}^* := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n:\|\boldsymbol{\beta}\|_{\text{group}}\le k} l(\boldsymbol{\beta})$$

Then,

$$l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*) \le \exp\left(-\frac{r}{k}\frac{\mu_{k+k'}}{L_2}\right)\left(l(\boldsymbol{\beta}^{(0)}) - l(\boldsymbol{\beta}^*)\right)$$

Proof. By Lemmas 16.4.7 and 16.4.8, we have that

$$\begin{aligned} l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^{(r+1)}) &\geq (2L_2)^{-1} \|\nabla l(\boldsymbol{\beta}^{(r)})|_{T_{i^*}}\|_2^2 - (1/2)L_2 \left\|\boldsymbol{\beta}^{\infty}|_{T_{j^*}}\right\|_2^2 \\ &\geq \frac{1}{k}\frac{\mu_{k+k'}}{L_2}\Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*)\Big) + \frac{1}{2}\left(\frac{(k'-k)\mu_{k+k'}^2}{kL_2} - L_2\right)\left\|\boldsymbol{\beta}^{\infty}|_{T_{j^*}}\right\|_2^2 \\ &\geq \frac{1}{k}\frac{\mu_{k+k'}}{L_2}\Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*)\Big), \end{aligned}$$

as long as $k' \geq k\big(L_2^2/\mu_{k+k'}^2 + 1\big)$. So,

$$\begin{aligned} l(\boldsymbol{\beta}^{(r+1)}) - l(\boldsymbol{\beta}^*) &= l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*) - \Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^{(r+1)})\Big) \\ &\leq l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*) - \frac{1}{k}\frac{\mu_{k+k'}}{L_2}\Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*)\Big) \\ &= \left(1 - \frac{1}{k}\frac{\mu_{k+k'}}{L_2}\right)\Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*)\Big) \\ &\leq \exp\left(-\frac{1}{k}\frac{\mu_{k+k'}}{L_2}\right)\Big(l(\boldsymbol{\beta}^{(r)}) - l(\boldsymbol{\beta}^*)\Big) \end{aligned}$$

Applying the above inductively proves the claim. □

Corollary 16.4.10 (Bicriteria sparsity with ε additive error)**.** After R iterations of Algorithm 9 with $k' \geq k\left(\frac{L_2^2}{\mu_{k+k'}^2} + 1\right)$, for

$$R \geq k \cdot \frac{L_2}{\mu_{k+k'}} \log \frac{l(\boldsymbol{\beta}^{(0)}) - l(\boldsymbol{\beta}^*)}{\varepsilon},$$

then $\boldsymbol{\beta}^{\infty}$ (Line 4) has group sparsity $\|\boldsymbol{\beta}^{\infty}\|_{\text{group}} \leq k'$ and satisfies

$$l(\boldsymbol{\beta}^{\infty}) \leq l(\boldsymbol{\beta}^*) + \varepsilon\,,$$

where $\mu_{k+k'}$ is a lower bound on the restricted strong convexity constant of l at group sparsity $k + k'$ and L_2 is an upper bound on the restricted smoothness constant of l at group sparsity 2 (see Definition 16.4.1).

Proof. This follows immediately from the bound of Lemma 16.4.9 and rearranging. □

16.5 Equivalence of Group Sequential Attention and Group Sequential LASSO

We generalize a result of [Hof17] to the group setting, which allows us to translate guarantees for Group Sequential LASSO (Algorithm 6) to Group Sequential Attention (Algorithm 8).

Lemma 16.5.1. Let $l : \mathbb{R}^n \to \mathbb{R}$ and $\lambda > 0$. Let $T_i \subseteq [n]$ for $i \in [t]$ form a partition of $[n]$. Let $S \subseteq [t]$. Then,

$$\inf_{\boldsymbol{\beta} \in \mathbb{R}^n} l(\boldsymbol{\beta}) + \lambda \sum_{i \in \overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2 = \inf_{\mathbf{w} \in \mathbb{R}^t, \boldsymbol{\beta} \in \mathbb{R}^n} l(\boldsymbol{\beta}_{\mathbf{w}}) + \frac{\lambda}{2} \left(\|\mathbf{w}|_{\overline{S}}\|_2^2 + \sum_{i \in \overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2^2 \right)$$

where $\boldsymbol{\beta}_{\mathbf{w}} \in \mathbb{R}^n$ is the vector such that $\boldsymbol{\beta}_{\mathbf{w}}|_{T_i} \coloneqq \mathbf{w}_i \cdot \boldsymbol{\beta}|_{T_i}$.

Proof. We have that

$$\inf_{\mathbf{w} \in \mathbb{R}^t, \boldsymbol{\beta} \in \mathbb{R}^n} l(\boldsymbol{\beta}_{\mathbf{w}}) + \frac{\lambda}{2} \left(\|\mathbf{w}|_{\overline{S}}\|_2^2 + \sum_{i \in \overline{S}} \|\boldsymbol{\beta}|_{T_i}\|_2^2 \right) = \inf_{\mathbf{w} \in \mathbb{R}^t, \mathbf{u} \in \mathbb{R}^n} l(\mathbf{u}) + \frac{\lambda}{2} \sum_{i \in \overline{S}} \mathbf{w}_i^2 + \frac{\|\mathbf{u}|_{T_i}\|_2^2}{\mathbf{w}_i^2}$$

Now note that for each $i \in \overline{S}$, we have that

$$\mathbf{w}_i^2 + \frac{\|\mathbf{u}|_{T_i}\|_2^2}{\mathbf{w}_i^2} \geq 2\|\mathbf{u}|_{T_i}\|_2$$

with equality if and only if $\mathbf{w}_i^2 = \|\mathbf{u}|_{T_i}\|_2$ by tightness of the AM-GM inequality. □

16.6 Experiments: feature selection via Sequential Attention

We present experimental results when running the Sequential Attention algorithm, as investigated in [YBC+23]. The code for these experiments can be found at https://github.com/google-research/google-research/tree/master/sequential_attention.

16.6.1 Small-scale experiments

We investigate the performance of Sequential Attention, as presented in Algorithm 8, through experiments on standard feature selection benchmarks for neural networks. In these experiments, we consider six datasets (see Table 16.1) used in experiments in [LRT21, BAZ19], and select $k = 50$ features using a one-layer neural network with hidden width 67 and ReLU activation (just as in these previous works). For more points of comparison, we also implement the attention-based feature selection algorithms of [BAZ19, LLY21] and the Group LASSO, which has been considered in many works that aim to sparsify neural networks. We also implement natural adaptations of the Sequential LASSO and OMP for neural networks and evaluate their performance.

In Figure 16.1, we see that Sequential Attention is competitive with or outperforms all feature selection algorithms on this benchmark suite. For each algorithm, we report the mean of the prediction accuracies averaged over five feature selection trials.

We note that our algorithm is considerably more efficient compared to prior feature selection algorithms, especially those designed for neural networks. This is because many of these prior algorithms introduce entire subnetworks to train [BAZ19, GGH19, WC20, LLY21], whereas Sequential Attention only adds d additional trainable variables. Furthermore, in these experiments, we implement an optimized version of Algorithm 8 that only trains one model rather than k models, by partitioning the training epochs into k parts and selecting one feature in each of these k parts. Combining these two aspects makes for an extremely efficient algorithm.

Table 16.1: Statistics on benchmark datasets.

Dataset	# Examples	# Features	# Classes	Type
Mice Protein	1,080	77	8	Biology
MNIST	60,000	784	10	Image
MNIST-Fashion	60,000	784	10	Image
ISOLET	7,797	617	26	Speech
COIL-20	1,440	400	20	Image
Activity	5,744	561	6	Sensor

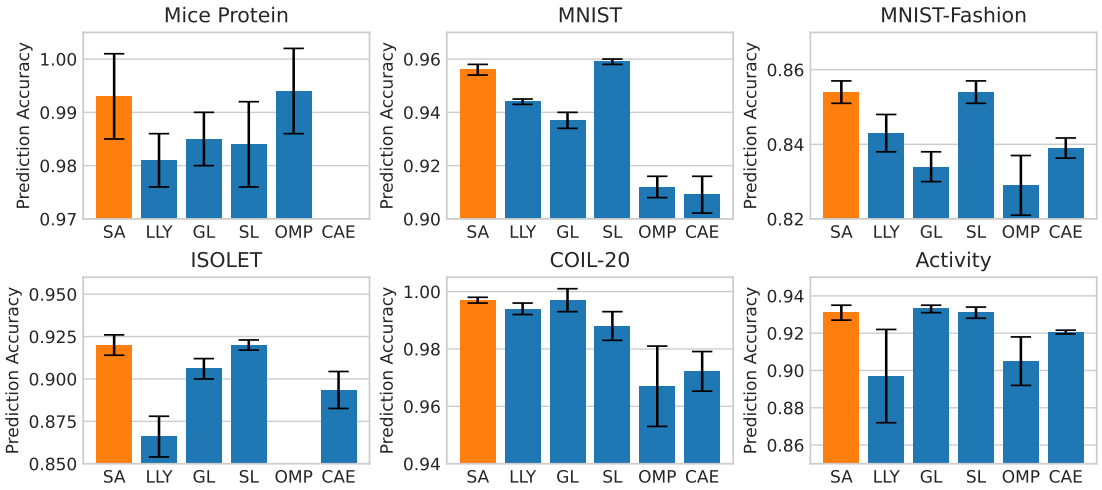

Figure 16.1: Feature selection results for small-scale neural network experiments. Here, SA = Sequential Attention, LLY = [LLY21], GL = Group LASSO, SL = Sequential LASSO, OMP = OMP, and CAE = Concrete Autoencoder [BAZ19].

16.6.2 Large-scale experiments

To demonstrate the scalability of our algorithm, we perform large-scale feature selection experiments on the Criteo click dataset, which consists of 39 features and over three billion examples for predicting click-through rates [DGL17]. Our results in Figure 16.2 show that Sequential Attention outperforms other methods when at least 15 features are selected. In particular, these plots highlight the fact that Sequential Attention excels at finding valuable features once a few features are already in the model, and that it has substantially less variance than LASSO-based feature selection algorithms.

16.6.3 Visualization of selected MNIST features

In Figure 16.3, we present visualizations of the features (i.e., pixels) selected by Sequential Attention and the baseline algorithms. This provides some intuition on the nature of the features that these algorithms select. Similar visualizations for MNIST can be found in works such as

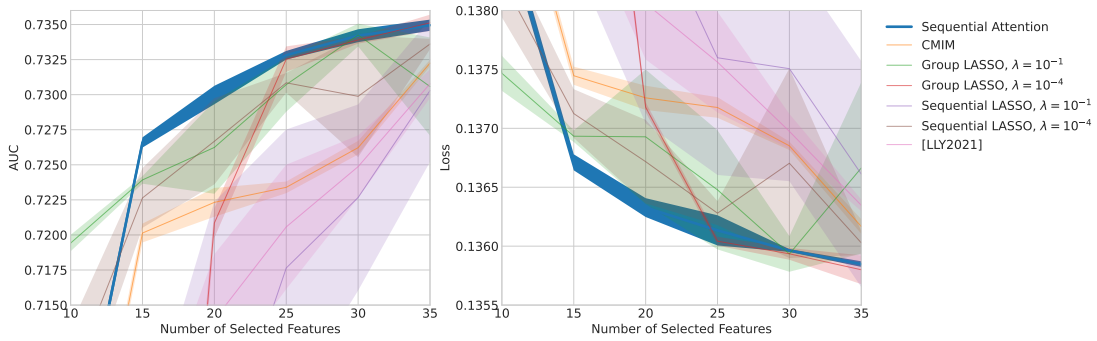

Figure 16.2: AUC and log loss when selecting $k \in \{10, 15, 20, 25, 30, 35\}$ features for Criteo dataset.

[BAZ19, GGH19, WC20, LRT21, LLY21]. Note that these visualizations serve as a basic sanity check about the kinds of pixels that these algorithms select. For instance, the degree to which the selected pixels are "clustered" can be used to informally assess the redundancy of features selected for image datasets, since neighboring pixels tend to represent redundant information. It is also useful at time to assess which regions of the image are selected. For example, the central regions of the MNIST images are more informative than the edges.

Sequential Attention selects a highly diverse set of pixels due to its adaptivity. Sequential LASSO also selects a very similar set of pixels, as suggested by our theoretical analysis in Section 16.3. Curiously, OMP does not yield a competitive set of pixels, which demonstrates that OMP does not generalize well from least squares regression and generalized linear models to deep neural networks.

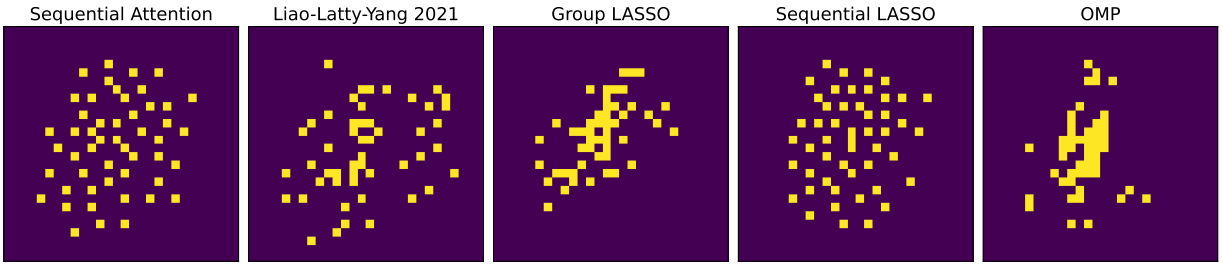

Figure 16.3: Visualizations of the $k = 50$ pixels selected by the feature selection algorithms on MNIST.

Chapter 17

Column subset selection with entrywise losses [WY23a]

As is the case for subspace embeddings and linear regression, the problem of low rank approximation is best understood when the norm under consideration is the ℓ_2 loss (which corresponds to the Frobenius norm in this case), and a long line of work has studied fast randomized algorithms for low rank approximation under the Frobenius norm [FKV04, DV06, DKM06a, DKM06b, DKM06c, DMM06b, CW13, MM15, CMM17, BW17]. However, when the input matrix is corrupted by heavy-tailed noise or include outliers, the ℓ_2 norm is not always the most desirable due to the fact that it tends to fit to the outliers too much. Thus, oftentimes, it is desirable to solve the low rank approximation problem under other error measures, especially those with slower growth than the ℓ_2 loss. One notable class of losses is the *entrywise ℓ_p loss*, and more generally, the *entrywise g loss*, where g can be an arbitrary loss function.

Definition 17.0.1 (Entrywise losses)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$. Then, we define the *entrywise g norm* of $\mathbf{A}$ as

$$\|\mathbf{A}\|_g \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{d} g(\mathbf{A}_{i,j}).$$

When $g(x) = |x|^p$, then we instead define

$$\|\mathbf{A}\|_{p,p} \coloneqq \left(\sum_{i=1}^{n}\sum_{j=1}^{d}|\mathbf{A}_{i,j}|^p\right)^{1/p}$$

to be the entrywise ℓ_p loss.

For $p \neq 2$, the entrywise loss low rank approximation is computationally hard to approximate under a variety of natural hardness assumptions [Mie09, GV18, DHJ+18, BBB+19, MW21] and thus we need to allow for an appropriate notion of approximation to obtain efficient algorithms. We study bicriteria approximation guarantees of the following form:

Definition 17.0.2 (Bicriteria coreset for low rank approximation)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, let k be a rank parameter, and let $\|\cdot\|$ be any loss function. Let $S \subseteq [d]$ a subset of columns, and write $\mathbf{A}|^S$ for

the $n \times S$ matrix formed by the columns of $\mathbf{A}$ indexed by S.[1] Then, S is a *bicriteria coreset* with distortion $\kappa \geq 1$ if

$$\min_{\mathbf{X} \in \mathbb{R}^{S \times d}} \|\mathbf{A} - \mathbf{A}|^S \mathbf{X}\| \leq \kappa \min_{\text{rank}(\mathbf{A}') \leq k} \|\mathbf{A} - \mathbf{A}'\|.$$

17.1 Algorithms for general entrywise losses

We begin by presenting our result on the entrywise g-norm low rank approximation problem, which was first considered by [SWZ19]. For our analysis, we will need to assume several natural properties on g, which have been considered in previous work [CW15b, CW15a, SWZ19, MMWY22] for obtaining provable guarantees for randomized numerical linear algebra under a broad class of loss functions:

Definition 17.1.1. Let $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$. Then:

- g satisfies the $\mathsf{ati}_{g,t}$*-approximate triangle inequality* if for any $x_1, x_2, \ldots, x_t$, $g(\sum x_i) \leq \mathsf{ati}_{g,t} \cdot \sum_i g(x_i)$.
- g is mon_g*-monotone* if for any $0 \leq |x| \leq |y|$, $g(x) \leq \mathsf{mon}_g \cdot g(y)$.
- g has at least lin_g*-linear growth* if for any $0 < |x| \leq |y|$, $g(y)/g(x) \geq \mathsf{lin}_g \cdot |y|/|x|$.

For example, popular functions that satisfy these bounds include the Huber loss, Fair loss, Cauchy loss, ℓ_1-ℓ_2 loss, and the quantile loss [SWZ19]. While the lin_g-linear growth bound excludes the Tukey loss, which grows quadratically near the origin and stays constant away from the origin, it allows for a modification of the Tukey loss where the constant away from the origin is replaced by an arbitrarily slow linear growth [CW15a].

[SWZ19] showed that, given an algorithm for solving linear regression in the g-norm with relative error reg_g, it is possible to compute a set of $O(k \log d)$ columns achieving an approximation ratio of

$$O(k \log k) \cdot \mathsf{reg}_g \cdot \mathsf{mon}_g \cdot \mathsf{ati}_{g,k+1}.$$

for g satisfying the mon_g-monotone and $\mathsf{ati}_{g,t}$-approximate triangle inequality properties. We show that for the slightly restricted family of g of at least lin_g-linear growth, which for example includes all convex g [CW15a], we obtain an improved approximation ratio of

$$O(\sqrt{k}) \cdot \frac{\mathsf{reg}_g \cdot \mathsf{ati}_{g,s+1}}{\mathsf{lin}_g}.$$

Our guarantee matches, and in fact improves a log factor, of the ℓ_1 column subset selection guarantee of [MW21], despite being a far more general result. Furthermore, our bound is tight, in the sense that the $\sqrt{k}$ cannot be improved to a smaller polynomial due to a matching lower bound for ℓ_1 column subset selection [SWZ17]. Our technique for removing the $\log k$ factor in the distortion is general, and can be used to improve prior results for ℓ_p column subset selection as well [CGK+17, DWZ+19, MW21].

[1] We allow for indexing matrices and vectors by arbitrary sets. For example, $\mathbb{R}^S$ is the set of vectors with entries indexed by elements s of S, and $\mathbb{R}^{S \times d}$ is the set of matrices with rows indexed by elements of S and columns indexed by $[d]$.

Theorem 17.1.2 (Improved guarantees for entrywise low rank approximation)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $k \geq 1$. Let $s = O(k)$. Let $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function satisfying the $\mathsf{ati}_{g,t}$-approximate triangle inequality for $t = s + 1$ and the lin_g-linear growth property. Furthermore, suppose that there is an algorithm outputting $\tilde{\mathbf{x}}$ such that

$$\|\mathbf{B}\tilde{\mathbf{x}} - \mathbf{b}\|_g \leq \mathsf{reg}_{g,s} \cdot \min_{\mathbf{x}\in\mathbb{R}^s} \|\mathbf{B}\tilde{\mathbf{x}} - \mathbf{b}\|_g$$

for any $\mathbf{B} \in \mathbb{R}^{n\times s}$ and $\mathbf{b} \in \mathbb{R}^n$. Then, there is an algorithm, Algorithm 10, which outputs a subset $S \subseteq [d]$ of $|S| = O(k(\log d)^2)$ columns and $\mathbf{X} \in \mathbb{R}^{t\times d}$ such that

$$\left\|\mathbf{A} - \mathbf{A}|^S\mathbf{X}\right\|_g \leq O(\sqrt{k})\frac{\mathsf{reg}_{g,O(s\log d)} \cdot \mathsf{ati}_{g,s+1}}{\mathsf{lin}_g} \min_{\mathrm{rank}(\mathbf{A}')\leq k} \|\mathbf{A} - \mathbf{A}'\|_g.$$

Algorithm 10 Column subset selection for M-estimators

input: Input matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, rank k, loss function g.
output: Subset $T \subseteq [d]$ of $O(k \log^2 d)$ columns.

$T_0 \leftarrow [d]$
$s \leftarrow O(k)$
while $|T_l| \geq 1000s$ **do**
  $t_l \leftarrow 160s \log_2 d_l$
  for $t = 1, 2, \ldots, O(\log\log d)$ **do**
    Sample $H \sim \binom{T_l}{t_l}$
    Let $\mathbf{x}^j$ minimize $\min_{\mathbf{x}} \|\mathbf{A}|^H\mathbf{x} - \mathbf{a}^j\|_g$ up to a reg_{g,t_l} factor for each $j \in T_l$
    Let $F_{l,t}$ be the $d_l/960 = |T_l|/960$ columns with smallest regression cost $\|\mathbf{A}|^H\mathbf{x}^j - \mathbf{a}^j\|_g$
    $C_{l,t} \leftarrow \sum_{j\in F_{l,t}} \|\mathbf{A}|^H\mathbf{x}^j - \mathbf{a}^j\|_g$
  Let t^* be the t with smallest $C_{l,t}$
  $T_{l+1} \leftarrow T_l \setminus F_{l,t^*}$

For the important case of the Huber loss, given by

$$H(x) = \begin{cases} |x|^2/2 & \text{if } |x| \leq 1 \\ |x| - 1/2 & \text{if } |x| > 1 \end{cases},$$

we specialize our technique to give the following optimized result:

Theorem 17.1.3 (Entrywise Huber low rank approximation)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $k \geq 1$. There is an algorithm which outputs a subset $S \subseteq [d]$ of $|S| = O(k \log d)$ columns and $\mathbf{X} \in \mathbb{R}^{S\times d}$ such that

$$\left\|\mathbf{A} - \mathbf{A}|^S\mathbf{X}\right\|_H \leq O(k) \min_{\mathrm{rank}(\mathbf{A}')\leq k} \|\mathbf{A} - \mathbf{A}'\|_H,$$

where $\|\cdot\|_H$ denotes the entrywise Huber loss.

The previous best known bound [SWZ19] gave a distortion of $\tilde{O}(k^2)$ for the same number of columns.

For both general entrywise low rank approximation as well as low rank approximation under the Huber loss, our new results are in fact based on constructions of well-conditioned spanning sets in Theorem 3.2.2.

17.1.1 An improved structural result on uniform sampling

We first give a slight more useful form of Theorem 3.2.2 to our setting.

