

Fast Randomized Algorithms for Convex Optimization and Statistical Estimation

Mert Pilanci

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2016-147

<http://www.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-147.html>

August 14, 2016



Copyright © 2016, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

**Fast Randomized Algorithms for Convex Optimization
and Statistical Estimation**

by

Mert Pilanci

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martin J. Wainwright, Co-chair
Professor Laurent El Ghaoui, Co-chair
Assistant Professor Aditya Guntuboyina

Summer 2016

Fast Randomized Algorithms for Convex Optimization
and Statistical Estimation

Copyright © 2016

by

Mert Pilanci

Abstract

Fast Randomized Algorithms for Convex Optimization and Statistical Estimation

by

Mert Pilanci

Doctor of Philosophy in Engineering – Electrical Engineering and Computer
Sciences

University of California, Berkeley

Professor Martin J. Wainwright, Co-chair

Professor Laurent El Ghaoui, Co-chair

With the advent of massive datasets, statistical learning and information processing techniques are expected to enable exceptional possibilities for engineering, data intensive sciences and better decision making. Unfortunately, existing algorithms for mathematical optimization, which is the core component in these techniques, often prove ineffective for scaling to the extent of all available data. In recent years, randomized dimension reduction has proven to be a very powerful tool for approximate computations over large datasets. In this thesis, we consider random projection methods in the context of general convex optimization problems on massive datasets. We explore many applications in machine learning, statistics and decision making and analyze various forms of randomization in detail. The central contributions of this thesis are as follows:

- We develop random projection methods for convex optimization problems and establish fundamental trade-offs between the size of the projection and accuracy of solution in convex optimization.
- We characterize information-theoretic limitations of methods that are based on random projection, which surprisingly shows that the most widely used form of random projection is, in fact, statistically sub-optimal.
- We present novel methods, which iteratively refine the solutions to achieve statistical optimality and enable solving large scale optimization and statistical inference problems orders-of-magnitude faster than existing methods.

- We develop new randomized methodologies for relaxing cardinality constraints in order to obtain checkable and more accurate approximations than the state of the art approaches.

To my family and Ilge

Contents

Contents	ii
List of Figures	viii
List of Tables	xii
Acknowledgements	xiii
1 Introduction	1
1.1 Motivation and background	1
1.1.1 Convex optimization	2
1.1.2 Empirical risk minimization	2
1.1.3 Minimax theory	3
1.1.4 Random projection	4
1.1.5 Sketching data streams and matrices	4
1.1.6 Different kinds of sketches	5
1.2 Goals and contributions of this thesis	7
1.2.1 Thesis organization and previously published work	8
1.2.2 Notation	8
2 Random projections of convex quadratic programs	9
2.1 Problem formulation	10
2.1.1 Guarantees for sub-Gaussian sketches	12
2.1.2 Guarantees for randomized orthogonal systems	14

2.2	Applications	14
2.2.1	Unconstrained least squares	15
2.2.2	ℓ_1 -constrained least squares	16
2.2.3	Compressed sensing and noise folding	21
2.2.4	Support vector machine classification	23
2.2.5	Matrix estimation with nuclear norm regularization	25
2.2.6	Group sparse regularization	28
2.3	Proofs of main results	29
2.3.1	Main argument	30
2.3.2	Proof of Theorem 1	32
2.3.3	Proof of Theorem 2	36
2.4	Techniques for sharpening bounds	40
2.4.1	Sharpening bounds for a subspace	41
2.4.2	Reduction to finite maximum	42
2.4.3	Sharpening ℓ_1 -based bounds	43
2.4.4	Sharpening nuclear norm bounds	47
2.5	Discussion	49
2.6	Proofs of technical results	50
2.6.1	Technical details for Corollary 3	50
2.6.2	Technical lemmas for Proposition 2	51
2.6.3	Proof of Lemma 7	51
2.6.4	Proof of Lemma 8	52
2.6.5	Proof of Lemma 9	53
2.6.6	A technical inclusion lemma	55
3	Iterative random projections and information theoretical bounds	57
3.1	Main results and consequences	60
3.1.1	Types of randomized sketches	60
3.1.2	Information-theoretical sub-optimality of the classical sketch	61
3.1.3	Introducing the Hessian sketch	66

3.1.4	Iterative Hessian sketch	68
3.1.5	Computational and space complexity	72
3.2	Consequences for concrete models	73
3.2.1	Unconstrained least squares	73
3.2.2	Sparse least-squares	75
3.2.3	Some larger-scale experiments	78
3.2.4	Matrix estimation with nuclear norm constraints	79
3.3	Discussion	82
3.4	Proof of lower bounds	84
3.4.1	Verification of condition (3.8)	84
3.4.2	Proof of Theorem 3	85
3.5	Proof of Proposition 3	87
3.6	Proof of Theorem 4	87
3.7	Maximum likelihood estimator and examples	88
3.7.1	Upper bound on MLE	88
3.7.2	Detailed calculations for illustrative examples	89
3.7.3	Proof of Proposition 4	91
4	Random projections for nonlinear optimization	94
4.1	Background	96
4.1.1	Classical version of Newton’s method	96
4.1.2	Approximate Newton methods	97
4.2	Newton Sketch and local convergence	98
4.2.1	Newton Sketch algorithm	98
4.2.2	Affine invariance of the Newton Sketch and sketched KKT systems	100
4.2.3	Some examples	100
4.2.4	Local convergence analysis using strong convexity	102
4.3	Newton Sketch for self-concordant functions	105
4.3.1	Unconstrained case	105
4.3.2	Newton Sketch with self-concordant barriers	107

4.3.3	Sketching with interior point methods	109
4.4	Applications and numerical results	110
4.4.1	Estimation in generalized linear models	111
4.4.2	Semidefinite programs	112
4.4.3	Portfolio optimization and SVMs	113
4.4.4	Unconstrained logistic regression with $d \ll n$	114
4.4.5	ℓ_1 -constrained logistic regression and data conditioning	115
4.4.6	A dual example: Lasso with $d \gg n$	116
4.5	Proofs of main results	117
4.5.1	Proof of Theorem 5	117
4.5.2	Proof of Theorem 6	121
4.5.3	Proof of Theorem 7	126
4.6	Discussion	126
4.7	Proofs of technical results	127
4.7.1	Technical results for Theorem 6	127
4.7.2	Proof of Lemma 24	132
4.7.3	Gaussian widths with ℓ_1 -constraints	132
5	Random projection, effective dimension and nonparametric regression	141
5.1	Problem formulation and background	143
5.1.1	Regression in reproducing kernel Hilbert spaces	143
5.1.2	Kernel ridge regression and its sketched form	144
5.1.3	Kernel complexity measures and statistical guarantees	145
5.2	Main results and their consequences	148
5.2.1	General conditions for sketched kernel optimality	149
5.2.2	Corollaries for randomized sketches	150
5.2.3	Comparison with Nyström-based approaches	153
5.3	Proofs of technical results	157
5.3.1	Proof of Theorem 10	157
5.4	Discussion	161

5.5	Subsampling sketches yield Nyström approximation	161
5.6	Proofs of technical results	162
5.6.1	Proof of Theorem 9	162
5.6.2	Proof of Lemma 28	165
5.6.3	Proof of Corollary 13	168
6	Relaxations of combinatorial optimization problems	171
6.1	General Sparse Learning as a Boolean Problem	172
6.1.1	Exact representation as a Boolean convex program	175
6.2	Convex-analytic conditions for IR exactness	177
6.2.1	Sufficient conditions for random ensembles	180
6.2.2	Analysis of randomized rounding	180
6.3	Penalized forms of cardinality	183
6.3.1	Reformulation as Boolean program	183
6.3.2	Least-squares regression	184
6.4	Numerical Results	186
6.4.1	Optimization techniques	187
6.4.2	Experiments on real datasets	187
6.5	Simplex Constrained Problems	188
6.6	Optimizing over sparse probability measures	189
6.6.1	Computing a bound on the quality of approximation	190
6.7	Recovering a Sparse Measure	191
6.7.1	An alternative minimal cardinality selection scheme	192
6.7.2	Noisy measure recovery	193
6.8	Convex Clustering	193
6.9	Algorithms	195
6.9.1	Exponentiated Gradient	195
6.10	Numerical Results	195
6.10.1	Recovering a Measure from Gaussian Measurements	195
6.10.2	Convex Clustering	197

6.11 Discussion	198
6.12 Proofs of technical results	199
6.12.1 Proof of Theorem 12	199
6.12.2 Proof of Theorem 13	203
Bibliography	205

List of Figures

2.1	Tangent cone at x^*	12
2.2	Comparison of Gaussian, Rademacher and randomized Hadamard sketches for unconstrained least squares. Each curve plots the approximation ratio $f(\hat{x})/f(x^*)$ versus the control parameter α , averaged over $T_{trial} = 100$ trials, for projection dimensions $m = \max\{1.5\alpha d, 1\}$ and for problem dimensions $d = 500$ and $n \in \{1024, 2048, 4096\}$	17
2.3	Comparison of Gaussian, Rademacher and randomized Hadamard sketches for the Lasso program (2.12). Each curve plots the approximation ratio $f(\hat{x})/f(x^*)$ versus the control parameter α , averaged over $T_{trial} = 100$ trials, for projection dimensions $m = \max\{4\alpha\ x^*\ _0 \log d, 1\}$, problem dimensions $(n, d) = (4096, 500)$, and ℓ_1 -constraint radius $R \in \{1, 5, 10, 20\}$	20
2.4	Comparison of Gaussian, Rademacher and randomized Hadamard sketches for the support vector machine (2.27). Each curve plots the approximation ratio $f(\hat{x})/f(x^*)$ versus the control parameter α , averaged over $T_{trial} = 100$ trials, for projection dimensions $m = \max\{5\alpha\ x^*\ _0 \log d, 1\}$, and problem dimensions $d \in \{1024, 2048, 4096\}$	24
3.1	Plots of mean-squared error versus the row dimension $n \in \{100, 200, 400, \dots, 25600\}$ for unconstrained least-squares in dimension $d = 10$	59
3.2	Simulations of the IHS algorithm for an unconstrained least-squares problem with noise variance $\sigma^2 = 1$, and of dimensions $(d, n) = (200, 6000)$	71
3.3	Simulations of the IHS algorithm for unconstrained least-squares.	75
3.4	Plots of the log error $\ x^t - x^{LS}\ _2$ (a) and $\ x^t - x^*\ _2$ (b) versus the iteration number t	76
3.5	Simulations of the IHS algorithm for ℓ_1 -constrained least-squares	77

3.6	Plots of the mean-squared prediction errors $\frac{\ A(\tilde{x}-x^*)\ _2^2}{n}$ versus the sample size $n \in 2^{\{9,10,\dots,19\}}$ for the original least-squares solution ($\tilde{x} = x^{\text{LS}}$ in blue) versus the sketched solution ($\hat{x} = x^{\text{LS}}$ in red).	80
3.7	Japanese Female Facial Expression (JAFFE) Database: The JAFFE database consists of 213 images of 7 different emotional facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models.	81
3.8	Simulations of the IHS algorithm for nuclear-norm constrained problems on the JAFFE dataset: Mean-squared error versus the row dimension $n \in [10, 100]$ for recovering a 20×20 matrix of rank $r=2$, using a sketch dimension $m = 60$ (a). Classification error rate versus regularization parameter $R \in \{1, \dots, 12\}$, with error bars corresponding to one standard deviation over the test set (b).	83
4.1	Comparisons of central paths for a simple linear program in two dimensions. Each row shows three independent trials for a given sketch dimension: across the rows, the sketch dimension ranges as $m \in \{d, 4d, 16d\}$. The black arrows show Newton steps taken by the standard interior point method, whereas red arrows show the steps taken by the sketched version. The green point at the vertex represents the optimum. In all cases, the sketched algorithm converges to the optimum, and as the sketch dimension m increases, the sketched central path converges to the standard central path.	134
4.2	Empirical illustration of the linear convergence of the Newton Sketch algorithm for an ensemble of portfolio optimization problems (4.15). In all cases, the algorithm was implemented using a sketch dimension $m = \lceil 4s \log d \rceil$, where s is an upper bound on the number of non-zeros in the optimal solution x^* ; this quantity satisfies the required lower bound (4.12), and consistent with the theory, the algorithm displays linear convergence.	135
4.3	Comparison of Newton Sketch with various other algorithms in the logistic regression problem with Gaussian data.	136
4.4	Comparison of Newton Sketch with other algorithms in the logistic regression problem with Student's t-distributed data	137
4.5	The performance of Newton Sketch is independent of condition numbers and problem related quantities. Plots of the number of iterations required to reach 10^{-6} accuracy in ℓ_1 -constrained logistic regression using Newton's Method and Projected Gradient Descent using line search.	138

4.6	Plots of the duality gap versus iteration number (top panel) and duality gap versus wall-clock time (bottom panel) for the original barrier method (blue) and sketched barrier method (red). The sketched interior point method is run 10 times independently yielding slightly different curves in red. While the sketched method requires more iterations, its overall wall-clock time is much smaller.	139
4.7	Plot of the wall-clock time in seconds for reaching a duality gap of 10^{-6} for the standard and sketched interior point methods as n increases (in log-scale). The sketched interior point method has significantly lower computation time compared to the original method.	140
5.1	Prediction error versus sample size for original KRR, Gaussian sketch, and ROS sketches for the Sobolev one kernel for the function $f^*(x) = 1.6 (x - 0.4)(x - 0.6) - 0.3$. In all cases, each point corresponds to the average of 100 trials, with standard errors also shown. (a) Squared prediction error $\ \hat{f} - f^*\ _n^2$ versus the sample size $n \in \{32, 64, 128, \dots, 16384\}$ for projection dimension $m = \lceil n^{1/3} \rceil$. (b) Rescaled prediction error $n^{2/3}\ \hat{f} - f^*\ _n^2$ versus the sample size. (c) Runtime versus the sample size. (d) Relative approximation error $\ \hat{f} - f^\diamond\ _n^2 / \ f^\diamond - f^*\ _n^2$ versus scaling parameter c for $n = 1024$ and $m = \lceil cn^{1/3} \rceil$ with $c \in \{0.5, 1, 2, \dots, 7\}$. The original KRR under $n = 8192$ and 16384 are not computed due to out-of-memory failures.	153
5.2	Prediction error versus sample size for original KRR, Gaussian sketch, and ROS sketches for the Gaussian kernel with the function $f^*(x) = 0.5e^{-x_1+x_2} - x_2x_3$. In all cases, each point corresponds to the average of 100 trials, with standard errors also shown. (a) Squared prediction error $\ \hat{f} - f^*\ _n^2$ versus the sample size $n \in \{32, 64, 128, \dots, 16384\}$ for projection dimension $m = \lceil 1.25(\log n)^{3/2} \rceil$. (b) Rescaled prediction error $\frac{n}{(\log n)^{3/2}}\ \hat{f} - f^*\ _n^2$ versus the sample size. (c) Runtime versus the sample size. (d) Relative approximation error $\ \hat{f} - f^\diamond\ _n^2 / \ f^\diamond - f^*\ _n^2$ versus scaling parameter c for $n = 1024$ and $m = \lceil c(\log n)^{3/2} \rceil$ with $c \in \{0.5, 1, 2, \dots, 7\}$. The original KRR under $n = 8192$ and 16384 are not computed due to out-of-memory failures.	154

5.3	Prediction error versus sample size for original KRR, Gaussian sketch, ROS sketch and Nyström approximation. Left panels (a) and (c) shows $\ \hat{f} - f^*\ _n^2$ versus the sample size $n \in \{32, 64, 128, 256, 512, 1024\}$ for projection dimension $m = \lceil 4\sqrt{\log n} \rceil$. In all cases, each point corresponds to the average of 100 trials, with standard errors also shown. Right panels (b) and (d) show the rescaled prediction error $\frac{n}{\sqrt{\log n}} \ \hat{f} - f^*\ _n^2$ versus the sample size. Top row correspond to covariates arranged uniformly on the unit interval, whereas bottom row corresponds to an irregular design (see text for details).	156
6.1	Problem of exact support recovery for the Lasso and the interval relaxation for different problem sizes $d \in \{64, 128, 256\}$. As predicted by theory, both methods undergo a phase transition from failure to success once the control parameter $\alpha := \frac{n}{k \log(d-k)}$ is sufficiently large. This behavior is confirmed for the interval relaxation in Theorem 12.	181
6.2	Plots of three different penalty functions as a function of $t \in \mathbb{R}$: reverse Huber (berhu) function $t \mapsto B(\frac{\sqrt{\rho}t}{\lambda})$ ℓ_1 -norm $t \mapsto \lambda t $ and the ℓ_0 -based penalty $t \mapsto \frac{t^2}{2} + \lambda\ t\ _0$	184
6.3	Objective value versus cardinality trade-off in a real dataset from cancer research. The proposed randomized rounding method considerably outperforms other methods by achieving lower objective value with smaller cardinality.	186
6.4	Classification accuracy versus cardinality in a real dataset from cancer research. The proposed method has considerably higher classification accuracy for a fixed cardinality.	188
6.5	Probability simplex and the reciprocal of the infinity norm . The sparsest probability distribution on the set C is x^* (green) which also minimizes $\frac{1}{\max_i x_i}$ on the intersection (red)	190
6.6	A comparison of the exact recovery probability in the noiseless setting (top) and estimation error in the noisy setting (bottom) of the proposed approach and the rescaled ℓ_1 heuristic	196
6.7	Proposed convex clustering scheme	197
6.8	Soft k-means algorithm	198

List of Tables

- 3.1 Running time comparison in seconds of the Baseline (homotopy method applied to original problem), IHS (homotopy method applied to sketched subproblems), and IHS plus sketching time. Each running time estimate corresponds to an average over 300 independent trials of the random sparse regression model described in the main text. . . . 79

Acknowledgements

This thesis owes its existence to the guidance, support and inspiration of several people. I am glad to acknowledge their contributions here, and apologize if I forgot to mention anyone.

Firstly, I am greatly indebted to my two advisors Martin Wainwright and Laurent El Ghaoui for guiding and supporting me throughout my graduate studies. They have set an example of excellence as professors, mentors and role models. They are extremely knowledgeable, friendly, patient and very enthusiastic about new ideas. They encouraged me to pursue diverse research directions, and I have been fortunate to freely choose research topics that interest me most.

I would like to express my sincere gratitude for Professor Orhan Arikan and Professor Erdal Arikan. I was very fortunate to have worked with them in Bilkent University. They have continued to be great mentors during my time at Berkeley and provided unparalleled perspective in our frequent discussions.

I would like to thank Professor Michael Jordan for being the chair of my Qualification Exam committee and also Professor Aditya Guntuboyina for valuable feedback and suggestions.

I am grateful for Microsoft Research for funding my studies through a generous MSR PhD Fellowship. I also had an enjoyable internship at Microsoft Research at Redmond. I would like thank my mentor Ofer Dekel for his guidance and valuable insights. I spent a wonderful summer at INRIA Research Center of Paris as a visiting researcher. I was very fortunate to work with Francis Bach who was a great mentor and colleague.

I have been extremely fortunate to be surrounded constantly by other wonderful students and colleagues at Berkeley. I would like to thank my peer fellows in the EECS and Statistics department: Nihar Shah, Rashmi Vinayak, Yuchen Zhang, Yun Yang, Venkat Chandrasekaran, Sivaraman Balakrishnan, Anh Pham, Vu Pham, Raaz Dwivedi, Ashwin Pananjady, Orhan Ocal, Andrew Godbehere and many others.

Finally, I would like to thank my parents, grandparents, my brother and my wife-to-be Ilge for their unconditional love and support. I undoubtedly could not have achieved this without them.

Chapter 1

Introduction

1.1 Motivation and background

As a result of the rapid growth of information sources, today's computing devices face unprecedented volumes of data. In fact, 90% of all the data in the world today has been generated within the last two years¹. With the advent of massive datasets, new possibilities for better decision making are unraveled via statistical learning and information processing techniques. Unfortunately, existing algorithms for mathematical optimization, which is the core component in these techniques, often prove ineffective for scaling to the extent of all available data. However, we can address problems at much larger scales by considering fundamental changes in how we access the data and design the underlying algorithms. For instance, we may prefer non-deterministic algorithms for better computational and statistical trade-offs compared to deterministic algorithms.

In this thesis we consider novel randomized algorithms and a theoretical framework that enable faster mathematical optimization and statistical estimation for large datasets. The key idea is to employ a carefully designed randomness in the data reading process to gather the *essence* of data without accessing it in entirety. We consider many applications in machine learning, data driven decision making and signal processing, then discuss theoretical and practical implications of the developed methods in detail.

¹ Big Data at the Speed of Business. IBM.com

1.1.1 Convex optimization

Mathematical optimization is a branch of applied mathematics focused on minimization or maximization of certain functions, potentially subject to given constraints. Convex optimization is a special class of mathematical optimization which has found wide applications in many areas of engineering and sciences including estimation, signal processing, control, data analysis and modeling, statistics and finance. The most basic advantage of convex optimization compared to other optimization problems is that any local minimum must be a global minimum. Hence the problems can be solved efficiently using specialized numerical methods for convex optimization. A very large class of inference, approximation, data analytics and engineering design problems can be formulated as convex optimization.

A function f is convex if it satisfies the inequality

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$.

A convex optimization problem is written as

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad i = 1, \dots, n \end{aligned}$$

where f and g_i are convex functions. Note that we can replace an affine constraint $h(x) = 0$ by a pair of inequality constraints $h(x) \leq 0$ and $h(x) \geq 0$ which are both convex constraints. Important examples are *linear programs* and *quadratic programs* where the objective and constraint functions are affine and quadratic respectively. In chapter 2 we describe how randomization can be used to solve quadratic programs with constraints approximately and faster. We review existing numerical methods and investigate novel fast randomized algorithms for solving general convex problems in Chapter 4.

1.1.2 Empirical risk minimization

In many machine learning, statistical estimation and decision making tasks, we frequently encounter the risk minimization problem

$$\min_{\theta \in \Theta} \mathbb{E}_w[\ell(\theta, w)]$$

where w is a random vector and ℓ is a loss function. The expected objective function is usually referred as the *population risk*. In general, minimizing the expected risk

is often intractable and the empirical risk minimization (ERM) method considers an empirical approximation of the risk using independent and identically distributed (i.i.d.) samples of w_1, \dots, w_n as follows

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(\theta, w_i).$$

In big data applications, the number of samples n can be very large and solving the above problem becomes a significant computational challenge. In the following three chapters of the thesis we will explore and theoretically analyze novel randomized algorithms in order to solve these problems faster than existing methods. In chapters 2, 3 and 4 we will consider instances of ERM including least-squares and logistic regression, support-vector machines and portfolio optimization.

1.1.3 Minimax theory

Minimax theory studies fundamental limits in statistical estimation and hypothesis testing problems. Here we only briefly review the basics of minimax theory which will play an essential role in Chapter 3 for designing better randomized sketching algorithms.

Suppose that we have samples w_1, \dots, w_n i.i.d. from a distribution $p_\theta \in \mathcal{P}$ where θ is a parameter which belongs to a known set Θ . In estimating θ from samples, we define the minimax risk as follows

$$\mathfrak{M}(\mathcal{P}, \Theta) := \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\|\hat{\theta} - \theta\|_2^2 \right],$$

where the infimum is taken over all estimators, i.e., functions of the observed data. The minimax risk can be interpreted in a game-theoretical setting: the statistician chooses an optimal estimator $\hat{\theta}$ based on the data, then the adversary chooses a worst-case parameter θ consistent with the observed data $w \sim p_\theta$.

In Chapter 3 we study the minimax risk in estimation problems when the data is sketched, i.e., randomly projected and we consider all estimators that are functions of the sketched data. Surprisingly, for most of the popular sketching matrices, we show the existence of a gap in terms of statistical estimation performance. Consequently in Chapter 3, we propose efficient iterative algorithms which obtain statistical minimax estimation error.

1.1.4 Random projection

A fundamental component of randomized algorithms considered in this thesis is randomized mechanisms for dimension reduction. Random projection is a mathematical technique to lower the dimensionality of a set of points lying in the Euclidean space. Here we briefly describe this simple but extremely powerful technique. Consider the set of points $\{x_1, \dots, x_N\}$ where each of which is an element of \mathbb{R}^n . We would like to obtain N points y_1, \dots, y_N each of which is in \mathbb{R}^m where $m \ll n$. The following lemma provides a randomized way to obtain such an embedding.

Lemma 1 (The Johnson-Lindenstrauss (J-L) lemma [70, 139]). *Given N points $\{x_i\}_{i=1}^N$, let $S^{m \times n}$ be a matrix such that $S_{kl} \sim \frac{1}{\sqrt{m}}N(0, 1)$ i.i.d. for all k, l . Define the points $y_i = Sx_i$. Then if $m \geq \frac{20 \log(N)}{\epsilon^2}$ for some $\epsilon \in (0, 1/2)$, then with probability at least $1/2$ it holds that*

$$(1 - \epsilon)\|x_i - x_j\|_2^2 \leq \|y_i - y_j\|_2^2 \leq (1 + \epsilon)\|x_i - x_j\|_2^2,$$

for all i and j .

Note that, in order to store the original points we need $O(Nn)$ space. The J-L lemma allows us to store the embedded points which needs only $O(N \log(N))$ space. Instead of using a i.i.d. Gaussian embedding matrix S we can also use an i.i.d. ± 1 matrix [1] which has computational advantages. Computing the embedding takes $O(Nmn)$ time. Recently, faster random projections which employ the Fast Fourier Transform (FFT) have been discovered which can reduce the embedding time to $O(Nn \log(m))$. In the sequel we will describe these fast embeddings which play a significant role in our development of fast optimization algorithms.

1.1.5 Sketching data streams and matrices

A *sketch* is a small data structure that is used to approximate high dimensional data streams or large matrices for approximate computing, querying and updating. Random projections provide a simple construction of linear sketches where we apply the random projection matrix $S \in \mathbb{R}^{m \times n}$ to a data vector $x \in \mathbb{R}^n$ to obtain the sketch Sx . In this context, the matrix S is referred as a *sketching matrix* and the vector x can be representing a data stream at a particular time instant.

One of the first uses of sketching in streaming algorithms have been approximating frequency moments [8]. When the vector $x \in \mathbb{R}^n$ contains number of occurrences of

objects and we would like to update x via $x' = x + \Delta$, we can use the linearity of the sketch $Sx' = Sx + S\Delta$ to update our approximation. Most importantly, we can approximate the second frequency moment $(\sum_{i=1}^n x_i^2)^{1/2} = \|x\|_2$ via the quantity $\|Sx\|_2$ using the J-L lemma without storing the entire data stream.

Sketching can also be used to obtain approximations of large data matrices. Consider $M \in \mathbb{R}^{n \times d}$ and the sketch $SM \in \mathbb{R}^{m \times d}$ where we can interpret it as randomly projecting each column Me_i of the matrix M . When $m \ll n$, the sketched matrix provides computational advantages in linear algebraic operations such as Singular Value Decomposition (SVD) or QR decomposition.

1.1.6 Different kinds of sketches

Given a sketching matrix $S \in \mathbb{R}^{m \times n}$, we use $\{s_i\}_{i=1}^m$ to denote the collection of its n -dimensional rows. We restrict our attention to sketch matrices that are zero-mean, and that are normalized so that $\mathbb{E}[S^T S/m] = I_n$. Various types of randomized sketches of matrices are possible, and we describe a few of them here.

1.1.6.0.1 Sub-Gaussian sketches The most classical sketch is based on a random matrix $S \in \mathbb{R}^{m \times n}$ with i.i.d. standard Gaussian entries, or somewhat more generally, sketch matrices based on i.i.d. sub-Gaussian rows. In particular, a zero-mean random vector $s \in \mathbb{R}^n$ is 1-sub-Gaussian if for any $u \in \mathbb{R}^n$, we have

$$\mathbb{P}[\langle s, u \rangle \geq \epsilon \|u\|_2] \leq e^{-\epsilon^2/2} \quad \text{for all } \epsilon \geq 0. \quad (1.1)$$

For instance, a vector with i.i.d. $N(0, 1)$ entries is 1-sub-Gaussian, as is a vector with i.i.d. Rademacher entries (uniformly distributed over $\{-1, +1\}$). We use the terminology *sub-Gaussian sketch* to mean a random matrix $S \in \mathbb{R}^{m \times n}$ with i.i.d. rows that are zero-mean, 1-sub-Gaussian, and with $\text{cov}(s) = I_n$.

From a theoretical perspective, sub-Gaussian sketches are attractive because of the well-known concentration properties of sub-Gaussian random matrices (e.g., [44, 140]). On the other hand, from a computational perspective, a disadvantage of sub-Gaussian sketches is that they require matrix-vector multiplications with unstructured random matrices. In particular, given a data matrix $A \in \mathbb{R}^{n \times d}$, computing its sketched version SA requires $\mathcal{O}(mnd)$ basic operations in general (using classical matrix multiplication).

1.1.6.0.2 Sketches based on randomized orthonormal systems (ROS) The second type of randomized sketch we consider is *randomized orthonormal system* (ROS), for which matrix multiplication can be performed much more efficiently. In

order to define a ROS sketch, we first let $H \in \mathbb{C}^{n \times n}$ be an orthonormal complex valued matrix with unit magnitude entries, i.e., $|H_{ij}| \in [-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$. Standard classes of such matrices are the Hadamard or Fourier bases, for which matrix-vector multiplication can be performed in $\mathcal{O}(n \log n)$ time via the fast Hadamard or Fourier transforms, respectively. Based on any such matrix, a sketching matrix $S \in \mathbb{C}^{m \times n}$ from a ROS ensemble is obtained by sampling i.i.d. rows of the form

$$s^T = \sqrt{n} e_j^T H D \quad \text{with probability } 1/n \text{ for } j = 1, \dots, n,$$

where the random vector $e_j \in \mathbb{R}^n$ is chosen uniformly at random from the set of all n canonical basis vectors, and $D = \text{diag}(\nu)$ is a diagonal matrix of i.i.d. Rademacher variables $\nu \in \{-1, +1\}^n$. Given a fast routine for matrix-vector multiplication, the sketch SM for a data matrix $M \in \mathbb{R}^{n \times d}$ can be formed in $\mathcal{O}(nd \log m)$ time (for instance, see the papers [5, 4, 55]). The fast matrix multiplication usually requires n to be a power of 2 (or power of r for a radix- r construction). However, in order to use the fast multiplication for an arbitrary n , we can augment the data matrix with a block of zero rows and do the same for the square root of the Hessian without changing the objective value.

1.1.6.0.3 Sketches based on random row sampling Given a probability distribution $\{p_j\}_{j=1}^n$ over $[n] = \{1, \dots, n\}$, another choice of sketch is to randomly sample the rows of a data matrix M a total of m times with replacement from the given probability distribution. Thus, the rows of S are independent and take on the values

$$s^T = \frac{e_j}{\sqrt{p_j}} \quad \text{with probability } p_j \text{ for } j = 1, \dots, n$$

where $e_j \in \mathbb{R}^n$ is the j^{th} canonical basis vector. Different choices of the weights $\{p_j\}_{j=1}^n$ are possible, including those based on the row ℓ_2 norms $p_j \propto \|M e_j\|_2^2$ and leverage values of M —i.e., $p_j \propto \|U e_j\|_2$ for $j = 1, \dots, n$, where $U \in \mathbb{R}^{n \times d}$ is the matrix of left singular vectors of M (e.g., see the paper [52]). When the matrix $M \in \mathbb{R}^{n \times d}$ corresponds to the adjacency matrix of a graph with d vertices and n edges, the leverage scores of M are also known as effective resistances which can be used to sub-sample edges of a given graph by preserving its spectral properties [129].

1.1.6.0.4 Sparse JL Sketches For sparse data matrices, the sketching operation can be done faster if the sketching matrix is chosen from a distribution over sparse matrices. Several works developed sparse JL embeddings [1, 42, 74] and sparse subspace embeddings [103]. Here we describe a construction given by [103, 74]. Given an integer s , each column of S is chosen to have exactly s non-zero entries in random locations, each equal to $\pm 1/\sqrt{s}$ uniformly at random. The column sparsity parameter s can be chosen $O(1/\epsilon)$ for subspace embeddings and $O(\log(1/\delta)/\epsilon)$ for sparse JL embeddings where δ is the failure probability.

1.2 Goals and contributions of this thesis

We can list the high level goals of this thesis as follows:

1. Developing random projection methods for convex optimization problems and characterizing fundamental trade-offs between the size of the projection and accuracy of solutions.
2. Analyzing information-theoretic limitations of random projection algorithms in statistics and optimization.
3. Designing computationally and statistically efficient statistical estimation algorithms when the sample size is very large. More precisely, the algorithm should run in linear time in the input data size and achieve statistical minimax optimality.
4. Developing new randomized methodologies for relaxing cardinality constraints in order to obtain better approximations than the state of the art approaches (e.g., ℓ_1 heuristic).

More specifically we can list the central contributions of this thesis as follows:

- Novel randomized algorithms for convex optimization: We develop a novel framework for general convex optimization problems which yields provably faster algorithms than currently available methods for large sample sizes. Specifically, the derived algorithms run in exactly linear time in the input data size. The algorithms significantly outperform existing methods on real-world large scale problems such as least-squares, logistic regression and linear, quadratic and semidefinite programming.
- Information-theoretical sub-optimality of traditional random projection methods: Using an information theoretical argument which is analogous to communication systems, we showed that these methods are sub-optimal in terms of a natural statistical error measure. Moreover, a novel alternative method is proposed which is proven to be statistically optimal and at the same time enjoys the same fast computation
- Novel convex relaxations with checkable optimality: We present a new framework which has several advantages over the well-known convex relaxations. In particular, the proposed approach produces bounds and checkable optimality without any assumptions on the data in contrast to known methods, such as ℓ_1 relaxation. Moreover, in many fundamental problems, such as estimation of a probability distribution, ℓ_1 relaxations are inapplicable while our methods were proven to be very effective in a variety of applications including data clustering.

- Privacy and accuracy trade-offs of random projections: We characterize a theoretical trade-off between the information theoretic amount of revealed data to an optimization service and the quality of optimization. Our theoretical results state that, privacy preserving optimization using a randomization method is possible depending on the geometric properties of the optimization constraint set. Interestingly, in many cases of interest, we need not know about the data to be able to optimize over it.

1.2.1 Thesis organization and previously published work

Several portions of this thesis are based on the previously published joint work with several collaborators. Chapter 2, 3 and 4 are based on joint work with Martin Wainwright [114, 115, 113]. Chapter 5 is based on a joint work with Yun Yang [151]. Chapter 6 is based on joint work with Laurent El Ghaoui [116] and Venkat Chandrasekaran [111].

1.2.2 Notation

For sequences $\{a_t\}_{t=0}^{\infty}$ and $\{b_t\}_{t=0}^{\infty}$, we use the notation $a_t \preceq b_t$ to mean that there is a constant (independent of t) such that $a_t \leq C b_t$ for all t . Equivalently, we write $b_t \succeq a_t$. We write $a_t \asymp b_t$ if $a_t \preceq b_t$ and $b_t \preceq a_t$. We use ℓ_p to denote the usual p -norms $\|x\|_p := (\sum_i x_i^p)^{1/p}$ and $\|x\|_{\infty} = \max_i |x_i|$. We use $e_i \in \mathbb{R}^n$, to denote the i 'th ordinary basis vector in \mathbb{R}^n . We use x_i to denote the i 'th index of a vector x and M_{ij} to denote the (i, j) 'th element of a matrix M . We use $\lambda_{min}(M)$ and $\lambda_{max}(M)$ to denote the minimum and maximum eigenvalue of a matrix $M \in \mathbb{R}^{n_1 \times n_2}$ respectively. For an integer i , $1 \leq i \leq rank(M)$, $\sigma_i(M)$ is the i 'th largest singular value of a matrix M . The Frobenius norm is defined by $\|M\|_F := \sqrt{\sum_i \sigma_i^2(M)}$ for a matrix. The ℓ_2 operator norm of a matrix M is defined by

$$\|M\|_2 := \max_{\|x\|_2 \leq 1} \|Mx\|_2 = \sigma_1.$$

The nuclear norm of a matrix is defined by $\|M\|_* := \sum_i \sigma_i(M)$. \mathbb{E} denotes the expectation of a random variable. The notation $(\cdot)_+$ denotes the positive part of a real scalar.

Chapter 2

Random projections of convex quadratic programs

Optimizing a convex function subject to convex constraints is fundamental to many disciplines in engineering, applied mathematics, and statistics [28, 104]. While most convex programs can be solved in polynomial time, the computational cost can still be prohibitive when the problem dimension and/or number of constraints are large. For instance, although many quadratic programs can be solved in cubic time, this scaling may be prohibitive when the dimension is on the order of millions. This type of concern is only exacerbated for more sophisticated cone programs, such as second-order cone and semidefinite programs. Consequently, it is of great interest to develop methods for approximately solving such programs, along with rigorous bounds on the quality of the resulting approximation.

In this section, we analyze a particular scheme for approximating a convex program defined by minimizing a convex quadratic objective function over an arbitrary convex set. The scheme is simple to describe and implement, as it is based on performing a random projection of the matrices and vectors defining the objective function. Since the underlying constraint set may be arbitrary, our analysis encompasses many problem classes including quadratic programs (with constrained or penalized least-squares as a particular case), as well as second-order cone programs and semidefinite programs (including low-rank matrix approximation as a particular case).

An interesting class of such optimization problems arise in the context of statistical estimation. Many such problems can be formulated as estimating an unknown parameter based on noisy linear measurements, along with the side information that the

true parameter belongs to a low-dimensional space. Examples of such low-dimensional structures include sparse vectors, low-rank matrices, discrete sets defined in a combinatorial manner, as well as algebraic sets, including norms for inducing shrinkage or smoothness. Convex relaxations provide a principled way of deriving polynomial-time methods for such problems [28], and their statistical performance has been extensively studied over the past decade (see the sources [30, 35, 144] for overviews). For many such problems, the ambient dimension of the parameter is very large, and the number of samples can also be large. In these contexts, convex programs may be difficult to solve exactly, and reducing the dimension and sample size by sketching is a very attractive option.

Our work is related to a line of work on sketching unconstrained least-squares problems (e.g., see the papers [123, 55, 90, 27] and references therein). The results given here generalize this line of work by providing guarantees for a broader class of constrained quadratic programs. In addition, our techniques are convex-analytic in nature, and by exploiting analytical tools from Banach space geometry and empirical process theory [45, 85, 84], lead to sharper bounds on the sketch size as well as sharper probabilistic guarantees. Our work also provides a unified view of both least-squares sketching [55, 90, 27] and compressed sensing [49, 51]. As we discuss in the sequel, various results in compressed sensing can be understood as special cases of sketched least-squares, in which the data matrix in the original quadratic program is the identity.

In addition to reducing computation and storage, random projection is also useful in the context of privacy preservation. Many types of modern data, including financial records and medical tests, have associated privacy concerns. Random projection allows for a sketched version of the data set to be stored, but such that there is a vanishingly small amount of information about any given data point. Our theory shows that this is still possible, while still solving a convex program defined by the data set up to δ -accuracy. In this way, we sharpen some results by Zhou and Wasserman [158] on privacy-preserving random projections for sparse regression. Our theory points to an interesting dichotomy in privacy-preserving optimization problems based on the trade-off between the complexity of the constraint set and mutual information between data and its sketch. We show that if the constraint set is *simple* enough in terms of a statistical measure, privacy-preserving optimization can be done with arbitrary accuracy.

2.1 Problem formulation

We begin by formulating the problem analyzed in this section, before turning to a statement of our main results.

Consider a convex program of the form

$$x^* \in \arg \min_{x \in \mathcal{C}} \underbrace{\|Ax - y\|_2^2}_{f(x)}, \quad (2.1)$$

where \mathcal{C} is some convex subset of \mathbb{R}^d , and $y \in \mathbb{R}^n$ $A \in \mathbb{R}^{n \times d}$ are a data vector and data matrix, respectively. Our goal is to obtain an δ -optimal solution to this problem in a computationally simpler manner, and we do so by projecting the problem into the lower dimensional space \mathbb{R}^m for $m < n$. In particular, given a *sketching matrix* $S \in \mathbb{R}^{m \times n}$. consider the *sketched problem*

$$\hat{x} \in \arg \min_{x \in \mathcal{C}} \underbrace{\|S(Ax - y)\|_2^2}_{g(x)}. \quad (2.2)$$

Note that by the optimality and feasibility of x^* and \hat{x} , respectively, for the original problem (2.1), we always have $f(x^*) \leq f(\hat{x})$. Accordingly, we say that \hat{x} is an δ -*optimal approximation* to the original problem (2.1) if

$$f(\hat{x}) \leq (1 + \delta)^2 f(x^*). \quad (2.3)$$

Our main result characterizes the number of projections m required to achieve this bound as a function of δ , and other problem parameters.

Our analysis involves a natural geometric object in convex analysis, namely the tangent cone of the constraint set \mathcal{C} at the optimum x^* , given by

$$\mathcal{K} := \text{clconv} \{ \Delta \in \mathbb{R}^d \mid \Delta = t(x - x^*) \text{ for some } t \geq 0 \text{ and } x \in \mathcal{C} \}, \quad (2.4)$$

where clconv denotes the closed convex hull. This set arises naturally in the convex optimality conditions for the original problem (2.1): any vector $\Delta \in \mathcal{K}$ defines a feasible direction at the optimal x^* , and optimality means that it is impossible to decrease the cost function by moving in directions belonging to the tangent cone. Figure 2.1 depicts an example of a tangent cone.

We use $A\mathcal{K}$ to denote the linearly transformed cone $\{A\Delta \in \mathbb{R}^n \mid \Delta \in \mathcal{K}\}$. Our main results involve measures of the “size” of this transformed cone when it is intersected with the Euclidean sphere $\mathcal{S}^{n-1} = \{z \in \mathbb{R}^n \mid \|z\|_2 = 1\}$. In particular, we define *Gaussian width* of the set $A\mathcal{K} \cap \mathcal{S}^{n-1}$ via

$$\mathbb{W}(A\mathcal{K}) := \mathbb{E}_g \left[\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} |\langle g, z \rangle| \right] \quad (2.5)$$

where $g \in \mathbb{R}^n$ is an i.i.d. sequence of $N(0, 1)$ variables. This complexity measure plays an important role in Banach space theory, learning theory and statistics (e.g., [117, 78, 85, 19]). As an example of a transformed tangent cone with small width, consider a low-rank matrix A where $r := \text{rank}(A) \ll d$, then the supremum in equation (2.5) is taken in an r -dimensional subspace. In this case, it can be shown that $\mathbb{W}(A\mathcal{K}) \leq \sqrt{r}$ —see Corollary 2 for details.

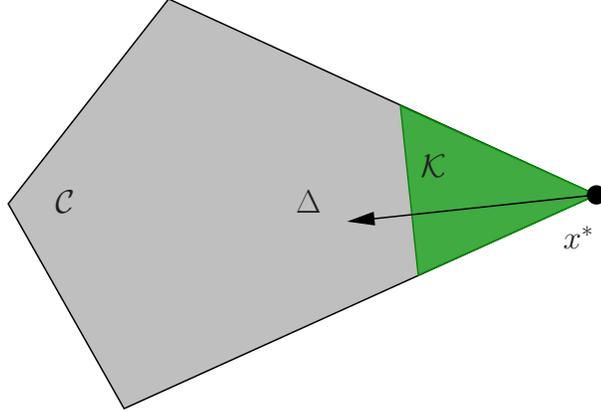


Figure 2.1: Tangent cone at x^*

2.1.1 Guarantees for sub-Gaussian sketches

Our first main result provides a relation between the sufficient sketch size and Gaussian complexity in the case of sub-Gaussian sketches.

Theorem 1 (Guarantees for sub-Gaussian projections). *Let $S \in \mathbb{R}^{m \times n}$ be drawn from a σ -sub-Gaussian ensemble. Then there are universal constants (c_0, c_1, c_2) such that, for any tolerance parameter $\delta \in (0, 1)$, given a sketch size lower bounded as*

$$m \geq \frac{c_0}{\delta^2} \mathbb{W}^2(A\mathcal{K}), \quad (2.6)$$

the approximate solution \hat{x} is guaranteed to be δ -optimal (2.3) for the original program with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.

As will be clarified in examples to follow, the squared width $\mathbb{W}^2(A\mathcal{K})$ scales proportionally to the effective dimension, or number of degrees of freedom in the set $A\mathcal{K} \cap \mathcal{S}^{n-1}$. Consequently, up to constant factors, Theorem 1 guarantees that we can project down to the effective dimension of the problem while preserving δ -optimality of the solution. Moreover, as we show in section 2.2-C, the sketch size lower-bound in Theorem 1 can not be improved substantially for arbitrary A and \mathcal{C} due to connections with Compressed Sensing and denoising.

This fact has an interesting corollary in the context of privacy-preserving optimization. Suppose that we model the data matrix $A \in \mathbb{R}^{n \times d}$ as being random, and

our goal is to solve the original convex program (2.1) up to δ -accuracy while revealing as little as possible about the individual entries of A . By Theorem 1, whenever the sketch dimension satisfies the lower bound (2.6), the sketched data matrix $SA \in \mathbb{R}^{m \times d}$ suffices to solve the original program up to δ -accuracy. We can thus ask about how much information per entry of A is retained by the sketched data matrix. One way in which to do so is by computing the mutual information per symbol, namely

$$\frac{I(SA; A)}{nd} = \frac{1}{nd} D(\mathbb{P}_{SA, A} \parallel \mathbb{P}_{SA} \mathbb{P}_A),$$

corresponding to the (renormalized) Kullback-Leibler divergence between the joint distribution over (SA, A) and the product of the marginals. Here we have chosen the renormalization (nd) since the matrix has dimensions $n \times d$. This question was studied by Zhou and Wasserman [158] in the context of privacy-preserving sparse regression, in which \mathcal{C} is an ℓ_1 -ball, to be discussed more at length in Section 2.2.2. In our setting, we have the following more generic corollary of Theorem 1:

Corollary 1. *Let the entries of A be drawn i.i.d. from a distribution with finite variance γ^2 . By using $m = \frac{c_0}{\delta^2} \mathbb{W}^2(A\mathcal{K})$ random Gaussian projections, we can ensure that*

$$\frac{I(SA; A)}{nd} \leq \frac{c_0}{\delta^2} \frac{\mathbb{W}^2(A\mathcal{K})}{n} \log(2\pi e \gamma^2), \quad (2.7)$$

and that the sketched solution is δ -optimal with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.

Note that the inequality $\mathbb{W}^2(A\mathcal{K}) \leq n$ always holds. However, for many problems, we have the much stronger guarantee $\mathbb{W}^2(A\mathcal{K}) = o(n)$, in which case the bound (2.7) guarantees that the mutual information per symbol is vanishing. There are various concrete problems, as discussed in Section 2.2, for which this type of scaling is reasonable. Thus, for any fixed $\delta \in (0, 1)$, we are guaranteed a δ -optimal solution with a vanishing mutual information per symbol.¹

Corollary 1 follows by a straightforward combination of past work with Theorem 1. In particular, Zhou and Wasserman [158] show that under the stated conditions, for a standard i.i.d. Gaussian sketching matrix S , the mutual information rate per symbol is upper bounded as

$$\frac{I(SA; A)}{nd} \leq \frac{m}{2n} \log(2\pi e \gamma^2).$$

Substituting in the stated choice of m and applying Theorem 1 yields the claim.

¹While this is a reasonable guarantee, we note that there are stronger measures of privacy than vanishing mutual information (e.g., differential privacy [56]).

2.1.2 Guarantees for randomized orthogonal systems

Our main result for randomized orthonormal systems involves the *S-Gaussian width* of the set $A\mathcal{K} \cap \mathcal{S}^{n-1}$, given by

$$\mathbb{W}_S(A\mathcal{K}) := \mathbb{E}_{g,S} \left[\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \left| \left\langle g, \frac{Sz}{\sqrt{m}} \right\rangle \right| \right]. \quad (2.8)$$

As will be clear in the corollaries to follow, in many cases, the *S-Gaussian width* is equivalent to the ordinary Gaussian width (2.5) up to numerical constants. It also involves the *Rademacher width* of the set $A\mathcal{K} \cap \mathcal{S}^{n-1}$, given by

$$\mathbb{R}(A\mathcal{K}) = \mathbb{E}_\varepsilon \left[\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} |\langle z, \varepsilon \rangle| \right], \quad (2.9)$$

where $\varepsilon \in \{-1, +1\}^n$ is an i.i.d. vector of Rademacher variables.

Theorem 2 (Guarantees for randomized orthonormal system). *Let $S \in \mathbb{R}^{m \times n}$ be drawn from a randomized orthonormal system (ROS). Then given a sample size m lower bounded as*

$$\frac{m}{\log m} > \frac{c_0}{\delta^2} (\mathbb{R}^2(A\mathcal{K}) + \log n) \mathbb{W}_S^2(A\mathcal{K}), \quad (2.10)$$

the approximate solution \hat{x} is guaranteed to be δ -optimal (2.3) for the original program with probability at least $1 - \frac{c_1}{(mn)^2} - c_1 \exp\left(-c_2 \frac{m\delta^2}{\mathbb{R}^2(A\mathcal{K}) + \log(mn)}\right)$.

The required projection dimension (2.10) for ROS sketches is in general larger than that required for sub-Gaussian sketches, due to the presence of the additional pre-factor $\mathbb{R}^2(A\mathcal{K}) + \log n$. For certain types of cones, we can use more specialized techniques to remove this pre-factor, so that it is not always required. The details of these arguments are given in Section 2.4, and we provide some illustrative examples of such sharpened results in the corollaries to follow. However, the potentially larger projection dimension is offset by the much lower computational complexity of forming matrix vector products using the ROS sketching matrix.

2.2 Applications

Our two main theorems are general results that apply to any choice of the convex constraint set \mathcal{C} . We now turn to some consequences of Theorems 1 and 2 for more specific classes of problems, in which the geometry enters in different ways.

2.2.1 Unconstrained least squares

We begin with the simplest possible choice, namely $\mathcal{C} = \mathbb{R}^d$, which leads to an unconstrained least squares problem. This class of problems has been studied extensively in past work on least-square sketching [90]; our derivation here provides a sharper result in a more direct manner. At least intuitively, given the data matrix $A \in \mathbb{R}^{n \times d}$, it should be possible to reduce the dimensionality to the rank of the data matrix A , while preserving the accuracy of the solution. In many cases, the quantity $\text{rank}(A)$ is substantially smaller than $\min\{n, d\}$. The following corollaries of Theorem 1 and 2 confirm this intuition:

Corollary 2 (Approximation guarantee for unconstrained least squares). *Consider the case of unconstrained least squares with $\mathcal{C} = \mathbb{R}^d$:*

- (a) *Given a sub-Gaussian sketch with dimension $m > c_0 \frac{\text{rank}(A)}{\delta^2}$, the sketched solution is δ -optimal (2.3) with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.*
- (b) *Given an ROS sketch with dimension $m > c'_0 \frac{\text{rank}(A)}{\delta^2} \log^4(n)$, the sketched solution is δ -optimal (2.3) with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.*

This corollary improves known results both in the probability estimate and required samples, in particular previous results hold only with constant probability; see the paper [90] for an overview of such results. Note that the total computational complexity of computing SA and solving the sketched least squares problem, for instance via QR decomposition [62], is of the order $\mathcal{O}(ndm + md^2)$ for sub-Gaussian sketches, and of the order $\mathcal{O}(nd \log(m) + md^2)$ for ROS sketches. Consequently, by using ROS sketches, the overall complexity of computing a δ -approximate least squares solution with exponentially high probability is $\mathcal{O}(\text{rank}(A)d^2 \log^4(n)/\delta^2 + nd \log(\text{rank}(A)/\delta^2))$. In many cases, this complexity is substantially lower than direct computation of the solution via QR decomposition, which would require $\mathcal{O}(nd^2)$ operations. We also note that the $\text{rank}(A)$ may not be known in advance. However in many applications such as polynomial and kernel regression, the matrix is approximately low rank. In such cases, standard bounds from matrix perturbation theory [132] can be applied to obtain an approximation bound via the decomposition $A = A_r + E$, where $\text{rank}(A_r) = r$ and $\|E\|_2$ is small.

Proof. Since $\mathcal{C} = \mathbb{R}^d$, the tangent cone \mathcal{K} is all of \mathbb{R}^d , and the set $A\mathcal{K}$ is the image of

A. Thus, we have

$$\mathbb{W}(A\mathcal{K}) = \mathbb{E}\left[\sup_{u \in \mathbb{R}^d} \frac{|\langle Au, g \rangle|}{\|Au\|_2}\right] \leq \sqrt{\text{rank}(A)}, \quad (2.11)$$

where the inequality follows from the fact that the image of A is at most $\text{rank}(A)$ -dimensional. Thus, the sub-Gaussian bound in part (a) is an immediate consequence of Theorem 1.

Turning to part (b), an application of Theorem 2 will lead to a sub-optimal result involving $(\text{rank}(A))^2$. In Section 2.4.1, we show how a refined argument will lead to bound stated here. \square

In order to investigate the theoretical prediction of Corollary 2, we performed some simple simulations on randomly generated problem instances. Fixing a dimension $d = 500$, we formed a random ensemble of least-squares problems by first generating a random data matrix $A \in \mathbb{R}^{n \times 500}$ with i.i.d. standard Gaussian entries. For a fixed random vector $x_0 \in \mathbb{R}^d$, we then computed the data vector $y = Ax_0 + w$, where the noise vector $w \sim N(0, \nu^2)$ where $\nu = \sqrt{0.2}$. Given this random ensemble of problems, we computed the projected data matrix-vector pairs (SA, Sy) using Gaussian, Rademacher, and randomized Hadamard sketching matrices, and then solved the projected convex program. We performed this experiment for a range of different problem sizes $n \in \{1024, 2048, 4096\}$. For any n in this set, we have $\text{rank}(A) = d = 500$, with high probability over the choice of randomly sampled A . Suppose that we choose a projection dimension of the form $m = \max\{1.5\alpha d, 1\}$, where the control parameter α ranges over the interval $[0, 1]$. Corollary 2 predicts that the approximation error should converge to 1 under this scaling, for each choice of n .

Figure 2.2 shows the results of these experiments, plotting the approximation ratio $f(\hat{x})/f(x^*)$ versus the control parameter α . Consistent with Corollary 2, regardless of the choice of n , once the projection dimension is a suitably large multiple of $\text{rank}(A) = 500$, the approximation quality becomes very good.

2.2.2 ℓ_1 -constrained least squares

We now turn to a constrained form of least-squares, in which the geometry of the tangent cone enters in a more interesting way. In particular, consider the ℓ_1 -

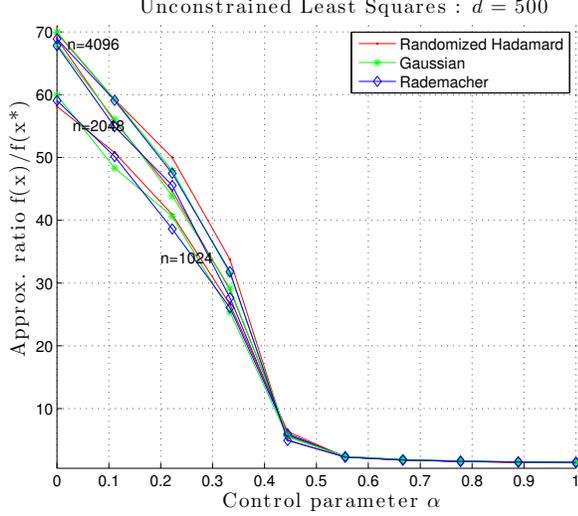


Figure 2.2: Comparison of Gaussian, Rademacher and randomized Hadamard sketches for unconstrained least squares. Each curve plots the approximation ratio $f(\hat{x})/f(x^*)$ versus the control parameter α , averaged over $T_{trial} = 100$ trials, for projection dimensions $m = \max\{1.5\alpha d, 1\}$ and for problem dimensions $d = 500$ and $n \in \{1024, 2048, 4096\}$.

constrained least squares program, known as the Lasso [36, 134], given by

$$x^* \in \arg \min_{\|x\|_1 \leq R} \|Ax - y\|_2^2. \quad (2.12)$$

It is widely used in signal processing and statistics for sparse signal recovery and approximation.

In this section, we show that as a corollary of Theorem 1, this quadratic program can be sketched logarithmically in dimension d when the optimal solution to the original problem is sparse. In particular, assuming that x^* is unique, we let k denote the number of non-zero coefficients of the unique solution to the above program. (When x^* is not unique, we let k denote the minimal cardinality among all optimal vectors). Define the ℓ_1 -restricted eigenvalues of the given data matrix A as

$$\gamma_k^-(A) := \min_{\substack{\|z\|_2=1 \\ \|z\|_1 \leq 2\sqrt{k}}} \|Az\|_2^2, \quad \text{and} \quad (2.13)$$

$$\gamma_k^+(A) := \max_{\substack{\|z\|_2=1 \\ \|z\|_1 \leq 2\sqrt{k}}} \|Az\|_2^2. \quad (2.14)$$

We note that our choice of introducing the factor of two in the constraint $\|z\|_1 \leq 2\sqrt{k}$ is for later theoretical convenience, due to the structure of the tangent cone associated with the ℓ_1 -norm. By rescaling as necessary, we may assume $\gamma_k^-(A) \leq 1$ without loss of generality.

Corollary 3 (Approximation guarantees for ℓ_1 -constrained least squares). *Consider the ℓ_1 -constrained least squares problem (2.12):*

(a) *For sub-Gaussian sketches, a sketch dimension lower bounded by*

$$m \geq \frac{c_0}{\delta^2} \min\left\{ \text{rank}(A), \max_{j \in [1:d]} \frac{\|a_j\|_2^2}{\gamma_k^-(A)} k \log(d) \right\} \quad (2.15)$$

guarantees that the sketched solution is δ -optimal (2.3) with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.

(b) *For ROS sketches, a sketch dimension lower bounded by*

$$m > \frac{c'_0}{\delta^2} \log^4(n) \min\left\{ \text{rank}(A) \frac{\left(\frac{\max_j \|a_j\|_2^2}{\gamma_k^-(A)} k \log(d)\right)^2}{\log^4(n)}, \left(\frac{\gamma_k^+(A) + 1}{\gamma_k^-(A)}\right)^2 k \log(d) \right\} \quad (2.16)$$

guarantees that the sketched solution is δ -optimal (2.3) with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.

We note that part (a) of this corollary improves the result of Zhou et al. [158], which establishes consistency of Lasso with a Gaussian sketch dimension of the order $k^2 \log(dnk)$, in contrast to the $k \log(d)$ requirement in the bound (2.15). To be more precise, these two results are slightly different, in that the result [158] focuses on support recovery, whereas Corollary 3 guarantees a δ -accurate approximation of the cost function.

Let us consider the complexity of solving the sketched problem using different methods. In the regime $n > d$, the complexity of solving the original Lasso problem as a linearly constrained quadratic program via interior point solvers is $\mathcal{O}(nd^2)$ per iteration (e.g., see Nesterov and Nemirovski [107]). Thus, computing the sketched data and solving the sketched Lasso problem requires $\mathcal{O}(ndm + md^2)$ operations for sub-Gaussian sketches, and $\mathcal{O}(nd \log(m) + md^2)$ for ROS sketches.

Another popular choice for solving the Lasso problem is to use a first-order algorithm [106]; such algorithms require $\mathcal{O}(nd)$ operations per iteration, and yield a solution that is $\mathcal{O}(1/T)$ -optimal within T iterations. If we apply such an algorithm to the sketched version for T steps, then we obtain a vector such that

$$f(\hat{x}) \leq (1 + \delta)^2 f(x^*) + \mathcal{O}\left(\frac{1}{T}\right).$$

Overall, obtaining this guarantee requires $\mathcal{O}(ndm + mdT)$ operations for sub-Gaussian sketches, and $\mathcal{O}(nd \log(m) + mdT)$ operations for ROS sketches.

Proof. Let S denote the support of the optimal solution x^* . The tangent cone to the ℓ_1 -norm constraint at the optimum x^* takes the form

$$\mathcal{K} = \{\Delta \in \mathbb{R}^d \mid \langle \Delta_S, \hat{z}_S \rangle + \|\Delta_{S^c}\|_1 \leq 0\}, \quad (2.17)$$

where Δ_S and Δ_{S^c} denote the restriction of the vector Δ to subsets S and S^c respectively and $\hat{z}_S := \text{sign}(x_S^*) \in \{-1, +1\}^k$ is the sign vector of the optimal solution on its support S . By the triangle inequality, any vector $\Delta \in \mathcal{K}$ satisfies the inequality

$$\|\Delta\|_1 \leq 2\|\Delta_S\|_1 \leq 2\sqrt{k}\|\Delta_S\|_2 \leq 2\sqrt{k}\|\Delta\|_2. \quad (2.18)$$

If $\|A\Delta\|_2 = 1$, then by the definition (2.13), we also have the upper bound $\|\Delta\|_2 \leq \frac{1}{\sqrt{\gamma_k^-(A)}}$, whence

$$\langle A\Delta, g \rangle \leq 2\sqrt{|S|} \|\Delta\|_2 \|A^T g\|_\infty \leq \frac{2\sqrt{|S|} \|A^T g\|_\infty}{\sqrt{\gamma_k^-(A)}}. \quad (2.19)$$

Note that $A^T g$ is a d -dimensional Gaussian vector, in which the j^{th} -entry has variance $\|a_j\|_2^2$. Consequently, inequality (2.19) combined with standard Gaussian tail bounds [85] imply that

$$\mathbb{W}(A\mathcal{K}) \leq 6\sqrt{k \log(d)} \max_{j=1, \dots, d} \frac{\|a_j\|_2}{\sqrt{\gamma_k^-(A)}}. \quad (2.20)$$

Combined with the bound from Corollary 2, also applicable in this setting, the claim (2.15) follows.

Turning to part (b), the first lower bound involving $\text{rank}(A)$ follows from Corollary 2. The second lower bound follows as a corollary of Theorem 2 in application to the Lasso; see Section 2.6.1 for the calculations. The third lower bound follows by a specialized argument given in Section 2.4.3.

□

In order to investigate the prediction of Corollary 3, we generated a random ensemble of sparse linear regression problems as follows. We first generated a data matrix $A \in \mathbb{R}^{4096 \times 500}$ by sampling i.i.d. standard Gaussian entries, and then a k' -sparse base vector $x_0 \in \mathbb{R}^d$ by choosing a uniformly random subset S of size $k' = d/10$, and setting its entries to in $\{-1, +1\}$ independent and equiprobably. Finally, we formed the data vector $y = Ax_0 + w$, where the noise vector $w \in \mathbb{R}^n$ has i.i.d. $N(0, \nu^2)$ entries with $\nu = \sqrt{0.2}$.

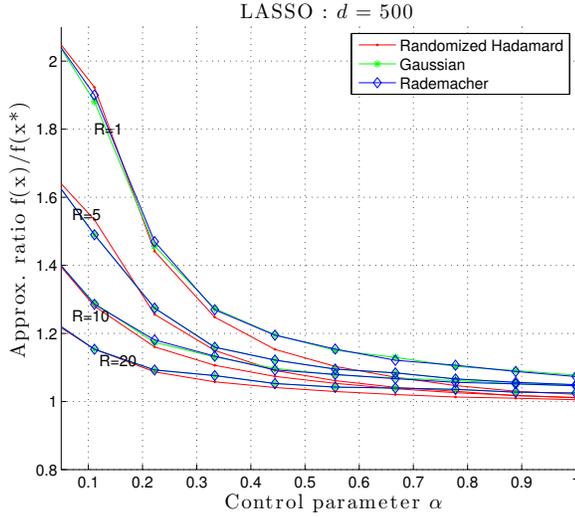


Figure 2.3: Comparison of Gaussian, Rademacher and randomized Hadamard sketches for the Lasso program (2.12). Each curve plots the approximation ratio $f(\hat{x})/f(x^*)$ versus the control parameter α , averaged over $T_{\text{trial}} = 100$ trials, for projection dimensions $m = \max\{4\alpha\|x^*\|_0 \log d, 1\}$, problem dimensions $(n, d) = (4096, 500)$, and ℓ_1 -constraint radius $R \in \{1, 5, 10, 20\}$.