Lemma 17.1.4. Let $\mathbf{A}_* \in \mathbb{R}^{n\times d}$ be a rank k matrix. Then, there exists a set $S \subseteq [d]$ of size $s = O(k)$ such that for every $j \in [d]$,

$$\|(\mathbf{A}_*|^S)^- \mathbf{a}_*^j\|_2^2 \leq O(1).$$

Proof. Since $\mathbf{A}_*$ has rank k, we can write $\mathbf{A}_* = \mathbf{QR}$ for some orthonormal $\mathbf{Q} \in \mathbb{R}^{n\times k}$ and $\mathbf{R} \in \mathbb{R}^{k\times d}$. Then by Theorem 3.2.2, there exists a set $S \subseteq [d]$ of size s such that for every $j \in H \cup \{i\}$, we have that $\|(\mathbf{R}|^S)^- \mathbf{r}^j\|_2^2 \leq O(1)$. The result then follows since

$$\begin{aligned}
\|(\mathbf{A}_*|^S)^- \mathbf{a}_*^j\|_2^2 &= (\mathbf{a}_*^j)^\top (\mathbf{A}_*|^S)^{-\top} (\mathbf{A}_*|^S)^- \mathbf{a}_*^j \\
&= (\mathbf{r}^j)^\top \mathbf{Q}^\top \mathbf{Q} (\mathbf{R}|^S)^{-\top} (\mathbf{R}|^S)^- \mathbf{Q}^\top \mathbf{Q} \mathbf{r}^j \\
&= (\mathbf{r}^j)^\top (\mathbf{R}|^S)^{-\top} (\mathbf{R}|^S)^- \mathbf{r}^j \\
&= \|(\mathbf{R}|^S)^- \mathbf{r}^j\|_2^2.
\end{aligned}$$

□

Using Lemma 17.1.4, we now obtain the following lemma, which gives an improved version of Lemmas 2.1 and 2.2 of [SWZ19].

Lemma 17.1.5. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $\mathbf{A}_* \in \mathbb{R}^{n\times d}$ be any rank k matrix and let $\mathbf{D} = \mathbf{A} - \mathbf{A}_*$. Let $s \geq O(k)$ and let $H \sim \binom{[d]}{2s}$ and let $i \sim [d] \setminus H$. Let $R = R(H \cup \{i\})$ be the set of size s given by Lemma 17.1.4 for $\mathbf{A}_*|^{H\cup\{i\}}$. The following hold:

- With probability at least $1/2$, $i \notin R$
- If $i \notin R$, then there is $\mathbf{x} \in \mathbb{R}^H$ such that

$$\min_{\mathbf{x}\in\mathbb{R}^H} \left\|\mathbf{A}|^H \mathbf{x} - \mathbf{a}^i\right\|_g^2 \leq O(1) \frac{\mathsf{ati}_{g,s+1}^2}{\mathsf{lin}_g^2} \sum_{j\in H\cup\{i\}} \left\|\mathbf{d}^j\right\|_g^2 \tag{17.1}$$

- With probability at least $1/4$ over $H \sim \binom{[d]}{2s}$,

$$|\{i \in [d] \setminus H : i \notin R(H \cup \{i\})\}| \geq \frac{d}{4}$$

Proof. By symmetry, i is a uniformly random index of $H\cup\{i\}$, so $\mathbf{Pr}\{i \notin R\} \geq 1 - s/(2s+1) > 1/2$, which gives the first conclusion.

Let α_j denote the jth entry of $(\mathbf{A}_*|^R)^- \mathbf{a}_*^i$ for each $j \in R$ and $\alpha_j = 0$ for $j \in H \setminus R$. We then have that

$$\min_{\mathbf{x}\in\mathbb{R}^H} \left\|\mathbf{A}|^H \mathbf{x} - \mathbf{a}^i\right\|_g \leq \left\|\sum_{j\in H} \alpha_j \mathbf{a}^j - \mathbf{a}^i\right\|_g$$

$$
\begin{aligned}
&\leq \Big\| \sum_{j \in H} \alpha_j (\mathbf{a}_*^j + \mathbf{d}^j) - (\mathbf{a}_*^i + \mathbf{d}^i) \Big\|_g \\
&= \Big\| \sum_{j \in R} \alpha_j \mathbf{d}^j - \mathbf{d}^i \Big\|_g && \text{since } \mathbf{A}_*|^R (\mathbf{A}_*|^R)^- \mathbf{a}_*^i = \mathbf{a}_*^i \\
&\leq \mathsf{ati}_{g,s+1} \Big(\sum_{j \in R} \|\alpha_j \mathbf{d}^j\|_g + \|\mathbf{d}^i\|_g \Big) && \text{approximate triangle inequality} \\
&\leq \frac{\mathsf{ati}_{g,s+1}}{\mathsf{lin}_g} \Big(\sum_{j \in R} \alpha_j \|\mathbf{d}^j\|_g + \|\mathbf{d}^i\|_g \Big) && \text{at least linear growth} \\
&\leq \frac{\mathsf{ati}_{g,s+1}}{\mathsf{lin}_g} \Big(\Big(\sum_{j \in R} \alpha_j^2 \Big)^{1/2} \Big(\sum_{j \in R} \|\mathbf{d}^j\|_g^2 \Big)^{1/2} + \|\mathbf{d}^i\|_g \Big) && \text{Cauchy–Schwarz} \\
&\leq O(1) \frac{\mathsf{ati}_{g,s+1}}{\mathsf{lin}_g} \Big(\Big(\sum_{j \in R} \|\mathbf{d}^j\|_g^2 \Big)^{1/2} + \|\mathbf{d}^i\|_g \Big).
\end{aligned}
$$

Squaring both sides yields the second conclusion.

The third conclusion follows from the same proof as Lemma 2.2 of [SWZ19]. □

17.1.2 Sharper guarantees for the [SWZ19] algorithm

We now use the result of Lemma 17.1.5 to improve the analysis of the [SWZ19] algorithm.

Level sets

Let $\mathbf{A} = \mathbf{A}_* + \mathbf{\Delta}$, where $\mathbf{A}_*$ is the best rank k approximation in the g-norm. Let the columns of $\mathbf{\Delta}$ be $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \ldots, \boldsymbol{\delta}^d$. To gain fine-grained control over the costs of the columns, we will need to consider a partition of the columns into $O(\log d)$ level sets based on $\|\boldsymbol{\delta}^j\|_g$.

Definition 17.1.6. Let $l \in \mathbb{N}$. Then:

- Let $s = O(k)$ denote the maximum size of an ℓ_2-well-conditioned subset given by Theorem 3.2.2 in k dimensions.
- Let $T_l \subseteq [d]$ denote the subset of columns surviving after the lth round of the algorithm. We assume without loss of generality that $T_l = [d_l]$ for some $d_l \leq d$. Furthermore, we assume without loss of generality that $\|\boldsymbol{\delta}^1\|_g \geq \|\boldsymbol{\delta}^2\|_g \geq \cdots \geq \|\boldsymbol{\delta}^{d_l}\|_g$.
- Let $\mathsf{Res}_l := \sum_{j=d_l/4}^{d_l} \|\boldsymbol{\delta}^j\|_g$ denote the residual cost, after restricting to the surviving columns and after removing the columns with cost in the top quarter.
- Let

$$
R_l^i := \begin{cases} \Big\{ j \in [d_l] \setminus [d_l/4] : \|\boldsymbol{\delta}^j\|_g \leq \frac{1}{d_l^2} \mathsf{Res}_l \Big\} & \text{if } i = \infty \\ \Big\{ j \in [d_l] \setminus [d_l/4] : 2^{-i} \cdot \mathsf{Res}_l < \|\boldsymbol{\delta}^j\|_g \leq 2^{-i+1} \cdot \mathsf{Res}_l \Big\} & \text{if } 0 < i < 2 \log_2 d_l \end{cases}
$$

Recall that our goal is to show that with constant probability, the $d_l/80$ columns with the smallest regression cost when fit on $\mathbf{A}|^H$ each have a cost of at most $O(\sqrt{k})\mathsf{Res}_l/d_l$. We first show that we may assume with out loss of generality that R_l^∞ is small in cardinality.

Lemma 17.1.7. If $|R_l^\infty| > d_l/4$, then with probability at least $1/6$ over the randomness of H,

$$\left|\left\{j \in T_l : \min_{\mathbf{x}\in\mathbb{R}^H} \left\||\mathbf{A}|^H\mathbf{x} - \mathbf{a}^j\right\|_g \le \frac{1}{d_l}\mathsf{Res}_l\right\}\right| \ge \frac{1}{80}d_l$$

Proof. Note that $\mathbf{E}|R_l^\infty \cap H| \ge 20s$. By Chernoff bounds, with probability at least $99/100$, we have that $|R_l^\infty \cap H| \ge 4s \ge 2k$. Then by conditioning on the size of $R_l^\infty \cap H$, we can apply the same proof from Lemma 2.5 of [SWZ19] restricted to R_l^∞ to show that with probability at least $1/5 - 1/100 \ge 1/6$ over the randomness of H,

$$\left|\left\{j \in T_l : \min_{\mathbf{x}\in\mathbb{R}^H} \left\||\mathbf{A}|^H\mathbf{x} - \mathbf{a}^j\right\|_g \le \frac{|H|}{d_l^2}\mathsf{Res}_l\right\}\right| \ge \frac{1}{20}|R_l^\infty| \ge \frac{1}{20}\cdot\frac{d_l}{4} = \frac{1}{80}d_l.$$

Note that $|H| \le d_l$, which gives the claimed result. □

By Lemma 17.1.7, we may assume that $|R_l^\infty| \le d_l/4$. In this case, we show that we must have many columns which belong to a large level set.

Lemma 17.1.8. Suppose that $|R_l^\infty| \le d_l/4$. Then, at least $d_l/4$ columns belong to a level set R_l^i such that $|R_l^i| \ge d_l/8\log_2 d_l$.

Proof. Note that the number of columns which can belong in a level set of size less than $d_l/8\log_2 d_l$ is less than

$$2(\log_2 d_l)\cdot\frac{d_l}{8\log_2 d_l} = \frac{d_l}{4}$$

since there are only $2\log_2 d_l$ level sets. Since there are at most $d_l/4$ columns in R_l^∞ and at most $d_l/4$ that are excluded for being in the top quarter, we conclude as desired. □

Fitting a constant fraction of columns

We will now show that we can fit a constant fraction of columns in a large level set with small cost. We first show the following lemma for a single level set:

Lemma 17.1.9. Let $i \in [2\log_2 d_l]$ be such that $|R_l^i| \ge d_l/8\log_2 d_l$. Then, with probability at least $1/6$, there are at least $|R_l^i|/20$ indices $j \in R_l^i$ such that there exists $\mathbf{x}$ satisfying

$$\min_{\mathbf{x}\in\mathbb{R}^H}\left\||\mathbf{A}|^H\mathbf{x} - \mathbf{a}^{j'}\right\|_g \le O(\sqrt{s})\frac{\mathsf{ati}_{g,s+1}}{\mathsf{lin}_g}\frac{\mathsf{Res}_l}{2^i}$$

Proof. The proof is based on adapting Lemmas 2.3, 2.4, and 2.5 of [SWZ19].

Note that $\mathbf{E}|R_l^i \cap H| \ge 20s$. By Chernoff bounds, with probability at least $99/100$, we have that $|R_l^\infty \cap H| \ge 4s$. We condition on this event. Then, let $H' \subseteq R_l^i \cap H$ be a uniformly random subset of $R_l^i \cap H$ of size $2s$. Then by Markov's inequality,

$$\Pr_{H'}\left\{\sum_{j\in H'}\left\|\boldsymbol{\delta}^j\right\|_g^2 \ge 40\frac{s}{|R_l^i|}\sum_{j\in R_l^i}\left\|\boldsymbol{\delta}^j\right\|_g^2\right\} \le \frac{\mathbf{E}\left[\sum_{j\in H'}\left\|\boldsymbol{\delta}^j\right\|_g^2\right]}{40\frac{s}{|R_l^i|}\sum_{j\in R_l^i}\left\|\boldsymbol{\delta}^j\right\|_g^2} \le \frac{\frac{2s}{|R_l^i|}\sum_{j\in R_l^i}\left\|\boldsymbol{\delta}^j\right\|_g^2}{40\frac{s}{|R_l^i|}\sum_{j\in R_l^i}\left\|\boldsymbol{\delta}^j\right\|_g^2} \le \frac{1}{20}$$

Furthermore, by an averaging argument, we have that

$$\left|\left\{j' \in R_l^i : \left\|\boldsymbol{\delta}^{j'}\right\|_g^2 \geq \frac{5}{|R_l^i|}\sum_{j\in R_l^i}\left\|\boldsymbol{\delta}^j\right\|_g^2\right\}\right| \leq \frac{1}{5}|R_l^i|$$

Now note that H' is a uniformly random subset of R_l^i of size $2s$. Then, by Lemma 17.1.5, we have that with probability at least $1/4$, there are at least $|R_l^i|/4$ indices $j' \in R_l^i$ for which (17.1) holds. Thus, for at least $|R_l^i|/4 - |R_l^i|/5 = |R_l^i|/20$ indices $j' \in R_l^i$, we have that

$$\begin{aligned}
\min_{\mathbf{x}\in\mathbb{R}^{H'}}\left\|\mathbf{A}|^{H'}\mathbf{x} - \mathbf{a}^{j'}\right\|_g^2 &\leq O(1)\frac{\mathsf{ati}_{g,s+1}^2}{\mathsf{lin}_g^2}\sum_{j\in H'\cup\{j'\}}\left\|\boldsymbol{\delta}^j\right\|_g^2 && \text{Lemma 17.1.5}\\
&\leq O(1)\frac{\mathsf{ati}_{g,s+1}^2}{\mathsf{lin}_g^2}\frac{20s+5}{|R_l^i|}\sum_{j\in R_l^i}\left\|\boldsymbol{\delta}^j\right\|_g^2 && \text{Lemma 17.1.9}\\
&\leq O(1)\frac{\mathsf{ati}_{g,s+1}^2}{\mathsf{lin}_g^2}\frac{s}{2^{2i}}\mathsf{Res}_l^2 && \text{Definition 17.1.6}
\end{aligned}$$

By padding $\mathbf{x}$ with zeros on $H \setminus H'$ and taking square roots, we get the desired result. □

Next, we apply an averaging argument to show that if we sum across all large level sets, we fit a constant fraction of columns all d_l with constant probability.

Lemma 17.1.10. Suppose that $|R_l^\infty| \leq d_l/4$. Then with probability at least $1/960$, there is a set of size $F \subseteq [d_l]$ such that $|F| \geq d_l/960$ and

$$\sum_{j\in F}\min_{\mathbf{x}\in\mathbb{R}^H}\|\mathbf{A}|^H\mathbf{x} - \mathbf{a}^j\|_g \leq O(\sqrt{s})\frac{\mathsf{ati}_{g,s+1}}{\mathsf{lin}_g}\cdot\mathsf{Res}_l$$

Proof. By Lemma 17.1.9, for a fixed level set i with $|R_l^i| \geq d_l/8\log_2 d_l$, with probability at least $1/6$, we fit at least $|R_l^i|/20$ columns with cost at most

$$O(\sqrt{s})\frac{\mathsf{ati}_{g,s+1}}{\mathsf{lin}_g}\frac{\mathsf{Res}_l}{2^i}$$

each. Then, let X_i be the random variable that represents the number of such columns in R_l^i, and define

$$X := \sum_{i:|R_l^i|\geq d_l/8\log_2 d_l} X_i$$

Note then that

$$\mathbf{E}[X] \geq \sum_{i:|R_l^i|\geq d_l/8\log_2 d_l}\frac{1}{6}\cdot\frac{1}{20}|R_l^i| \geq \frac{1}{6\cdot 20\cdot 4}d_l = \frac{1}{480}d_l$$

where the last inequality is by Lemma 17.1.8. Then by a standard averaging argument,

$$\frac{1}{480}d_l \leq d_l\cdot\mathbf{Pr}\{X \geq d_l/960\} + \frac{d_l}{960}\mathbf{Pr}\{X < d_l/960\}$$

$$\leq d_l \cdot \mathbf{Pr}\{X \geq d_l/960\} + \frac{d_l}{960}$$

so X is at least $d_l/960$ with probability at least $1/960$. Furthermore, the total cost of all of the columns which are fit well is at most

$$\sum_i O(\sqrt{s}) \frac{\mathsf{ati}_{g,s+1}}{\mathsf{lin}_g} \frac{\mathsf{Res}_l}{2^i} \cdot |R_l^i| \leq O(\sqrt{s}) \frac{\mathsf{ati}_{g,s+1}}{\mathsf{lin}_g} \cdot \mathsf{Res}_l. \qquad \square$$

Proof of Theorem 17.1.2

We now give proofs for the various guarantees of our algorithm.

Proof of Theorem 17.1.2. Note first that the algorithm decreases the size of T_l by a $(1 - 1/960)$ factor at each iteration. Thus, the algorithm makes at most $L = O(\log d)$ iterations of the outer loop. By Lemma 17.1.10, we have a constant probability of success of choosing $d_l/960$ columns such that the total cost is at most

$$O(\sqrt{s}) \frac{\mathsf{ati}_{g,s+1}}{\mathsf{lin}_g} \cdot \mathsf{Res}_l.$$

Since we repeat $O(\log L) = O(\log \log d)$ times and use an reg_{g,t_l}-approximate regression algorithm, we with probability at least $1 - 1/100L$, we find $d_l/960$ columns $F_l \subseteq T_l$ and corresponding coefficients $\mathbf{X}$ such that

$$\left\| \mathbf{A}|^{F_l} - \mathbf{A}|^{S_l} \mathbf{X} \right\|_g \leq O(\sqrt{s}) \frac{\mathsf{reg}_{g,t_l} \cdot \mathsf{ati}_{g,s+1}}{\mathsf{lin}_g} \cdot \mathsf{Res}_l.$$

Thus, our total cost is

$$\sum_{l=1}^{O(\log d)} O(\sqrt{s}) \frac{\mathsf{reg}_{g,t_l} \cdot \mathsf{ati}_{g,s+1}}{\mathsf{lin}_g} \cdot \mathsf{Res}_l.$$

Finally, as argued in [SWZ19, MW21], we show that $\sum_l \mathsf{Res}_l = O(\|\mathbf{\Delta}\|_g)$. Note that if a column j contributes to Res_l, then it must be in the bottom $3/4$ fraction of the $\|\boldsymbol{\delta}^j\|_g$ in round l. Then since the bottom $1/960$ fraction of $\|\boldsymbol{\delta}^j\|_g$ is fitted and removed in each round, $\|\boldsymbol{\delta}^j\|_g$ can only contribute to Res_l in $O(1)$ rounds. Thus, the sum is bounded by $O(1) \sum_j \|\boldsymbol{\delta}^j\|_g = O(\|\mathbf{\Delta}\|_g)$.

The total number of columns selected is $O(s \log d)$ in each of the $O(\log d)$ rounds, for a total of $O(s \log^2 d)$. $\square$

17.2 Huber column subset selection

For the important case of the Huber loss, the result of Theorem 17.1.2 only yields a distortion of $\tilde{O}(k^{3/2})$, due to a k factor loss from the approximate triangle inequality term. We further optimize our argument specifically for the Huber loss and obtain a distortion of $O(k)$ instead.

Theorem 17.1.3 (Entrywise Huber low rank approximation)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $k \geq 1$. There is an algorithm which outputs a subset $S \subseteq [d]$ of $|S| = O(k \log d)$ columns and $\mathbf{X} \in \mathbb{R}^{S\times d}$ such that

$$\left\|\mathbf{A} - \mathbf{A}|^S\mathbf{X}\right\|_H \leq O(k) \min_{\mathrm{rank}(\mathbf{A}')\leq k} \|\mathbf{A} - \mathbf{A}'\|_H,$$

where $\|\cdot\|_H$ denotes the entrywise Huber loss.

Our improvement comes from the following structural result, which yields Theorem 17.1.3 when combined with Theorem 17.4.3:

Lemma 17.2.1. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $\mathbf{A}_*$ denote the optimal rank k approximation to $\mathbf{A}$ in the entrywise Huber norm. Then, there exists a set $S \subseteq [d]$ of $O(k)$ columns of $\mathbf{A}$ and $\mathbf{X} \in \mathbb{R}^{S\times d}$ such that

$$\left\|\mathbf{A} - \mathbf{A}|^S\mathbf{X}\right\|_H \leq O(d)\|\mathbf{A} - \mathbf{A}_*\|_H.$$

Proof. Let $S \subseteq [d]$ be an ℓ_2-well-conditioned coreset for the columns of $\mathbf{A}_*$, given by Theorem 3.2.2. For each $j \notin S$, we let the jth column of $\mathbf{X}$ be the coefficient vector for fitting $\mathbf{a}_*^j$ by $\mathbf{A}_*|^S$.

Following [CW15b, Lemma 37], we have that for any $\mathbf{x} \in \mathbb{R}^d$,

$$H(\|\mathbf{x}\|_2) \leq \frac{\|\mathbf{x}\|_2^2}{\|\mathbf{x}\|_\infty^2} H(\|\mathbf{x}\|_\infty) = \sum_{j=1}^d \frac{\mathbf{x}_i^2}{\|\mathbf{x}\|_\infty^2} H(\|\mathbf{x}\|_\infty) \leq \sum_{j=1}^d H(\mathbf{x}_i) = \|\mathbf{x}\|_H.$$

Then,

$$\begin{aligned}\left\|\mathbf{A} - \mathbf{A}|^S\mathbf{X}\right\|_H &= \left\|(\mathbf{A}^* + \boldsymbol{\Delta}) - (\mathbf{A}^* + \boldsymbol{\Delta})|^S\mathbf{X}\right\|_H \\ &= \left\|\boldsymbol{\Delta} - \boldsymbol{\Delta}|^S\mathbf{X}\right\|_H \\ &\leq O(1)(\|\boldsymbol{\Delta}\|_H + \left\|\boldsymbol{\Delta}|^S\mathbf{X}\right\|_H)\end{aligned}$$

so it suffices to bound $\left\|\boldsymbol{\Delta}|^S\mathbf{X}\right\|_H$. We have

$$\begin{aligned}\left\|\boldsymbol{\Delta}|^S\mathbf{X}\right\|_H &= \sum_{j=1}^d \sum_{i=1}^n H(\mathbf{e}_i^\top \boldsymbol{\Delta}|^S\mathbf{x}^j) \\ &\leq \sum_{j=1}^d \sum_{i=1}^n H(\left\|\mathbf{e}_i^\top \boldsymbol{\Delta}|^S\right\|_2 \left\|\mathbf{x}^j\right\|_2) && \text{Cauchy–Schwarz} \\ &\leq O(1) \sum_{j=1}^d \sum_{i=1}^n H(\left\|\mathbf{e}_i^\top \boldsymbol{\Delta}|^S\right\|_2) \\ &\leq O(1) \sum_{j=1}^d \sum_{i=1}^n \left\|\mathbf{e}_i^\top \boldsymbol{\Delta}|^S\right\|_H \\ &\leq O(1) \sum_{j=1}^d \left\|\boldsymbol{\Delta}|^S\right\|_H\end{aligned}$$

$$\leq O(d)\|\mathbf{\Delta}\|_H$$

as claimed. □

17.3 Algorithms for the entrywise ℓ_p norm

For $p \neq 2$, efficient bicriteria approximations for entrywise ℓ_p low rank approximation were obtained in a line of work initiated by [SWZ17], who studied the case of $p = 1$. For other $p \neq 2$, [CGK+17, DWZ+19] gave algorithms selecting $O(k \log d)$ columns achieving a distortion of $\tilde{O}(k^{1/p})$ for $p < 2$ and $\tilde{O}(k^{1-1/p})$ for $p > 2$, and a hardness result showing that any approximation spanned by k columns must have distortion at least

$$\Omega(k^{1-1/p}) \tag{17.2}$$

Perhaps surprisingly, [MW21] then showed that the lower bound of (17.2) could be circumvented when $p < 2$, by giving an algorithm which selected $\tilde{O}(k \log d)$ columns and achieved a distortion of $\tilde{O}(k^{1/p-1/2})$. Note that this does not contradict the lower bound, since the hardness result of (17.2) applies only when *exactly* k columns are selected. It was also shown that this result was optimal for such bicriteria algorithms, with a lower bound ruling out $k^{1/p-1/2-o(1)}$ approximations for any algorithm selecting $\tilde{O}(k)$ columns, based on a result of [SWZ17] which ruled out $k^{1/2-o(1)}$ approximations for any set of $\text{poly}(k)$ columns for $p = 1$.

Unfortunately, the algorithmic result of [MW21] uses p-stable random variables [Nol20] which only exist for $p \leq 2$, and similar improvements were not given for $p > 2$. Similarly, the hardness results also rely on specific properties of $p < 2$, and do not apply to $p > 2$. This motivates the following question:

Question 17.3.1. What distortions are possible for entrywise ℓ_p low rank approximation, if $O(k \log d)$ columns can be selected?

Our main result for entrywise ℓ_p low rank approximation is an algorithm which achieves the natural analogue of the algorithmic result of [MW21], which circumvents (17.2):

Theorem 17.3.2 (Entrywise ℓ_p low rank approximation [WY23a])**.** Let $p \in [2, \infty]$, let $\mathbf{A} \in \mathbb{R}^{n \times d}$, and let $k \geq 1$. There is an algorithm which outputs a subset $S \subseteq [d]$ of $O(k \log d)$ columns and $\mathbf{X} \in \mathbb{R}^{S \times d}$ such that

$$\left\|\mathbf{A} - \mathbf{A}|^S\mathbf{X}\right\|_{p,p} \leq O(k^{1/2-1/p}) \min_{\text{rank}(\mathbf{A}') \leq k} \|\mathbf{A} - \mathbf{A}'\|_{p,p}.$$

17.3.1 Improved existential result

Our main improvement comes from the following lemma, which is inspired by the techniques of [MW21]. Rather than relying p-stable variables as in [MW21, Theorem 2.4], we instead make use of ℓ_p Lewis weights in our argument. We refer to Chapter 6 for a comprehensive discussion of ℓ_p Lewis weights and their applications.

Lemma 17.3.3. Let $2 \le p \le \infty$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $\mathbf{A}_*$ denote the optimal rank k approximation to $\mathbf{A}$ in the entrywise ℓ_p norm. Then, there exists a set $S \subseteq [d]$ of $O(k)$ columns of $\mathbf{A}$ and $\mathbf{R} \in \mathbb{R}^{k\times d}$ such that

$$\left\|\mathbf{A} - \mathbf{A}|^S\mathbf{R}\right\|_{p,p} \le O(k^{1/2-1/p})\|\mathbf{A} - \mathbf{A}_*\|_{p,p}. \tag{17.3}$$

Proof. Let $\mathbf{A}_* = \mathbf{U}\mathbf{V}^\top$ for some $\mathbf{U} \in \mathbb{R}^{n\times k}$ and $\mathbf{V}^\top \in \mathbb{R}^{k\times d}$. Now let $\mathbf{w}$ be the ℓ_p Lewis weights of $\mathbf{V}$ and let $\hat{\mathbf{X}}$ minimize

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times k}} \left\|(\mathbf{A} - \mathbf{X}\mathbf{V}^\top)\mathbf{W}^{1/2-1/p}\right\|_{p,2}$$

up to a factor of 2. We have

$$\begin{aligned}
\|\mathbf{A} - \hat{\mathbf{X}}\mathbf{V}^\top\|_{p,p} &\le \|\mathbf{A} - \mathbf{U}\mathbf{V}^\top\|_{p,p} + \|\mathbf{U}\mathbf{V}^\top - \hat{\mathbf{X}}\mathbf{V}^\top\|_{p,p} \\
&\le \|\mathbf{A} - \mathbf{U}\mathbf{V}^\top\|_{p,p} + \|(\mathbf{U}\mathbf{V}^\top - \hat{\mathbf{X}}\mathbf{V}^\top)\mathbf{W}^{1/2-1/p}\|_{p,2} && \text{Lemma 6.2.3} \\
&\le \|\mathbf{A} - \mathbf{U}\mathbf{V}^\top\|_{p,p} + \|(\mathbf{U}\mathbf{V}^\top - \mathbf{A})\mathbf{W}^{1/2-1/p}\|_{p,2} \\
&\qquad + \|(\mathbf{A} - \hat{\mathbf{X}}\mathbf{V}^\top)\mathbf{W}^{1/2-1/p}\|_{p,2} \\
&\le \|\mathbf{A} - \mathbf{U}\mathbf{V}^\top\|_{p,p} + 3\|(\mathbf{U}\mathbf{V}^\top - \mathbf{A})\mathbf{W}^{1/2-1/p}\|_{p,2} && \text{near optimality} \\
&\le \|\mathbf{A} - \mathbf{U}\mathbf{V}^\top\|_{p,p} + 3k^{1/2-1/p}\|\mathbf{U}\mathbf{V}^\top - \mathbf{A}\|_{p,p} && \text{Lemma 6.2.2} \\
&= O(k^{1/2-1/p})\|\mathbf{A} - \mathbf{U}\mathbf{V}^\top\|_{p,p}.
\end{aligned}$$

Thus, we have reduced the problem to an ℓ_2 problem, at a cost of $O(k^{1/2-1/p})$ distortion. Lemma 27 of [CW15b] then shows that if $\mathbf{S}^\top$ is an ℓ_2 sparsifier for $\mathbf{V}^\top\mathbf{W}^{1/2-1/p}$ which samples $O(k)$ columns (see [SWZ19, Lemma C.25], based on [BSS12, Theorem 3.1]), then a minimizer $\hat{\mathbf{U}}$ of

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times k}} \left\|(\mathbf{A} - \mathbf{X}\mathbf{V}^\top)\mathbf{W}^{1/2-1/p}\mathbf{S}^\top\right\|_{p,2}$$

satisfies

$$\left\|(\mathbf{A} - \hat{\mathbf{U}}\mathbf{V}^\top)\mathbf{W}^{1/2-1/p}\right\|_{p,2} \le 2\min_{\mathbf{X}\in\mathbb{R}^{n\times k}} \left\|(\mathbf{A} - \mathbf{X}\mathbf{V}^\top)\mathbf{W}^{1/2-1/p}\right\|_{p,2}.$$

It follows that

$$\|\mathbf{A} - \hat{\mathbf{U}}\mathbf{V}^\top\|_{p,p} \le O(k^{1/2-1/p})\|\mathbf{A} - \mathbf{U}\mathbf{V}^\top\|_{p,p}.$$

Finally, note that $\hat{\mathbf{U}}$ can be written as

$$\hat{\mathbf{U}} = \mathbf{A}\mathbf{W}^{1/2-1/p}\mathbf{S}^\top(\mathbf{V}^\top\mathbf{W}^{1/2-1/p}\mathbf{S}^\top)^-.$$

Thus, there exists an $O(k^{1/2-1/p})$-approximate solution with a left factor formed by $O(k)$ columns of $\mathbf{A}$. □

With Lemma 17.3.3 in hand, we can now apply the existential-to-algorithmic reduction of Theorem 17.4.3 to obtain the following:

Theorem 17.3.4. Let $2 \le p < \infty$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and let $k \ge 1$. There is an algorithm which outputs a subset $S \subseteq [d]$ of $|S| = O(k \log d)$ columns and $\mathbf{X} \in \mathbb{R}^{S\times d}$ such that

$$\left\|\mathbf{A} - \mathbf{A}|^S\mathbf{X}\right\|_{p,p} \le O(k^{1/2-1/p}) \min_{\mathrm{rank}(\hat{\mathbf{A}})\le k} \|\mathbf{A} - \hat{\mathbf{A}}\|_{p,p}.$$

We note that by setting $p = O(\log n)$, we also obtain a result for $p = \infty$.

Theorem 17.3.5. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and let $k \geq 1$. There is an algorithm which outputs a subset $S \subseteq [d]$ of $|S| = O(k \log d)$ columns and $\mathbf{X} \in \mathbb{R}^{S \times d}$ such that

$$\left\|\mathbf{A} - \mathbf{A}|^S \mathbf{X}\right\|_{\infty,\infty} \leq O(k^{1/2}) \min_{\mathrm{rank}(\hat{\mathbf{A}}) \leq k} \|\mathbf{A} - \hat{\mathbf{A}}\|_{\infty,\infty}.$$

17.3.2 Lower bounds

For $p = \infty$, we show that Theorem 17.3.2 is tight by showing that any set of at most $\mathrm{poly}(k)$ columns cannot achieve a distortion better than $k^{1/2-o(1)}$.

Our result is based on a variation on the ideas of Theorem 1.4 of [SWZ17].

Definition 17.3.6 (Hard distribution)**.** Let $c \geq 1$ be any constant and let $r = k^c$. We then define a distribution $\mathcal{D}$ over $(k + 2^r) \times r$ matrix as follows. We let the first k rows have entries drawn independently from $\mathcal{N}(0, \mathbf{I}_r)$ and scaled by k, and we let the last 2^r rows be the 2^r vectors in $\{\pm 1\}^r$.

We will argue that with high probability, no matrix in the column span of $r/2$ columns of $\mathbf{A} \sim \mathcal{D}$ can approximate $\mathbf{A}$ by better than a $\sqrt{k}$ factor. The optimal rank k approximation of any matrix drawn from the distribution in Definition 17.3.6 has ℓ_∞ has cost at most 1, by setting the rank k approximation to be the first k rows:

Lemma 17.3.7. Let $\mathbf{A} \sim \mathcal{D}$ for $\mathcal{D}$ defined in Definition 17.3.6. Then, with probability 1,

$$\min_{\mathrm{rank}(\hat{\mathbf{A}}) \leq k} \|\mathbf{A} - \hat{\mathbf{A}}\|_{\infty,\infty} \leq 1.$$

Furthermore, the addition of the 2^r hypercube vectors to the matrix gives the following property:

Lemma 17.3.8. Let $S \subseteq [r]$. Then, for any $\mathbf{X} \in \mathbb{R}^{S \times r}$,

$$\left\|\mathbf{A} - \mathbf{A}|^S \mathbf{X}\right\|_{\infty,\infty} \geq \max_{j=1}^{r} \|\mathbf{X}\mathbf{e}_j\|_1 - 1$$

Proof. Let $j \in [r]$. Then, there exists a row i of $\mathbf{A}|^S$ such that for each $j' \in S$, $\mathbf{A}_{i,j'} = \mathrm{sgn}(\mathbf{X}_{j',j})$, since $\mathbf{A}$ contains all sign vectors. Thus,

$$\mathbf{e}_i^\top \mathbf{A}|^S \mathbf{X}\mathbf{e}_j = \sum_{j' \in S} \mathbf{A}_{i,j'} \mathbf{X}_{j',j} = \sum_{j' \in S} \mathrm{sgn}(\mathbf{X}_{j',j}) \mathbf{X}_{j',j} = \|\mathbf{X}\mathbf{e}_j\|_1.$$

On the other hand, $\mathbf{A}$ has absolute value at most 1 on this coordinate, thus yielding the claim. □

With these insights in hand, the proof now essentially follows that of [SWZ17, Theorem G.28]; it is shown in [SWZ17] that if $\mathbf{x} \in \mathbb{R}^S$ fits the first k rows well in ℓ_1 norm, then it must satisfy $\|\mathbf{x}\|_1 = \Omega(k^{0.5-o(1)})$. Since we scale the first k rows by k, this means that we either have a high ℓ_∞ cost in the first k rows, or a high ℓ_∞ cost in the bottom 2^r rows.

Theorem 17.3.9. Let $\alpha \in (0, 0.5)$, $k \in \mathbb{N}$, and $r = \text{poly}(k)$. Then, there exists a $(k + r) \times r$ matrix $\mathbf{A}$ such that

$$\min_{\text{rank}(\hat{\mathbf{A}}) \le k} \|\mathbf{A} - \hat{\mathbf{A}}\|_{\infty,\infty} \le 1$$

and for any $S \subseteq [r]$ with $|S| \le r/2$,

$$\min_{\mathbf{X} \in \mathbb{R}^{S \times r}} \|\mathbf{A} - \mathbf{A}|^S \mathbf{X}\|_{\infty,\infty} \ge \Omega(k^{0.5-\alpha}).$$

Proof. The proof closely follows [SWZ17, Theorem G.28]. For $\mathbf{B} \sim \mathcal{N}(0, 1)^{k \times s}$ and scalars $\beta, \gamma > 0$, we say the event $\mathcal{E}(\mathbf{B}, \beta, \gamma)$ holds if

- $\|\mathbf{B}\|_2 \le O(\sqrt{s})$
- $\mathbf{Bx}$ has at most $O(k/\log k)$ coordinates with absolute value at least $\Omega(1/\log k)$, whenever $\|\mathbf{x}\|_1 \le O(k^{\gamma})$ and $\|\mathbf{x}\|_\infty \le O(k^{-\beta})$

(see [SWZ17, Definition G.19]). It is shown in [SWZ17, Lemma G.20] that if $k \le s \le \text{poly}(k)$, $\beta > \gamma > 0$, and $\beta + \gamma < 1$, then $\mathbf{Pr}\{\mathcal{E}(\mathbf{B}, \beta, \gamma)\} \ge 1 - \exp(-\Theta(k))$. We will apply this to the first k rows $\mathbf{A}|_{[k]}$ of $\mathbf{A}$ scaled down by k, as well as to restrictions $\mathbf{A}|^S_{[k]}$ of these rows to columns $S \subseteq [r]$.

It is shown in [SWZ17, Claim G.29] that for any $S \subseteq [r]$,

$$\mathbf{Pr}\left\{ \mathcal{E}\left(\frac{1}{k}\mathbf{A}|^S_{[k]}, 0.5 + \alpha/2, 0.5 - \alpha\right) \middle| \mathcal{E}\left(\frac{1}{k}\mathbf{A}|_{[k]}, 0.5 + \alpha/2, 0.5 - \alpha\right) \right\} = 1$$

We thus condition on $\mathcal{E}(\frac{1}{k}\mathbf{A}|_{[k]}, 0.5 + \alpha/2, 0.5 - \alpha)$, which implies $\mathcal{E}(\frac{1}{k}\mathbf{A}|^S_{[k]}, 0.5 + \alpha/2, 0.5 - \alpha)$ for every $S \subseteq [r]$. Then by [SWZ17, Lemma G.22], for any $S \subseteq [r]$ of size at most $r/2$, with probability at least $1 - \exp(-\Theta(rk))$, a constant fraction of the $r/2$ remaining rows $l \in [r] \setminus S$ satisfies that

$$\min_{\mathbf{x} \in \mathbb{R}^S} \left\| \frac{1}{k}\mathbf{A}_{[k]}|^S \mathbf{x} - \mathbf{A}\mathbf{e}_l \right\|_1 + \|\mathbf{x}\|_1 = \Omega(k^{0.5-\alpha})$$

By relating the ℓ_1 and ℓ_∞ norms up to a factor of k for the first term and by using Lemma 17.3.8 for the second term, this gives a lower bound of $\Omega(k^{0.5-\alpha})$ on some entry of $\mathbf{A} - \mathbf{A}|^S \mathbf{X}$ for any $\mathbf{X}$, for this fixed S. The failure rate of $\exp(-\Theta(rk))$ is small enough for us to union bound over all choices of $S \subseteq [r]$ of size at most $r/2$, thus proving the theorem. □

17.4 Reduction from existential to algorithmic column subset selection

We show an improvement and generalization of the argument of [MW21], which shows that an existential result showing the existence of $s = s(k)$ columns with a distortion of $\kappa(d)$ on any $n \times d$ instance for rank k approximation implies an algorithmic version which selects $O(s \log d)$ columns with a distortion of $O(\kappa(2s + 1))$. Note that the number of columns can only depend on k, whereas the distortion can depend on d.

Definition 17.4.1. Let $\mathbf{A} = \mathbf{A}_* + \mathbf{\Delta}$, where $\mathbf{A}_*$ is the best rank k approximation in the entrywise g norm, that is,

$$\|\mathbf{\Delta}\|_g = \min_{\text{rank}(\hat{\mathbf{A}}) \leq k} \left\|\mathbf{A} - \hat{\mathbf{A}}\right\|_g.$$

Let the columns of $\mathbf{\Delta}$ be $\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \ldots, \boldsymbol{\delta}^d$.

Definition 17.4.2. Let $l \in \mathbb{N}$. Then:

- Let $s(k)$ denote the maximum size of a set of columns S for any $n \times d$ instance $\mathbf{B}$ for rank k approximation in the entrywise g-norm that can achieve a $\kappa(d)$ approximation, that is, there exists a set $S \subseteq [d]$ such that

$$\min_{\mathbf{X} \in \mathbb{R}^{S \times d}} \left\|\mathbf{B} - \mathbf{B}|^S \mathbf{X}\right\|_g \leq \kappa(d) \|\mathbf{\Delta}\|_g \tag{17.4}$$

- Let $T_l \subseteq [d]$ denote the subset of columns surviving after the lth round of the algorithm. We assume without loss of generality that $T_l = [d_l]$ for some $d_l \leq d$. Furthermore, we assume without loss of generality that $\|\boldsymbol{\delta}^1\|_g \geq \|\boldsymbol{\delta}^2\|_g \geq \cdots \geq \|\boldsymbol{\delta}^{d_l}\|_g$.
- Let $\mathsf{Res}_l := \sum_{j=d_l/4}^{d_l} \|\boldsymbol{\delta}^j\|_g$ denote the residual cost, after restricting to the surviving columns and after removing the columns with cost in the top quarter.

Algorithm 11 Column subset selection for M-estimators

input: Input matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, rank k, loss function g, parameter s.
output: Subset $T \subseteq [d]$ of $O(s \log d)$ columns.

$T_0 \leftarrow [d]$
while $|T_l| \geq 1000s$ **do**
  $t_l \leftarrow 30s$
  for $t = 1, 2, \ldots, O(\log \log d)$ **do**
    Sample $H \sim \binom{T_l}{t_l}$
    Let $\mathbf{x}^j$ minimize $\min_{\mathbf{x}} \|\mathbf{A}|^H \mathbf{x} - \mathbf{a}^j\|_g$ up to a reg_{g,t_l} factor for each $j \in T_l$
    Let $F_{l,t}$ be the $d_l/20 = |T_l|/20$ columns with smallest regression cost $\|\mathbf{A}|^H \mathbf{x}^j - \mathbf{a}^j\|_g$
    $C_{l,t} \leftarrow \sum_{j \in F_{l,t}} \|\mathbf{A}|^H \mathbf{x}^j - \mathbf{a}^j\|_g$
  Let t^* be the t with smallest $C_{l,t}$
  $T_{l+1} \leftarrow T_l \setminus F_{l,t^*}$

Theorem 17.4.3 (Generalization and improvement of [MW21])**.** Consider the definitions in Definition 17.4.2. Suppose that there is an algorithm outputting $\tilde{\mathbf{x}}$ such that

$$\|\mathbf{B}\tilde{\mathbf{x}} - \mathbf{b}\|_g \leq \mathsf{reg}_{g,s} \cdot \min_{\mathbf{x} \in \mathbb{R}^s} \|\mathbf{B}\tilde{\mathbf{x}} - \mathbf{b}\|_g$$

for any $\mathbf{B} \in \mathbb{R}^{n \times s}$ and $\mathbf{b} \in \mathbb{R}^n$. Then, Algorithm 11 outputs a subset $S \subseteq [d]$ of $|S| = O(s \log d)$ columns and $\mathbf{X} \in \mathbb{R}^{S \times d}$ such that

$$\left\|\mathbf{A} - \mathbf{A}|^S \mathbf{X}\right\|_g \leq O(\kappa) \mathsf{reg}_{g,O(s)} \min_{\text{rank}(\hat{\mathbf{A}}) \leq k} \|\mathbf{A} - \hat{\mathbf{A}}\|_g$$

We present the following main lemma, which follows [MW21, Claim 2.6] but also makes some additional improvements to remove a log factor:

Lemma 17.4.4. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Let $s = s(k)$ and $\kappa = \kappa(2s+1)$. Let $H \sim \binom{[d]}{2s}$ and let $i \sim [d] \setminus H$. Then,

$$\mathbf{Pr}\left\{\min_{\mathbf{x}\in\mathbb{R}^H} \left\|\mathbf{a}^i - \mathbf{A}|^H\mathbf{x}\right\|_g \le \frac{600\kappa}{d_l}\mathsf{Res}_l\right\} \ge \frac{1}{10}$$

Proof. Let $G \coloneqq [d_l] \setminus [d_l/4]$. Note that $\mathbf{E}|G\cap H| \ge 20s$. By Chernoff bounds, with probability at least $99/100$, we have that $|G\cap H| \ge 4s$. We conditioned on this event.

Let H' be a uniformly random subset of $G\cap H$ of size $2s$. Let $R = R(H'\cup\{i\})$ be the set of $s(k)$ columns satisfying (17.4). Then by Markov's inequality,

$$\mathop{\mathbf{Pr}}_{H'}\left\{\sum_{j\in H'}\left\|\boldsymbol{\delta}^j\right\|_g \ge 20\frac{s}{|G|}\sum_{j\in G}\left\|\boldsymbol{\delta}^j\right\|_g\right\} \le \frac{\mathbf{E}_{H'}\left[\sum_{j\in H'}\left\|\boldsymbol{\delta}^j\right\|_g\right]}{20\frac{s}{|G|}\sum_{j\in G}\left\|\boldsymbol{\delta}^j\right\|_g} \le \frac{1}{10}$$

and similarly,

$$\mathop{\mathbf{Pr}}_{i}\left\{\left\|\boldsymbol{\delta}^i\right\|_g \ge \frac{10}{|G|}\sum_{j\in G}\left\|\boldsymbol{\delta}^j\right\|_g\right\} \le \frac{\mathbf{E}_i\left[\left\|\boldsymbol{\delta}^i\right\|_g\right]}{\frac{5}{|G|}\sum_{j\in G}\left\|\boldsymbol{\delta}^j\right\|_g} \le \frac{1}{10}$$

Now note that conditioned on the choice of $H'\cup\{i\}$, i is a uniformly random element of $H'\cup\{i\}$, so $\mathbf{Pr}\{i\notin R\} \ge 1/2$. Furthermore,

$$\min_{\mathbf{X}\in\mathbb{R}^{R\times(2s+1)}}\left\|\mathbf{A}|^{H'\cup\{i\}} - \mathbf{A}|^R\mathbf{X}\right\|_g \le \kappa \min_{\mathrm{rank}(\hat{\mathbf{A}})\le k}\left\|\mathbf{A}|^{H'\cup\{i\}} - \hat{\mathbf{A}}\right\|_g \le \kappa\cdot\left\|\Delta|^{H'\cup\{i\}}\right\|_g$$

so by Markov's inequality,

$$\min_{\mathbf{x}\in\mathbb{R}^R}\left\|\mathbf{a}^i - \mathbf{A}|^R\mathbf{x}\right\| \le \frac{10\kappa}{s}\left\|\Delta|^{H'\cup\{i\}}\right\|_g$$

with probability at least $9/10$. By a union bound, we have that with probability at least

$$1 - \frac{1}{100} - \frac{1}{10} - \frac{1}{10} - \frac{1}{10} \ge \frac{1}{10},$$

we have

$$\min_{\mathbf{x}\in\mathbb{R}^R}\left\|\mathbf{a}^i - \mathbf{A}|^R\mathbf{x}\right\|_g \le \frac{10\kappa}{s}\left(\frac{10}{|G|}\sum_{j\in G}\left\|\boldsymbol{\delta}^j\right\|_g + 20\frac{s}{|G|}\sum_{j\in G}\left\|\boldsymbol{\delta}^j\right\|_g\right) \le \frac{400\kappa}{|G|}\sum_{j\in G}\left\|\boldsymbol{\delta}^j\right\|_g.$$

To conclude, note that $|G| = d_l - d_l/4 = 3d_l/4$ and that we can pad $\mathbf{x}$ with zeros on coordinates in $H\setminus R$. □

We then just mimic the proof of Theorem 17.1.2 to complete the proof.