In our experiments, we solved the Lasso (2.12) with a choice of radius parameter $R \in \{1, 5, 10, 20\}$, and set $k = \|x^*\|_0$. We then set the projection dimension

$m = \max\{4\alpha k \log d, 1\}$ where $\alpha \in (0, 1)$ is a control parameter, and solved the sketched Lasso for Gaussian, Rademacher and randomized Hadamard sketching matrices. Our theory predicts that the approximation ratio tends to one as the control parameter α increases. The results are plotted in Figure 2.3, and confirm this qualitative prediction.

2.2.3 Compressed sensing and noise folding

It is worth noting that various compressed sensing results can be recovered as a special case of Corollary 3—more precisely, one in which the “data matrix” A is simply the identity (so that $n = d$). With this choice, the original problem (2.1) corresponds to the classical denoising problem, namely

$$x^* = \arg \min_{x \in \mathcal{C}} \|x - y\|_2^2, \quad (2.21)$$

so that the cost function is simply $f(x) = \|x - y\|_2^2$. With the choice of constraint set $\mathcal{C} = \{\|x\|_1 \leq R\}$, the optimal solution x^* to the original problem is unique, and can be obtained by performing a coordinate-wise soft-thresholding operation on the data vector y . For this choice, the sketched version of the de-noising problem (2.21) is given by

$$\hat{x} = \arg \min_{x \in \mathcal{C}} \|Sx - Sy\|_2^2 \quad (2.22)$$

2.2.3.0.5 Noiseless version: In the noiseless version of compressed sensing, we have $y = \bar{x} \in \mathcal{C}$, and hence the optimal solution to the original “denoising” problem (2.21) is given by $x^* = \bar{x}$, with optimal value

$$f(x^*) = \|x^* - \bar{x}\|_2^2 = 0.$$

Using the sketched data vector $S\bar{x} \in \mathbb{R}^m$, we can solve the sketched program (2.22). If doing so yields a δ -approximation \hat{x} , then in this special case, we are guaranteed that

$$\|\hat{x} - \bar{x}\|_2^2 = f(\hat{x}) \leq (1 + \delta)^2 f(x^*) = 0, \quad (2.23)$$

which implies that we have exact recovery—that is, $\hat{x} = \bar{x}$.

2.2.3.0.6 Noisy versions: In a more general setting, we observe the vector $y = \bar{x} + w$, where $\bar{x} \in \mathcal{C}$ and $w \in \mathbb{R}^n$ is some type of observation noise. The sketched observation model then takes the form

$$Sy = S\bar{x} + Sw,$$

so that the sketching matrix is applied to both the true vector \bar{x} and the noise vector w . This set-up corresponds to an instance of compressed sensing with “folded” noise (e.g., see the papers [12, 2]), which some argue is a more realistic set-up for compressed sensing. In this context, our results imply that the sketched version satisfies the bound

$$\|\hat{x} - y\|_2^2 \leq (1 + \delta)^2 \|x^* - y\|_2^2. \quad (2.24)$$

If we think of y as an approximately sparse vector and x^* as the best approximation to y from the ℓ_1 -ball, then this bound (2.24) guarantees that we recover a δ -approximation to the best sparse approximation. Moreover, this bound shows that the compressed sensing error should be closely related to the error in denoising, as has been made precise in recent work [51]. Moreover, this connection and information theoretic lower-bounds for Compressed Sensing (see e.g., [2]) also imply that our approximation results in Theorems 1 and 2 can not be improved substantially.

Let us summarize these conclusions in a corollary:

Corollary 4. *Consider an instance of the denoising problem (2.21) when $\mathcal{C} = \{x \in \mathbb{R}^n \mid \|x\|_1 \leq R\}$.*

- (a) *For sub-Gaussian sketches with projection dimension $m \geq \frac{c_0}{\delta^2} \|x^*\|_0 \log d$, we are guaranteed exact recovery in the noiseless case (2.23), and δ -approximate recovery (2.24) in the noisy case, both with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.*
- (b) *For ROS sketches, the same conclusions hold with probability $1 - e^{-c_1 \frac{m \delta^2}{\log^4 n}}$ using a sketch dimension*

$$m \geq \frac{c_0}{\delta^2} \min \{ \|x^*\|_0 \log^5 d, \|x^*\|_0^2 \log d \}. \quad (2.25)$$

Of course, a more general version of this corollary holds for any convex constraint set \mathcal{C} , involving the Gaussian/Rademacher width functions. In this more setting, the corollary generalizes results by Chandrasekaran et al. [35], who studied randomized Gaussian sketches in application to atomic norms, to other types of sketching matrices and other types of constraints. They provide a number of calculations of widths for various atomic norm constraint sets, including permutation and orthogonal matrices, and cut polytopes, which can be used in conjunction with the more general form of Corollary 4.

2.2.4 Support vector machine classification

Our theory also has applications to learning linear classifiers based on labeled samples. In the context of binary classification, a labeled sample is a pair (a_i, z_i) , where the vector $a_i \in \mathbb{R}^n$ represents a collection of features, and $z_i \in \{-1, +1\}$ is the associated class label. A linear classifier is specified by a function $a \mapsto \text{sign}(\langle w, a \rangle) \in \{-1, +1\}$, where $w \in \mathbb{R}^n$ is a weight vector to be estimated.

Given a set of labelled patterns $\{a_i, z_i\}_{i=1}^d$, the support vector machine [40, 131] estimates the weight vector w^* by minimizing the function

$$w^* = \arg \min_{w \in \mathbb{R}^n} \left\{ \frac{1}{2C} \sum_{i=1}^d g(z_i, \langle w, a_i \rangle) + \frac{1}{2} \|w\|_2^2 \right\}. \quad (2.26)$$

In this formulation, the squared hinge loss $g(w) := (1 - y_i \langle w, a_i \rangle)_+^2$ is used to measure the performance of the classifier on sample i , and the quadratic penalty $\|w\|_2^2$ serves as a form of regularization.

By considering the dual of this problem, we arrive at a least-squares problem that is amenable to our sketching techniques. Let $A \in \mathbb{R}^{n \times d}$ be a matrix with $a_i \in \mathbb{R}^n$ as its i^{th} column, let $D = \text{diag}(z) \in \mathbb{R}^{d \times d}$ be a diagonal matrix, and define $B^T = [(AD)^T \frac{1}{C} I]$. With this notation, the associated dual problem (e.g. see the paper [86]) takes the form

$$x^* := \arg \min_{x \in \mathbb{R}^d} \|Bx\|_2^2 \text{ s.t. } x \geq 0 \text{ and } \sum_{i=1}^d x_i = 1. \quad (2.27)$$

The optimal solution $x^* \in \mathbb{R}^d$ corresponds to a vector of weights associated with the samples: it specifies the optimal SVM weight vector via $w^* = \sum_{i=1}^d x_i^* z_i a_i$. It is often the case that the dual solution x^* has relatively few non-zero coefficients, corresponding to samples that lie on the so-called margin of the support vector machine.

The sketched version is then given by

$$\hat{x} := \arg \min_{x \in \mathbb{R}^d} \|SBx\|_2^2 \text{ s.t. } x \geq 0 \text{ and } \sum_{i=1}^d x_i = 1. \quad (2.28)$$

The simplex constraint in the quadratic program (2.27), although not identical to an ℓ_1 -constraint, leads to similar scaling in terms of the sketch dimension.

Corollary 5 (Sketch dimensions for support vector machines). *Given a collection of labeled samples $\{(a_i, z_i)\}_{i=1}^d$, let $\|x^*\|_0$ denote the number of samples on the margin in the SVM solution (2.27). Then given a sub-Gaussian sketch with dimension*

$$m \geq \frac{c_0}{\delta^2} \|x^*\|_0 \log(d) \max_{j=1, \dots, d} \frac{\|a_j\|_2^2}{\gamma_k^-(A)}, \quad (2.29)$$

the sketched solution (2.28) is δ -optimal with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.

We omit the proof, as the calculations specializing from Theorem 1 are essentially the same as those of Corollary 3. The computational complexity of solving the SVM problem as a linearly constrained quadratic problem is same with the Lasso problem, so that the same conclusions apply.

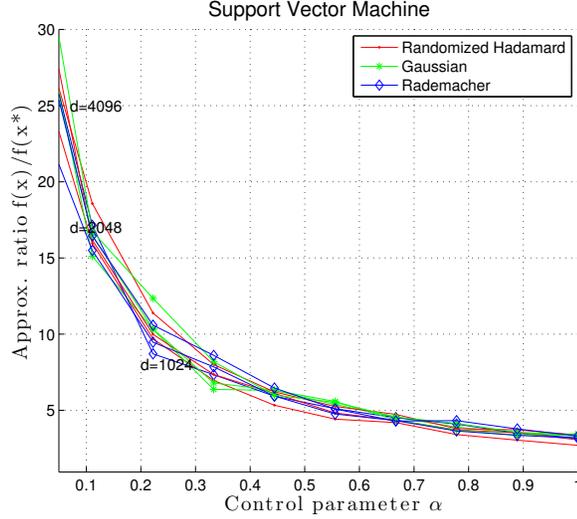


Figure 2.4: Comparison of Gaussian, Rademacher and randomized Hadamard sketches for the support vector machine (2.27). Each curve plots the approximation ratio $f(\hat{x})/f(x^*)$ versus the control parameter α , averaged over $T_{trial} = 100$ trials, for projection dimensions $m = \max\{5\alpha\|x^*\|_0 \log d, 1\}$, and problem dimensions $d \in \{1024, 2048, 4096\}$.

In order to study the prediction of Corollary 5, we generated some classification experiments, and tested the performance of the sketching procedure. Consider a two-component Gaussian mixture model, based on the component distributions $N(\mu_0, I)$ and $N(\mu_1, I)$, where μ_0 and μ_1 are uniformly distributed in $[-3, 3]$. Placing equal weights on each component, we draw d samples from this mixture distribution, and then use the resulting data to solve the SVM dual program (2.27), thereby obtaining an optimal linear decision boundary specified by the vector x^* . The number of non-zero entries $\|x^*\|_0$ corresponds to the number of examples on the decision boundary, known as support vectors. We then solve the sketched version (2.28), using either Gaussian, Rademacher or randomized Hadamard sketches, and using a projection

dimension scaling as $m = \max\{5\alpha\|x^*\|_0 \log d, 1\}$, where $\alpha \in [0, 1]$ is a control parameter. We repeat this experiment for problem dimensions $d \in \{1024, 2048, 4096\}$, performing $T_{trial} = 100$ trials for each choice of (α, d) .

Figure 2.4 shows plots of the approximation ratio versus the control parameter. Each bundle of curves corresponds to a different problem dimension, and has three curves for the three different sketch types. Consistent with the theory, in all cases, the approximation error approaches one as α scales upwards.

It is worthwhile noting that similar sketching techniques can be applied to other optimization problems that involve the unit simplex as a constraint. Another instance is the Markowitz formulation of the portfolio optimization problem [91]. Here the goal is to estimate a vector $x \in \mathbb{R}^d$ in the unit simplex, corresponding to non-negative weights associated with each of d possible assets, so as to minimize the variance of the return subject to a lower bound on the expected return. More precisely, we let $\mu \in \mathbb{R}^d$ denote a vector corresponding to mean return associated with the assets, and we let $\Sigma \in \mathbb{R}^{d \times d}$ be a symmetric, positive semidefinite matrix, corresponding to the covariance of the returns. Typically, the mean vector and covariance matrix are estimated from data. Given the pair (μ, Σ) , the Markowitz allocation is given by

$$x^* = \arg \min_{x \in \mathbb{R}^d} x^T \Sigma x \quad \text{such that } \langle \mu, x \rangle \geq \gamma, x \geq 0 \text{ and } \sum_{j=1}^d x_j = 1. \quad (2.30)$$

Note that this problem can be written in the same form as the SVM, since the covariance matrix $\Sigma \succeq 0$ can be factorized as $\Sigma = A^T A$. Whenever the expected return constraint $\langle \mu, x \rangle \geq \gamma$ is active at the solution, the tangent cone is given by

$$\mathcal{K} = \left\{ \Delta \in \mathbb{R}^d \mid \langle \mu, \Delta \rangle \geq 0, \sum_{j=1}^d \Delta_j = 0, \Delta_{S^c} \geq 0 \right\}$$

where S is the support of x^* . This tangent cone is a subset of the tangent cone for the SVM, and hence the bounds of Corollary 5 also apply to the portfolio optimization problem.

2.2.5 Matrix estimation with nuclear norm regularization

We now turn to the use of sketching for matrix estimation problems, and in particular those that involve nuclear norm constraints. Let $\mathcal{C} \subset \mathbb{R}^{d_1 \times d_2}$ be a convex subset of the space of all $d_1 \times d_2$ matrices. Many matrix estimation problems can be written in the general form

$$\min_{X \in \mathcal{C}} \|y - \mathcal{A}(X)\|_2^2$$

where $y \in \mathbb{R}^n$ is a data vector, and \mathcal{A} is a linear operator from $\mathbb{R}^{d_1 \times d_2}$ to \mathbb{R}^n . Letting vec denote the vectorized form of a matrix, we can write $\mathcal{A}(X) = A \text{vec}(X)$ for a suitably defined matrix $A \in \mathbb{R}^{n \times D}$, where $D = d_1 d_2$. Consequently, our general sketching techniques are again applicable.

In many matrix estimation problems, of primary interest are matrices of relatively low rank. Since rank constraints are typically computationally intractable, a standard convex surrogate is the nuclear norm of matrix, given by the sum of its singular values

$$\|X\|_* = \sum_{j=1}^{\min\{d_1, d_2\}} \sigma_j(X). \quad (2.31)$$

As an illustrative example, let us consider the problem of weighted low-rank matrix approximation. Suppose that we wish to approximate a given matrix $Z \in \mathbb{R}^{d_1 \times d_2}$ by a low-rank matrix X of the same dimensions, where we measure the quality of approximation using a weighted Frobenius norm

$$\|Z - X\|_\omega^2 = \sum_{j=1}^{d_2} \omega_j^2 \|z_j - x_j\|_2^2, \quad (2.32)$$

where z_j and x_j are the j^{th} columns of Z and X respectively, and $\omega \in \mathbb{R}^{d_2}$ is a vector of non-negative weights. If the weight vector is uniform ($\omega_j = c$ for all $j = 1, \dots, d$), then the norm $\|\cdot\|_\omega$ is simply the usual Frobenius norm, a low-rank minimizer can be obtained by computing a partial singular value decomposition of the data matrix Y . For non-uniform weights, it is no longer easy to solve the rank-constrained minimization problem. Accordingly, it is natural to consider the convex relaxation

$$X^* := \arg \min_{\|X\|_* \leq R} \|Z - X\|_\omega^2, \quad (2.33)$$

in which the rank constraint is replaced by the nuclear norm constraint $\|X\|_* \leq R$. This program can be written in an equivalent vectorized form in dimension $D = d_1 d_2$ by defining the block-diagonal matrix $A = \text{blkdiag}(\omega_1 I, \dots, \omega_{d_2} I)$, as well as the vector $y \in \mathbb{R}^D$ whose j^{th} block is given by $\omega_j y_j$. We can then consider the equivalent problem $X^* := \arg \min_{\|X\|_* \leq R} \|y - A \text{vec}(X)\|_2^2$, as well as its sketched version

$$\hat{X} := \arg \min_{\|X\|_* \leq R} \|S y - S A \text{vec}(X)\|_2^2. \quad (2.34)$$

Suppose that the original optimum X^* has rank r : it then be described using at $\mathcal{O}(r(d_1 + d_2))$ real numbers. Intuitively, it should be possible to project the original problem down to this dimension while still guaranteeing an accurate solution. The following corollary provides a rigorous confirmation of this intuition:

Corollary 6 (Sketch dimensions for weighted low-rank approximation). *Consider the weighted low-rank approximation problem (2.33) based on a weight vector with condition number $\kappa^2(\omega) = \frac{\max_{j=1,\dots,d} \omega_j^2}{\min_{j=1,\dots,d} \omega_j^2}$, and suppose that the optimal solution has rank $r = \text{rank}(X^*)$.*

(a) *For sub-Gaussian sketches, a sketch dimension lower bounded by*

$$m \geq \frac{c_0}{\delta^2} \kappa^2(\omega) r (d_1 + d_2) \quad (2.35)$$

guarantees that the sketched solution (2.34) is δ -optimal (2.3) with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.

(b) *For ROS sketches, a sketch dimension lower bounded by*

$$m > \frac{c'_0}{\delta^2} \kappa^2(\omega) r (d_1 + d_2) \log^4(d_1 d_2). \quad (2.36)$$

guarantees that the sketched solution (2.34) is δ -optimal (2.3) with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.

For this particular application, the use of sketching is not likely to lead to substantial computational savings, since the optimization space remains $d_1 d_2$ dimensional in both the original and sketched versions. However, the lower dimensional nature of the sketched data can be still very useful in reducing storage requirements and privacy-preserving optimization.

Proof. We prove part (a) here, leaving the proof of part (b) to Section 2.4.4. Throughout the proof, we adopt the shorthand notation $\omega_{\min} = \min_{j=1,\dots,d} \omega_j$ and $\omega_{\max} = \max_{j=1,\dots,d} \omega_j$. As shown in past work on nuclear norm regularization (see Lemma 1 in the paper [101]), the tangent cone of the nuclear norm constraint $\|X\|_* \leq R$ at a rank r matrix is contained within the cone

$$\mathcal{K}' = \{ \Delta \in \mathbb{R}^{d_1 \times d_2} \mid \|\Delta\|_* \leq 2\sqrt{r} \|\Delta\|_F \}. \quad (2.37)$$

For any matrix Δ with $\|A \text{vec}(\Delta)\|_2 = 1$, we must have $\|\Delta\|_{\text{F}} = \|\text{vec}(\Delta)\|_2 \leq \frac{1}{\omega_{\min}}$. By definition of the Gaussian width, we then have

$$\mathbb{W}(A\mathcal{K}) \leq \frac{1}{\omega_{\min}} \mathbb{E} \left[\sup_{\|\Delta\|_* \leq 2\sqrt{r}} |\langle A^T g, \text{vec}(\Delta) \rangle| \right].$$

Since A^T is a diagonal matrix, the vector $A^T g$ has independent entries with maximal variance ω_{\max}^2 . Letting $G \in \mathbb{R}^{d_1 \times d_2}$ denote the matrix formed by segmenting the vector $A^T g$ into d_2 blocks of length d_1 , we have

$$\begin{aligned} \mathbb{W}(A\mathcal{K}) &\leq \frac{1}{\omega_{\min}} \mathbb{E} \left[\sup_{\|\Delta\|_* \leq 2\sqrt{r}} |\text{trace}(G\Delta)| \right] \\ &\leq \frac{2\sqrt{r}}{\omega_{\min}} \mathbb{E} [\|G\|_2] \end{aligned}$$

where we have used the duality between the operator and nuclear norms. By standard results on operator norms of Gaussian random matrices [44], we have $\mathbb{E}[\|G\|_2] \leq \omega_{\max}(\sqrt{d_1} + \sqrt{d_2})$, and hence

$$\mathbb{W}(A\mathcal{K}) \leq 2 \frac{\omega_{\max}}{\omega_{\min}} \sqrt{r} (\sqrt{d_1} + \sqrt{d_2}).$$

Thus, the bound (2.35) follows as a corollary of Theorem 1. \square

2.2.6 Group sparse regularization

As a final example, let us consider optimization problems that involve constraints to enforce group sparsity. This notion is a generalization of elementwise sparsity, defined in terms of a partition \mathcal{G} of the index set $[d] = \{1, 2, \dots, d\}$ into a collection of non-overlapping subsets, referred to as groups. Given a group $g \in \mathcal{G}$ and a vector $x \in \mathbb{R}^d$, we use $x_g \in \mathbb{R}^{|g|}$ to denote the sub-vector indexed by elements of g . A basic form of the group Lasso norm [154] is given by

$$\|x\|_{\mathcal{G}} = \sum_{g \in \mathcal{G}} \|x_g\|_2. \quad (2.38)$$

Note that in the special case that \mathcal{G} consists of d groups, each of size 1, this norm reduces to the usual ℓ_1 -norm. More generally, with non-trivial grouping, it defines a

second-order cone constraint [28]. Bach et al. [17] provide an overview of the group Lasso norm (2.38), as well as more exotic choices for enforcing group sparsity.

Here let us consider the problem of sketching the second-order cone program (SOCP)

$$x^* = \arg \min_{\|x\|_{\mathcal{G}} \leq R} \|Ax - y\|_2^2. \quad (2.39)$$

We let k denote the number of active groups in the optimal solution x^* —that is, the number of groups for which $x_g^* \neq 0$. For any group $g \in \mathcal{G}$, we use A_g to denote the $n \times |g|$ sub-matrix with columns indexed by g . In analogy to the sparse RE condition (2.13), we define the group-sparse restricted eigenvalue $\gamma_{k,\mathcal{G}}^-(A) := \min_{\substack{\|z\|_2=1 \\ \|z\|_{\mathcal{G}} \leq 2\sqrt{k}}} \|Az\|_2^2$.

Corollary 7 (Guarantees for group-sparse least-squares squares). *For the group Lasso program (2.39) with maximum group size $M = \max_{g \in \mathcal{G}} |g|$, a projection dimension lower bounded as*

$$m \geq \frac{c_0}{\delta^2} \min \left\{ \text{rank}(A), \max_{g \in \mathcal{G}} \frac{\|A_g\|_2}{\gamma_{k,\mathcal{G}}^-(A)} (k \log |\mathcal{G}| + kM) \right\} \quad (2.40)$$

guarantees that the sketched solution is δ -optimal (2.3) with probability at least $1 - c_1 e^{-c_2 m \delta^2}$.

Note that this is a generalization of Corollary 3 on sketching the ordinary Lasso. Indeed, when we have $|\mathcal{G}| = d$ groups, each of size $M = 1$, then the lower bound (2.40) reduces to the lower bound (2.15). As might be expected, the proof of Corollary 7 is similar to that of Corollary 3. It makes use of some standard results on the expected maxima of χ^2 -variables to upper bound the Gaussian complexity; see the paper [100] for more details on this calculation.

2.3 Proofs of main results

We now turn to the proofs of our main results, namely Theorem 1 on sub-Gaussian sketching, and Theorem 2 on sketching with randomized orthogonal systems. At a high level, the proofs consists of two parts. The first part is a deterministic argument, using convex optimality conditions. The second step is probabilistic, and depends on the particular choice of random sketching matrices.

2.3.1 Main argument

Central to the proofs of both Theorem 1 and 2 are the following two variational quantities:

$$Z_1(A\mathcal{K}) := \inf_{v \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \frac{1}{m} \|Sv\|_2^2, \quad \text{and} \quad (2.41a)$$

$$Z_2(A\mathcal{K}) := \sup_{v \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \left| \langle u, \left(\frac{S^T S}{m} - I \right) v \rangle \right|, \quad (2.41b)$$

where we recall that \mathcal{S}^{n-1} is the Euclidean unit sphere in \mathbb{R}^n , and in equation (2.41b), the vector $u \in \mathcal{S}^{n-1}$ is fixed but arbitrary. These are deterministic quantities for any fixed choice of sketching matrix S , but random variables for randomized sketches. As it will be illustrated by our subsequent analysis, these quantities isolate the stochastic nature of the random sketch S and are considerably easier to analyze owing to connections with some well-studied sub-Gaussian empirical processes (e.g. see [97]). The following lemma demonstrates the significance of these two quantities:

Lemma 2. *For any sketching matrix $S \in \mathbb{R}^{m \times n}$, we have*

$$f(\hat{x}) \leq \left\{ 1 + 2 \frac{Z_2(A\mathcal{K})}{Z_1(A\mathcal{K})} \right\}^2 f(x^*) \quad (2.42)$$

Consequently, we see that in order to establish that \hat{x} is δ -optimal, we need to control the ratio $Z_2(A\mathcal{K})/Z_1(A\mathcal{K})$.

Proof. Define the error vector $\hat{e} := \hat{x} - x^*$. We first assume $f(x^*) = \|Ax^* - y\|_2^2 > 0$ and we shall return to this case later. By the triangle inequality, we have

$$\|A\hat{x} - y\|_2 \leq \|Ax^* - y\|_2 + \|A\hat{e}\|_2 \quad (2.43)$$

$$= \|Ax^* - y\|_2 \left\{ 1 + \frac{\|A\hat{e}\|_2}{\|Ax^* - y\|_2} \right\}. \quad (2.44)$$

Squaring both sides yields

$$f(\hat{x}) \leq \left(1 + \frac{\|A\hat{e}\|_2}{\|Ax^* - y\|_2} \right)^2 f(x^*).$$

Consequently, it suffices to control the ratio $\frac{\|A\hat{e}\|_2}{\|Ax^* - y\|_2}$, and we use convex optimality conditions to do so. If $\|A\hat{e}\|_2 = 0$, the claim (2.42) is trivially true, hence we assume $\|A\hat{e}\|_2 > 0$ without loss of generality.

Since \hat{x} and x^* are optimal and feasible, respectively, for the sketched problem (2.2), we have $g(\hat{x}) \leq g(x^*)$, or equivalently

$$\frac{1}{2m} \|SA\hat{e} + SAx^* - Sy\|_2^2 \leq \frac{1}{2m} \|SAx^* - Sy\|_2^2.$$

Expanding the left-hand-side and subtracting $\frac{1}{2m} \|SAx^* - Sy\|_2^2$ from both sides yields

$$\begin{aligned} \frac{1}{2m} \|SA\hat{e}\|_2^2 &\leq -\langle Ax^* - y, \frac{1}{m} S^T S A\hat{e} \rangle \\ &= -\langle Ax^* - y, (\frac{1}{m} S^T S - I) A\hat{e} \rangle - \langle Ax^* - y, A\hat{e} \rangle, \end{aligned}$$

where we have added and subtracted $\langle Ax^* - y, A\hat{e} \rangle$. Now by the optimality of x^* for the original problem (2.1), we have

$$\langle (Ax^* - y), A\hat{e} \rangle = \langle A^T(Ax^* - y), \hat{x} - x^* \rangle \geq 0,$$

and hence

$$\frac{1}{2m} \|SA\hat{e}\|_2^2 \leq \left| \langle Ax^* - y, (\frac{1}{m} S^T S - I) A\hat{e} \rangle \right|. \quad (2.45)$$

Letting $\{s_i\}_{i=1}^m$ correspond to the rows of S , note that the first term in the above right-hand side contains the random matrix

$$\frac{1}{m} S^T S - I = \frac{1}{m} \sum_{i=1}^m s_i s_i^T - I.$$

Since $\mathbb{E}s_1 s_1^T = I$, this random matrix is zero-mean and it should be possible to control its fluctuations as a function of m , and the two vectors $Ax^* - y$ and $A\hat{e}$ that also arise in the inequality (2.45). Whereas the vector $Ax^* - y$ is non-random, the challenge here is that \hat{e} is a random vector that also depends on the sketch matrix. For this reason, we need to prove a form of uniform law of large numbers of this term. In this context, the previously defined quantities $Z_1(A\mathcal{K})$ and $Z_2(A\mathcal{K})$ play the role of

uniform lower and upper bounds on appropriately scaled form of the left-hand-side and right-hand side (respectively) of the inequality (2.45). Renormalizing the right-hand side of inequality (2.45), we find that

$$\frac{1}{2m} \|SA\hat{e}\|_2^2 \leq \|Ax^* - y\|_2 \|A\hat{e}\|_2 \quad \left| \left\langle \frac{Ax^* - y}{\|Ax^* - y\|_2}, \left(\frac{1}{m} S^T S - I \right) \frac{A\hat{e}}{\|A\hat{e}\|_2} \right\rangle \right|.$$

By the optimality of \hat{x} , we have $A\hat{e} \in \mathcal{AK}$ and $\frac{Ax^* - y}{\|Ax^* - y\|_2}$ is a fixed unit-norm vector, whence the basic inequality (2.46) and definitions (2.41a) and (2.41b) imply that

$$\frac{1}{2} Z_1(\mathcal{AK}) \|A\hat{e}\|_2^2 \leq \|A\hat{e}\|_2 \|Ax^* - y\|_2 Z_2(\mathcal{AK})$$

Cancelling terms yields the inequality

$$\frac{\|A\hat{e}\|_2}{\|Ax^* - y\|_2} \leq 2 \frac{Z_2(\mathcal{AK})}{Z_1(\mathcal{AK})}.$$

Combined with our earlier inequality (2.43), the claim (2.42) follows for $\|Ax^* - y\|_2^2 > 0$.

Finally, consider the special case $f(x^*) = \|Ax^* - y\|_2^2 = 0$, and show that $f(\hat{x}) = 0$. Since inequality (2.45) still holds, we find that

$$\frac{1}{2m} \|SA\hat{e}\|_2^2 \leq 0.$$

Combined with the definition (2.41a) of $Z_1(\mathcal{AK})$, we see that $\frac{1}{2} Z_1(\mathcal{AK}) \|A\hat{e}\|_2^2 \leq 0$. As long as $Z_1(\mathcal{AK}) > 0$, we are thus guaranteed that $\|A\hat{e}\|_2 = 0$. Since $\|A\hat{x} - y\|_2 \leq \|A\hat{e}\|_2$, we conclude that $f(\hat{x}) = \|A\hat{x} - y\|_2^2 = 0$ as claimed. \square

2.3.2 Proof of Theorem 1

In order to complete the proof of Theorem 1, we need to upper bound the ratio $Z_2(\mathcal{AK})/Z_1(\mathcal{AK})$. The following lemmas provide such control in the sub-Gaussian case. As usual, we let $S \in \mathbb{R}^{m \times n}$ denote the matrix with the vectors $\{s_i\}_{i=1}^m$ as its rows.

Lemma 3 (Lower bound on $Z_1(A\mathcal{K})$). *Under the conditions of Theorem 1, for i.i.d. σ -sub-Gaussian vectors $\{s_i\}_{i=1}^m$, we have*

$$\underbrace{\inf_{v \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \frac{1}{m} \|Sv\|_2^2}_{Z_1(A\mathcal{K})} \geq 1 - \delta \quad (2.46)$$

with probability at least $1 - \exp(-c_1 \frac{m\delta^2}{\sigma^4})$.

Lemma 4 (Upper bound on $Z_2(A\mathcal{K})$). *Under the conditions of Theorem 1, for i.i.d. σ -sub-Gaussian vectors $\{s_i\}_{i=1}^m$ and any fixed vector $u \in \mathcal{S}^{n-1}$, we have*

$$\underbrace{\sup_{v \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \left| \langle u, (\frac{1}{m} S^T S - I) v \rangle \right|}_{Z_2(A\mathcal{K})} \leq \delta \quad (2.47)$$

with probability at least $1 - 6 \exp(-c_1 \frac{m\delta^2}{\sigma^4})$.

Taking these two lemmas as given, we can complete the proof of Theorem 1. As long as $\delta \in (0, 1/2)$, they imply that

$$2 \frac{Z_2(A\mathcal{K})}{Z_1(A\mathcal{K})} \leq \frac{2\delta}{1-\delta} \leq 4\delta \quad (2.48)$$

with probability at least $1 - 4 \exp(-c_1 \frac{m\delta^2}{\sigma^4})$. The rescaling $4\delta \mapsto \delta$, with appropriate changes of the universal constants, yields the result.

It remains to prove the two lemmas. In the sub-Gaussian case, both of these results exploit a result due to Mendelson et al. [97]:

Proposition 1. *Let $\{s_i\}_{i=1}^m$ be i.i.d. samples from a zero-mean σ -sub-Gaussian distribution with $\text{cov}(s_i) = I_{n \times n}$. Then there are universal constants such that for any subset $\mathcal{Y} \subseteq \mathcal{S}^{n-1}$, we have*

$$\sup_{y \in \mathcal{Y}} \left| y^T \left(\frac{S^T S}{m} - I_{n \times n} \right) y \right| \leq c_1 \frac{\mathbb{W}(\mathcal{Y})}{\sqrt{m}} + \delta \quad (2.49)$$

with probability at least $1 - e^{-\frac{c_2 m \delta^2}{\sigma^4}}$.

This claim follows from their Theorem D, using the linear functions $f_y(s) = \langle s, y \rangle$.

2.3.2.1 Proof of Lemma 3

Lemma 3 follows immediately from Proposition 1: in particular, the bound (2.49) with the set $\mathcal{Y} = \mathcal{AK} \cap \mathcal{S}^{n-1}$ ensures that

$$\inf_{v \in \mathcal{AK} \cap \mathcal{S}^{n-1}} \frac{\|Sv\|_2^2}{m} \geq 1 - c_1 \frac{\mathbb{W}(\mathcal{Y})}{\sqrt{m}} - \frac{\delta}{2} \stackrel{(i)}{\geq} 1 - \delta,$$

where inequality (i) follows as long as $m > \frac{c_0}{\delta^2} \mathbb{W}(\mathcal{AK})$ for a sufficiently large universal constant.

2.3.2.2 Proof of Lemma 4

The proof of this claim is more involved. Let us partition the set $\mathcal{V} = \mathcal{AK} \cap \mathcal{S}^{n-1}$ into two disjoint subsets, namely

$$\begin{aligned} \mathcal{V}_+ &= \{v \in \mathcal{V} \mid \langle u, v \rangle \geq 0\}, \quad \text{and} \\ \mathcal{V}_- &= \{v \in \mathcal{V} \mid \langle u, v \rangle < 0\}. \end{aligned}$$

Introducing the shorthand $Q = \frac{S^T S}{m} - I$, we then have

$$Z_2(\mathcal{AK}) \leq \sup_{v \in \mathcal{V}_+} |u^T Qv| + \sup_{v \in \mathcal{V}_-} |u^T Qv|,$$

and we bound each of these terms in turn.

Beginning with the first term, for any $v \in \mathcal{V}_+$, the triangle inequality implies that

$$\begin{aligned} |u^T Qv| &\leq \frac{1}{2} |(u+v)^T Q(u+v)| + \frac{1}{2} |u^T Qu| \\ &\quad + \frac{1}{2} |v^T Qv|. \end{aligned} \tag{2.50}$$

Defining the set $\mathcal{U}_+ := \{\frac{u+v}{\|u+v\|_2} \mid v \in \mathcal{V}_+\}$, we apply Proposition 1 three times in succession, with the choices $\mathcal{Y} = \mathcal{U}_+$, $\mathcal{Y} = \mathcal{V}_+$ and $\mathcal{Y} = \{u\}$ respectively, which yields

$$\sup_{v \in \mathcal{V}_+} \frac{|(u+v)^T Q(u+v)|}{\|u+v\|_2^2} \leq c_1 \frac{\mathbb{W}(\mathcal{U}_+)}{\sqrt{m}} + \delta \tag{2.51a}$$

$$\sup_{v \in \mathcal{AK} \cap \mathcal{S}^{n-1}} |v^T Qv| \leq c_1 \frac{\mathbb{W}(\mathcal{AK} \cap \mathcal{S}^{n-1})}{\sqrt{m}} + \delta, \tag{2.51b}$$

$$|u^T Qu| \leq c_1 \frac{\mathbb{W}(\{u\})}{\sqrt{m}} + \delta. \tag{2.51c}$$

All three bounds hold with probability at least $1 - 3e^{-c_2 m \delta^2 / \sigma^4}$. Note that $\|u+v\|_2^2 \leq 4$, so that the bound (2.51a) implies that $|(u+v)^T Q(u+v)| \leq 4c_1 \mathbb{W}(\mathcal{U}_+) + 4\delta$ for all $v \in \mathcal{V}_+$. Thus, when inequalities (2.51a) through (2.51c) hold, the decomposition (2.50) implies that

$$\begin{aligned} & |u^T Q u| \\ & \leq \frac{c_1}{2} \{4\mathbb{W}(\mathcal{U}_+) + \mathbb{W}(A\mathcal{K} \cap \mathcal{S}^{n-1}) + \mathbb{W}(\{u\})\} + 3\delta. \end{aligned} \quad (2.52)$$

It remains to simplify the sum of the three Gaussian complexity terms. An easy calculation gives $\mathbb{W}(\{u\}) \leq \sqrt{2/\pi} \leq \mathbb{W}(A\mathcal{K} \cap \mathcal{S}^{n-1})$. In addition, we claim that

$$\mathbb{W}(\mathcal{U}_+) \leq \mathbb{W}(\{u\}) + \mathbb{W}(A\mathcal{K} \cap \mathcal{S}^{n-1}). \quad (2.53)$$

Given any $v \in \mathcal{V}_+$, let $\Pi(v)$ denote its projection onto the subspace orthogonal to u . We can then write $v = \alpha u + \Pi(v)$ for some scalar $\alpha \in [0, 1]$, where $\|\Pi(v)\|_2 = \sqrt{1 - \alpha^2}$. In terms of this decomposition, we have

$$\begin{aligned} \|u+v\|_2^2 &= \|(1+\alpha)u + \Pi(v)\|_2^2 \\ &= (1+\alpha)^2 + 1 - \alpha^2 \\ &= 2 + 2\alpha. \end{aligned}$$

Consequently, we have

$$\begin{aligned} \left| \left\langle g, \frac{u+v}{\|u+v\|_2} \right\rangle \right| &= \left| \frac{(1+\alpha)}{\sqrt{2(1+\alpha)}} \langle g, u \rangle + \frac{1}{\sqrt{2(1+\alpha)}} \langle g, \Pi(v) \rangle \right| \\ &\leq |\langle g, u \rangle| + |\langle g, \Pi(v) \rangle|. \end{aligned} \quad (2.54)$$

For any pair $v, v' \in \mathcal{V}_+$, note that

$$\begin{aligned} \text{var}(\langle g, \Pi(v) \rangle - \langle g, \Pi(v') \rangle) &= \|\Pi(v) - \Pi(v')\|_2^2 \leq \|v - v'\|_2^2 \\ &= \text{var}(\langle g, v \rangle - \langle g, v' \rangle). \end{aligned}$$

where the inequality follows by the non-expansiveness of projection. Consequently, by the Sudakov-Fernique comparison inequality [85], we have

$$\mathbb{E} \left[\sup_{v \in \mathcal{V}_+} |\langle g, \Pi(v) \rangle| \right] \leq \mathbb{E} \left[\sup_{v \in \mathcal{V}_+} |\langle g, v \rangle| \right] = \mathbb{W}(\mathcal{V}_+).$$

Since $\mathcal{V}_+ \subseteq A\mathcal{K} \cap \mathcal{S}^{n-1}$, we have $\mathbb{W}(\mathcal{V}_+) \leq \mathbb{W}(A\mathcal{K} \cap \mathcal{S}^{n-1})$. Combined with our earlier inequality (2.54), we have shown that

$$\mathbb{W}(\mathcal{U}_+) \leq \mathbb{W}(\{u\}) + \mathbb{W}(A\mathcal{K} \cap \mathcal{S}^{n-1}) \leq 2\mathbb{W}(A\mathcal{K} \cap \mathcal{S}^{n-1}).$$

Substituting back into our original upper bound (2.52), we have established that

$$\begin{aligned} & \sup_{v \in \mathcal{V}_+} |u^T Q v| \\ & \leq \frac{c_1}{2\sqrt{m}} \{8\mathbb{W}(A\mathcal{K} \cap \mathcal{S}^{n-1}) + 2\mathbb{W}(A\mathcal{K} \cap \mathcal{S}^{n-1})\} + 3\delta \end{aligned} \quad (2.55)$$

$$= \frac{5c_1}{\sqrt{m}} \mathbb{W}(A\mathcal{K} \cap \mathcal{S}^{n-1}) + 3\delta. \quad (2.56)$$

with high probability.

As for the supremum over \mathcal{V}_- , in this case, we use the decomposition

$$u^T Q v = \frac{1}{2} \left\{ v^T Q v + u^T Q u - (v - u)^T Q (v - u) \right\}.$$

The analogue of \mathcal{U}_+ is the set $\mathcal{U}_- = \left\{ \frac{v-u}{\|v-u\|_2} \mid v \in \mathcal{V}_- \right\}$. Since $\langle -u, v \rangle \geq 0$ for all $v \in \mathcal{V}_-$, the same argument as before can be applied to show that $\sup_{v \in \mathcal{V}_-} |u^T Q v|$ satisfies the same bound (2.55) with high probability.

Putting together the pieces, we have established that, with probability at least $1 - 6e^{-c_2 m \delta^2 / \sigma^4}$, we have

$$\begin{aligned} Z_2(A\mathcal{K}) &= \sup_{v \in A\mathcal{K} \cap \mathcal{S}^{n-1}} |u^T Q v| \\ &\leq \frac{10c_1}{\sqrt{m}} \mathbb{W}(A\mathcal{K} \cap \mathcal{S}^{n-1}) + 6\delta \\ &\stackrel{(i)}{\leq} 9\delta, \end{aligned}$$

where inequality (i) makes use of the assumed lower bound on the projection dimension. The claim follows by rescaling δ and redefining the universal constants appropriately.

2.3.3 Proof of Theorem 2

We begin by stating two technical lemmas that provide control on the random variables $Z_1(A\mathcal{K})$ and $Z_2(A\mathcal{K})$ for randomized orthogonal systems. These results involve the S -Gaussian width previously defined in equation (2.8); we also recall the Rademacher width

$$\mathbb{R}(A\mathcal{K}) := \mathbb{E}_\varepsilon \left[\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} |\langle z, \varepsilon \rangle| \right]. \quad (2.57)$$

Lemma 5 (Lower bound on $Z_1(A\mathcal{K})$). *Given a projection size m satisfying the bound (2.10) for a sufficiently large universal constant c_0 , we have*

$$\underbrace{\inf_{v \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \frac{1}{m} \|Sv\|_2^2}_{Z_1(A\mathcal{K})} \geq 1 - \delta \quad (2.58)$$

with probability at least $1 - \frac{c_1}{(mn)^2} - c_1 \exp\left(-c_2 \frac{m\delta^2}{\mathbb{R}^2(A\mathcal{K}) + \log(mn)}\right)$.

Lemma 6 (Upper bound on $Z_2(A\mathcal{K})$). *Given a projection size m satisfying the bound (2.10) for a sufficiently large universal constant c_0 , we have*

$$\underbrace{\sup_{v \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \left| \left\langle u, \left(\frac{S^T S}{m} - I \right) v \right\rangle \right|}_{Z_2(A\mathcal{K})} \leq \delta \quad (2.59)$$

with probability at least $1 - \frac{c_1}{(mn)^2} - c_1 \exp\left(-c_2 \frac{m\delta^2}{\mathbb{R}^2(A\mathcal{K}) + \log(mn)}\right)$.

Taking them as given, the proof of Theorem 2 is easily completed. Based on a combination of the two lemmas, for any $\delta \in [0, 1/2]$, we have

$$2 \frac{Z_2(A\mathcal{K})}{Z_1(A\mathcal{K})} \leq \frac{2\delta}{1 - \delta} \leq 4\delta,$$

with probability at least $1 - \frac{c_1}{(mn)^2} - c_1 \exp\left(-c_2 \frac{m\delta^2}{\mathbb{R}^2(A\mathcal{K}) + \log(mn)}\right)$. The claimed form of the bound follows via the rescaling $\delta \mapsto 4\delta$, and suitable adjustments of the universal constants.

In the following, we use $\mathbb{B}_2^n = \{z \in \mathbb{R}^n \mid \|z\|_2 \leq 1\}$ to denote the Euclidean ball of radius one in \mathbb{R}^n .

Proposition 2. *Let $\{s_i\}_{i=1}^m$ be i.i.d. samples from a randomized orthogonal system.*

Then for any subset $\mathcal{Y} \subseteq \mathbb{B}_2^n$ and any $\delta \in [0, 1]$ and $\kappa > 0$, we have

$$\begin{aligned} & \sup_{y \in \mathcal{Y}} \left| y^T \left(\frac{S^T S}{m} - I \right) y \right| \\ & \leq 8 \left\{ \mathbb{R}(\mathcal{Y}) + \sqrt{2(1 + \kappa) \log(mn)} \right\} \frac{\mathbb{W}_S(\mathcal{Y})}{\sqrt{m}} + \frac{\delta}{2} \end{aligned} \quad (2.60)$$

with probability at least $1 - \frac{c_1}{(mn)^\kappa} - c_1 \exp\left(-c_2 \frac{m\delta^2}{\mathbb{R}^2(\mathcal{Y}) + \log(mn)}\right)$.

2.3.3.1 Proof of Lemma 5

This lemma is an immediate consequence of Proposition 2 with $\mathcal{Y} = A\mathcal{K} \cap \mathcal{S}^{n-1}$ and $\kappa = 2$. In particular, with a sufficiently large constant c_0 , the lower bound (2.10) on the projection dimension ensures that $8\left\{\mathbb{R}(\mathcal{Y}) + \sqrt{6 \log(mn)}\right\} \leq \frac{\delta}{2}$, from which the claim follows.

2.3.3.2 Proof of Lemma 6

We again introduce the convenient shorthand $Q = \frac{S^T S}{m} - I$. For any subset $\mathcal{Y} \subseteq \mathbb{B}_2^n$, define the random variable $Z_0(\mathcal{Y}) = \sup_{y \in \mathcal{Y}} |y^T Q y|$. Note that Proposition 2 provides control on any such random variable. Now given the fixed unit-norm vector $u \in \mathbb{R}^n$, define the set

$$\mathcal{V} = \frac{1}{2}\{u + v \mid v \in A\mathcal{K} \cap \mathcal{S}^{n-1}\}.$$

Since $\|u + v\|_2 \leq \|u\|_2 + \|v\|_2 = 2$, we have the inclusion $\mathcal{V} \subseteq \mathbb{B}_2^n$. For any $v \in A\mathcal{K} \cap \mathcal{S}^{n-1}$, the triangle inequality implies that

$$\begin{aligned} & |u^T Q v| \\ &= 4\left|\left(\frac{u+v}{2}\right)^T Q \left(\frac{u+v}{2}\right)\right| + |v^T Q v| + |u^T Q u| \\ &\leq 4Z_0(\mathcal{V}) + Z_0(A\mathcal{K} \cap \mathcal{S}^{n-1}) + Z_0(\{u\}). \end{aligned}$$

We now apply Proposition 2 in three times in succession with the sets $\mathcal{Y} = \mathcal{V}$, $\mathcal{Y} = A\mathcal{K} \cap \mathcal{S}^{n-1}$ and $\mathcal{Y} = \{u\}$, thereby finding that

$$\begin{aligned} & |u^T Q v| \\ &\leq \frac{1}{\sqrt{m}}\left\{4\Phi(\mathcal{V}) + \Phi(A\mathcal{K} \cap \mathcal{S}^{n-1}) + \Phi(\{u\})\right\} + 3\delta, \end{aligned}$$

where we have defined the set-based function

$$\Phi(\mathcal{Y}) = 8\left\{\mathbb{R}(\mathcal{Y}) + \sqrt{6 \log(mn)}\right\} \mathbb{W}_S(\mathcal{Y})$$

By inspection, we have $\mathbb{R}(\{u\}) \leq 1 \leq 2\mathbb{R}(A\mathcal{K} \cap \mathcal{S}^{n-1})$ and $\mathbb{W}_S(\{u\}) \leq 1 \leq 2\mathbb{W}_S(A\mathcal{K})$, and hence $\Phi(\{u\}) \leq 2\Phi(A\mathcal{K} \cap \mathcal{S}^{n-1})$. Moreover, by the triangle inequality, we have

$$\begin{aligned} \mathbb{R}(\mathcal{V}) &\leq \mathbb{E}_\varepsilon |\langle \varepsilon, u \rangle| + \mathbb{E}_\varepsilon \left[\sup_{v \in A\mathcal{K} \cap \mathcal{S}^{n-1}} |\langle \varepsilon, v \rangle| \right] \\ &\leq 1 + \mathbb{R}(A\mathcal{K} \cap \mathcal{S}^{n-1}) \leq 4\mathbb{R}(A\mathcal{K} \cap \mathcal{S}^{n-1}). \end{aligned}$$

A similar argument yields $\mathbb{W}_S(\mathcal{V}) \leq 3\mathbb{W}_S(A\mathcal{K})$, and putting together the pieces yields

$$\begin{aligned} \Phi(\mathcal{V}) &\leq 8\{3\mathbb{R}(A\mathcal{K} \cap \mathcal{S}^{n-1}) + \sqrt{6\log(mn)}\} (3\mathbb{W}_S(A\mathcal{K})) \\ &\leq 9\Phi(A\mathcal{K} \cap \mathcal{S}^{n-1}). \end{aligned}$$

Putting together the pieces, we have shown that for any $v \in A\mathcal{K} \cap \mathcal{S}^{n-1}$,

$$|u^T Qv| \leq \frac{39}{\sqrt{m}}\Phi(A\mathcal{K} \cap \mathcal{S}^{n-1}) + 3\delta.$$

Using the lower bound (2.10) on the projection dimension, we are have $\frac{39}{\sqrt{m}}\Phi(A\mathcal{K} \cap \mathcal{S}^{n-1}) \leq \delta$, and hence $Z_2(A\mathcal{K}) \leq 4\delta$ with probability at least $1 - \frac{c_1}{(mn)^2} - c_1 \exp\left(-c_2 \frac{m\delta^2}{\mathbb{R}^2(A\mathcal{K}) + \log(mn)}\right)$. A rescaling of δ , along with suitable modification of the numerical constants, yields the claim.

2.3.3.3 Proof of Proposition 2

We first fix the diagonal matrix $D = \text{diag}(\nu)$, and compute probabilities over the randomness in the vectors $\tilde{s}_i = \sqrt{n}H^T p_i$, where the picking vector p_i is chosen uniformly at random from the canonical basis in \mathbb{R}^n . Using \mathbb{P}_P to denote probability taken over these i.i.d. choices, we define an i.i.d. copy S' of the sketching matrix S . Then following the classical symmetrization argument (see [118], p. 14) yields

$$\begin{aligned} \mathbb{P}_P[Z_0 \geq t] &= \mathbb{P}_P\left[\sup_{z \in \mathcal{Y}} \left| z^T \left(\frac{1}{m} S^T S - \frac{1}{m} \mathbb{E} S'^T S' \right) z \right| \right] \\ &\leq 4 \mathbb{P}_{\varepsilon, P} \left[\underbrace{\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle \tilde{s}_i, Dz \rangle \right|^2}_{Z'_0} \geq \frac{t}{4} \right], \end{aligned}$$

where $\{\varepsilon_i\}_{i=1}^m$ is an i.i.d. sequence of Rademacher variables. Now define the function $g : \{-1, 1\}^d \rightarrow \mathbb{R}$ via

$$g(\nu) := \mathbb{E}_{\varepsilon, P} \left[\sup_{y \in \mathcal{Y}} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle \tilde{s}_i, \text{diag}(\nu)y \rangle \right| \right]. \quad (2.61)$$

Note that $\mathbb{E}[g(\nu)] = \mathbb{W}_S(\mathcal{Y})$ by construction since the randomness in S consists of the choice of ν and the picking matrix P . For a truncation level $\tau > 0$ to be chosen, define the events

$$\begin{aligned} \mathcal{G}_1 &:= \left\{ \max_{j=1, \dots, n} \sup_{y \in \mathcal{Y}} |\langle \sqrt{n}h_j, \text{diag}(\nu)y \rangle| \leq \tau \right\}, \\ \mathcal{G}_2 &:= \left\{ g(\nu) \leq \mathbb{W}_S(\mathcal{Y}) + \frac{\delta}{32\tau} \right\}. \end{aligned}$$

To be clear, the only randomness involved in either event is over the Rademacher vector $\nu \in \{-1, +1\}^n$. We then condition on the event $\mathcal{G} = \mathcal{G}_1 \cap \mathcal{G}_2$ and its complement to obtain

$$\begin{aligned} \mathbb{P}_{\varepsilon, P, \nu}[Z'_0 \geq t] &= \mathbb{E}\left\{\mathbb{I}[Z'_0 \geq t] \mathbb{I}[\mathcal{G}] + \mathbb{I}[Z'_0 \geq t] \mathbb{I}[\mathcal{G}^c]\right\} \\ &\leq \mathbb{P}_{\varepsilon, P}[Z'_0 \geq t \mid \nu \in \mathcal{G}] \mathbb{P}_\nu[\mathcal{G}] + \mathbb{P}_\nu[\mathcal{G}^c]. \end{aligned}$$

We bound each of these two terms in turn.

Lemma 7. *For any $\delta \in [0, 1]$, we have*

$$\mathbb{P}_{\varepsilon, P}\left[Z'_0 \geq 2\tau \mathbb{W}_S(\mathcal{Y}) + \frac{\delta}{16} \mid \mathcal{G}\right] \mathbb{P}_D[\mathcal{G}] \leq c_1 e^{-c_2 \frac{m\delta^2}{\tau^2}}. \quad (2.62)$$

Lemma 8. *With truncation level $\tau = \mathbb{R}(\mathcal{Y}) + \sqrt{2(1 + \kappa) \log(mn)}$ for some $\kappa > 0$, we have*

$$\mathbb{P}_\nu[\mathcal{G}^c] \leq \frac{1}{(mn)^\kappa} + e^{-\frac{m\delta^2}{4096\tau^2}}. \quad (2.63)$$

See Section 2.6.2 for the proof of these two claims.

Combining Lemmas 7 and 8, we conclude that

$$\begin{aligned} &\mathbb{P}_{P, \nu}[Z \geq 8\tau \mathbb{W}_S(\mathcal{Y}) + \frac{\delta}{2}] \\ &\leq 4\mathbb{P}_{\varepsilon, P, \nu}[Z'_0 \geq 2\tau \mathbb{W}_S(\mathcal{Y}) + \frac{\delta}{8}] \\ &\leq c_1 e^{-c_2 \frac{m\delta^2}{\tau^2}} + \frac{1}{(mn)^\kappa}, \end{aligned}$$

as claimed.

2.4 Techniques for sharpening bounds

In this section, we provide some technique for obtaining sharper bounds for randomized orthonormal systems when the underlying tangent cone has particular structure. In particular, this technique can be used to obtain sharper bounds for subspaces, ℓ_1 -induced cones, as well as nuclear norm cones.

2.4.1 Sharpening bounds for a subspace

As a warm-up, we begin by showing how to obtain sharper bounds when \mathcal{K} is a subspace. For instance, this allows us to obtain the result stated in Corollary 2(b). Consider the random variable

$$\begin{aligned} Z(A\mathcal{K}) &:= \sup_{z \in A\mathcal{K} \cap \mathbb{B}_2} |z^T Q z| \\ &\geq \sup_{z \in A\mathcal{K} \cap \mathbb{S}^{n-1}} |z^T Q z|, \quad \text{where } Q = \frac{S^T S}{m} - I. \end{aligned}$$

For a parameter $\epsilon \in (0, 1)$ to be chosen, let $\{z^1, \dots, z^M\}$ be an ϵ -cover of the set $A\mathcal{K} \cap \mathbb{B}_2$. For any $z \in A\mathcal{K} \cap \mathbb{B}_2$, there is some $j \in [M]$ such that $z = z^j + \Delta$, where $\|\Delta\|_2 \leq \epsilon$. Consequently, we can write

$$|z^T Q z| \leq |(z^j)^T Q z^j| + 2|\Delta^T Q z^j| + |\Delta^T Q \Delta|.$$

Since $A\mathcal{K}$ is a subspace, the difference vector Δ also belongs to $A\mathcal{K}$. Consequently, we have

$$\begin{aligned} &|\Delta^T Q z^j| \\ &\leq \epsilon \sup_{z, z' \in A\mathcal{K} \cap \mathbb{B}_2} |z^T Q z'| \\ &= \epsilon \sup_{z, z' \in A\mathcal{K} \cap \mathbb{B}_2} \frac{1}{2} \left| 4 \left(\frac{z + z'}{2} \right)^T Q \left(\frac{z + z'}{2} \right) - z^T Q z - (z')^T Q z' \right| \\ &\leq \epsilon \sup_{z \in A\mathcal{K} \cap \mathbb{B}_2} \frac{4}{2} |z^T Q z| + \epsilon \sup_{z \in A\mathcal{K} \cap \mathbb{B}_2} |z^T Q z| + \epsilon \sup_{z \in A\mathcal{K} \cap \mathbb{B}_2} |z^T Q z| \\ &= 4\epsilon \sup_{z \in A\mathcal{K} \cap \mathbb{B}_2} |z^T Q z|. \end{aligned}$$

Noting also that $|\Delta^T Q \Delta| \leq \epsilon^2 Z(A\mathcal{K})$, we have shown that

$$(1 - 4\epsilon - \epsilon^2) Z(A\mathcal{K}) \leq \max_{j=1, \dots, M} |(z^j)^T Q z^j|.$$

Setting $\epsilon = 1/16$ yields that $Z(A\mathcal{K}) \leq \frac{3}{2} \max_{j=1, \dots, M} |(z^j)^T R z^j|$.

Having reduced the problem to a finite maximum, we can now make use of JL-embedding property of a randomized orthogonal system proven in Theorem 3.1 of Krahmer and Ward [80]: in particular, their theorem implies that for any collection of M fixed points $\{z^1, \dots, z^M\}$ and $\delta \in (0, 1)$, an ROS sketching matrix $S \in \mathbb{R}^{m \times n}$ satisfies the bounds

$$\begin{aligned} (1 - \delta) \|z^j\|_2^2 &\leq \frac{1}{m} \|S z^j\|_2^2 \leq (1 + \delta) \|z^j\|_2^2 \\ &\text{for all } j = 1, \dots, M \end{aligned} \tag{2.64}$$

with probability $1 - \eta$ if $m \geq \frac{c}{\delta^2} \log^4(n) \log(\frac{M}{\eta})$. For our chosen collection, we have $\|z^j\|_2 = 1$ for all $j = 1, \dots, M$, so that our discretization plus this bound implies that $Z(A\mathcal{K}) \leq \frac{3}{2}\delta$. Setting $\eta = e^{-c_2 m \delta^2}$ for a sufficiently small constant c_2 yields that this bound holds with probability $1 - e^{-c_2 m \delta^2}$.

The only remaining step is to relate $\log M$ to the Gaussian width of the set. By the Sudakov minoration [85] and recalling that $\epsilon = 1/16$, there is a universal constant $c > 0$ such that

$$\sqrt{\log M} \leq c \mathbb{W}(A\mathcal{K}) \stackrel{(i)}{\leq} c \sqrt{\text{rank}(A)},$$

where the final inequality (i) follows from our previous calculation (2.11) in the proof of Corollary 2.

2.4.2 Reduction to finite maximum

The preceding argument suggests a general scheme for obtaining sharper results, namely by reducing to finite maxima. In this section, we provide a more general form of this scheme. It applies to random variables of the form

$$Z(\mathcal{Y}) = \sup_{y \in \mathcal{Y}} |y^T (\frac{A^T S^T S A}{m} - I)y|, \text{ where } \mathcal{Y} \subseteq \mathbb{R}^d. \quad (2.65)$$

For any set \mathcal{Y} , we define the first and second set differences as

$$\begin{aligned} \partial[\mathcal{Y}] &:= \mathcal{Y} - \mathcal{Y} = \{y - y' \mid y, y' \in \mathcal{Y}\}, \text{ and} \\ \partial^2[\mathcal{Y}] &:= \partial[\partial[\mathcal{Y}]]. \end{aligned}$$

Note that $\mathcal{Y} \subseteq \partial[\mathcal{Y}]$ whenever $0 \in \mathcal{Y}$. Let $\Pi(\mathcal{Y})$ denote the projection of \mathcal{Y} onto the Euclidean sphere \mathcal{S}^{d-1} .

With this notation, the following lemma shows how to reduce bounding $Z(\mathcal{Y})$ to taking a finite maximum over a cover of a related set.

Lemma 9. *Consider a pair of sets \mathcal{Y}_0 and \mathcal{Y}_1 such that $0 \in \mathcal{Y}_0$, the set \mathcal{Y}_1 is convex, and for some constant $\alpha \geq 1$, we have*

$$(a) \mathcal{Y}_1 \subseteq \text{clconv}(\mathcal{Y}_0), \quad (2.66)$$

$$(b) \partial^2[\mathcal{Y}_0] \subseteq \alpha \mathcal{Y}_1, \quad \text{and} \quad (2.67)$$

$$(c) \Pi(\partial^2[\mathcal{Y}_0]) \subseteq \alpha \mathcal{Y}_1. \quad (2.68)$$

Let $\{z^1, \dots, z^M\}$ be an ϵ -covering of the set $\partial[\mathcal{Y}_0]$ in Euclidean norm for some $\epsilon \in (0, \frac{1}{27\alpha^2}]$. Then for any symmetric matrix Q , we have

$$\sup_{z \in \mathcal{Y}_1} |z^T Q z| \leq 3 \max_{j=1, \dots, M} |(z^j)^T Q z^j|. \quad (2.69)$$

See Section 2.6.5 for the proof of this lemma. In the following subsections, we demonstrate how this auxiliary result can be used to obtain sharper results for various special cases.

2.4.3 Sharpening ℓ_1 -based bounds

The sharpened bounds in Corollary 3 are based on the following lemma. It applies to the tangent cone \mathcal{K} of the ℓ_1 -norm at a vector x^* with ℓ_0 -norm equal to k , as defined in equation (2.17).

Lemma 10. *For any $\delta \in (0, 1)$, a projection dimension lower bounded as $m \geq \frac{c_0}{\delta^2} \left(\frac{\gamma_k^+(A)+1}{\gamma_k^-(A)} \right)^2 k \log^5(d)$ guarantees that*

$$\sup_{v \in AK \cap \mathcal{S}^{n-1}} |v \left(\frac{S^T S}{m} - I \right) v| \leq \delta \quad (2.70)$$

with probability at least $1 - e^{-c_1 \frac{m\delta^2}{\log^4 n}}$.

Proof. Any $v \in AK \cap \mathcal{S}^{n-1}$ has the form $v = Au$ for some $u \in \mathcal{K}$. Any $u \in \mathcal{K}$ satisfies the inequality $\|u\|_1 \leq 2\sqrt{k}\|u\|_2$, so that by definition of the ℓ_1 -restricted eigenvalue (2.13), we are guaranteed that $\gamma_k^-(A)\|u\|_2^2 \leq \|Au\|_2^2 = 1$. Putting together the pieces, we conclude that

$$\begin{aligned} & \sup_{v \in AK \cap \mathcal{S}^{n-1}} |v^T (S^T S - I) v| \\ & \leq \frac{1}{\gamma_k^-(A)} \sup_{y \in \mathcal{Y}_1} \left| y \left(\frac{A^T S^T S A}{m} - A^T A \right) y \right| \\ & = \frac{1}{\gamma_k^-(A)} Z(\mathcal{Y}_1), \end{aligned}$$

where

$$\begin{aligned}\mathcal{Y}_1 &= \mathbb{B}_2(1) \cap \mathbb{B}_1(2\sqrt{k}) \\ &= \{\Delta \in \mathbb{R}^d \mid \|\Delta\|_1 \leq 2\sqrt{k}, \|\Delta\|_2 \leq 1\}.\end{aligned}$$

Now consider the set

$$\begin{aligned}\mathcal{Y}_0 &= \mathbb{B}_2(3) \cap \mathbb{B}_0(4k) \\ &= \{\Delta \in \mathbb{R}^d \mid \|\Delta\|_0 \leq 4k, \|\Delta\|_2 \leq 3\},\end{aligned}$$

We claim that the pair $(\mathcal{Y}_0, \mathcal{Y}_1)$ satisfy the conditions of Lemma 9 with $\alpha = 24$. The inclusion (2.66)(a) follows from Lemma 11 in the paper [87]; it is also a consequence of a more general result to be stated in the sequel as Lemma 14. Turning to the inclusion (2.66)(b), any vector $v \in \partial^2[\mathcal{Y}_0]$ can be written as $y - y' - (x - x')$ with $x, x', y, y' \in \mathcal{Y}_0$, whence $\|v\|_0 \leq 16k$ and $\|v\|_2 \leq 12$. Consequently, we have $\|v\|_1 \leq 4\sqrt{k}\|v\|_2$. Rescaling by $1/12$ shows that $\partial^2[\mathcal{Y}_0] \subseteq 24\mathcal{Y}_1$. A similar argument shows that $\Pi(\partial^2[\mathcal{Y}_0])$ satisfies the same containment.

Consequently, applying Lemma 9 with the symmetric matrix $R = \frac{A^T S^T S A}{m} - A^T A$ implies that

$$Z(\mathcal{Y}_1) \leq 3 \max_{j=1, \dots, M} |(z^j)^T R z^j|,$$

where $\{z^1, \dots, z^M\}$ is an $\frac{1}{27\alpha^2}$ covering of the set $\partial[\mathcal{Y}_0]$. By the JL-embedding result of Kraher and Ward [80], taking $m > \frac{c}{\delta^2} \log^4 d \log(M/\eta)$ samples suffices to ensure that, with probability at least $1 - \eta$, we have

$$\max_{j=1, \dots, M} |(z^j)^T R z^j| \leq \delta \max_{j=1, \dots, M} \|A z^j\|_2^2. \quad (2.71)$$

By the Sudakov minoration [85] and recalling that $\epsilon = \frac{1}{27\alpha^2}$ is a fixed quantity, we have

$$\sqrt{\log M} \leq c' \mathbb{W}(\mathcal{Y}_0) \leq c'' \sqrt{k \log d}, \quad (2.72)$$

where the final step follows by an easy calculation. Since $\|z^j\|_2 = 1$ for all $j \in [M]$, we are guaranteed that $\max_{j=1,\dots,M} \|Az^j\|_2^2 \leq \gamma_k^+(A)$, so that our earlier bound (2.71) implies that as long as $m > \frac{c}{\delta^2} k \log(d) \log^4 n$, we have

$$\sup_{v \in AK \cap \mathcal{S}^{n-1}} \left| v \left(\frac{S^T S}{m} - I \right) v \right| \leq 3\delta \frac{\gamma_k^+(A)}{\gamma_k^-(A)}$$

with high probability. Applying the rescaling $\delta \mapsto \frac{\gamma_k^-(A)}{\gamma_k^+(A)} \delta$ yields the claim. \square

Lemma 11. *Let $u \in \mathcal{S}^{d-1}$ be a fixed vector. Under the conditions of Lemma 10, we have*

$$\max_{v \in AK \cap \mathcal{S}^{n-1}} \left| u \left(\frac{S^T S}{m} - I \right) v \right| \leq \delta \tag{2.73}$$

with probability at least $1 - e^{-c_1 \frac{m\delta^2}{\log^4 n}}$.

Proof. Throughout this proof, we make use of the convenient shorthand $Q = \frac{S^T S}{m} - I$. Choose the sets \mathcal{Y}_0 and \mathcal{Y}_1 as in Lemma 10. Any $v \in AK \cap \mathcal{S}^{n-1}$ can be written as $v = Az$ for some $z \in \mathcal{K}$, and for which $\|z\|_2 \leq \frac{\|Az\|_2}{\sqrt{\gamma_k^-(A)}}$. Consequently, using the definitions of \mathcal{Y}_0 and \mathcal{Y}_1 , we have

$$\begin{aligned} & \max_{v \in AK \cap \mathcal{S}^{n-1}} |u^T Q v| \\ & \leq \frac{1}{\sqrt{\gamma_k^-(A)}} \max_{z \in \mathcal{Y}_1} |u^T Q A z| \end{aligned} \tag{2.74}$$

$$\begin{aligned} & \leq \frac{1}{\sqrt{\gamma_k^-(A)}} \max_{z \in \text{clconv}(\mathcal{Y}_0)} |u^T Q A z| \\ & = \frac{1}{\sqrt{\gamma_k^-(A)}} \max_{z \in \mathcal{Y}_0} |u^T Q A z|, \end{aligned} \tag{2.75}$$

where the last equality follows since the supremum is attained at an extreme point of \mathcal{Y}_0 .

For a parameter $\epsilon \in (0, 1)$ to be chosen, let $\{z^1, \dots, z^M\}$ be a ϵ -covering of the set

\mathcal{Y}_0 in the Euclidean norm. Using this covering, we can write

$$\begin{aligned}
& \sup_{z \in \mathcal{Y}_0} |u^T QAz| \\
& \leq \max_{j \in [M]} |u^T QAz^j| + \sup_{\Delta \in \partial[\mathcal{Y}_0], \|\Delta\|_2 \leq \epsilon} |u^T QA\Delta| \\
& = \max_{j \in [M]} |u^T QAz^j| + \epsilon \sup_{\Delta \in \Pi(\partial[\mathcal{Y}_0])} |u^T QA\Delta| \\
& \leq \max_{j \in [M]} |u^T QAz^j| + \epsilon\alpha \sup_{\Delta \in \mathcal{Y}_1} |u^T QA\Delta|.
\end{aligned}$$

Combined with equation (2.75), we conclude that

$$\begin{aligned}
& \sup_{z \in AK \cap \mathcal{S}^{n-1}} |u^T QAz| \\
& \leq \frac{1}{(1 - \epsilon\alpha)\sqrt{\gamma_k^-(A)}} \max_{j \in [M]} |u^T QAz^j|. \tag{2.76}
\end{aligned}$$

For each $j \in [M]$, we have the upper bound

$$\begin{aligned}
|u^T QAz^j| & \leq |(Az^j + u)^T Q(Az^j + u)| \\
& \quad + |(Az^j)^T QAz^j| + |u^T Qu|. \tag{2.77}
\end{aligned}$$

Based on this decomposition, we apply the JL-embedding property [80] to ROS matrices to the collection of $2M + 1$ points given by $\cup_{j \in [M]} \{Az^j, Az^j + u, \} \cup \{u\}$. Doing so ensures that, for any fixed $\delta \in (0, 1)$, we have

$$\max_{j \in [M]} |u^T QAz^j| \leq \delta (\|Az^j + u\|_2^2 + \|Az^j\|_2^2 + \|u\|_2^2).$$

with probability $1 - \eta$ as long as $m > \frac{c_0}{\delta^2} \log^4(n) \log\left(\frac{2M+1}{\eta}\right)$. Now observe that

$$\begin{aligned}
\|Az^j + u\|_2^2 + \|Az^j\|_2^2 + \|u\|_2^2 & \leq 3\|Az^j\|_2^2 + 3\|u\|_2^2 \\
& \leq 3(3\gamma_k^+(A) + 1),
\end{aligned}$$

where the final inequality follows by noting

$$\max_{z \in \mathcal{Y}_0} \|Az\|_2^2 \leq \max_{\substack{\|z\|_2 \leq 3 \\ \|z\|_1 \leq 6\sqrt{k}}} \|Az\|_2^2 \leq 3\gamma_k^+(A).$$

Consequently, we have $\max_{j \in [M]} |u^T Q A z^j| \leq 9\delta \left(\gamma_k^+(A) + 1 \right)$. Setting $\epsilon = \frac{1}{2\alpha}$, $\eta = e^{-c_1 \frac{m\delta^2}{\log^4(n)}}$ and with our earlier bound (2.76), we conclude that

$$\sup_{v \in \mathcal{AK} \cap \mathcal{S}^{n-1}} \left| u^T \left(\frac{S^T S}{m} - I \right) A v \right| \leq 18\delta \frac{(\gamma_k^+(A) + 1)}{\sqrt{\gamma_k^-(A)}} \quad (2.78)$$

$$\leq 18\delta \frac{(\gamma_k^+(A) + 1)}{\gamma_k^-(A)} \quad (2.79)$$

with probability $1 - e^{-c_1 \frac{m\delta^2}{\log^4 n}}$ where the last inequality follows from the assumption $\gamma_k^-(A) \leq 1$. Combined with the covering number estimate from equation (2.72), the claim follows. □

2.4.4 Sharpening nuclear norm bounds

We now show how the same approach may also be used to derive sharper bounds on the projection dimension for nuclear norm regularization. As shown in Lemma 1 in the paper [101], for the nuclear norm ball $\|X\|_* \leq R$, the tangent cone at any rank r matrix is contained within the set

$$\mathcal{K} := \left\{ \Delta \in \mathbb{R}^{d_1 \times d_2} \mid \|\Delta\|_* \leq 2\sqrt{r} \|\Delta\|_F \right\}, \quad (2.80)$$

and accordingly, our analysis focuses on the set $\mathcal{AK} \cap \mathcal{S}^{n-1}$, where $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^n$ is a general linear operator.

In analogy with the sparse restricted eigenvalues (2.13), we define the rank-constrained eigenvalues of the general operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^n$ as follows:

$$\gamma_r^-(\mathcal{A}) := \min_{\substack{\|Z\|_F=1 \\ \|Z\|_* \leq 2\sqrt{r}}} \|\mathcal{A}(Z)\|_2^2, \quad \text{and} \quad (2.81)$$

$$\gamma_r^+(\mathcal{A}) := \max_{\substack{\|Z\|_F=1 \\ \|Z\|_* \leq 2\sqrt{r}}} \|\mathcal{A}(Z)\|_2^2. \quad (2.82)$$

Lemma 12. *Suppose that the optimum X^* has rank at most r . For any $\delta \in (0, 1)$, an ROS sketch dimension lower bounded as $m \geq \frac{c_0}{\delta^2} \left(\frac{\gamma_r^+(\mathcal{A})}{\gamma_r^-(\mathcal{A})} \right)^2 r(d_1 + d_2) \log^4(d_1 d_2)$ ensures that*

$$\sup_{z \in \mathcal{AK} \cap \mathcal{S}^{n-1}} \left| z \left(\frac{S^T S}{m} - I \right) z \right| \leq \delta \quad (2.83)$$

with probability at least $1 - e^{-c_1 \frac{m\delta^2}{\log^4(d_1 d_2)}}$.

Proof. For an integer $r \geq 1$, consider the sets

$$\begin{aligned} \mathcal{Y}_1(r) &= \mathbb{B}_F(1) \cap \mathbb{B}_*(2\sqrt{r}) \\ &= \{\Delta \in \mathbb{R}^{d_1 \times d_2} \mid \|\Delta\|_* \leq 2\sqrt{r}, \|\Delta\|_F \leq 1\}, \end{aligned} \tag{2.84a}$$

$$\begin{aligned} \mathcal{Y}_0(r) &= \{\mathbb{B}_F(3) \cap \mathbb{B}_{\text{rank}}(4r)\} \\ &= \{\Delta \in \mathbb{R}^{n_1 \times n_2} \mid \|\Delta\|_0 \leq 4r, \|\Delta\|_F \leq 3\}. \end{aligned} \tag{2.84b}$$

In order to apply Lemma 9 with this pair, we must first show that the inclusions (2.66) hold. Inclusions (b) and (c) hold with $\alpha = 12$, as in the preceding proof of Lemma 10. Moreover, inclusion (a) also holds, but this is a non-trivial claim stated and proved separately as Lemma 14 in Section 2.6.6.

Consequently, an application of Lemma 9 with the symmetric matrix $Q = \frac{A^* S^T S A}{m} - \mathcal{A}^* \mathcal{A}$ in dimension $d_1 d_2$ guarantees that

$$Z(\mathcal{Y}_1(r)) \leq 3 \max_{j=1, \dots, M} |(z^j)^T Q z^j|,$$

where $\{z^1, \dots, z^M\}$ is a $\frac{1}{27\alpha^2}$ -covering of the set $\mathcal{Y}_0(r)$. By arguing as in the preceding proof of Lemma 10, the proof is then reduced to upper bounding the Gaussian complexity of $\mathcal{Y}_0(r)$. Letting $G \in \mathbb{R}^{d_1 \times d_2}$ denote a matrix of i.i.d. $N(0, 1)$ variates, we have

$$\begin{aligned} \mathbb{W}(\mathcal{Y}_0(r)) &= \mathbb{E} \left[\sup_{\Delta \in \mathcal{Y}_0(r)} \langle G, \Delta \rangle \right] \\ &\leq 6\sqrt{r} \mathbb{E}[\|G\|_2] \\ &\leq 6\sqrt{r} (\sqrt{d_1} + \sqrt{d_2}), \end{aligned}$$

where the final line follows from standard results [44] on the operator norms of Gaussian random matrices. \square

Lemma 13. *Let $u \in \mathcal{S}^{n-1}$ be a fixed vector. Under the assumptions of Lemma 12, we have*

$$\sup_{z \in \mathcal{AK} \cap \mathcal{S}^{n-1}} \left| u \left(\frac{S^T S}{m} - I \right) z \right| \leq \delta \quad (2.85)$$

with probability at least $1 - e^{-c_1 \frac{m\delta^2}{\log^4(d_1 d_2)}}$.

The proof parallels the proof of Lemma 11, and hence is omitted. Finally the sharpened bounds follow from the above lemmas and the deterministic bound (2.48).

2.5 Discussion

In this chapter, we have analyzed random projection methods for computing approximation solutions to convex programs. Our theory applies to any convex program based on a linear/quadratic objective functions, and involving arbitrary convex constraint set. Our main results provide lower bounds on the projection dimension that suffice to ensure that the optimal solution to sketched problem provides a δ -approximation to the original problem. In the sub-Gaussian case, this projection dimension can be chosen proportional to the square of the Gaussian width of the tangent cone, and in many cases, the same results hold (up to logarithmic factors) for sketches based on randomized orthogonal systems. This width depends both on the geometry of the constraint set, and the associated structure of the optimal solution to the original convex program. We also provided numerical simulations to illustrate the corollaries of our theorems in various concrete settings.

It is also worthwhile to make some comments about the practical uses of our guarantees. In some cases, our lower bounds on the required projection dimension m involve quantities of the unknown optimal solution x^* —for instance, its sparsity in an ℓ_1 -constrained problem, or its rank as a matrix in nuclear norm constrained problem. We note that it always suffices to choose m proportional to the dimension d , since we always have $\mathbb{W}^2(\mathcal{AK}) \leq \text{rank}(A) \leq d$ for any constraint set and optimal solution. However depending on the regularization parameters, i.e., radius of the constraint set, or some additional information, a practitioner can choose a smaller value of m depending on the application. In certain scenarios, it is known a priori that that optimal solution x^* has a bounded sparsity: for instance, this is the case in decoding sparse least-squares superposition codes [10, 72], in which the sparsity $\|x^*\|_0$ relates to the rate of the code. There is also a recent line of work in sparse learning literature aimed towards bounding the support of the optimal solution x^* before solving an ℓ_1 penalized convex optimization problem. Such bounds can be computed in $\mathcal{O}(nd)$

time, and have been shown to be accurate for real datasets [58]. In conjunction with such bounds, our theory provides practical choices for choosing m . Another possibility is based on a form of cross-validation: over a sequence of projection dimensions, one could solve a small subset of sketched problems, and choose a reliable dimension based on the (lack of) variability in the subset. (Once the projection dimension satisfies our bounds, our theory guarantees that the solutions from two independent sketches will be extremely close with very high probability.) research.

2.6 Proofs of technical results

2.6.1 Technical details for Corollary 3

In this section, we show how the second term in the bound (2.16) follows as a corollary of Theorem 2. From our previous calculations in the proof of Corollary 3(a), we have

$$\mathbb{R}(A\mathcal{K}) \leq \mathbb{E}_\varepsilon \left[\sup_{\substack{\|u\|_1 \leq 2\sqrt{k}\|u\|_2 \\ \|Au\|_2=1}} |\langle u, A^T \varepsilon \rangle| \right] \quad (2.86)$$

$$\leq \frac{2\sqrt{k}}{\sqrt{\gamma_k^-(A)}} \mathbb{E}[\|A^T \varepsilon\|_\infty] \quad (2.87)$$

$$\leq 6\sqrt{k \log d} \max_{j=1, \dots, d} \frac{\|a_j\|_2}{\sqrt{\gamma_k^-(A)}}. \quad (2.88)$$

Turning to the S -Gaussian width, we have

$$\begin{aligned} \mathbb{W}_S(A\mathcal{K}) &= \mathbb{E}_{g,S} \left[\sup_{\substack{\|u\|_1 \leq 2\sqrt{k}\|u\|_2 \\ \|Au\|_2=1}} \left| \left\langle g, \frac{SAu}{\sqrt{m}} \right\rangle \right| \right] \\ &\leq \frac{2\sqrt{k}}{\sqrt{\gamma_k^-(A)}} \mathbb{E}_{g,S} \left\| \frac{A^T S^T g}{\sqrt{m}} \right\|_\infty. \end{aligned}$$

Now the vector $S^T g / \sqrt{m}$ is zero-mean Gaussian with covariance $S^T S / m$. Consequently

$$\mathbb{E}_g \left\| \frac{A^T S^T g}{\sqrt{m}} \right\|_\infty \leq 4 \max_{j=1, \dots, d} \frac{\|Sa_j\|_2}{\sqrt{m}} \sqrt{\log d}.$$

Define the event $\mathcal{E} = \left\{ \frac{\|Sa_j\|_2}{\sqrt{m}} \leq 2\|a_j\|_2 \text{ for } j = 1, \dots, d \right\}$. By the JL embedding theorem of Krahmer and Ward [80], as long as $m > c_0 \log^5(n) \log(d)$, we can ensure

that $\mathbb{P}[\mathcal{E}^c] \leq \frac{1}{n}$. Since we always have $\|Sa_j\|_2/\sqrt{m} \leq \|a_j\|_2\sqrt{n}$, we can condition on \mathcal{E} and its complement, thereby obtaining that

$$\begin{aligned} & \mathbb{E}_{g,S} \left[\left\| \frac{A^T S^T g}{\sqrt{m}} \right\|_\infty \right] \\ & \leq 8 \max_{j=1,\dots,d} \|a_j\|_2 \sqrt{\log d} + 4 \mathbb{P}[\mathcal{E}^c] \sqrt{n} \max_{j=1,\dots,d} \|a_j\|_2 \sqrt{\log d} \\ & \leq 12 \max_{j=1,\dots,d} \|a_j\|_2 \sqrt{\log d}. \end{aligned}$$

Combined with our earlier calculation, we conclude that

$$\mathbb{W}_S(A\mathcal{K}) \leq \frac{\max_{j=1,\dots,d} \|a_j\|_2}{\sqrt{\gamma_k^-(A)}} \sqrt{k \log d}.$$

Substituting this upper bound, along with our earlier upper bound on the Rademacher width (2.86), yields the claim as a consequence of Theorem 2.

2.6.2 Technical lemmas for Proposition 2

In this section, we prove the two technical lemmas, namely Lemma 7 and 8, that underlie the proof of Proposition 2.

2.6.3 Proof of Lemma 7

Fixing some $D = \text{diag}(\nu) \in \mathcal{G}$, we first bound the deviations of Z'_0 above its expectation using Talagrand's theorem on empirical processes (e.g., see Massart [92] for one version with reasonable constants). Define the random vector $\tilde{s} = \sqrt{n}h$, where h is a randomly selected row, as well as the functions $g_y(\varepsilon, \tilde{s}) = \varepsilon \langle \tilde{s}, \text{diag}(\nu)y \rangle^2$, we have $\|g_z\|_\infty \leq \tau^2$ for all $y \in \mathcal{Y}$. Letting $\tilde{s} = \sqrt{n}h$ for a randomly chosen row h , we have

$$\text{var}(g_y) \leq \tau^2 \mathbb{E}[\langle \tilde{s}, \text{diag}(\nu)y \rangle^2] = \tau^2,$$

also uniformly over $y \in \mathcal{Y}$. Thus, for any $\nu \in \mathcal{G}$, Talagrand's theorem [92] implies that

$$\mathbb{P}_{\varepsilon,P} \left[Z'_0 \geq \mathbb{E}_{\varepsilon,P}[Z'_0] + \frac{\delta}{16} \right] \leq c_1 e^{-c_2 \frac{m\delta^2}{\tau^2}} \quad \text{for all } \delta \in [0, 1].$$

It remains to bound the expectation. By the Ledoux-Talagrand contraction for Rademacher processes [85], for any $\nu \in \mathcal{G}$, we have

$$\begin{aligned} \mathbb{E}_{\varepsilon, P}[Z'_0] &\stackrel{(i)}{\leq} 2\tau \mathbb{E}_{\varepsilon, P}\left[\sup_{y \in \mathcal{Y}} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle s_i, y \rangle \right|\right] \\ &\stackrel{(ii)}{\leq} 2\tau \left\{ \mathbb{W}_S(\mathcal{Y}) + \frac{\delta}{32\tau} \right\} \\ &= 2\tau \mathbb{W}_S(\mathcal{Y}) + \frac{\delta}{16}, \end{aligned}$$

where inequality (i) uses the inclusion $\nu \in \mathcal{G}_1$, and step (ii) relies on the inclusion $\nu \in \mathcal{G}_2$. Putting together the pieces yields the claim (2.62).

2.6.4 Proof of Lemma 8

It suffices to show that

$$\mathbb{P}[\mathcal{G}_1^c] \leq \frac{1}{(mn)^\kappa} \quad \text{and} \quad \mathbb{P}[\mathcal{G}_2^c] \leq c_1 e^{-c_2 m \delta^2}.$$

We begin by bounding $\mathbb{P}[\mathcal{G}_1^c]$. Recall $s_i^T = \sqrt{np_i^T} H \text{diag}(\nu)$, where $\nu \in \{-1, +1\}^n$ is a vector of i.i.d. Rademacher variables. Consequently, we have $\langle s_i, y \rangle = \sum_{j=1}^n (\sqrt{n} H_{ij}) \nu_j y_j$. Since $|\sqrt{n} H_{ij}| = 1$ for all (i, j) , the random variable $\langle s_i, y \rangle$ is equal in distribution to the random variable $\langle \nu, y \rangle$. Consequently, we have the equality in distribution

$$\sup_{y \in \mathcal{Y}} |\langle \sqrt{np_i^T} H \text{diag}(\nu), y \rangle| \stackrel{d}{=} \underbrace{\sup_{y \in \mathcal{Y}} |\langle \nu, y \rangle|}_{f(\nu)}.$$

Since this equality in distribution holds for each $i = 1, \dots, n$, the union bound guarantees that

$$\mathbb{P}[\mathcal{G}_1^c] \leq n \mathbb{P}[f(\nu) > \tau].$$

Accordingly, it suffices to obtain a tail bound on f . By inspection, the function f is convex in ν , and moreover $|f(\nu) - f(\nu')| \leq \|\nu - \nu'\|_2$, so that it is 1-Lipschitz. Therefore, by standard concentration results [84], we have

$$\mathbb{P}[f(\nu) \geq \mathbb{E}[f(\nu)] + t] \leq e^{-\frac{t^2}{2}}. \tag{2.89}$$

By definition, $\mathbb{E}[f(\nu)] = \mathbb{R}(\mathcal{Y})$, so that setting $t = \sqrt{2(1 + \kappa) \log(mn)}$ yields the bound tail bound $\mathbb{P}[\mathcal{G}_1^c] \leq \frac{1}{(mn)^\kappa}$, as claimed.

Next we control the probability of the event \mathcal{G}_2^c . The function g from equation (2.61) is clearly convex in the vector ν ; we now show that it is also Lipschitz with constant $1/\sqrt{m}$. Indeed, for any two vectors $\nu, \nu' \in \{-1, 1\}^d$, we have

$$|g(\nu) - g(\nu')| \leq \mathbb{E}_{\varepsilon, P} \left[\sup_{y \in \mathcal{Y}} \left| \left\langle \frac{1}{m} \sum_{i=1}^m \varepsilon_i \text{diag}(\nu - \nu') \sqrt{n} H^T p_i, y \right\rangle \right| \right] \leq \mathbb{E}_{\varepsilon, P} \|(\text{diag}(\nu - \nu'))\| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \sqrt{n} H$$

where the first inequality follows from triangle inequality and the definition (2.61) and the second inequality follows from Cauchy-Schwartz inequality since $\|y\|_2 \leq 1$. Introducing the shorthand $\Delta = \text{diag}(\nu - \nu')$ and $\tilde{s}_i = \sqrt{n} H^T p_i$, Jensen's inequality yields

$$\begin{aligned} |g(\nu) - g(\nu')|^2 &\leq \frac{1}{m^2} \mathbb{E}_{\varepsilon, P} \|\Delta \sum_{i=1}^m \varepsilon_i \tilde{s}_i\|_2^2 \\ &= \frac{1}{m^2} \text{trace} \left(\Delta \mathbb{E}_P \left[\sum_{i=1}^m \tilde{s}_i \tilde{s}_i^T \right] \Delta \right) \\ &= \frac{1}{m} \text{trace} \left(\Delta^2 \text{diag} \left(\mathbb{E}_P \left[\frac{1}{m} \sum_{i=1}^m \tilde{s}_i \tilde{s}_i^T \right] \right) \right). \end{aligned}$$

By construction, we have $|\tilde{s}_{ij}| = 1$ for all (i, j) , whence $\text{diag}(\mathbb{E}_P[\frac{1}{m} \sum_{i=1}^m \tilde{s}_i \tilde{s}_i^T]) = I_{n \times n}$. Since $\text{trace}(\Delta^2) = \|\nu - \nu'\|_2^2$, we have established that $|g(\nu) - g(\nu')|^2 \leq \frac{\|\nu - \nu'\|_2^2}{m}$, showing that g is a $1/\sqrt{m}$ -Lipschitz function. By standard concentration results [84], we conclude that

$$\mathbb{P}[\mathcal{G}_2^c] = \mathbb{P} \left[g(\nu) \geq \mathbb{E}[g(\nu)] + \frac{\delta}{32\tau} \right] \leq e^{-\frac{m\delta^2}{4096\tau^2}},$$

as claimed.

2.6.5 Proof of Lemma 9

By the inclusion (2.66)(a), we have $\sup_{z \in \mathcal{Y}_1} |z^T Q z| \leq \sup_{z \in \text{clconv}(\mathcal{Y}_0)} |z^T Q z|$. Any vector $v \in \text{conv}(\mathcal{Y}_0)$ can be written as a convex combination of the form $v = \sum_{i=1}^T \alpha_i z_i$, where the vectors $\{z_i\}_{i=1}^T$ belong to \mathcal{Y}_0 and the non-negative weights $\{\alpha_i\}_{i=1}^T$ sum to one, whence

$$\begin{aligned} |v^T Q v| &\leq \sum_{i=1}^T \sum_{j=1}^T \alpha_i \alpha_j |z_i^T Q z_j| \\ &\leq \frac{1}{2} \max_{i, j \in [T]} |(z_i + z_j)^T Q (z_i + z_j) - z_i^T Q z_i - z_j^T Q z_j| \\ &\leq \frac{3}{2} \sup_{z \in \partial[\mathcal{Y}_0]} |z^T Q z|. \end{aligned}$$

Since this upper bound applies to any vector $v \in \text{conv}(\mathcal{Y}_0)$, it also applies to any vector in the closure, whence

$$\sup_{z \in \mathcal{Y}_1} |z^T Q z| \leq \sup_{z \in \text{clconv}(\mathcal{Y}_0)} |z^T Q z| \quad (2.90)$$

$$\leq \frac{3}{2} \sup_{z \in \partial[\mathcal{Y}_0]} |z^T Q z|. \quad (2.91)$$

Now for some $\epsilon \in (0, 1]$ to be chosen, let $\{z^1, \dots, z^M\}$ be an ϵ -covering of the set $\partial[\mathcal{Y}_0]$ in Euclidean norm. Any vector $z \in \partial[\mathcal{Y}_0]$ can be written as $z = z^j + \Delta$ for some $j \in [M]$, and some vector with Euclidean norm at most ϵ . Moreover, the vector $\Delta \in \partial^2[\mathcal{Y}_0]$, whence

$$\begin{aligned} & \sup_{z \in \partial[\mathcal{Y}_0]} |z^T Q z| \\ & \leq \max_{j \in [M]} |(z^j)^T Q z^j| + 2 \sup_{\substack{\Delta \in \partial^2[\mathcal{Y}_0] \\ \|\Delta\|_2 \leq \epsilon}} \max_{j \in [M]} |\Delta^T Q z^j| \\ & \quad + \sup_{\substack{\Delta \in \partial^2[\mathcal{Y}_0] \\ \|\Delta\|_2 \leq \epsilon}} |\Delta^T Q \Delta|. \end{aligned} \quad (2.92)$$

Since $z^j \in \mathcal{Y}_0 \subseteq \partial^2[\mathcal{Y}_0]$, we have

$$\begin{aligned} & \sup_{z \in \partial[\mathcal{Y}_0]} |z^T Q z| \\ & \leq \max_{j \in [M]} |(z^j)^T Q z^j| + 2 \sup_{\substack{\Delta, \Delta' \in \partial^2[\mathcal{Y}_0] \\ \|\Delta\|_2 \leq \epsilon}} |\Delta^T Q \Delta'| \\ & \quad + \sup_{\substack{\Delta \in \partial^2[\mathcal{Y}_0] \\ \|\Delta\|_2 \leq \epsilon}} |\Delta^T Q \Delta| \\ & \leq \max_{j \in [M]} |(z^j)^T Q z^j| + 3 \sup_{\substack{\Delta, \Delta' \in \partial^2[\mathcal{Y}_0] \\ \|\Delta\|_2 \leq \epsilon}} |\Delta^T Q \Delta'| \\ & \leq \max_{j \in [M]} |(z^j)^T Q z^j| + 3\epsilon \sup_{\substack{\Delta \in \Pi(\partial^2[\mathcal{Y}_0]) \\ \Delta' \in \partial^2[\mathcal{Y}_0]}} |\Delta^T Q \Delta'| \\ & \leq \max_{j \in [M]} |(z^j)^T Q z^j| + 3\epsilon \sup_{\Delta, \Delta' \in \alpha \mathcal{Y}_1} |\Delta^T Q \Delta'|, \end{aligned}$$

where the final inequality makes use of the inclusions (2.66)(b) and (c). Finally, we observe that

$$\begin{aligned} \sup_{\Delta, \Delta' \in \alpha \mathcal{Y}_1} |\Delta^T Q \Delta'| &= \sup_{\Delta, \Delta' \in \alpha \mathcal{Y}_1} \frac{1}{2} |(\Delta + \Delta')^T Q (\Delta + \Delta')^T - \Delta^T Q \Delta - \Delta'^T Q \Delta'| \\ &\leq \frac{1}{2} \{4 + 1 + 1\} \sup_{\Delta \in \alpha \mathcal{Y}_1} |\Delta^T Q \Delta| \\ &= 3\alpha^2 \sup_{z \in \mathcal{Y}_1} |z^T Q z|, \end{aligned}$$

where we have used the fact that $\frac{\Delta+\Delta'}{2} \in \alpha\mathcal{Y}_1$, by convexity of the set $\alpha\mathcal{Y}_1$.

Putting together the pieces, we have shown that

$$\sup_{z \in \mathcal{Y}_1} |z^T Q z| \leq \frac{3}{2} \left\{ \max_{j \in [M]} |(z^j)^T Q z^j| + 9\epsilon\alpha^2 \sup_{\Delta \in \mathcal{Y}_1} |\Delta^T Q \Delta| \right\}.$$

Setting $\epsilon = \frac{1}{27\alpha^2}$ ensures that $9\epsilon\alpha^2 < 1/3$, and hence the claim (2.69) follows after some simple algebra.

2.6.6 A technical inclusion lemma

Recall the sets $\mathcal{Y}_1(r)$ and $\mathcal{Y}_0(r)$ previously defined in equations (2.84a) and (2.84b).

Lemma 14. *We have the inclusion*

$$\mathcal{Y}_1(r) \subseteq \text{clconv}(\mathcal{Y}_0(r)), \quad (2.93)$$

where clconv denotes the closed convex hull.

Proof. Define the support functions $\phi_0(X) = \sup_{\Delta \in \mathcal{Y}_0} \langle\langle X, \Delta \rangle\rangle$ and $\phi_1(X) = \sup_{\Delta \in \mathcal{Y}_1} \langle\langle X, \Delta \rangle\rangle$ where $\langle\langle X, \Delta \rangle\rangle := \text{trace}(X^T \Delta)$ stands for the standard inner product. It suffices to show that $\phi_1(X) \leq 3\phi_0(X)$ for each $X \in \mathcal{S}^{d \times d}$. The Frobenius norm, nuclear norm and rank are all invariant to unitary transformation, so we may take X to be diagonal without loss of generality. In this case, we may restrict the optimization to diagonal matrices Δ , and note that

$$\|\Delta\|_{\text{F}} = \sqrt{\sum_{j=1}^d \Delta_{jj}^2}, \quad \text{and} \quad \|\Delta\|_* = \sum_{j=1}^d |\Delta_{jj}|.$$

Let S be the indices of the $[r]$ diagonal elements that are largest in absolute value.

It is easy to see that

$$\phi_0(X) = \sqrt{\sum_{j \in S} X_{jj}^2}.$$

On the other hand, for any index $k \notin S$, we have $|X_{kk}| \leq |X_{jj}|$ for $j \in S$, and hence

$$\max_{k \notin S} |X_{kk}| \leq \frac{1}{\lfloor r \rfloor} \sum_{j \in S} |X_{jj}| \leq \frac{1}{\sqrt{\lfloor r \rfloor}} \sqrt{\sum_{j \in S} X_{jj}^2}$$

Using this fact, we can write

$$\begin{aligned} \phi_1(X) &\leq \sup_{\sum_{j \in S} \Delta_{jj}^2 \leq 1} \sum_{j \in S} \Delta_{jj} X_{jj} \\ &\quad + \sup_{\sum_{k \notin S} |\Delta_{kk}| \leq \sqrt{r}} \sum_{k \notin S} \Delta_{kk} X_{kk} \\ &= \sqrt{\sum_{j \in S} X_{jj}^2} + \sqrt{r} \max_{k \notin S} |X_{kk}| \\ &\leq \left(1 + \frac{\sqrt{r}}{\sqrt{\lfloor r \rfloor}}\right) \sqrt{\sum_{j \in S} X_{jj}^2} \\ &\leq 3\phi_0(X), \end{aligned}$$

as claimed. □

Chapter 3

Iterative random projections and information theoretical bounds

Randomized sketches are a well-established way of obtaining an approximate solutions to a variety of problems, and there is a long line of work on their uses (e.g., see the books and papers by [139, 26, 90, 55, 73], as well as references therein). In application to the least-squares problem we considered in the previous chapter,

$$x^{\text{LS}} := \arg \min_{x \in \mathcal{C}} f(x) \quad \text{where } f(x) := \frac{1}{2n} \|Ax - y\|_2^2. \quad (3.1)$$

sketching methods involves using a random matrix $S \in \mathbb{R}^{m \times n}$ to project the data matrix A and/or data vector y to a lower dimensional space ($m \ll n$), and then solving the approximated least-squares problem. In this chapter we explore alternative approximation properties of various sketches from a statistical perspective. There are many choices of random sketching matrices; see Section 3.1.1 for discussion of a few possibilities. Given some choice of random sketching matrix S , the most well-studied form of sketched least-squares is based on solving the problem

$$\tilde{x} := \arg \min_{x \in \mathcal{C}} \left\{ \frac{1}{2n} \|SAx - Sy\|_2^2 \right\}, \quad (3.2)$$

in which the data matrix-vector pair (A, y) are approximated by their sketched versions (SA, Sy) . Note that the sketched program is an m -dimensional least-squares problem, involving the new data matrix $SA \in \mathbb{R}^{m \times d}$. Thus, in the regime $n \gg d$, this approach can lead to substantial computational savings as long as the projection dimension m can be chosen substantially less than n . A number of authors

(e.g., [123, 26, 55, 90, 114]) have investigated the properties of this sketched solution (3.2), and accordingly, we refer to it as the *classical least-squares sketch*.

There are various ways in which the quality of the approximate solution \tilde{x} can be assessed. One standard way is in terms of the minimizing value of the quadratic cost function f defining the original problem (3.1), which we refer to as *cost approximation*. In terms of f -cost, the approximate solution \tilde{x} is said to be δ -optimal if

$$f(x^{\text{LS}}) \leq f(\tilde{x}) \leq (1 + \delta)^2 f(x^{\text{LS}}). \quad (3.3)$$

For example, in the case of unconstrained least-squares ($\mathcal{C} = \mathbb{R}^d$) with $n > d$, it is known that with Gaussian random sketches, a sketch size $m \gtrsim \frac{1}{\delta^2} d$ suffices to guarantee that \tilde{x} is δ -optimal with high probability (for instance, see the papers by [123] and [90], as well as references therein). Similar guarantees can be established for sketches based on sampling according to the statistical leverage scores [54, 52]. Sketching can also be applied to problems with constraints: [26] prove analogous results for the case of non-negative least-squares considering the sketch in equation (3.2), whereas our own past work [114] provides sufficient conditions for δ -accurate cost approximation of least-squares problems over arbitrary convex sets based also on the form in (3.2).

It should be noted, however, that other notions of “approximation goodness” are possible. In many applications, it is the least-squares minimizer x^{LS} itself—as opposed to the cost value $f(x^{\text{LS}})$ —that is of primary interest. In such settings, a more suitable measure of approximation quality would be the ℓ_2 -norm $\|\tilde{x} - x^{\text{LS}}\|_2$, or the prediction (semi)-norm

$$\|\tilde{x} - x^{\text{LS}}\|_A := \frac{1}{\sqrt{n}} \|A(\tilde{x} - x^{\text{LS}})\|_2. \quad (3.4)$$

We refer to these measures as *solution approximation*.

Now of course, a cost approximation bound (3.3) can be used to derive guarantees on the solution approximation error. However, it is natural to wonder whether or not, for a reasonable sketch size, the resulting guarantees are “good”. For instance, using arguments from [55], for the problem of unconstrained least-squares, it can be shown that the same conditions ensuring a δ -accurate cost approximation also ensure that

$$\|\tilde{x} - x^{\text{LS}}\|_A \leq \delta \sqrt{f(x^{\text{LS}})}. \quad (3.5)$$

Given lower bounds on the singular values of the data matrix A , this bound also yields control of the ℓ_2 -error.

In certain ways, the bound (3.5) is quite satisfactory: given our normalized definition (3.1) of the least-squares cost f , the quantity $f(x^{\text{LS}})$ remains an order one quantity as the sample size n grows, and the multiplicative factor δ can be reduced

by increasing the sketch dimension m . But how small should δ be chosen? In many applications of least-squares, each element of the response vector $y \in \mathbb{R}^n$ corresponds to an observation, and so as the sample size n increases, we expect that x^{LS} provides a more accurate approximation to some underlying population quantity, say $x^* \in \mathbb{R}^d$. As an illustrative example, in the special case of unconstrained least-squares, the accuracy of the least-squares solution x^{LS} as an estimate of x^* scales as $\|x^{\text{LS}} - x^*\|_A \asymp \frac{\sigma^2 d}{n}$. Consequently, in order for our sketched solution to have an accuracy of the same order as the least-square estimate, we must set $\delta^2 \asymp \frac{\sigma^2 d}{n}$. Combined with our earlier bound on the projection dimension, this calculation suggests that a projection dimension of the order

$$m \gtrsim \frac{d}{\delta^2} \gtrsim \frac{n}{\sigma^2}$$

is required. This scaling is undesirable in the regime $n \gg d$, where the whole point of sketching is to have the sketch dimension m much lower than n .

Now the alert reader will have observed that the preceding argument was only rough and heuristic. However, the first result of this chapter (Theorem 3) provides a rigorous confirmation of the conclusion: whenever $m \ll n$, the classical least-squares sketch (3.2) is sub-optimal as a method for solution approximation. Figure 3.1 provides an empirical demonstration of the poor behavior of the classical least-squares sketch for an unconstrained problem.

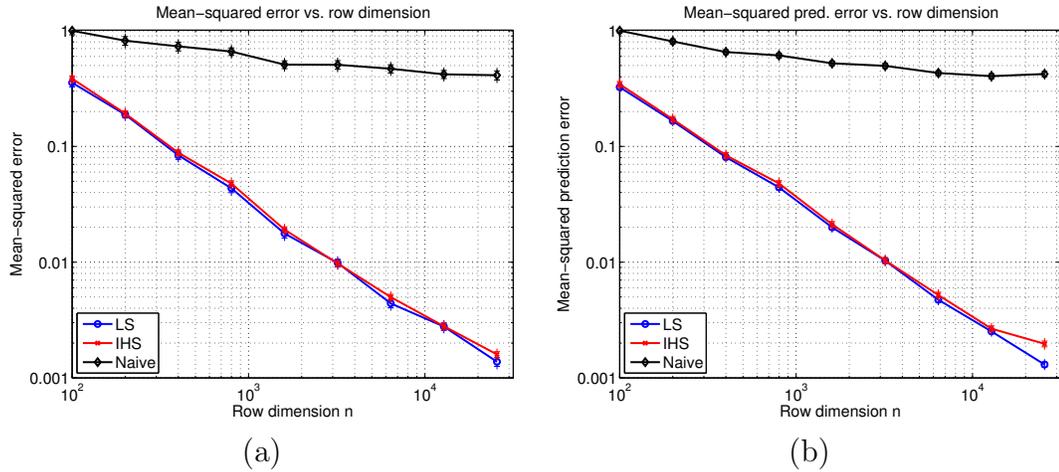


Figure 3.1: Plots of mean-squared error versus the row dimension $n \in \{100, 200, 400, \dots, 25600\}$ for unconstrained least-squares in dimension $d = 10$.

This sub-optimality holds not only for unconstrained least-squares but also more generally for a broad class of constrained problems. Actually, Theorem 3 is a more

general claim: *any estimator* based only on the pair (SA, Sy) —an infinite family of methods including the standard sketching algorithm as a particular case—is sub-optimal relative to the original least-squares estimator in the regime $m \ll n$. We are thus led to a natural question: can this sub-optimality be avoided by a different type of sketch that is nonetheless computationally efficient? Motivated by this question, our second main result (Theorem 4) is to propose an alternative method—known as the iterative Hessian sketch—and prove that it yields optimal approximations to the least-squares solution using a projection size that scales with the intrinsic dimension of the underlying problem, along with a logarithmic number of iterations. The main idea underlying iterative Hessian sketch is to obtain multiple sketches of the data (S^1A, \dots, S^NA) and iteratively refine the solution where N can be chosen logarithmic in n .

The remainder of this chapter is organized as follows. In Section 3.1, we begin by introducing some background on classes of random sketching matrices, before turning to the statement of our lower bound (Theorem 3) on the classical least-squares sketch (3.2). We then introduce the Hessian sketch, and show that an iterative version of it can be used to compute ε -accurate solution approximations using $\log(1/\varepsilon)$ -steps (Theorem 4). In Section 3.2, we illustrate the consequences of this general theorem for various specific classes of least-squares problems, and we conclude with a discussion in Section 3.3.

3.1 Main results and consequences

In this section, we begin with background on different classes of randomized sketches, including those based on random matrices with sub-Gaussian entries, as well as those based on randomized orthonormal systems and random sampling. In Section 3.1.2, we prove a general lower bound on the solution approximation accuracy of any method that attempts to approximate the least-squares problem based on observing only the pair (SA, Sy) . This negative result motivates the investigation of alternative sketching methods, and we begin this investigation by introducing the Hessian sketch in Section 3.1.3. It serves as the basic building block of the iterative Hessian sketch (IHS), which can be used to construct an iterative method that is optimal up to logarithmic factors.

3.1.1 Types of randomized sketches

In the following section, we present a lower bound that applies to all the three kinds of sketching matrices described in this thesis including Sub-Gaussian sketches,

ROS sketches and random row sampling. For sketches based on random row sampling, we assume that the weights are α -balanced, meaning that

$$\max_{j=1,\dots,n} p_j \leq \frac{\alpha}{n} \quad (3.6)$$

for some constant α independent of n .

3.1.2 Information-theoretical sub-optimality of the classical sketch

We begin by proving a lower bound on any estimator that is a function of the pair (SA, Sy) . In order to do so, we consider an ensemble of least-squares problems, namely those generated by a noisy observation model of the form

$$y = Ax^* + w, \quad \text{where } w \sim N(0, \sigma^2 I_n), \quad (3.7)$$

the data matrix $A \in \mathbb{R}^{n \times d}$ is fixed, and the unknown vector x^* belongs to some set \mathcal{C}_0 that is star-shaped around zero.¹ In this case, the constrained least-squares estimate x^{LS} from equation (3.1) corresponds to a constrained form of maximum-likelihood for estimating the unknown regression vector x^* . In Section 3.7, we provide a general upper bound on the error $\mathbb{E}[\|x^{\text{LS}} - x^*\|_A^2]$ in the least-squares solution as an estimate of x^* . This result provides a baseline against which to measure the performance of a sketching method: in particular, our goal is to characterize the minimal projection dimension m required in order to return an estimate \tilde{x} with an error guarantee $\|\tilde{x} - x^{\text{LS}}\|_A \approx \|x^{\text{LS}} - x^*\|_A$. The result to follow shows that unless $m \geq n$, then *any method* based on observing *only* the pair (SA, Sy) necessarily has a substantially larger error than the least-squares estimate. In particular, our result applies to an arbitrary measurable function $(SA, Sy) \mapsto x^\dagger$, which we refer to as an *estimator*.

More precisely, our lower bound applies to any random matrix $S \in \mathbb{R}^{m \times n}$ for which

$$\|\mathbb{E}[S^T(SS^T)^{-1}S]\|_2 \leq \eta \frac{m}{n}, \quad (3.8)$$

where η is a constant independent of n and m , and $\|A\|_2$ denotes the ℓ_2 -operator norm (maximum eigenvalue for a symmetric matrix). In Section 3.4.1, we show that these conditions hold for various standard choices, including most of those discussed in the previous section. Letting $\mathbb{B}_A(1)$ denote the unit ball defined by the semi-norm

¹Explicitly, this star-shaped condition means that for any $x \in \mathcal{C}_0$ and scalar $t \in [0, 1]$, the point tx also belongs to \mathcal{C}_0 .

$\|\cdot\|_A$, our lower bound also involves the complexity of the set $\mathcal{C}_0 \cap \mathbb{B}_A(1)$, which we measure in terms of its metric entropy. In particular, for a given tolerance $\delta > 0$, the δ -packing number M_δ of the set $\mathcal{C}_0 \cap \mathbb{B}_A(1)$ with respect to $\|\cdot\|_A$ is the largest number of vectors $\{x^j\}_{j=1}^M \subset \mathcal{C}_0 \cap \mathbb{B}_A(1)$ such that $\|x^j - x^k\|_A > \delta$ for all distinct pairs $j \neq k$.

With this set-up, we have the following result:

Theorem 3 (Sub-optimality). *For any random sketching matrix $S \in \mathbb{R}^{m \times n}$ satisfying condition (3.8), any estimator $(SA, Sy) \mapsto x^\dagger$ has MSE lower bounded as*

$$\sup_{x^* \in \mathcal{C}_0} \mathbb{E}_{S,w} [\|x^\dagger - x^*\|_A^2] \geq \frac{\sigma^2}{128\eta} \frac{\log(\frac{1}{2}M_{1/2})}{\min\{m, n\}} \quad (3.9)$$

where $M_{1/2}$ is the 1/2-packing number of $\mathcal{C}_0 \cap \mathbb{B}_A(1)$ in the semi-norm $\|\cdot\|_A$.

The proof, given in Section 3.4, is based on a reduction from statistical minimax theory combined with information-theoretic bounds. The lower bound is best understood by considering some concrete examples:

Example 1 (Sub-optimality for ordinary least-squares). We begin with the simplest case—namely, in which $\mathcal{C} = \mathbb{R}^d$. With this choice and for any data matrix A with $\text{rank}(A) = d$, it is straightforward to show that the least-squares solution x^{LS} has its prediction mean-squared error at most

$$\mathbb{E}[\|x^{\text{LS}} - x^*\|_A^2] \lesssim \frac{\sigma^2 d}{n}. \quad (3.10a)$$

On the other hand, with the choice $\mathcal{C}_0 = \mathbb{B}_2(1)$, we can construct a 1/2-packing with $M = 2^d$ elements, so that Theorem 3 implies that any estimator x^\dagger based on (SA, Sy) has its prediction MSE lower bounded as

$$\mathbb{E}_{S,w} [\|\hat{x} - x^*\|_A^2] \gtrsim \frac{\sigma^2 d}{\min\{m, n\}}. \quad (3.10b)$$

Consequently, the sketch dimension m must grow proportionally to n in order for the sketched solution to have a mean-squared error comparable to the original

least-squares estimate. This is highly undesirable for least-squares problems in which $n \gg d$, since it should be possible to sketch down to a dimension proportional to $\text{rank}(A) = d$. Thus, Theorem 3 this reveals a surprising gap between the classical least-squares sketch (3.2) and the accuracy of the original least-squares estimate.

In contrast, the sketching method we describe now, known as iterative Hessian sketching (IHS), matches the optimal mean-squared error using a sketch of size $d + \log(n)$ in each round, and a total of $\log(n)$ rounds; see Corollary 9 for a precise statement. The red curves in Figure 3.1 show that the mean-squared errors ($\|\hat{x} - x^*\|_2^2$ in panel (a), and $\|\hat{x} - x^*\|_A^2$ in panel (b)) of the IHS method using this sketch dimension closely track the associated errors of the full least-squares solution (blue curves). Consistent with our previous discussion, both curves drop off at the n^{-1} rate.

Since the IHS method with $\log(n)$ rounds uses a total of $T = \log(n)\{d + \log(n)\}$ sketches, a fair comparison is to implement the classical method with T sketches in total. The black curves show the MSE of the resulting sketch: as predicted by our theory, these curves are relatively flat as a function of sample size n . Indeed, in this particular case, the lower bound (3.9)

$$\mathbb{E}_{S,w}[\|\tilde{x} - x^*\|_A^2] \gtrsim \frac{\sigma^2 d}{m} \gtrsim \frac{\sigma^2}{\log^2(n)},$$

showing we can expect (at best) an inverse logarithmic drop-off. \diamond

This sub-optimality can be extended to other forms of constrained least-squares estimates as well, such as those involving sparsity constraints.

Example 2 (Sub-optimality for sparse linear models). We now consider the sparse variant of the linear regression problem, which involves the ℓ_0 -“ball”

$$\mathbb{B}_0(k) := \left\{x \in \mathbb{R}^d \mid \sum_{j=1}^d \mathbb{I}[x_j \neq 0] \leq k\right\},$$

corresponding to the set of all vectors with at most k non-zero entries. Fixing some radius $R \geq \sqrt{k}$, consider a vector $x^* \in \mathcal{C}_0 := \mathbb{B}_0(k) \cap \{\|x\|_1 = R\}$, and suppose that we make noisy observations of the form $y = Ax^* + w$.

Given this set-up, one way in which to estimate x^* is by computing the least-squares estimate x^{LS} constrained² to the ℓ_1 -ball $\mathcal{C} = \{x \in \mathbb{R}^n \mid \|x\|_1 \leq R\}$. This estimator is a form of the Lasso [134]: as shown in Section 3.7.2, when the design matrix A satisfies the restricted isometry property (see [34] for a definition), then it has MSE at most

$$\mathbb{E}[\|x^{\text{LS}} - x^*\|_A^2] \lesssim \frac{\sigma^2 k \log\left(\frac{ed}{k}\right)}{n}. \quad (3.11a)$$

On the other hand, the $\frac{1}{2}$ -packing number M of the set \mathcal{C}_0 can be lower bounded as $\log M \gtrsim k \log\left(\frac{ed}{k}\right)$; see Section 3.7.2 for the details of this calculation. Consequently, in application to this particular problem, Theorem 3 implies that any estimator x^\dagger based on the pair (SA, Sy) has mean-squared error lower bounded as

$$\mathbb{E}_{w,S}[\|x^\dagger - x^*\|_A^2] \gtrsim \frac{\sigma^2 k \log\left(\frac{ed}{k}\right)}{\min\{m, n\}}. \quad (3.11b)$$

Again, we see that the projection dimension m must be of the order of n in order to match the mean-squared error of the constrained least-squares estimate x^{LS} up to constant factors. By contrast, in this special case, the sketching method we describe in this section matches the error $\|x^{\text{LS}} - x^*\|_2$ using a sketch dimension that scales only as $k \log\left(\frac{ed}{k}\right) + \log(n)$; see Corollary 10 for the details of a more general result. \diamond

Example 3 (Sub-optimality for low-rank matrix estimation). In the problem of multivariate regression, the goal is to estimate a matrix $X^* \in \mathbb{R}^{d_1 \times d_2}$ model based on

²This set-up is slightly unrealistic, since the estimator is assumed to know the radius $R = \|x^*\|_1$. In practice, one solves the least-squares problem with a Lagrangian constraint, but the underlying arguments are basically the same.

observations of the form

$$Y = AX^* + W, \quad (3.12)$$

where $Y \in \mathbb{R}^{n \times d_1}$ is a matrix of observed responses, $A \in \mathbb{R}^{n \times d_1}$ is a data matrix, and $W \in \mathbb{R}^{n \times d_2}$ is a matrix of noise variables. One interpretation of this model is as a collection of d_2 regression problems, each involving a d_1 -dimensional regression vector, namely a particular column of X^* . In many applications, among them reduced rank regression, multi-task learning and recommender systems (e.g., [130, 154, 101, 31]), it is reasonable to model the matrix X^* as having a low-rank. Note a rank constraint on matrix X be written as an ℓ_0 -“norm” constraint on its singular values: in particular, we have

$$\text{rank}(X) \leq r \quad \text{if and only if} \quad \sum_{j=1}^{\min\{d_1, d_2\}} \mathbb{I}[\gamma_j(X) > 0] \leq r,$$

where $\gamma_j(X)$ denotes the j^{th} singular value of X . This observation motivates a standard relaxation of the rank constraint using the nuclear norm $\|X\|_* := \sum_{j=1}^{\min\{d_1, d_2\}} \gamma_j(X)$.

Accordingly, let us consider the constrained least-squares problem

$$X^{\text{LS}} = \arg \min_{X \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2} \|Y - AX\|_{\text{F}}^2 \right\} \quad \text{such that } \|X\|_* \leq R, \quad (3.13)$$

where $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm on matrices, or equivalently the Euclidean norm on its vectorized version. Let \mathcal{C}_0 denote the set of matrices with rank $r < \frac{1}{2} \min\{d_1, d_2\}$, and Frobenius norm at most one. In this case, we show in Section 3.7 that the constrained least-squares solution X^{LS} satisfies the bound

$$\mathbb{E} \left[\|X^{\text{LS}} - X^*\|_A^2 \right] \lesssim \frac{\sigma^2 r (d_1 + d_2)}{n}. \quad (3.14a)$$

On the other hand, the $\frac{1}{2}$ -packing number of the set \mathcal{C}_0 is lower bounded as $\log M \gtrsim r(d_1 + d_2)$, so that Theorem 3 implies that any estimator X^\dagger based on the

pair (SA, SY) has MSE lower bounded as

$$\mathbb{E}_{w,S} [\|X^\dagger - X^*\|_A^2] \gtrsim \frac{\sigma^2 r(d_1 + d_2)}{\min\{m, n\}}. \quad (3.14b)$$

As with the previous examples, we see the sub-optimality of the sketched approach in the regime $m < n$. In contrast, for this class of problems, our sketching method matches the error $\|X^{\text{LS}} - X^*\|_A$ using a sketch dimension that scales only as $\{r(d_1 + d_2) + \log(n)\} \log(n)$. See Corollary 11 for further details.

◇

3.1.3 Introducing the Hessian sketch

As will be revealed during the proof of Theorem 3, the sub-optimality is in part due to sketching the response vector—i.e., observing Sy instead of y . It is thus natural to consider instead methods that sketch *only* the data matrix A , as opposed to both the data matrix and data vector y . In abstract terms, such methods are based on observing the pair $(SA, A^T y) \in \mathbb{R}^{m \times d} \times \mathbb{R}^d$. One such approach is what we refer to as the *Hessian sketch*—namely, the sketched least-squares problem

$$\hat{x} := \arg \min_{x \in \mathcal{C}} \left\{ \underbrace{\frac{1}{2} \|SAx\|_2^2 - \langle A^T y, x \rangle}_{g_S(x)} \right\}. \quad (3.15)$$

As with the classical least-squares sketch (3.2), the quadratic form is defined by the matrix $SA \in \mathbb{R}^{m \times d}$, which leads to computational savings. Although the Hessian sketch on its own does not provide an optimal approximation to the least-squares solution, it serves as the building block for an iterative method that can obtain an ε -accurate solution approximation in $\log(1/\varepsilon)$ iterations.

In controlling the error with respect to the least-squares solution x^{LS} the set of possible descent directions $\{x - x^{\text{LS}} \mid x \in \mathcal{C}\}$ plays an important role. In particular, we now define the *transformed tangent cone*

$$\mathcal{K}_A^{\text{LS}} = \{v \in \mathbb{R}^d \mid v = tA(x - x^{\text{LS}}) \text{ for some } t \geq 0 \text{ and } x \in \mathcal{C}\}. \quad (3.16)$$

Note that the error vector $\hat{v} := A(\hat{x} - x^{\text{LS}})$ of interest belongs to this cone. Our

approximation bound is a function of the quantities

$$Z_1(A\mathcal{K})(S) := \inf_{v \in \mathcal{K}_A^{\text{LS}} \cap \mathcal{S}^{n-1}} \frac{1}{m} \|Sv\|_2^2 \quad \text{and} \quad (3.17a)$$

$$Z_2(A\mathcal{K})(S) := \sup_{v \in \mathcal{K}_A^{\text{LS}} \cap \mathcal{S}^{n-1}} \left| \langle u, \left(\frac{S^T S}{m} - I_n \right) v \rangle \right|, \quad (3.17b)$$

where u is a fixed unit-norm vector. These variables played an important role in our previous analysis [114] of the classical sketch (3.2). The following bound applies in a deterministic fashion to any sketching matrix.

Proposition 3 (Bounds on Hessian sketch). *For any convex set \mathcal{C} and any sketching matrix $S \in \mathbb{R}^{m \times n}$, the Hessian sketch solution \hat{x} satisfies the bound*

$$\|\hat{x} - x^{\text{LS}}\|_A \leq \frac{Z_2(A\mathcal{K})}{Z_1(A\mathcal{K})} \|x^{\text{LS}}\|_A. \quad (3.18)$$

For random sketching matrices, Proposition 3 can be combined with probabilistic analysis to obtain high probability error bounds. For a given tolerance parameter $\rho \in (0, \frac{1}{2}]$, consider the “good event”

$$\mathcal{E}(\rho) := \left\{ Z_1(A\mathcal{K}) \geq 1 - \rho, \text{ and } Z_2(A\mathcal{K}) \leq \frac{\rho}{2} \right\}. \quad (3.19a)$$

Conditioned on this event, Proposition 3 implies that

$$\|\hat{x} - x^{\text{LS}}\|_A \leq \frac{\rho}{2(1-\rho)} \|x^{\text{LS}}\|_A \leq \rho \|x^{\text{LS}}\|_A, \quad (3.19b)$$

where the final inequality holds for all $\rho \in (0, 1/2]$.

Thus, for a given family of random sketch matrices, we need to choose the projection dimension m so as to ensure the event $\mathcal{E}\rho$ holds for some ρ . For future reference, let us state some known results for the cases of sub-Gaussian and ROS sketching matrices. We use (c_0, c_1, c_2) to refer to numerical constants, and we let $D = \dim(\mathcal{C})$ denote the dimension of the space \mathcal{C} . In particular, we have $D = d$ for vector-valued estimation, and $D = d_1 d_2$ for matrix problems.

Our bounds involve the “size” of the cone $\mathcal{K}_A^{\text{LS}}$ previously defined (3.16), as measured in terms of its *Gaussian width*

$$\mathcal{W}(\mathcal{K}_A^{\text{LS}}) := \mathbb{E}_g \left[\sup_{v \in \mathcal{K}_A^{\text{LS}} \cap \mathbb{B}_2(1)} |\langle g, v \rangle| \right], \quad (3.20)$$

where $g \sim N(0, I_n)$ is a standard Gaussian vector. With this notation, we have the following:

Lemma 15 (Sufficient conditions on sketch dimension [114]).

(a) For sub-Gaussian sketch matrices, given a sketch size $m > \frac{c_0}{\rho^2} \mathcal{W}^2(\mathcal{K}_A^{\text{LS}})$, we have

$$\mathbb{P}[\mathcal{E}(\rho)] \geq 1 - c_1 e^{-c_2 m \delta^2}. \quad (3.21a)$$

(b) For randomized orthogonal system (ROS) sketches (sampled with replacement) over the class of self-bounding cones, given a sketch size $m > \frac{c_0 \log^4(D)}{\rho^2} \mathcal{W}^2(\mathcal{K}_A^{\text{LS}})$, we have

$$\mathbb{P}[\mathcal{E}(\rho)] \geq 1 - c_1 e^{-c_2 \frac{m \rho^2}{\log^4(D)}}. \quad (3.21b)$$

The class of self-bounding cones is described more precisely in Lemma 8 of [114]. It includes among other special cases the cones generated by unconstrained least-squares (Example 1), ℓ_1 -constrained least squares (Example 2), and least squares with nuclear norm constraints (Example 3). For these cones, given a sketch size $m > \frac{c_0 \log^4(D)}{\rho^2} \mathcal{W}^2(\mathcal{K}_A^{\text{LS}})$, the Hessian sketch applied with ROS matrices is guaranteed to return an estimate \hat{x} such that

$$\|\hat{x} - x^{\text{LS}}\|_A \leq \rho \|x^{\text{LS}}\|_A \quad (3.22)$$

with high probability. More recent work by [25] has established sharp bounds for various forms of sparse Johnson-Lindenstrauss transforms [73]. As a corollary of their results, a form of the guarantee (3.22) also holds for such random projections.

Returning to the main thread, the bound (3.22) is an analogue of our earlier bound (3.5) for the classical sketch with $\sqrt{f(x^{\text{LS}})}$ replaced by $\|x^{\text{LS}}\|_A$. For this reason, we see that the Hessian sketch alone suffers from the same deficiency as the classical sketch: namely, it will require a sketch size $m \asymp n$ in order to mimic the $\mathcal{O}(n^{-1})$ accuracy of the least-squares solution.

3.1.4 Iterative Hessian sketch

Despite the deficiency of the Hessian sketch itself, it serves as the building block for a novel scheme—known as the iterative Hessian sketch—that can be used to match the accuracy of the least-squares solution using a reasonable sketch dimension. Let

begin by describing the underlying intuition. As summarized by the bound (3.19b), conditioned on the good event $\mathcal{E}(\rho)$, the Hessian sketch returns an estimate with error within a ρ -factor of $\|x^{\text{LS}}\|_A$, where x^{LS} is the solution to the original unsketched problem. As show by Lemma 15, as long as the projection dimension m is sufficiently large, we can ensure that $\mathcal{E}(\rho)$ holds for some $\rho \in (0, 1/2)$ with high probability. Accordingly, given the current iterate x^t , *suppose that we can construct a new least-squares problem* for which the optimal solution is $x^{\text{LS}} - x^t$. Applying the Hessian sketch to this problem will then produce a new iterate x^{t+1} whose distance to x^{LS} has been reduced by a factor of ρ . Repeating this procedure N times will reduce the initial approximation error by a factor ρ^N .