Proof of Theorem 17.4.3. Note first that the algorithm decreases the size of T_l by a $(1 - 1/20)$ factor at each iteration. Thus, the algorithm makes at most $L = O(\log d)$ iterations of the outer loop. By averaging Lemma 17.4.4 over the $3d_l/4$ bottom columns, we have a probability of at least $1/20$ of choosing $d_l/20$ columns such that the total cost is at most

$$O(\kappa) \cdot \mathsf{Res}_l.$$

Since we repeat $O(\log L) = O(\log \log d)$ times and use an reg_{g,t_l}-approximate regression algorithm, we with probability at least $1 - 1/100L$, we find $d_l/20$ columns $F_l \subseteq T_l$ and corresponding coefficients $\mathbf{X}$ such that

$$\left\|\mathbf{A}|^{F_l} - \mathbf{A}|^{S_l}\mathbf{X}\right\|_g \leq O(\kappa)\mathsf{reg}_{g,t_l}\mathsf{Res}_l.$$

Thus, our total cost is

$$\sum_{l=1}^{O(\log d)} O(\kappa)\mathsf{reg}_{g,t_l}\mathsf{Res}_l.$$

Finally, as argued in [SWZ19, MW21], we show that $\sum_l \mathsf{Res}_l = O(\|\mathbf{\Delta}\|_g)$. Note that if a column j contributes to Res_l, then it must be in the bottom $3/4$ fraction of the $\|\boldsymbol{\delta}^j\|_g$ in round l. Then since the bottom $1/20$ fraction of $\|\boldsymbol{\delta}^j\|_g$ is fitted and removed in each round, $\|\boldsymbol{\delta}^j\|_g$ can only contribute to Res_l in $O(1)$ rounds. Thus, the sum is bounded by $O(1)\sum_j \|\boldsymbol{\delta}^j\|_g = O(\|\mathbf{\Delta}\|_g)$.

The total number of columns selected is $O(s)$ in each of the $O(\log d)$ rounds, for a total of $O(s \log d)$. □

Chapter 18

Spectral low rank approximation for sparse singular vectors [WY22b]

In this section, we study algorithms for the classical problem of low rank approximation under the *spectral norm*.

Definition 18.0.1. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then, we define the spectral norm of $\mathbf{A}$ to be

$$\|\mathbf{A}\|_2 \coloneqq \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

Because the spectral norm is unitarily invariant, the classical Eckhart–Young–Minrky theorem [EY36, Mir60] shows that the singular value decomposition yields the optimal rank k approximation, for all k. While the singular value decomposition (SVD) can be expensive to compute for large matrices, the recent results in randomized numerical linear algebra have achieved substantial developments in fast approximation algorithms for the SVD, culminating in the following result of [MM15]:

Theorem 18.0.2 (Approximate spectral SVD [MM15])**.** Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then, there is an algorithm which computes a rank k orthogonal projection matrix $\mathbf{P} \in \mathbb{R}^{d \times d}$ such that

$$\|\mathbf{A} - \mathbf{A}\mathbf{P}\|_2 \leq (1 + \varepsilon) \min_{\mathrm{rank}(\mathbf{A}') \leq k} \|\mathbf{A} - \mathbf{A}'\|_2$$

which runs in time at most $O(\varepsilon^{-1/2}\, \mathsf{nnz}(\mathbf{A}) k \log d)$.

A natural question is whether this running time can be improved or not, under natural assumptions. One common assumption which often arises in practice is to assume that the top k singular vectors of $\mathbf{A}$ are *sparse*, i.e., there are only s nonzero values in the singular vectors. This scenario is a phenomenon known as *localization* of eigenvectors, and occurs frequently in many applications [HBCY21, ZYC+21], for example in quantum many-body problems [LVW09, NH15] and network analysis [PC18].

This question was studied in the work of [HBCY21] and a followup work of [ZYC+21], which studied algorithms for computing eigenvectors in symmetric matrices with localized eigenvectors. In [HBCY21], the authors study an algorithm for finding a small submatrix containing the supports

of the leading eigenvectors by greedily adding rows and columns without formal guarantees, and [ZYC+21] seek to improve this approach using reinforcement learning techniques.

In our work of [WY22b], we obtain one of the first provable speedups over [MM15] under a sparse singular vector assumption:

Theorem 18.0.3 (Approximate spectral SVD for sparse singular vectors [WY22b])**.** Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ whose top k left and right singular vectors have at most s nonzero entries. Then, there is an algorithm which computes a rank k orthogonal projection matrix $\mathbf{P} \in \mathbb{R}^{d\times d}$ such that

$$\|\mathbf{A} - \mathbf{AP}\|_2 \leq (1+\varepsilon) \min_{\mathrm{rank}(\mathbf{A}')\leq k} \|\mathbf{A} - \mathbf{A}'\|_2$$

which runs in time at most

$$O\left(\frac{\mathsf{nnz}(\mathbf{A})}{\sqrt{\varepsilon}} + \frac{n}{\varepsilon}\right) \log \frac{sdk \log n}{\varepsilon} + \mathrm{poly}\left(s, k, \frac{1}{\varepsilon}, \log n\right).$$

At a high level, our idea is to first identify a set of around $O(sk)$ (or a slightly larger number of) coordinates which contains the support of the top k singular vectors, at which point we can just output the SVD of this submatrix, padded with zeros. Thus, the difficulty lies in identifying this subset of $O(sk)$ coordinates. The work of [MM15] shows that if we know the value of the $(k+1)$th singular value $\boldsymbol{\sigma}_{k+1}$, then we can use a Chebyshev polynomial approximation of degree roughly $q = 1/\sqrt{\varepsilon}$ to identify singular vectors with singular values larger than $(1+\varepsilon)\boldsymbol{\sigma}_{k+1}$ from the vectors $\mathbf{Ag}, (\mathbf{AA}^\top)\mathbf{Ag}, \ldots, (\mathbf{AA}^\top)^q\mathbf{Ag}$, known as the *Krylov subspace*. Thus, the main problem to tackle is to find an algorithm to determine the value of $\boldsymbol{\sigma}_{k+1}$, up to a $(1+\varepsilon)$ factor. To do this, we introduce a two-stage algorithm. In the first step, we identify the value of $\boldsymbol{\sigma}_{k+1}$ up to a factor of $(1+\sqrt{\varepsilon})$ using a combination of naive power iteration together with an efficient binary searching technique over the singular values. In the second step, we know the value of $\boldsymbol{\sigma}_{k+1}$ up to a value of $(1+\sqrt{\varepsilon})$, and thus we can afford to make $1/\sqrt{\varepsilon}$ guesses to the value of $\boldsymbol{\sigma}_{k+1}$ in powers of $(1+\varepsilon)$, and add $O(sk)$ entries to our superset of the support of the sparse singular vectors for each one of the $1/\sqrt{\varepsilon}$ guesses. Then, one of these guesses will guess the right value of $\boldsymbol{\sigma}_{k+1}$, and in total, the size of our support superset is just $O(sk/\sqrt{\varepsilon})$. Our result of Theorem 18.0.3 follows.

Remark 18.0.4. Note that the problem we study differs from the related problem of *sparse low rank approximation*, where we seek a low rank approximation with sparse factors for an arbitrary matrix $\mathbf{A}$. Unlike our problem, this problem is intractable under standard complexity assumptions [MWA06, Mag17, CPR16, LRG23].

18.1 Technical overview

Our first idea is that with a budget of $\mathsf{nnz}(\mathbf{A})/\sqrt{\varepsilon}$ running time, we can run naïve power method for $1/\sqrt{\varepsilon}$ iterations initialized with a single random Gaussian vector $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_d)$ to compute $(\mathbf{AA}^\top)^{1/\sqrt{\varepsilon}}\mathbf{Ag}$. Using the SVD $\mathbf{A} = \mathbf{U\Sigma V}^\top$ of $\mathbf{A}$, we may write this as $\mathbf{U\Sigma}^{O(1/\sqrt{\varepsilon})}\mathbf{V}^\top\mathbf{g}$. Then by the rotational invariance of the Gaussian, this random vector is distributed as a random linear combination of the left singular vectors of $\mathbf{A}$, where the ith left singular vector is scaled by roughly $\sigma_i^{1/\sqrt{\varepsilon}}$. Then if $i \in [k]$ is such that $\sigma_i \geq (1+\sqrt{\varepsilon})\sigma_{k+1}$, then $\sigma_i^{1/\sqrt{\varepsilon}}$ is a constant factor larger than

$\sigma_{k+1}^{1/\sqrt{\varepsilon}}$, so the s entries corresponding to the ith left singular vector stand out in $(\mathbf{A}\mathbf{A}^\top)^{1/\sqrt{\varepsilon}}\mathbf{A}\mathbf{g}$. Thus, selecting the top sk entries with largest absolute value in $(\mathbf{A}\mathbf{A}^\top)^{1/\sqrt{\varepsilon}}\mathbf{A}\mathbf{g}$ retrieves a superset of the support of the left singular vectors with singular value σ_i for which $\sigma_i \geq (1+\sqrt{\varepsilon})\sigma_{k+1}$. We can repeat on the right side as well to obtain the support of the large right singular vectors.

The above approach is enough to find the large singular components with singular value at least $(1+\sqrt{\varepsilon})\sigma_{k+1}$, but we must find singular values all the way down to $(1+\varepsilon)\sigma_{k+1}$ for a $(1+\varepsilon)$ relative error approximation. To do this, we use the approach of [MM15] of using Chebyshev polynomials, which, given a location parameter α, gives us a degree $1/\sqrt{\varepsilon}$ polynomial p for which $p(x)$, for all $x \geq (1+\varepsilon)\alpha$, is a constant times greater than any $p(x)$ for $x \leq \alpha$ (see Lemma 18.2.9 for the mathematical statement). If we knew the location $\alpha = \sigma_{k+1}$, then we could compute $p(\mathbf{A})\mathbf{g}$ in $\mathsf{nnz}(\mathbf{A})/\sqrt{\varepsilon}$ time and use the same approach as before to find the support of all singular components i for which $\sigma_i \geq (1+\varepsilon)\sigma_{k+1}$. The challenge, of course, is that we do not know σ_{k+1}.

We first show how to find the value of σ_{k+1} up to a $(1+\sqrt{\varepsilon})$ factor. To this end, we first show that if σ_i for $i \in [k]$ is large, i.e. $\sigma_i \geq (1+\sqrt{\varepsilon})\sigma_{k+1}$, then we can find σ_i up to a $(1+\varepsilon)$ factor using the set of sk large coordinates on the left and right located before, using the power method. However, note that we do not know for which i this is true. That is, if we let $\hat{\mathbf{A}}$ be the $sk \times sk$ submatrix supported on the large coordinates identified using the power method, we expect the large singular values of $\hat{\mathbf{A}}$ to be good estimates of the large singular values of $\mathbf{A}$, but we do not know which of them are large enough to actually be good estimates.

To address this, let $i \in [k]$, and first note that $\sigma_i(\hat{\mathbf{A}})$ is always a lower bound on $\sigma_i(\mathbf{A})$ by the Cauchy interlacing theorem. Furthermore, suppose that $\hat{\mathbf{B}}$ is a rank $i-1$ approximation to $\hat{\mathbf{A}}$. Then, $\left\|\mathbf{A}-\hat{\mathbf{B}}\right\|_2$ serves as an upper bound for $\sigma_i(\mathbf{A})$, as

$$\sigma_i(\mathbf{A}) = \min_{\text{rank } i-1\ \mathbf{C}} \|\mathbf{A}-\mathbf{C}\|_2 \leq \left\|\mathbf{A}-\hat{\mathbf{B}}\right\|_2.$$

We show that for $i \in [k]$ such that $\sigma_i \geq (1+\sqrt{\varepsilon})\sigma_{k+1}$, these are good lower and upper bounds on the singular value $\sigma_i(\mathbf{A})$, i.e., they are within $(1+\varepsilon)$ factors of each other. Furthermore, they can both be computed in time roughly

$$\frac{\mathsf{nnz}(\mathbf{A})}{\sqrt{\varepsilon}} + \mathrm{poly}(s,k,\varepsilon^{-1}).$$

Thus, we have an extremely efficient way to certify our estimates to the singular values $\sigma_i(\mathbf{A})$, if they are large enough. We then consider the following binary search strategy over the singular values: if the upper and lower bounds are within $(1+\varepsilon)$ factors of each other, then we keep searching lower, and otherwise, we search higher. If the $\sigma_{i_*}(\mathbf{A})$ found is such that $\sigma_{i_*}(\mathbf{A}) \geq (1+\sqrt{\varepsilon})\sigma_{k+1}(\mathbf{A})$, then the top i_* singular components are found in the initial power method step accurately enough so that $\left\|\mathbf{A}-\hat{\mathbf{B}}\right\|_2$ is close to $\sigma_{i_*+1}(\mathbf{A}) \leq (1+\sqrt{\varepsilon})\sigma_{k+1}(\mathbf{A})$, where $\hat{\mathbf{B}}$ is a rank i_* approximation $\hat{\mathbf{B}}$ of $\hat{\mathbf{A}}$. Otherwise, $\sigma_{i_*}(\mathbf{A})$ itself is within a $(1+\sqrt{\varepsilon})$ factor of $\sigma_{k+1}(\mathbf{A})$.

Now that we are within a $(1+\sqrt{\varepsilon})$ factor of $\sigma_{k+1}(\mathbf{A})$, we just need $1/\sqrt{\varepsilon}$ guesses in powers of $(1+\varepsilon)$ in order to guess $\sigma_{k+1}(\mathbf{A})$ up to a factor of $(1+\varepsilon)$. We can in fact afford to guess all of these locations α, compute the corresponding Chebyshev polynomial p, compute $p(\mathbf{A})\mathbf{g}$ from precomputed Krylov iterates, select the top sk entries, and then add the entries to the support that we consider.

With the support superset in hand, we finish the algorithm by performing an approximate SVD on this submatrix. Our full discussion can be found in Section 18.2.

18.2 Proof of Theorem 18.0.3

In this section, we discuss our results on performing an approximate SVD with relative spectral norm error, when we are promised that the input matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ has top k left and right singular vectors that are s-sparse.

18.2.1 Approximating singular components

To carry out our plan as described in the introduction (Section 18.1), we first calculate the magnitude of coordinates that we need to capture in order to achieve a relative error spectral approximation. We follow [MM15] and make use of the fact that additive Frobenius norm low rank approximation implies additive spectral norm low rank approximation, originally due to [Gu15].

Lemma 18.2.1 (Theorem 3.4 of [Gu15])**.** For any $\mathbf{A} \in \mathbb{R}^{n \times d}$, let $\mathbf{B} \in \mathbb{R}^{n \times d}$ be any rank k matrix satisfying $\|\mathbf{A} - \mathbf{B}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \eta$. Then,

$$\|\mathbf{A} - \mathbf{B}\|_2^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_2^2 + \eta.$$

By the above result, it suffices to find a rank k matrix $\mathbf{B}$ such that

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \varepsilon\sigma_{k+1}^2.$$

Using this, we show that it suffices to find all coordinates of the top left singular vectors $\mathbf{U}\mathbf{e}_j$ such that

$$|\mathbf{e}_i^\top \mathbf{U}\mathbf{e}_j| \geq \frac{\varepsilon}{k\sqrt{sr}} \frac{\sigma_{k+1}}{\sigma_j},$$

and similarly, all coordinates of the top right singular vectors $\mathbf{V}\mathbf{e}_j$ such that

$$|\mathbf{e}_i^\top \mathbf{V}\mathbf{e}_j| \geq \frac{\varepsilon}{k\sqrt{sr}} \frac{\sigma_{k+1}}{\sigma_j}.$$

Lemma 18.2.2. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ have rank r with singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, and let $\varepsilon \in (0, 1/2)$. Let $S \subset [n]$ and $T \subset [d]$ be a set of coordinates such that

$$S \supset \bigcup_{j \in [r]} \left\{ i \in [n] : |\mathbf{e}_i^\top \mathbf{U}\mathbf{e}_j| \geq \frac{\varepsilon}{k\sqrt{sr}} \frac{\sigma_{k+1}}{\sigma_j} \right\}$$

$$T \supset \bigcup_{j \in [r]} \left\{ i \in [d] : |\mathbf{e}_i^\top \mathbf{V}\mathbf{e}_j| \geq \frac{\varepsilon}{k\sqrt{sr}} \frac{\sigma_{k+1}}{\sigma_j} \right\}$$

Let $\mathbf{B}$ be a rank k matrix such that

$$\|\mathbf{P}_S \mathbf{A} \mathbf{P}_T - \mathbf{B}\|_F^2 \leq \min_{\text{rank } k\ \mathbf{C}} \|\mathbf{P}_S \mathbf{A} \mathbf{P}_T - \mathbf{C}\|_F^2 + \eta.$$

Then,

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \le \|\mathbf{A}-\mathbf{A}_k\|_F^2 + 8\varepsilon\sigma_{k+1}^2 + \eta.$$

Proof. Note first that

$$\|\mathbf{A}-\mathbf{A}_k\|_F^2 = \sum_{t=k+1}^{r} \sigma_t^2(\mathbf{A}) \le \sum_{t=k+1}^{r} \sigma_{k+1}^2(\mathbf{A}) \le \sigma_{k+1}^2(\mathbf{A})r.$$

Then,

$$\begin{aligned}
\|\mathbf{A}-\mathbf{B}\|_F^2 &= \|\mathbf{A}-\mathbf{P}_S\mathbf{A}\mathbf{P}_T\|_F^2 + \|\mathbf{P}_S\mathbf{A}\mathbf{P}_T-\mathbf{B}\|_F^2 \\
&\overset{(1)}{\le} \|\mathbf{A}-\mathbf{P}_S\mathbf{A}\mathbf{P}_T\|_F^2 + \|\mathbf{P}_S\mathbf{A}\mathbf{P}_T-\mathbf{A}_k\|_F^2 + \eta \\
&\overset{(2)}{=} \|\mathbf{A}-\mathbf{P}_S\mathbf{A}\mathbf{P}_T\|_F^2 + \|\mathbf{P}_S\mathbf{A}\mathbf{P}_T-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F^2 + \|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F^2 + \eta \\
&\overset{(3)}{=} \|\mathbf{A}-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F^2 + \|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F^2 + \eta \\
&\overset{(4)}{\le} \left(\|\mathbf{A}-\mathbf{A}_k\|_F + \|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F\right)^2 + \|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F^2 + \eta \\
&= \|\mathbf{A}-\mathbf{A}_k\|_F^2 + 2\|\mathbf{A}-\mathbf{A}_k\|_F\|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F + 2\|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F^2 + \eta \\
&\le \|\mathbf{A}-\mathbf{A}_k\|_F^2 + 2\sigma_{k+1}\sqrt{r}\|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F + 2\|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F^2 + \eta
\end{aligned}$$

In the above, the inequality (1) is due to the approximate optimality of $\mathbf{B}$, the identities (2) and (3) are by the Pythagorean theorem, and inequality (4) is the triangle inequality. Finally, we calculate that

$$\begin{aligned}
\|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F &\le \|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\|_F + \|\mathbf{P}_S\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F \\
&= \left\|\sum_{j=1}^{k}\sigma_j\mathbf{P}_{\overline{S}}\mathbf{U}\mathbf{e}_j(\mathbf{V}\mathbf{e}_j)^\top\right\|_F + \left\|\sum_{j=1}^{k}\sigma_j\mathbf{P}_S\mathbf{U}\mathbf{e}_j(\mathbf{V}\mathbf{e}_j)^\top\mathbf{P}_{\overline{T}}\right\|_F \\
&\le \sum_{j=1}^{k}\sigma_j\|\mathbf{P}_{\overline{S}}\mathbf{U}\mathbf{e}_j\|_2\|\mathbf{V}\mathbf{e}_j\|_2 + \sigma_j\|\mathbf{P}_S\mathbf{U}\mathbf{e}_j\|_2\|\mathbf{P}_{\overline{T}}\mathbf{V}\mathbf{e}_j\|_2 \\
&\le \sum_{j=1}^{k} 2\sigma_j\left(\frac{\varepsilon}{k\sqrt{sr}}\frac{\sigma_{k+1}}{\sigma_j}\right)\sqrt{s} \\
&= \frac{2\varepsilon}{\sqrt{r}}\sigma_{k+1}
\end{aligned}$$

so the previous bound is

$$\begin{aligned}
\|\mathbf{A}-\mathbf{B}\|_F^2 &\le \|\mathbf{A}-\mathbf{A}_k\|_F^2 + 4\sigma_{k+1}\sqrt{r}\|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F + 2\|\mathbf{A}_k-\mathbf{P}_S\mathbf{A}_k\mathbf{P}_T\|_F^2 + \eta \\
&\le \|\mathbf{A}-\mathbf{A}_k\|_F^2 + 4\varepsilon\sigma_{k+1}^2 + \frac{8\varepsilon^2}{r}\sigma_{k+1}^2 + \eta \\
&\le \|\mathbf{A}-\mathbf{A}_k\|_F^2 + 8\varepsilon\sigma_{k+1}^2 + \eta.
\end{aligned}$$

□

18.2.2 Finding the support of singular vectors with large singular value

We next show how to find all large coordinates of singular vectors whose singular values σ_j are at least a $(1+\sqrt{\varepsilon})$ factor larger than σ_{k+1}. By the results of the previous section, we seek to find all of the large coordinates of the top sparse singular vectors, which have absolute value at least

$$\tau_j \coloneqq \frac{\varepsilon}{k\sqrt{sr}}\frac{\sigma_{k+1}}{\sigma_j}$$

for the jth singular vector.