With this intuition in place, we now turn a precise formulation of the *iterative Hessian sketch*. Consider the optimization problem

$$\hat{u} = \arg \min_{u \in \mathcal{C} - x^t} \left\{ \frac{1}{2} \|Au\|_2^2 - \langle A^T(y - Ax^t), u \rangle \right\}, \quad (3.23)$$

where x^t is the iterate at step t . By construction, the optimum to this problem is given by $\hat{u} = x^{\text{LS}} - x^t$. We then apply to Hessian sketch to this optimization problem (3.23) in order to obtain an approximation $x^{t+1} = x^t + \hat{u}$ to the original least-squares solution x^{LS} that is more accurate than x^t by a factor $\rho \in (0, 1/2)$. Recursing this procedure yields a sequence of iterates whose error decays geometrically in ρ .

Formally, the iterative Hessian sketch algorithm takes the following form:

Iterative Hessian sketch (IHS): Given an iteration number $N \geq 1$:

- (1) Initialize at $x^0 = 0$.
- (2) For iterations $t = 0, 1, 2, \dots, N - 1$, generate an independent sketch matrix $S^{t+1} \in \mathbb{R}^{m \times n}$, and perform the update

$$x^{t+1} = \arg \min_{x \in \mathcal{C}} \left\{ \frac{1}{2m} \|S^{t+1}A(x - x^t)\|_2^2 - \langle A^T(y - Ax^t), x \rangle \right\}. \quad (3.24)$$

- (3) Return the estimate $\hat{x} = x^N$.

The following theorem summarizes the key properties of this algorithm. It involves the sequence $\{Z_1(A\mathcal{K})(S^t), Z_2(A\mathcal{K})(S^t)\}_{t=1}^N$, where the quantities $Z_1(A\mathcal{K})$ and $Z_2(A\mathcal{K})$

were previously defined in equations (3.17a) and (3.17b). In addition, as a generalization of the event (3.19a), we define the sequence of “good” events

$$\mathcal{E}^t(\rho) := \left\{ Z_1(A\mathcal{K})(S^t) \geq 1 - \rho, \text{ and } Z_2(A\mathcal{K})(S^t) \leq \frac{\rho}{2} \right\} \quad \text{for } t = 1, \dots, N. \quad (3.25)$$

With this notation, we have the following guarantee:

Theorem 4 (Guarantees for iterative Hessian sketch). *The final solution $\hat{x} = x^N$ satisfies the bound*

$$\|\hat{x} - x^{LS}\|_A \leq \left\{ \prod_{t=1}^N \frac{Z_2(A\mathcal{K})(S^t)}{Z_1(A\mathcal{K})(S^t)} \right\} \|x^{LS}\|_A. \quad (3.26a)$$

Consequently, conditioned on the event $\cap_{t=1}^N \mathcal{E}^t(\rho)$ for some $\rho \in (0, 1/2)$, we have

$$\|\hat{x} - x^{LS}\|_A \leq \rho^N \|x^{LS}\|_A. \quad (3.26b)$$

Note that for any $\rho \in (0, 1/2)$, then event $\mathcal{E}^t(\rho)$ implies that $\frac{Z_2(A\mathcal{K})(S^t)}{Z_1(A\mathcal{K})(S^t)} \leq \rho$, so that the bound (3.26b) is an immediate consequence of the product bound (3.26a).

Remark. For unconstrained problems, $S_t = S$ can be generated once and fixed for all iterations and the guarantees of the theorem still hold. This follows from a simple modification of the proof of Theorem 4.

Lemma 15 can be combined with the union bound in order to ensure that the compound event $\cap_{t=1}^N \mathcal{E}^t(\rho)$ holds with high probability over a sequence of N iterates, as long as the sketch size is lower bounded as $m \geq \frac{c_0}{\rho^2} \mathbb{W}^2(\mathcal{K}_A^{LS}) \log^4(D) + \log N$. Based on the bound (3.26b), we then expect to observe geometric convergence of the iterates.

In order to test this prediction, we implemented the IHS algorithm using Gaussian sketch matrices, and applied it to an unconstrained least-squares problem based on a data matrix with dimensions $(d, n) = (200, 6000)$ and noise variance $\sigma^2 = 1$. As shown in Section 3.7.2, the Gaussian width of \mathcal{K}_A^{LS} is proportional to d , so that Lemma 15 shows that it suffices to choose a projection dimension $m \gtrsim \gamma d$ for a sufficiently large constant γ . Panel (a) of Figure 3.2 illustrates the resulting convergence rate of the IHS algorithm, measured in terms of the error $\|x^t - x^{LS}\|_A$, for different values $\gamma \in \{4, 6, 8\}$. As predicted by Theorem 4, the convergence rate is geometric (linear on the log scale shown), with the rate increasing as the parameter γ is increased.

Assuming that the sketch dimension has been chosen to ensure geometric convergence, Theorem 4 allows us to specify, for a given target accuracy $\varepsilon \in (0, 1)$, the number of iterations required.

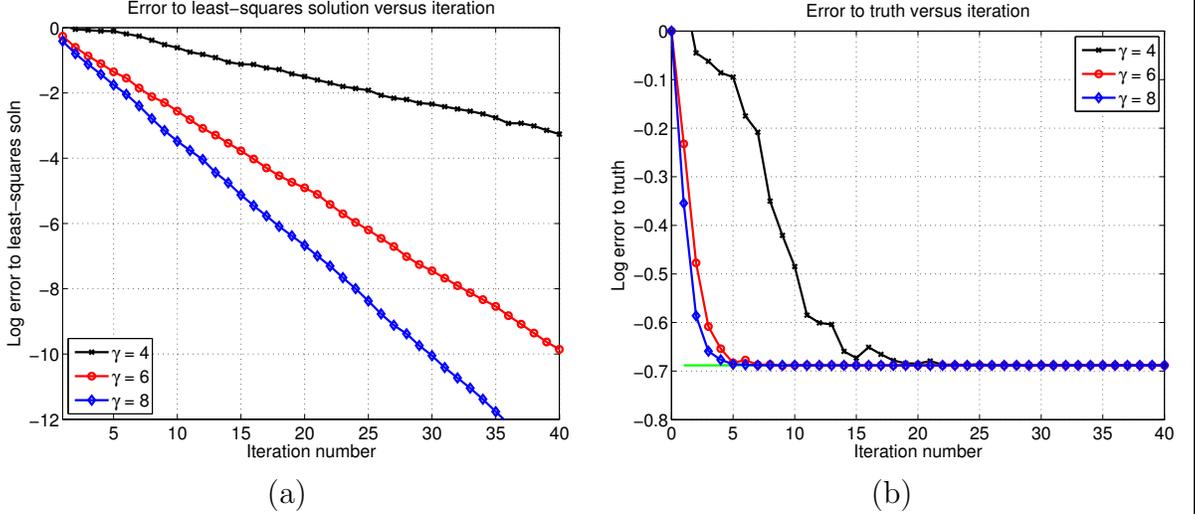


Figure 3.2: Simulations of the IHS algorithm for an unconstrained least-squares problem with noise variance $\sigma^2 = 1$, and of dimensions $(d, n) = (200, 6000)$.

Corollary 8. Fix some $\rho \in (0, 1/2)$, and choose a sketch dimension $m > \frac{c_0 \log^4(D)}{\rho^2} \mathcal{W}^2(\mathcal{K}_A^{LS})$. If we apply the IHS algorithm for $N(\rho, \varepsilon) := 1 + \frac{\log(1/\varepsilon)}{\log(1/\rho)}$ steps, then the output $\hat{x} = x^N$ satisfies the bound

$$\frac{\|\hat{x} - x^{LS}\|_A}{\|x^{LS}\|_A} \leq \varepsilon \quad (3.27)$$

with probability at least $1 - c_1 N(\rho, \varepsilon) e^{-c_2 \frac{m \rho^2}{\log^4(D)}}$.

This corollary is an immediate consequence of Theorem 4 combined with Lemma 15, and it holds for both ROS and sub-Gaussian sketches. (In the latter case, the additional $\log(D)$ terms may be omitted.) Combined with bounds on the width function $\mathcal{W}(\mathcal{K}_A^{LS})$, it leads to a number of concrete consequences for different statistical models, as we illustrate in the following section.

One way to understand the improvement of the IHS algorithm over the classical sketch is as follows. Fix some error tolerance $\varepsilon \in (0, 1)$. Disregarding logarithmic factors, our previous results [114] on the classical sketch then imply that a sketch size $m \gtrsim \varepsilon^{-2} \mathcal{W}^2(\mathcal{K}_A^{LS})$ is sufficient to produce a ε -accurate solution approximation. In contrast, Corollary 8 guarantees that a sketch size $m \gtrsim \log(1/\varepsilon) \mathcal{W}^2(\mathcal{K}_A^{LS})$ is sufficient. Thus, the benefit is the reduction from ε^{-2} to $\log(1/\varepsilon)$ scaling of the required sketch size.

It is worth noting that in the absence of constraints, the least-squares problem reduces to solving a linear system, so that alternative approaches are available. For instance, one can use a randomized sketch to obtain a preconditioner, which can then be used within the conjugate gradient method. As shown in past work [122, 14], two-step methods of this type can lead to same reduction of ε^{-2} dependence to $\log(1/\varepsilon)$. However, a method of this type is very specific to unconstrained least-squares, whereas the procedure described here is generally applicable to least-squares over any compact, convex constraint set.

3.1.5 Computational and space complexity

Let us now make a few comments about the computational and space complexity of implementing the IHS algorithm using the fast Johnson-Lindenstrauss (ROS) sketches, such as those based on the fast Hadamard transform. For a given sketch size m , the IHS algorithm requires $\mathcal{O}(nd \log(m))$ basic operations to compute the data sketch $S^{t+1}A$ at iteration t ; in addition, it requires $\mathcal{O}(nd)$ operations to compute $A^T(y - Ax^t)$. Consequently, if we run the algorithm for N iterations, then the overall complexity scales as

$$\mathcal{O}\left(N (nd \log(m) + C(m, d))\right), \quad (3.28)$$

where $C(m, d)$ is the complexity of solving the $m \times d$ dimensional problem in the update (3.24). Also note that, in problems where the data matrix A is sparse, $S^{t+1}A$ can be computed in time proportional to the number of non-zero elements in A using Gaussian sketching matrices. The space used by the sketches SA scales as $\mathcal{O}(md)$. To be clear, note that the IHS algorithm also requires access to the data via matrix-vector multiplies for forming $A^T(y - Ax^t)$. In limited memory environments, computing matrix-vector multiplies is considerably easier via distributed or interactive computation. For example, they can be efficiently implemented for multiple large datasets which can be loaded to memory only one at a time.

If we want to obtain estimates with accuracy ε , then we need to perform $N \asymp \log(1/\varepsilon)$ iterations in total. Moreover, for ROS sketches, we need to choose $m \gtrsim \mathbb{W}^2(\mathcal{K}_A^{\text{LS}}) \log^4(d)$. Consequently, it only remains to bound the Gaussian width \mathbb{W} in order to specify complexities that depend only on the pair (n, d) , and properties of the solution x^{LS} .

For an unconstrained problem with $n > d$, the Gaussian width can be bounded as $\mathbb{W}^2(\mathcal{K}_A^{\text{LS}}) \lesssim d$, and the complexity of the solving the sub-problem (3.24) can be bounded as d^3 . Thus, the overall complexity of computing an ε -accurate solution scales as $\mathcal{O}(nd \log(d) + d^3) \log(1/\varepsilon)$, and the space required is $\mathcal{O}(d^2)$.

As will be shown in Section 3.2.2, in certain cases, the cone $\mathcal{K}_A^{\text{LS}}$ can have substantially lower complexity than the unconstrained case. For instance, if the solution is sparse, say with k non-zero entries and the least-squares program involves an ℓ_1 -constraint, then we have $\mathbb{W}^2(\mathcal{K}_A^{\text{LS}}) \lesssim k \log d$. Using a standard interior point method to solve the sketched problem, the total complexity for obtaining an ε -accurate solution is upper bounded by $\mathcal{O}((nd \log(k) + k^2 d \log^2(d)) \log(1/\varepsilon))$. Although the sparsity k is not known a priori, there are bounds on it that can be computed in $\mathcal{O}(nd)$ time (for instance, see [58]).

3.2 Consequences for concrete models

In this section, we derive some consequences of Corollary 8 for particular classes of least-squares problems. Our goal is to provide empirical confirmation of the sharpness of our theoretical predictions, namely the minimal sketch dimension required in order to match the accuracy of the original least-squares solution.

3.2.1 Unconstrained least squares

We begin with the simplest case, namely the unconstrained least-squares problem ($\mathcal{C} = \mathbb{R}^d$). For a given pair (n, d) with $n > d$, we generated a random ensemble of least-square problems according to the following procedure:

- first, generate a random data matrix $A \in \mathbb{R}^{n \times d}$ with i.i.d. $N(0, 1)$ entries
- second, choose a regression vector x^* uniformly at random from the sphere \mathcal{S}^{d-1}
- third, form the response vector $y = Ax^* + w$, where $w \sim N(0, \sigma^2 I_n)$ is observation noise with $\sigma = 1$.

As discussed following Lemma 15, for this class of problems, taking a sketch dimension $m \gtrsim \frac{d}{\rho^2}$ guarantees ρ -contractivity of the IHS iterates with high probability. Consequently, we can obtain a ε -accurate approximation to the original least-squares solution by running roughly $\log(1/\varepsilon)/\log(1/\rho)$ iterations.

Now how should the tolerance ε be chosen? Recall that the underlying reason for solving the least-squares problem is to approximate x^* . Given this goal, it is natural to measure the approximation quality in terms of $\|x^t - x^*\|_A$. Panel (b) of Figure 3.2 shows the convergence of the iterates to x^* . As would be expected, this measure of

error levels off at the ordinary least-squares error

$$\|x^{\text{LS}} - x^*\|_A^2 \asymp \frac{\sigma^2 d}{n} \approx 0.10.$$

Consequently, it is reasonable to set the tolerance parameter proportional to $\sigma^2 \frac{d}{n}$, and then perform roughly $1 + \frac{\log(1/\varepsilon)}{\log(1/\rho)}$ steps. The following corollary summarizes the properties of the resulting procedure:

Corollary 9. *For some given $\rho \in (0, 1/2)$, suppose that we run the IHS algorithm for*

$$N = 1 + \lceil \frac{\log \sqrt{n} \frac{\|x^{\text{LS}}\|_A}{\sigma}}{\log(1/\rho)} \rceil$$

iterations using $m = \frac{c_0}{\rho^2} d$ projections per round. Then the output \hat{x} satisfies the bounds

$$\|\hat{x} - x^{\text{LS}}\|_A \leq \sqrt{\frac{\sigma^2 d}{n}}, \quad \text{and} \quad \|x^N - x^*\|_A \leq \sqrt{\frac{\sigma^2 d}{n}} + \|x^{\text{LS}} - x^*\|_A \quad (3.29)$$

with probability greater than $1 - c_1 N e^{-c_2 \frac{m \rho^2}{\log^4(d)}}$.

In order to confirm the predicted bound (3.29) on the error $\|\hat{x} - x^{\text{LS}}\|_A$, we performed a second experiment. Fixing $n = 100d$, we generated $T = 20$ random least squares problems from the ensemble described above with dimension d ranging over $\{32, 64, 128, 256, 512\}$. By our previous choices, the least-squares estimate should have error $\|x^{\text{LS}} - x^*\|_2 \approx \sqrt{\frac{\sigma^2 d}{n}} = 0.1$ with high probability, independently of the dimension d . This predicted behavior is confirmed by the blue bars in Figure 3.3; the bar height corresponds to the average over $T = 20$ trials, with the standard errors also marked. On these same problem instances, we also ran the IHS algorithm using $m = 6d$ samples per iteration, and for a total of

$$N = 1 + \lceil \frac{\log(\sqrt{\frac{n}{d}})}{\log 2} \rceil = 4 \quad \text{iterations.}$$

Since $\|x^{\text{LS}} - x^*\|_A \asymp \sqrt{\frac{\sigma^2 d}{n}} \approx 0.10$, Corollary 9 implies that with high probability, the sketched solution $\hat{x} = x^N$ satisfies the error bound

$$\|\hat{x} - x^*\|_2 \leq c'_0 \sqrt{\frac{\sigma^2 d}{n}}$$

for some constant $c'_0 > 0$. This prediction is confirmed by the green bars in Figure 3.3, showing that $\|\hat{x} - x^*\|_A \approx 0.11$ across all dimensions. Finally, the red bars show the results of running the classical sketch with a sketch dimension of $(6 \times 4)d = 24d$ sketches, corresponding to the total number of sketches used by the IHS algorithm. Note that the error is roughly twice as large.

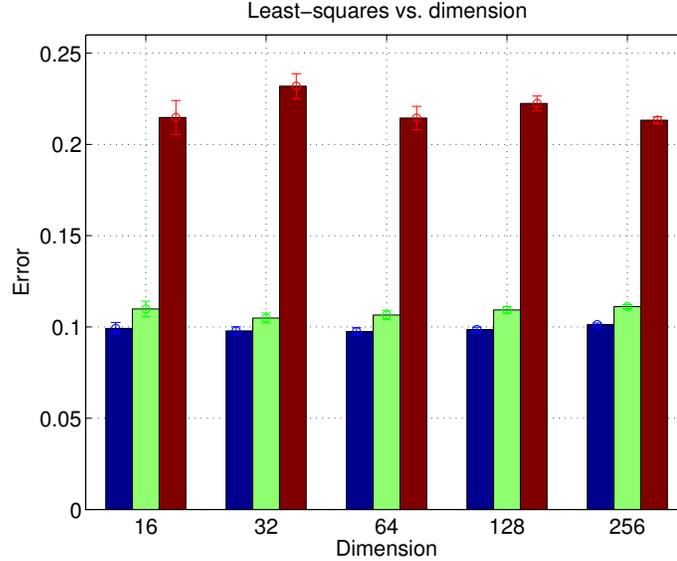


Figure 3.3: Simulations of the IHS algorithm for unconstrained least-squares.

3.2.2 Sparse least-squares

We now turn to a study of an ℓ_1 -constrained form of least-squares, referred to as the Lasso or relaxed basis pursuit program [36, 134]. In particular, consider the convex program

$$x^{\text{LS}} = \arg \min_{\|x\|_1 \leq R} \left\{ \frac{1}{2} \|y - Ax\|_2^2 \right\}, \quad (3.30)$$

where $R > 0$ is a user-defined radius. This estimator is well-suited to the problem of sparse linear regression, based on the observation model $y = Ax^* + w$, where x^* has at most k non-zero entries, and $A \in \mathbb{R}^{n \times d}$ has i.i.d. $N(0, 1)$ entries. For the purposes of this illustration, we assume³ that the radius is chosen such that $R = \|x^*\|_1$.

Under these conditions, the proof of Corollary 10 shows that a sketch size $m \geq \gamma k \log\left(\frac{ed}{k}\right)$ suffices to guarantee geometric convergence of the IHS updates. Panel (a) of Figure 3.4 illustrates the accuracy of this prediction, showing the resulting convergence rate of the the IHS algorithm, measured in terms of the error $\|x^t - x^{\text{LS}}\|_A$, for different values $\gamma \in \{2, 5, 25\}$. As predicted by Theorem 4, the convergence rate is geometric (linear on the log scale shown), with the rate increasing as the parameter γ is increased.

³In practice, this unrealistic assumption of exactly knowing $\|x^*\|_1$ is avoided by instead considering the ℓ_1 -penalized form of least-squares, but we focus on the constrained case to keep this illustration as simple as possible.

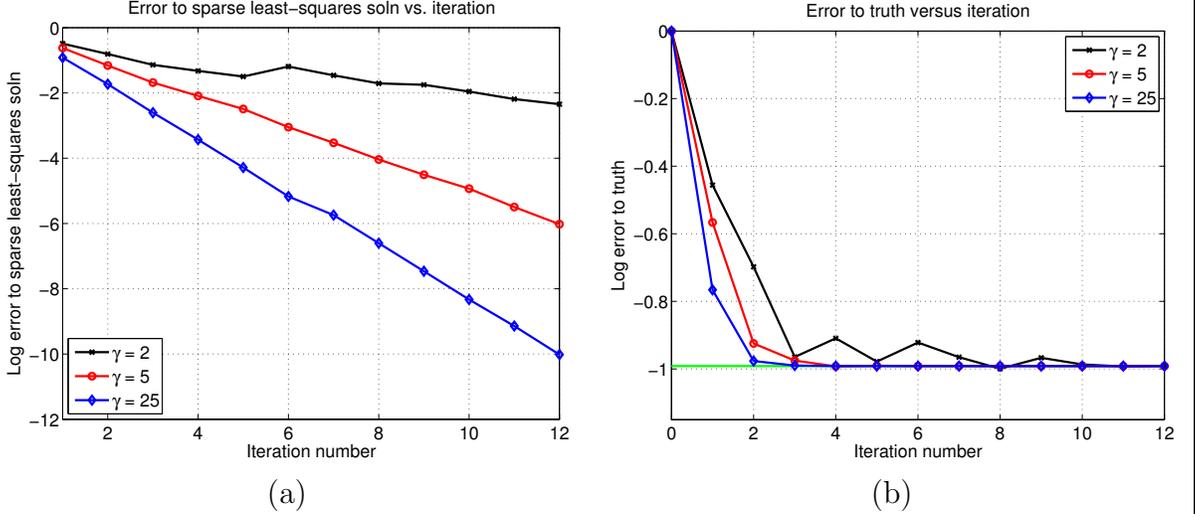


Figure 3.4: Plots of the log error $\|x^t - x^{\text{LS}}\|_2$ (a) and $\|x^t - x^*\|_2$ (b) versus the iteration number t .

As long as $n \gtrsim k \log\left(\frac{ed}{k}\right)$, it also follows as a corollary of Proposition 4 that

$$\|x^{\text{LS}} - x^*\|_A^2 \lesssim \frac{\sigma^2 k \log\left(\frac{ed}{k}\right)}{n}. \quad (3.31)$$

with high probability. This bound suggests an appropriate choice for the tolerance parameter ε in Theorem 4, and leads us to the following guarantee.

Corollary 10. *For the stated random ensemble of sparse linear regression problems, suppose that we run the IHS algorithm for $N = 1 + \lceil \frac{\log \sqrt{n} \|x^{\text{LS}}\|_A}{\log(1/\rho)} \rceil$ iterations using $m = \frac{c_0}{\rho^2} k \log\left(\frac{ed}{k}\right)$ projections per round. Then with probability greater than $1 - c_1 N e^{-c_2 \frac{m \rho^2}{\log^4(d)}}$, the output \hat{x} satisfies the bounds*

$$\|\hat{x} - x^{\text{LS}}\|_A \leq \sqrt{\frac{\sigma^2 k \log\left(\frac{ed}{k}\right)}{n}} \quad \text{and} \quad \|x^N - x^*\|_A \leq \sqrt{\frac{\sigma^2 k \log\left(\frac{ed}{k}\right)}{n}} + \|x^{\text{LS}} - x^*\|_A. \quad (3.32)$$

In order to verify the predicted bound (3.32) on the error $\|\hat{x} - x^{\text{LS}}\|_A$, we performed a second experiment. Fixing $n = 100k \log\left(\frac{ed}{k}\right)$, we generated $T = 20$ random least squares problems (as described above) with the regression dimension ranging as $d \in$

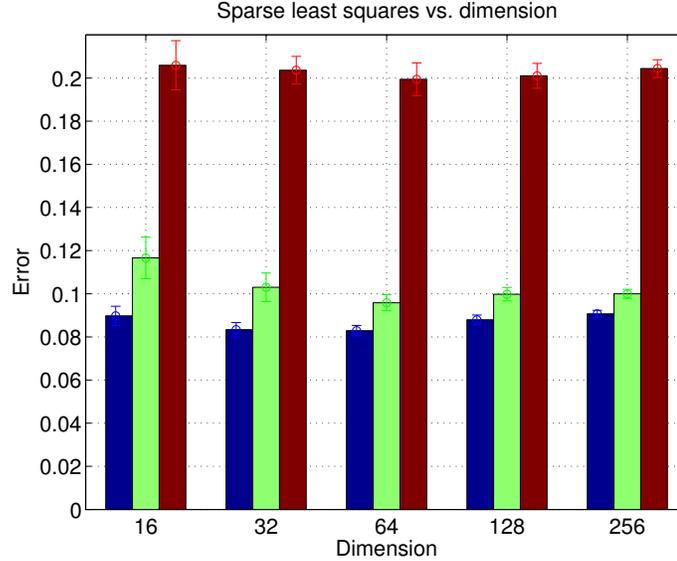


Figure 3.5: Simulations of the IHS algorithm for ℓ_1 -constrained least-squares

$\{32, 64, 128, 256\}$, and sparsity $k = \lceil 2\sqrt{d} \rceil$. Based on these choices, the least-squares estimate should have error $\|x^{\text{LS}} - x^*\|_A \approx \sqrt{\frac{\sigma^2 k \log(\frac{ed}{k})}{n}} = 0.1$ with high probability, independently of the pair (k, d) . This predicted behavior is confirmed by the blue bars in Figure 3.5; the bar height corresponds to the average over $T = 20$ trials, with the standard errors also marked.

On these same problem instances, we also ran the IHS algorithm using $N = 4$ iterations with a sketch size $m = 4k \log(\frac{ed}{k})$. Together with our earlier calculation of $\|x^{\text{LS}} - x^*\|_A$, Corollary 9 implies that with high probability, the sketched solution $\hat{x} = x^N$ satisfies the error bound

$$\|\hat{x} - x^*\|_A \leq c_0 \sqrt{\frac{\sigma^2 k \log(\frac{ed}{k})}{n}} \quad (3.33)$$

for some constant $c_0 \in (1, 2]$. This prediction is confirmed by the green bars in Figure 3.5, showing that $\|\hat{x} - x^*\|_A \lesssim 0.11$ across all dimensions. Finally, the red bars in Figure 3.5 show the error based on using the naive sketch estimate with a total of $M = Nm$ random projections in total; as with the case of ordinary least-squares, the resulting error is roughly twice as large. We also note that a similar bound also applies to problems where a parameter constrained to unit simplex is estimated, such as in portfolio analysis and density estimation [91, 111].

3.2.3 Some larger-scale experiments

In order to further explore the computational gains guaranteed by IHS, we performed some larger scale experiments on sparse regression problems, with the sample size n ranging over the set $\{2^{12}, 2^{13}, \dots, 2^{19}\}$ with a fixed input dimension $d = 500$. As before, we generate observations from the linear model $y = Ax^* + w$, where x^* has at most k non-zero entries, and each row of the data matrix $A \in \mathbb{R}^{n \times d}$ is distributed i.i.d. according to a $N(1_d, \Sigma)$ distribution. Here the d -dimensional covariance matrix Σ has entries $\Sigma_{jk} = 2 \times 0.9^{|j-k|}$, so that the columns of the matrix A will be correlated. Setting a sparsity $k = \lceil 3 \log(d) \rceil$, we chose the unknown regression vector x^* with its support uniformly random with entries $\pm \frac{1}{\sqrt{k}}$ with equal probability.

Baseline: In order to provide a baseline for comparison, we used the homotopy algorithm—that is, the Lasso modification of the LARS updates [110, 57]—to solve the original ℓ_1 constrained problem with ℓ_1 -ball radius $R = \sqrt{k}$. The homotopy algorithm is especially efficient when the Lasso solution x^{LS} is sparse. Since the columns of A are correlated in our ensemble, standard first-order algorithms—among them iterative soft-thresholding, FISTA, spectral projected gradient methods, as well as (block) coordinate descent methods, see, e.g., [20, 149]—performed poorly relative to the homotopy algorithm in terms of computation time; see [18] for observations of this phenomenon in past work.

IHS implementation: For comparison, we implemented the IHS algorithm with a projection dimension $m = \lfloor 4k \log(d) \rfloor$. After projecting the data, we then used the homotopy method to solve the projected sub-problem at each step. In each trial, we ran the IHS algorithm for $N = \lceil \log n \rceil$ iterations.

Table 3.1 provides a summary comparison of the running times for the baseline method (homotopy method on the original problem), versus the IHS method (running time for computing the iterates using the homotopy method), and IHS method plus sketching time. Note that with the exception of the smallest problem size ($n = 4096$), the IHS method including sketching time is the fastest, and it is more than two times faster for large problems. The gains are somewhat more significant if we remove the sketching time from the comparison.

One way in which to measure the quality of the least-squares solution x^{LS} as an estimate of x^* is via its mean-squared (in-sample) prediction error $\|x^{\text{LS}} - x^*\|_A^2 = \frac{\|A(x^{\text{LS}} - x^*)\|_2^2}{n}$. For the random ensemble of problems that we have generated, the bound (3.33) guarantees that the squared error should decay at the rate $1/n$ as the sample size n is increased with the dimension d and sparsity k fixed. Figure 3.6 compares the prediction MSE of x^{LS} versus the analogous quantity $\|\hat{x} - x^*\|_A^2$ for the sketched solution. Note that the two curves are essentially indistinguishable, showing

Samples n	4096	8192	16384	32768	65536	131072	262144	524288
Baseline	0.0840	0.1701	0.3387	0.6779	1.4083	2.9052	6.0163	12.0969
IHS	0.0783	0.0993	0.1468	0.2174	0.3601	0.6846	1.4748	3.1593
IHS+Sketch	0.0877	0.1184	0.1887	0.3222	0.5814	1.1685	2.5967	5.5792

Table 3.1: Running time comparison in seconds of the Baseline (homotopy method applied to original problem), IHS (homotopy method applied to sketched subproblems), and IHS plus sketching time. Each running time estimate corresponds to an average over 300 independent trials of the random sparse regression model described in the main text.

that the sketched solution provides an estimate of x^* that is as good as the original least-squares estimate.

3.2.4 Matrix estimation with nuclear norm constraints

We now turn to the study of nuclear-norm constrained form of least-squares matrix regression. This class of problems has proven useful in many different application areas, among them matrix completion, collaborative filtering, multi-task learning and control theory (e.g., [59, 153, 15, 121, 102]). In particular, let us consider the convex program

$$X^{\text{LS}} = \arg \min_{X \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2} \|Y - AX\|_{\text{F}}^2 \right\} \quad \text{such that } \|X\|_* \leq R, \quad (3.34)$$

where $R > 0$ is a user-defined radius as a regularization parameter.

3.2.4.1 Simulated data

Recall the linear observation model previously introduced in Example 3: we observe the pair (Y, A) linked according to the linear $Y = AX^* + W$, where the unknown matrix $X^* \in \mathbb{R}^{d_1 \times d_2}$ is an unknown matrix of rank r . The matrix W is observation noise, formed with i.i.d. $N(0, \sigma^2)$ entries. This model is a special case of the more general class of matrix regression problems [102]. As shown in Section 3.7.2, if we solve the nuclear-norm constrained problem with $R = \|X^*\|_*$, then it produces a solution such that $\mathbb{E}[\|X^{\text{LS}} - X^*\|_{\text{F}}^2] \lesssim \sigma^2 \frac{r(d_1 + d_2)}{n}$. The following corollary characterizes

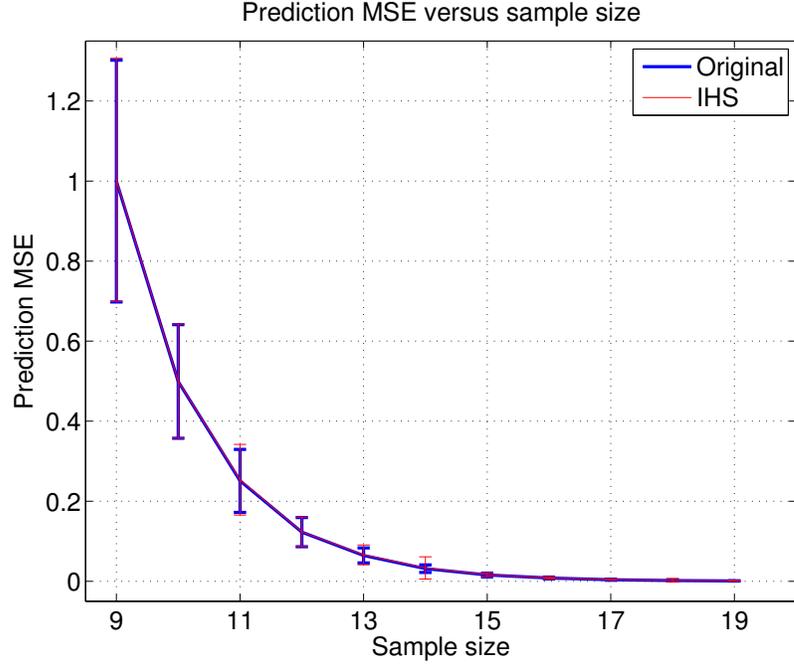


Figure 3.6: Plots of the mean-squared prediction errors $\frac{\|A(\tilde{x}-x^*)\|_2^2}{n}$ versus the sample size $n \in 2^{\{9,10,\dots,19\}}$ for the original least-squares solution ($\tilde{x} = x^{LS}$ in blue) versus the sketched solution ($\hat{x} = x^{LS}$ in red).

the sketch dimension and iteration number required for the IHS algorithm to match this scaling up to a constant factor.

Corollary 11 (IHS for nuclear-norm constrained least squares). *Suppose that we run the IHS algorithm for $N = 1 + \lceil \frac{\log \sqrt{n} \|X^{LS}\|_A}{\log(1/\rho)} \rceil$ iterations using $m = c_0 \rho^2 r (d_1 + d_2)$ projections per round. Then with probability greater than $1 - c_1 N e^{-c_2 \frac{m \rho^2}{\log^4(d_1 d_2)}}$, the output X^N satisfies the bound*

$$\|X^N - X^*\|_A \leq \sqrt{\frac{\sigma^2 r (d_1 + d_2)}{n}} + \|X^{LS} - X^*\|_A. \quad (3.35)$$

We have also performed simulations for low-rank matrix estimation, and observed that the IHS algorithm exhibits convergence behavior qualitatively similar to that

shown in Figures 3.3 and 3.5. Similarly, panel (a) of Figure 3.8 compares the performance of the IHS and classical methods for sketching the optimal solution over a range of row sizes n . As with the unconstrained least-squares results from Figure 3.1, the classical sketch is very poor compared to the original solution whereas the IHS algorithm exhibits near optimal performance.

3.2.4.2 Application to multi-task learning

To conclude, let us illustrate the use of the IHS algorithm in speeding up the training of a classifier for facial expressions. In particular, suppose that our goal is to separate a collection of facial images into different groups, corresponding either to distinct individuals or to different facial expressions. One approach would be to learn a different linear classifier ($a \mapsto \langle a, x \rangle$) for each separate task, but since the classification problems are so closely related, the optimal classifiers are likely to share structure. One way of capturing this shared structure is by concatenating all the different linear classifiers into a matrix, and then estimating this matrix in conjunction with a nuclear norm penalty [9, 11].



Figure 3.7: Japanese Female Facial Expression (JAFFE) Database: The JAFFE database consists of 213 images of 7 different emotional facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models.

In more detail, we performed a simulation study using the The Japanese Female Facial Expression (JAFFE) database [89]. It consists of $N = 213$ images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 different Japanese female models; see Figure 3.7 for a few example images. We performed an approximately 80 : 20 split of the data set into $n_{\text{train}} = 170$ training and $n_{\text{test}} = 43$ test images respectively. Then we consider classifying each facial expression and each female model as a separate task which gives a total of $d_{\text{task}} = 17$ tasks. For each task $j = 1, \dots, d_{\text{task}}$, we construct a linear classifier of the form $a \mapsto \text{sign}(\langle a, x_j \rangle)$,

where $a \in \mathbb{R}^d$ denotes the vectorized image features given by Local Phase Quantization [108]. In our implementation, we fixed the number of features $d = 32$. Given this set-up, we train the classifiers in a joint manner, by optimizing simultaneously over the matrix $X \in \mathbb{R}^{d \times d_{\text{task}}}$ with the classifier vector $x_j \in \mathbb{R}^d$ as its j^{th} column. The image data is loaded into the matrix $A \in \mathbb{R}^{n_{\text{train}} \times d}$, with image feature vector $a_i \in \mathbb{R}^d$ in column i for $i = 1, \dots, n_{\text{train}}$. Finally, the matrix $Y \in \{-1, +1\}^{n_{\text{train}} \times d_{\text{task}}}$ encodes class labels for the different classification problems. These instantiations of the pair (Y, X) give us an optimization problem of the form (3.34), and we solve it over a range of regularization radii R .

More specifically, in order to verify the classification accuracy of the classifier obtained by IHT algorithm, we solved the original convex program, the classical sketch based on ROS sketches of dimension $m = 100$, and also the corresponding IHS algorithm using ROS sketches of size 20 in each of 5 iterations. In this way, both the classical and IHS procedures use the same total number of sketches, making for a fair comparison. We repeated each of these three procedures for all choices of the radius $R \in \{1, 2, 3, \dots, 12\}$, and then applied the resulting classifiers to classify images in the test dataset. For each of the three procedures, we calculated the classification error rate, defined as the total number of mis-classified images divided by $n_{\text{test}} \times d_{\text{task}}$. Panel (b) of Figure 3.8 plots the resulting classification errors versus the regularization parameter. The error bars correspond to one standard deviation calculated over the randomness in generating sketching matrices. The plots show that the IHS algorithm yields classifiers with performance close to that given by the original solution over a range of regularizer parameters, and is superior to the classification sketch. The error bars also show that the IHS algorithm has less variability in its outputs than the classical sketch.

3.3 Discussion

In chapter, we focused on the problem of solution approximation (as opposed to cost approximation) for a broad class of constrained least-squares problem. We began by showing that the classical sketching methods are sub-optimal, from an information-theoretic point of view, for the purposes of solution approximation. We then proposed a novel iterative scheme, known as the iterative Hessian sketch, for deriving ε -accurate solution approximations. We proved a general theorem on the properties of this algorithm, showing that the sketch dimension per iteration need grow only proportionally to the statistical dimension of the optimal solution, as measured by the Gaussian width of the tangent cone at the optimum. By taking $\log(1/\varepsilon)$ iterations, the IHS algorithm is guaranteed to return an ε -accurate solution approximation with exponentially high probability.

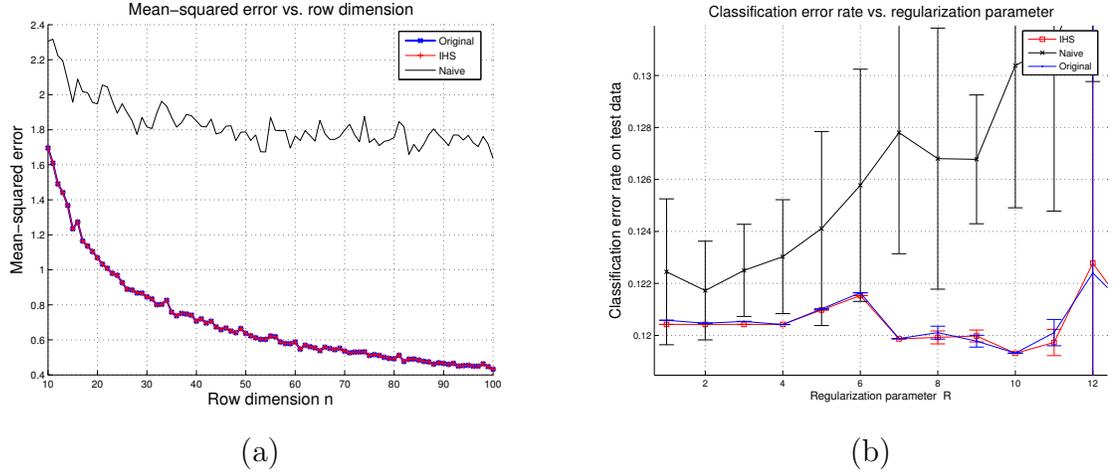


Figure 3.8: Simulations of the IHS algorithm for nuclear-norm constrained problems on the JAFFE dataset: Mean-squared error versus the row dimension $n \in [10, 100]$ for recovering a 20×20 matrix of rank r_2 , using a sketch dimension $m = 60$ (a). Classification error rate versus regularization parameter $R \in \{1, \dots, 12\}$, with error bars corresponding to one standard deviation over the test set (b).

In addition to these theoretical results, we also provided empirical evaluations that reveal the sub-optimality of the classical sketch, and show that the IHS algorithm produces near-optimal estimators. Finally, we applied our methods to a problem of facial expression using a multi-task learning model applied to the JAFFE face database. We showed that IHS algorithm applied to a nuclear-norm constrained program produces classifiers with considerably better classification accuracy compared to the naive sketch.

There are many directions for further research, but we only list here some of them. The idea behind iterative sketching can also be applied to problems beyond minimizing a least-squares objective function subject to convex constraints. Examples include penalized forms of regression, e.g., see the recent work [151], and various other cost functions. An important class of such problems are ℓ_p -norm forms of regression, based on the convex program

$$\min_{x \in \mathbb{R}^d} \|Ax - y\|_p^p \quad \text{for some } p \in [1, \infty].$$

The case of ℓ_1 -regression ($p = 1$) is an important special case, known as robust regression; it is especially effective for data sets containing outliers [69]. Recent

work [37] has proposed to find faster solutions of the ℓ_1 -regression problem using the classical sketch (i.e., based on (SA, Sy)) but with sketching matrices based on Cauchy random vectors. Our iterative technique might be useful in obtaining sharper bounds for solution approximation in this setting as well. In the following section we will show how these result can be generalized to sketching for general convex objective functions.

3.4 Proof of lower bounds

This section is devoted to the verification of condition (3.8) for different model classes, followed by the proof of Theorem 3.

3.4.1 Verification of condition (3.8)

We verify the condition for three different types of sketches.

3.4.1.1 Gaussian sketches:

First, let $S \in \mathbb{R}^{m \times n}$ be a random matrix with i.i.d. Gaussian entries. We use the singular value decomposition to write $S = U\Lambda V^T$ where both U and V are orthonormal matrices of left and right singular vectors. By rotation invariance, the columns $\{v_i\}_{i=1}^m$ are uniformly distributed over the sphere \mathcal{S}^{n-1} . Consequently, we have

$$\mathbb{E}_S[S^T(SS^T)^{-1}S] = \mathbb{E} \sum_{i=1}^m v_i v_i^T = \frac{m}{n} I_n, \quad (3.36)$$

showing that condition (3.8) holds with $\eta = 1$.

3.4.1.2 ROS sketches (sampled without replacement):

In this case, we have $S = \sqrt{n}PHD$, where $P \in \mathbb{R}^{m \times n}$ is a random picking matrix with each row being a standard basis vector sampled without replacement. We then have $SS^T = nI_m$ and also $\mathbb{E}_P[P^T P] = \frac{m}{n} I_n$, so that

$$\mathbb{E}_S[S^T(SS^T)^{-1}S] = \mathbb{E}_{D,P}[DH^T P^T PHD] = \mathbb{E}_D[DH^T (\frac{m}{n} I_n) HD] = \frac{m}{n} I_n,$$

showing that the condition holds with $\eta = 1$.

3.4.1.3 Weighted row sampling:

Finally, suppose that we sample m rows independently using a distribution $\{p_j\}_{j=1}^n$ on the rows of the data matrix that is α -balanced (3.6). Letting $\mathcal{R} \subseteq \{1, 2, \dots, n\}$ be the subset of rows that are sampled, and let N_j be the number of times each row is sampled. We then have

$$\mathbb{E}\left[S^T(SS^T)^{-1}S\right] = \sum_{j \in \mathcal{R}} \mathbb{E}[e_j e_j^T] = D,$$

where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with entries $D_{jj} = \mathbb{P}[j \in \mathcal{R}]$. Since the trials are independent, the j^{th} row is sampled at least once in m trials with probability $q_j = 1 - (1 - p_j)^m$, and hence

$$\mathbb{E}_S[S^T(SS^T)^{-1}S] = \text{diag}(\{1 - (1 - p_i)^m\}_{i=1}^n) \preceq (1 - (1 - p_\infty)^m)I_n \preceq mp_\infty,$$

where $p_\infty = \max_{j \in [n]} p_j$. Consequently, as long as the row weights are α -balanced (3.6) so that $p_\infty \leq \frac{\alpha}{n}$, we have

$$\|\mathbb{E}_S[S^T(SS^T)^{-1}S]\|_2 \leq \alpha \frac{m}{n}$$

showing that condition (3.8) holds with $\eta = \alpha$, as claimed.

3.4.2 Proof of Theorem 3

Let $\{z^j\}_{j=1}^M$ be a $1/2$ -packing of $\mathcal{C}_0 \cap \mathbb{B}_A(1)$ in the semi-norm $\|\cdot\|_A$, and for a fixed $\delta \in (0, 1/4)$, define $x^j = 4\delta z^j$. Since $4\delta \in (0, 1)$, the star-shaped assumption guarantees that each x^j belongs to \mathcal{C}_0 . We thus obtain a collection of M vectors in \mathcal{C}_0 such that

$$2\delta \leq \underbrace{\frac{1}{\sqrt{n}} \|A(x^j - x^k)\|_2}_{\|x^j - x^k\|_A} \leq 8\delta \quad \text{for all } j \neq k.$$

Letting J be a random index uniformly distributed over $\{1, \dots, M\}$, suppose that conditionally on $J = j$, we observe the sketched observation vector $Sy = SAx^j + Sw$, as well as the sketched matrix SA . Conditioned on $J = j$, the random vector Sy follows a $N(SAx^j, \sigma^2 SS^T)$ distribution, denoted by \mathbb{P}_{x^j} . We let \bar{Y} denote the resulting mixture variable, with distribution $\frac{1}{M} \sum_{j=1}^M \mathbb{P}_{x^j}$.

Consider the multiway testing problem of determining the index J based on observing \bar{Y} . With this set-up, a standard reduction in statistical minimax

(e.g., [23, 152]) implies that, for any estimator x^\dagger , the worst-case mean-squared error is lower bounded as

$$\sup_{x^* \in \mathcal{C}} \mathbb{E}_{S,w} \|x^\dagger - x^*\|_A^2 \geq \delta^2 \inf_{\psi} \mathbb{P}[\psi(\bar{Y}) \neq J], \quad (3.37)$$

where the infimum ranges over all testing functions ψ . Consequently, it suffices to show that the testing error is lower bounded by $1/2$.

In order to do so, we first apply Fano's inequality [39] conditionally on the sketching matrix S to see that

$$\mathbb{P}[\psi(\bar{Y}) \neq J] = \mathbb{E}_S \left\{ \mathbb{P}[\psi(\bar{Y}) \neq J \mid S] \right\} \geq 1 - \frac{\mathbb{E}_S [I_S(\bar{Y}; J)] + \log 2}{\log M}, \quad (3.38)$$

where $I_S(\bar{Y}; J)$ denotes the mutual information between \bar{Y} and J with S fixed. Our next step is to upper bound the expectation $\mathbb{E}_S [I(\bar{Y}; J)]$.

Letting $D(\mathbb{P}_{x^j} \parallel \mathbb{P}_{x^k})$ denote the Kullback-Leibler divergence between the distributions \mathbb{P}_{x^j} and \mathbb{P}_{x^k} , the convexity of Kullback-Leibler divergence implies that

$$I_S(\bar{Y}; J) = \frac{1}{M} \sum_{j=1}^M D(\mathbb{P}_{x^j} \parallel \frac{1}{M} \sum_{k=1}^M \mathbb{P}_{x^k}) \leq \frac{1}{M^2} \sum_{j,k=1}^M D(\mathbb{P}_{x^j} \parallel \mathbb{P}_{x^k}).$$

Computing the KL divergence for Gaussian vectors yields

$$I_S(\bar{Y}; J) \leq \frac{1}{M^2} \sum_{j,k=1}^M \frac{1}{2\sigma^2} (x^j - x^k)^T A^T \left[S^T (S S^T)^{-1} S \right] A (x^j - x^k).$$

Thus, using condition (3.8), we have

$$\mathbb{E}_S [I(\bar{Y}; J)] \leq \frac{1}{M^2} \sum_{j,k=1}^M \frac{m \eta}{2 n \sigma^2} \|A(x^j - x^k)\|_2^2 \leq \frac{32 m \eta}{\sigma^2} \delta^2,$$

where the final inequality uses the fact that $\|x^j - x^k\|_A \leq 8\delta$ for all pairs.

Combined with our previous bounds (3.37) and (3.38), we find that

$$\sup_{x^* \in \mathcal{C}} \mathbb{E} \|\hat{x} - x^*\|_2^2 \geq \delta^2 \left\{ 1 - \frac{32 \frac{m \eta \delta^2}{\sigma^2} + \log 2}{\log M} \right\}.$$

Setting $\delta = \frac{\sigma^2 \log(M/2)}{64 \eta m}$ yields the lower bound (3.9).

3.5 Proof of Proposition 3

Since \hat{x} and x^{LS} are optimal and feasible, respectively, for the Hessian sketch program (3.15), we have

$$\langle A^T S^T (SA\hat{x} - y), x^{\text{LS}} - \hat{x} \rangle \geq 0 \quad (3.39a)$$

Similarly, since x^{LS} and \hat{x} are optimal and feasible, respectively, for the original least squares program

$$\langle A^T (Ax^{\text{LS}} - y), \hat{x} - x^{\text{LS}} \rangle \geq 0. \quad (3.39b)$$

Adding these two inequalities and performing some algebra yields the basic inequality

$$\frac{1}{m} \|SA\Delta\|_2^2 \leq \left| (Ax^{\text{LS}})^T \left(I_n - \frac{S^T S}{m} \right) A\Delta \right|. \quad (3.40)$$

Since Ax^{LS} is independent of the sketching matrix and $A\Delta \in \mathcal{K}_A^{\text{LS}}$, we have

$$\frac{1}{m} \|SA\Delta\|_2^2 \geq Z_1(A\mathcal{K}) \|A\Delta\|_2^2, \quad \text{and} \quad \left| (Ax^{\text{LS}})^T \left(I_n - \frac{S^T S}{m} \right) A\Delta \right| \leq Z_2(A\mathcal{K}) \|Ax^{\text{LS}}\|_2 \|A\Delta\|_2, \blacksquare$$

using the definitions (3.17a) and (3.17b) of the random variables $Z_1(A\mathcal{K})$ and $Z_2(A\mathcal{K})$ respectively. Combining the pieces yields the claim.

3.6 Proof of Theorem 4

It suffices to show that, for each iteration $t = 0, 1, 2, \dots$, we have

$$\|x^{t+1} - x^{\text{LS}}\|_A \leq \frac{Z_2(A\mathcal{K})(S^{t+1})}{Z_1(A\mathcal{K})(S^{t+1})} \|x^t - x^{\text{LS}}\|_A. \quad (3.41)$$

The claimed bounds (3.26a) and (3.26b) then follow by applying the bound (3.41) successively to iterates 1 through N .

For simplicity in notation, we abbreviate S^{t+1} to S and x^{t+1} to \hat{x} . Define the error vector $\Delta = \hat{x} - x^{\text{LS}}$. With some simple algebra, the optimization problem (3.24) that underlies the update $t + 1$ can be re-written as

$$\hat{x} = \arg \min_{x \in \mathcal{C}} \left\{ \frac{1}{2m} \|SAx\|_2^2 - \langle A^T \tilde{y}, x \rangle \right\},$$

where $\tilde{y} := y - \left[I - \frac{S^T S}{m} \right] Ax^t$. Since \hat{x} and x^{LS} are optimal and feasible respectively, the usual first-order optimality conditions imply that

$$\left\langle A^T \frac{S^T S}{m} Ax - A^T \tilde{y}, x^{\text{LS}} - \hat{x} \right\rangle \geq 0.$$

As before, since x^{LS} is optimal for the original program, we have

$$\langle A^T(Ax^{\text{LS}} - \tilde{y} + [I - \frac{S^T S}{m}]Ax^t), \hat{x} - x^{\text{LS}} \rangle \geq 0.$$

Adding together these two inequalities and introducing the shorthand $\Delta = \hat{x} - x^{\text{LS}}$ yields

$$\frac{1}{m} \|SA\Delta\|_2^2 \leq \left| (A(x^{\text{LS}} - x^t))^T [I - \frac{S^T S}{m}] A\Delta \right| \quad (3.42)$$

Note that the vector $A(x^{\text{LS}} - x^t)$ is independent of the randomness in the sketch matrix S^{t+1} . Moreover, the vector $A\Delta$ belongs to the cone \mathcal{K} , so that by the definition of $Z_2(A\mathcal{K})(S^{t+1})$, we have

$$\left| (A(x^{\text{LS}} - x^t))^T [I - \frac{S^T S}{m}] A\Delta \right| \leq \|A(x^{\text{LS}} - x^t)\|_2 \|A\Delta\|_2 Z_2(A\mathcal{K})(S^{t+1}). \quad (3.43a)$$

Similarly, note the lower bound

$$\frac{1}{m} \|SA\Delta\|_2^2 \geq \|A\Delta\|_2^2 Z_1(A\mathcal{K})(S^{t+1}). \quad (3.43b)$$

Combining the two bounds (3.43a) and (3.43b) with the earlier bound (3.42) yields the claim (3.41).

3.7 Maximum likelihood estimator and examples

In this section, we a general upper bound on the error of the constrained least-squares estimate. We then use it (and other results) to work through the calculations underlying Examples 1 through 3 from Section 3.1.2.

3.7.1 Upper bound on MLE

The accuracy of x^{LS} as an estimate of x^* depends on the “size” of the star-shaped set

$$\mathcal{K}(x^*) = \left\{ v \in \mathbb{R}^d \mid v = \frac{t}{\sqrt{n}} A(x - x^*) \text{ for some } t \in [0, 1] \text{ and } x \in \mathcal{C} \right\}. \quad (3.44)$$

When the vector x^* is clear from context, we use the shorthand notation \mathcal{K}^* for this set. By taking a union over all possible $x^* \in \mathcal{C}_0$, we obtain the set $\bar{\mathcal{K}} := \bigcup_{x^* \in \mathcal{C}_0} \mathcal{K}(x^*)$,

which plays an important role in our bounds. The complexity of these sets can be measured of their *localized Gaussian widths*. For any radius $\varepsilon > 0$ and set $\Theta \subseteq \mathbb{R}^n$, the Gaussian width of the set $\Theta \cap \mathbb{B}_2(\varepsilon)$ is given by

$$\mathcal{W}_\varepsilon(\Theta) := \mathbb{E}_g \left[\sup_{\substack{\theta \in \Theta \\ \|\theta\|_2 \leq \varepsilon}} |\langle w, \theta \rangle| \right], \quad (3.45a)$$

where $g \sim N(0, I_{n \times n})$ is a standard Gaussian vector. Whenever the set Θ is star-shaped, then it can be shown that, for any $\sigma > 0$ and positive integer ℓ , the inequality

$$\frac{\mathcal{W}_\varepsilon(\Theta)}{\varepsilon \sqrt{\ell}} \leq \frac{\varepsilon}{\sigma} \quad (3.45b)$$

has a smallest positive solution, which we denote by $\varepsilon_\ell(\Theta; \sigma)$. We refer the reader to [19] for further discussion of such localized complexity measures and their properties.

The following result bounds the mean-squared error associated with the constrained least-squares estimate:

Proposition 4. *For any set \mathcal{C} containing x^* , the constrained least-squares estimate (3.1) has mean-squared error upper bounded as*

$$\mathbb{E}_w [\|x^{LS} - x^*\|_A^2] \leq c_1 \left\{ \varepsilon_n^2(\mathcal{K}^*) + \frac{\sigma^2}{n} \right\} \leq c_1 \left\{ \varepsilon_n^2(\bar{\mathcal{K}}) + \frac{\sigma^2}{n} \right\}. \quad (3.46)$$

We provide the proof of this claim in Section 3.7.3.

3.7.2 Detailed calculations for illustrative examples

In this section, we collect together the details of calculations used in our illustrative examples from Section 3.1.2. In all cases, we make use of the convenient shorthand $\tilde{A} = A/\sqrt{n}$.

3.7.2.1 Unconstrained least squares: Example 1

By definition of the Gaussian width, we have

$$\mathcal{W}_\delta(\mathcal{K}^*) = \mathbb{E}_g \left[\sup_{\|\tilde{A}(x-x^*)\|_2 \leq \delta} |\langle g, \tilde{A}(x-x^*) \rangle| \right] \leq \delta \sqrt{d}$$

since the vector $\tilde{A}(x-x^*)$ belongs to a subspace of dimension $\text{rank}(A) = d$. The claimed upper bound (3.10a) thus follows as a consequence of Proposition 4.

3.7.2.2 Sparse vectors: Example 2

The RIP property of order $8k$ implies that

$$\frac{\|\Delta\|_2^2}{2} \stackrel{(i)}{\leq} \|\tilde{A}\Delta\|_2^2 \stackrel{(ii)}{\leq} 2\|\Delta\|_2^2 \quad \text{for all vectors with } \|\Delta\|_0 \leq 8k,$$

a fact which we use throughout the proof. By definition of the Gaussian width, we have

$$\mathcal{W}_\delta(\mathcal{K}^*) = \mathbb{E}_g \left[\sup_{\substack{\|x\|_1 \leq \|x^*\|_1 \\ \|\tilde{A}(x-x^*)\|_2 \leq \delta}} |\langle g, \tilde{A}(x-x^*) \rangle| \right].$$

Since $x^* \in \mathbb{B}_0(k)$, it can be shown (e.g., see the proof of Corollary 3 in [114]) that for any vector $\|x\|_1 \leq \|x^*\|_1$, we have $\|x - x^*\|_1 \leq 2\sqrt{k}\|x - x^*\|_2$. Thus, it suffices to bound the quantity

$$F(\delta; k) := \mathbb{E}_g \left[\sup_{\substack{\|\Delta\|_1 \leq 2\sqrt{k}\|\Delta\|_2 \\ \|\tilde{A}\Delta\|_2 \leq \delta}} |\langle g, \tilde{A}\Delta \rangle| \right].$$

By Lemma 11 in [87], we have

$$\mathbb{B}_1(\sqrt{s}) \cap \mathbb{B}_2(1) \subseteq 3 \text{ clconv} \left\{ \mathbb{B}_0(s) \cap \mathbb{B}_2(1) \right\},$$

where clconv denotes the closed convex hull. Applying this lemma with $s = 4k$, we have

$$F(\delta; k) \leq 3 \left[\sup_{\substack{\|\Delta\|_0 \leq 4k \\ \|\tilde{A}\Delta\|_2 \leq \delta}} |\langle g, \tilde{A}\Delta \rangle| \right] \leq 3 \mathbb{E} \left[\sup_{\substack{\|\Delta\|_0 \leq 4k \\ \|\Delta\|_2 \leq 2\delta}} |\langle g, \tilde{A}\Delta \rangle| \right],$$

using the lower RIP property (i). By the upper RIP property, for any pair of vectors Δ, Δ' with ℓ_0 -norms at most $4k$, we have

$$\text{var}(\langle g, \tilde{A}\Delta \rangle - \langle g, \tilde{A}\Delta' \rangle) \leq 2\|\Delta - \Delta'\|_2^2 = 2 \text{var}(\langle g, \Delta - \Delta' \rangle)$$

Consequently, by the Sudakov-Fernique comparison [85], we have

$$\mathbb{E} \left[\sup_{\substack{\|\Delta\|_0 \leq 4k \\ \|\Delta\|_2 \leq 2\delta}} |\langle g, \tilde{A}\Delta \rangle| \right] \leq 2 \mathbb{E} \left[\sup_{\substack{\|\Delta\|_0 \leq 4k \\ \|\Delta\|_2 \leq 2\delta}} |\langle g, \Delta \rangle| \right] \leq c \delta \sqrt{k \log\left(\frac{ed}{k}\right)},$$

where the final inequality standard results on Gaussian widths [63]. All together, we conclude that

$$\varepsilon_n^2(\mathcal{K}^*; \sigma) \leq c_1 \sigma^2 \frac{k \log\left(\frac{ed}{k}\right)}{n}.$$

Combined with Proposition 4, the claimed upper bound (3.11a) follows.

In the other direction, a straightforward argument (e.g., [119]) shows that there is a universal constant $c > 0$ such that $\log M_{1/2} \geq ck \log\left(\frac{ed}{k}\right)$, so that the stated lower bound follows from Theorem 3.

3.7.2.3 Low rank matrices: Example 3:

By definition of the Gaussian width, we have

$$\mathcal{W}_\delta(\mathcal{K}^*) = \mathbb{E}_g \left[\sup_{\substack{\|\tilde{A}(X-X^*)\|_F \leq \delta \\ \|X\|_* \leq \|X^*\|_*}} |\langle \tilde{A}^T G, (X - X^*) \rangle| \right],$$

where $G \in \mathbb{R}^{n \times d_2}$ is a Gaussian random matrix, and $\langle C, D \rangle$ denotes the trace inner product between matrices C and D . Since X^* has rank at most r , it can be shown that $\|X - X^*\|_* \leq 2\sqrt{r}\|X - X^*\|_F$; for instance, see Lemma 1 in [101]. Recalling that $\gamma_{\min}(\tilde{A})$ denotes the minimum singular value, we have

$$\|X - X^*\|_F \leq \frac{1}{\gamma_{\min}(\tilde{A})} \|\tilde{A}(X - X^*)\|_F \leq \frac{\delta}{\gamma_{\min}(\tilde{A})}.$$

Thus, by duality between the nuclear and operator norms, we have

$$\mathbb{E}_g \left[\sup_{\substack{\|\tilde{A}(X-X^*)\|_F \leq \delta \\ \|X\|_* \leq \|X^*\|_*}} |\langle G, \tilde{A}(X - X^*) \rangle| \right] \leq \frac{2\sqrt{r}\delta}{\gamma_{\min}(A)} \mathbb{E}[\|\tilde{A}^T G\|_2].$$

Now consider the matrix $A^T G \in \mathbb{R}^{d_1 \times d_2}$. For any fixed pair of vectors $(u, v) \in \mathcal{S}^{d_1-1} \times \mathcal{S}^{d_2-1}$, the random variable $Z = u^T \tilde{A}^T G v$ is zero-mean Gaussian with variance at most $\gamma_{\max}^2(\tilde{A})$. Consequently, by a standard covering argument in random matrix theory [140], we have $\mathbb{E}[\|\tilde{A}^T G\|_2] \lesssim \gamma_{\max}(\tilde{A})(\sqrt{d_1 + d_2})$. Putting together the pieces, we conclude that

$$\varepsilon_n^2 \preceq \sigma^2 \frac{\gamma_{\max}^2(A)}{\gamma_{\min}^2(A)} r (d_1 + d_2),$$

so that the upper bound (3.14a) follows from Proposition 4.

3.7.3 Proof of Proposition 4

Throughout this proof, we adopt the shorthand $\varepsilon_n = \varepsilon_n(\mathcal{K}^*)$. Our strategy is to prove the following more general claim: for any $t \geq \varepsilon_n$, we have

$$\mathbb{P}_{S,w} [\|x^{\text{LS}} - x^*\|_A^2 \geq 16t\varepsilon_n] \leq c_1 e^{-c_2 \frac{nt\varepsilon_n}{\sigma^2}}. \quad (3.47)$$

A simple integration argument applied to this tail bound implies the claimed bound (3.46) on the expected mean-squared error.