Our identification of the large coordinates of the top sparse singular vectors starts from the standard analysis of the power method (see also, e.g., the overview of [MM15]). If we run power method starting from a random Gaussian vector $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_d)$, that is, we compute $(\mathbf{A}\mathbf{A}^\top)^q\mathbf{A}\mathbf{g}$ for some $q \in \mathbb{N}$, then we retrieve a random Gaussian linear combination of the left singular vectors $\mathbf{U}\mathbf{e}_j$, each scaled by σ_j^{2q+1}. This is a simple consequence of the rotational invariance of the Gaussian:

Lemma 18.2.3. Let $\mathbf{g}' \sim \mathcal{N}(0, \mathbf{I}_d)$ and let $q \in \mathbb{N}$. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a rank r matrix and let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be its singular value decomposition. Then, $(\mathbf{A}\mathbf{A}^\top)^q\mathbf{A}\mathbf{g}'$ has the same distribution as

$$\mathbf{U}\mathbf{\Sigma}^{2q+1}\mathbf{g} = \sum_{j=1}^{r} \mathbf{g}_j\sigma_j^{2q+1}\mathbf{U}\mathbf{e}_j$$

for $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_r)$.

Note then that for $\sigma_j \geq (1+\sqrt{\varepsilon})\sigma_{k+1}$, the jth singular vector is scaled more than the $(k+1)$-st singular vector by a factor of at least $(\sigma_j/\sigma_{k+1})^{2q+1}$. For q roughly order $1/\sqrt{\varepsilon}$, this separates all large coordinates of the jth singular vector from the coordinates of the $(k+1)$-st singular vector.

Lemma 18.2.4. For

$$q = O\left(\frac{1}{\sqrt{\varepsilon}}\log\frac{sk^2\sqrt{sr\log n}}{\varepsilon}\right),$$

the sk coordinates of $(\mathbf{A}\mathbf{A}^\top)^q\mathbf{A}\mathbf{g}$ with largest absolute value are guaranteed to contain all entries $i \in [n]$ for which there exists a $j \in [k]$ with $\sigma_j \geq (1+\sqrt{\varepsilon})\sigma_{k+1}$ and

$$\left|\mathbf{e}_i^\top\mathbf{U}\mathbf{e}_j\right| \geq \tau_j.$$

Proof. For

$$q = O\left(\frac{1}{\sqrt{\varepsilon}}\log\frac{sk^2\sqrt{sr\log n}}{\varepsilon}\right),$$

the blow up factor $(\sigma_j/\sigma_{k+1})^{2q+1}$ is at least

$$\left(\frac{\sigma_j}{\sigma_{k+1}}\right)^{2q+1} \geq (1+\sqrt{\varepsilon})^{2q}\frac{\sigma_j}{\sigma_{k+1}} = \Theta\left(\frac{sk^2\sqrt{sr\log n}}{\varepsilon}\right)\frac{\sigma_j}{\sigma_{k+1}} = \Theta\left(\frac{sk\sqrt{\log n}}{\tau_j}\right)$$

for the jth singular component. The time required to compute this vector $(\mathbf{A}\mathbf{A}^\top)^q\mathbf{A}\mathbf{g}$ is

$$O\left(\frac{\mathsf{nnz}(\mathbf{A})}{\sqrt{\varepsilon}}\log\frac{sk^2\sqrt{sr\log n}}{\varepsilon}\right) = O\left(\frac{\mathsf{nnz}(\mathbf{A})}{\sqrt{\varepsilon}}\log\frac{srk\log n}{\varepsilon}\right)$$

Now note that for each $i \in [n]$, we have that

$$\mathbf{e}_i^\top \mathbf{U}\mathbf{\Sigma}^{2q+1}\mathbf{g} \sim \mathcal{N}\Big(0, \big\|\mathbf{e}_i^\top \mathbf{U}\mathbf{\Sigma}^{2q+1}\big\|_2^2\Big).$$

Since the maximum absolute value among n Gaussians is $O(\sqrt{\log n})$ with constant probability, we have

$$\big|\mathbf{e}_i^\top \mathbf{U}\mathbf{\Sigma}^{2q+1}\mathbf{g}\big| \le O(\sqrt{\log n})\big\|\mathbf{e}_i^\top \mathbf{U}\mathbf{\Sigma}^{2q+1}\big\|_2.$$

Furthermore, if we consider all i in the support of the top k singular vectors, which is at most sk coordinates, then the minimum absolute value among the sk Gaussians is

$$\big|\mathbf{e}_i^\top \mathbf{U}\mathbf{\Sigma}^{2q+1}\mathbf{g}\big| \ge \Omega\left(\frac{1}{sk}\right)\big\|\mathbf{e}_i^\top \mathbf{U}\mathbf{\Sigma}^{2q+1}\big\|_2.$$

Now consider a coordinate $i \in [n]$ such that

$$\big|\mathbf{e}_i^\top \mathbf{U}\mathbf{e}_j\big| \ge \tau_j$$

for some $j \in [k]$ such that $\sigma_j \ge (1+\sqrt{\varepsilon})\sigma_{k+1}$. Then by the previous results,

$$\begin{aligned}
\big|\mathbf{e}_i^\top \mathbf{U}\mathbf{\Sigma}^{2q+1}\mathbf{g}\big| &\ge \Omega\left(\frac{1}{sk}\right)\big\|\mathbf{e}_i^\top \mathbf{U}\mathbf{\Sigma}^{2q+1}\big\|_2 \\
&\ge \Omega\left(\frac{1}{sk}\right)\sigma_j^{2q+1}\tau_j \\
&= \Omega\left(\frac{1}{sk}\right)\sigma_{k+1}^{2q+1}\left(\frac{\sigma_j}{\sigma_{k+1}}\right)^{2q+1}\tau_j \\
&\ge \Omega\left(\frac{1}{sk}\right)\sigma_{k+1}^{2q+1}\Theta\left(\frac{sk\sqrt{\log n}}{\tau_j}\right)\tau_j \\
&= \Omega\Big(\sigma_{k+1}^{2q+1}\sqrt{\log n}\Big).
\end{aligned}$$

On the other hand, for any $i \in [n]$ that is outside of the at most sk coordinates of the support of the top k singular vectors, then

$$\big|\mathbf{e}_i^\top \mathbf{U}\mathbf{\Sigma}^{2q+1}\mathbf{g}\big| \le O(\sqrt{\log n})\big\|\mathbf{e}_i^\top \mathbf{U}\mathbf{\Sigma}^{2q+1}\big\|_2 \le O(\sigma_{k+1}^{2q+1}\sqrt{\log n}).$$

We thus conclude as desired. □

In other words, we can identify a set of sk coordinates that contains all large entries of left singular vectors j for which $\sigma_j \ge (1+\sqrt{\varepsilon})\sigma_{k+1}$. Repeating for the right singular vectors, we may identify the sets S and T as required by Lemma 18.2.2.

18.2.3 Approximating large singular values

Our next task is to compute the singular values of $\mathbf{A}$ with $\sigma_j(\mathbf{A}) \ge (1+\sqrt{\varepsilon})\sigma_{k+1}(\mathbf{A})$, up to $(1+\varepsilon)$ factors. We first show that approximating the singular values of $\mathbf{P}_S\mathbf{A}\mathbf{P}_T$ directly approximates the singular values of $\mathbf{A}$, when the singular values are sufficiently large.

Lemma 18.2.5. Let m be the number of singular values of $\mathbf{A}$ such that $\sigma_j(\mathbf{A}) \geq (1+\sqrt{\varepsilon})\sigma_{k+1}(\mathbf{A})$. Let $S \subset [n]$ and $T \subset [d]$ be sets satisfying the hypotheses of Lemma 18.2.2. Then for each $l \in [m]$,

$$(1-8\varepsilon)\sigma_l^2(\mathbf{A}) \leq \sigma_l^2(\mathbf{P}_S\mathbf{A}\mathbf{P}_T) \leq \sigma_l^2(\mathbf{A}).$$

Proof. Recall the Cauchy interlacing theorem:

Theorem 18.2.6 (Cauchy interlacing theorem)**.** Let $\mathbf{M}$ be a symmetric matrix and let $\mathbf{N}$ be a principal submatrix of size $l \times l$. Then for all $j \in [l]$,

$$\lambda_j(\mathbf{M}) \geq \lambda_j(\mathbf{N}) \geq \lambda_{n-l+j}(\mathbf{M}).$$

Then applying the interlacing theorem to $\mathbf{M} = \mathbf{A}\mathbf{A}^\top$ and $\mathbf{N} = \mathbf{P}_S\mathbf{A}\mathbf{A}^\top\mathbf{P}_S^\top$, we find that the singular values of $\mathbf{P}_S\mathbf{A}$ uniformly bound the top sk singular values of $\mathbf{A}$ from below, and similarly, the singular values of $\mathbf{P}_S\mathbf{A}\mathbf{P}_T$ uniformly bound the singular values of $\mathbf{P}_S\mathbf{A}$ from below. We thus have that

$$\sigma_j(\mathbf{A}) \geq \sigma_j(\mathbf{P}_S\mathbf{A}) \geq \sigma_j(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)$$

for all $j \in [sk]$. Furthermore, we know by Lemma 18.2.2 that for each $l \in [m]$,

$$\|\mathbf{A}-\mathbf{A}_l\|_F^2 \leq \|\mathbf{A}-(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l\|_F^2 \leq \|\mathbf{A}-\mathbf{A}_l\|_F^2 + 8\varepsilon\sigma_{l+1}(\mathbf{A})^2 \tag{18.1}$$

where $(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l$ is the best rank l approximation $\mathbf{P}_S\mathbf{A}\mathbf{P}_T$. Now note that

$$\|\mathbf{A}\|_F^2 - \|\mathbf{A}-\mathbf{A}_l\|_F^2 = \|\mathbf{A}_l\|_F^2$$

and

$$\begin{aligned}\langle \mathbf{A}-(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l, (\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l\rangle &= \langle \mathbf{A}-\mathbf{P}_S\mathbf{A}\mathbf{P}_T, (\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l\rangle \\ &\quad + \langle \mathbf{P}_S\mathbf{A}\mathbf{P}_T-(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l, (\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l\rangle = 0\end{aligned}$$

so

$$\|\mathbf{A}\|_F^2 - \|\mathbf{A}-(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l\|_F^2 = \|(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l\|_F^2$$

by the Pythagorean theorem. Then subtracting the inequalities of Equation 18.1 from $\|\mathbf{A}\|_F^2$, we have that

$$\|\mathbf{A}_l\|_F^2 - 8\varepsilon\sigma_{l+1}(\mathbf{A})^2 \leq \|(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l\|_F^2 \leq \|\mathbf{A}_l\|_F^2.$$

Then,

$$\begin{aligned}\sigma_l^2(\mathbf{P}_S\mathbf{A}\mathbf{P}_T) &= \|(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_l\|_F^2 - \|(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)_{l-1}\|_F^2 \\ &\geq \|\mathbf{A}_l\|_F^2 - 8\varepsilon\sigma_{l+1}(\mathbf{A})^2 - \|\mathbf{A}_{l-1}\|_F^2 \\ &= \sigma_l^2(\mathbf{A}) - 8\varepsilon\sigma_{l+1}(\mathbf{A})^2 \\ &\geq (1-8\varepsilon)\sigma_l^2(\mathbf{A})\end{aligned}$$

as desired. □

We may use the existing results of [MM15] to find $(1+\varepsilon)$ factor approximations to the top k singular values of $\mathbf{P}_S\mathbf{A}\mathbf{P}_T$ in time

$$O\left(\frac{\mathsf{nnz}(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)k}{\sqrt{\varepsilon}}\log(sk)\right) = O\left(\frac{s^2k^3}{\sqrt{\varepsilon}}\log(sk)\right).$$

However, note that given estimates for the singular values of $\mathbf{P}_S\mathbf{A}\mathbf{P}_T$, we do not know which ones are within a $(1+\varepsilon)$ factor of the singular values of $\mathbf{A}$, since we do not know the number m of singular values j with $\sigma_j(\mathbf{A}) \geq (1+\sqrt{\varepsilon})\sigma_{k+1}(\mathbf{A})$. However, by the Cauchy interlacing theorem, the singular values of $\mathbf{P}_S\mathbf{A}\mathbf{P}_T$ are always a lower bound on the singular values of $\mathbf{A}$, so it suffices to compute an upper bound for the singular values of $\mathbf{A}$ that are at most a $(1+\varepsilon)$ factor larger than the lower bound. We obtain such an upper bound on the singular values of $\mathbf{A}$ by approximating $\|\mathbf{A}-\mathbf{B}\|_2$ for a rank l matrix $\mathbf{B}$. Indeed, if $\mathbf{B}$ is rank l, then

$$\|\mathbf{A}-\mathbf{B}\|_2^2 \geq \min_{\text{rank } l\ \mathbf{C}} \|\mathbf{A}-\mathbf{C}\|_F^2 = \sigma_{l+1}(\mathbf{A})^2.$$

This idea is executed in the following lemma.

Lemma 18.2.7. Let $S \subset [n]$ and $T \subset [d]$ be sets of size sk each that satisfy the hypotheses of Lemma 18.2.2. Given such S and T and an index $j \in [k]$, there is a randomized algorithm that runs in time

$$O\left(\frac{\mathsf{nnz}(\mathbf{A}) + s^2k^3}{\sqrt{\varepsilon}}\log(sk)\right)$$

and outputs numbers U and L such that

$$L \leq \sigma_j^2(\mathbf{A}) \leq U$$

with probability at least 0.99. Furthermore, if $j \in [m]$, where m is the number of singular values j with $\sigma_j \geq (1+\sqrt{\varepsilon})\sigma_{k+1}$, we have that

$$\frac{U}{L} \leq \frac{1+10\varepsilon}{1-9\varepsilon} \leq 1+20\varepsilon.$$

Proof. We first show how to obtain the lower bound L. By the Cauchy interlacing theorem (as in Lemma 18.2.5), we have that

$$\sigma_j(\mathbf{P}_S\mathbf{A}\mathbf{P}_T) \leq \sigma_j(\mathbf{A}).$$

Then by the randomized block Krylov algorithm of [MM15] (see Theorem 18.0.2), we may find an estimate L to $\sigma_j(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)$ such that

$$(1-\varepsilon)\sigma_j(\mathbf{P}_S\mathbf{A}\mathbf{P}_T) \leq L \leq \sigma_j(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)$$

in time

$$O\left(\frac{\mathsf{nnz}(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)k}{\sqrt{\varepsilon}}\log(sk)\right) = O\left(\frac{s^2k^3}{\sqrt{\varepsilon}}\log(sk)\right).$$

Furthermore, if $j \in [m]$, then by Lemma 18.2.5,

$$L \geq (1-\varepsilon)\sigma_j(\mathbf{P}_S\mathbf{A}\mathbf{P}_T) \geq (1-\varepsilon)(1-8\varepsilon)\sigma_j(\mathbf{A}) \geq (1-9\varepsilon)\sigma_j(\mathbf{A}).$$

For the upper bound, we use the rank j approximation $\mathbf{B}$ obtained by running the randomized block Krylov algorithm of [MM15] on $\mathbf{P}_S\mathbf{A}\mathbf{P}_T$. Note that

$$\sigma_j(\mathbf{A}) = \min_{\text{rank } j\ \mathbf{C}} \|\mathbf{A}-\mathbf{C}\|_2 \le \|\mathbf{A}-\mathbf{B}\|_2$$

for any rank $j-1$ matrix $\mathbf{B}$. By the results of [MM15], we may compute an estimate U such that

$$(1+\varepsilon)\|\mathbf{A}-\mathbf{B}\|_2 \ge U \ge \|\mathbf{A}-\mathbf{B}\|_2$$

in time

$$O\left(\frac{\mathsf{nnz}(\mathbf{A}-\mathbf{B})}{\sqrt{\varepsilon}}\right) = O\left(\frac{\mathsf{nnz}(\mathbf{A})+s^2k^2}{\sqrt{\varepsilon}}\right).$$

Furthermore, for $j \in [m]$, if we find a rank $j-1$ matrix $\mathbf{B}$ such that

$$\begin{aligned}\|\mathbf{P}_S\mathbf{A}\mathbf{P}_T-\mathbf{B}\|_F^2 &\le \min_{\text{rank } j-1\ \mathbf{C}} \|\mathbf{P}_S\mathbf{A}\mathbf{P}_T-\mathbf{C}\|_F^2 + \varepsilon\sigma_j(\mathbf{P}_S\mathbf{A}\mathbf{P}_T)^2 \\ &\le \min_{\text{rank } j-1\ \mathbf{C}} \|\mathbf{P}_S\mathbf{A}\mathbf{P}_T-\mathbf{C}\|_F^2 + \varepsilon\sigma_j(\mathbf{A})^2,\end{aligned}$$

which we can by the results of [MM15] as before, then by Lemma 18.2.2,

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \le \|\mathbf{A}-\mathbf{A}_{j-1}\|_F^2 + 9\varepsilon\sigma_j^2(\mathbf{A}).$$

By Lemma 18.2.1, this implies that

$$\|\mathbf{A}-\mathbf{B}\|_2^2 \le \|\mathbf{A}-\mathbf{A}_{j-1}\|_2^2 + 9\varepsilon\sigma_j^2(\mathbf{A}) = (1+9\varepsilon)\sigma_j^2(\mathbf{A}). \qquad \square$$

We now show how to use the above result to efficiently find a $(1+\sqrt{\varepsilon})$ factor approximation to $\sigma_{k+1}(\mathbf{A})$ using binary search.

Lemma 18.2.8. There is a randomized algorithm that runs in time

$$O\left(\frac{\mathsf{nnz}(\mathbf{A})+s^2k^3}{\sqrt{\varepsilon}}\log(sk)(\log k)\right)$$

that finds a $(1+\sqrt{\varepsilon})$ factor approximation to $\sigma_{k+1}(\mathbf{A})$.

Proof. If $\sigma_k(\mathbf{A}) \ge (1+\sqrt{\varepsilon})\sigma_{k+1}(\mathbf{A})$, then deflating off the top k components already gives a $(1+\varepsilon)$ factor approximation to $\sigma_{k+1}(\mathbf{A})$. Otherwise, we proceed with binary search as follows.

Suppose we consider $j \in [k]$. If the upper and lower bounds for $\sigma_j(\mathbf{A})$ in Lemma 18.2.7 are within a $(1+O(\varepsilon))$ factor, then we know that $\sigma_{k+1}(\mathbf{A})$ is smaller than this, up to a $(1\pm O(\varepsilon))$ factor. On the other hand, if the upper and lower bounds for $\sigma_j(\mathbf{A})$ are further than a $(1+O(\varepsilon))$ factor, then $\sigma_j(\mathbf{A}) \le (1+\sqrt{\varepsilon})\sigma_{k+1}(\mathbf{A})$, since otherwise the upper and lower bounds for $\sigma_j(\mathbf{A})$ would have matched up to a $(1\pm O(\varepsilon))$ factor by the second guarantee of Lemma 18.2.7. Thus, we may use binary search over the at most k singular values in at most $O(\log k)$ calls to the algorithm of Lemma 18.2.7. $\square$

18.2.4 Approximating small singular values

With a $(1+\sqrt{\varepsilon})$ factor approximation to $\sigma_{k+1}(\mathbf{A})$ in hand, we now zoom into the singular values between $\sigma_{k+1}(\mathbf{A})$ and $(1+\sqrt{\varepsilon})\sigma_{k+1}(\mathbf{A})$. We consider partitioning this $(1+\sqrt{\varepsilon})$ factor window into $O(1/\sqrt{\varepsilon})$ buckets that increase in powers of $(1+\varepsilon)$, that is

$$L, L(1+\varepsilon), L(1+\varepsilon)^2, L(1+\varepsilon)^3, \ldots, L(1+\varepsilon)^{O(1/\sqrt{\varepsilon})} = (1+\sqrt{\varepsilon})L$$

where L is a lower bound on $\sigma_{k+1}(\mathbf{A})$, up to a $(1+\sqrt{\varepsilon})$ factor. Our idea now is to simply enumerate over these $O(1/\sqrt{\varepsilon})$ guesses to a $(1\pm\varepsilon)$-approximation of $\sigma_{k+1}(\mathbf{A})$, and then choose the best result.

With only a $(1+\varepsilon)$ factor gap in the singular values, using power method as before will require roughly (ignoring log factors) $1/\varepsilon$ iterations, which takes time roughly $\mathsf{nnz}(\mathbf{A})/\varepsilon$ to separate out the singular components, which is above our target budget. However, using Chebyshev polynomials, it is known that a $(1+\varepsilon)$ factor gap in the singular values can be separated with only roughly $1/\sqrt{\varepsilon}$ iterations [MM15] which takes time only $\mathsf{nnz}(\mathbf{A})/\sqrt{\varepsilon}$. The main lemma for this technique is the following:

Lemma 18.2.9 (Lemma 5, [MM15])**.** Given a specified value $\alpha > 0$, gap $\gamma \in (0,1]$, and $q \geq 1$, there exists a degree q polynomial $p(x)$ such that:

1. $p((1+\gamma)\alpha) = (1+\gamma)\alpha$
2. $p(x) \geq x$ for all $x \geq (1+\gamma)\alpha$
3. $|p(x)| \leq \frac{\alpha}{2^{q\sqrt{\gamma}-1}}$ for all $x \in [0,\alpha]$

Furthermore, when q is odd, the polynomial only contains odd powered monomials.

In words, the above lemma states that there is a polynomial that "jumps" by a factor of $2^{q\sqrt{\gamma}-1}$ in a window of size $(1+\gamma)$ at a specified location α. The difference between this lemma and our power method analysis from before is that we must specify the location of our "jump", α, in order to use the above polynomial in the Krylov method, whereas in the power method, the polynomial $p(x) = x^q$ had the "jump" property at any location α. Thus, in order to use the above lemma, we must *first* specify our jump location α, and then proceed with our previous techniques.

Our procedure is thus as follows. We first compute Krylov iterates $(\mathbf{A}\mathbf{A}^\top)^i\mathbf{A}\mathbf{g}$ for $i \in [q]$, where $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_d)$ and

$$q = O\left(\frac{1}{\sqrt{\varepsilon}} \log \frac{sk^2\sqrt{sr\log n}}{\varepsilon}\right).$$

We then proceed with our enumeration procedure. We guess a bucket $\alpha = L(1+\varepsilon)^t$ for some $t \in [O(1/\sqrt{\varepsilon})]$, and then consider the degree q polynomial $p_\alpha(x)$ that jumps by a $2^{q\sqrt{\varepsilon}-1}$ factor at α by Lemma 18.2.9. Then, we may compute the vector $\mathbf{U}p_\alpha(\mathbf{\Sigma})\mathbf{V}^\top\mathbf{g}$ as a linear combination of the Krylov iterates

$$(\mathbf{A}\mathbf{A}^\top)^i\mathbf{A}\mathbf{g} = \mathbf{U}\mathbf{\Sigma}^{2i+1}\mathbf{V}^\top\mathbf{g}$$

where the coefficients of the linear combination are the coefficients of the polynomial p_α. Next, we take the top sk entries of $\mathbf{U}p_\alpha(\mathbf{\Sigma})\mathbf{V}^\top\mathbf{g}$ as sets S_α and T_α, combine them with the sk entries S and T obtained earlier by the power method, and then take our new subset of entries to be

$$S' := S \cup \bigcup_\alpha S_\alpha$$

$$T' := T \cup \bigcup_{\alpha} T_\alpha$$

Finally we compute a rank k matrix $\mathbf{B}$ such that

$$\|\mathbf{P}_{S'}\mathbf{A}\mathbf{P}_{T'} - \mathbf{B}\|_F^2 \leq \min_{\text{rank } k\ \mathbf{C}} \|\mathbf{P}_{S'}\mathbf{A}\mathbf{P}_{T'} - \mathbf{C}\|_F^2 + \varepsilon\sigma_{k+1}^2(\mathbf{P}_{S'}\mathbf{A}\mathbf{P}_{T'})$$

using the results of [MM15].