Since x^* and x^{LS} are feasible and optimal, respectively, for the optimization problem (3.1), we have the basic inequality

$$\frac{1}{2n} \|y - Ax^{\text{LS}}\|_2^2 \leq \frac{1}{2n} \|y - Ax^*\|_2^2 = \frac{1}{2n} \|w\|_2^2.$$

Introducing the shorthand $\Delta = x^{\text{LS}} - x^*$ and re-arranging terms yields

$$\frac{1}{2} \|\Delta\|_A^2 = \frac{1}{2n} \|A\Delta\|_2^2 \leq \frac{\sigma}{n} \left| \sum_{i=1}^n \langle g, A\Delta \rangle_i \right|, \quad (3.48)$$

where $g \sim N(0, I_n)$ is a standard normal vector.

For a given $u \geq \varepsilon_n$, define the “bad” event

$$\mathcal{B}(u) := \left\{ \exists z \in \mathcal{C} - x^* \text{ with } \|z\|_A \geq u, \text{ and } \left| \frac{\sigma}{n} \sum_{i=1}^n g_i (Az)_i \right| \geq 2u \|z\|_A \right\}$$

The following lemma controls the probability of this event:

Lemma 16. *For all $u \geq \varepsilon_n$, we have $\mathbb{P}[\mathcal{B}(u)] \leq e^{-\frac{nu^2}{2\sigma^2}}$.*

Returning to prove this lemma momentarily, let us prove the bound (3.47). For any $t \geq \varepsilon_n$, we can apply Lemma 16 with $u = \sqrt{t\varepsilon_n}$ to find that

$$\mathbb{P}[\mathcal{B}^c(\sqrt{t\varepsilon_n})] \geq 1 - e^{-\frac{nt\varepsilon_n}{2\sigma^2}}.$$

If $\|\Delta\|_A < \sqrt{t\varepsilon_n}$, then the claim is immediate. Otherwise, we have $\|\Delta\|_A \geq \sqrt{t\varepsilon_n}$. Since $\Delta \in \mathcal{C} - x^*$, we may condition on $\mathcal{B}^c(\sqrt{t\varepsilon_n})$ so as to obtain the bound

$$\left| \frac{\sigma}{n} \sum_{i=1}^n g_i (A\Delta)_i \right| \leq 2 \|\Delta\|_A \sqrt{t\varepsilon_n}.$$

Combined with the basic inequality (3.48), we see that

$$\frac{1}{2} \|\Delta\|_A^2 \leq 2 \|\Delta\|_A \sqrt{t\varepsilon_n}, \quad \text{or equivalently } \|\Delta\|_A^2 \leq 16t\varepsilon_n,$$

a bound that holds with probability greater than $1 - e^{-\frac{nt\varepsilon_n}{2\sigma^2}}$ as claimed.

It remains to prove Lemma 16. Our proof involves the auxiliary random variable

$$V_n(u) := \sup_{\substack{z \in \text{star}(\mathcal{C} - x^*) \\ \|z\|_A \leq u}} \left| \frac{\sigma}{n} \sum_{i=1}^n g_i (Az)_i \right|,$$

Inclusion of events: We first claim that $\mathcal{B}(u) \subseteq \{V_n(u) \geq 2u^2\}$. Indeed, if $\mathcal{B}(u)$ occurs, then there exists some $z \in \mathcal{C} - x^*$ with $\|z\|_A \geq u$ and

$$\left| \frac{\sigma}{n} \sum_{i=1}^n g_i (Az)_i \right| \geq 2u \|z\|_A. \quad (3.49)$$

Define the rescaled vector $\tilde{z} = \frac{u}{\|z\|_A} z$. Since $z \in \mathcal{C} - x^*$ and $\frac{u}{\|z\|_A} \leq 1$, the vector $\tilde{z} \in \text{star}(\mathcal{C} - x^*)$. Moreover, by construction, we have $\|\tilde{z}\|_A = u$. When the inequality (3.49) holds, the vector \tilde{z} thus satisfies $\left| \frac{\sigma}{n} \sum_{i=1}^n g_i (A\tilde{z})_i \right| \geq 2u^2$, which certifies that $V_n(u) \geq 2u^2$, as claimed.

Controlling the tail probability: The final step is to control the probability of the event $\{V_n(u) \geq 2u^2\}$. Viewed as a function of the standard Gaussian vector (g_1, \dots, g_n) , it is easy to see that $V_n(u)$ is Lipschitz with constant $L = \frac{\sigma u}{\sqrt{n}}$. Consequently, by concentration of measure for Lipschitz Gaussian functions, we have

$$\mathbb{P}[V_n(u) \geq \mathbb{E}[V_n(u)] + u^2] \leq e^{-\frac{nu^2}{2\sigma^2}}. \quad (3.50)$$

In order to complete the proof, it suffices to show that $\mathbb{E}[V_n(u)] \leq u^2$. By definition, we have $\mathbb{E}[V_n(u)] = \frac{\sigma}{\sqrt{n}} \mathcal{W}_u(\mathcal{K}^*)$. Since \mathcal{K}^* is a star-shaped set, the function $v \mapsto \mathcal{W}_v(\mathcal{K}^*)/v$ is non-increasing [19]. Since $u \geq \varepsilon_n$, we have

$$\sigma \frac{\mathcal{W}_u(\mathcal{K}^*)}{u} \leq \sigma \frac{\mathcal{W}_{\varepsilon_n}(\mathcal{K}^*)}{\varepsilon_n} \leq \varepsilon_n.$$

where the final step follows from the definition of ε_n . Putting together the pieces, we conclude that $\mathbb{E}[V_n(u)] \leq \varepsilon_n u \leq u^2$ as claimed.

Chapter 4

Random projections for nonlinear optimization

Relative to first-order methods, second-order methods for convex optimization enjoy superior convergence in both theory and practice. For instance, Newton’s method converges at a quadratic rate for strongly convex and smooth problems. Even for functions that are weakly convex—that is, convex but not strongly convex—modifications of Newton’s method have super-linear convergence (for instance, see the paper [150] for an analysis of the Levenberg-Marquardt Method). This rate is faster than the $1/T^2$ convergence rate that can be achieved by a first-order method like accelerated gradient descent, with the latter rate known to be unimprovable (in general) for first-order methods [104]. Yet another issue in first-order methods is the tuning of step size, whose optimal choice depends on the strong convexity parameter and/or smoothness of the underlying problem. For example, consider the problem of optimizing a function of the form $x \mapsto g(Ax)$, where $A \in \mathbb{R}^{n \times d}$ is a “data matrix”, and $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is a twice-differentiable function. Here the performance of first-order methods will depend on both the convexity/smoothness of g , as well as the conditioning of the data matrix. In contrast, whenever the function g is self-concordant, then Newton’s method with suitably damped steps has a global complexity guarantee that is provably independent of such problem-dependent parameters.

On the other hand, each step of Newton’s method requires solving a linear system defined by the Hessian matrix. For instance, in application to the problem family just described involving an $n \times d$ data matrix, each of these steps has complexity scaling as $\mathcal{O}(nd^2)$. For this reason, both forming the Hessian and solving the corresponding linear system pose a tremendous numerical challenge for large values of (n, d) —for instance, values of thousands to millions, as is common in big data applications. In order to address this issue, a wide variety of different approximations to Newton’s method have been proposed and studied. The general class of quasi-Newton methods are based on estimating the inverse Hessian using successive evaluations of the gradient

vectors. Examples of such quasi-Newton methods include DFP and BFGS schemes as well their limited memory versions; see the book by Wright and Nocedal [148] and references therein for further details. A disadvantage of such first-order Hessian approximations is that the associated convergence guarantees are typically weaker than those of Newton’s method and require stronger assumptions.

In this chapter, we propose and analyze a randomized approximation of Newton’s method, known as the *Newton Sketch*. Instead of explicitly computing the Hessian, the Newton Sketch method approximates it via a random projection of dimension m . When these projections are carried out using the fast Johnson-Lindenstrauss (JL) transform, say based on Hadamard matrices, each iteration has complexity $\mathcal{O}(nd \log(m) + dm^2)$. Our results show that it is always sufficient to choose m proportional to $\min\{d, n\}$, and moreover, that the sketch dimension m can be much smaller for certain types of constrained problems. Thus, in the regime $n > d$ and with $m \asymp d$, the complexity per iteration can be substantially lower than the $\mathcal{O}(nd^2)$ complexity of each Newton step. For instance, for an objective function of the form $f(x) = g(Ax)$ in the regime $n \geq d^2$, the complexity of Newton Sketch per iteration is $\mathcal{O}(nd \log d)$, which (modulo the logarithm) is linear in the input data size nd . Thus, the computational complexity per iteration is comparable to first-order methods that have access only to the gradient $A^T g'(Ax)$. In contrast to first-order methods, we show that for self-concordant functions, the total complexity of obtaining a δ -optimal solution is $\mathcal{O}(nd(\log d) \log(1/\delta))$, and without any dependence on constants such as strong convexity or smoothness parameters. Moreover, for problems with $d > n$, we provide a dual strategy that effectively has the same guarantees with roles of d and n exchanged.

We also consider other random projection matrices and sub-sampling strategies, including partial forms of random projection that exploit known structure in the Hessian. For self-concordant functions, we provide an affine invariant analysis proving that the convergence is linear-quadratic and the guarantees are independent of various problem parameters, such as condition numbers of matrices involved in the objective function. Finally, we describe an interior point method to deal with arbitrary convex constraints, which combines the Newton sketch with the barrier method. We provide an upper bound on the total number of iterations required to obtain a solution with a pre-specified target accuracy.

The remainder of this chapter is organized as follows. We begin in Section 4.1 with some background on the classical form of Newton’s method, past work on approximate forms of Newton’s method, random matrices for sketching, and Gaussian widths as a measure of the size of a set. In Section 4.2, we formally introduce the Newton Sketch, including both fully and partially sketched versions for unconstrained and constrained problems. We provide some illustrative examples in Section 4.2.3 before turning to local convergence theory in Section 4.2.4. Section 4.3 is devoted to global convergence results for self-concordant functions, in both the constrained and

unconstrained settings. In Section 4.4, we consider a number of applications and provide additional numerical results. The bulk of our proofs are given in Section 4.5, with some more technical aspects deferred to later sections.

4.1 Background

We begin with some background material on the standard form of Newton’s method, past work on approximate or stochastic forms of Newton’s method, the basics of random sketching, and the notion of Gaussian width as a complexity measure.

4.1.1 Classical version of Newton’s method

In this section, we briefly review the convergence properties and complexity of the classical form of Newton’s method; see the sources [148, 28, 104] for further background. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a closed, convex and twice-differentiable function that is bounded below. Given a convex and closed set \mathcal{C} , we assume that the constrained minimizer

$$x^* := \arg \min_{x \in \mathcal{C}} f(x) \tag{4.1}$$

exists and is uniquely defined. We define the minimum and maximum eigenvalues $\gamma = \lambda_{\min}(\nabla^2 f(x^*))$ and $\beta = \lambda_{\max}(\nabla^2 f(x^*))$ of the Hessian evaluated at the minimum.

We assume moreover that the Hessian map $x \mapsto \nabla^2 f(x)$ is Lipschitz continuous with modulus L , meaning that

$$\|\nabla^2 f(x + \Delta) - \nabla^2 f(x)\|_2 \leq L \|\Delta\|_2. \tag{4.2}$$

Under these conditions and given an initial point $\tilde{x}^0 \in \mathcal{C}$ such that $\|\tilde{x}^0 - x^*\|_2 \leq \frac{\gamma}{2L}$, the Newton updates are guaranteed to converge quadratically—viz.

$$\|\tilde{x}^{t+1} - x^*\|_2 \leq \frac{2L}{\gamma} \|\tilde{x}^t - x^*\|_2^2,$$

This result is classical: for instance, see Boyd and Vandenberghe [28] for a proof. Newton’s method can be slightly modified to be globally convergent by choosing the step sizes via a simple backtracking line-search procedure.

The following result characterizes the complexity of Newton’s method when applied to self-concordant functions and is central in the development of interior point methods (for instance, see the books [107, 28]). We defer the definitions of self-concordance and the line-search procedure in the following sections. The number

of iterations needed to obtain a δ approximate minimizer of a strictly convex self-concordant function f is bounded by

$$\frac{20 - 8a}{ab(1 - 2a)} (f(x^0) - f(x^*)) + \log_2 \log_2(1/\delta),$$

where a, b are constants in the line-search procedure.¹

4.1.2 Approximate Newton methods

Given the complexity of the exact Newton updates, various forms of approximate and stochastic variants of Newton’s method have been proposed, which we discuss here. In general, inexact solutions of the Newton updates can be used to guarantee convergence while reducing overall computational complexity [47, 48]. In the unconstrained setting, the Newton update corresponds to solving a linear system of equations, and one approximate approach is truncated Newton’s method: it involves applying the conjugate gradient (CG) method for a specified number of iterations, and then using the solution as an approximate Newton step [48]. In applying this method, the Hessian need not be formed since the CG updates only need access to matrix-vector products with the Hessian. While this strategy is popular, theoretical analysis of inexact Newton methods typically need strong assumptions on the eigenvalues of the Hessian [47]. Since the number of steps of CG for reaching a certain residual error necessarily depends on the condition number, the overall complexity of truncated Newton’s Method is problem-dependent; the condition numbers can be arbitrarily large, and in general are unknown *a priori*. Ill-conditioned Hessian systems are common in applications of Newton’s method within interior point methods. Consequently, software toolboxes typically perform approximate Newton steps using CG updates in earlier iterations, but then shift to exact Newton steps via Cholesky or QR decompositions in later iterations.

A more recent line of work, inspired by the success of stochastic first-order algorithms for large scale machine learning applications, has focused on stochastic forms of second-order optimization algorithms (e.g., [126, 24, 32, 33]). Schraudolph et al. [126] use online limited memory BFGS-like updates to maintain an inverse Hessian approximation. Byrd et al. [33, 32] propose stochastic second-order methods that use batch sub-sampling in order to obtain curvature information in a computationally inexpensive manner. These methods are numerically effective in problems in which objective consists of a sum of a large number of individual terms; however, their theoretical analysis again involves strong assumptions on the eigenvalues of the Hessian. Moreover, such second-order methods do not retain the affine invariance of the original Newton’s method, which guarantees iterates are independent of the coordinate system and conditioning. When simple stochastic schemes like sub-sampling are used

¹Typical values of these constants are $a = 0.1$ and $b = 0.5$ in practice.

to approximate the Hessian, affine invariance is lost, since subsampling is coordinate and conditioning dependent. In contrast, the stochastic form of Newton’s method we proposed is constructed so as to retain this affine invariance property, and thus not depend on the problem conditioning.

4.2 Newton Sketch and local convergence

With the basic background in place, let us now introduce the Newton sketch algorithm, and then develop a number of convergence guarantees associated with it. It applies to an optimization problem of the form $\min_{x \in \mathcal{C}} f(x)$, where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a twice-differentiable convex function, and $\mathcal{C} \subseteq \mathbb{R}^d$ is a closed and convex constraint set.

4.2.1 Newton Sketch algorithm

In order to motivate the Newton Sketch algorithm, recall the standard form of Newton’s algorithm: given a current iterate $\tilde{x}^t \in \mathcal{C}$, it generates the new iterate \tilde{x}^{t+1} by performing a constrained minimization of the second order Taylor expansion—viz.

$$\tilde{x}^{t+1} = \arg \min_{x \in \mathcal{C}} \left\{ \frac{1}{2} \langle x - \tilde{x}^t, \nabla^2 f(\tilde{x}^t) (x - \tilde{x}^t) \rangle + \langle \nabla f(\tilde{x}^t), x - \tilde{x}^t \rangle \right\}. \quad (4.3a)$$

In the unconstrained case—that is, when $\mathcal{C} = \mathbb{R}^d$ —it takes the simpler form

$$\tilde{x}^{t+1} = \tilde{x}^t - [\nabla^2 f(\tilde{x}^t)]^{-1} \nabla f(\tilde{x}^t). \quad (4.3b)$$

Now suppose that we have available a Hessian matrix square root $\nabla^2 f(x)^{1/2}$ —that is, a matrix $\nabla^2 f(x)^{1/2}$ of dimensions $n \times d$ such that

$$(\nabla^2 f(x)^{1/2})^T \nabla^2 f(x)^{1/2} = \nabla^2 f(x) \quad \text{for some integer } n \geq \text{rank}(\nabla^2 f(x)).$$

In many cases, such a matrix square root can be computed efficiently. For instance, consider a function of the form $f(x) = g(Ax)$ where $A \in \mathbb{R}^{n \times d}$, and the function $g : \mathbb{R}^n \rightarrow \mathbb{R}$ has the separable form $g(Ax) = \sum_{i=1}^n g_i(\langle a_i, x \rangle)$. In this case, a suitable Hessian matrix square root is given by the $n \times d$ matrix $\nabla^2 f(x)^{1/2} := \text{diag}\{g_i''(\langle a_i, x \rangle)^{1/2}\}_{i=1}^n A$. In Section 4.2.3, we discuss various concrete instantiations of such functions.

In terms of this notation, the ordinary Newton update can be re-written as

$$\tilde{x}^{t+1} = \arg \min_{x \in \mathcal{C}} \underbrace{\left\{ \frac{1}{2} \|\nabla^2 f(\tilde{x}^t)^{1/2} (x - \tilde{x}^t)\|_2^2 + \langle \nabla f(\tilde{x}^t), x - \tilde{x}^t \rangle \right\}}_{\tilde{\Phi}(x)},$$

and the Newton Sketch algorithm is most easily understood based on this form of the updates. More precisely, for a sketch dimension m to be chosen, let $S \in \mathbb{R}^{m \times n}$ be a sub-Gaussian, ROS, sparse-JL sketch or subspace embedding (when \mathcal{C} is a subspace), satisfying the relation $\mathbb{E}[S^T S] = I_n$. The *Newton Sketch algorithm* generates a sequence of iterates $\{x^t\}_{t=0}^\infty$ according to the recursion

$$x^{t+1} \in \arg \min_{x \in \mathcal{C}} \left\{ \underbrace{\frac{1}{2} \|S^t \nabla^2 f(x^t)^{1/2} (x - x^t)\|_2^2 + \langle \nabla f(x^t), x - x^t \rangle}_{\Phi(x; S^t)} \right\}, \quad (4.4)$$

where $S^t \in \mathbb{R}^{m \times d}$ is an independent realization of a sketching matrix. When the problem is unconstrained, i.e., $\mathcal{C} = \mathbb{R}^d$ and the matrix $\nabla^2 f(x^t)^{1/2} (S^t)^T S^t \nabla^2 f(x^t)^{1/2}$ is invertible, the Newton Sketch update takes the simpler form

$$x^{t+1} = x^t - (\nabla^2 f(x^t)^{1/2} (S^t)^T S^t \nabla^2 f(x^t)^{1/2})^{-1} \nabla f(x^t). \quad (4.5)$$

The intuition underlying the Newton Sketch updates is as follows: the iterate x^{t+1} corresponds to the constrained minimizer of the random objective function $\Phi(x; S^t)$ whose expectation $\mathbb{E}[\Phi(x; S^t)]$, taking averages over the isotropic sketch matrix S^t , is equal to the original Newton objective $\tilde{\Phi}(x)$. Consequently, it can be seen as a stochastic form of the Newton update, which minimizes a random quadratic approximation at each iteration.

We also analyze a *partially sketched Newton update*, which takes the following form. Given an additive decomposition of the form $f = f_0 + g$, we perform a sketch of the Hessian $\nabla^2 f_0$ while retaining the exact form of the Hessian $\nabla^2 g$. This splitting leads to the partially sketched update

$$x^{t+1} := \arg \min_{x \in \mathcal{C}} \left\{ \frac{1}{2} (x - x^t)^T Q^t (x - x^t) + \langle \nabla f(x^t), x - x^t \rangle \right\}, \quad (4.6)$$

where $Q^t := (S^t \nabla^2 f_0(x^t)^{1/2})^T S^t \nabla^2 f_0(x^t)^{1/2} + \nabla^2 g(x^t)$.

For either the fully sketched (4.4) or partially sketched updates (4.6), our analysis shows that there are many settings in which the sketch dimension m can be chosen to be substantially smaller than n , in which cases the sketched Newton updates will be much cheaper than a standard Newton update. For instance, the unconstrained update (4.5) can be computed in at most $\mathcal{O}(md^2)$ time, as opposed to the $\mathcal{O}(nd^2)$ time of the standard Newton update. In constrained settings, we show that the sketch dimension m can often be chosen even smaller—even $m \ll d$ —which leads to further savings.

4.2.2 Affine invariance of the Newton Sketch and sketched KKT systems

A desirable feature of the Newton Sketch is that, similar to the original Newton's method, both of its forms remain (statistically) invariant under an affine transformation. In other words, if we apply Newton Sketch on an affine transformation of a particular function, the statistics of the iterates are related by the same transformation. As a concrete example, consider the problem of minimizing a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ subject to equality constraints $Cx = e$, for some matrix $C \in \mathbb{R}^{n \times d}$ and vector $e \in \mathbb{R}^n$. For this particular problem, the Newton Sketch update takes the form

$$x^{t+1} := \arg \min_{Cx=d} \left\{ \frac{1}{2} \|S^t \nabla^2 f(x^t)^{1/2} (x - x^t)\|_2^2 + \langle \nabla f(x^t), x - x^t \rangle \right\}. \quad (4.7)$$

Equivalently, by introducing Lagrangian dual variables for the linear constraints, it is equivalent to solve the following *sketched KKT system*

$$\begin{bmatrix} (\nabla^2 f(x^t)^{1/2})^T (S^t)^T S^t \nabla^2 f(x^t)^{1/2} & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\text{NSK}} \\ w_{\text{NSK}} \end{bmatrix} = - \begin{bmatrix} \nabla f(x^t) \\ 0 \end{bmatrix}$$

where $\Delta x_{\text{NSK}} = x^{t+1} - x^t \in \mathbb{R}^d$ is the sketched Newton step where x^t is assumed feasible, and $w_{\text{NSK}} \in \mathbb{R}^n$ is the optimal dual variable for the stochastic quadratic approximation.

Now fix the random sketching matrix S^t and consider the transformed objective function $\widehat{f}(y) := f(By)$, where $B \in \mathbb{R}^{d \times d}$ is an invertible matrix. If we apply the Newton Sketch algorithm to the transformed problem involving \widehat{f} , the sketched Newton step Δy_{NSK} is given by the solution to the system

$$\begin{bmatrix} B^T (\nabla^2 f(x^t)^{1/2})^T (S^t)^T S^t \nabla^2 f(x^t)^{1/2} B & B^T C^T \\ CB & 0 \end{bmatrix} \begin{bmatrix} \Delta y_{\text{NSK}} \\ \widehat{w}_{\text{NSK}} \end{bmatrix} = - \begin{bmatrix} B^T \nabla f(x^t) \\ 0 \end{bmatrix},$$

which shows that $B \Delta y_{\text{NSK}} = \Delta x_{\text{NSK}}$. Note that the upper-left block in the above matrix is has rank at most m , and consequently the above 2×2 block matrix has rank at most $m + \text{rank}(C)$.

4.2.3 Some examples

In order to provide some intuition, let us provide some simple examples to which the sketched Newton updates can be applied.

Example: Newton Sketch for LP solving
Consider a linear program (LP) in the standard form

$$\min_{Ax \leq b} \langle c, x \rangle \quad (4.8)$$

where $A \in \mathbb{R}^{n \times d}$ is a given constraint matrix. We assume that the polytope $\{x \in \mathbb{R}^d \mid Ax \leq b\}$ is bounded so that the minimum is achieved. A barrier method approach to this LP is based on solving a sequence of problems of the form

$$\min_{x \in \mathbb{R}^d} \left\{ \tau \langle c, x \rangle - \underbrace{\sum_{i=1}^n \log(b_i - \langle a_i, x \rangle)}_{f(x)} \right\},$$

where $a_i \in \mathbb{R}^d$ denotes the i^{th} row of A , and $\tau > 0$ is a weight parameter that is adjusted during the algorithm. By inspection, the function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is twice-differentiable, and its Hessian is given by $\nabla^2 f(x) = A^T \text{diag}\left\{\frac{1}{(b_i - \langle a_i, x \rangle)^2}\right\} A$. A Hessian square root is given by $\nabla^2 f(x)^{1/2} := \text{diag}\left(\frac{1}{|b_i - \langle a_i, x \rangle|}\right) A$, which allows us to compute the sketched version

$$S \nabla^2 f(x)^{1/2} = S \text{diag}\left(\frac{1}{|b_i - \langle a_i, x \rangle|}\right) A.$$

With a ROS sketch matrix, computing this matrix requires $\mathcal{O}(nd \log(m))$ basic operations. The complexity of each Newton Sketch iteration scales as $\mathcal{O}(md^2)$, where m is at most $\mathcal{O}(d)$. In contrast, the standard unsketched form of the Newton update has complexity $\mathcal{O}(nd^2)$, so that the sketched method is computationally cheaper whenever there are many more constraints than dimensions ($n > d$).

By increasing the barrier parameter τ , we obtain a sequence of solutions that approach the optimum to the LP, which we refer to as the central path. As a simple illustration, Figure 4.1 compares the central paths generated by the ordinary and sketched Newton updates for a polytope defined by $n = 32$ constraints in dimension $d = 2$. Each row shows three independent trials of the method for a given sketch dimension m ; the top, middle and bottom rows correspond to sketch dimensions $m \in \{d, 4d, 16d\}$ respectively. Note that as the sketch dimension m is increased, the central path taken by the sketched updates converges to the standard central path.

As a second example, we consider the problem of maximum likelihood estimation for generalized linear models.

Example: Newton Sketch for maximum likelihood estimation

The class of generalized linear models (GLMs) is used to model a wide variety of prediction and classification problems, in which the goal is to predict some output variable $y \in \mathcal{Y}$ on the basis of a covariate vector $a \in \mathbb{R}^d$. It includes as special cases the standard linear Gaussian model (in which $\mathcal{Y} = \mathbb{R}$), as well as logistic models for classification (in which $\mathcal{Y} = \{-1, +1\}$), as well as Poisson models for count-valued responses (in which $\mathcal{Y} = \{0, 1, 2, \dots\}$). See the book [94] for further details and applications.

Given a collection of n observations $\{(y_i, a_i)\}_{i=1}^n$ of response-covariate pairs from some GLM, the problem of constrained maximum likelihood estimation be written in the form

$$\min_{x \in \mathcal{C}} \left\{ \underbrace{\sum_{i=1}^n \psi(\langle a_i, x \rangle, y_i)}_{f(x)} \right\}, \quad (4.9)$$

where $\psi : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$ is a given convex function, and $\mathcal{C} \subset \mathbb{R}^d$ is a convex constraint set, chosen by the user to enforce a certain type of structure in the solution. Important special cases of GLMs include the linear Gaussian model, in which $\psi(u, y) = \frac{1}{2}(y - u)^2$, and the problem (4.9) corresponds to a regularized form of least-squares, as well as the problem of logistic regression, obtained by setting $\psi(u, y) = \log(1 + \exp(-yu))$.

Letting $A \in \mathbb{R}^{n \times d}$ denote the data matrix with $a_i \in \mathbb{R}^d$ as its i^{th} row, the Hessian of the objective (4.9) takes the form

$$\nabla^2 f(x) = A^T \text{diag}(\psi''(a_i^T x))_{i=1}^n A$$

Since the function ψ is convex, we are guaranteed that $\psi''(a_i^T x) \geq 0$, and hence the $n \times d$ matrix $\text{diag}(\psi''(a_i^T x))^{1/2} A$ can be used as a matrix square-root. We return to explore this class of examples in more depth in Section 4.4.1.

4.2.4 Local convergence analysis using strong convexity

Returning now to the general setting, we begin by proving a local convergence guarantee for the sketched Newton updates. In particular, this theorem provides insight into how large the sketch dimension m must be in order to guarantee good local behavior of the sketched Newton algorithm.

Our analysis involves the geometry of the tangent cone of the optimal vector x^* which was first introduced in Section 2. Let us recall the definition in this context: Given a constraint set \mathcal{C} and the minimizer $x^* := \arg \min_{x \in \mathcal{C}} f(x)$ the tangent cone at x^* is given by

$$\mathcal{K} = \{\Delta \in \mathbb{R}^d \mid x^* + t\Delta \in \mathcal{C} \text{ for some } t > 0\}. \quad (4.10)$$

The local analysis to be given in this section involves the *cone-constrained eigenvalues of the Hessian* $\nabla^2 f(x^*)$, defined as

$$\gamma = \inf_{z \in \mathcal{K} \cap \mathcal{S}^{d-1}} \langle z, \nabla^2 f(x^*) z \rangle, \quad \text{and} \quad \beta = \sup_{z \in \mathcal{K} \cap \mathcal{S}^{d-1}} \langle z, \nabla^2 f(x^*) z \rangle. \quad (4.11)$$

In the unconstrained case ($\mathcal{C} = \mathbb{R}^d$), we have $\mathcal{K} = \mathbb{R}^d$, and so that γ and β reduce to the minimum and maximum eigenvalues of the Hessian $\nabla^2 f(x^*)$. In the classical analysis

of Newton’s method, these quantities measure the strong convexity and smoothness parameters of the function f . Note that the condition $\gamma > 0$ is much weaker than strong convexity as it can hold for Hessian matrices that are rank-deficient, as long as the tangent cone \mathcal{K} is suitably small.

Recalling the definition of the Gaussian width from Section 2, our choice of the sketch dimension m depends on the width of the renormalized tangent cone. In particular, for the following theorem, we require it to be lower bounded as

$$m \geq \frac{c}{\epsilon^2} \max_{x \in \mathcal{C}} \mathbb{W}^2(\nabla^2 f(x)^{1/2} \mathcal{K}), \quad (4.12)$$

where $\epsilon \in (0, \frac{\gamma}{9\beta})$ is a user-defined tolerance, and c is a universal constant. Since the Hessian square-root $\nabla^2 f(x)^{1/2}$ has dimensions $n \times d$, this squared Gaussian width is at most $\min\{n, d\}$. This worst-case bound is achieved for an unconstrained problem (in which case $\mathcal{K} = \mathbb{R}^d$), but the Gaussian width can be substantially smaller for constrained problems. For instance, consider an equality constrained problem with affine constraint $Cx = b$. For such a problem, the tangent cone lies within the nullspace of the matrix C —say it is d_C -dimensional. It then follows that the squared Gaussian width (4.12) is also bounded by d_C ; see the example following Theorem 5 for a concrete illustration. Other examples in which the Gaussian width can be substantially smaller include problems involving simplex constraints (portfolio optimization), or ℓ_1 -constraints (sparse regression).

With this set-up, the following theorem is applicable to any twice-differentiable objective f with cone-constrained eigenvalues (γ, β) defined in equation (4.11), and with Hessian that is L -Lipschitz continuous, as defined in equation (4.2).

Theorem 5 (Local convergence of Newton Sketch). *For a given tolerance $\epsilon \in (0, \frac{2\gamma}{9\beta})$, consider the Newton Sketch updates (4.4) based on an initialization x^0 such that $\|x^0 - x^*\|_2 \leq \frac{\gamma}{8L}$, and a sketch dimension m satisfying the lower bound (4.12). Then with probability at least $1 - c_1 N e^{-c_2 m}$, the Euclidean error satisfies the bound*

$$\|x^{t+1} - x^*\|_2 \leq \epsilon \frac{\beta}{\gamma} \|x^t - x^*\|_2 + \frac{4L}{\gamma} \|x^t - x^*\|_2^2, \quad \text{for iterations } t = 0, \dots, N - 1. \quad (4.13)$$

The bound (4.13) shows that when ϵ is small enough—say $\epsilon = \beta/4\gamma$ —then the optimization error $\Delta^t = x^t - x^*$ decays at a linear-quadratic convergence rate. More specifically, the rate is initially quadratic—that is, $\|\Delta^{t+1}\|_2 \approx \frac{4L}{\gamma} \|\Delta^t\|_2^2$ when $\|\Delta^t\|_2$ is large. However, as the iterations progress and $\|\Delta^t\|_2$ becomes substantially less than 1, then the rate becomes linear—meaning that $\|\Delta^{t+1}\|_2 \approx \epsilon \frac{\beta}{\gamma} \|\Delta^t\|_2$ —since the term $\frac{4L}{\gamma} \|\Delta^t\|_2^2$ becomes negligible compared to $\epsilon \frac{\beta}{\gamma} \|\Delta^t\|_2$. Unwrapping the recursion for all

N steps, the linear rate guarantees the conservative error bounds

$$\|x^N - x^*\|_2 \leq \frac{\gamma}{8L} \left(\frac{1}{2} + \epsilon \frac{\beta}{\gamma}\right)^N, \quad \text{and} \quad f(x^N) - f(x^*) \leq \frac{\beta\gamma}{8L} \left(\frac{1}{2} + \epsilon \frac{\beta}{\gamma}\right)^N. \quad (4.14)$$

A notable feature of Theorem 5 is that, depending on the structure of the problem, the linear-quadratic convergence can be obtained using a sketch dimension m that is substantially smaller than $\min\{n, d\}$. As an illustrative example, we performed simulations for some instantiations of a portfolio optimization problem: it is a linearly-constrained quadratic program of the form

$$\min_{\substack{x \geq 0 \\ \sum_{j=1}^d x_j = 1}} \left\{ \frac{1}{2} x^T A^T A x - \langle c, x \rangle \right\}, \quad (4.15)$$

where $A \in \mathbb{R}^{n \times d}$ and $c \in \mathbb{R}^d$ are matrices and vectors that arise from data (see Section 4.4.3 for more details). We used the Newton Sketch to solve different sizes of this problem $d \in \{10, 20, 30, 40, 50, 60\}$, and with $n = d^3$ in each case. Each problem was constructed so that the optimal vector $x^* \in \mathbb{R}^d$ had at most $k = \lceil 2 \log(d) \rceil$ non-zero entries. A calculation of the Gaussian width for this problem (see Section 4.7.3 for the details) shows that it suffices to take a sketch dimension $m \gtrsim s \log d$, and we implemented the algorithm with this choice. Figure 4.2 shows the convergence rate of the Newton Sketch algorithm for the six different problem sizes: consistent with our theory, the sketch dimension $m \ll \min\{d, n\}$ suffices to guarantee linear convergence in all cases.

It is also possible obtain an asymptotically super-linear rate by using an iteration-dependent sketching accuracy $\epsilon = \epsilon(t)$. The following corollary summarizes one such possible guarantee:

Corollary 12. *Consider the Newton Sketch iterates using the iteration-dependent sketching accuracy $\epsilon(t) = \frac{1}{\log(1+t)}$. Then with the same probability as in Theorem 5, we have*

$$\|x^{t+1} - x^*\|_2 \leq \frac{1}{\log(1+t)} \frac{\beta}{\gamma} \|x^t - x^*\|_2 + \frac{4L}{\gamma} \|x^t - x^*\|_2^2,$$

and consequently, super-linear convergence is obtained—namely, $\lim_{t \rightarrow \infty} \frac{\|x^{t+1} - x^*\|_2}{\|x^t - x^*\|_2} = 0$.

Note that the price for this super-linear convergence is that the sketch size is inflated by the factor $\epsilon^{-2}(t) = \log^2(1+t)$, so it is only logarithmic in the iteration number.

4.3 Newton Sketch for self-concordant functions

The analysis and complexity estimates given in the previous section involve the curvature constants (γ, β) and the Lipschitz constant L , which are seldom known in practice. Moreover, as with the analysis of classical Newton method, the theory is local, in that the linear-quadratic convergence takes place once the iterates enter a suitable basin of the origin.

In this section, we seek to obtain global convergence results that do not depend on unknown problem parameters. As in the classical analysis, the appropriate setting in which to seek such results is for self-concordant functions, and using an appropriate form of backtracking line search. We begin by analyzing the unconstrained case, and then discuss extensions to constrained problems with self-concordant barriers. In each case, we show that given a suitable lower bound on the sketch dimension, the sketched Newton updates can be equipped with global convergence guarantees that hold with exponentially high probability. Moreover, the total number of iterations does not depend on any unknown constants such as strong convexity and Lipschitz parameters.

4.3.1 Unconstrained case

In this section, we consider the unconstrained optimization problem $\min_{x \in \mathbb{R}^d} f(x)$, where f is a closed convex self-concordant function that is bounded below. A closed convex function $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is said to be *self-concordant* if

$$|\phi'''(x)| \leq 2 (\phi''(x))^{3/2}. \quad (4.16)$$

This definition can be extended to a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ by imposing this requirement on the univariate functions $\phi_{x,y}(t) := f(x+ty)$, for all choices of x, y in the domain of f . Examples of self-concordant functions include linear and quadratic functions and negative logarithm. Moreover, the property of self-concordance is preserved under addition and affine transformations.

Our main result provide a bound on the total number of Newton Sketch iterations required to obtain a δ -accurate solution without imposing any sort of initialization condition, as was done in our previous analysis. This bound scales proportionally to $\log(1/\delta)$ and inversely in a parameter ν that depends on sketching accuracy $\epsilon \in (0, \frac{1}{4})$ and backtracking parameters (a, b) via

$$\nu = ab \frac{\eta^2}{1 + (\frac{1+\epsilon}{1-\epsilon})\eta} \quad \text{where} \quad \eta = \frac{1}{8} \frac{1 - \frac{1}{2}(\frac{1+\epsilon}{1-\epsilon})^2 - a}{(\frac{1+\epsilon}{1-\epsilon})^3}. \quad (4.17)$$

With this set-up, we have the following guarantee:

Algorithm 1 Unconstrained Newton Sketch with backtracking line search

Require: Starting point x^0 , tolerance $\delta > 0$, (a, b) line-search parameters, sketching matrices $\{S^t\}_{t=0}^\infty \in \mathbb{R}^{m \times n}$.
1: Compute approximate Newton step Δx^t and approximate Newton decrement $\lambda(x)$

$$\begin{aligned}\Delta x^t &:= \arg \min_{\Delta} \langle \nabla f(x^t), \Delta \rangle + \frac{1}{2} \|S^t (\nabla^2 f(x^t))^{1/2} \Delta\|_2^2; \\ \tilde{\lambda}_f(x^t) &:= \nabla f(x^t)^T \Delta x^t.\end{aligned}$$

2: Quit if $\tilde{\lambda}(x^t)^2/2 \leq \delta$.

3: Line search: choose μ : while $f(x^t + \mu \Delta x^t) > f(x^t) + a\mu\lambda(x^t)$, $\mu \leftarrow b\mu$

4: Update: $x^{t+1} = x^t + \mu \Delta x^t$

Ensure: minimizer x^t , optimality gap $\lambda(x^t)$

Theorem 6. *Let f be a strictly convex self-concordant function. Given a sketching matrix $S \in \mathbb{R}^{m \times n}$ with $m = \frac{c_3}{\epsilon^2} \max_{x \in \mathcal{C}} \text{rank}(\nabla^2 f(x))$, the number of total iterations T for obtaining an δ approximate solution in function value via Algorithm 1 is bounded by*

$$N = \frac{f(x^0) - f(x^*)}{\nu} + 0.65 \log_2\left(\frac{1}{16\delta}\right), \quad (4.18)$$

with probability at least $1 - c_1 N e^{-c_2 m}$.

The iteration bound (4.18) shows that the convergence of the Newton Sketch is independent of the properties of the function f and problem parameters, similar to classical Newton's method. Note that for problems with $n > d$, the complexity of each Newton Sketch step is at most $\mathcal{O}(d^3 + nd \log d)$, which is smaller than that of Newton's Method ($\mathcal{O}(nd^2)$), and also smaller than typical first-order optimization methods ($\mathcal{O}(nd)$) whenever $n > d^2$.

4.3.1.1 Rank-deficient Hessians

As stated, Theorem 6 requires the function to be strictly convex. However, by exploiting the affine invariance of the Newton Sketch updates, we can also obtain guarantees of the form (4.18) for the Newton sketch applied to problems with singular Hessians. As a concrete example, given a matrix $A \in \mathbb{R}^{n \times d}$ that is rank-deficient—that is, with $\text{rank}(A) = r < \min\{n, d\}$ —consider a function of the form $f(x) = g(Ax)$, where $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is strictly convex and self-concordant. Due to the rank-deficiency of A , the Hessian of f will also be rank-deficient, so that Theorem 6 does not directly apply. However, suppose that we let $A = U\Sigma V^T$ be the full singular value decomposition of A , where Σ is a diagonal matrix with $\Sigma_{jj} = 0$ for all indices $j > r$. With this notation, define the function $\hat{f}(y) = g(AVy)$, corresponding to the invertible transformation $x = Vy$. We then have

$$\hat{f}(y) = g(U\Sigma y) = g(U\Sigma_{1:r} y_{1:r}),$$

where $y_{1:r} \in \mathbb{R}^r$ denotes the subvector of the first r entries of y . Hence, viewed as a function on \mathbb{R}^r , the transformed function \widehat{f} is strictly convex and self-concordant, so that Theorem 6 can be applied. By the affine invariance property, the Newton Sketch applied to the original function f has the same convergence guarantees (and transformed iterates) as the reduced strictly convex function. Consequently, the sketch size choice $m = \frac{c}{\epsilon^2} \text{rank}(A)$ is sufficient. Note that in many applications, the rank of A can be much smaller than $\min(n, d)$, and so that the Newton Sketch complexity $\mathcal{O}(m^2 d)$ is correspondingly smaller, relative to other schemes that do not exploit the low-rank structure. Some optimization methods can exploit low-rankness when a factorization of the form $A = LR$ is available. However, note that the cost of computing such a low rank factorization scales as $\mathcal{O}(nd^2)$, which dominates the overall complexity of Newton Sketch, including sketching time.

4.3.2 Newton Sketch with self-concordant barriers

We now turn to the more general constrained case. Given a closed, convex self-concordant function $f_0 : \mathbb{R}^d \rightarrow \mathbb{R}$, let \mathcal{C} be a convex subset of \mathbb{R}^d , and consider the constrained optimization problem $\min_{x \in \mathcal{C}} f_0(x)$. If we are given a convex self-concordant barrier function $g(x)$ for the constraint set \mathcal{C} , it is customary to consider the unconstrained and penalized problem

$$\min_{x \in \mathbb{R}^d} \left\{ \underbrace{f_0(x) + g(x)}_{f(x)} \right\},$$

which approximates the original problem. One way in which to solve this unconstrained problem is by sketching the Hessian of both f_0 and g , in which case the theory of the previous section is applicable. However, there are many cases in which the constraints describing \mathcal{C} are relatively simple, and so the Hessian of g is highly-structured. For instance, if the constraint set is the usual simplex (i.e., $x \geq 0$ and $\langle 1, x \rangle \leq 1$), then the Hessian of the associated log barrier function is a diagonal matrix plus a rank one matrix. Other examples include problems for which g has a separable structure; such functions frequently arise as regularizers for ill-posed inverse problems. Examples of such regularizers include ℓ_2 regularization $g(x) = \frac{1}{2} \|x\|_2^2$, graph regularization $g(x) = \frac{1}{2} \sum_{i,j \in E} (x_i - x_j)^2$ induced by an edge set E (e.g., finite differences) and also other differentiable norms $g(x) = \left(\sum_{i=1}^d x_i^p \right)^{1/p}$ for $1 < p < \infty$.

In all such cases, an attractive strategy is to apply a *partial Newton Sketch*, in which we sketch the Hessian term $\nabla^2 f_0(x)$ and retain the exact Hessian $\nabla^2 g(x)$, as in the previously described updates (4.6). More formally, Algorithm 2 provides a summary of the steps, including the choice of the line search parameters. The main result of this section provides a guarantee on this algorithm, assuming that the sequence of sketch dimensions $\{m^t\}_{t=0}^\infty$ is appropriately chosen.

Algorithm 2 Newton Sketch with self-concordant barriers

Require: Starting point x^0 , constraint \mathcal{C} , corresponding barrier function g such that $f = f_0 + g$, tolerance $\delta > 0$, (α, β) line-search parameters, sketching matrices $S^t \in \mathbb{R}^{m \times n}$.

1: Compute approximate Newton step Δx^t and approximate Newton decrement $\tilde{\lambda}_f$.

$$\begin{aligned}\Delta x^t &:= \arg \min_{x^t + \Delta \in \mathcal{C}} \langle \nabla f(x^t), \Delta \rangle + \frac{1}{2} \|S^t (\nabla^2 f_0(x^t))^{1/2} \Delta\|_2^2 + \frac{1}{2} \Delta^T \nabla^2 g(x^t) \Delta; \\ \tilde{\lambda}_f(x^t) &:= \nabla f(x^t)^T \Delta x^t\end{aligned}$$

2: Quit if $\tilde{\lambda}(x^t)^2/2 \leq \delta$.

3: Line search: choose μ : while $f(x^t + \mu \Delta x^t) > f(x^t) + \alpha \mu \lambda(x^t)$, $\mu \leftarrow \beta \mu$.

4: Update: $x^{t+1} = x^t + \mu \Delta x^t$.

Ensure: minimizer x^t , optimality gap $\lambda(x^t)$.

The choice of sketch dimensions depends on the tangent cones defined by the iterates, namely the sets

$$\mathcal{K}^t := \{\Delta \in \mathbb{R}^d \mid x^t + \alpha \Delta \in \mathcal{C} \text{ for some } \alpha > 0\}.$$

For a given sketch accuracy $\epsilon \in (0, 1)$, we require that the sequence of sketch dimensions satisfies the lower bound

$$m^t \geq \frac{c_3}{\epsilon^2} \max_{x \in \mathcal{C}} \mathbb{W}^2(\nabla^2 f(x)^{1/2} \mathcal{K}^t). \quad (4.19)$$

Finally, the reader should recall the parameter ν was defined in equation (4.17), which depends only on the sketching accuracy ϵ and the line search parameters. Given this set-up, we have the following guarantee:

Theorem 7. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex and self-concordant function, and let $g : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a convex and self-concordant barrier for the convex set \mathcal{C} . Suppose that we implement Algorithm 2 with sketch dimensions $\{m^t\}_{t \geq 0}$ satisfying the lower bound (4.19). Then performing*

$$N = \frac{f(x^0) - f(x^*)}{\nu} + 0.65 \log_2 \left(\frac{1}{16\delta} \right) \quad \textit{iterations}$$

suffices to obtain δ -approximate solution in function value with probability at least $1 - c_1 N e^{-c_2 m}$.

Thus, we see that the Newton Sketch method can also be used with self-concordant barrier functions, which considerably extends its scope. In the above theorem, note that we can isolate affine constraints from \mathcal{C} and enforce them at each Newton step. Section 4.4.6 provides a numerical illustration of its performance in this context. As we discuss in the next section, there is a flexibility in choosing the decomposition f_0 and g corresponding to objective and barrier, which enables us to also sketch the constraints.

4.3.3 Sketching with interior point methods

In this section, we discuss the application of Newton Sketch to a form of barrier or interior point methods. In particular we discuss two different strategies and provide rigorous worst-case complexity results when the functions in the objective and constraints are self-concordant. More precisely, let us consider a problem of the form

$$\min_{x \in \mathbb{R}^d} f_0(x) \quad \text{subject to} \quad g_j(x) \leq 0 \quad \text{for } j = 1, \dots, r, \quad (4.20)$$

where f_0 and $\{g_j\}_{j=1}^r$ are twice-differentiable convex functions. We assume that there exists a unique solution x^* to the above problem.

The barrier method for computing x^* is based on solving a sequence of problems of the form

$$\hat{x}(\tau) := \arg \min_{x \in \mathbb{R}^d} \left\{ \tau f_0(x) - \sum_{j=1}^r \log(-g_j(x)) \right\}, \quad (4.21)$$

for increasing values of the parameter $\tau \geq 1$. The family of solutions $\{\hat{x}(\tau)\}_{\tau \geq 1}$ trace out what is known as the central path. A standard bound (e.g., [28]) on the sub-optimality of $\hat{x}(\tau)$ is given by

$$f_0(\hat{x}(\tau)) - f_0(x^*) \leq \frac{r}{\tau}.$$

The barrier method successively updates the penalty parameter τ and also the starting points supplied to Newton's method using previous solutions.

Since Newton's method lies at the heart of the barrier method, we can obtain a fast version by replacing the exact Newton minimization with the Newton Sketch. Algorithm 3 provides a precise description of this strategy. As noted in Step 1, there are two different strategies in dealing with the convex constraints $g_j(x) \leq 0$ for $j = 1, \dots, r$:

- *Full sketch:* Sketch the full Hessian of the objective function (4.21) using Algorithm 1,
- *Partial sketch:* Sketch only the Hessians corresponding to a subset of the functions $\{f_0, g_j, j = 1, \dots, r\}$, and use exact Hessians for the other functions. Apply Algorithm 2.

As shown by our theory, either approach leads to the same convergence guarantees, but the associated computational complexity can vary depending both on how data enters the objective and constraints, as well as the Hessian structure arising from particular functions. The following theorem is an application of the classical results

Algorithm 3 Interior point methods using Newton Sketch

Require: Strictly feasible starting point x^0 , initial parameter τ^0 s.t. $\tau := \tau^0 > 0$, $\mu > 1$, tolerance $\delta > 0$.

- 1: Centering step: Compute $\hat{x}(\tau)$ by Newton Sketch with backtracking line-search initialized at x using Algorithm 1 or Algorithm 2.
- 2: Update $x := \hat{x}(\tau)$.
- 3: Quit if $r/\tau \leq \delta$.
- 4: Increase τ by $\tau := \mu\tau$.

Ensure: minimizer $\hat{x}(\tau)$.

on the barrier method tailored for Newton Sketch using any of the above strategies (e.g., see Boyd and Vandenberghe [28]). As before, the key parameter ν was defined in Theorem 6.

Theorem 8 (Newton Sketch complexity for interior point methods). *For a given target accuracy $\delta \in (0, 1)$ and any $\mu > 1$, the total number of Newton Sketch iterations required to obtain a δ -accurate solution using Algorithm 3 is at most*

$$\frac{\log(r/(\tau^0\delta))}{\log \mu} \left(\frac{r(\mu - 1 - \log \mu)}{\nu} + 0.65 \log_2\left(\frac{1}{16\delta}\right) \right). \quad (4.22)$$

If the parameter μ is set to minimize the above upper-bound, the choice $\mu = 1 + \frac{1}{r}$ yields $\mathcal{O}(\sqrt{r})$ iterations. However, this “optimal” choice is typically not used in practice when applying the standard Newton method; instead, it is common to use a fixed value of $\mu \in [2, 100]$. In experiments, experience suggests that the number of Newton iterations needed is a constant independent of r and other parameters. Theorem 8 allows us to obtain faster interior point solvers with rigorous worst-case complexity results. We show different applications of Algorithm 3 in the following section.

4.4 Applications and numerical results

In this section, we discuss some applications of the Newton Sketch to different optimization problems. In particular, we show various forms of Hessian structure that arise in applications, and how the Newton sketch can be computed. When the objective and/or the constraints contain more than one term, the barrier method with Newton Sketch has some flexibility in sketching. We discuss the choices of partial Hessian sketching strategy in the barrier method. It is also possible to apply the sketch in the primal or dual form, and we provide illustrations of both strategies here.

4.4.1 Estimation in generalized linear models

Recall the problem of (constrained) maximum likelihood estimation for a generalized linear model, as previously introduced in Example 4.2.3. It leads to the family of optimization problems (4.9): here $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is a given convex function arising from the probabilistic model, and $\mathcal{C} \subseteq \mathbb{R}^d$ is a closed convex set that is used to enforce a certain type of structure in the solution. Popular choices of such constraints include ℓ_1 -balls (for enforcing sparsity in a vector), nuclear norms (for enforcing low-rank structure in a matrix), and other non-differentiable semi-norms based on total variation (e.g., $\sum_{j=1}^{d-1} |x_{j+1} - x_j|$), useful for enforcing smoothness or clustering constraints.

Suppose that we apply the Newton Sketch algorithm to the optimization problem (4.9). Given the current iterate x^t , computing the next iterate x^{t+1} requires solving the constrained quadratic program

$$\min_{x \in \mathcal{C}} \left\{ \frac{1}{2} \|S \text{diag}(\psi''(\langle a_i, x^t \rangle, y_i))^{1/2} A(x - x^t)\|_2^2 + \sum_{i=1}^n \langle x, \psi'(\langle a_i, x^t \rangle, y_i) \rangle \right\}. \quad (4.23)$$

When the constraint \mathcal{C} is a scaled version of the ℓ_1 -ball—that is, $\mathcal{C} = \{x \in \mathbb{R}^d \mid \|x\|_1 \leq R\}$ for some radius $R > 0$ —the convex program (4.23) is an instance of the Lasso program [134], for which there is a very large body of work. For small values of R , where the cardinality of the solution x is very small, an effective strategy is to apply a homotopy type algorithm, also known as LARS [57, 66], which solves the optimality conditions starting from $R = 0$. For other sets \mathcal{C} , another popular choice is projected gradient descent, which is efficient when projection onto \mathcal{C} is computationally simple.

Focusing on the ℓ_1 -constrained case, let us consider the problem of choosing a suitable sketch dimension m . Our choice involves the ℓ_1 -restricted minimal eigenvalue of the data matrix A , which is defined by (2.13) in Section 2. Note that we are always guaranteed that $\gamma_k^-(A) \geq \lambda_{\min}(A^T A)$. Our result also involves certain quantities that depend on the function ψ , namely

$$\psi''_{\min} := \min_{x \in \mathcal{C}} \min_{i=1, \dots, n} \psi''(\langle a_i, x \rangle, y_i), \quad \text{and} \quad \psi''_{\max} := \max_{x \in \mathcal{C}} \max_{i=1, \dots, n} \psi''(\langle a_i, x \rangle, y_i),$$

where $a_i \in \mathbb{R}^d$ is the i^{th} row of A . With this set-up, supposing that the optimal solution x^* has cardinality at most $\|x^*\|_0 \leq k$, then it can be shown (see Lemma 25 in Section 4.7.3) that it suffices to take a sketch size

$$m = c_0 \frac{\psi''_{\max} \max_{j=1, \dots, d} \|A_j\|_2^2}{\psi''_{\min} \gamma_k^-(A)} k \log d, \quad (4.24)$$

where c_0 is a universal constant. Let us consider some examples to illustrate:

- Least-Squares regression: $\psi(u) = \frac{1}{2}u^2$, $\psi''(u) = 1$ and $\psi''_{\min} = \psi''_{\max} = 1$.

- Poisson regression: $\psi(u) = e^u$, $\psi''(u) = e^u$ and $\frac{\psi''_{\max}}{\psi''_{\min}} = \frac{e^{RA_{\max}}}{e^{-RA_{\min}}}$
- Logistic regression: $\psi(u) = \log(1 + e^u)$, $\psi''(u) = \frac{e^u}{(e^u+1)^2}$ and $\frac{\psi''_{\max}}{\psi''_{\min}} = \frac{e^{RA_{\min}}}{e^{-RA_{\max}}} \frac{(e^{-RA_{\max}}+1)^2}{(e^{RA_{\min}}+1)^2}$,

where $A_{\max} := \max_{i=1,\dots,n} \|a_i\|_{\infty}$, and $A_{\min} := \min_{i=1,\dots,n} \|a_i\|_{\infty}$.

For typical distributions of the data matrices, the sketch size choice given in equation (4.24) scales as $\mathcal{O}(k \log d)$. As an example, consider data matrices $A \in \mathbb{R}^{n \times d}$ where each row is independently sampled from a sub-Gaussian distribution with parameter one (see equation (1.1)). Then standard results on random matrices [140] show that $\gamma_k^-(A) > 1/2$ with high probability as long as $n > c_1 k \log d$ for a sufficiently large constant c_1 . In addition, we have $\max_{j=1,\dots,d} \|A_j\|_2^2 = \mathcal{O}(n)$, as well as $\frac{\psi''_{\max}}{\psi''_{\min}} = \mathcal{O}(\log(n))$. For such problems, the per iteration complexity of Newton Sketch update scales as $\mathcal{O}(k^2 d \log^2(d))$ using standard Lasso solvers (e.g., [75]) or as $\mathcal{O}(kd \log(d))$ using projected gradient descent. Both of these scalings are substantially smaller than conventional algorithms that fail to exploit the small intrinsic dimension of the tangent cone.

4.4.2 Semidefinite programs

The Newton Sketch can also be applied to semidefinite programs. As one illustration, let us consider a metric learning problem studied in machine learning. Suppose that we are given d -dimensional feature vectors $\{a_i\}_{i=1}^n$ and a collection of $\binom{n}{2}$ binary indicator variables $y_{ij} \in \{-1, +1\}^n$ given by

$$y_{ij} = \begin{cases} +1 & \text{if } a_i \text{ and } a_j \text{ belong to the same class} \\ -1 & \text{otherwise,} \end{cases}$$

defined for all distinct indices $i, j \in \{1, \dots, n\}$. The task is to estimate a positive semidefinite matrix X such that the semi-norm $\|(a_i - a_j)\|_X := \sqrt{\langle a_i - a_j, X(a_i - a_j) \rangle}$ is a good predictor of whether or not vectors i and j belong to the same class. Using the least-squares loss, one way in which to do so is by solving the semidefinite program (SDP)

$$\min_{X \succeq 0} \left\{ \sum_{i \neq j}^{\binom{n}{2}} (\langle X, (a_i - a_j)(a_i - a_j)^T \rangle - y_{ij})^2 + \lambda \text{trace}(X) \right\}.$$

Here the term $\text{trace}(X)$, along with its multiplicative pre-factor $\lambda > 0$ that can be adjusted by the user, is a regularization term for encouraging a relatively low-rank

solution. Using the standard self-concordant barrier $X \mapsto \log \det(X)$ for the PSD cone, the barrier method involves solving a sequence of sub-problems of the form

$$\min_{X \in \mathbb{R}^{d \times d}} \left\{ \underbrace{\tau \sum_{i=1}^n (\langle X, a_i a_i^T \rangle - y_i)^2 + \tau \lambda \text{trace} X - \log \det(X)}_{f(\text{vec}(X))} \right\}.$$

Now the Hessian of the function $\text{vec}(X) \mapsto f(\text{vec}(X))$ is a $d^2 \times d^2$ matrix given by

$$\nabla^2 f(\text{vec}(X)) = \tau \sum_{i \neq j}^{\binom{n}{2}} \text{vec}(A_{ij}) \text{vec}(A_{ij})^T + X^{-1} \otimes X^{-1},$$

where $A_{ij} := (a_i - a_j)(a_i - a_j)^T$. Then we can apply the barrier method with partial Hessian sketch on the first term, $\{S_{ij} \text{vec}(A_{ij})\}_{i \neq j}$ and exact Hessian for the second term. Since the vectorized decision variable is $\text{vec}(X) \in \mathbb{R}^{d^2}$ the complexity of Newton Sketch is $\mathcal{O}(m^2 d^2)$ while the complexity of a classical SDP interior-point solver is $\mathcal{O}(nd^4)$ in practice.

4.4.3 Portfolio optimization and SVMs

Here we consider the Markowitz formulation of the portfolio optimization problem [91]. The objective is to find a vector $x \in \mathbb{R}^d$ belonging to the unit simplex, corresponding to non-negative weights associated with each of d possible assets, so as to maximize the expected return minus a coefficient times the variance of the return. Letting $\mu \in \mathbb{R}^d$ denote a vector corresponding to mean return of the assets, and we let $\Sigma \in \mathbb{R}^{d \times d}$ be a symmetric, positive semidefinite matrix, covariance of the returns. The optimization problem is given by

$$\max_{x \geq 0, \sum_{j=1}^d x_j \leq 1} \left\{ \langle \mu, x \rangle - \lambda \frac{1}{2} x^T \Sigma x \right\}. \quad (4.25)$$

The covariance of returns is often estimated from past stock data via an empirical covariance matrix of the form $\Sigma = A^T A$; here columns of A are time series corresponding to assets normalized by \sqrt{n} , where n is the length of the observation window.

The barrier method can be used solve the above problem by solving penalized problems of the form

$$\min_{x \in \mathbb{R}^d} \left\{ \underbrace{-\tau \mu^T x + \tau \lambda \frac{1}{2} x^T A^T A x - \sum_{i=1}^d \log(\langle e_i, x \rangle) - \log(1 - \langle 1, x \rangle)}_{f(x)} \right\},$$

where $e_i \in \mathbb{R}^d$ is the i^{th} element of the canonical basis and $\mathbf{1}$ is a row vector of all-ones. Then the Hessian of the above barrier penalized formulation can be written as

$$\nabla^2 f(x) = \tau\lambda A^T A + (\text{diag}\{x_i^2\}_{i=1}^d)^{-1} + \mathbf{1}\mathbf{1}^T.$$

Consequently, we can sketch the data dependent part of the Hessian via $\tau\lambda SA$ which has at most rank m and keep the remaining terms in the Hessian exact. Since the matrix $\mathbf{1}\mathbf{1}^T$ is rank one, the resulting sketched estimate is therefore diagonal plus rank $(m+1)$ where the matrix inversion lemma [62] can be applied for efficient computation of the Newton Sketch update. Therefore, as long as $m \leq d$, the complexity per iteration scales as $\mathcal{O}(md^2)$, which is cheaper than the $\mathcal{O}(nd^2)$ per step complexity associated with classical interior point methods. We also note that support vector machine classification problems with squared hinge loss also has the same form as in equation (4.25), so that the same strategy can be applied.

4.4.4 Unconstrained logistic regression with $d \ll n$

Let us now turn to some numerical comparisons of the Newton Sketch with other popular optimization methods for large-scale instances of logistic regression. More specifically, we generated a data matrix $A \in \mathbb{R}^{n \times d}$ with $d = 100$ features and $n = 65536$ observations. Each row $a_i \in \mathbb{R}^d$ was generated from the d -variate Gaussian distribution $N(0, \Sigma)$ where the covariance matrix Σ has 1 on diagonals and ρ on off-diagonals. As shown in Figures 4.3 and 4.3, the convergence of the algorithm per iteration is very similar to Newton’s method. Besides the original Newton’s method, the other algorithms compared are

- Gradient Descent (GD) with backtracking line search
- Stochastic Average Gradient (SAG) with line search
- Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) (MATLAB R2015a implementation)
- Truncated Newton’s Method (*trunNewt*)

We ran the Newton Sketch algorithm with ROS sketch and sketch size $m = 4d$ and plot iterates over 10 independent trials. The gradient method is using backtracking line search. For the Truncated Newton’s Method, we first performed experiments by setting the maximum CG iteration number in the range $\{\log(d), 2\log(d), 3\log(d), \dots, 10\log(d)\}$, and then also implemented the residual stopping rule with accuracy $1/t$ as suggested in [48]. The best choice among these parameters is shown as *trunNewt* in the plots. All algorithms are implemented in MATLAB

(R2015a). In the plots, each iteration of the SAG algorithm corresponds to a pass over the data, which is of comparable complexity to a single iteration of GD. In order to keep the plots relatively uncluttered, we have excluded Stochastic Gradient Descent since it is dominated by another stochastic first-order method (SAG), and Accelerated Gradient Method as it is quite similar to Gradient Descent. In Figure 4.3, panels (a) and (b) show the case with no correlation ($\rho = 0$), panels (c) and (d) show the case with correlation $\rho = 0.5$ and panels (e) and (f) shows the case with correlation $\rho = 0.9$. Plots on the left in Figure 4.3—that is panels (a), (c) and (e)—show the log duality gap versus the number of iterations: as expected, on this scale, the classical form of Newton’s method is the fastest. However, when the log optimality gap is plotted versus the wall-clock time (right-side panels (b), (d) and (f)), we now see that the Newton sketch is the fastest.