Note that if the α we choose satisfies $\alpha \in [\sigma_{k+1}(\mathbf{A}), (1+\varepsilon)\sigma_{k+1}(\mathbf{A})]$, then all singular values j that are at least a $(1+\varepsilon)$ factor larger than α and at most $\Theta(1)\sigma_{k+1}(\mathbf{A})$ are scaled by at least a factor of

$$2^{q\sqrt{\varepsilon}-1} = \Theta\left(\frac{sk^2\sqrt{sr\log n}}{\varepsilon}\right) = \Theta\left(\frac{sk^2\sqrt{sr\log n}}{\varepsilon}\right)\frac{\sigma_j}{\sigma_{k+1}} = \Theta\left(\frac{sk\sqrt{\log n}}{\tau_j}\right),$$

which means we may recover all coordinates of the jth singular vectors that are at least τ_j for these singular values, as done in the analyses in Section 18.2.2. Thus, we have that

$$\begin{aligned}
\|\mathbf{A} - \mathbf{B}\|_2^2 &\leq \|\mathbf{A} - \mathbf{A}_l\|_2^2 + 8\varepsilon\sigma_{l+1}^2(\mathbf{A}) + \varepsilon\sigma_{l+1}^2(\mathbf{P}_{S'}\mathbf{A}\mathbf{P}_{T'}) \\
&\leq \|\mathbf{A} - \mathbf{A}_l\|_2^2 + 9\varepsilon\sigma_{l+1}^2(\mathbf{A}) \\
&= (1+9\varepsilon)\sigma_{l+1}^2(\mathbf{A}) \\
&\leq (1+9\varepsilon)(1+\varepsilon)\sigma_{k+1}^2(\mathbf{A}) \\
&\leq (1+11\varepsilon)\sigma_{k+1}^2(\mathbf{A})
\end{aligned}$$

by Lemma 18.2.2, where $l \in [k]$ is such that $\sigma_{k+1}^2(\mathbf{A}) \leq \sigma_{l+1}^2(\mathbf{A}) \leq (1+\varepsilon)\sigma_{k+1}^2(\mathbf{A})$.

The initial computation of the Krylov iterates takes time

$$O(\mathsf{nnz}(\mathbf{A})q) = O\left(\frac{\mathsf{nnz}(\mathbf{A})}{\sqrt{\varepsilon}}\log\frac{srk\log n}{\varepsilon}\right)$$

and a single guess of α takes time

$$O(nq) = O\left(\frac{n}{\sqrt{\varepsilon}}\log\frac{srk\log n}{\varepsilon}\right)$$

which we repeat $O(1/\sqrt{\varepsilon})$ times, so the total running time in this section is

$$O\left(\left(\frac{\mathsf{nnz}(\mathbf{A})}{\sqrt{\varepsilon}} + \frac{n}{\varepsilon}\right)\log\frac{srk\log n}{\varepsilon}\right).$$

We then additionally run an approximate SVD using Theorem 18.0.2 on the $O(sk/\sqrt{\varepsilon}) \times O(sk/\sqrt{\varepsilon})$ matrix, which adds an $s^2k^3(\log(sk))/\varepsilon^{3/2}$ term, for a running time of

$$O\left(\left(\frac{\mathsf{nnz}(\mathbf{A})}{\sqrt{\varepsilon}} + \frac{n}{\varepsilon}\right)\log\frac{srk\log n}{\varepsilon} + \frac{s^2k^3}{\varepsilon^{3/2}}\log(sk)\right).$$

This dominates the running times of the previous steps and thus is the running time of our entire algorithm.

Chapter 19

Future directions for sparse optimization

We conclude Part III of this thesis with several open directions arising from our investigations in this area.

Greedy algorithms for column subset selection. Our first question is to obtain an optimal understanding of the greedy algorithm for column subset selection with the Frobenius norm. Consider the greedy algorithm that iteratively updates a subset S of columns by setting $S \leftarrow S \cup \{i\}$, where i is the column which minimizes $\min_{\mathbf{X}} \|\mathbf{A}|^{S\cup\{i\}}\mathbf{X} - \mathbf{A}\|_F^2$. We ask whether this algorithm results in a nearly optimal column subset in the following sense:

Question 19.0.1. Let $k \in \mathbb{N}$. Does the greedy algorithm for column subset selection output a subset of $\tilde{O}(k/\varepsilon)$ columns such that

$$\min_{\mathbf{X}} \|\mathbf{A}|^{S}\mathbf{X} - \mathbf{A}\|_F^2 \leq (1+\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

Currently, it is known that similar bounds can be obtained up to a logarithmic factor in a condition number-type parameter [ABF+16], or polynomial factors in k and ε [DV06, BRW21]. It is interesting to determine whether the lower bound of $\Omega(k/\varepsilon)$ [DV06] can be achieved, or if there exists a matrix $\mathbf{A}$ for which the greedy algorithm fails to achieve the above guarantee with $\tilde{O}(k/\varepsilon)$ columns. In fact, to the best of our knowledge, it is not known whether there is any efficient deterministic column subset selection algorithm that achieves this guarantee with $\tilde{O}(k/\varepsilon)$ columns.

More broadly, we raise the question of whether greedy algorithms can replace other techniques in matrix approximation in greater generality.

Question 19.0.2. Can sparse optimization techniques, especially greedy algorithms, replace sketching and sampling techniques in matrix approximation?

Although randomized algorithms based on sketching and sampling have proven to be a highly successful development in algorithms research, they often lack the simplicity, ease of implementation, practical performance, and interpretability of plain greedy algorithms, which often prove to be the preferred choice in practical applications. Indeed, arguably the most natural algorithms for matrix approximation are accomplished by greedy algorithms, from the singular value decomposition to the first proposed algorithms for column subset selection [Cha86] to the

most popular approaches for neural network compression [FC19]. Thus, as algorithms researchers, one of the most important directions is to establish how good these greedy algorithms are in the context of matrix approximation.

There have already been a few fruitful lines of work establishing the near-optimality of greedy algorithms for matrix approximation. A variation on a greedy algorithm has been shown to be nearly optimal for column subset selection under the entrywise ℓ_p loss for certain values of p, as shown by [SWZ17, MW21, WY23a] (see also Chapter 17). Our nearly optimal online coreset algorithm for John ellipsoids in Chapter 11 may also be viewed as a greedy algorithm. Yet another recent positive result for greedy algorithms is the result that when selecting a maximum volume subset of points, the greedy algorithm is nearly optimal in the composable coreset model [ÇM09, IMGR19, IMGR20, GMS23]. We hope that greedy algorithms will continue to prove to be the "right" algorithm in matrix approximation problems.

Column subset selection for the entrywise ℓ_p loss. The second question we raise is on settling the trade-offs for column subset selection problem for the entrywise ℓ_p loss. It is known that for $p < 2$, there is always a subset S of $|S| = \tilde{O}(k)$ columns satisfying

$$\min_{\mathbf{X}} \|\mathbf{A}|^S\mathbf{X} - \mathbf{A}\|_{p,p}^p \leq \tilde{O}(k^{1/p-1/2}) \min_{\mathrm{rank}(\hat{\mathbf{A}})\leq k} \|\mathbf{A} - \hat{\mathbf{A}}\|_{p,p}^p$$

and that any column subset with a distortion factor of $O(k^{1/p-1/2-o(1)})$ must contain at least $k(\log k)^{\omega(1)}$ columns [MW21]. In Chapter 17, we have shown a similar upper bound for $p > 2$, showing that there is always a subset S of $|S| = \tilde{O}(k)$ columns satisfying

$$\min_{\mathbf{X}} \|\mathbf{A}|^S\mathbf{X} - \mathbf{A}\|_{p,p}^p \leq \tilde{O}(k^{1/2-1/p}) \min_{\mathrm{rank}(\hat{\mathbf{A}})\leq k} \|\mathbf{A} - \hat{\mathbf{A}}\|_{p,p}^p.$$

However, we only have a nearly matching lower bound for $p = \infty$, which shows that a column subset with a distortion factor of $O(k^{1/2-o(1)})$ must contain at least $k^{\omega(1)}$ columns. This leads to the following question:

Question 19.0.3. Let $k \in \mathbb{N}$ and $2 < p < \infty$. What is the minimum possible distortion κ such that for any $\mathbf{A}$, there exists a subset S of $|S| = \tilde{O}(k)$ columns such that

$$\min_{\mathbf{X}} \|\mathbf{A}|^S\mathbf{X} - \mathbf{A}\|_{p,p}^p \leq \kappa \min_{\mathrm{rank}(\hat{\mathbf{A}})\leq k} \|\mathbf{A} - \hat{\mathbf{A}}\|_{p,p}^p.$$

Bibliography

[AAHS99] Pankaj K. Agarwal, Boris Aronov, Sariel Har-Peled, and Micha Sharir. Approximation and exact algorithms for minimum-width annuli and shells. In Victor Milenkovic, editor, *Proceedings of the Fifteenth Annual Symposium on Computational Geometry, Miami Beach, Florida, USA, June 13-16, 1999*, pages 380–389. ACM, 1999. 11.4.5

[ABF+16] Jason M. Altschuler, Aditya Bhaskara, Gang Fu, Vahab S. Mirrokni, Afshin Rostamizadeh, and Morteza Zadimoghaddam. Greedy column subset selection: New bounds and distributed algorithms. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2539–2548. JMLR.org, 2016. 1.4.2, 16.1.2, 19

[AHV04] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, 2004. 1.3.4, 11, 11.4, 11.4.2, 11.4.5

[AHV05] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52(1-30):3, 2005. 1.3.4, 11, 11.4, 11.4.1, 11.4.2

[AHY08] Pankaj K. Agarwal, Sariel Har-Peled, and Hai Yu. Robust shape fitting via peeling and grating coresets. *Discret. Comput. Geom.*, 39(1-3):38–58, 2008. 11.4.1, 11.4.2, 11.4.1, 11.4.1, 1

[AKPS19] Deeksha Adil, Rasmus Kyng, Richard Peng, and Sushant Sachdeva. Iterative refinement for ℓ_p-norm regression. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 1405–1424. SIAM, 2019. 12.3.3

[AKY23] Sepehr Assadi, Michael Kapralov, and Huacheng Yu. On constructing spanners from random gaussian projections. In Nicole Megow and Adam D. Smith, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2023, September 11-13, 2023, Atlanta, Georgia, USA*, volume 275 of *LIPIcs*, pages 57:1–57:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023. 1.1.1

[ALO15] Zeyuan Allen-Zhu, Zhenyu Liao, and Lorenzo Orecchia. Spectral sparsification and

regret minimization beyond matrix multiplicative updates. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 237–245. ACM, 2015. 15.1, 15.1

[AM15] Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 775–783, 2015. 14.1.5

[AP21] Sercan Ö Arik and Tomas Pfister. TabNet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021. 16.1.2

[AS98] Pankaj K. Agarwal and Micha Sharir. Efficient algorithms for geometric optimization. *ACM Comput. Surv.*, 30(4):412–458, 1998. 11.4.5

[AS15] Pankaj K. Agarwal and R. Sharathkumar. Streaming algorithms for extent problems in high dimensions. *Algorithmica*, 72(1):83–98, 2015. 1.3.4, 11, 11.4, 11.4.1

[AS16] Jose M Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. *Advances in neural information processing systems*, 29, 2016. 16.1

[AS20] Kyriakos Axiotis and Maxim Sviridenko. Sparse convex optimization via adaptively regularized hard thresholding. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 452–462. PMLR, 2020. 16.1.2, 16.1.4, 16.4.1

[AS21] Kyriakos Axiotis and Maxim Sviridenko. Local search algorithms for rank-constrained convex optimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 16.1.5

[AS22] Kyriakos Axiotis and Maxim Sviridenko. Iterative hard thresholding with adaptive regularization: Sparser solutions without sacrificing runtime. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 1175–1197. PMLR, 2022. 16.1.4

[ASW13] Haim Avron, Vikas Sindhwani, and David P. Woodruff. Sketching structured matrices for faster nonlinear regression. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2994–3002, 2013. 7.2

[Aue30] Herman Auerbach. *On the area of convex curves with conjugate diameters*. PhD

thesis, PhD thesis, University of Lwów, 1930. 1.5.2, 3.1

[AVu97] Noga Alon and Văn H. Vũ. Anti-Hadamard matrices, coin weighing, threshold gates, and indecomposable hypergraphs. *J. Combin. Theory Ser. A*, 79(1):133–160, 1997. 1.3.3

[AW20a] Ehsan Amid and Manfred K. Warmuth. Reparameterizing mirror descent as gradient descent. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 16.1.2

[AW20b] Ehsan Amid and Manfred K. Warmuth. Winnowing with gradient descent. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 163–182. PMLR, 2020. 16.1.2

[AY23] Kyriakos Axiotis and Taisuke Yasuda. Performance of ℓ_1 regularization for sparse convex optimization. *CoRR*, abs/2307.07405, 2023. (document), 1.4.2, 1.4.3, 16

[Bak99] Sergey Bakin. *Adaptive regression and model selection in data mining problems*. PhD dissertation, The Australian National University, 1999. 7

[BAZ19] Muhammed Fatih Balın, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *International conference on machine learning*, pages 444–453. PMLR, 2019. (document), 16.6.1, 16.6.1, 16.1, 16.6.3

[BBB+19] Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, and David P. Woodruff. A PTAS for ℓ_p-low rank approximation. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 747–766. SIAM, 2019. 17

[BBK+18] Avrim Blum, Vladimir Braverman, Ananya Kumar, Harry Lang, and Lin F. Yang. Approximate convex hull of data streams. In Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella, editors, *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic*, volume 107 of *LIPIcs*, pages 21:1–21:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. 11.4.2

[BCFS14] Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1556–1564, 2014. 1.4.1, 16.1.1

[BDM11] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near optimal column-based matrix reconstruction. In Rafail Ostrovsky, editor, *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA,*

October 22-25, 2011, pages 305–314. IEEE Computer Society, 2011. 16.1.2

[BDM+20] Vladimir Braverman, Petros Drineas, Cameron Musco, Christopher Musco, Jalaj Upadhyay, David P. Woodruff, and Samson Zhou. Near optimal linear algebra in the online and sliding window models. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, pages 517–528. IEEE, 2020. 1.3.3, 1.3.4, 6.6, 6.6.3, 14.1, 14.1, 14.1.2, 14.1.2, 14.6.1, 14.6.2

[BDN15] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 499–508. ACM, 2015. 1.2.2

[Ber63] Claude Berge. *Topological spaces: Including a treatment of multi-valued functions, vector spaces and convexity*. Oliver & Boyd, 1963. 16.2.5

[BFL16] Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coreset constructions. *CoRR*, abs/1612.00889, 2016. 1.3.1, 1.3.1

[BJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004. 3

[BJM+11] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, 5:19–53, 2011. 8

[BK15] András A. Benczúr and David R. Karger. Randomized approximation schemes for cuts and flows in capacitated graphs. *SIAM J. Comput.*, 44(2):290–319, 2015. 1.3.2

[BKLZ20] Aditya Bhaskara, Amin Karbasi, Silvio Lattanzi, and Morteza Zadimoghaddam. Online MAP inference of determinantal point processes. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 11.4.4, 2

[BLM89] J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Math.*, 162(1-2):73–141, 1989. 1.5.1, 1.5.1, 4.2.5, 6.1.3, 6.1.4, 6.1.3, 6.1.3, 6.4.2, 6.4.5, 7.1, 8.1.3, 10.2, 12.1, 12.5.2, 14.3.2

[BLVZ19] Aditya Bhaskara, Silvio Lattanzi, Sergei Vassilvitskii, and Morteza Zadimoghaddam. Residual based sampling for online low rank approximation. In David Zuckerman, editor, *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 1596–1614. IEEE Computer Society, 2019. 14.1, 14.1.2

[BMD09] Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In Claire Mathieu,

editor, *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2009, New York, NY, USA, January 4-6, 2009*, pages 968–977. SIAM, 2009. 16.1.2

[BMN01] Aharon Ben-Tal, Tamar Margalit, and Arkadi Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM J. Optim.*, 12(1):79–108, 2001. 12.3.3

[BMV23] Aditya Bhaskara, Sepideh Mahabadi, and Ali Vakilian. Tight bounds for volumetric spanners and applications. *CoRR*, abs/2310.00175, 2023. 3.2

[BO10] Vladimir Braverman and Rafail Ostrovsky. Zero-one frequency laws. In Leonard J. Schulman, editor, *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 281–290. ACM, 2010. 10

[BPC+11] Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011. 8

[BRT09] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. 1.4.1, 16.1.1

[BRW21] Aditya Bhaskara, Aravinda Kanchana Ruwanpathirana, and Maheshakya Wijewardena. Additive error guarantees for weighted low rank approximation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 874–883. PMLR, 2021. 19

[BSS12] Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012. 15.1, 15.1, 17.3.1

[BSST13] Joshua D. Batson, Daniel A. Spielman, Nikhil Srivastava, and Shang-Hua Teng. Spectral sparsification of graphs: theory and algorithms. *Commun. ACM*, 56(8):87–94, 2013. 1.3.2

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004. 16.2.1, 16.3.1

[BW17] Christos Boutsidis and David P. Woodruff. Optimal CUR matrix decompositions. *SIAM J. Comput.*, 46(2):543–589, 2017. 16.1.2, 17

[BWZ16] Christos Boutsidis, David P. Woodruff, and Peilin Zhong. Optimal principal component analysis in distributed and streaming models. In Daniel Wichs and Yishay Mansour, editors, *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 236–249. ACM, 2016. 1.2.1, 1.4.2

[Can08] Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris*, 346(9-10):589–592, 2008. 1.4.1, 16.1.1

[CCKW22] Nadiia Chepurko, Kenneth L. Clarkson, Praneeth Kacham, and David P. Woodruff.

Near-optimal algorithms for linear algebra in the current matrix multiplication time. In Joseph (Seffi) Naor and Niv Buchbinder, editors, *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, pages 3043–3068. SIAM, 2022. 1, 1.2.2

[CCLY19] Michael B. Cohen, Ben Cousins, Yin Tat Lee, and Xin Yang. A near-optimal algorithm for approximating the John ellipsoid. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 849–873. PMLR, 2019. 1.1.1, 6.1.3

[CD21] Xue Chen and Michal Derezinski. Query complexity of least absolute deviation regression via robust uniform convergence. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 1144–1179. PMLR, 2021. 7.1, 12.1, 12.6, 12.6.1, 12.6.1, 12.6.1, 12.6.1, 13.1.5

[CDDR23] Shabarish Chenakkod, Michal Derezinski, Xiaoyu Dong, and Mark Rudelson. Optimal embedding dimension for sparse subspace embeddings. *CoRR*, abs/2311.10680, 2023. 1.1.1, 1, 1.2.2

[CDM+16] Kenneth L. Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, Xiangrui Meng, and David P. Woodruff. The fast Cauchy transform and faster robust linear regression. *SIAM J. Comput.*, 45(3):763–810, 2016. 3

[CDS98] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998. 1.4.1, 16.1.1

[CEM+15] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 163–172. ACM, 2015. 1.1.1, 14.1, 14.1, 4, 4, 16.1.2

[CGK+17] Flavio Chierichetti, Sreenivas Gollapudi, Ravi Kumar, Silvio Lattanzi, Rina Panigrahy, and David P. Woodruff. Algorithms for ℓ_p low-rank approximation. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 806–814. PMLR, 2017. 1.4.2, 5, 16.1.2, 17.1, 17.3

[Cha86] Tony F Chan. Alternative to the SVD: rank revealing QR-factorizations. In *Advanced Algorithms and Architectures for Signal Processing I*, volume 696, pages 31–38. SPIE, 1986. 1.4.2, 19

[Cha02] Timothy M. Chan. Approximating the diameter, width, smallest enclosing cylinder, and minimum-width annulus. *Int. J. Comput. Geom. Appl.*, 12(1-2):67–85, 2002.

11.4.5

[Cha06] Timothy M. Chan. Faster core-set constructions and data-stream algorithms in fixed dimensions. *Comput. Geom.*, 35(1-2):20–35, 2006. 11.4, 11.4.2, 11.4.4, 11.4.5

[Cha16] Timothy M. Chan. Dynamic streaming algorithms for epsilon-kernels. In Sándor P. Fekete and Anna Lubiw, editors, *32nd International Symposium on Computational Geometry, SoCG 2016, June 14-18, 2016, Boston, MA, USA*, volume 51 of *LIPIcs*, pages 27:1–27:11. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016. 11.4.2

[Chi22] Lénaïc Chizat. Convergence rates of gradient methods for convex optimization in the space of measures. *Open J. Math. Optim.*, 3:Art. No. 8, 19, 2022. 16.1.2

[Çiv14] Ali Çivril. Column subset selection problem is UG-hard. *J. Comput. Syst. Sci.*, 80(4):849–859, 2014. 16.1.2

[CKL22] Yu Chen, Sanjeev Khanna, and Huan Li. On weighted graph sparsification by linear sketching. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, October 31 - November 3, 2022*, pages 474–485. IEEE, 2022. 1.1.1

[Cla05] Kenneth L. Clarkson. Subgradient and sampling algorithms for ℓ_1 regression. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, pages 257–266, USA, 2005. Society for Industrial and Applied Mathematics. 3.1, 6.1.2, 6.1.3, 6.1.2

[CLM+15] Michael B. Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In Tim Roughgarden, editor, *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, ITCS 2015, Rehovot, Israel, January 11-13, 2015*, pages 181–190. ACM, 2015. 1.3.2, 6.1.1, 6.6.3, 8, 4

[CLS22] Cheng Chen, Yi Li, and Yiming Sun. Online active regression. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 3320–3335. PMLR, 2022. 6.6, 12.1

[ÇM09] Ali Çivril and Malik Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theor. Comput. Sci.*, 410(47-49):4801–4811, 2009. 19

[ÇM12] Ali Çivril and Malik Magdon-Ismail. Column subset selection via sparse approximation of SVD. *Theor. Comput. Sci.*, 421:1–14, 2012. 16.1.2

[CMM17] Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In Philip N. Klein, editor, *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 1758–1777. SIAM, 2017. 14.1, 14.1, 4, 14.1.5, 4, 14.1.1, 14.1.2, 14.3.5, 16.1.2, 17

[CMP16] Michael B. Cohen, Cameron Musco, and Jakub Pachocki. Online row sampling. In

Klaus Jansen, Claire Mathieu, José D. P. Rolim, and Chris Umans, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2016, September 7-9, 2016, Paris, France*, volume 60 of *LIPIcs*, pages 7:1–7:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016. 1.3.3, 1.3.3, 1.3.3, 1.3.8, 6.1.3, 6.6

[CMP20] Michael B. Cohen, Cameron Musco, and Jakub Pachocki. Online row sampling. *Theory Comput.*, 16:1–25, 2020. 1.3.3, 1.3.3, 1.3.4, 1.3.3, 1.3.8, 6.1.3, 6.6, 6.6.3, 11.1

[CNW16] Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. In Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi, editors, *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, volume 55 of *LIPIcs*, pages 11:1–11:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016. 1.1.1, 13.1.1

[Coh16] Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In Robert Krauthgamer, editor, *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 278–287. SIAM, 2016. 1.2.2, 14.3.2

[CP15] Michael B. Cohen and Richard Peng. L_p row sampling by lewis weights. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 183–192. ACM, 2015. 1.5.1, 6.1.3, 6.1.3, 6.1.6, 6.1.3, 6.1.3, 6.1.8, 6.1.3, 6.1.9, 6.3.1, 6.3.2, 6.3.3, 9.1, 9.1, 10.2, 12.4, 4, 14.3.2

[CP19] Xue Chen and Eric Price. Active regression via linear-sample sparsification. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 663–695. PMLR, 2019. 1.3.4, 12.1

[CPR16] Siu On Chan, Dimitris Papailliopoulos, and Aviad Rubinstein. On the approximability of sparse PCA. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 623–646. JMLR.org, 2016. 18.0.4

[CRT06] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006. 1.4.1, 16.1.1

[CSS21] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. Improved coresets and sublinear algorithms for power means in euclidean spaces. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 21085–21098, Virtual, 2021. 1.3.4, 12.6, 13.1.5,

13.1.5

[CSWZ23] Yeshwanth Cherapanamjeri, Sandeep Silwal, David P. Woodruff, and Samson Zhou. Optimal algorithms for linear algebra in the current matrix multiplication time. In Nikhil Bansal and Viswanath Nagarajan, editors, *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023*, pages 4026–4049. SIAM, 2023. 1.1.1, 1, 1.2.2, 14.1

[CT07] Emmanuel Candès and Terence Tao. Rejoinder: "The Dantzig selector: statistical estimation when p is much larger than n" [Ann. Statist. **35** (2007), no. 6, 2313–2351; mr2382644]. *Ann. Statist.*, 35(6):2392–2404, 2007. 1.4.1, 16.1.1

[CW09] Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In Michael Mitzenmacher, editor, *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 205–214. ACM, 2009. 1.2.1

[CW13] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 81–90. ACM, 2013. 1.1.1, 1, 1.2.2, 1.3.2, 4.1, 4.1.1, 6.1.1, 8, 13.1.1, 4, 17

[CW15a] Kenneth L. Clarkson and David P. Woodruff. Input sparsity and hardness for robust subspace approximation. In Venkatesan Guruswami, editor, *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 310–329. IEEE Computer Society, 2015. 1.1.1, 4.1, 8, 10, 10.1, 10.1.2, 10.2, 10.2, 10.2.1, 13.1.6, 14.1, 14.1, 14.1, 17.1, 17.1

[CW15b] Kenneth L. Clarkson and David P. Woodruff. Sketching for M-estimators: A unified approach to robust regression. In Piotr Indyk, editor, *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 921–939. SIAM, 2015. 1.2.1, 10.2, 17.1, 17.2, 17.3.1

[CW22] Moses Charikar and Erik Waingarten. The Johnson-Lindenstrauss lemma for clustering and subspace approximation: From coresets to dimension reduction. *CoRR*, abs/2205.00371, 2022. 14.1

[CWW19] Kenneth L. Clarkson, Ruosong Wang, and David P. Woodruff. Dimensionality reduction for Tukey regression. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1262–1271. PMLR, 2019. 10, 10.2, 10.2.1

[CWZ23] Vincent Cohen-Addad, David P. Woodruff, and Samson Zhou. Streaming euclidean k-median and k-means with o(log n) space. *CoRR*, abs/2310.02882, 2023. 1.3.3, 14.1, 14.1.2, 14.6.2

[DDH+09] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W.

Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009. 3.1, 3.1.1, 6.1.2, 6.1.3, 6.1.2, 10.2, 12.2

[DDWY23] Gregory Dexter, Petros Drineas, David P. Woodruff, and Taisuke Yasuda. Sketching algorithms for sparse dictionary learning: PTAS and turnstile streaming. *CoRR*, abs/2310.19068, 2023. 1.1.1

[DG03] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003. 3

[DGL17] Diemert Eustache, Meynet Julien, Pierre Galland, and Damien Lefortier. Attribution modeling increases efficiency of bidding in display advertising. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, Halifax, NS, Canada, August, 14, 2017*. ACM, 2017. 16.6.2

[DHJ+18] Chen Dan, Kristoffer Arnsfelt Hansen, He Jiang, Liwei Wang, and Yuchen Zhou. Low rank approximation of binary matrices: Column subset selection and generalizations. In Igor Potapov, Paul G. Spirakis, and James Worrell, editors, *43rd International Symposium on Mathematical Foundations of Computer Science, MFCS 2018, August 27-31, 2018, Liverpool, UK*, volume 117 of *LIPIcs*, pages 41:1–41:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. 17

[DK11] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1057–1064, 2011. 1.4.1, 8, 16.1.5

[DKM06a] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices I: approximating matrix multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006. 1.3.2, 17

[DKM06b] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices II: computing a low-rank approximation to a matrix. *SIAM J. Comput.*, 36(1):158–183, 2006. 1.3.2, 1.4.2, 17

[DKM06c] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices III: computing a compressed approximate matrix decomposition. *SIAM J. Comput.*, 36(1):184–206, 2006. 17

[DMM06a] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006, Miami, Florida, USA, January 22-26, 2006*, pages 1127–1136. ACM Press, 2006. 1.1.2, 1.3.2, 1.3.3, 6.1.3, 12.1, 4

[DMM06b] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Subspace sampling and relative-error matrix approximation: Column-row-based methods. In Yossi Azar and Thomas Erlebach, editors, *Algorithms - ESA 2006, 14th Annual European Symposium, Zurich, Switzerland, September 11-13, 2006, Proceedings*, volume 4168 of *Lecture Notes in Computer Science*, pages 304–314. Springer, 2006. 1.3.2, 1.4.2, 14.1, 4, 17

[DMM08] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM J. Matrix Anal. Appl.*, 30(2):844–881, 2008. 1.3.2, 1.4.2, 14.1, 4, 16.1.2

[DMMW12] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13:3475–3506, 2012. 1.1.1, 1, 1.3.2, 6.1.1, 8, 4

[DP22] Amit Deshpande and Rameshwar Pratap. One-pass additive-error subset selection for ℓ_p subspace approximation. In Mikolaj Bojanczyk, Emanuela Merelli, and David P. Woodruff, editors, *49th International Colloquium on Automata, Languages, and Programming, ICALP 2022, July 4-8, 2022, Paris, France*, volume 229 of *LIPIcs*, pages 51:1–51:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. 14.1

[DR10] Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 329–338. IEEE Computer Society, 2010. 16.1.2

[DRVW06] Amit Deshpande, Luis Rademacher, Santosh S. Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006, Miami, Florida, USA, January 22-26, 2006*, pages 1117–1126. ACM Press, 2006. 14.1

[DS89] David L. Donoho and Philip B. Stark. Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, 49(3):906–931, 1989. 1.4.1, 16.1.1

[DTV11] Amit Deshpande, Madhur Tulsiani, and Nisheeth K. Vishnoi. Algorithms and hardness for subspace approximation. In Dana Randall, editor, *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011, San Francisco, California, USA, Janu-ary 23-25, 2011*, pages 482–496. SIAM, 2011. 14.1, 14.1.1, 14.3.2

[Dud67] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis*, 1:290–330, 1967. 2.3.3

[DV06] Amit Deshpande and Santosh S. Vempala. Adaptive sampling and fast low-rank matrix approximation. In Josep Díaz, Klaus Jansen, José D. P. Rolim, and Uri Zwick, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 9th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems, APPROX 2006 and 10th International Workshop on Randomization and Computation, RANDOM 2006, Barcelona, Spain, August 28-30 2006, Proceedings*, volume 4110 of *Lecture Notes in Computer Science*, pages 292–303. Springer, 2006. 1.4.2, 13.1.6, 13.6.3, 16.1.2, 17, 19

[DV07] Amit Deshpande and Kasturi R. Varadarajan. Sampling-based dimension reduction for subspace approximation. In David S. Johnson and Uriel Feige, editors, *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego,*

California, USA, June 11-13, 2007, pages 641–650. ACM, 2007. 13.1.6, 14.1, 14.1, 14.1, 15.2

[Dvo61] Aryeh Dvoretzky. Some results on convex bodies and Banach spaces. In *Proc. Internat. Sympos. Linear Spaces (Jerusalem, 1960)*, pages 123–160. Jerusalem Academic Press, Jerusalem; Pergamon, Oxford, 1961. 13.1.5, 4, 14.1.1, 14.3.1

[DWZ+19] Chen Dan, Hong Wang, Hongyang Zhang, Yuchen Zhou, and Pradeep Ravikumar. Optimal analysis of subset-selection based l_p low-rank approximation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2537–2548, 2019. 1.4.2, 5, 16.1.2, 17.1, 17.3

[DWZ23] Ran Duan, Hongxun Wu, and Renfei Zhou. Faster matrix multiplication via asymmetric hashing. In *64th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2023, Santa Cruz, CA, USA, November 6-9, 2023*, pages 2129–2138. IEEE, 2023. 1

[EHJT04] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors. 16.1.3

[EKDN18] Ethan R Elenberg, Rajiv Khanna, Alexandros G Dimakis, and Sahand Negahban. Restricted strong convexity implies weak submodularity. *The Annals of Statistics*, 46(6B):3539–3568, 2018. 1.4.1, 16.1, 16.1.2, 16.1.4, 16.1.5, 16.4

[EY36] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. 18

[FC19] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 19

[FEGK15] Ahmed K. Farahat, Ahmed Elgohary, Ali Ghodsi, and Mohamed S. Kamel. Greedy column subset selection for large-scale data sets. *Knowl. Inf. Syst.*, 45(1):1–34, 2015. 16.1.2

[Fel20] Dan Feldman. Core-sets: An updated survey. *WIREs Data Mining Knowl. Discov.*, 10(1), 2020. 1.3.1

[FGK11] Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel. An efficient greedy method for unsupervised feature selection. In Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu, editors, *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, pages 161–170. IEEE Computer Society, 2011. 16.1.2

[FGK13] Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel. Efficient greedy feature selection for unsupervised learning. *Knowl. Inf. Syst.*, 35(2):285–310, 2013. 16.1.2

[FGM17] Robert M. Freund, Paul Grigas, and Rahul Mazumder. A new perspective on boosting in linear regression via subgradient optimization and relatives. *Ann. Statist.*, 45(6):2328–2364, 2017. 16.1.3

[FHT10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010. 8

[FKT15] Dean P. Foster, Howard J. Karloff, and Justin Thaler. Variable selection is hard. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 696–709. JMLR.org, 2015. 1.4.1, 7, 16.1.1

[FKV04] Alan M. Frieze, Ravi Kannan, and Santosh S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, 2004. 1.4.2, 17

[FKW21] Zhili Feng, Praneeth Kacham, and David P. Woodruff. Dimensionality reduction for the sum-of-distances metric. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 3220–3229. PMLR, 2021. 13.1.6, 14.1, 14.1, 4, 4, 4, 14.1.1, 14.3.2

[FL11] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In Lance Fortnow and Salil P. Vadhan, editors, *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 569–578. ACM, 2011. 1.3.1, 1.3.1, 1.3.1, 6.1.1, 13.1.6, 14.1, 14.1, 14.1, 4, 4

[FLM77] T. Figiel, J. Lindenstrauss, and V. D. Milman. The dimension of almost spherical sections of convex bodies. *Acta Math.*, 139(1-2):53–94, 1977. 4, 14.1.1, 14.3.1

[FLPS22] Maryam Fazel, Yin Tat Lee, Swati Padmanabhan, and Aaron Sidford. Computing lewis weights to high precision. In Joseph (Seffi) Naor and Niv Buchbinder, editors, *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, pages 2723–2742. SIAM, 2022. 6.1.3

[FMSW10] Dan Feldman, Morteza Monemizadeh, Christian Sohler, and David P. Woodruff. Coresets and sketches for high dimensional subspace approximation problems. In Moses Charikar, editor, *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 630–649. SIAM, 2010. 1.1.1, 14.1

[Fre75] David A. Freedman. On tail probabilities for martingales. *Ann. Probability*, 3:100–118, 1975. 4.3.2

[FSS20] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM J. Comput.*, 49(3):601–657, 2020. 1.3.1, 1.3.1, 14.1

[GE96] Ming Gu and Stanley C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM J. Sci. Comput.*, 17(4):848–869, 1996. 1.4.2

[GGH19] Ning Gui, Danni Ge, and Ziyin Hu. AFS: An attention-based mechanism for supervised feature selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3705–3713, 2019. 16.6.1, 16.6.3

[GK10] Eugene Gover and Nishan Krikorian. Determinants and the volumes of parallelotopes and zonotopes. *Linear Algebra Appl.*, 433(1):28–40, 2010. 11.1, 11.1.5

[GMS23] Siddharth Gollapudi, Sepideh Mahabadi, and Varun Sivashankar. Composable coresets for determinant maximization: Greedy is almost optimal. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 19

[GPV21] Mehrdad Ghadiri, Richard Peng, and Santosh S Vempala. Faster p-norm regression using sparsity. *arXiv preprint arXiv:2109.11537*, 2021. 8

[GRSW12] Venkatesan Guruswami, Prasad Raghavendra, Rishi Saket, and Yi Wu. Bypassing UGC from some optimal geometric inapproximability results. In Yuval Rabani, editor, *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 699–717. SIAM, 2012. 14.1

[Gu15] Ming Gu. Subspace iteration randomization and singular value problems. *SIAM J. Sci. Comput.*, 37(3), 2015. 18.2.1, 18.2.1

[GV18] Nicolas Gillis and Stephen A. Vavasis. On the complexity of robust PCA and ℓ_1-norm low-rank matrix approximation. *Math. Oper. Res.*, 43(4):1072–1084, 2018. 17

[GV21] Aparna Gupte and Vinod Vaikuntanathan. The fine-grained hardness of sparse linear regression. *CoRR*, abs/2106.03131, 2021. 1.4.1, 7, 16.1.1

[GVL13] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013. 1.1, 1.4.2

[GVR10] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*, 2010. 16.1.2

[GZ84] Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *Ann. Probab.*, 12(4):929–998, 1984. With discussion. 2.3.1

[HBCY21] Taylor M. Hernandez, Roel Van Beeumen, Mark A. Caprio, and Chao Yang. A greedy algorithm for computing eigenvalues of a symmetric matrix with localized eigenvectors. *Numer. Linear Algebra Appl.*, 28(2), 2021. 18

[HK16] Elad Hazan and Zohar S. Karnin. Volumetric spanners: An efficient exploration basis for learning. *J. Mach. Learn. Res.*, 17:119:1–119:34, 2016. 3.2, 3.2.2

[HMR23] Hussein Hazimeh, Rahul Mazumder, and Peter Radchenko. Grouped variable selection with discrete optimization: computational and statistical perspectives. *Ann. Statist.*, 51(1):1–32, 2023. 8

[Hof17] Peter D Hoff. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017. 16.1.2, 16.5

[HRR22] Laurel Heck, Victor Reis, and Thomas Rothvoss. The vector balancing constant for zonotopes. *CoRR*, abs/2210.16460, 2022. 6

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer, 2009. 8

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015. 7

[HV20] Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 1416–1429. ACM, 2020. 13.1.6, 13.5, 14.1, 14.1, 14.1, 4, 4, 4, 14.1.2, 15.2, 15.2

[HW20] Daniel Hug and Wolfgang Weil. *Lectures on convex geometry*, volume 286 of *Graduate Texts in Mathematics*. Springer, Cham, [2020] ©2020. 11.4.2, 11.4.4

[IMGR19] Piotr Indyk, Sepideh Mahabadi, Shayan Oveis Gharan, and Alireza Rezaei. Composable core-sets for determinant maximization: A simple near-optimal algorithm. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4254–4263. PMLR, 2019. 11.4.4, 11.4.4, 3, 19

[IMGR20] Piotr Indyk, Sepideh Mahabadi, Shayan Oveis Gharan, and Alireza Rezaei. Composable core-sets for determinant maximization problems via spectral spanners. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 1675–1694. SIAM, 2020. 11.4.1, 11.4.4, 19

[Ind06] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006. 3, 3.0.1

[IW05] Piotr Indyk and David P. Woodruff. Optimal approximations of the frequency moments of data streams. In Harold N. Gabow and Ronald Fagin, editors, *Proceedings of the 37th Annual ACM Symposium on Theory of Computing, Baltimore, MD, USA, May 22-24, 2005*, pages 202–208. ACM, 2005. 3, 4.1

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven,*

Conn., 1982), volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984. 1.2.2

[JLL+21] Shuli Jiang, Dennis Li, Irene Mengze Li, Arvind V. Mahankali, and David P. Woodruff. Streaming and distributed algorithms for robust column subset selection. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4971–4981. PMLR, 2021. 1.4.2, 5, 16.1.2

[JLLS23] Arun Jambulapati, James R. Lee, Yang P. Liu, and Aaron Sidford. Sparsifying generalized linear models. *CoRR*, abs/2311.18145, 2023. 10.2

[JLS22] Arun Jambulapati, Yang P. Liu, and Aaron Sidford. Improved iteration complexities for overconstrained p-norm regression. In Stefano Leonardi and Anupam Gupta, editors, *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 529–542. ACM, 2022. 6.1.3, 6.1.7, 6.1.3, 6.1.8, 9.2

[Joh48] Fritz John. Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948*, pages 187–204. Interscience Publishers, Inc., New York, N. Y., 1948. 1.5.2, 11

[JTD11] Prateek Jain, Ambuj Tewari, and Inderjit S. Dhillon. Orthogonal matching pursuit with replacement. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1215–1223, 2011. 16.1.2, 16.1.4

[JTK14] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in neural information processing systems*, 27, 2014. 16.1.4

[KKB07] Kwangmoo Koh, Seung-Jean Kim, and Stephen P. Boyd. An interior-point method for large-scale ℓ_1-regularized logistic regression. *J. Mach. Learn. Res.*, 8:1519–1555, 2007. 8

[KKMR21] Jonathan A. Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. On the power of preconditioning in sparse linear regression. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 550–561. IEEE, 2021. 16.1.1

[KKMR22] Jonathan A. Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. Distributional hardness against preconditioned lasso via erasure-robust designs. *CoRR*, abs/2203.02824, 2022. 16.1.1

[KKMR23] Jonathan A. Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. Feature adaptation for sparse linear regression. *CoRR*, abs/2305.16892, 2023. 16.1.1

[KMN11] Daniel M. Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit johnson-

lindenstrauss families. In Leslie Ann Goldberg, Klaus Jansen, R. Ravi, and José D. P. Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 14th International Workshop, APPROX 2011, and 15th International Workshop, RANDOM 2011, Princeton, NJ, USA, August 17-19, 2011. Proceedings*, volume 6845 of *Lecture Notes in Computer Science*, pages 628–639. Springer, 2011. 3

[KNR99] Ilan Kremer, Noam Nisan, and Dana Ron. On randomized one-round communication complexity. *Comput. Complex.*, 8(1):21–49, 1999. 2.2.1

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*, volume 9851 of *Lecture Notes in Computer Science*, pages 795–811. Springer, 2016. 16.1.5

[KW22] Praneeth Kacham and David P. Woodruff. Sketching algorithms and lower bounds for ridge regression. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 10539–10556. PMLR, 2022. 1.2.1

[KY05] Piyush Kumar and E Alper Yildirim. Minimum-volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and applications*, 126(1):1–21, 2005. 3.2

[LC14] Shan Luo and Zehua Chen. Sequential lasso cum EBIC for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*, 109(507):1229–1240, 2014. 16.1.1

[Lee16] Yin Tat Lee. *Faster algorithms for convex and combinatorial optimization*. PhD thesis, Massachusetts Institute of Technology, 2016. 6.1.3

[Lew78] D. R. Lewis. Finite dimensional subspaces of L_p. *Studia Mathematica*, 63(2):207–212, 1978. 1.5.2, 6.1.3, 6.1.4, 6.1.3, 6.1.6, 10.3

[LKD+17] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *5th International Conference on Learning Representations (ICLR)*, 2017. 16.1

[LL16] Vadim Lebedev and Victor Lempitsky. Fast convnets using group-wise brain damage. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2554–2564, 2016. 16.1

[LLAN06] Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y. Ng. Efficient L1 regularized logistic regression. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 401–408. AAAI Press, 2006. 8

[LLW23] Yi Li, Honghao Lin, and David P. Woodruff. The ℓ_p-subspace sketch problem in small dimensions with applications to support vector machines. In Nikhil Bansal and Viswanath Nagarajan, editors, *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023*, pages 850–877. SIAM, 2023. 6.1.3, 13.1.5

[LLY21] Yiwen Liao, Raphaël Latty, and Bin Yang. Feature selection using batch-wise attenuation and feature mask normalization. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021. (document), 16.1.2, 16.6.1, 16.6.1, 16.1, 16.6.3

[LMP13] Mu Li, Gary L. Miller, and Richard Peng. Iterative row sampling. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 127–136. IEEE Computer Society, 2013. 1.3.2, 6.1.1, 8

[LRG23] Quoc-Tung Le, Elisa Riccietti, and Rémi Gribonval. Spurious valleys, np-hardness, and tractability of sparse matrix factorization with fixed support. *SIAM J. Matrix Anal. Appl.*, 44(2):503–529, 2023. 18.0.4

[LRT21] Ismael Lemhadri, Feng Ruan, and Rob Tibshirani. Lassonet: Neural networks with feature sparsity. In *International Conference on Artificial Intelligence and Statistics*, pages 10–18. PMLR, 2021. 16.6.1, 16.6.3

[LS10] Michael Langberg and Leonard J. Schulman. Universal epsilon-approximators for integrals. In Moses Charikar, editor, *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 598–607. SIAM, 2010. 1.3.1, 1.3.1, 1.3.1, 6.1.1, 4

[LS15] Yin Tat Lee and He Sun. Constructing linear-sized spectral sparsification in almost-linear time. In Venkatesan Guruswami, editor, *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 250–269. IEEE Computer Society, 2015. 12.1, 15.1

[LS17] Edo Liberty and Maxim Sviridenko. Greedy minimization of weakly supermodular set functions. In Klaus Jansen, José D. P. Rolim, David Williamson, and Santosh S. Vempala, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2017, August 16-18, 2017, Berkeley, CA, USA*, volume 81 of *LIPIcs*, pages 19:1–19:11. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. 1.4.1, 1.4.2, 16.1, 16.1.2, 16.1.2, 16.1.4, 16.4

[LSW18] Roie Levin, Anish Prasad Sevekari, and David P. Woodruff. Robust subspace approximation in a stream. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10706–10716, 2018. 14.1

[LSY19] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architec-

ture search. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 16.1.2

[LT80] D. R. Lewis and Nicole Tomczak-Jaegermann. Hilbertian and complemented finite-dimensional subspaces of Banach lattices and unitary ideals. *J. Functional Analysis*, 35(2):165–190, 1980. 1.5.1, 1.5.1, 9

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 1991. 1.5.1, 1.5.1, 6.1.3, 6.1.4, 6.1.3, 6.1.3, 7.1, 8.1.3, 12.1

[LVW09] Aart Lagendijk, Bart Van-Tiggelen, and Diederik S Wiersma. Fifty years of anderson localization. *Phys. Today*, 62(8):24–29, 2009. 18