On the other hand, Figure 4.4 reveals the sensitivity of first order methods to data conditioning. For these experiments, we generated a feature matrix A with $d = 100$ features and $n = 65536$ observations where each row $a_i \in \mathbb{R}^d$ was generated from the Student’s t-distribution with covariance Σ . The covariance matrix Σ has 1 on diagonals and ρ on off-diagonals. In Figure 4.4, panels (a) and (b) show the case with no correlation ($\rho = 0$), panels (c) and (d) show the case with correlation $\rho = 0.5$ and panels (e) and (f) shows the case with correlation $\rho = 0.9$. As it can be seen in Figure 4.4, SAG and GD perform quite poor. As predicted by theory, Newton Sketch performs well even with high correlations and non-Gaussian data while first order algorithms perform poorly.

4.4.5 ℓ_1 -constrained logistic regression and data conditioning

Next we provide some numerical comparisons of Newton Sketch, Newton’s Method and Projected Gradient Descent when applied to an ℓ_1 -constrained form of logistic regression. More specifically, we first generate a feature matrix $A \in \mathbb{R}^{n \times d}$ based on $d = 100$ features and $n = 1000$ observations. Each row $a_i \in \mathbb{R}^d$ is drawn from the d -variate Gaussian distribution $N(0, \Sigma)$; the covariance matrix has entries of the form $\Sigma_{ij} = 2|\rho|^{i-j}$, where $\rho \in [0, 1)$ is a parameter controlling the correlation, and hence the condition number of the data. For 10 different values of ρ we solved the ℓ_1 -constrained problem ($\|x\|_1 \leq 0.1$), performing 200 independent trials (regenerating the data and sketching matrices randomly each time). The Newton and sketched Newton steps are solved exactly using the homotopy algorithm—that is, the Lasso modification of the LARS updates [110, 57]. The homotopy method is very effective when the solution is very sparse. The ROS sketch with a sketch size of $m = \lceil 4 \times 10 \log d \rceil$ is used where 10 is the estimated cardinality of solution. As shown in Figure 4.5, Newton Sketch converges in about 6 (± 2) iterations independent of data conditioning while the exact Newton’s method converges in 3 (± 1) iterations. However the number of iterations needed for projected gradient with line search increases steeply as ρ increases. Note

that, ignoring logarithmic terms, the projected gradient and Newton Sketch have similar computational complexity ($\mathcal{O}(nd)$) per iteration while the Newton's method has higher computational complexity ($\mathcal{O}(nd^2)$).

4.4.6 A dual example: Lasso with $d \gg n$

The regularized Lasso problem takes the form $\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{2} \|Ax - y\|_2^2 + \lambda \|x\|_1 \right\}$, where $\lambda > 0$ is a user-specified regularization parameter. In this section, we consider efficient sketching strategies for this class of problems in the regime $d \gg n$. In particular, let us consider the corresponding dual program, given by

$$\max_{\|A^T w\|_\infty \leq \lambda} \left\{ -\frac{1}{2} \|y - w\|_2^2 \right\}.$$

By construction, the number of constraints d in the dual program is larger than the number of optimization variables n . If we apply the barrier method to solve this dual formulation, then we need to solve a sequence of problems of the form

$$\min_{w \in \mathbb{R}^n} \left\{ \underbrace{\tau \|y - w\|_2^2 - \sum_{j=1}^d \log(\lambda - \langle A_j, w \rangle) - \sum_{j=1}^d \log(\lambda + \langle A_j, w \rangle)}_{f(x)} \right\},$$

where $A_j \in \mathbb{R}^n$ denotes the j^{th} column of A . The Hessian of the above barrier penalized formulation can be written as

$$\nabla^2 f(w) = \tau I_n + \text{Adiag} \left(\frac{1}{(\lambda - \langle A_j, w \rangle)^2} \right) A^T + \text{Adiag} \left(\frac{1}{(\lambda + \langle A_j, w \rangle)^2} \right) A^T,$$

Consequently we can keep the first term in the Hessian, τI exact and apply partial sketching to the Hessians of the last two terms via

$$S \text{diag} \left(\frac{1}{|\lambda - \langle A_j, w \rangle|} + \frac{1}{|\lambda + \langle A_j, w \rangle|} \right) A^T.$$

Since the partially sketched Hessian is of the form $tI_n + VV^T$, where V is rank at most m , we can use matrix inversion lemma for efficiently calculating Newton Sketch updates. The complexity of the above strategy for $d > n$ is $\mathcal{O}(nm^2)$, where m is at most n , whereas traditional interior point solvers are typically $\mathcal{O}(dn^2)$ per iteration.

In order to test this algorithm, we generated a feature matrix $A \in \mathbb{R}^{n \times d}$ with $d = 4096$ features and $n = 50$ observations. Each row $a_i \in \mathbb{R}^d$ was generated from the multivariate Gaussian distribution $N(0, \Sigma)$ with $\Sigma_{ij} = 2 * |0.5|^{i-j}$. For a given problem instance, we ran 10 independent trials of the sketched barrier method with

$m = 4d$ and ROS sketch, and compared the results to the original barrier method. Figure 4.6 shows the the duality gap versus iteration number (top panel) and versus the wall-clock time (bottom panel) for the original barrier method (blue) and sketched barrier method (red): although the sketched algorithm requires more iterations, these iterations are cheaper, leading to a smaller wall-clock time. This point is reinforced by Figure 4.7, where we plot the wall-clock time required to reach a duality gap of 10^{-6} versus the number of features n in problem families of increasing size. Note that the sketched barrier method outperforms the original barrier method, with significantly less computation time for obtaining similar accuracy.

4.5 Proofs of main results

We now turn to the proofs of our theorems, with more technical details deferred to later sections.

4.5.1 Proof of Theorem 5

For any $x \in \text{dom}(f)$, and $r \in \mathbb{R}^d \setminus \{0\}$, we define the following pair of random variables

$$\begin{aligned} Z_u(S; x, r) &:= \sup_{w \in \nabla^2 f(x)^{1/2} \mathcal{K} \cap \mathcal{S}^{n-1}} \left\langle w, (S^T S - I) \frac{r}{\|r\|_2} \right\rangle, \\ Z_\ell(S; x) &:= \inf_{w \in \nabla^2 f(x)^{1/2} \mathcal{K} \cap \mathcal{S}^{n-1}} \|Sw\|_2^2. \end{aligned}$$

Of particular interest to us in analyzing the sketched Newton updates are the sequence of random variables

$$Z_u^t := Z_u(S^t; x^t, \nabla^2 f(x^t)^{1/2} \Delta^t), \quad \text{and} \quad Z_\ell^t := Z_\ell(S^t; x^t).$$

For a given tolerance parameter $\epsilon \in (0, \frac{2\gamma}{9\beta}]$, we define the “good event”

$$\mathcal{E}^t := \left\{ Z_1(A\mathcal{K})^t \leq \frac{\epsilon}{2}, \text{ and } Z_2(A\mathcal{K})^t \geq 1 - \epsilon \right\}. \quad (4.26)$$

The following result gives sufficient conditions on the sketch dimension for this event to hold with high probability:

Lemma 17 (Sufficient conditions on sketch dimension [114]). *(a) For sub-Gaussian sketch matrices, given a sketch size $m > \frac{c_0}{\epsilon^2} \max_{x \in \mathcal{C}} \mathbb{W}^2(\nabla^2 f(x)^{1/2} \mathcal{K})$, we have*

$$\mathbb{P}[\mathcal{E}^t] \geq 1 - c_1 e^{-c_2 m \epsilon^2}. \quad (4.27)$$

(b) For randomized orthogonal system (ROS) sketches and JL embeddings, over the class of self-bounding cones, given a sketch size $m > \frac{c_0 \log^4 n}{\epsilon^2} \max_{x \in \mathcal{C}} \mathbb{W}^2(\nabla^2 f(x)^{1/2} \mathcal{K})$, we have

$$\mathbb{P}[\mathcal{E}^t] \geq 1 - c_1 e^{-c_2 \frac{m \epsilon^2}{\log^4 n}}. \quad (4.28)$$

The remainder of our proof is based on showing that given any initialization x^0 such that $\|x^0 - x^*\|_2 \leq \frac{\gamma}{8L}$, then whenever the event $\cap_{t=1}^N \mathcal{E}^t$ holds, the error vectors $\Delta^t = x^t - x^*$ satisfy the recursion

$$\|\Delta^{t+1}\|_2 \leq \frac{Z_2(A\mathcal{K})^t}{Z_1(A\mathcal{K})^t} \frac{\beta}{7\gamma} \|\Delta^t\|_2 + \frac{1}{Z_1(A\mathcal{K})^t} \frac{8L}{7\gamma} \|\Delta^t\|_2^2 \quad \text{for all } t = 0, 1, \dots, N-1. \quad (4.29)$$

Since we have $\frac{Z_2(A\mathcal{K})^t}{Z_1(A\mathcal{K})^t} \leq \epsilon$ and $\frac{1}{Z_1(A\mathcal{K})^t} \leq 2$ whenever the event $\cap_{t=1}^N \mathcal{E}^t$ holds, the bound (4.13) stated in the theorem then follows. Applying Lemma 17 yields the stated probability bound.

Accordingly, it remains to prove the recursion (4.29), and we do so via a basic inequality argument. Recall the function $x \mapsto \Phi(x; S^t)$ that underlies the sketch Newton update (4.4) in moving from iterate x^t to iterate x^{t+1} . Since the vectors x^{t+1} and x^* are optimal and feasible, respectively, for the constrained optimization problem, the error vector $\Delta^{t+1} := x^{t+1} - x^*$ satisfies the inequality $\langle \nabla \Phi(x^{t+1}; S^t), -\Delta^{t+1} \rangle \geq 0$, or equivalently

$$\langle (S^t \nabla^2 f(x^t)^{1/2})^T S^t \nabla^2 f(x^t)^{1/2} (\Delta^{t+1} - \Delta^t) + \nabla f(x^t), -\Delta^{t+1} \rangle \geq 0.$$

Similarly, since x^* and x^{t+1} are optimal and feasible, respectively, for the minimization of f , we have

$$\langle f(x^*), \Delta^{t+1} \rangle \geq 0.$$

Adding these two inequalities and re-arranging leads to the *basic inequality*

$$\underbrace{\|S^t \nabla^2 f(x^t)^{1/2} \Delta^{t+1}\|_2^2}_{\text{LHS}} \leq \underbrace{\langle S^t \nabla^2 f(x^t)^{1/2} \Delta^{t+1}, S^t \nabla^2 f(x^t)^{1/2} \Delta^t \rangle - \langle \nabla f(x^t) - \nabla f(x^*), \Delta^{t+1} \rangle}_{\text{RHS}} \quad (4.30) \quad \blacksquare$$

This inequality forms the core of our argument: in particular, the bulk of our proof is devoted to establishing the following bounds:

Lemma 18 (Upper and lower bounds). *We have*

$$LHS \geq Z_1(A\mathcal{K})^t \left\{ \gamma - L\|\Delta^t\|_2 \right\} \|\Delta^{t+1}\|_2^2, \quad \text{and} \quad (4.31a)$$

$$RHS \leq Z_2(A\mathcal{K})^t \left\{ \beta + L\|\Delta^t\|_2 \right\} \|\Delta^t\|_2 \|\Delta^{t+1}\|_2 + L\|\Delta^t\|_2^2 \|\Delta^{t+1}\|_2. \quad (4.31b)$$

Taking this lemma as given for the moment, let us complete the proof of the recursion (4.29). Our proof consists of two steps:

- we first show that bound (4.29) holds for Δ^{t+1} whenever $\|\Delta^t\|_2 \leq \frac{\gamma}{8L}$.
- we then show by induction that, conditioned on the event $\cap_{t=1}^N \mathcal{E}^t$, the bound $\|\Delta^t\|_2 \leq \frac{\gamma}{8L}$ holds for all iterations $t = 0, 1, \dots, N$.

Assuming that $\|\Delta^t\|_2 \leq \frac{\gamma}{8L}$, then our basic inequality (4.30) combined with Lemma 18 implies that

$$\|\Delta^{t+1}\|_2 \leq \frac{Z_2(A\mathcal{K})^t \{ \beta + L\|\Delta^t\|_2 \}}{Z_1(A\mathcal{K})^t \{ \gamma - L\|\Delta^t\|_2 \}} \|\Delta^t\|_2 + \frac{L}{Z_1(A\mathcal{K})^t \{ \gamma - L\|\Delta^t\|_2 \}} \|\Delta^t\|_2^2.$$

We have $L\|\Delta^t\|_2 \leq \gamma/8 \leq \beta/8$, and $(\gamma - L\|\Delta^t\|_2)^{-1} \leq \frac{8}{7\gamma}$ hence

$$\|\Delta^{t+1}\|_2 \leq \frac{Z_2(A\mathcal{K})^t}{Z_1(A\mathcal{K})^t} \frac{9\beta}{7\gamma} \|\Delta^t\|_2 + \frac{1}{Z_1(A\mathcal{K})^t} \frac{8L}{7\gamma} \|\Delta^t\|_2^2, \quad (4.32)$$

thereby verifying the claim (4.29).

Now we need to check for any iteration t , the bound $\|\Delta^t\|_2 \leq \frac{\gamma}{8L}$ holds. We do so by induction. The base case is trivial since $\|\Delta^0\|_2 \leq \frac{\gamma}{8L}$ by assumption. Supposing that the bound holds at time t , by our argument above, inequality (4.32) holds, and hence

$$\|\Delta^{t+1}\|_2 \leq \frac{9}{56} \frac{\beta Z_2(A\mathcal{K})^t}{L Z_1(A\mathcal{K})^t} + \frac{16L}{7\gamma Z_1(A\mathcal{K})^t} \frac{\gamma^2}{64L^2} = \frac{Z_2(A\mathcal{K})^t}{Z_1(A\mathcal{K})^t} \frac{9}{28} \frac{\beta}{L} + \frac{1}{Z_1(A\mathcal{K})^t} \frac{1}{28} \frac{\gamma}{L}.$$

Whenever \mathcal{E}^t holds, we have $\frac{Z_2(A\mathcal{K})^t}{Z_1(A\mathcal{K})^t} \leq \frac{2\gamma}{9\beta}$ and $\frac{1}{Z_1(A\mathcal{K})^t} \leq \frac{1}{2}$, whence $\|\Delta^{t+1}\|_2 \leq \left(\frac{1}{28} + \frac{1}{14} \right) \frac{\gamma}{L} \leq \frac{\gamma}{8L}$, as claimed.

The final remaining detail is to prove Lemma 18.

4.5.1.0.1 Proof of Lemma 18: We first prove the lower bound (4.31a) on the LHS. Since $\nabla^2 f(x^t)^{1/2} \Delta^{t+1} \in \nabla^2 f(x^t)^{1/2} \mathcal{K}$, the definition of $Z_1(A\mathcal{K})^t$ ensures that

$$\begin{aligned} \text{LHS} &= \|S^t \nabla^2 f(x^t)^{1/2} \Delta^{t+1}\|_2^2 \geq Z_1(A\mathcal{K})^t \|\nabla^2 f(x^t)^{1/2} \Delta^{t+1}\|_2^2 \\ &\stackrel{(i)}{=} Z_1(A\mathcal{K})^t (\Delta^{t+1})^T \nabla^2 f(x^t) \Delta^{t+1} \\ &= Z_1(A\mathcal{K})^t \{(\Delta^{t+1})^T \nabla^2 f(x^*) \Delta^{t+1} + (\Delta^{t+1})^T (\nabla^2 f(x^t) - \nabla^2 f(x^*)) \Delta^{t+1}\} \\ &\stackrel{(ii)}{\geq} Z_1(A\mathcal{K})^t \{\gamma \|\Delta^{t+1}\|_2^2 - L \|\Delta^{t+1}\|_2^2 \|\Delta^t\|_2\} \end{aligned}$$

where step (i) follows since $(\nabla^2 f(x^t)^{1/2})^T \nabla^2 f(x^t)^{1/2} = \nabla^2 f(x^t)$, and step (ii) follows from the definitions of γ and L .

Next we prove the upper bound (4.31b) on the RHS. Throughout this proof, we write S instead of S^t so as to simplify notation. By the integral form of Taylor series, we have

$$\begin{aligned} \text{RHS} &= \int_0^1 (\Delta^t)^T [(S \nabla^2 f(x^t)^{1/2})^T S \nabla^2 f(x^t)^{1/2} - \nabla^2 f(x^t + u(x^* - x^t))] \Delta^{t+1} du \\ &= T_1 + T_2 \end{aligned}$$

where

$$T_1 := (\Delta^t)^T [(S \nabla^2 f(x^t)^{1/2})^T S \nabla^2 f(x^t)^{1/2} - \nabla^2 f(x^t)] \Delta^{t+1}, \quad \text{and} \quad (4.33a)$$

$$T_2 := \int_0^1 (\Delta^t)^T [-\nabla^2 f(x^t + u(x^* - x^t)) + \nabla^2 f(x^t)] \Delta^{t+1} du. \quad (4.33b)$$

Here the decomposition into T_1 and T_2 follows by adding and subtracting the term $(\Delta^t)^T \nabla^2 f(x^t) \Delta^{t+1}$.

We begin by upper bounding the term T_1 . By the definition of $Z_2(A\mathcal{K})^t$, we have

$$T_1 \leq \left| (\Delta^t)^T Q^T(x^t) \left[\frac{S^T S}{m} - I \right] \nabla^2 f(x^t)^{1/2} \Delta^{t+1} \right| \leq Z_2 \|\nabla^2 f(x^t)^{1/2} \Delta^t\|_2 \|\nabla^2 f(x^t)^{1/2} \Delta^{t+1}\|_2. \quad \blacksquare$$

By adding and subtracting terms, we have

$$\begin{aligned} \|\nabla^2 f(x^t)^{1/2} \Delta^t\|_2^2 &= (\Delta^t)^T \nabla^2 f(x^t) \Delta^t = (\Delta^t)^T \nabla^2 f(x^*) \Delta^t + (\Delta^t)^T [\nabla^2 f(x^t) - \nabla^2 f(x^*)] \Delta^t \\ &\leq \beta \|\Delta^t\|_2^2 + L \|\Delta^t\|_2^3 = \|\Delta^t\|_2^2 (\beta + L \|\Delta^t\|_2), \quad \blacksquare \end{aligned}$$

where the final step follows from the definitions of β and L . A similar argument yields

$$\|\nabla^2 f(x^t)^{1/2} \Delta^{t+1}\|_2^2 \leq \|\Delta^{t+1}\|_2^2 (\beta + L \|\Delta^t\|_2).$$

Overall, we have shown that

$$T_1 \leq Z_2(A\mathcal{K})^t (\beta + L \|\Delta^t\|_2) \|\Delta^t\|_2 \|\Delta^{t+1}\|_2. \quad (4.34)$$

Turning to the quantity T_2 , we have

$$\begin{aligned} T_2 &\leq \left\{ \int_0^1 \sup_{v, \tilde{v} \in \mathcal{K} \cap \mathcal{S}^{d-1}} |v^T [\nabla^2 f(x^t + u(x^* - x^t)) - \nabla^2 f(x^t)] \tilde{v}| du \right\} \|\Delta^t\|_2 \|\Delta^{t+1}\|_2 \\ &\leq L \|\Delta^t\|_2^2 \|\Delta^{t+1}\|_2, \end{aligned} \quad (4.35)$$

where the final step uses the local Lipschitz property again. Combining the bound (4.34) with the bound (4.35) yields the bound (4.31b) on the RHS.

4.5.2 Proof of Theorem 6

Recall that in this case, we assume that f is a self-concordant strictly convex function. We adopt the following notation and conventions from the book [107]. For a given $x \in \mathbb{R}^d$, we define the pair of dual norms

$$\|u\|_x := \langle \nabla^2 f(x) u, u \rangle^{1/2}, \quad \text{and} \quad \|v\|_x^* := \langle \nabla^2 f(x)^{-1} v, v \rangle^{1/2},$$

as well as the Newton decrement

$$\lambda_f(x) = \langle \nabla^2 f(x)^{-1} \nabla f(x), \nabla f(x) \rangle^{1/2} = \|\nabla^2 f(x)^{-1} \nabla f(x)\|_x = \|\nabla^2 f(x)^{-1/2} \nabla f(x)\|_2. \blacksquare$$

Note that $\nabla^2 f(x)^{-1}$ is well-defined for strictly convex self-concordant functions. In terms of this notation, the exact Newton update is given by $x \mapsto x_{\text{NE}} := x + v$, where

$$v_{\text{NE}} := \arg \min_{z \in \mathcal{C} - x} \underbrace{\left\{ \frac{1}{2} \|\nabla^2 f(x)^{1/2} z\|_2^2 + \langle z, \nabla f(x) \rangle \right\}}_{\Phi(z)}, \quad (4.36)$$

whereas the Newton Sketch update is given by $x \mapsto x_{\text{NSK}} := x + v_{\text{NSK}}$, where

$$v_{\text{NSK}} := \arg \min_{z \in \mathcal{C} - x} \left\{ \frac{1}{2} \|S \nabla^2 f(x)^{1/2} z\|_2^2 + \langle z, \nabla f(x) \rangle \right\}. \quad (4.37)$$

The proof of Theorem 6 given in this section involves the unconstrained case ($\mathcal{C} = \mathbb{R}^d$), whereas the proofs of later theorems involve the more general constrained case. In the unconstrained case, the two updates take the simpler forms

$$x_{\text{NE}} = x - (\nabla^2 f(x))^{-1} \nabla f(x), \quad \text{and} \quad x_{\text{NSK}} = x - (\nabla^2 f(x)^{1/2} S^T S \nabla^2 f(x)^{1/2})^{-1} \nabla f(x). \blacksquare$$

For a self-concordant function, the sub-optimality of the Newton iterate x_{NE} in function value satisfies the bound

$$f(x_{\text{NE}}) - \underbrace{\min_{x \in \mathbb{R}^d} f(x)}_{f(x^*)} \leq [\lambda_f(x_{\text{NE}})]^2.$$

This classical bound is not directly applicable to the Newton Sketch update, since it involves the *approximate* Newton decrement $\tilde{\lambda}_f(x)^2 = -\langle \nabla f(x), v_{\text{NSK}} \rangle$, as opposed to the exact one $\lambda_f(x)^2 = -\langle \nabla f(x), v_{\text{NE}} \rangle$. Thus, our strategy is to prove that with high probability over the randomness in the sketch matrix, the approximate Newton decrement can be used as an exit condition.

Recall the definitions (4.36) and (4.37) of the exact v_{NE} and sketched Newton v_{NSK} update directions, as well as the definition of the tangent cone \mathcal{K} at $x \in \mathcal{C}$. Let \mathcal{K}^t be the tangent cone at x^t . The following lemma provides a high probability bound on their difference:

Lemma 19. *Let $S \in \mathbb{R}^{m \times n}$ be a sub-Gaussian, ROS or JL sketching matrix and consider any fixed vector $x \in \mathcal{C}$ independent of the sketch matrix. If $m \geq c_0 \frac{\mathbb{W}(\nabla^2 f(x)^{1/2} \mathcal{K}^t)^2}{\epsilon^2}$, then*

$$\|\nabla^2 f(x)^{1/2}(v_{\text{NSK}} - v_{\text{NE}})\|_2 \leq \epsilon \|\nabla^2 f(x)^{1/2} v_{\text{NE}}\|_2 \quad (4.38)$$

with probability at least $1 - c_1 e^{-c_2 m \epsilon^2}$.

Similar to the standard analysis of Newton's method, our analysis of the Newton Sketch algorithm is split into two phases defined by the magnitude of the decrement $\tilde{\lambda}_f(x)$. In particular, the following lemma constitute the core of our proof:

Lemma 20. *For $\epsilon \in (0, 1/2)$, there exist constants $\nu > 0$ and $\eta \in (0, 1/16)$ such that:*

- (a) *If $\tilde{\lambda}_f(x) > \eta$, then $f(x_{\text{NSK}}) - f(x) \leq -\nu$ with probability at least $1 - c_1 e^{-c_2 m \epsilon^2}$.*
- (b) *Conversely, if $\tilde{\lambda}_f(x) \leq \eta$, then*

$$\tilde{\lambda}_f(x_{\text{NSK}}) \leq \tilde{\lambda}_f(x), \quad \text{and} \quad (4.39a)$$

$$\lambda_f(x_{\text{NSK}}) \leq \left(\frac{16}{25}\right) \lambda_f(x), \quad (4.39b)$$

where both bounds hold with probability $1 - c_1 e^{c_2 m \epsilon^2}$.

Using this lemma, let us now complete the proof of the theorem, dividing our analysis into the two phases of the algorithm.

4.5.2.0.2 First phase analysis: By Lemma 20(a) each iteration in the first phase decreases the function value by at least $\nu > 0$, the number of first phase iterations N_1 is at most

$$N_1 := \frac{f(x^0) - f(x^*)}{\nu},$$

with probability at least $1 - N_1 c_1 e^{-c_2 m}$.

4.5.2.0.3 Second phase analysis: Next, let us suppose that at some iteration t , the condition $\tilde{\lambda}_f(x^t) \leq \eta$ holds, so that part (b) of Lemma 20 can be applied. In fact, the bound (4.39a) then guarantees that $\tilde{\lambda}_f(x^{t+1}) \leq \eta$, so that we may apply the contraction bound (4.39b) repeatedly for N_2 rounds so as to obtain that

$$\lambda_f(x^{t+N_2}) \leq \left(\frac{16}{25}\right)^{N_2} \lambda_f(x^t)$$

with probability $1 - N_2 c_1 e^{c_2 m}$.

Since $\lambda_f(x^t) \leq \eta \leq 1/16$ by assumption, the self-concordance of f then implies that

$$f(x^{t+k}) - f(x^*) \leq \left(\frac{16}{25}\right)^k \frac{1}{16}.$$

Therefore, in order to ensure that and consequently for achieving $f(x^{t+k}) - f(x^*) \leq \epsilon$, it suffices to the number of second phase iterations lower bounded as $N_2 \geq 0.65 \log_2(\frac{1}{16\epsilon})$.

Putting together the two phases, we conclude that the total number of iterations N required to achieve ϵ - accuracy is at most

$$N = N_1 + N_2 \leq \frac{f(x^0) - f(x^*)}{\gamma} + 0.65 \log_2\left(\frac{1}{16\epsilon}\right),$$

and moreover, this guarantee holds with probability at least $1 - N c_1 e^{-c_2 m \epsilon^2}$.

The final step in our proof of the theorem is to establish Lemma 20, and we do in the next two subsections.

4.5.2.1 Proof of Lemma 20(a)

Our proof of this part is performed conditionally on the event $\mathcal{D} := \{\tilde{\lambda}_f(x) > \eta\}$. Our strategy is to show that the backtracking line search leads to a stepsize $s > 0$ such that function decrement in moving from the current iterate x to the new sketched iterate $x_{\text{NSK}} = x + s v_{\text{NSK}}$ is at least

$$f(x_{\text{NSK}}) - f(x) \leq -\nu \quad \text{with probability at least } 1 - c_1 e^{-c_2 m}. \quad (4.40)$$

The outline of our proof is as follows. Defining the univariate function $g(u) := f(x + u v_{\text{NSK}})$ and $\epsilon' = \frac{2\epsilon}{1-\epsilon}$, we first show that $\hat{u} = \frac{1}{1+(1+\epsilon')\tilde{\lambda}_f(x)}$ satisfies the bound

$$g(\hat{u}) \leq g(0) - a \hat{u} \tilde{\lambda}_f(x)^2, \quad (4.41a)$$

which implies that \hat{u} satisfies the exit condition of backtracking line search. Therefore, the stepsize s must be lower bounded as $s \geq b\hat{u}$, which then implies that the updated solution $x_{\text{NSK}} = x + sv_{\text{NSK}}$ satisfies the decrement bound

$$f(x_{\text{NSK}}) - f(x) \leq -ab \frac{\tilde{\lambda}_f(x)^2}{1 + (1 + \frac{2\epsilon}{1-\epsilon})\tilde{\lambda}_f(x)}. \quad (4.41b)$$

Since $\tilde{\lambda}_f(x) > \eta$ by assumption and the function $u \rightarrow \frac{u^2}{1+(1+\frac{2\epsilon}{1-\epsilon})u}$ is monotone increasing, this bound implies that inequality (4.40) holds with $\nu = ab \frac{\eta^2}{1+(1+\frac{2\epsilon}{1-\epsilon})\eta}$.

It remains to prove the claims (4.41a) and (4.41b), for which we make use of the following auxiliary lemma:

Lemma 21. *For $u \in \text{dom } g \cap \mathbb{R}^+$, we have the decrement bound*

$$g(u) \leq g(0) + u \langle \nabla f(x), v_{\text{NSK}} \rangle - u \|\nabla^2 f(x)\|^{1/2} \|v_{\text{NSK}}\|_2 - \log(1 - u \|\nabla^2 f(x)\|^{1/2} \|v_{\text{NSK}}\|_2). \quad (4.42)$$

provided that $u \|\nabla^2 f(x)\|^{1/2} \|v_{\text{NSK}}\|_2 < 1$.

Lemma 22. *With probability at least $1 - c_1 e^{-c_2 m}$, we have*

$$\|\nabla^2 f(x)\|^{1/2} \|v_{\text{NSK}}\|_2^2 \leq \left(\frac{1+\epsilon}{1-\epsilon}\right)^2 [\tilde{\lambda}_f(x)]^2. \quad (4.43)$$

The proof of these lemmas are provided in Sections 4.7.1.2 and 4.7.1.3. Using them, let us prove the claims (4.41a) and (4.41b). Recalling our shorthand $\epsilon' := \frac{1+\epsilon}{1-\epsilon} - 1 = \frac{2\epsilon}{1-\epsilon}$, substituting inequality (4.43) into the decrement formula (4.42) yields

$$\begin{aligned} g(u) &\leq g(0) - u\tilde{\lambda}_f(x)^2 - u(1+\epsilon')\tilde{\lambda}_f(x) - \log(1 - u(1+\epsilon')\tilde{\lambda}_f(x)) \\ &= g(0) - \left\{ u(1+\epsilon')^2\tilde{\lambda}_f(x)^2 + u(1+\epsilon')\tilde{\lambda}_f(x) + \log(1 - u(1+\epsilon')\tilde{\lambda}_f(x)) \right\} \\ &\quad + u((1+\epsilon')^2 - 1)\tilde{\lambda}_f(x)^2 \end{aligned} \quad (4.44)$$

where we added and subtracted $u(1+\epsilon')^2\tilde{\lambda}_f(x)^2$ so as to obtain the final equality.

We now prove inequality (4.41a). Now setting $u = \hat{u} := \frac{1}{1+(1+\epsilon')\tilde{\lambda}_f(x)}$, which satisfies the conditions of Lemma 21 yields

$$g(\hat{u}) \leq g(0) - (1+\epsilon')\tilde{\lambda}_f(x) + \log(1 + (1+\epsilon')\tilde{\lambda}_f(x)) + \frac{(\epsilon'^2 + 2\epsilon')\tilde{\lambda}_f(x)^2}{1 + (1+\epsilon')\tilde{\lambda}_f(x)}.$$

Making use of the standard inequality $-u + \log(1 + u) \leq -\frac{\frac{1}{2}u^2}{(1+u)}$ (for instance, see the book [28]), we find that

$$\begin{aligned} g(\widehat{u}) &\leq g(0) - \frac{\frac{1}{2}(1 + \epsilon')^2 \widetilde{\lambda}_f(x)^2}{1 + (1 + \epsilon') \widetilde{\lambda}_f(x)} + \frac{(\epsilon'^2 + 2\epsilon') \widetilde{\lambda}_f(x)^2}{1 + (1 + \epsilon') \widetilde{\lambda}_f(x)} \\ &= g(0) - \left(\frac{1}{2} - \frac{1}{2}\epsilon'^2 - \epsilon'\right) \widetilde{\lambda}_f(x)^2 \widehat{u} \\ &\leq g(0) - \alpha \widetilde{\lambda}_f(x)^2 \widehat{u}, \end{aligned}$$

where the final inequality follows from our assumption $\alpha \leq \frac{1}{2} - \frac{1}{2}\epsilon'^2 - \epsilon'$. This completes the proof of the bound (4.41a). Finally, the lower bound (4.41b) follows by setting $u = b\widehat{u}$ into the decrement inequality (4.42).

4.5.2.2 Proof of Lemma 20(b)

The proof of this part hinges on the following auxiliary lemma:

Lemma 23. *For all $\epsilon \in (0, 1/2)$, we have*

$$\lambda_f(x_{NSK}) \leq \frac{(1 + \epsilon)\lambda_f^2(x) + \epsilon\lambda_f(x)}{\left(1 - (1 + \epsilon)\lambda_f(x)\right)^2}, \quad \text{and} \quad (4.45a)$$

$$(1 - \epsilon)\lambda_f(x) \leq \widetilde{\lambda}_f(x) \leq (1 + \epsilon)\lambda_f(x), \quad (4.45b)$$

where all bounds hold with probability at least $1 - c_1 e^{-c_2 m \epsilon^2}$.

See Section 4.7.1.4 for the proof.

We now use Lemma 23 to prove the two claims in the lemma statement.

4.5.2.2.1 Proof of the bound (4.39a): Recall from the theorem statement that $\eta := \frac{1}{8} \frac{1 - \frac{1}{2}(\frac{1+\epsilon}{1-\epsilon})^2 - a}{(\frac{1+\epsilon}{1-\epsilon})^3}$. By examining the roots of a polynomial in ϵ , it can be seen that $\eta \leq \frac{1-\epsilon}{1+\epsilon} \frac{1}{16}$. By applying the inequalities (4.45b), we have

$$(1 + \epsilon)\lambda_f(x) \leq \frac{1 + \epsilon}{1 - \epsilon} \widetilde{\lambda}_f(x) \leq \frac{1 + \epsilon}{1 - \epsilon} \eta \leq \frac{1}{16} \quad (4.46)$$

whence inequality (4.45a) implies that

$$\lambda_f(x_{NSK}) \leq \frac{\frac{1}{16}\lambda_f(x) + \epsilon\lambda_f(x)}{\left(1 - \frac{1}{16}\right)^2} \leq \left(\frac{16}{225} + \frac{256}{225}\epsilon\right) \lambda_f(x) \leq \frac{16}{25} \lambda_f(x). \quad (4.47)$$

Here the final inequality holds for all $\epsilon \in (0, 1/2)$. Combining the bound (4.45b) with inequality (4.47) yields

$$\tilde{\lambda}_f(x_{\text{NSK}}) \leq (1 + \epsilon)\lambda_f(x_{\text{NSK}}) \leq (1 + \epsilon)\left(\frac{16}{25}\right)\tilde{\lambda}_f(x) \leq \tilde{\lambda}_f(x),$$

where the final inequality again uses the condition $\epsilon \in (0, \frac{1}{2})$. This completes the proof of the bound (4.39a).

4.5.2.2 Proof of the bound (4.39b): This inequality has been established as a consequence of proving the bound (4.47).

4.5.3 Proof of Theorem 7

Given the proof of Theorem 6, it remains only to prove the following modified version of Lemma 19. It applies to the exact and sketched Newton directions $v_{\text{NE}}, v_{\text{NSK}} \in \mathbb{R}^d$ that are defined as follows

$$v_{\text{NE}} := \arg \min_{z \in \mathcal{C}-x} \left\{ \frac{1}{2} \|\nabla^2 f(x)^{1/2} z\|_2^2 + \langle z, \nabla f(x) \rangle + \frac{1}{2} \langle z, \nabla^2 g(x) z \rangle \right\}, \quad (4.48a)$$

$$v_{\text{NSK}} = \arg \min_{z \in \mathcal{C}-x} \underbrace{\left\{ \frac{1}{2} \|S \nabla^2 f(x)^{1/2} z\|_2^2 + \langle z, \nabla f(x) \rangle + \frac{1}{2} \langle z, \nabla^2 g(x) z \rangle \right\}}_{\Psi(z; S)}. \quad (4.48b)$$

Thus, the only difference is that the Hessian $\nabla^2 f(x)$ is sketched, whereas the term $\nabla^2 g(x)$ remains unsketched. Also note that since the function g is a self-concordant barrier for the set \mathcal{C} , we can safely omit the constraint \mathcal{C} in the definitions of sketched and original Newton steps.

Lemma 24. *Let $S \in \mathbb{R}^{m \times n}$ be a sub-Gaussian, ROS or JL sketching matrix, and let $x \in \mathbb{R}^d$ be a (possibly random) vector independent of S . If $m \geq c_0 \max_{x \in \mathcal{C}} \frac{\mathbb{W}(\nabla^2 f(x)^{1/2} \mathcal{K})^2}{\epsilon^2}$, then*

$$\|\nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \leq \epsilon \|\nabla^2 f(x)^{1/2} v_{\text{NE}}\|_2 \quad (4.49)$$

with probability at least $1 - c_1 e^{-c_2 m \epsilon^2}$.

4.6 Discussion

In this chapter we introduced and analyzed the Newton Sketch, a randomized approximation to the classical Newton updates. This algorithm is a natural generalization of the Iterative Hessian Sketch (IHS) updates analyzed in the previous chapter.

The IHS applies only to constrained least-squares problems (for which the Hessian is independent of the iteration number), whereas the Newton Sketch applies to twice differentiable convex functions, minimized over a closed and convex set. We described various applications of the Newton Sketch, including its use with barrier methods to solve various forms of constrained problems. For the minimization of self-concordant functions, the combination of the Newton Sketch within interior point updates leads to much faster algorithms for an extensive body of convex optimization problems.

Each iteration of the Newton Sketch has lower computational complexity than classical Newton’s method. Moreover, ignoring logarithmic factors, it has lower overall computational complexity than first-order methods when either $n \geq d^2$, when applied in the primal form, or $d \geq n^2$, when applied in the dual form; here n and d denote the dimensions of the data matrix A . In the context of barrier methods, the parameters n and d typically correspond to the number of constraints and number of variables, respectively. In many “big data” problems, one of the dimensions is much larger than the other, in which case the Newton Sketch is advantageous. Moreover, sketches based on the randomized Hadamard transform are well-suited to in parallel environments: in this case, the sketching step can be done in $\mathcal{O}(\log m)$ time with $\mathcal{O}(nd)$ processors. This scheme significantly decreases the amount of central computation—namely, from $\mathcal{O}(m^2d + nd \log m)$ to $\mathcal{O}(m^2d + \log d)$.

There are a number of open problems associated with the Newton Sketch. Here we focused our analysis on the cases of sub-Gaussian, randomized orthogonal system (ROS) sketches and JL embeddings. It would also be interesting to analyze sketches based on row sampling and leverage scores. Such techniques preserve the sparsity of the Hessian, and can be used in conjunction with sparse KKT system solvers. Finally, it would be interesting to explore the problem of lower bounds on the sketch dimension m . In particular, is there a threshold below which any algorithm that has access only to gradients and m -sketched Hessians must necessarily converge at a sub-linear rate, or in a way that depends on the strong convexity and smoothness parameters? Such a result would clarify whether or not the guarantees we obtained are improvable.

4.7 Proofs of technical results

4.7.1 Technical results for Theorem 6

In this section, we collect together various technical results and proofs that are required in the proof of Theorem 6.

4.7.1.1 Proof of Lemma 19

Let u be a unit-norm vector independent of S , and consider the random quantities

$$Z_1(A\mathcal{K})(S, x) := \inf_{v \in \nabla^2 f(x)^{1/2} \mathcal{K}^t \cap \mathcal{S}^{n-1}} \|Sv\|_2^2 \quad \text{and} \quad (4.50a)$$

$$Z_2(A\mathcal{K})(S, x) := \sup_{v \in \nabla^2 f(x)^{1/2} \mathcal{K}^t \cap \mathcal{S}^{n-1}} \left| \langle u, (S^T S - I_n) v \rangle \right|. \quad (4.50b)$$

By the optimality and feasibility of v_{NSK} and v_{NE} (respectively) for the sketched Newton update (4.37), we have

$$\frac{1}{2} \|S \nabla^2 f(x)^{1/2} v_{\text{NSK}}\|_2^2 - \langle v_{\text{NSK}}, \nabla f(x) \rangle \leq \frac{1}{2} \|\nabla^2 f(x)^{1/2} v_{\text{NE}}\|_2^2 - \langle v_{\text{NE}}, \nabla f(x) \rangle.$$

Defining the difference vector $\hat{e} := v_{\text{NSK}} - v_{\text{NE}}$, some algebra leads to the basic inequality

$$\frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 \leq -\langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, S^T S \nabla^2 f(x)^{1/2} \hat{e} \rangle + \langle \hat{e}, \nabla f(x) \rangle. \quad (4.51)$$

Moreover, by the optimality and feasibility of v_{NE} and v_{NSK} for the exact Newton update (4.36), we have

$$\langle \nabla^2 f(x) v_{\text{NE}} - \nabla f(x), \hat{e} \rangle = \langle \nabla^2 f(x) v_{\text{NE}} - \nabla f(x), v_{\text{NSK}} - v_{\text{NE}} \rangle \geq 0. \quad (4.52)$$

Consequently, by adding and subtracting $\langle \nabla^2 f(x) v_{\text{NE}}, \hat{e} \rangle$, we find that

$$\frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 \leq \left| \langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, (I_n - S^T S) \nabla^2 f(x)^{1/2} \hat{e} \rangle \right|. \quad (4.53)$$

By definition, the error vector \hat{e} belongs to the cone \mathcal{K}^t and the vector $\nabla^2 f(x)^{1/2} v_{\text{NE}}$ is fixed and independent of the sketch. Consequently, invoking definitions (4.50a) and (4.50b) of the random variables $Z_1(A\mathcal{K})$ and $Z_2(A\mathcal{K})$ yields

$$\begin{aligned} \frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 &\geq \frac{Z_1(A\mathcal{K})}{2} \|\nabla^2 f(x)^{1/2} \hat{e}\|_2^2, \\ \left| \langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, (I_n - S^T S) \nabla^2 f(x)^{1/2} \hat{e} \rangle \right| &\leq Z_2(A\mathcal{K}) \|\nabla^2 f(x)^{1/2} v_{\text{NE}}\|_2 \|\nabla^2 f(x)^{1/2} \hat{e}\|_2, \end{aligned}$$

Putting together the pieces, we find that

$$\|\nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \leq \frac{2Z_2(A\mathcal{K})(S, x)}{Z_1(A\mathcal{K})(S, x)} \|\nabla^2 f(x)^{1/2} (v_{\text{NE}})\|_2. \quad (4.54)$$

Finally, for any $\delta \in (0, 1)$, let us define the event $\mathcal{E}(\delta) = \{Z_1(A\mathcal{K}) \geq 1 - \delta, \text{ and } Z_2(A\mathcal{K}) \leq \delta\}$. By Lemma 4 and Lemma 5 of [114], we are guaranteed that $\mathbb{P}[\mathcal{E}(\delta)] \geq 1 - c_1 e^{-c_2 m \delta^2}$. Conditioned on the event $\mathcal{E}(\delta)$, the bound (4.54) implies that

$$\|\nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \leq \frac{2\delta}{1 - \delta} \|\nabla^2 f(x)^{1/2} (v_{\text{NE}})\|_2.$$

By setting $\delta = \frac{\epsilon}{4}$, the claim follows.

4.7.1.2 Proof of Lemma 21

By construction, the function $g(u) = f(x + uv_{\text{NSK}})$ is strictly convex and self-concordant. Consequently, it satisfies the bound $\frac{d}{du} (g''(u)^{-1/2}) \leq 1$, whence

$$g''(s)^{-1/2} - g''(0)^{-1/2} = \int_0^s \frac{d}{du} (g''(u)^{-1/2}) du \leq s.$$

or equivalently $g''(s) \leq \frac{g''(0)}{(1 - sg''(0)^{1/2})^2}$ for $s \in \text{dom } g \cap [0, g''(0)^{-1/2})$. Integrating this inequality twice yields the bound

$$g(u) \leq g(0) + ug'(0) - ug''(0)^{1/2} - \log(1 - ug''(0)^{1/2}). \quad (4.55)$$

Since $g'(u) = \langle \nabla f(x + uv_{\text{NSK}}), v_{\text{NSK}} \rangle$ and $g''(u) = \langle v_{\text{NSK}}, \nabla^2 f(x + uv_{\text{NSK}}) v_{\text{NSK}} \rangle$, the decrement bound (4.42) follows.

4.7.1.3 Proof of Lemma 22

We perform this analysis conditional on the bound (4.38) from Lemma 19. We begin by observing that

$$\begin{aligned} \|\nabla^2 f(x)^{1/2} v_{\text{NSK}}\|_2 &\leq \|\nabla^2 f(x)^{1/2} v_{\text{NE}}\|_2 + \|\nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \\ &= \lambda_f(x) + \|\nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2. \end{aligned} \quad (4.56)$$

Lemma 19 implies that $\|\nabla^2 [f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \leq \epsilon \|\nabla^2 [f(x)]^{1/2} v_{\text{NE}}\|_2 = \epsilon \lambda_f(x)$. In conjunction with the bound (4.56), we see that

$$\|\nabla^2 f(x)^{1/2} v_{\text{NSK}}\|_2 \leq (1 + \epsilon) \lambda_f(x). \quad (4.57)$$

Our next step is to lower bound the term $\langle \nabla f(x), v_{\text{NSK}} \rangle$: in particular, by adding and subtracting a factor of the original Newton step v_{NE} , we find that

$$\begin{aligned} \langle \nabla f(x), v_{\text{NSK}} \rangle &= \langle [\nabla^2 f(x)]^{-1/2} \nabla f(x), \nabla^2 [f(x)]^{1/2} v_{\text{NSK}} \rangle \\ &= \langle [\nabla^2 f(x)]^{-1/2} \nabla f(x), \nabla^2 [f(x)]^{1/2} v_{\text{NE}} \rangle + \langle [\nabla^2 f(x)]^{-1/2} \nabla f(x), \nabla^2 [f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \rangle \\ &= -\|\nabla^2 [f(x)]^{-1/2} \nabla f(x)\|_2^2 + \langle [\nabla^2 f(x)]^{-1/2} \nabla f(x), \nabla^2 [f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \rangle \\ &\leq -\|\nabla^2 [f(x)]^{-1/2} \nabla f(x)\|_2^2 + \|\nabla^2 f(x)^{-1/2} \nabla f(x)\|_2 \|\nabla^2 [f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \\ &= -\lambda_f(x)^2 + \lambda_f(x) \|\nabla^2 [f(x)]^{1/2} (v_{\text{NSK}} - v_{\text{NE}})\|_2 \\ &\leq -\lambda_f(x)^2 (1 - \epsilon), \end{aligned} \quad (4.58)$$

where the final step again makes use of Lemma 19. Repeating the above argument in the reverse direction yields the lower bound $\langle \nabla f(x), v_{\text{NSK}} \rangle \geq -\lambda_f(x)^2 (1 + \epsilon)$, so that we may conclude that

$$|\tilde{\lambda}_f(x) - \lambda_f(x)| \leq \epsilon \lambda_f(x). \quad (4.59)$$

Finally, by squaring both sides of the inequality (4.56) and combining with the above bounds gives

$$\|[\nabla^2 f(x)]^{1/2} v_{\text{NSK}}\|_2^2 \leq \frac{-(1+\epsilon)^2}{1-\epsilon} \langle \nabla f(x), v_{\text{NSK}} \rangle = \frac{(1+\epsilon)^2}{1-\epsilon} \tilde{\lambda}_f^2(x) \leq \left(\frac{1+\epsilon}{1-\epsilon}\right)^2 \tilde{\lambda}_f^2(x),$$

as claimed.

4.7.1.4 Proof of Lemma 23

We have already proved the bound (4.45b) during our proof of Lemma 22—in particular, see equation (4.59). Accordingly, it remains only to prove the inequality (4.45a).

Introducing the shorthand $\tilde{\lambda} := (1+\epsilon)\lambda_f(x)$, we first claim that the Hessian satisfies the sandwich relation

$$(1-s\alpha)^2 \nabla^2 f(x) \preceq \nabla^2 f(x + sv_{\text{NSK}}) \preceq \frac{1}{(1-s\alpha)^2} \nabla^2 f(x), \quad (4.60)$$

for $|1-s\alpha| < 1$ where $\alpha = (1+\epsilon)\lambda_f(x)$, with probability at least $1 - c_1 e^{-c_2 m \epsilon^2}$. Let us recall Theorem 4.1.6 of Nesterov [104]: it guarantees that

$$(1-s\|v_{\text{NSK}}\|_x)^2 \nabla^2 f(x) \preceq \nabla^2 f(x + sv_{\text{NSK}}) \preceq \frac{1}{(1-s\|v_{\text{NSK}}\|_x)^2} \nabla^2 f(x). \quad (4.61)$$

Now recall the bound (4.38) from Lemma 19: combining it with an application of the triangle inequality (in terms of the semi-norm $\|v\|_x = \|\nabla^2 f(x)^{1/2} v\|_2$) yields

$$\|\nabla^2 f(x)^{1/2} v_{\text{NSK}}\|_2 \leq (1+\epsilon) \|\nabla^2 f(x)^{1/2} v_{\text{NE}}\|_2 = (1+\epsilon) \|v_{\text{NE}}\|_x,$$

with probability at least $1 - e^{-c_1 m \epsilon^2}$, and substituting this inequality into the bound (4.61) yields the sandwich relation (4.60) for the Hessian.

Using this sandwich relation (4.60), the Newton decrement can be bounded as

$$\begin{aligned} \lambda_f(x_{\text{NSK}}) &= \|\nabla^2 f(x_{\text{NSK}})^{-1/2} \nabla f(x_{\text{NSK}})\|_2 \\ &\leq \frac{1}{(1-(1+\epsilon)\lambda_f(x))} \|\nabla^2 f(x)^{-1/2} \nabla f(x_{\text{NSK}})\|_2 \\ &= \frac{1}{(1-(1+\epsilon)\lambda_f(x))} \left\| \nabla^2 f(x)^{-1/2} \left(\nabla f(x) + \int_0^1 \nabla^2 f(x + sv_{\text{NSK}}) v_{\text{NSK}} ds \right) \right\|_2 \\ &= \frac{1}{(1-(1+\epsilon)\lambda_f(x))} \left\| \nabla^2 f(x)^{-1/2} \left(\nabla f(x) + \int_0^1 \nabla^2 f(x + sv_{\text{NSK}}) v_{\text{NE}} ds + \Delta \right) \right\|_2, \end{aligned}$$

where we have defined $\Delta = \int_0^1 \nabla^2 f(x + sv_{\text{NSK}})(v_{\text{NSK}} - v_{\text{NE}}) ds$. By the triangle inequality, we can write $\lambda_f(x_{\text{NSK}}) \leq \frac{1}{(1-(1+\epsilon)\lambda_f(x))} (M_1 + M_2)$, where

$$M_1 := \left\| \nabla^2 f(x)^{-1/2} \left(\nabla f(x) + \int_0^1 \nabla^2 f(x + tv_{\text{NSK}})v_{\text{NE}} dt \right) \right\|_2, \quad \text{and} \quad M_2 := \left\| \nabla^2 f(x)^{-1/2} \Delta \right\|_2. \quad \blacksquare$$

In order to complete the proof, it suffices to show that

$$M_1 \leq \frac{(1+\epsilon)\lambda_f(x)^2}{1-(1+\epsilon)\lambda_f(x)}, \quad \text{and} \quad M_2 \leq \frac{\epsilon\lambda_f(x)}{1-(1+\epsilon)\lambda_f(x)}.$$

4.7.1.4.1 Bound on M_1 : Re-arranging and then invoking the Hessian sandwich relation (4.60) yields

$$\begin{aligned} M_1 &= \left\| \int_0^1 (\nabla^2 f(x)^{-1/2} \nabla^2 f(x + sv_{\text{NSK}}) \nabla^2 f(x)^{-1/2} - I) ds (\nabla^2 f(x)^{1/2} v_{\text{NE}}) \right\|_2 \\ &\leq \left| \int_0^1 \left(\frac{1}{(1-s(1+\epsilon)\lambda_f(x))^2} - 1 \right) ds \right| (\nabla^2 f(x)^{1/2} v_{\text{NE}}) \Big\|_2 \\ &= \frac{(1+\epsilon)\lambda_f(x)}{1-(1+\epsilon)\lambda_f(x)} \left\| \nabla^2 f(x)^{1/2} v_{\text{NE}} \right\|_2 \\ &= \frac{(1+\epsilon)\lambda_f^2(x)}{1-(1+\epsilon)\lambda_f(x)}. \end{aligned}$$

4.7.1.4.2 Bound on M_2 : We have

$$\begin{aligned} M_2 &= \left\| \int_0^1 \nabla^2 f(x)^{-1/2} \nabla^2 f(x + sv_{\text{NSK}}) \nabla^2 f(x)^{-1/2} ds \nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \right\|_2 \\ &\leq \left\| \int_0^1 \frac{1}{(1-s(1+\epsilon)\lambda_f(x))^2} ds \nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \right\|_2 \\ &= \frac{1}{1-(1+\epsilon)\lambda_f(x)} \left\| \nabla^2 f(x)^{1/2} (v_{\text{NSK}} - v_{\text{NE}}) \right\|_2 \\ &\stackrel{(i)}{\leq} \frac{1}{1-(1+\epsilon)\lambda_f(x)} \epsilon \left\| \nabla^2 f(x)^{1/2} v_{\text{NE}} \right\|_2 \\ &= \frac{\epsilon\lambda_f(x)}{1-(1+\epsilon)\lambda_f(x)}, \end{aligned}$$

where the inequality in step (i) follows from Lemma 19.

4.7.2 Proof of Lemma 24

The proof follows the basic inequality argument of the proof of Lemma 19. Since v_{NSK} and v_{NE} are optimal and feasible (respectively) for the sketched Newton problem (4.48b), we have $\Psi(v_{\text{NSK}}; S) \leq \Psi(v_{\text{NE}}; S)$. Defining the difference vector $\hat{e} := v_{\text{NSK}} - v_{\text{NE}}$, some algebra leads to the basic inequality

$$\begin{aligned} \frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 + \frac{1}{2} \langle \hat{e}, \nabla^2 g(x) \hat{e} \rangle &\leq -\langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, S^T S \nabla^2 f(x)^{1/2} \hat{e} \rangle \\ &\quad + \langle \hat{e}, (\nabla f(x) - \nabla^2 g(x)) v_{\text{NE}} \rangle. \end{aligned}$$

On the other hand since v_{NE} and v_{NSK} are optimal and feasible (respectively) for the Newton step (4.48a), we have

$$\langle \nabla^2 f(x) v_{\text{NE}} + \nabla^2 g(x) v_{\text{NE}} - \nabla f(x), \hat{e} \rangle \geq 0.$$

Consequently, by adding and subtracting $\langle \nabla^2 f(x) v_{\text{NE}}, \hat{e} \rangle$, we find that

$$\frac{1}{2} \|S \nabla^2 f(x)^{1/2} \hat{e}\|_2^2 + \frac{1}{2} \langle v_{\text{NE}}, \nabla^2 g(x) v_{\text{NE}} \rangle \leq \left| \langle \nabla^2 f(x)^{1/2} v_{\text{NE}}, (I_n - S^T S) \nabla^2 f(x)^{1/2} \hat{e} \rangle \right|. \quad (4.62)$$

We next define the matrix $\bar{H}(x)^{1/2} := \begin{bmatrix} \nabla^2 f(x)^{1/2} \\ \nabla^2 g(x)^{1/2} \end{bmatrix}$ and the augmented sketching matrix $\bar{S} := \begin{bmatrix} S & 0 \\ 0 & I_q \end{bmatrix}$ where $q = 2n$. Then we can rewrite the inequality (4.62) as follows

$$\frac{1}{2} \|\bar{S} \bar{H}(x)^{1/2} \hat{e}\|_2^2 \leq \left| \langle \bar{H}(x)^{1/2} v_{\text{NE}}, (I_q - \bar{S}^T \bar{S}) \bar{H}(x)^{1/2} \hat{e} \rangle \right|.$$

Note that the modified sketching matrix \bar{S} also satisfies the conditions (4.50a) and (4.50b). Consequently the remainder of the proof follows as in the proof of Lemma 19.

4.7.3 Gaussian widths with ℓ_1 -constraints

In this section, we state and prove an elementary lemma that bounds for the Gaussian width for a broad class of ℓ_1 -constrained problems. In particular, given a twice-differentiable convex function ψ , a vector $c \in \mathbb{R}^d$, a radius R and a collection of d -vectors $\{a_i\}_{i=1}^n$, consider a convex program of the form

$$\min_{x \in \mathcal{C}} \left\{ \sum_{i=1}^n \psi(\langle a_i, x \rangle) + \langle c, x \rangle \right\}, \quad \text{where } \mathcal{C} = \{x \in \mathbb{R}^d \mid \|x\|_1 \leq R\}. \quad (4.63)$$

Lemma 25. *Suppose that the ℓ_1 -constrained program (4.63) has a unique optimal solution x^* such that $\|x^*\|_0 \leq s$ for some integer k . Then denoting the tangent cone at x^* by \mathcal{K} , then*

$$\max_{x \in \mathcal{C}} \mathbb{W}(\nabla^2 f(x)^{1/2} \mathcal{K}) \leq 6\sqrt{k \log d} \sqrt{\frac{\psi''_{\max}}{\psi''_{\min}}} \frac{\max_{j=1, \dots, d} \|A_j\|_2}{\sqrt{\gamma_k^-(A)}},$$

where

$$\psi''_{\min} = \min_{x \in \mathcal{C}} \min_{i=1, \dots, n} \psi''(\langle a_i, x \rangle, y_i), \quad \text{and} \quad \psi''_{\max} = \max_{x \in \mathcal{C}} \max_{i=1, \dots, n} \psi''(\langle a_i, x \rangle, y_i).$$

Proof. It is well-known (e.g., [67, 114]) that the tangent cone of the ℓ_1 -norm at any k -sparse solution is a subset of the cone $\{z \in \mathbb{R}^d \mid \|z\|_1 \leq 2\sqrt{k}\|z\|_2\}$. Using this fact, we have the following sequence of upper bounds

$$\begin{aligned} \mathbb{W}(\nabla^2 f(x)^{1/2} \mathcal{K}) &= \mathbb{E}_w \max_{\substack{z^T \nabla^2 f(x) z = 1, \\ z \in \mathcal{K}}} \langle w, \nabla^2 f(x)^{1/2} z \rangle \\ &= \mathbb{E}_w \max_{\substack{z^T A^T \text{diag}(\psi''(\langle a_i, x \rangle, y_i)) A z = 1, \\ z \in \mathcal{K}}} \langle w, \text{diag}(\psi''(\langle a_i, x \rangle, y_i))^{1/2} A z \rangle \\ &\leq \mathbb{E}_w \max_{\substack{z^T A^T A z \leq 1/\psi''_{\min} \\ z \in \mathcal{K}}} \langle w, \text{diag}(\psi''(\langle a_i, x \rangle, y_i))^{1/2} A z \rangle \\ &\leq \mathbb{E}_w \max_{\|z\|_1 \leq \frac{2\sqrt{k}}{\sqrt{\gamma_k^-(A)}} \frac{1}{\sqrt{\psi''_{\min}}}} \langle w, \text{diag}(\psi''(\langle a_i, x \rangle, y_i))^{1/2} A z \rangle \\ &= \frac{2\sqrt{k}}{\sqrt{\gamma_k^-(A)}} \frac{1}{\sqrt{\psi''_{\min}}} \mathbb{E}_w \|A^T \text{diag}(\psi''(\langle a_i, x \rangle, y_i))^{1/2} w\|_{\infty} \\ &= \frac{2\sqrt{s}}{\sqrt{\gamma_k^-(A)}} \frac{1}{\sqrt{\psi''_{\min}}} \mathbb{E}_w \max_{j=1, \dots, d} \left| \underbrace{\sum_{i=1, \dots, n} w_i A_{ij} \psi''(\langle a_i, x \rangle, y_i)^{1/2}}_{Q_j} \right|. \end{aligned}$$

Here the random variables Q_j are zero-mean Gaussians with variance at most

$$\sum_{i=1, \dots, n} A_{ij}^2 \psi''(\langle a_i, x \rangle, y_i) \leq \psi''_{\max} \|A_j\|_2^2.$$

Consequently, applying standard bounds on the suprema of Gaussian variates [85], we obtain

$$\mathbb{E}_w \max_{j=1, \dots, d} \left| \sum_{i=1, \dots, n} w_i A_{ij} \psi''(\langle a_i, x \rangle, y_i)^{1/2} \right| \leq 3\sqrt{\log d} \sqrt{\psi''_{\max}} \max_{j=1, \dots, d} \|A_j\|_2.$$

When combined with the previous inequality, the claim follows. \square

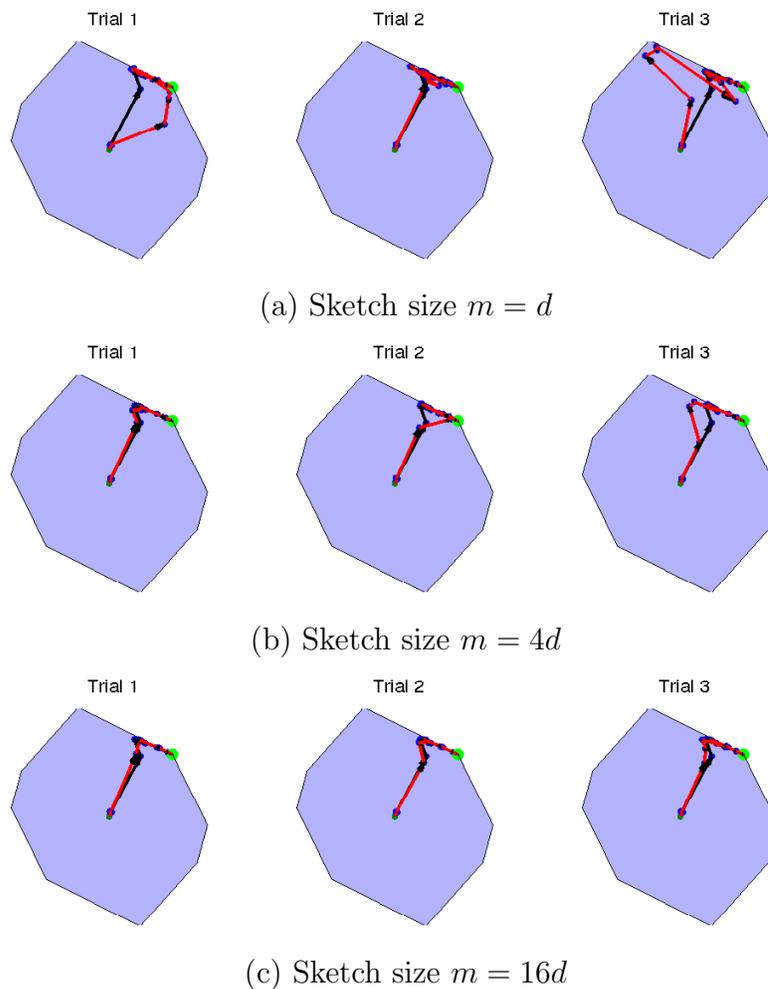


Figure 4.1: Comparisons of central paths for a simple linear program in two dimensions. Each row shows three independent trials for a given sketch dimension: across the rows, the sketch dimension ranges as $m \in \{d, 4d, 16d\}$. The black arrows show Newton steps taken by the standard interior point method, whereas red arrows show the steps taken by the sketched version. The green point at the vertex represents the optimum. In all cases, the sketched algorithm converges to the optimum, and as the sketch dimension m increases, the sketched central path converges to the standard central path.

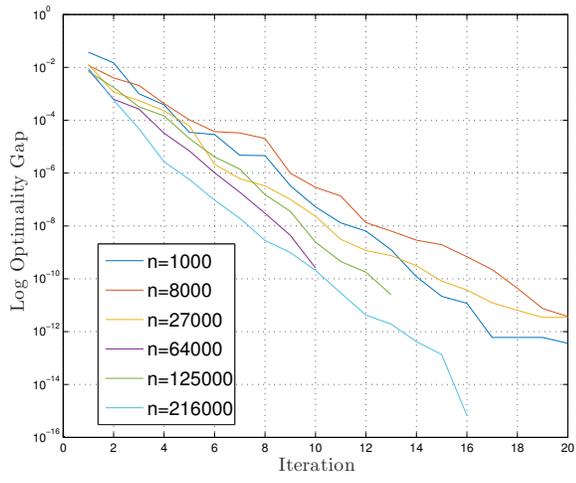


Figure 4.2: Empirical illustration of the linear convergence of the Newton Sketch algorithm for an ensemble of portfolio optimization problems (4.15). In all cases, the algorithm was implemented using a sketch dimension $m = \lceil 4s \log d \rceil$, where s is an upper bound on the number of non-zeros in the optimal solution x^* ; this quantity satisfies the required lower bound (4.12), and consistent with the theory, the algorithm displays linear convergence.

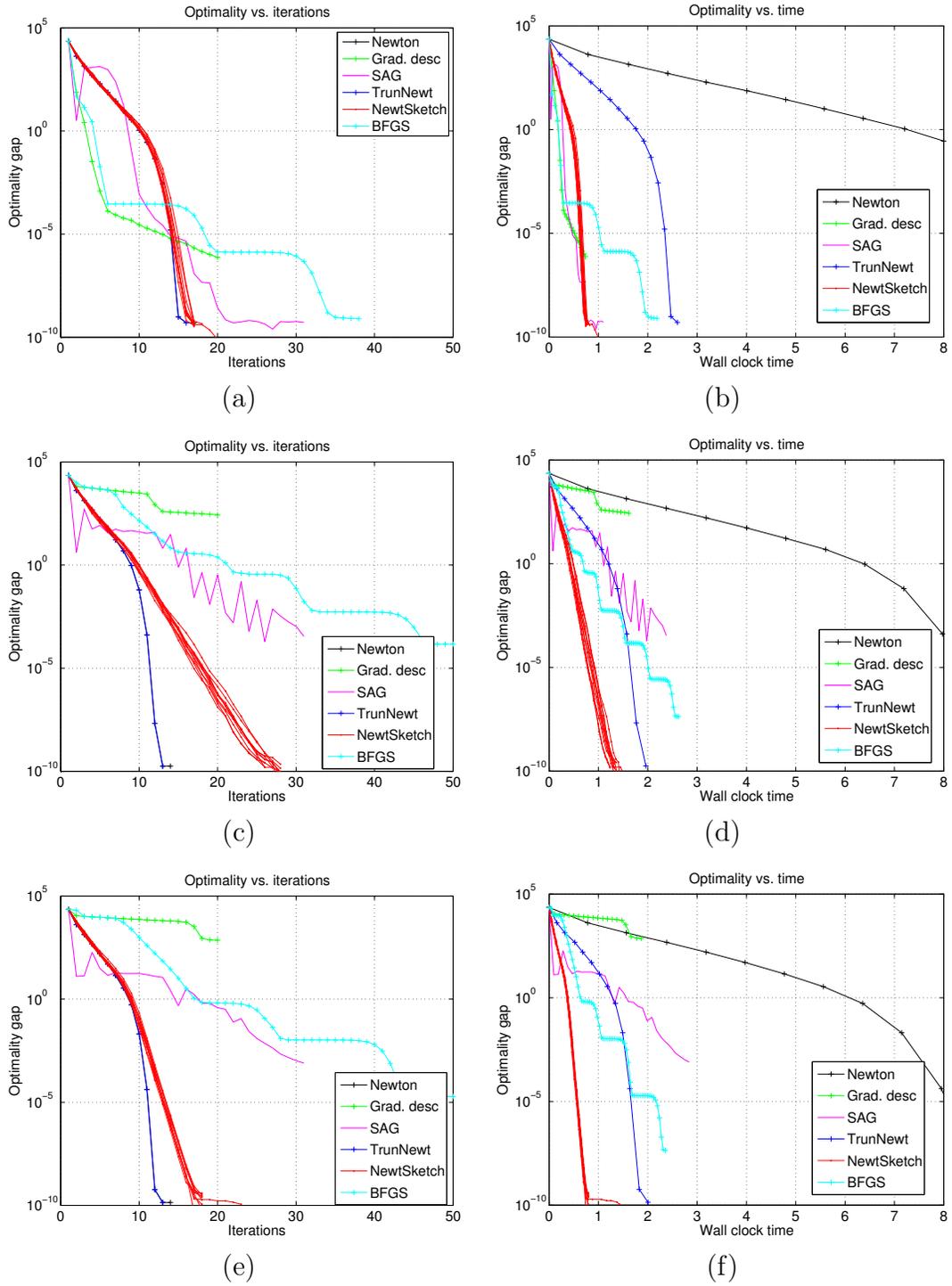


Figure 4.3: Comparison of Newton Sketch with various other algorithms in the logistic regression problem with Gaussian data.

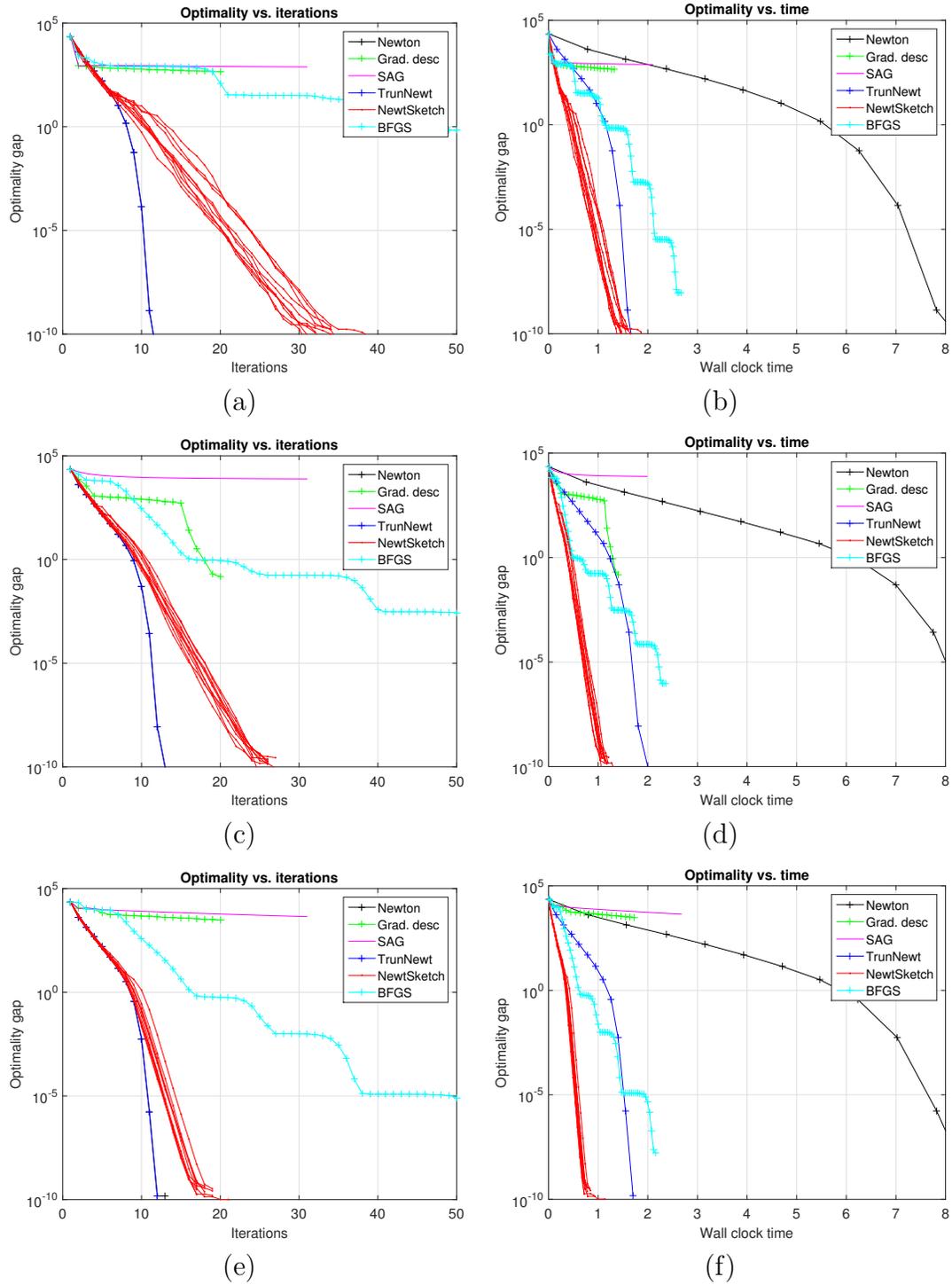


Figure 4.4: Comparison of Newton Sketch with other algorithms in the logistic regression problem with Student's t-distributed data

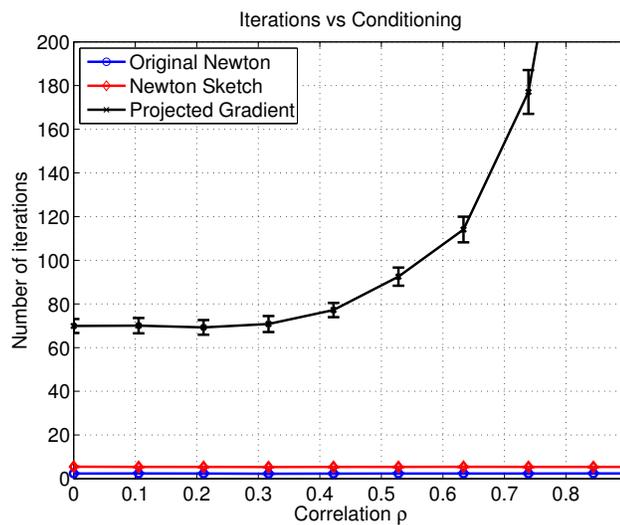


Figure 4.5: The performance of Newton Sketch is independent of condition numbers and problem related quantities. Plots of the number of iterations required to reach 10^{-6} accuracy in ℓ_1 -constrained logistic regression using Newton’s Method and Projected Gradient Descent using line search.

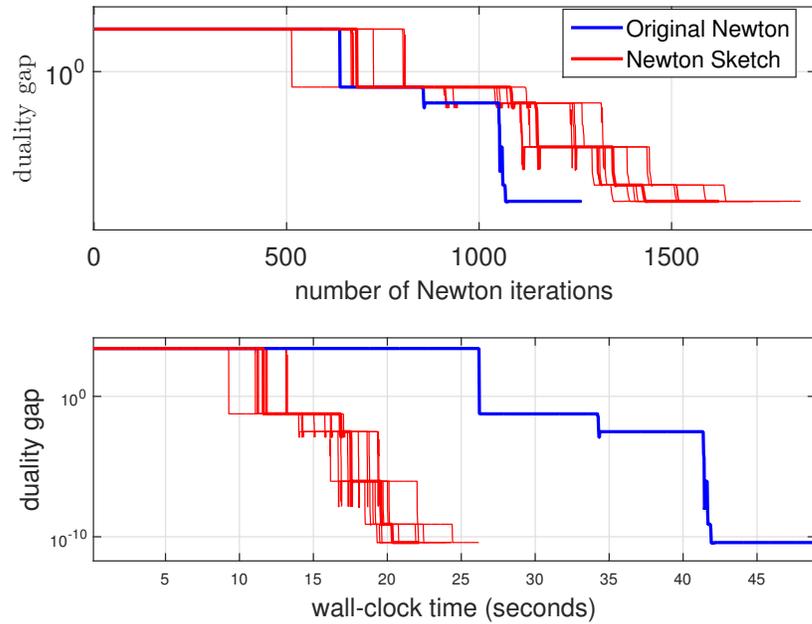


Figure 4.6: Plots of the duality gap versus iteration number (top panel) and duality gap versus wall-clock time (bottom panel) for the original barrier method (blue) and sketched barrier method (red). The sketched interior point method is run 10 times independently yielding slightly different curves in red. While the sketched method requires more iterations, its overall wall-clock time is much smaller.

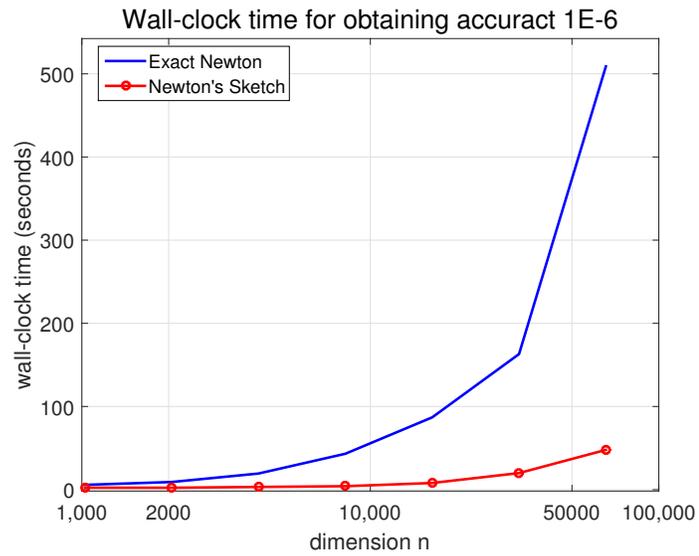


Figure 4.7: Plot of the wall-clock time in seconds for reaching a duality gap of 10^{-6} for the standard and sketched interior point methods as n increases (in log-scale). The sketched interior point method has significantly lower computation time compared to the original method.

Chapter 5

Random projection, effective dimension and nonparametric regression

The goal of non-parametric regression is to make predictions of a response variable $Y \in \mathbb{R}$ based on observing a covariate vector $X \in \mathcal{X}$. In practice, we are given a collection of n samples, say $\{(x_i, y_i)\}_{i=1}^n$ of covariate-response pairs and our goal is to estimate the regression function $f^*(x) = \mathbb{E}[Y \mid X = x]$. In the standard Gaussian model, it is assumed that the covariate-response pairs are related via the model

$$y_i = f^*(x_i) + \sigma w_i, \quad \text{for } i = 1, \dots, n \quad (5.1)$$

where the sequence $\{w_i\}_{i=1}^n$ consists of i.i.d. standard Gaussian variates. It is typical to assume that the regression function f^* has some regularity properties, and one way of enforcing such structure is to require f^* to belong to a reproducing kernel Hilbert space, or RKHS for short [13, 141, 65]). Given such an assumption, it is natural to estimate f^* by minimizing a combination of the least-squares fit to the data and a penalty term involving the squared Hilbert norm, leading to an estimator known *kernel ridge regression*, or KRR for short [68, 127]). From a statistical point of view, the behavior of KRR can be characterized using existing results on M -estimation and empirical processes (e.g. [79, 96, 138]). When the regularization parameter is set appropriately, it is known to yield a function estimate with minimax prediction error for various classes of kernels.

Despite these attractive statistical properties, the computational complexity of computing the KRR estimate prevents it from being routinely used in large-scale problems. More precisely, in a standard implementation [124], the time complexity and space complexity of KRR scales as $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$, respectively, where n refers to the number of samples. As a consequence, it becomes important to design methods for computing approximate forms of the KRR estimate, while retaining guarantees

of optimality in terms of statistical minimaxity. Various authors have taken different approaches to this problem. Zhang et al. [155] analyze a distributed implementation of KRR, in which a set of t machines each compute a separate estimate based on a random t -way partition of the full data set, and combine it into a global estimate by averaging. This divide-and-conquer approach has time complexity and space complexity $\mathcal{O}(n^3/t^3)$ and $\mathcal{O}(n^2/t^2)$, respectively. Zhang et al. [155] give conditions on the number of splits t , as a function of the kernel, under which minimax optimality of the resulting estimator can be guaranteed. More closely related to our methods that are based on forming a low-rank approximation to the n -dimensional kernel matrix, such as the Nyström methods (e.g. [53, 61]). The time complexity by using a low-rank approximation is either $\mathcal{O}(nr^2)$ or $\mathcal{O}(n^2r)$, depending on the specific approach (excluding the time for factorization), where r is the maintained rank, and the space complexity is $\mathcal{O}(nr)$. Some recent work [16, 7] analyzes the tradeoff between the rank r and the resulting statistical performance of the estimator, and we discuss this line of work at more length in Section 5.2.3.

We will consider approximations to KRR based on random projections, also known as sketches, of the data. Random projections are a classical way of performing dimensionality reduction, and are widely used in many algorithmic contexts (e.g., see the book [139] and references therein). Our proposal is to approximate n -dimensional kernel matrix by projecting its row and column subspaces to a randomly chosen m -dimensional subspace with $m \ll n$. By doing so, an approximate form of the KRR estimate can be obtained by solving an m -dimensional quadratic program, which involves time and space complexity $\mathcal{O}(m^3)$ and $\mathcal{O}(m^2)$. Computing the approximate kernel matrix is a pre-processing step that has time complexity $\mathcal{O}(n^2 \log(m))$ for suitably chosen projections; this pre-processing step is trivially parallelizable, meaning it can be reduced to $\mathcal{O}(n^2 \log(m)/t)$ by using $t \leq n$ clusters.

Given such an approximation, we pose the following question: how small can the projection dimension m be chosen while still retaining minimax optimality of the approximate KRR estimate? We answer this question by connecting it to the *statistical dimension* d_n of the n -dimensional kernel matrix, a quantity that measures the effective number of degrees of freedom. (See Section 5.1.3 for a precise definition.) From the results of earlier work on random projections for constrained Least Squares estimators (e.g., see [114, 112]), it is natural to conjecture that it should be possible to project the kernel matrix down to the statistical dimension while preserving minimax optimality of the resulting estimator. The main contribution of this chapter is to confirm this conjecture for several classes of random projection matrices.

It is worth mentioning that our sketching approach is radically different from the classical least-squares sketch—the former applies random projection to reduce the parameter dimension while the latter reduce the number of observations. As shown in [112], although the classical least-squares sketch approximates the value of the quadratic objective function, it is sub-optimal for approximating the solution in terms

of some distance measure between the approximate minimizer and the true minimizer. However, our sketching approach retains minimax optimality of the approximate KRR estimate.

The remainder of this chapter is organized as follows. Section 5.1 is devoted to further background on non-parametric regression, reproducing kernel Hilbert spaces and associated measures of complexity, as well as the notion of statistical dimension of a kernel. In Section 5.2, we turn to statements of our main results. Theorem 10 provides a general sufficient condition on a random sketch for the associated approximate form of KRR to achieve the minimax risk. In Corollary 13, we derive some consequences of this general result for particular classes of random sketch matrices, and confirm these theoretical predictions with some simulations. We also compare at more length to methods based on the Nyström approximation in Section 5.2.3. Section 5.3 is devoted to the proofs of our main results. We conclude with a discussion in Section 5.4.

5.1 Problem formulation and background

We begin by introducing some background on nonparametric regression and reproducing kernel Hilbert spaces, before formulating the main problem.

5.1.1 Regression in reproducing kernel Hilbert spaces

Given n samples $\{(x_i, y_i)\}_{i=1}^n$ from the non-parametric regression model (5.1), our goal is to estimate the unknown regression function f^* . The quality of an estimate \hat{f} can be measured in different ways: for consistency with our earlier results, we will focus on the squared $L^2(\mathbb{P}_n)$ error

$$\|\hat{f} - f^*\|_n^2 := \frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))^2. \quad (5.2)$$

Naturally, the difficulty of non-parametric regression is controlled by the structure in the function f^* , and one way of modeling such structure is within the framework of a reproducing kernel Hilbert space (or RKHS for short). Here we provide a very brief introduction referring the reader to the books [21, 65, 141] for more details and background.

Given a space \mathcal{X} endowed with a probability distribution \mathbb{P} , the space $L^2(\mathbb{P})$ consists of all functions that are square-integrable with respect to \mathbb{P} . In abstract terms, a space $\mathcal{H} \subset L^2(\mathbb{P})$ is an RKHS if for each $x \in \mathcal{X}$, the evaluation function $f \mapsto f(x)$ is a bounded linear functional. In more concrete terms, any RKHS is generated by a positive semidefinite (PSD) kernel function in the following way. A

PSD kernel function is a symmetric function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that, for any positive integer N , collections of points $\{v_1, \dots, v_N\}$ and weight vector $\omega \in \mathbb{R}^N$, the sum $\sum_{i,j=1}^N \omega_i \omega_j \mathcal{K}(v_i, v_j)$ is non-negative. Suppose moreover that for each fixed $v \in \mathcal{X}$, the function $u \mapsto \mathcal{K}(u, v)$ belongs to $L^2(\mathbb{P})$. We can then consider the vector space of all functions $g : \mathcal{X} \rightarrow \mathbb{R}$ of the form

$$g(\cdot) = \sum_{i=1}^N \omega_i \mathcal{K}(\cdot, v_i)$$

for some integer N , points $\{v_1, \dots, v_N\} \subset \mathcal{X}$ and weight vector $w \in \mathbb{R}^N$. By taking the closure of all such linear combinations, it can be shown [13] that we generate an RKHS, and one that is uniquely associated with the kernel \mathcal{K} . We provide some examples of various kernels and the associated function classes in Section 5.1.3 to follow.

5.1.2 Kernel ridge regression and its sketched form

Given the dataset $\{(x_i, y_i)\}_{i=1}^n$, a natural method for estimating unknown function $f^* \in \mathcal{H}$ is known as kernel ridge regression (KRR): it is based on the convex program

$$f^\diamond := \arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{H}}^2 \right\}, \quad (5.3)$$

where λ_n is a regularization parameter corresponding to the Hilbert space norm $\|\cdot\|_{\mathcal{H}}$.

As stated, this optimization problem can be infinite-dimensional in nature, since it takes place over the Hilbert space. However, as a straightforward consequence of the representer theorem [76], the solution to this optimization problem can be obtained by solving the n -dimensional convex program. In particular, let us define the *empirical kernel matrix*, namely the n -dimensional symmetric matrix K with entries $K_{ij} = n^{-1} \mathcal{K}(x_i, x_j)$. Here we adopt the n^{-1} scaling for later theoretical convenience. In terms of this matrix, the KRR estimate can be obtained by first solving the quadratic program

$$\omega^\dagger = \arg \min_{\omega \in \mathbb{R}^n} \left\{ \frac{1}{2} \omega^T K^2 \omega - \omega^T \frac{Ky}{\sqrt{n}} + \lambda_n \omega^T K \omega \right\}, \quad (5.4a)$$

and then outputting the function

$$f^\diamond(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \omega_i^\dagger \mathcal{K}(\cdot, x_i). \quad (5.4b)$$

In principle, the original KRR optimization problem (5.4a) is simple to solve: it is an n dimensional quadratic program, and can be solved exactly using $\mathcal{O}(n^3)$ via a

QR decomposition. However, in many applications, the number of samples may be large, so that this type of cubic scaling is prohibitive. In addition, the n -dimensional kernel matrix K is dense in general, and so requires storage of order n^2 numbers, which can also be problematic in practice.

We consider an approximation based on limiting the original parameter $\omega \in \mathbb{R}^n$ to an m -dimensional subspace of \mathbb{R}^n , where $m \ll n$ is the *projection dimension*. We define this approximation via a sketch matrix $S \in \mathbb{R}^{m \times n}$, such that the m -dimensional subspace is generated by the row span of S . More precisely, the *sketched kernel ridge regression* estimate is given by first solving

$$\hat{\alpha} = \arg \min_{\theta \in \mathbb{R}^m} \left\{ \frac{1}{2} \alpha^T (SK)(KS^T) \alpha - \alpha^T S \frac{Ky}{\sqrt{n}} + \lambda_n \alpha^T SKS^T \alpha \right\}, \quad (5.5a)$$

and then outputting the function

$$\hat{f}(\cdot) := \frac{1}{\sqrt{n}} \sum_{i=1}^n (S^T \hat{\alpha})_i \mathcal{K}(\cdot, x_i). \quad (5.5b)$$

Note that the sketched program (5.5a) is a quadratic program in m dimensions: it takes as input the m -dimensional matrices (SK^2S^T, SKS^T) and the m -dimensional vector SKy . Consequently, it can be solved efficiently via QR decomposition with computational complexity $\mathcal{O}(m^3)$. Moreover, the computation of the sketched kernel matrix $SK = [SK_1, \dots, SK_n]$ in the input can be parallelized across its columns.

In this section, we analyze various forms of randomized sketching matrices. In section 5.5, we show that the sketched KRR estimate (5.5a) based on a sub-sampling sketch matrix is equivalent to the Nyström approximation.

5.1.3 Kernel complexity measures and statistical guarantees

So as to set the stage for later results, let us characterize an appropriate choice of the regularization parameter λ , and the resulting bound on the prediction error $\|f^\diamond - f^*\|_n$. Recall the empirical kernel matrix K defined in the previous section: since it is symmetric and positive definite, it has an eigendecomposition of the form $K = UDU^T$, where $U \in \mathbb{R}^{n \times n}$ is an orthonormal matrix, and $D \in \mathbb{R}^{n \times n}$ is diagonal with elements $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \dots \geq \hat{\mu}_n \geq 0$. Using these eigenvalues, consider the *kernel complexity function*

$$\hat{\mathcal{R}}(\delta) = \sqrt{\frac{1}{n} \sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}}, \quad (5.6)$$

corresponding to a rescaled sum of the eigenvalues, truncated at level δ^2 . This function arises via analysis of the local Rademacher complexity of the kernel class (e.g., [19,

79, 96, 120]). For a given kernel matrix and noise variance $\sigma > 0$, the *critical radius* is defined to be the smallest positive solution $\delta_n > 0$ to the inequality

$$\frac{\widehat{\mathcal{R}}(\delta)}{\delta} \leq \frac{\delta}{\sigma}. \quad (5.7)$$

Note that the existence and uniqueness of this critical radius is guaranteed for any kernel class [19].

5.1.3.0.3 Bounds on ordinary KRR: The significance of the critical radius is that it can be used to specify bounds on the prediction error in kernel ridge regression. More precisely suppose that we compute the KRR estimate (5.3) with any regularization parameter $\lambda \geq 2\delta_n^2$. Then with probability at least $1 - c_1 e^{-c_2 n \delta_n^2}$, we are guaranteed that

$$\|f^\diamond - f^*\|_n^2 \leq c_u \{\lambda_n + \delta_n^2\}, \quad (5.8)$$

where $c_u > 0$ is a universal constant (independent of n , σ and the kernel). This known result follows from standard techniques in empirical process theory (e.g., [138, 19]); we also note that it can be obtained as a corollary of our more general theorem on sketched KRR estimates to follow (viz. Theorem 10).

To illustrate, let us consider a few examples of reproducing kernel Hilbert spaces, and compute the critical radius in different cases. In working through these examples, so as to determine explicit rates, we assume that the design points $\{x_i\}_{i=1}^n$ are sampled i.i.d. from some underlying distribution \mathbb{P} , and we make use of the useful fact that, up to constant factors, we can always work with the population-level kernel complexity function

$$\mathcal{R}(\delta) = \sqrt{\frac{1}{n} \sum_{j=1}^{\infty} \min\{\delta^2, \mu_j\}}, \quad (5.9)$$

where $\{\mu_j\}_{j=1}^{\infty}$ are the eigenvalues of the kernel integral operator (assumed to be uniformly bounded). This equivalence follows from standard results on the population and empirical Rademacher complexities [96, 19].

Example 4 (Polynomial kernel). For some integer $D \geq 1$, consider the kernel function on $[0, 1] \times [0, 1]$ given by $\mathcal{K}_{\text{poly}}(u, v) = (1 + \langle u, v \rangle)^D$. For $D = 1$, it generates the class of all linear functions of the form $f(x) = a_0 + a_1 x$ for some scalars (a_0, a_1) , and corresponds to a linear kernel. More generally, for larger integers D , it generates the class of all polynomial functions of degree at most D —that is, functions of the form $f(x) = \sum_{j=0}^D a_j x^j$.

Let us now compute a bound on the critical radius δ_n . It is straightforward to show that the polynomial kernel is of finite rank at most $D + 1$, meaning that the kernel

matrix K always has at most $\min\{D + 1, n\}$ non-zero eigenvalues. Consequently, as long $n > D + 1$, there is a universal constant c such that

$$\widehat{\mathcal{R}}(\delta) \leq c \sqrt{\frac{D+1}{n}} \delta,$$

which implies that $\delta_n^2 \lesssim \sigma^2 \frac{D+1}{n}$. Consequently, we conclude that the KRR estimate satisfies the bound $\|\widehat{f} - f^*\|_n^2 \lesssim \sigma^2 \frac{D+1}{n}$ with high probability. Note that this bound is intuitive, since a polynomial of degree D has $D + 1$ free parameters.

Example 5 (Gaussian kernel). The Gaussian kernel with bandwidth $h > 0$ takes the form $\mathcal{K}_{\text{Gau}}(u, v) = e^{-\frac{1}{2h^2}(u-v)^2}$. When defined with respect to Lebesgue measure on the real line, the eigenvalues of the kernel integral operator scale as $\mu_j \asymp \exp(-\pi h^2 j^2)$ as $j \rightarrow \infty$. Based on this fact, it can be shown that the critical radius scales as $\delta_n^2 \asymp \frac{\sigma^2}{n} \sqrt{\log\left(\frac{n}{\sigma^2}\right)}$. Thus, even though the Gaussian kernel is non-parametric (since it cannot be specified by a fixed number of parameters), it is still a relatively small function class.

Example 6 (First-order Sobolev space). As a final example, consider the kernel defined on the unit square $[0, 1] \times [0, 1]$ given by $\mathcal{K}_{\text{sob}}(u, v) = \min\{u, v\}$. It generates the function class

$$\begin{aligned} \mathcal{H}^1[0, 1] = \{ & f : [0, 1] \rightarrow \mathbb{R} \mid f(0) = 0, \\ & \text{and } f \text{ is abs. cts. with } \int_0^1 [f'(x)]^2 dx < \infty \}, \end{aligned} \quad (5.10)$$

a class that contains all Lipschitz functions on the unit interval $[0, 1]$. Roughly speaking, we can think of the first-order Sobolev class as functions that are almost everywhere differentiable with derivative in $L^2[0, 1]$. Note that this is a much larger kernel class than the Gaussian kernel class. The first-order Sobolev space can be generalized to higher order Sobolev spaces, in which functions have additional smoothness. See the book [65] for further details on these and other reproducing kernel Hilbert spaces.

If the kernel integral operator is defined with respect to Lebesgue measure on the unit interval, then the population level eigenvalues are given by $\mu_j = \left(\frac{2}{(2j-1)\pi}\right)^2$ for $j = 1, 2, \dots$. Given this relation, some calculation shows that the critical radius scales as $\delta_n^2 \asymp \left(\frac{\sigma^2}{n}\right)^{2/3}$. This is the familiar minimax risk for estimating Lipschitz functions in one dimension [133].

5.1.3.0.4 Lower bounds for non-parametric regression: For future reference, it is also convenient to provide a lower bound on the prediction error achievable by *any estimator*. In order to do so, we first define the *statistical dimension* of the kernel as

$$d_n := \min \{j \in [n] : \widehat{\mu}_j \leq \delta_n^2\}, \quad (5.11)$$

and $d_n = n$ if no such index j exists. By definition, we are guaranteed that $\widehat{\mu}_j > \delta_n^2$ for all $j \in \{1, 2, \dots, d_n\}$. In terms of this statistical dimension, we have

$$\widehat{\mathcal{R}}(\delta_n) = \left[\frac{d_n}{n} \delta_n^2 + \frac{1}{n} \sum_{j=d_n+1}^n \widehat{\mu}_j \right]^{1/2},$$

showing that the statistical dimension controls a type of bias-variance tradeoff.

It is reasonable to expect that the critical rate δ_n should be related to the statistical dimension as $\delta_n^2 \asymp \frac{\sigma^2 d_n}{n}$. This scaling relation holds whenever the tail sum satisfies a bound of the form $\sum_{j=d_n+1}^n \widehat{\mu}_j \lesssim d_n \delta_n^2$. Although it is possible to construct pathological examples in which this scaling relation does not hold, it is true for most kernels of interest, including all examples considered in this section. For any such regular kernel, the critical radius provides a fundamental lower bound on the performance of *any estimator*, as summarized in the following theorem:

Theorem 9 (Critical radius and minimax risk). *Given n i.i.d. samples $\{(y_i, x_i)\}_{i=1}^n$ from the standard non-parametric regression model over any regular kernel class, any estimator \widetilde{f} has prediction error lower bounded as*

$$\sup_{\|f^*\|_{\mathcal{H}} \leq 1} \mathbb{E} \|\widetilde{f} - f^*\|_n^2 \geq c_\ell \delta_n^2, \quad (5.12)$$

where $c_\ell > 0$ is a numerical constant, and δ_n is the critical radius (5.7).

The proof of this claim, provided in Section 5.6.1, is based on a standard application of Fano's inequality, combined with a random packing argument. It establishes that the critical radius is a fundamental quantity, corresponding to the appropriate benchmark to which sketched kernel regression estimates should be compared.

5.2 Main results and their consequences

We now turn to statements of our main theorems on kernel sketching, as well as a discussion of some of their consequences. We first introduce the notion of a K -satisfiable sketch matrix, and then show (in Theorem 10) that any sketched KRR estimate based on a K -satisfiable sketch also achieves the minimax risk. We illustrate this achievable result with several corollaries for different types of randomized sketches. For Gaussian and ROS sketches, we show that choosing the sketch dimension proportional to the statistical dimension of the kernel (with additional log factors in the ROS case) is sufficient to guarantee that the resulting sketch will be K -satisfiable with high probability. In addition, we illustrate the sharpness of our theoretical predictions via some experimental simulations.

5.2.1 General conditions for sketched kernel optimality

Recall the definition (5.11) of the statistical dimension d_n , and consider the eigen-decomposition $K = UDU^T$ of the kernel matrix, where $U \in \mathbb{R}^{n \times n}$ is an orthonormal matrix of eigenvectors, and $D = \text{diag}\{\widehat{\mu}_1, \dots, \widehat{\mu}_n\}$ is a diagonal matrix of eigenvalues. Let $U_1 \in \mathbb{R}^{n \times d_n}$ denote the left block of U , and similarly, $U_2 \in \mathbb{R}^{n \times (n-d_n)}$ denote the right block. Note that the columns of the left block U_1 correspond to the eigenvectors of K associated with the leading d_n eigenvalues, whereas the columns of the right block U_2 correspond to the eigenvectors associated with the remaining $n - d_n$ smallest eigenvalues. Intuitively, a sketch matrix $S \in \mathbb{R}^{m \times n}$ is “good” if the sub-matrix $SU_1 \in \mathbb{R}^{m \times d_n}$ is relatively close to an isometry, whereas the sub-matrix $SU_2 \in \mathbb{R}^{m \times (n-d_n)}$ has a relatively small operator norm.

This intuition can be formalized in the following way. For a given kernel matrix K , a sketch matrix S is said to be K -satisfiable if there is a universal constant c such that

$$\|(SU_1)^T SU_1 - I_{d_n}\|_2 \leq 1/2, \quad \text{and} \quad \|SU_2 D_2^{1/2}\|_2 \leq c \delta_n, \quad (5.13)$$

where $D_2 = \text{diag}\{\widehat{\mu}_{d_n+1}, \dots, \widehat{\mu}_n\}$.

Given this definition, the following theorem shows that any sketched KRR estimate based on a K -satisfiable matrix achieves the minimax risk (with high probability over the noise in the observation model):

Theorem 10 (Upper bound). *Given n i.i.d. samples $\{(y_i, x_i)\}_{i=1}^n$ from the standard non-parametric regression model, consider the sketched KRR problem (5.5a) based on a K -satisfiable sketch matrix S . Then any for $\lambda_n \geq 2\delta_n^2$, the sketched regression estimate \widehat{f} from equation (5.5b) satisfies the bound*

$$\|\widehat{f} - f^*\|_n^2 \leq c_u \{\lambda_n + \delta_n^2\}$$

with probability greater than $1 - c_1 e^{-c_2 n \delta_n^2}$.

We emphasize that in the case of fixed design regression and for a fixed sketch matrix, the K -satisfiable condition on the sketch matrix S is a deterministic statement: apart from the sketch matrix, it only depends on the properties of the kernel function \mathcal{K} and design variables $\{x_i\}_{i=1}^n$. Thus, when using randomized sketches, the algorithmic randomness can be completely decoupled from the randomness in the noisy observation model (5.1).

5.2.1.0.5 Proof intuition: The proof of Theorem 10 is given in Section 5.3.1. At a high-level, it is based on an upper bound on the prediction error $\|\widehat{f} - f^*\|_n^2$ that involves two sources of error: the *approximation error* associated with solving

a zero-noise version of the KRR problem in the projected m -dimensional space, and the *estimation error* between the noiseless and noisy versions of the projected problem. In more detail, letting $z^* := (f^*(x_1), \dots, f^*(x_n))$ denote the vector of function evaluations defined by f^* , consider the quadratic program

$$\alpha^\dagger := \arg \min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{2n} \|z^* - \sqrt{n}KS^T\alpha\|_2^2 + \lambda_n \|K^{1/2}S^T\alpha\|_2^2 \right\}, \quad (5.14)$$

as well as the associated fitted function $f^\dagger = \frac{1}{\sqrt{n}} \sum_{i=1}^n (S\alpha^\dagger)_i \mathcal{K}(\cdot, x_i)$. The vector $\alpha^\dagger \in \mathbb{R}^m$ is the solution of the sketched problem in the case of zero noise, whereas the fitted function f^\dagger corresponds to the best penalized approximation of f^* within the range space of S^T .

Given this definition, we then have the elementary inequality

$$\frac{1}{2} \|\widehat{f} - f^*\|_n^2 \leq \underbrace{\|f^\dagger - f^*\|_n^2}_{\text{Approximation error}} + \underbrace{\|f^\dagger - \widehat{f}\|_n^2}_{\text{Estimation error}}. \quad (5.15)$$

For a fixed sketch matrix, the approximation error term is deterministic: it corresponds to the error induced by approximating f^* over the range space of S^T . On the other hand, the estimation error depends both on the sketch matrix and the observation noise. In Section 5.3.1, we state and prove two lemmas that control the approximation and error terms respectively.

As a corollary, Theorem 10 implies the stated upper bound (5.8) on the prediction error of the original (unsketched) KRR estimate (5.3). Indeed, this estimator can be obtained using the “sketch matrix” $S = I_{n \times n}$, which is easily seen to be K -satisfiable. In practice, however, we are interested in $m \times n$ sketch matrices with $m \ll n$, so as to achieve computational savings. In particular, a natural conjecture is that it should be possible to efficiently generate K -satisfiable sketch matrices with the projection dimension m proportional to the statistical dimension d_n of the kernel. Of course, one such K -satisfiable matrix is given by $S = U_1^T \in \mathbb{R}^{d_n \times n}$, but it is not easy to generate, since it requires computing the eigendecomposition of K . Nonetheless, as we now show, there are various randomized constructions that lead to K -satisfiable sketch matrices with high probability.

5.2.2 Corollaries for randomized sketches

When combined with additional probabilistic analysis, Theorem 10 implies that various forms of randomized sketches achieve the minimax risk using a sketch dimension proportional to the statistical dimension d_n . Here we analyze the Gaussian and ROS families of random sketches, as previously defined in Section 5.1.2. Throughout our analysis, we require that the sketch dimension satisfies a lower bound of the form

$$m \geq \begin{cases} c d_n & \text{for Gaussian sketches, and} \\ c d_n \log^4(n) & \text{for ROS sketches,} \end{cases} \quad (5.16a)$$

where d_n is the *statistical dimension* as previously defined in equation (5.11). Here it should be understood that the constant c can be chosen sufficiently large (but finite). In addition, for the purposes of stating high probability results, we define the function

$$\phi(m, d_n, n) := \begin{cases} c_1 e^{-c_2 m} & \text{for Gaussian sketches, and} \\ c_1 \left[e^{-c_2 \frac{m}{d_n \log^2(n)}} + e^{-c_2 d_n \log^2(n)} \right] & \text{for ROS sketches,} \end{cases} \quad (5.16b)$$

where c_1, c_2 are universal constants. With this notation, the following result provides a high probability guarantee for both Gaussian and ROS sketches:

Corollary 13 (Guarantees for Gaussian and ROS sketches). *Given n i.i.d. samples $\{(y_i, x_i)\}_{i=1}^n$ from the standard non-parametric regression model (5.1), consider the sketched KRR problem (5.5a) based on a sketch dimension m satisfying the lower bound (5.16a). Then there is a universal constant c'_u such that for any $\lambda_n \geq 2\delta_n^2$, the sketched regression estimate (5.5b) satisfies the bound*

$$\|\hat{f} - f^*\|_n^2 \leq c'_u \{\lambda_n + \delta_n^2\}$$

with probability greater than $1 - \phi(m, d_n, n) - c_3 e^{-c_4 n \delta_n^2}$.

In order to illustrate Corollary 13, let us return to the three examples previously discussed in Section 5.1.3. To be concrete, we derive the consequences for Gaussian sketches, noting that ROS sketches incur only an additional $\log^4(n)$ overhead.

- for the D^{th} -order polynomial kernel from Example 4, the statistical dimension d_n for any sample size n is at most $D + 1$, so that a sketch size of order $D + 1$ is sufficient. This is a very special case, since the kernel is finite rank and so the required sketch dimension has no dependence on the sample size.
- for the Gaussian kernel from Example 5, the statistical dimension satisfies the scaling $d_n \asymp \sqrt{\log n}$, so that it suffices to take a sketch dimension scaling logarithmically with the sample size.
- for the first-order Sobolev kernel from Example 6, the statistical dimension scales as $d_n \asymp n^{1/3}$, so that a sketch dimension scaling as the cube root of the sample size is required.

Remark. In practice, the target sketch dimension m is only known up to a multiplicative constant. To determine this multiplicative constant, one can implement the randomized algorithm in an adaptive fashion where the multiplicative constant is increased until the squared $L^2(\mathbb{P}_n)$ norm of the change in the fitted function \hat{f} falls below a desired tolerance. This adaptive procedure only slightly increases the time complexity—when increasing the sketch dimension from m to m' , we only need to sample additional $m' - m$ rows to form the new sketch matrix S' for any of the three random sketch schemes described in Section 5.1.2. Correspondingly, to form the new sketched kernel matrix $S'K$, we only need to compute the product of the new rows of S' and the kernel matrix K . Fig. 5.1(d) and Fig. 5.2(d) below show that the relative approximation error $\|\hat{f} - f^\diamond\|_n^2 / \|f^\diamond - f^*\|_n^2$ has a rapid decay as the projection dimension m grows, which justifies the validity of the adaptive procedure.

In order to illustrate these theoretical predictions, we performed some simulations. Beginning with the Sobolev kernel $\mathcal{K}_{\text{sob}}(u, v) = \min\{u, v\}$ on the unit square, as introduced in Example 6, we generated n i.i.d. samples from the model (5.1) with noise standard deviation $\sigma = 0.5$, the unknown regression function

$$f^*(x) = 1.6 |(x - 0.4)(x - 0.6)| - 0.3, \quad (5.17)$$

and uniformly spaced design points $x_i = \frac{i}{n}$ for $i = 1, \dots, n$. By construction, the function f^* belongs to the first-order Sobolev space with $\|f^*\|_{\mathcal{H}} \approx 1.3$. As suggested by our theory for the Sobolev kernel, we set the projection dimension $m = \lceil n^{1/3} \rceil$, and then solved the sketched version of kernel ridge regression, for both Gaussian sketches and ROS sketches based on the fast Hadamard transform. We performed simulations for n in the set $\{32, 64, 128, \dots, 16384\}$ so as to study scaling with the sample size. As noted above, our theory predicts that the squared prediction loss $\|\hat{f} - f^*\|_n^2$ should tend to zero at the same rate $n^{-2/3}$ as that of the unsketched estimator f^\diamond . Figure 5.1 confirms this theoretical prediction. In panel (a), we plot the squared prediction error versus the sample size, showing that all three curves (original, Gaussian sketch and ROS sketch) tend to zero. Panel (b) plots the *rescaled* prediction error $n^{2/3} \|\hat{f} - f^*\|_n^2$ versus the sample size, with the relative flatness of these curves confirming the $n^{-2/3}$ decay predicted by our theory. Panel (c) plots the running time versus the sample size and the squared prediction error, showing that kernel sketching considerably speeds up KRR.

In our second experiment, we repeated the same set of simulations this time for the 3-d Gaussian kernel $\mathcal{K}_{\text{Gau}}(u, v) = e^{-\frac{1}{2h^2}\|u-v\|_2^2}$ with bandwidth $h = 1$, and the function $f^*(x) = 0.5 e^{-x_1+x_2} - x_2x_3$. In this case, as suggested by our theory, we choose the sketch dimension $m = \lceil 1.25(\log n)^{3/2} \rceil$. Figure 5.2 shows the same types of plots with the prediction error. In this case, we expect that the squared prediction error will decay at the rate $\frac{(\log n)^{3/2}}{n}$. This prediction is confirmed by the plot in panel (b),

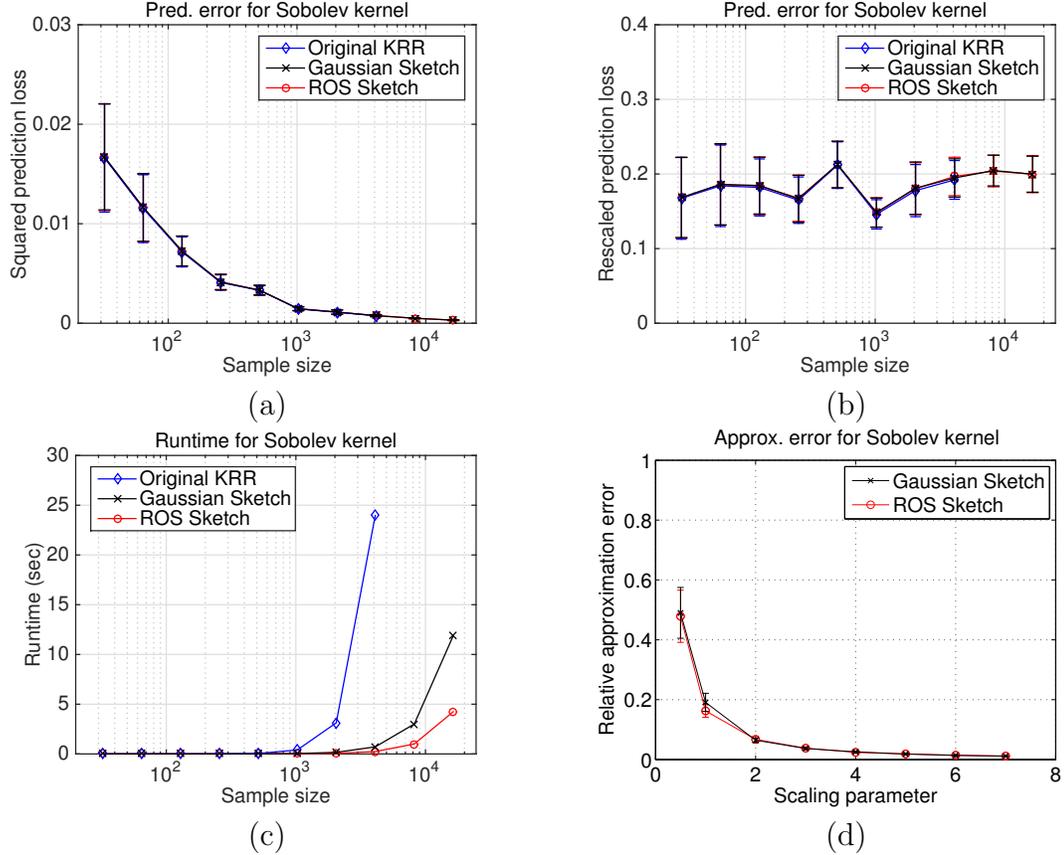


Figure 5.1: Prediction error versus sample size for original KRR, Gaussian sketch, and ROS sketches for the Sobolev one kernel for the function $f^*(x) = 1.6|(x - 0.4)(x - 0.6)| - 0.3$. In all cases, each point corresponds to the average of 100 trials, with standard errors also shown. (a) Squared prediction error $\|\hat{f} - f^*\|_n^2$ versus the sample size $n \in \{32, 64, 128, \dots, 16384\}$ for projection dimension $m = \lceil n^{1/3} \rceil$. (b) Rescaled prediction error $n^{2/3}\|\hat{f} - f^*\|_n^2$ versus the sample size. (c) Runtime versus the sample size. (d) Relative approximation error $\|\hat{f} - f^\diamond\|_n^2 / \|f^\diamond - f^*\|_n^2$ versus scaling parameter c for $n = 1024$ and $m = \lceil cn^{1/3} \rceil$ with $c \in \{0.5, 1, 2, \dots, 7\}$. The original KRR under $n = 8192$ and 16384 are not computed due to out-of-memory failures.

showing that the rescaled error $\frac{n}{(\log n)^{3/2}}\|\hat{f} - f^*\|_n^2$, when plotted versus the sample size, remains relatively constant over a wide range.

5.2.3 Comparison with Nyström-based approaches

It is interesting to compare the convergence rate and computational complexity of our methods with guarantees based on the Nyström approximation. As shown in

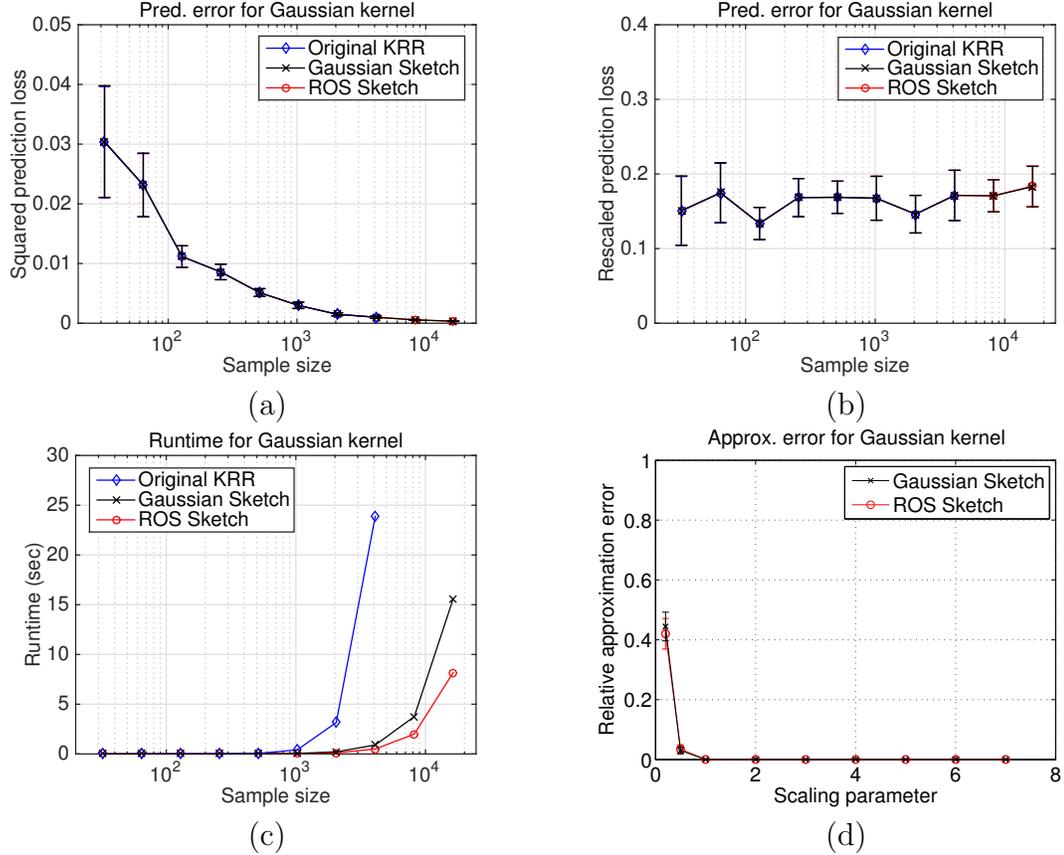


Figure 5.2: Prediction error versus sample size for original KRR, Gaussian sketch, and ROS sketches for the Gaussian kernel with the function $f^*(x) = 0.5 e^{-x_1+x_2} - x_2x_3$. In all cases, each point corresponds to the average of 100 trials, with standard errors also shown. (a) Squared prediction error $\|\hat{f} - f^*\|_n^2$ versus the sample size $n \in \{32, 64, 128, \dots, 16384\}$ for projection dimension $m = \lceil 1.25(\log n)^{3/2} \rceil$. (b) Rescaled prediction error $\frac{n}{(\log n)^{3/2}} \|\hat{f} - f^*\|_n^2$ versus the sample size. (c) Runtime versus the sample size. (d) Relative approximation error $\|\hat{f} - f^\diamond\|_n^2 / \|f^\diamond - f^*\|_n^2$ versus scaling parameter c for $n = 1024$ and $m = \lceil c(\log n)^{3/2} \rceil$ with $c \in \{0.5, 1, 2, \dots, 7\}$. The original KRR under $n = 8192$ and 16384 are not computed due to out-of-memory failures.

Section 5.5, this Nyström approximation approach can be understood as a particular form of our sketched estimate, one in which the sketch corresponds to a random row-sampling matrix.

Bach [16] analyzed the prediction error of the Nyström approximation to KRR based on uniformly sampling a subset of p -columns of the kernel matrix K , leading to an overall computational complexity of $\mathcal{O}(np^2)$. In order for the approximation

to match the performance of KRR, the number of sampled columns must be lower bounded as

$$p \gtrsim n \|\text{diag}(K(K + \lambda_n I)^{-1})\|_\infty \log n,$$

a quantity which can be substantially larger than the statistical dimension required by our methods. Moreover, as shown in the following example, there are many classes of kernel matrices for which the performance of the Nyström approximation will be poor.

Example 7 (Failure of Nyström approximation). Given a sketch dimension $m \leq n \log 2$, consider an empirical kernel matrix K that has a block diagonal form $\text{diag}(K_1, K_2)$, where $K_1 \in \mathbb{R}^{(n-k) \times (n-k)}$ and $K_2 \in \mathbb{R}^{k \times k}$ for any integer $k \leq \frac{n}{m} \log 2$. Then the probability of not sampling any of the last k columns/rows is at least $1 - (1 - k/n)^m \geq 1 - e^{-km/n} \geq 1/2$. This means that with probability at least $1/2$, the sub-sampling sketch matrix can be expressed as $S = (S_1, 0)$, where $S_1 \in \mathbb{R}^{m \times (n-k)}$. Under such an event, the sketched KRR (5.5a) takes on a degenerate form, namely

$$\hat{\alpha} = \arg \min_{\theta \in \mathbb{R}^m} \left\{ \frac{1}{2} \alpha^T S_1 K_1^2 S_1^T \alpha - \alpha^T S_1 \frac{K_1 y_1}{\sqrt{n}} + \lambda_n \alpha^T S_1 K_1 S_1^T \alpha \right\},$$

and objective that depends only on the first $n - k$ observations. Since the values of the last k observations can be arbitrary, this degeneracy has the potential to lead to substantial approximation error.

The previous example suggests that the Nyström approximation is likely to be very sensitive to non-inhomogeneity in the sampling of covariates. In order to explore this conjecture, we performed some additional simulations, this time comparing both Gaussian and ROS sketches with the uniform Nyström approximation sketch. Returning again to the Gaussian kernel $\mathcal{K}_{\text{Gau}}(u, v) = e^{-\frac{1}{2h^2}(u-v)^2}$ with bandwidth $h = 0.25$, and the function $f^*(x) = -1 + 2x^2$, we first generated n i.i.d. samples that were uniform on the unit interval $[0, 1]$. We then implemented sketches of various types (Gaussian, ROS or Nyström) using a sketch dimension $m = \lceil 4\sqrt{\log n} \rceil$. As shown in the top row (panels (a) and (b)) of Figure 5.3, all three sketch types perform very well for this regular design, with prediction error that is essentially indistinguishable from the original KRR estimate. Keeping the same kernel and function, we then considered an irregular form of design, namely with $k = \lceil \sqrt{n} \rceil$ samples perturbed as follows:

$$x_i \sim \begin{cases} \text{Unif}[0, 1/2] & \text{if } i = 1, \dots, n - k \\ 1 + z_i & \text{for } i = k + 1, \dots, n \end{cases}$$

where each $z_i \sim N(0, 1/n)$. The performance of the sketched estimators in this case are shown in the bottom row (panels (c) and (d)) of Figure 5.3. As before, both

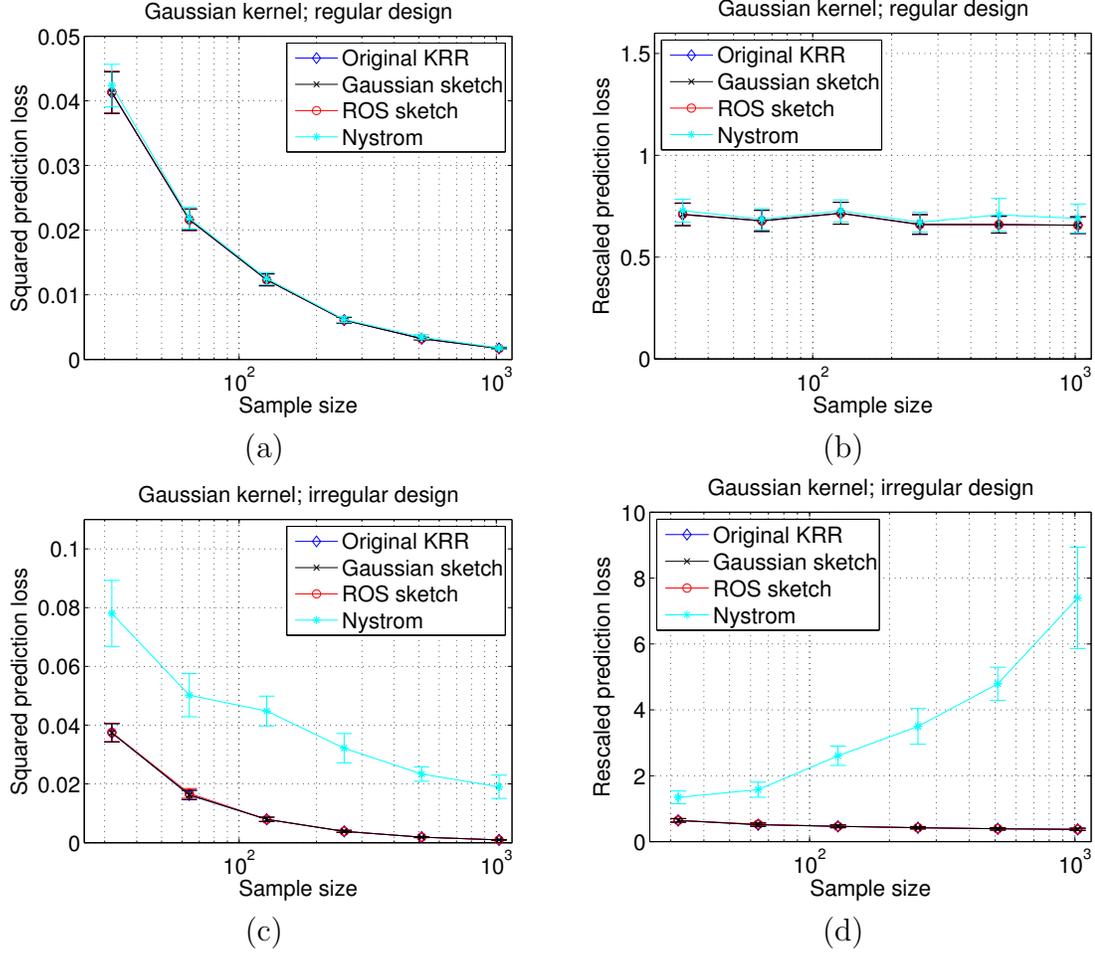


Figure 5.3: Prediction error versus sample size for original KRR, Gaussian sketch, ROS sketch and Nyström approximation. Left panels (a) and (c) shows $\|\hat{f} - f^*\|_n^2$ versus the sample size $n \in \{32, 64, 128, 256, 512, 1024\}$ for projection dimension $m = \lceil 4\sqrt{\log n} \rceil$. In all cases, each point corresponds to the average of 100 trials, with standard errors also shown. Right panels (b) and (d) show the rescaled prediction error $\frac{n}{\sqrt{\log n}} \|\hat{f} - f^*\|_n^2$ versus the sample size. Top row correspond to covariates arranged uniformly on the unit interval, whereas bottom row corresponds to an irregular design (see text for details).

the Gaussian and ROS sketches track the performance of the original KRR estimate very closely; in contrast, the Nyström approximation behaves very poorly for this regression problem, consistent with the intuition suggested by the preceding example.

As is known from general theory on the Nyström approximation, its performance can be improved by knowledge of the so-called leverage scores of the underlying matrix. In this vein, recent work by Alaoui and Mahoney [7] suggests a Nyström approx-

imation non-uniform sampling of the columns of kernel matrix involving the leverage scores. Assuming that the leverage scores are known, they show that their method matches the performance of original KRR using a non-uniform sub-sample of the order $\text{trace}(K(K + \lambda_n I)^{-1}) \log n$ columns. When the regularization parameter λ_n is set optimally—that is, proportional to δ_n^2 —then apart from the extra logarithmic factor, this sketch size scales with the statistical dimension, as defined here. However, the leverage scores are *not known*, and their method for obtaining a sufficiently approximation requires sampling \tilde{p} columns of the kernel matrix K , where

$$\tilde{p} \gtrsim \lambda_n^{-1} \text{trace}(K) \log n.$$

For a typical (normalized) kernel matrix K , we have $\text{trace}(K) \gtrsim 1$; moreover, in order to achieve the minimax rate, the regularization parameter λ_n should scale with δ_n^2 . Putting together the pieces, we see that the sampling parameter \tilde{p} must satisfy the lower bound $\tilde{p} \gtrsim \delta_n^{-2} \log n$. This requirement is much larger than the statistical dimension, and prohibitive in many cases:

- for the Gaussian kernel, we have $\delta_n^2 \asymp \frac{\sqrt{\log(n)}}{n}$, and so $\tilde{p} \gtrsim n \log^{1/2}(n)$, meaning that all rows of the kernel matrix are sampled. In contrast, the statistical dimension scales as $\sqrt{\log n}$.
- for the first-order Sobolev kernel, we have $\delta_n^2 \asymp n^{-2/3}$, so that $\tilde{p} \gtrsim n^{2/3} \log n$. In contrast, the statistical dimension for this kernel scales as $n^{1/3}$.

It remains an open question as to whether a more efficient procedure for approximating the leverage scores might be devised, which would allow a method of this type to be statistically optimal in terms of the sampling dimension.

5.3 Proofs of technical results

In this section, we provide the proofs of our main theorems. Some technical proofs of the intermediate results are provided in later sections.

5.3.1 Proof of Theorem 10

Recall the definition (5.14) of the estimate f^\dagger , as well as the upper bound (5.15) in terms of approximation and estimation error terms. The remainder of our proof consists of two technical lemmas used to control these two terms.

Lemma 26 (Control of estimation error). *Under the conditions of Theorem 10, we have*

$$\|f^\dagger - \hat{f}\|_n^2 \leq c \delta_n^2 \tag{5.18}$$

with probability at least $1 - c_1 e^{-c_2 n \delta_n^2}$.