[LWW21] Yi Li, Ruosong Wang, and David P. Woodruff. Tight bounds for the subspace sketch problem with applications. *SIAM J. Comput.*, 50(4):1287–1335, 2021. 1.5.2, 6.1.3, 6.1.5, 8.0.4, 9, 9, 11.3, 13.1.3, 14.1, 14.1.1, 15.1, 15.2

[LWY21] Yi Li, David P. Woodruff, and Taisuke Yasuda. Exponentially improved dimensionality reduction for ℓ_1: Subspace embeddings and independence testing. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 3111–3195. PMLR, 2021. (document), 1.2.1, 1.2.3, 4, 4, 4.0.2, 5, 13.1.5

[Mag10] Malik Magdon-Ismail. Row sampling for matrix algorithms via a non-commutative bernstein bound. *CoRR*, abs/1008.0587, 2010. 1.3.2, 1.3.3, 1.3.2, 6.1.3

[Mag17] Malik Magdon-Ismail. Np-hardness and inapproximability of sparse PCA. *Inf. Process. Lett.*, 126:35–38, 2017. 18.0.4

[Mah11] Michael W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, 2011. 1.1

[MBLS18] Gonzalo E. Mena, David Belanger, Scott W. Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 16.1.2

[Mie09] Pauli Miettinen. *Matrix decomposition methods for data mining: Computational complexity and algorithms*. PhD thesis, University of Helsinki, 2009. 17

[Mil71] V. D. Milman. A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Priložen.*, 5(4):28–37, 1971. 13.1.5, 14.3.1

[Mir60] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *Quart. J. Math. Oxford Ser. (2)*, 11:50–59, 1960. 18

[MK21] Kakeru Mitsuno and Takio Kurita. Filter pruning using hierarchical group sparse regularization for deep convolutional neural networks. In *25th international conference on pattern recognition (ICPR)*, pages 1089–1095. IEEE, 2021. 16.1

[MM13] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In Dan Boneh, Tim

Roughgarden, and Joan Feigenbaum, editors, *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 91–100. ACM, 2013. 1.2.1, 3, 3, 3.1

[MM15] Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1396–1404, 2015. 1.4.3, 17, 18, 18.0.2, 18, 18, 18.1, 18.2.1, 18.2.2, 18.2.3, 18.2.3, 18.2.4, 18.2.9, 18.2.4

[MM20] Cameron Musco and Christopher Musco. Projection-cost-preserving sketches: Proof strategies and constructions. *CoRR*, abs/2004.08434, 2020. 14.1

[MMK20] Kakeru Mitsuno, Jun'ichi Miyao, and Takio Kurita. Hierarchical group sparse regularization for deep convolutional neural networks. In *International Joint Conference on Neural Networks IJCNN*, pages 1–8. IEEE, 2020. 16.1

[MMM+22] Raphael A. Meyer, Cameron Musco, Christopher Musco, David P. Woodruff, and Samson Zhou. Fast regression for structured inputs. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 7.2, 7.2.1, 7.2

[MMMW21] Raphael A. Meyer, Cameron Musco, Christopher Musco, and David P. Woodruff. Hutch++: Optimal stochastic trace estimation. In Hung Viet Le and Valerie King, editors, *4th Symposium on Simplicity in Algorithms, SOSA 2021, Virtual Conference, January 11-12, 2021*, pages 142–155. SIAM, 2021. 1.1.1

[MMO22] Yury Makarychev, Naren Sarayu Manoj, and Max Ovsiankin. Streaming algorithms for ellipsoidal approximation of convex polytopes. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 3070–3093. PMLR, 2022. 11

[MMO23] Yury Makarychev, Naren Sarayu Manoj, and Max Ovsiankin. Near-optimal streaming ellipsoidal rounding for general convex polytopes. *CoRR*, abs/2311.09460, 2023. 11

[MMR19] Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn. Performance of johnson-lindenstrauss transform for k-means and k-medians clustering. In Moses Charikar and Edith Cohen, editors, *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019*, pages 1027–1038. ACM, 2019. 1.1.1, 2.1.2

[MMR21] Tung Mai, Cameron Musco, and Anup Rao. Coresets for classification - simplified and strengthened. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11643–11654,

2021. 10

[MMWY22] Cameron Musco, Christopher Musco, David P. Woodruff, and Taisuke Yasuda. Active linear regression for ℓ_p norms and beyond. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, October 31 - November 3, 2022*, pages 744–753. IEEE, 2022. (document), 1.3.5, 6.1.3, 10, 10, 10.0.1, 12, 12.1, 12.1.3, 13.1.5, 13.4, 14.1, 14.3.2, 14.3.2, 17.1

[MO23] Naren Sarayu Manoj and Max Ovsiankin. The change-of-measure method, block lewis weights, and approximating matrix block norms. *CoRR*, abs/2311.10013, 2023. 7.1

[MOW21] Alexander Munteanu, Simon Omlor, and David P. Woodruff. Oblivious sketching for logistic regression. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7861–7871. PMLR, 2021. 1.2.1

[MOW23] Alexander Munteanu, Simon Omlor, and David Woodruff. Almost linear constant-factor sketching for ℓ_1 and logistic regression. In *The Eleventh International Conference on Learning Representations*, 2023. 1.2.1, 4

[MRWZ20] Sepideh Mahabadi, Ilya P. Razenshteyn, David P. Woodruff, and Samson Zhou. Non-adaptive adaptive sampling on turnstile streams. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 1251–1264. ACM, 2020. 11.4.1, 11.4.4, 11.4.12, 3, 11.4.14, 3, 14.1

[MSS10] Asish Mukhopadhyay, Animesh Sarker, and Tom Switzer. Approximate ellipsoid in the streaming model. In Weili Wu and Ovidiu Daescu, editors, *Combinatorial Optimization and Applications - 4th International Conference, COCOA 2010, Kailua-Kona, HI, USA, December 18-20, 2010, Proceedings, Part II*, volume 6509 of *Lecture Notes in Computer Science*, pages 401–413. Springer, 2010. 1.3.4, 11

[MSSW18] Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6562–6571, 2018. 8, 10

[MT20] Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numer.*, 29:403–572, 2020. 1.1

[MvdGB08] Lukas Meier, Sara van de Geer, and Peter Bühlmann. The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(1):53–71, 2008. 16.1.1, 8

[MW21] Arvind V. Mahankali and David P. Woodruff. Optimal ℓ_1 column subset selection and a fast PTAS for low rank approximation. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual*

Conference, January 10 - 13, 2021, pages 560–578. SIAM, 2021. 1.4.2, 5, 16.1.2, 17, 17.1, 17.1.2, 17.3, 17.3, 17.3.1, 17.4, 17.4.3, 17.4, 17.4, 19, 19

[MWA06] Baback Moghaddam, Yair Weiss, and Shai Avidan. Generalized spectral bounds for sparse LDA. In William W. Cohen and Andrew W. Moore, editors, *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 641–648. ACM, 2006. 18.0.4

[MWZ24] Arvind V. Mahankali, David P. Woodruff, and Ziyu Zhang. Near-linear time and fixed-parameter tractable algorithms for tensor decompositions. In Venkatesan Guruswami, editor, *15th Innovations in Theoretical Computer Science Conference, ITCS 2024, January 30 to February 2, 2024, Berkeley, CA, USA*, volume 287 of *LIPIcs*, pages 79:1–79:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024. 1.1.1

[Nat95] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995. 1.4.1, 7

[NH15] Rahul Nandkishore and David A Huse. Many-body localization and thermalization in quantum statistical mechanics. *Annu. Rev. Condens. Matter Phys.*, 6(1):15–38, 2015. 18

[NLS16] Arvind Neelakantan, Quoc V. Le, and Ilya Sutskever. Neural programmer: Inducing latent programs with gradient descent. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. 16.1.2

[NN13] Jelani Nelson and Huy L. Nguyen. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 117–126. IEEE Computer Society, 2013. 1.2.2, 3.1, 14.3.2

[NN14] Jelani Nelson and Huy L. Nguyên. Lower bounds for oblivious subspace embeddings. In Javier Esparza, Pierre Fraigniaud, Thore Husfeldt, and Elias Koutsoupias, editors, *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, volume 8572 of *Lecture Notes in Computer Science*, pages 883–894. Springer, 2014. 1.2.2

[Nol20] John P. Nolan. *Univariate stable distributions: models for heavy tailed data*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, [2020] ©2020. 3.0.1, 3.3, 5, 17.3

[NUD17] Sharan Narang, Eric Undersander, and Gregory F. Diamos. Block-sparse recurrent neural networks. *CoRR*, abs/1711.02782, 2017. 16.1

[OPT00] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2):319–337, 2000. 8

[PC18] Romualdo Pastor-Satorras and Claudio Castellano. Eigenvector localization in real networks and its implications for epidemic spreading. *Journal of Statistical Physics*,

173(3):1110–1123, 2018. 18

[Pie80] Albrecht Pietsch. *Operator ideals*, volume 20 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam-New York, 1980. 1.5.1

[PPP21] Aditya Parulekar, Advait Parulekar, and Eric Price. L1 regression with Lewis weights subsampling. In Mary Wootters and Laura Sanità, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2021, August 16-18, 2021, University of Washington, Seattle, Washington, USA (Virtual Conference)*, volume 207 of *LIPIcs*, pages 49:1–49:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. 12.1, 12.6, 13.1.5

[PRK93] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993. 16.1.1

[PRPG22] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 17359–17371. PMLR, 2022. 16.1

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014. 16.1

[PSZ22] Eric Price, Sandeep Silwal, and Samson Zhou. Hardness and algorithms for robust and sparse optimization. In *International Conference on Machine Learning*, pages 17926–17944. PMLR, 2022. 1.4.1, 7, 16.1.1

[PTB13] Udaya Parampalli, Xiaohu Tang, and Serdar Boztas. On the construction of binary sequence families with low correlation and large sizes. *IEEE Trans. Inf. Theory*, 59(2):1082–1089, 2013. 11.3.2

[Puk06] Friedrich Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006. 12.1

[PVZ17] Grigoris Paouris, Petros Valettas, and Joel Zinn. Random version of Dvoretzky's theorem in ℓ_p^n. *Stochastic Process. Appl.*, 127(10):3187–3227, 2017. 4, 14.1.1, 14.3.1

[PWZ23] Swati Padmanabhan, David P. Woodruff, and Qiuyi (Richard) Zhang. Computing approximate ℓ_p sensitivities. *CoRR*, abs/2311.04158, 2023. 6.1.1, 6

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*,

52(3):471–501, 2010. 16.1.5

[Roc70] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970. 16.2.4

[RPYU18] Mengye Ren, Andrei Pokrovsky, Bin Yang, and Raquel Urtasun. Sbnet: Sparse blocks network for fast inference. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8711–8720. Computer Vision Foundation / IEEE Computer Society, 2018. 16.1

[RV07] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4):21, 2007. 1.3.2, 1.3.3, 1.3.2, 6.1.3

[RV09] Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009. 14.5.2, 14.5.2

[RWY10] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, 2010. 1.4.1, 16.1.1

[RZH04] Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *J. Mach. Learn. Res.*, 5:941–973, 2004. 16.1.3

[Sar06] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 143–152. IEEE Computer Society, 2006. 1.1.2, 1.2.2, 1.2.3, 1.2.2, 12.1

[Sch87] Gideon Schechtman. More on embedding subspaces of L_p in l_r^n. *Compositio Math.*, 61(2):159–169, 1987. 1.5.1

[Sch07] Gideon Schechtman. Aimpl: Fourier analytic methods in convex geometry, available at http://aimpl.org/fourierconvex/1/, 2007. 6

[Sch11] Gideon Schechtman. Tight embedding of subspaces of L_p in ℓ_p^n for even p. *Proc. Amer. Math. Soc.*, 139(12):4419–4421, 2011. 1.5.1, 1

[SCHU17] Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017. 16.1

[SFR07] Mark Schmidt, Glenn Fung, and Rómer Rosales. Fast optimization methods for L1 regularization: A comparative study and two new approaches. In Joost N. Kok, Jacek Koronacki, Ramón López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *Machine Learning: ECML 2007, 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007, Proceedings*, volume 4701 of *Lecture Notes in Computer Science*, pages 286–297. Springer, 2007. 8

[SS02] Michael E. Saks and Xiaodong Sun. Space lower bounds for distance approximation

in the data stream model. In John H. Reif, editor, *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 360–369. ACM, 2002. 3

[SS11] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011. 1.3.2, 6.1.1, 8, 4

[SSZ10] Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM J. Optim.*, 20(6):2807–2832, 2010. 1.4.1, 16.1, 16.1.2, 16.1.4, 16.4

[SV12] Nariankadu D. Shyamalkumar and Kasturi R. Varadarajan. Efficient subspace approximation algorithms. *Discret. Comput. Geom.*, 47(1):44–63, 2012. 13.1.6, 13.1.6, 14.1, 14.1, 14.1, 15.2

[SVW15] Maxim Sviridenko, Jan Vondrák, and Justin Ward. Optimal approximation for submodular and supermodular optimization with bounded curvature. In Piotr Indyk, editor, *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 1134–1148. SIAM, 2015. 1.4.2, 16.1.2

[SW11] Christian Sohler and David P. Woodruff. Subspace embeddings for the l_1-norm with applications. In Lance Fortnow and Salil P. Vadhan, editors, *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 755–764. ACM, 2011. 1.2.1, 1.5.2, 3, 3.1, 3.3, 3.3, 4, 4.1

[SW18] Christian Sohler and David P. Woodruff. Strong coresets for k-median and subspace approximation: Goodbye dimension. In Mikkel Thorup, editor, *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 802–813. IEEE Computer Society, 2018. 13.1.6, 14.1, 14.1, 14.1, 4, 4, 4, 14.1.1, 14.1.1, 5, 14.2, 14.2, 14.2.1, 14.2.1, 14.2.1, 14.2.2, 14.2.2

[SWMW89] Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989. 12.1

[SWY75] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975. 16.1

[SWY+19] Zhao Song, Ruosong Wang, Lin F. Yang, Hongyang Zhang, and Peilin Zhong. Efficient symmetric norm regression via linear sketching. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 828–838, 2019. 10.2

[SWZ17] Zhao Song, David P. Woodruff, and Peilin Zhong. Low rank approximation with entrywise ℓ_1-norm error. In Hamed Hatami, Pierre McKenzie, and Valerie King, editors, *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 688–701. ACM, 2017. 1.4.2, 5, 16.1.2, 17.1, 17.3, 17.3, 17.3.2, 17.3.2, 17.3.2, 19

[SWZ19] Zhao Song, David P. Woodruff, and Peilin Zhong. Towards a zero-one law for column subset selection. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 6120–6131, 2019. (document), 1.4.2, 10.3, 16.1.2, 17.1, 17.1, 17.1, 17.1.1, 17.1.1, 17.1.2, 17.1.2, 17.1.2, 17.1.2, 17.3.1, 17.4

[SZ01] Gideon Schechtman and Artem Zvavitch. Embedding subspaces of l_p into l_p^n, $0 < p < 1$. *Mathematische Nachrichten*, 227(1):133–142, 2001. 1.5.1, 1.5.1, 6.1.3, 6.1.4, 6.1.3, 10.3, 12.1

[Tal90] Michel Talagrand. Embedding subspaces of L_1 into l_1^N. *Proc. Amer. Math. Soc.*, 108(2):363–369, 1990. 1.5.1, 1.5.1, 6.1.3, 6.1.3

[Tal95] Michel Talagrand. Embedding subspaces of L_p in l_p^N. In *Geometric aspects of functional analysis (Israel, 1992–1994)*, volume 77 of *Oper. Theory Adv. Appl.*, pages 311–325. Birkhäuser, Basel, 1995. 1.5.1, 1.5.1, 6.1.3, 6.1.3

[TB97] Lloyd N. Trefethen and David Bau, III. *Numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. 1.1

[TBF+12] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(2):245–266, 2012. 8, 2, 16.1.2

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. 1.4.1, 7

[Tib15] Ryan J. Tibshirani. A general framework for fast stagewise algorithms. *J. Mach. Learn. Res.*, 16:2543–2588, 2015. 16.1.3

[TMF20] Murad Tukan, Alaa Maalouf, and Dan Feldman. Coresets for near-convex functions. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 10, 10.2.5

[Tod16] Michael J. Todd. *Minimum volume ellipsoids - theory and algorithms*, volume 23 of *MOS-SIAM Series on Optimization*. SIAM, 2016. 3.2, 11, 11.4.2, 11.4.4, 11.4.3, 11.4.3

[Tro04] Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inf. Theory*, 50(10):2231–2242, 2004. 16.1.1

[Tro06] Joel A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inf. Theory*, 52(3):1030–1051, 2006. 1.4.1, 16.1.1

[TT11] Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011. 8

[Ver18] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in*

Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018. 2.3.6

[vH14] Ramon van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014. 2.3.3, 2.3.3

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 16.1.2

[VX12] Kasturi R. Varadarajan and Xin Xiao. On the sensitivity of shape fitting problems. In Deepak D'Souza, Telikepalli Kavitha, and Jaikumar Radhakrishnan, editors, *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2012, December 15-17, 2012, Hyderabad, India*, volume 18 of *LIPIcs*, pages 486–497. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2012. 14.1, 4, 4, 14.1.1

[WC20] Maksymilian Wojtas and Ke Chen. Feature importance ranking for deep learning. *Advances in Neural Information Processing Systems*, 33:5105–5114, 2020. 16.6.1, 16.6.3

[WDL+09] Kilian Q. Weinberger, Anirban Dasgupta, John Langford, Alexander J. Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 1113–1120. ACM, 2009. 16.1

[Wel74] Lloyd R. Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Trans. Inf. Theory*, 20(3):397–399, 1974. 11.3.5

[Woj91] P. Wojtaszczyk. *Banach spaces for analysts*, volume 25 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1991. 9.2

[Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1-2):1–157, 2014. 1.1, 3.1

[WW19] Ruosong Wang and David P. Woodruff. Tight bounds for ℓ_p oblivious subspace embeddings. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 1825–1843. SIAM, 2019. 1.2.1, 3, 3.0.2, 3, 3.0.3, 3, 3.1, 3.3, 3.3.1, 4, 4.0.1, 4.1, 4.2.5, 4.4, 4.4.2, 5, 13.1.5

[WW22] Ruosong Wang and David P. Woodruff. Tight bounds for ℓ_1 oblivious subspace embeddings. *ACM Trans. Algorithms*, 18(1):8:1–8:32, 2022. 3, 3.0.2, 3, 3.0.3, 3, 3.1, 3.3, 4, 4.0.1, 4.1, 4.4, 4.4.2, 5

[WWW+16] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *Advances in Neural Information Processing Systems*, 29, 2016. 16.1

[WXXZ23] Virginia Vassilevska Williams, Yinzhan Xu, Zixuan Xu, and Renfei Zhou. New

bounds for matrix multiplication: from alpha to omega. *CoRR*, abs/2307.07970, 2023. 1

[WY22a] David P. Woodruff and Taisuke Yasuda. High-dimensional geometric streaming in polynomial space. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, October 31 - November 3, 2022*, pages 732–743. IEEE, 2022. (document), 1.1.1, 1.3.3, 9, 9, 9.0.2, 9.0.3, 11, 11.0.1

[WY22b] David P. Woodruff and Taisuke Yasuda. Improved algorithms for low rank approximation from sparsity. In Joseph (Seffi) Naor and Niv Buchbinder, editors, *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, pages 2358–2403. SIAM, 2022. (document), 1.3.3, 1.3.5, 6.1.7, 6.6.3, 18, 18, 18.0.3

[WY23a] David P. Woodruff and Taisuke Yasuda. New subset selection algorithms for low rank approximation: Offline and online. In Barna Saha and Rocco A. Servedio, editors, *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023*, pages 1802–1813. ACM, 2023. (document), 1.2.1, 1.2.3, 1.3.5, 1.4.2, 1.4.3, 3, 3, 3.0.4, 3.2, 3.2.1, 12, 12.1, 12.1.3, 13.1.6, 14, 14.1, 14.1, 14.1.1, 14.1.2, 5, 5, 14.3.2, 14.3.2, 14.6.1, 15.2, 15.2, 15.2, 17, 17.3.2, 19

[WY23b] David P. Woodruff and Taisuke Yasuda. Online lewis weight sampling. In Nikhil Bansal and Viswanath Nagarajan, editors, *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22-25, 2023*, pages 4622–4666. SIAM, 2023. (document), 1.3.5, 6, 6.1.3, 6.1.3, 6.1.3, 6.1.11, 6.3.3, 12.4, 4, 14.3.2, 14.6.1, 16.1.2

[WY23c] David P. Woodruff and Taisuke Yasuda. Sharper bounds for ℓ_p sensitivity sampling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 37238–37272. PMLR, 2023. (document), 1.3.5, 4.1.2, 6.4.2, 7, 7.1, 7.1.2, 8, 8, 8.0.2, 8.1.3, 12.5.2, 14.5.1, 6

[WY24a] David P. Woodruff and Taisuke Yasuda. Coresets for multiple ℓ_p regression. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2024*, Proceedings of Machine Learning Research. PMLR, 2024. (document), 1.3.5, 12.1, 13, 15.2

[WY24b] David P. Woodruff and Taisuke Yasuda. Nearly linear sparsification of ℓ_p subspace approximation, 2024. (document), 1.3.5, 8, 8.0.1, 8, 8.0.3, 13.1.6, 14, 15.2, 15.2

[WZ13] David P. Woodruff and Qin Zhang. Subspace embeddings and ℓ_p-regression using exponential random variables. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 546–567. JMLR.org, 2013. 1.2.1, 3, 3, 3.1, 3.3

[YBC+23] Taisuke Yasuda, Mohammad Hossein Bateni, Lin Chen, Matthew Fahrbach, Gang

Fu, and Vahab Mirrokni. Sequential attention for feature selection. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. (document), 1.4.3, 16, 16.1.1, 8, 8, 16.1.2, 16.1.2, 16.1.2, 16.1.2, 16.1.2, 16.6

[YL06] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006. 7

[ZC06] Hamid Zarrabi-Zadeh and Timothy M. Chan. A simple streaming algorithm for minimum enclosing balls. In *Proceedings of the 18th Annual Canadian Conference on Computational Geometry, CCCG 2006, August 14-16, 2006, Queen's University, Ontario, Canada*, 2006. 11.4

[Zho09] Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009. 1.4.1, 16.1.1

[Zou12] Anastasios Zouzias. A matrix hyperbolic cosine algorithm and applications. In Artur Czumaj, Kurt Mehlhorn, Andrew M. Pitts, and Roger Wattenhofer, editors, *Automata, Languages, and Programming - 39th International Colloquium, ICALP 2012, Warwick, UK, July 9-13, 2012, Proceedings, Part I*, volume 7391 of *Lecture Notes in Computer Science*, pages 846–858. Springer, 2012. 15.1, 15.1

[Zva00] A. Zvavitch. More on embedding subspaces of L_p into l_p^N, $0 < p < 1$. In *Geometric aspects of functional analysis*, volume 1745 of *Lecture Notes in Math.*, pages 269–280. Springer, Berlin, 2000. 1.5.1, 6.1.3

[ZY06] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006. 16.1.1

[ZYC+21] Li Zhou, Lihao Yan, Mark A. Caprio, Weiguo Gao, and Chao Yang. Solving the k-sparse eigenvalue problem with reinforcement learning. *CSIAM Transactions on Applied Mathematics*, 2(4):697–723, 2021. 18