Lemma 27 (Control of approximation error). *For any K -satisfiable sketch matrix S , we have*

$$\|f^\dagger - f^*\|_n^2 \leq c \{\lambda_n + \delta_n^2\} \quad \text{and} \quad \|f^\dagger\|_{\mathcal{H}} \leq c \left\{1 + \frac{\delta_n^2}{\lambda_n}\right\}. \quad (5.19)$$

These two lemmas, in conjunction with the upper bound (5.15), yield the claim in the theorem statement. Accordingly, it remains to prove the two lemmas.

5.3.1.1 Proof of Lemma 26

So as to simplify notation, we assume throughout the proof that $\sigma = 1$. (A simple rescaling argument can be used to recover the general statement). Since α^\dagger is optimal for the quadratic program (5.14), it must satisfy the zero gradient condition

$$-SK\left(\frac{1}{\sqrt{n}}f^* - KS^T\alpha^\dagger\right) + 2\lambda_n SKS^T\alpha^\dagger = 0. \quad (5.20)$$

By the optimality of $\hat{\alpha}$ and feasibility of α^\dagger for the sketched problem (5.5a), we have

$$\begin{aligned} & \frac{1}{2}\|KS^T\hat{\alpha}\|_2^2 - \frac{1}{\sqrt{n}}y^T KS^T\hat{\alpha} + \lambda_n\|K^{1/2}S^T\hat{\alpha}\|_2^2 \\ & \leq \frac{1}{2}\|KS^T\alpha^\dagger\|_2^2 - \frac{1}{\sqrt{n}}y^T KS^T\alpha^\dagger + \lambda_n\|K^{1/2}S^T\alpha^\dagger\|_2^2 \end{aligned}$$

Defining the error vector $\hat{\Delta} := S^T(\hat{\alpha} - \alpha^\dagger)$, some algebra leads to the following inequality

$$\frac{1}{2}\|K\hat{\Delta}\|_2^2 \leq -\langle K\hat{\Delta}, KS^T\alpha^\dagger \rangle + \frac{1}{\sqrt{n}}y^T K\hat{\Delta} + \lambda_n\|K^{1/2}S^T\alpha^\dagger\|_2^2 - \lambda_n\|K^{1/2}S^T\hat{\alpha}\|_2^2. \quad (5.21)$$

Consequently, by plugging in $y = z^* + w$ and applying the optimality condition (5.20), we obtain the basic inequality

$$\frac{1}{2}\|K\hat{\Delta}\|_2^2 \leq \left| \frac{1}{\sqrt{n}}w^T K\hat{\Delta} \right| - \lambda_n\|K^{1/2}\hat{\Delta}\|_2^2. \quad (5.22)$$

The following lemma provides control on the right-hand side:

Lemma 28. *With probability at least $1 - c_1 e^{-c_2 n \delta_n^2}$, we have*

$$\left| \frac{1}{\sqrt{n}}w^T K\hat{\Delta} \right| \leq \begin{cases} 6\delta_n\|K\hat{\Delta}\|_2 + 2\delta_n^2 & \text{for all } \|K^{1/2}\hat{\Delta}\|_2 \leq 1, \\ 2\delta_n\|K\hat{\Delta}\|_2 + 2\delta_n^2\|K^{1/2}\hat{\Delta}\|_2 + \frac{1}{16}\delta_n^2 & \text{for all } \|K^{1/2}\hat{\Delta}\|_2 \geq 1. \end{cases} \quad (5.23)$$

See Section 5.6.2 for the proof of this lemma.

Based on this auxiliary result, we divide the remainder of our analysis into two cases:

5.3.1.1.1 Case 1: If $\|K^{1/2}\widehat{\Delta}\|_2 \leq 1$, then the basic inequality (5.22) and the top inequality in Lemma 28 imply

$$\frac{1}{2}\|K\widehat{\Delta}\|_2^2 \leq \left| \frac{1}{\sqrt{n}}w^TK\widehat{\Delta} \right| \leq 6\delta_n\|K\widehat{\Delta}\|_2 + 2\delta_n^2 \quad (5.24)$$

with probability at least $1 - c_1e^{-c_2n\delta_n^2}$. Note that we have used that fact that the randomness in the sketch matrix S is independent of the randomness in the noise vector w . The quadratic inequality (5.24) implies that $\|K\widehat{\Delta}\|_2 \leq c\delta_n$ for some universal constant c .

5.3.1.1.2 Case 2: If $\|K^{1/2}\widehat{\Delta}\|_2 > 1$, then the basic inequality (5.22) and the bottom inequality in Lemma 28 imply

$$\frac{1}{2}\|K\widehat{\Delta}\|_2^2 \leq 2\delta_n\|K\widehat{\Delta}\|_2 + 2\delta_n^2\|K^{1/2}\widehat{\Delta}\|_2 + \frac{1}{16}\delta_n^2 - \lambda_n\|K^{1/2}\widehat{\Delta}\|_2^2$$

with probability at least $1 - c_1e^{-c_2n\delta_n^2}$. If $\lambda_n \geq 2\delta_n^2$, then under the assumed condition $\|K^{1/2}\widehat{\Delta}\|_2 > 1$, the above inequality gives

$$\frac{1}{2}\|K\widehat{\Delta}\|_2^2 \leq 2\delta_n\|K\widehat{\Delta}\|_2 + \frac{1}{16}\delta_n^2 \leq \frac{1}{4}\|K\widehat{\Delta}\|_2^2 + 4\delta_n^2 + \frac{1}{16}\delta_n^2.$$

By rearranging terms in the above, we obtain $\|K\widehat{\Delta}\|_2^2 \leq c\delta_n^2$ for a universal constant, which completes the proof.

5.3.1.2 Proof of Lemma 27

Our goal is to show that the bound

$$\frac{1}{2n}\|z^* - \sqrt{n}KS^T\alpha^\dagger\|_2^2 + \lambda_n\|K^{1/2}S^T\alpha^\dagger\|_2^2 \leq c\{\lambda_n + \delta_n^2\}.$$

In fact, since α^\dagger is a minimizer, it suffices to exhibit some $\alpha \in \mathbb{R}^m$ for which this inequality holds. Recalling the eigendecomposition $K = UDU^T$, it is equivalent to exhibit some $\alpha \in \mathbb{R}^m$ such that

$$\frac{1}{2}\|\theta^* - D\widetilde{S}^T\alpha\|_2^2 + \lambda_n\alpha^T\widetilde{S}D\widetilde{S}^T\alpha \leq c\{\lambda_n + \delta_n^2\}, \quad (5.25)$$

where $\tilde{S} = SU$ is the transformed sketch matrix, and the vector $\theta^* = n^{-1/2}Uz^* \in \mathbb{R}^n$ satisfies the ellipse constraint $\|D^{-1/2}\theta^*\|_2 \leq 1$.

We do so via a constructive procedure. First, we partition the vector $\theta^* \in \mathbb{R}^n$ into two sub-vectors, namely $\theta_1^* \in \mathbb{R}^{d_n}$ and $\theta_2^* \in \mathbb{R}^{n-d_n}$. Similarly, we partition the diagonal matrix D into two blocks, D_1 and D_2 , with dimensions d_n and $n - d_n$ respectively. Under the condition $m > d_n$, we may let $\tilde{S}_1 \in \mathbb{R}^{m \times d_n}$ denote the left block of the transformed sketch matrix, and similarly, let $\tilde{S}_2 \in \mathbb{R}^{m \times (n-d_n)}$ denote the right block. In terms of this notation, the assumption that S is K -satisfiable corresponds to the inequalities

$$\|\tilde{S}_1^T \tilde{S}_1 - I_{d_n}\|_2 \leq \frac{1}{2}, \quad \text{and} \quad \|\tilde{S}_2 \sqrt{D_2}\|_2 \leq c\delta_n. \quad (5.26)$$

As a consequence, we are guaranteed that the matrix $\tilde{S}_1^T \tilde{S}_1$ is invertible, so that we may define the m -dimensional vector

$$\hat{\alpha} = \tilde{S}_1(\tilde{S}_1^T \tilde{S}_1)^{-1}(D_1)^{-1}\theta_1^* \in \mathbb{R}^m,$$

Recalling the disjoint partition of our vectors and matrices, we have

$$\|\theta^* - D\tilde{S}^T \hat{\alpha}\|_2^2 = \underbrace{\|\theta_1^* - D_1 \tilde{S}_1^T \hat{\alpha}\|_2^2}_{=0} + \underbrace{\|\theta_2^* - D_2 \tilde{S}_2^T \tilde{S}_1(\tilde{S}_1^T \tilde{S}_1)^{-1} D_1^{-1} \theta_1^*\|_2^2}_{T_1^2} \quad (5.27a)$$

By the triangle inequality, we have

$$\begin{aligned} T_1 &\leq \|\theta_2^*\|_2 + \|D_2 \tilde{S}_2^T \tilde{S}_1(\tilde{S}_1^T \tilde{S}_1)^{-1} D_1^{-1} \theta_1^*\|_2 \\ &\leq \|\theta_2^*\|_2 + \|D_2 \tilde{S}_2^T\|_2 \|\tilde{S}_1\|_2 \|(\tilde{S}_1^T \tilde{S}_1)^{-1}\|_2 \|D_1^{-1/2}\|_2 \|D_1^{-1/2} \theta_1^*\|_2 \\ &\leq \|\theta_2^*\|_2 + \|\sqrt{D_2}\|_2 \|\tilde{S}_2 \sqrt{D_2}\|_2 \|\tilde{S}_1\|_2 \|(\tilde{S}_1^T \tilde{S}_1)^{-1}\|_2 \|D_1^{-1/2}\|_2 \|D_1^{-1/2} \theta_1^*\|_2. \end{aligned}$$

Since $\|D^{-1/2}\theta^*\|_2 \leq 1$, we have $\|D_1^{-1/2}\theta_1^*\|_2 \leq 1$ and moreover

$$\|\theta_2^*\|_2^2 = \sum_{j=d_n+1}^n (\theta_j^*)^2 \leq \delta_n^2 \sum_{j=d_n+1}^n \frac{(\theta_j^*)^2}{\hat{\mu}_j} \leq \delta_n^2,$$

since $\hat{\mu}_j \leq \delta_n^2$ for all $j \geq d_n + 1$. Similarly, we have $\|\sqrt{D_2}\|_2 \leq \sqrt{\hat{\mu}_{d_n+1}} \leq \delta_n$, and $\|D_1^{-1/2}\|_2 \leq \delta_n^{-1}$. Putting together the pieces, we have

$$T_1 \leq \delta_n + \|\tilde{S}_2 \sqrt{D_2}\|_2 \|\tilde{S}_1\|_2 \|(\tilde{S}_1^T \tilde{S}_1)^{-1}\|_2 \leq (c\delta_n) \sqrt{\frac{3}{2}} 2 = c'\delta_n, \quad (5.27b)$$

where we have invoked the K -satisfiability of the sketch matrix to guarantee the bounds $\|\tilde{S}_1\|_2 \leq \sqrt{3/2}$, $\|(\tilde{S}_1^T \tilde{S}_1)\|_2 \geq 1/2$ and $\|\tilde{S}_2 \sqrt{D_2}\|_2 \leq c\delta_n$. Bounds (5.27a) and (5.27b) in conjunction guarantee that

$$\|\theta^* - D\tilde{S}^T \hat{\alpha}\|_2^2 \leq c\delta_n^2, \quad (5.28a)$$

where the value of the universal constant c may change from line to line.

Turning to the remaining term on the left-side of inequality (5.25), applying the triangle inequality and the previously stated bounds leads to

$$\begin{aligned} \widehat{\alpha}^T \widetilde{S} D \widetilde{S}^T \widehat{\alpha} &\leq \|D_1^{-1/2} \theta_1^*\|_2^2 + \|D_2^{1/2} \widetilde{S}_2^T\|_2 \|\widetilde{S}_1\|_2 \\ &\quad \cdot \|(\widetilde{S}_1^T \widetilde{S}_1)^{-1}\|_2 \|D_1^{-1/2}\|_2 \|D_1^{-1/2} \theta_1^*\|_2 \\ &\leq 1 + (c\delta_n) \sqrt{3/2} \frac{1}{2} \delta_n^{-1} (1) \leq c'. \end{aligned} \tag{5.28b}$$

Combining the two bounds (5.28a) and (5.28b) yields the claim (5.25).

5.4 Discussion

In this chapter, we have analyzed randomized sketching methods for kernel ridge regression. Our main theorem gives sufficient conditions on any sketch matrix for the sketched estimate to achieve the minimax risk for non-parametric regression over the underlying kernel class. We specialized this general result to two broad classes of sketches, namely those based on Gaussian random matrices and randomized orthogonal systems (ROS), for which we proved that a sketch size proportional to the statistical dimension is sufficient to achieve the minimax risk. More broadly, we suspect that sketching methods of the type analyzed here have the potential to save time and space in other forms of statistical computation, and we hope that the results given here are useful for such explorations.

5.5 Subsampling sketches yield Nyström approximation

In this section, we show that the the sub-sampling sketch matrix described at the end of Section 5.1.2 coincides with applying Nyström approximation [147] to the kernel matrix.

We begin by observing that the original KRR quadratic program (5.4a) can be written in the equivalent form $\min_{\omega \in \mathbb{R}^n, u \in \mathbb{R}^n} \left\{ \frac{1}{2n} \|u\|^2 + \lambda_n \omega^T K \omega \right\}$ such that $y - \sqrt{n} K \omega = u$.

The dual of this constrained quadratic program (QP) is given by

$$\xi^\dagger = \arg \max_{\xi \in \mathbb{R}^n} \left\{ -\frac{n}{4\lambda_n} \xi^T K \xi + \xi^T y - \frac{1}{2} \xi^T \xi \right\}. \tag{5.29}$$

The KRR estimate f^\dagger and the original solution ω^\dagger can be recovered from the dual solution ξ^\dagger via the relation $f^\dagger(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \omega_i^\dagger \mathcal{K}(\cdot, x_i)$ and $\omega^\dagger = \frac{\sqrt{n}}{2\lambda_n} \xi^\dagger$.

Now turning to the the sketched KRR program (5.5a), note that it can be written in the equivalent form $\min_{\alpha \in \mathbb{R}^n, u \in \mathbb{R}^n} \left\{ \frac{1}{2n} \|u\|^2 + \lambda_n \alpha^T SKS^T \alpha \right\}$ subject to the constraint $y - \sqrt{n}KS^T \alpha = u$. The dual of this constrained QP is given by

$$\xi^\dagger = \arg \max_{\xi \in \mathbb{R}^n} \left\{ -\frac{n}{4\lambda_n} \xi^T \tilde{K} \xi + \xi^T y - \frac{1}{2} \xi^T \xi \right\}, \quad (5.30)$$

where $\tilde{K} = KS^T(SKS^T)^{-1}SK$ is a rank- m matrix in $\mathbb{R}^{n \times n}$. In addition, the sketched KRR estimate \hat{f} , the original solution $\hat{\alpha}$ and the dual solution ξ^\dagger are related by $\hat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (S^T \hat{\alpha})_i \mathcal{K}(\cdot, x_i)$ and $\hat{\alpha} = \frac{\sqrt{n}}{2\lambda_n} (SKS^T)^{-1} SK \xi^\dagger$.

When S is the sub-sampling sketch matrix, the matrix $\tilde{K} = KS^T(SKS^T)^{-1}SK$ is known as the Nyström approximation [147]. Consequently, the dual formulation of sketched KRR based on a sub-sampling matrix can be viewed as the Nyström approximation as applied to the dual formulation of the original KRR problem.

5.6 Proofs of technical results

5.6.1 Proof of Theorem 9

We begin by converting the problem to an instance of the normal sequence model [71]. Recall that the kernel matrix can be decomposed as $K = U^T D U$, where $U \in \mathbb{R}^{n \times n}$ is orthonormal, and $D = \text{diag}\{\hat{\mu}_1, \dots, \hat{\mu}_n\}$. Any function $f^* \in \mathcal{H}$ can be decomposed as

$$f^* = \frac{1}{\sqrt{n}} \sum_{j=1}^n \mathcal{K}(\cdot, x_j) (U^T \beta^*)_j + g, \quad (5.31)$$

for some vector $\beta^* \in \mathbb{R}^n$, and some function $g \in \mathcal{H}$ is orthogonal to $\text{span}\{\mathcal{K}(\cdot, x_j), j = 1, \dots, n\}$. Consequently, the inequality $\|f^*\|_{\mathcal{H}} \leq 1$ implies that

$$\left\| \frac{1}{\sqrt{n}} \sum_{j=1}^n \mathcal{K}(\cdot, x_j) (U^T \beta^*)_j \right\|_{\mathcal{H}}^2 = (U^T \beta^*)^T U^T D U (U^T \beta^*) = \|\sqrt{D} \beta^*\|_2^2 \leq 1.$$

Moreover, we have $f^*(x_1^n) = \sqrt{n} U^T D \beta^*$, and so the original observation model (5.1) has the equivalent form $y = \sqrt{n} U^T \theta^* + w$, where $\theta^* = D \beta^*$. In fact, due to the rotation invariance of the Gaussian, it is equivalent to consider the normal sequence model

$$\tilde{y} = \theta^* + \frac{w}{\sqrt{n}}. \quad (5.32)$$

Any estimate $\tilde{\theta}$ of θ^* defines the function estimate $\tilde{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathcal{K}(\cdot, x_i) (U^T D^{-1} \tilde{\theta})_i$, and by construction, we have $\|\tilde{f} - f^*\|_n^2 = \|\tilde{\theta} - \theta^*\|_2^2$. Finally, the original constraint $\|\sqrt{D} \beta^*\|_2^2 \leq 1$ is equivalent to $\|D^{-1/2} \theta^*\|_2 \leq 1$. Thus, we have a version of the normal sequence model subject to an ellipse constraint.

After this reduction, we can assume that we are given n i.i.d. observations $\tilde{y}_1^n = \{\tilde{y}_1, \dots, \tilde{y}_n\}$, and our goal is to lower bound the Euclidean error $\|\tilde{\theta} - \theta^*\|_2^2$ of any estimate of θ^* . In order to do so, we first construct a $\delta/2$ -packing of the set $\mathcal{B} = \{\theta \in \mathbb{R}^n \mid \|D^{-1/2} \theta\|_2 \leq 1\}$, say $\{\theta^1, \dots, \theta^M\}$. Now consider the random ensemble of regression problems in which we first draw an index A uniformly at random from the index set $[M]$, and then conditioned on $A = a$, we observe n i.i.d. samples from the non-parametric regression model with $f^* = f^a$. Given this set-up, a standard argument using Fano's inequality implies that

$$\mathbb{P}[\|\tilde{f} - f^*\|_n^2 \geq \frac{\delta^2}{4}] \geq 1 - \frac{I(\tilde{y}_1^n; A) + \log 2}{\log M},$$

where $I(\tilde{y}_1^n; A)$ is the mutual information between the samples \tilde{y}_1^n and the random index A . It remains to construct the desired packing and to upper bound the mutual information.

For a given $\delta > 0$, define the ellipse

$$\mathcal{E}(\delta) := \left\{ \theta \in \mathbb{R}^n \mid \underbrace{\sum_{j=1}^n \frac{\theta_j^2}{\min\{\delta^2, \hat{\mu}_j\}}}_{\|\theta\|_{\mathcal{E}}^2} \leq 1 \right\}. \quad (5.33)$$

By construction, observe that $\mathcal{E}(\delta)$ is contained within Hilbert ball of unit radius. Consequently, it suffices to construct a $\delta/2$ -packing of this ellipse in the Euclidean norm.

Lemma 29. *For any $\delta \in (0, \delta_n]$, there is a $\delta/2$ -packing of the ellipse $\mathcal{E}(\delta)$ with cardinality*

$$\log M = \frac{1}{64} d_n. \quad (5.34)$$

Taking this packing as given, note that by construction, we have

$$\|\theta^a\|_2^2 = \delta^2 \sum_{j=1}^n \frac{(\theta^a_j)^2}{\delta^2} \leq \delta^2, \quad \text{and hence} \quad \|\theta^a - \theta^b\|_2^2 \leq 4\delta^2.$$

In conjunction with concavity of the KL divergence, we have

$$I(y_1^n; J) \leq \frac{1}{M^2} \sum_{a,b=1}^M D(\mathbb{P}^a \parallel \mathbb{P}^b) = \frac{1}{M^2} \frac{n}{2\sigma^2} \sum_{a,b=1}^M \|\theta^a - \theta^b\|_2^2 \leq \frac{2n}{\sigma^2} \delta^2$$

For any δ such that $\log 2 \leq \frac{2n}{\sigma^2} \delta^2$ and $\delta \leq \delta_n$, we have

$$\mathbb{P}\left[\|\tilde{f} - f^*\|_n^2 \geq \frac{\delta^2}{4}\right] \geq 1 - \frac{4n\delta^2/\sigma^2}{d_n/64}.$$

Moreover, since the kernel is regular, we have $\sigma^2 d_n \geq cn\delta_n^2$ for some positive constant c . Thus, setting $\delta^2 = \frac{c\delta_n^2}{512}$ yields the claim.

5.6.1.0.1 Proof of Lemma 29: It remains to prove the lemma, and we do so via the probabilistic method. Consider a random vector $\theta \in \mathbb{R}^n$ of the form

$$\theta = \left[\frac{\delta}{\sqrt{2d_n}} w_1 \quad \frac{\delta}{\sqrt{2d_n}} w_2 \quad \cdots \quad \frac{\delta}{\sqrt{2d_n}} w_{d_n} \quad 0 \quad \cdots \quad 0 \right], \quad (5.35)$$

where $w = (w_1, \dots, w_{d_n})^T \sim N(0, I_{d_n})$ is a standard Gaussian vector. We claim that a collection of M such random vectors $\{\theta^1, \dots, \theta^M\}$, generated in an i.i.d. manner, defines the required packing with high probability.

On one hand, for each index $a \in [M]$, since $\delta^2 \leq \delta_n^2 \leq \hat{\mu}_j$ for each $j \leq d_n$, we have $\|\theta^a\|_{\mathcal{E}}^2 = \frac{\|w^a\|_2^2}{2d_n}$, corresponding to a normalized χ^2 -variate. Consequently, by a combination of standard tail bounds and the union bound, we have

$$\mathbb{P}\left[\|\theta^a\|_{\mathcal{E}}^2 \leq 1 \quad \text{for all } a \in [M]\right] \geq 1 - M e^{-\frac{d_n}{16}}.$$

Now consider the difference vector $\theta^a - \theta^b$. Since the underlying Gaussian noise vectors w^a and w^b are independent, the difference vector $w^a - w^b$ follows a $N(0, 2I_m)$ distribution. Consequently, the event $\|\theta^a - \theta^b\|_2 \geq \frac{\delta}{2}$ is equivalent to the event $\sqrt{2}\|\theta\|_2 \geq \frac{\delta}{2}$, where θ is a random vector drawn from the original ensemble. Note that $\|\theta\|_2^2 = \delta^2 \frac{\|w\|_2^2}{2d_n}$. Then a combination of standard tail bounds for χ^2 -distributions and the union bound argument yields

$$\mathbb{P}\left[\|\theta^a - \theta^b\|_2^2 \geq \frac{\delta^2}{4} \quad \text{for all } a, b \in [M]\right] \geq 1 - M^2 e^{-\frac{d_n}{16}}.$$

Combining the last two display together, we obtain

$$\begin{aligned} & \mathbb{P}\left[\|\theta^a\|_{\mathcal{E}}^2 \leq 1 \text{ and } \|\theta^a - \theta^b\|_2^2 \geq \frac{\delta^2}{4} \quad \text{for all } a, b \in [M]\right] \\ & \geq 1 - M e^{-\frac{d_n}{16}} - M^2 e^{-\frac{d_n}{16}}. \end{aligned}$$

This probability is positive for $\log M = d_n/64$.

5.6.2 Proof of Lemma 28

For use in the proof, for each $\delta > 0$, let us define the random variable

$$Z_n(\delta) = \sup_{\substack{\|K^{1/2}\Delta\|_2 \leq 1 \\ \|K\Delta\|_2 \leq \delta}} \left| \frac{1}{\sqrt{n}} w^T K \Delta \right|. \quad (5.36)$$

5.6.2.0.2 Top inequality in the bound (5.23): If the top inequality is violated, then we claim that we must have $Z_n(\delta_n) > 2\delta_n^2$. On one hand, if the bound (5.23) is violated by some vector $\Delta \in \mathbb{R}^n$ with $\|K\Delta\|_2 \leq \delta_n$, then we have

$$2\delta_n^2 \leq \left| \frac{1}{\sqrt{n}} w^T K \Delta \right| \leq Z_n(\delta_n).$$

On the other hand, if the bound is violated by some function with $\|K\Delta\|_2 > \delta_n$, then we can define the rescaled vector $\tilde{\Delta} = \frac{\delta_n}{\|K\Delta\|_2} \Delta$, for which we have

$$\|K\tilde{\Delta}\|_2 = \delta_n, \quad \text{and} \quad \|K^{1/2}\tilde{\Delta}\|_2 = \frac{\delta_n}{\|K\Delta\|_2} \|K^{1/2}\Delta\|_2 \leq 1$$

showing that $Z_n(\delta_n) \geq 2\delta_n^2$ as well.

When viewed as a function of the standard Gaussian vector $w \in \mathbb{R}^n$, it is easy to see that $Z_n(\delta_n)$ is Lipschitz with parameter δ_n/\sqrt{n} . Consequently, by concentration of measure for Lipschitz functions of Gaussians [84], we have

$$\mathbb{P}[Z_n(\delta_n) \geq \mathbb{E}[Z_n(\delta_n)] + t] \leq e^{-\frac{nt^2}{2\delta_n^2}}. \quad (5.37)$$

Moreover, we claim that

$$\mathbb{E}[Z_n(\delta_n)] \stackrel{(i)}{\leq} \underbrace{\sqrt{\frac{1}{n} \sum_{i=1}^n \min\{\delta_n^2, \hat{\mu}_j\}}}_{\hat{\mathcal{R}}(\delta_n)} \stackrel{(ii)}{\leq} \delta_n^2 \quad (5.38)$$

where inequality (ii) follows by definition of the critical radius (recalling that we have set $\sigma = 1$ by a rescaling argument). Setting $t = \delta_n^2$ in the tail bound (5.37), we see that $\mathbb{P}[Z_n(\delta_n) \geq 2\delta_n^2] \leq e^{n\delta_n^2/2}$, which completes the proof of the top bound.

It only remains to prove inequality (i) in equation (5.38). The kernel matrix K can be decomposed as $K = U^T D U$, where $D = \text{diag}\{\hat{\mu}_1, \dots, \hat{\mu}_n\}$, and U is a unitary matrix. Defining the vector $\beta = D U \Delta$, the two constraints on Δ can be expressed as

$\|D^{-1/2}\beta\|_2 \leq 1$ and $\|\beta\|_2 \leq \delta$. Note that any vector satisfying these two constraints must belong to the ellipse

$$\mathcal{E} := \left\{ \beta \in \mathbb{R}^n \mid \sum_{j=1}^n \frac{\beta_j^2}{\nu_j} \leq 2 \quad \text{where } \nu_j = \max\{\delta_n^2, \hat{\mu}_j\} \right\}.$$

Consequently, we have

$$\mathbb{E}[Z_n(\delta_n)] \leq \mathbb{E} \left[\sup_{\beta \in \mathcal{E}} \frac{1}{\sqrt{n}} |\langle U^T w, \beta \rangle| \right] = \mathbb{E} \left[\sup_{\beta \in \mathcal{E}} \frac{1}{\sqrt{n}} |\langle w, \beta \rangle| \right],$$

since $U^T w$ also follows a standard normal distribution. By the Cauchy-Schwarz inequality, we have

$$\mathbb{E} \left[\sup_{\beta \in \mathcal{E}} \frac{1}{\sqrt{n}} |\langle w, \beta \rangle| \right] \leq \frac{1}{\sqrt{n}} \mathbb{E} \sqrt{\sum_{j=1}^n \nu_j w_j^2} \leq \underbrace{\frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^n \nu_j}}_{\widehat{\mathcal{R}}(\delta_n)},$$

where the final step follows from Jensen's inequality.

5.6.2.0.3 Bottom inequality in the bound (5.23): We now turn to the proof of the bottom inequality. We claim that it suffices to show that

$$\left| \frac{1}{\sqrt{n}} w^T K \widetilde{\Delta} \right| \leq 2 \delta_n \|K \widetilde{\Delta}\|_2 + 2 \delta_n^2 + \frac{1}{16} \|K \widetilde{\Delta}\|_2^2 \quad (5.39)$$

for all $\widetilde{\Delta} \in \mathbb{R}^n$ such that $\|K^{1/2} \widetilde{\Delta}\|_2 = 1$. Indeed, for any vector $\Delta \in \mathbb{R}^n$ with $\|K^{1/2} \Delta\|_2 > 1$, we can define the rescaled vector $\widetilde{\Delta} = \Delta / \|K^{1/2} \Delta\|_2$, for which we have $\|K^{1/2} \widetilde{\Delta}\|_2 = 1$. Applying the bound (5.39) to this choice and then multiplying both sides by $\|K^{1/2} \Delta\|_2$, we obtain

$$\begin{aligned} \left| \frac{1}{\sqrt{n}} w^T K \Delta \right| &\leq 2 \delta_n \|K \Delta\|_2 + 2 \delta_n^2 \|K^{1/2} \Delta\|_2 + \frac{1}{16} \frac{\|K \Delta\|_2^2}{\|K^{1/2} \Delta\|_2} \\ &\leq 2 \delta_n \|K \Delta\|_2 + 2 \delta_n^2 \|K^{1/2} \Delta\|_2 + \frac{1}{16} \|K \Delta\|_2^2, \end{aligned}$$

as required.

Recall the family of random variables Z_n previously defined (5.36). For any $u \geq \delta_n$, we have

$$\mathbb{E}[Z_n(u)] = \widehat{\mathcal{R}}(u) = u \frac{\widehat{\mathcal{R}}(u)}{u} \stackrel{(i)}{\leq} u \frac{\widehat{\mathcal{R}}(\delta_n)}{\delta_n} \stackrel{(ii)}{\leq} u \delta_n,$$

where inequality (i) follows since the function $u \mapsto \frac{\widehat{\mathcal{R}}(u)}{u}$ is non-increasing, and step (ii) follows by our choice of δ_n . Setting $t = \frac{u^2}{32}$ in the concentration bound (5.37), we conclude that

$$\mathbb{P}\left[Z_n(u) \geq u\delta_n + \frac{u^2}{64}\right] \leq e^{-cnu^2} \quad \text{for each } u \geq \delta_n. \quad (5.40)$$

We are now equipped to prove the bound (5.39) via a ‘‘peeling’’ argument. Let \mathcal{E} denote the event that the bound (5.39) is violated for some vector $\widetilde{\Delta}$ with $\|K^{1/2}\widetilde{\Delta}\|_2 = 1$. For real numbers $0 \leq a < b$, let $\mathcal{E}(a, b)$ denote the event that it is violated for some vector with $\|K^{1/2}\Delta\|_2 = 1$ and $\|K\widetilde{\Delta}\|_2 \in [a, b]$. For $m = 0, 1, 2, \dots$, define $u_m = 2^m\delta_n$. We then have the decomposition $\mathcal{E} = \mathcal{E}(0, u_0) \cup \left(\bigcup_{m=0}^{\infty} \mathcal{E}(u_m, u_{m+1})\right)$ and hence by union bound,

$$\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{E}(0, u_0)] + \sum_{m=0}^{\infty} \mathbb{P}[\mathcal{E}(u_m, u_{m+1})]. \quad (5.41)$$

The final step is to bound each of the terms in this summation, Since $u_0 = \delta_n$, we have

$$\mathbb{P}[\mathcal{E}(0, u_0)] \leq \mathbb{P}[Z_n(\delta_n) \geq 2\delta_n^2] \leq e^{-cn\delta_n^2}. \quad (5.42)$$

On the other hand, suppose that $\mathcal{E}(u_m, u_{m+1})$ holds, meaning that there exists some vector $\widetilde{\Delta}$ with $\|K^{1/2}\widetilde{\Delta}\|_2 = 1$ and $\|K\widetilde{\Delta}\|_2 \in [u_m, u_{m+1}]$ such that

$$\begin{aligned} \left| \frac{1}{\sqrt{n}} w^T K \widetilde{\Delta} \right| &\geq 2\delta_n \|K\widetilde{\Delta}\|_2 + 2\delta_n^2 + \frac{1}{16} \|K\widetilde{\Delta}\|_2^2 \\ &\geq 2\delta_n u_m + 2\delta_n^2 + \frac{1}{16} u_m^2 \\ &\geq \delta_n u_{m+1} + \frac{1}{64} u_{m+1}^2, \end{aligned}$$

where the second inequality follows since $\|K\widetilde{\Delta}\|_2 \geq u_m$; and the third inequality follows since $u_{m+1} = 2u_m$. This lower bound implies that $Z_n(u_{m+1}) \geq \delta_n u_{m+1} + \frac{u_{m+1}^2}{64}$, whence the bound (5.40) implies that

$$\mathbb{P}[\mathcal{E}(u_m, u_{m+1})] \leq e^{-cnu_{m+1}^2} \leq e^{-cn2^{2m}\delta_n^2}.$$

Combining this tail bound with our earlier bound (5.42) and substituting into the union bound (5.41) yields

$$\mathbb{P}[\mathcal{E}] \leq e^{-cn\delta_n^2} + \sum_{m=0}^{\infty} \exp(-cn2^{2m}\delta_n^2) \leq c_1 e^{-c_2 n \delta_n^2},$$

as claimed.

5.6.3 Proof of Corollary 13

Based on Theorem 10, we need to verify that the stated lower bound (5.16a) on the projection dimension is sufficient to guarantee that a random sketch matrix is K -satisfiable with high probability. In particular, let us state this guarantee as a formal claim:

Lemma 30. *Under the lower bound (5.16a) on the sketch dimension, a {Gaussian, ROS} random sketch is K -satisfiable with probability at least $\phi(m, d_n, n)$.*

We split our proof into two parts, one for each inequality in the definition (5.13) of K -satisfiability.

5.6.3.1 Proof of inequality (i):

We need to bound the operator norm of the matrix $Q = U_1^T S^T S U_1 - I_{d_n}$, where the matrix $U_1 \in \mathbb{R}^{n \times d_n}$ has orthonormal columns. Let $\{v^1, \dots, v^N\}$ be a $1/2$ -cover of the Euclidean sphere \mathcal{S}^{d_n-1} ; by standard arguments [93], we can find such a set with $N \leq e^{2d_n}$ elements. Using this cover, a straightforward discretization argument yields

$$\|Q\|_2 \leq 4 \max_{j,k=1,\dots,N} \langle v^j, Qv^k \rangle = 4 \max_{j,k=1,\dots,N} (\tilde{v})^j \left\{ S^T S - I_n \right\} \tilde{v}^k,$$

where $\tilde{v}^j := U_1 v^j \in \mathcal{S}^{n-1}$, and $\tilde{Q} = S^T S - I_n$. In the Gaussian case, standard sub-exponential bounds imply that $\mathbb{P}[(\tilde{v})^j \tilde{Q} \tilde{v}^k \geq 1/8] \leq c_1 e^{-c_2 m}$, and consequently, by the union bound, we have

$$\mathbb{P}[\|Q\|_2 \geq 1/2] \leq c_1 e^{-c_2 m + 4d_n} \leq c_1 e^{-c'_2 m},$$

where the second and third steps use the assumed lower bound on m . In the ROS case, results of Krahmer and Ward [80] imply that

$$\mathbb{P}[\|Q\|_2 \geq 1/2] \leq c_1 e^{-c_2 \frac{m}{\log^4(n)}}.$$

where the final step uses the assumed lower bound on m .

5.6.3.2 Proof of inequality (ii):

We split this claim into two sub-parts: one for Gaussian sketches, and the other for ROS sketches. Throughout the proof, we make use of the $n \times n$ diagonal matrix $\bar{D} = \text{diag}(0_{d_n}, D_2)$, with which we have $SU_2 D_2^{1/2} = S\bar{D}^{1/2}$.

5.6.3.2.1 Gaussian case: By the definition of the matrix spectral norm, we know

$$\|SU\bar{D}^{1/2}\|_2 := \sup_{\substack{u \in \mathcal{S}^{m-1} \\ v \in \mathcal{E}}} \langle u, Sv \rangle, \quad (5.43)$$

where $\mathcal{E} = \{v \in \mathbb{R}^n \mid \|U\bar{D}v\|_2 \leq 1\}$, and $\mathcal{S}^{m-1} = \{u \in \mathbb{R}^m \mid \|u\|_2 = 1\}$.

We may choose a $1/2$ -cover $\{u^1, \dots, u^M\}$ of the set \mathcal{S}^{m-1} of the set with $\log M \leq 2m$ elements. We then have

$$\begin{aligned} \|SU\bar{D}^{1/2}\|_2 &\leq \max_{j \in [M]} \sup_{v \in \mathcal{E}} \langle u^j, Sv \rangle + \frac{1}{2} \sup_{\substack{u \in \mathcal{S}^{d_n-1} \\ v \in \mathcal{E}}} \langle u, Sv \rangle \\ &= \max_{j \in [M]} \sup_{v \in \mathcal{E}} \langle u^j, Sv \rangle + \frac{1}{2} \|SU\bar{D}^{1/2}\|_2, \end{aligned}$$

and re-arranging implies that

$$\|SU\bar{D}^{1/2}\|_2 \leq 2 \underbrace{\max_{j \in [M]} \sup_{v \in \mathcal{E}} \langle u^j, \tilde{S}v \rangle}_{\tilde{Z}}.$$

For each fixed $u^j \in \mathcal{S}^{d_n-1}$, consider the random variable $Z^j := \sup_{v \in \mathcal{E}} \langle u^j, Sv \rangle$. It is equal in distribution to the random variable $V(g) = \frac{1}{\sqrt{m}} \sup_{v \in \mathcal{E}} \langle g, v \rangle$, where $g \in \mathbb{R}^n$ is a standard Gaussian vector. For $g, g' \in \mathbb{R}^n$, we have

$$\begin{aligned} |V(g) - V(g')| &\leq \frac{2}{\sqrt{m}} \sup_{v \in \mathcal{E}} |\langle g - g', v \rangle| \\ &\leq \frac{2\|D_2^{1/2}\|_2}{\sqrt{m}} \|g - g'\|_2 \leq \frac{2\delta_n}{\sqrt{m}} \|g - g'\|_2, \end{aligned}$$

where we have used the fact that $\hat{\mu}_j \leq \delta_n^2$ for all $j \geq d_n + 1$. Consequently, by concentration of measure for Lipschitz functions of Gaussian random variables [84], we have

$$\mathbb{P}[V(g) \geq \mathbb{E}[V(g)] + t] \leq e^{-\frac{mt^2}{8\delta_n^2}}. \quad (5.44)$$

Turning to the expectation, we have

$$\mathbb{E}[V(g)] = \frac{2}{\sqrt{m}} \mathbb{E}\|D_2^{1/2}g\|_2 \leq 2\sqrt{\frac{\sum_{j=d_n+1}^n \mu_j}{m}} = 2\sqrt{\frac{n}{m}} \sqrt{\frac{\sum_{j=d_n+1}^n \mu_j}{n}} \leq 2\delta_n \quad (5.45)$$

where the last inequality follows since $m \geq n\delta_n^2$ and $\sqrt{\frac{\sum_{j=d_n+1}^n \mu_j}{n}} \leq \delta_n$. Combining the pieces, we have shown that $\mathbb{P}[Z^j \geq c_0(1 + \epsilon)\delta_n] \leq e^{-c_2m}$ for each $j = 1, \dots, M$. Finally, setting $t = c\delta_n$ in the tail bound (5.44) for a constant $c \geq 1$ large enough to ensure that $\frac{c_2m}{8} \geq 2 \log M$. Taking the union bound over all $j \in [M]$ yields

$$\mathbb{P}[\|SU\bar{D}^{1/2}\|_2 \geq 8c\delta_n] \leq c_1 e^{-\frac{c_2m}{8} + \log M} \leq c_1 e^{-c'_2m}$$

which completes the proof.

5.6.3.2.2 ROS case: Here we pursue a matrix Chernoff argument analogous to that in the paper [137]. Letting $r \in \{-1, 1\}^n$ denote an i.i.d. sequence of Rademacher variables, the ROS sketch can be written in the form $S = PH\text{diag}(r)$, where P is a partial identity matrix scaled by n/m , and the matrix H is orthonormal with elements bounded as $|H_{ij}| \leq c/\sqrt{n}$ for some constant c . With this notation, we can write

$$\|PH\text{diag}(r)\bar{D}^{1/2}\|_2^2 = \left\| \frac{1}{m} \sum_{i=1}^m v_i v_i^T \right\|_2,$$

where $v_i \in \mathbb{R}^n$ are random vectors of the form $\sqrt{n}\bar{D}^{1/2}\text{diag}(r)He$, where $e \in \mathbb{R}^n$ is chosen uniformly at random from the standard Euclidean basis.

We first show that the vectors $\{v_i\}_{i=1}^m$ are uniformly bounded with high probability. Note that we certainly have $\max_{i \in [m]} \|v_i\|_2 \leq \max_{j \in [n]} F_j(r)$, where

$$F_j(r) := \sqrt{n} \|\bar{D}^{1/2}\text{diag}(r)He_j\|_2 = \sqrt{n} \|\bar{D}^{1/2}\text{diag}(He_j)r\|_2.$$

Beginning with the expectation, define the vector $\tilde{r} = \text{diag}(He_j)r$, and note that it has entries bounded in absolute value by c/\sqrt{n} . Thus we have,

$$\mathbb{E}[F_j(r)] \leq \left[n \mathbb{E}[\tilde{r}^T \bar{D} \tilde{r}] \right]^{1/2} \leq c \sqrt{\sum_{j=d_n+1}^n \hat{\mu}_j} \leq c\sqrt{n}\delta_n^2$$

For any two vectors $r, r' \in \mathbb{R}^n$, we have

$$\left| F(r) - F(r') \right| \leq \sqrt{n} \|r - r'\|_2 \|\bar{D}^{1/2}\text{diag}(He_j)\|_2 \leq \delta_n.$$

Consequently, by concentration results for convex Lipschitz functions of Rademacher variables [84], we have

$$\mathbb{P}\left[F_j(r) \geq c_0 \sqrt{n} \delta_n^2 \log n \right] \leq c_1 e^{-c_2 n \delta_n^2 \log^2 n}.$$

Taking the union bound over all n rows, we see that

$$\max_{i \in [n]} \|v_i\|_2 \leq \max_{j \in [n]} F_j(r) \leq 4\sqrt{n}\delta_n^2 \log(n)$$

with probability at least $1 - c_1 e^{-c_2 n \delta_n^2 \log^2(n)}$. Finally, a simple calculation shows that $\|\mathbb{E}[v_1 v_1^T]\|_2 \leq \delta_n^2$. Consequently, by standard matrix Chernoff bounds [135, 137], we have

$$\mathbb{P}\left[\left\| \frac{1}{m} \sum_{i=1}^m v_i v_i^T \right\|_2 \geq 2\delta_n^2 \right] \leq c_1 e^{-c_2 \frac{m\delta_n^2}{n\delta_n^4 \log^2(n)}} + c_1 e^{-c_2 n \delta_n^2 \log^2(n)}, \quad (5.46)$$

from which the claim follows.

Chapter 6

Relaxations of combinatorial optimization problems

Over the past several decades, the rapid increase of data dimensionality and complexity has led a tremendous surge of interest of models for high-dimensional data that incorporate some type of low-dimensional structure. Sparsity is a canonical way of imposing low-dimensional structure, and has received considerable attention in many fields, including statistics, signal processing, machine learning and applied mathematics [49, 134, 144]. Sparse models often typically more interpretable from the scientific standpoint, and they are also desirable from a computational perspective.

The most direct approach to enforcing sparsity in a learning problem is by controlling the ℓ_0 -“norm” of the solution, which counts the number of non-zero entries in a vector. Unfortunately, at least in general, optimization problems involving such an ℓ_0 -constraint are known to be computationally intractable. The classical approach of circumventing this difficulty while still promoting sparsity in the solution is to replace the ℓ_0 -constraint with an ℓ_1 -constraint, or alternatively to augment the objective function with an ℓ_1 -penalty. This approach is well-known and analyzed various assumptions on the data generating mechanisms (e.g., [34, 49, 30, 144]). However, in a typical statistical setting, these mechanisms are not under the user’s control, and it is difficult to verify post hoc that an ℓ_1 -based solution is of suitably high quality.

The main contribution of this chapter is to provide novel frameworks for obtaining approximate solutions to cardinality-constrained problems, and one in which the quality can be easily verified. Our first approach is based on showing a broad class of cardinality-constrained (or penalized) problems can be expressed equivalently as convex programs involving Boolean variables. This reformulation allows us to apply various standard hierarchies of relaxations for Boolean programs, among them Sherali-Adams or Lasserre hierarchies [128, 82, 83, 145]. When the solution of any such relaxation is integral—i.e., belongs to the Boolean hypercube—then it must be

an optimal solution to the original problem. Otherwise, any non-integral solution still provides a lower bound on the minimum over all Boolean solutions.

The simplest relaxation is the first-order one, based on relaxing each Boolean variable to the unit interval $[0, 1]$. We provide an in-depth analysis of the necessary and sufficient conditions for this first-order relaxation to have an integral solution. In the case of least-squares regression, and for a random ensemble of problems of the compressed sensing type [34, 49], we show that the relaxed solution is integral with high probability once the sample size exceeds a critical threshold. In this regime, like ℓ_1 -relaxations, our first-order method recovers the support of sparse vector exactly, but *unlike* ℓ_1 -relaxations, the integral solution also certifies that it has recovered the sparsest solution. Finally, there are many settings in which the first-order relaxation might not be integral. For such cases, we study a form of randomized rounding for generating feasible solutions, and we prove a result that controls the approximation ratio. Our framework also allows to specify a target cardinality unlike methods based on ℓ_1 regularization. This feature is desirable for many applications including portfolio optimization [91], machine learning [46, 111] and control theory [28].

The remainder of this chapter is organized as follows. We begin in Section 6.1 by introducing the problem of sparse learning, and then showing how the constrained version can be reformulated as a convex program in Boolean variables. In Section 6.2, we study the first-order relaxation in some detail, including conditions for exactness as well as analysis of randomized rounding procedures. Section 6.3 is devoted to discuss of the penalized form of sparse learning problems, whereas Section 6.4 discusses numerical issues and applications to real-world data sets. In Section 6.5, we describe a novel relaxation approach for optimization problems with simplex constraints and present applications and numerical simulations.

6.1 General Sparse Learning as a Boolean Problem

We consider a learning problem based on samples of the form $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$. This set-up is flexible enough to model various problems, including regression problems (output space $\mathcal{Y} = \mathbb{R}$), binary classification problems (output space $\mathcal{Y} = \{-1, +1\}$), and so on. Given a collection of n samples $\{(x_i, y_i)\}_{i=1}^n$, our goal is to learn a linear function $x \mapsto \langle x, w \rangle$ that can be used to predict or classify future (unseen) outputs. In order to learn the weight vector $w \in \mathbb{R}^d$, we consider a cardinality-constrained

program of the form

$$P^* := \min_{\substack{w \in \mathbb{R}^d \\ \|w\|_0 \leq k}} \underbrace{\left\{ \sum_{i=1}^n f(\langle x_i, w \rangle; y_i) + \frac{1}{2} \rho \|w\|_2^2 \right\}}_{F(w)} \quad (6.1)$$

As will be clarified, the additional regularization term $\frac{1}{2} \rho \|w\|_2^2$ is useful for convex-analytic reasons, in particular in ensuring strong convexity and coercivity of the objective, and thereby the existence of a unique optimal solution $w^* \in \mathbb{R}^d$. Our results also involve the Legendre-Fenchel conjugate of the function $t \mapsto f(t; y)$, given by (for each fixed $y \in \mathcal{Y}$)

$$f^*(s; y) := \sup_{t \in \mathbb{R}} \{s t - f(t; y)\}. \quad (6.2)$$

Let us consider some examples to illustrate.

Example 8 (Least-squares regression). In the problem of least-squares regression, the outputs are real-valued (see e.g., [28]). Adopting the cost function $f(t, y) = \frac{1}{2}(t - y)^2$ leads to ℓ_0 -constrained problem

$$P^* := \min_{\substack{w \in \mathbb{R}^d \\ \|w\|_0 \leq k}} \underbrace{\left\{ \frac{1}{2} \sum_{i=1}^n (\langle x_i, w \rangle - y_i)^2 + \frac{1}{2} \rho \|w\|_2^2 \right\}}_{F_{LS}(w)} \quad (6.3)$$

This formulation, while close in spirit to elastic net [159], is based on imposing the cardinality constraint exactly, as opposed to in a relaxed form via ℓ_1 -regularization. However, in contrast to the elastic net, it is a nonconvex problem, so that we need to study relaxations of it. A straightforward calculation yields the conjugate dual function

$$f^*(s; y) = \frac{s^2}{2} + s y, \quad (6.4)$$

which will play a role in our relaxations of the nonconvex problem (6.3). ■

The preceding example has a natural extension in terms of generalized linear models:

Example 9 (Generalized linear models). In a generalized linear model, the output $y \in \mathcal{Y}$ is related to the covariate $x \in \mathbb{R}^d$ via a conditional distribution in the exponential form (see e.g. [94, 99])

$$\mathbb{P}_w(y | x) = h(y) \exp(y \langle x, w \rangle - \psi(\langle x, w \rangle)). \quad (6.5)$$

Here $h : \mathbb{R}^d \rightarrow \mathbb{R}_+$ is some fixed function, and $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is the cumulant generating function, given by $\psi(t) = \log \int_{\mathcal{Y}} e^{ty} h(y) dy$. Letting $f(\langle x, w \rangle; y)$ be the negative log-likelihood associated with this family, we obtain the general family of cardinality-constrained likelihood estimates

$$\min_{\substack{w \in \mathbb{R}^d \\ \|w\|_0 \leq k}} \underbrace{\left\{ \sum_{i=1}^n \{ \psi(\langle x_i, w \rangle) - y_i \langle x_i, w \rangle \} + \frac{1}{2} \rho \|w\|_2^2 \right\}}_{F_{\text{GR}}(w)} \quad (6.6)$$

Specifically, least-squares regression is a particular case of the problem (6.6), corresponding to the choice $\psi(t) = t^2/2$. Similarly, logistic regression for binary responses $y \in \{0, 1\}$ can be obtained by setting $\psi(t) = \log(1 + e^t)$.

In the likelihood formulation (6.6), we have $f(t; y) = \psi(t) - yt$, whence conjugate dual takes the form

$$f^*(s; y) = \sup_{t \in \mathbb{R}} \{ st - \psi(t) + yt \} = \psi^*(s + y), \quad (6.7)$$

where ψ^* denotes the conjugate dual of ψ . As particular examples, in the case of logistic regression, the dual of the logistic function $\psi(t) = \log(1 + e^t)$ takes the form $\psi^*(s) = s \log s + (1 - s) \log(1 - s)$ for $s \in [0, 1]$, and takes the value infinity otherwise. ■

As a final example, let us consider a cardinality-constrained version of the support vector machine:

Example 10 (Support vector machine classification). In this case, the outputs are binary $y \in \{-1, 1\}$, and our goal is to learn a linear classifier $x \mapsto \text{sign}(\langle x, w \rangle) \in \{-1, 1\}$ [40]. The cardinality-constrained version of the support vector machine (SVM) is based on minimizing the objective function

$$\min_{\substack{w \in \mathbb{R}^d \\ \|w\|_0 \leq k}} \underbrace{\left\{ \sum_{i=1}^n \phi(y_i \langle x_i, w \rangle) + \frac{1}{2} \rho \|w\|_2^2 \right\}}_{F_{\text{SVM}}(w)}, \quad (6.8)$$

where $\phi(t) = \max\{1 - t, 0\}$ is known as the hinge loss function. The conjugate dual of the hinge loss takes the form

$$\phi^*(s) = \begin{cases} s & \text{if } s \in [-1, 0] \\ \infty & \text{otherwise.} \end{cases}$$

■

Having considered various examples of sparse learning, we now turn to developing an exact Boolean representation that is amenable to various relaxations.

6.1.1 Exact representation as a Boolean convex program

Let us now show how the cardinality-constrained program (6.1) can be represented exactly as a convex program in Boolean variables. This representation, while still nonconvex, is useful because it immediately leads to a hierarchy of relaxations. Given the collection of covariates $\{x_i\}_{i=1}^n$, we let $X \in \mathbb{R}^{n \times d}$ denote the design matrix with $x_i^T \in \mathbb{R}^d$ as its i^{th} row.

Theorem 11 (Exact representation). *Suppose that for each $y \in \mathcal{Y}$, the function $t \mapsto f(t; y)$ is closed and convex. Then for any $\rho > 0$, the cardinality-constrained program (6.1) can be represented exactly as the Boolean convex program*

$$P^* = \min_{\substack{u \in \{0,1\}^d \\ \sum_{j=1}^d u_j \leq k}} \max_{v \in \mathbb{R}^n} \underbrace{\left\{ -\frac{1}{2\rho} v^T X D(u) X^T v - \sum_{i=1}^n f^*(v_i; y_i) \right\}}_{G(u)}, \quad (6.9)$$

where $D(u) := \text{diag}(u) \in \mathbb{R}^{d \times d}$ is a diagonal matrix.

The function $u \mapsto G(u)$ —in particular, defined by maximizing over $v \in \mathbb{R}^n$ —is a maximum of a family of functions that are linear in the vector u , and hence is convex. Thus, apart from the Boolean constraint, all other quantities in the program (6.9) are relatively simple: a linear constraint and a convex objective function. Consequently, we can obtain tractable approximations by relaxing the Boolean constraint. The simplest such approach is to replace the Boolean hypercube $\{0,1\}^d$ with the unit hypercube $[0,1]^d$. Doing so leads the *interval relaxation* of the exact Boolean representation, namely the convex relaxation

$$P_{\text{IR}} = \min_{\substack{u \in [0,1]^d \\ \sum_{j=1}^d u_j \leq k}} \max_{v \in \mathbb{R}^n} \underbrace{\left\{ -\frac{1}{2\rho} v^T X D(u) X^T v - \sum_{i=1}^n f^*(v_i; y_i) \right\}}_{G(u)}. \quad (6.10)$$

Note that this is a convex program, and so can be solved by standard methods. In particular the sub-gradient descent method (e.g., see [105]) can be applied directly if a closed form solution, or a solver for the inner maximization problem is available. In Section 6.2, we return to analyze when the interval relaxation is tight—that is, when $P_{\text{IR}} = P^*$.

In the case of least-squares regression, Theorem 11 and the interval relaxation take an especially simple form, which we state as a corollary.

Corollary 14. *The cardinality constrained problem is equivalent to the Boolean SDP*

$$P^* = \min_{\substack{(u,t) \in \{0,1\}^d \times \mathbb{R}_+ \\ \sum_{j=1}^d u_j \leq k}} t \quad \text{such that} \quad \begin{bmatrix} I_n + \frac{1}{\rho} X D(u) X^T & y \\ y^T & t \end{bmatrix} \succeq 0. \quad (6.11)$$

Thus, the interval relaxation (6.10) is an ordinary SDP in variables $(u, t) \in [0, 1]^d \times \mathbb{R}_+$.

Proof. As discussed in Example 8, the conjugate dual of the least-squares loss $t \mapsto f(t; y) = \frac{1}{2}(t - y)^2$ is given by $f^*(s; y) = \frac{s^2}{2} + sy$. Substituting this dual function into equation (6.9), we find that

$$G(u) = \max_{v \in \mathbb{R}^n} \left\{ -\frac{1}{2}v^T \left(\frac{XD(u)X^T}{\rho} + I \right) v - \langle v, y \rangle \right\},$$

where we have defined the diagonal matrix $D(u) := \text{diag}(u) \in \mathbb{R}^{d \times d}$. Taking derivatives shows that the optimum is achieved at

$$\hat{v} = -\left(\frac{XD(u)X^T}{\rho} + I \right)^{-1} y, \quad (6.12)$$

and substituting back into equation (6.9) and applying Theorem 11 yield the representation

$$P^* = \min_{\substack{u \in \{0,1\}^d \\ \sum_{j=1}^d u_j \leq k}} \left\{ y^T \left(\frac{1}{\rho} XD(u)X^T + I_n \right)^{-1} y \right\}. \quad (6.13)$$

By introducing a slack variable $t \in \mathbb{R}_+$ and using the Schur complement formula (see e.g. [28]), some further calculation shows that this Boolean problem (6.13) is equivalent to the Boolean SDP (6.11), as claimed. \square

We now present the proof of Theorem 11.

Proof. Recalling that $D(u) := \text{diag}(u)$ is a diagonal matrix, for each fixed $u \in \{0, 1\}^d$, consider the change of variable $w \mapsto D(u)w$. With this notation, the original problem (6.1) is equivalent to

$$P^* = \min_{\|D(u)w\|_0 \leq k} \left\{ \sum_{i=1}^n f(\langle D(u)x_i, w \rangle; y_i) + \frac{1}{2}\rho \|D(u)w\|_2^2 \right\}. \quad (6.14)$$

Noting that we can take $w_i = 0$ when $u_i = 0$ and vice-versa, the original problem (6.1) becomes

$$P^* = \min_{\substack{u \in \{0,1\}^d \\ \sum_{j=1}^d u_j \leq k}} \min_{w \in \mathbb{R}^d} \left\{ \sum_{i=1}^n f(\langle D(u)x_i, w \rangle; y_i) + \frac{1}{2}\rho \|w\|_2^2 \right\}. \quad (6.15)$$

It remains to prove that, for each fixed Boolean vector $u \in \{0, 1\}^d$, we have

$$\min_{w \in \mathbb{R}^d} \left\{ \sum_{i=1}^n f(\langle D(u)x_i, w \rangle; y_i) + \frac{1}{2}\rho\|w\|_2^2 \right\} = \max_{v \in \mathbb{R}^n} \left\{ -\frac{1}{2\rho}\|D(u)X^T v\|_2^2 - \sum_{i=1}^n f^*(v_i; y_i) \right\}. \quad (6.16)$$

From the conjugate representation of f , we find that

$$\min_{w \in \mathbb{R}^d} \max_{v \in \mathbb{R}^n} \left\{ \sum_{i=1}^n v_i \langle D(u)x_i, w \rangle - f^*(v_i; y_i) + \frac{1}{2}\rho\|w\|_2^2 \right\}.$$

Under the stated assumptions, strong duality must hold, so that it is permissible to exchange the order of the minimum and maximum. Doing so yields

$$\max_{v \in \mathbb{R}^n} \min_{w \in \mathbb{R}^d} \left\{ \sum_{i=1}^n v_i \langle D(u)x_i, w \rangle - f^*(v_i; y_i) + \frac{1}{2}\rho\|w\|_2^2 \right\}.$$

Finally, strong convexity ensures that the minimum over w is unique: more specifically, it is given by $w^* = \frac{1}{\rho} \sum_{i=1}^n D(u)x_i v_i$. Substituting this optimum yields the claimed equality (6.16). \square

6.2 Convex-analytic conditions for IR exactness

We now turn to analysis of the interval relaxation (6.10), and in particular, determining when it is exact. Note that by strong convexity, the original cardinality-constrained problem (6.1) has a unique solution, say $w^* \in \mathbb{R}^d$. Let S denote the support set of w^* , and let u^* be a Boolean indicator vector for membership in S —that is, $u_j^* = 1$ if $j \in S$ and zero otherwise.

An attractive feature of the IR relaxation is that integrality of an optimal solution \hat{u} to the relaxed problem provides a *certificate of exactness*—that is, if the interval relaxation (6.10) has an optimal solution $\hat{u} \in \{0, 1\}^d$, then it must be the case that $\hat{u} = u^*$ (so that we recover the support set of w^*), and moreover that

$$P_{\text{IR}} = P^*. \quad (6.17)$$

In this case, we are guaranteed to recover the optimal solution w^* of the original problem (6.1) by solving the constrained problem with $w_j = 0$ for all $j \notin S$.

In contrast, methods based on ℓ_1 -relaxations do not provide such certificates of exactness. In the least-squares regression, the use of ℓ_1 -relaxation is known as the Lasso [134], and there is an extensive literature devoted to conditions on the design

matrix $X \in \mathbb{R}^{n \times d}$ under which the ℓ_1 -relaxation provides a “good” solution. Unfortunately, these conditions are either computationally infeasible to check (e.g., restricted eigenvalue, isometry and nullspace conditions [22, 43] and the related irrepresentability conditions for support recovery [60, 95, 157]). Although polynomial-time checkable conditions do exist (such as pairwise incoherence conditions [136, 50, 60]), they provide weak guarantees, only holding for sample sizes much larger than the threshold at which the ℓ_1 -relaxation begins to work. In addition, most of the previous work on analyzing ℓ_1 relaxations considered a statistical data model where there exists a true sparse coefficient generating the response. However in many applications such assumptions do not necessarily hold and it is unclear whether ℓ_1 regularization provides a good optimization heuristic for an arbitrary input data.

It is thus of interest to investigate conditions under which the relaxation (IR) is guaranteed to have an integer solution and hence be tight. The following result provides an if-and-only if characterization.

Proposition 5. *The interval relaxation is tight—that is, $P_{IR} = P^*$ —if and only if there exist a pair $(\lambda, \hat{v}) \in \mathbb{R}_+ \times \mathbb{R}^n$ such that*

$$\hat{v} \in \arg \max_{v \in \mathbb{R}^n} \left\{ -\frac{1}{2\rho} v^T X_S X_S^T v - \sum_{i=1}^n f^*(v_i; y_i) \right\}, \quad \text{and} \quad (6.18a)$$

$$|\langle X_j, \hat{v} \rangle| > \lambda \quad \text{for all } j \in S, \quad \text{and} \quad |\langle X_j, \hat{v} \rangle| < \lambda \quad \text{for all } j \notin S, \quad (6.18b)$$

where $X_j \in \mathbb{R}^n$ denotes the j^{th} column of the design matrix, S denotes the support of the unique optimal solution w^* to the original problem (6.1).

Proof. Beginning with the saddle-point representation from equation (6.10), we apply the first-order convex optimality condition for constrained minimization. More precisely, the relaxed solution \hat{u} is optimal if and only if the following inclusion holds:

$$0 \in \left\{ \partial_u \max_{v \in \mathbb{R}^n} \left\{ -\frac{1}{2\rho} v^T X D(u) X^T v - \sum_{i=1}^n f^*(v_i; y_i) \right\} + \mathbb{N} \right\},$$

where \mathbb{N} denotes the normal cone of the constraint set $\left\{ u \in [0, 1]^d \mid \sum_{j=1}^d u_j \leq k \right\}$. Note that the subgradient with respect to u_j is given by $-(\langle X_j, \hat{v} \rangle)^2$, where the vector \hat{v} was defined in equation (6.18a). Using representation of the normal cone at the integral point u^* and associating $\lambda \geq 0$ as the dual parameter corresponding to constraint $\sum_{j=1}^d u_j$, we arrive at the stated condition (6.18b). \square

In the case of least-squares regression, the conditions of Proposition 5 can be simplified substantially. Recall that interval relaxation for least-squares regression is

given by

$$P_{IR} = \min_{\substack{u \in [0,1]^d \\ \sum_{j=1}^d u_j \leq k}} \left\{ y^T \left(\frac{1}{\rho} XD(u)X^T + I_n \right)^{-1} y \right\}. \quad (6.19)$$

Let S denote the support of the unique optimal solution w^* to the original least-squares problem (6.3), say of cardinality k , and define the $n \times n$ matrix

$$M := (I_n + \rho^{-1} X_S X_S^T)^{-1} \quad (6.20)$$

With this notation, we have:

Corollary 15. *The interval relaxation of cardinality-constrained least-squares is exact ($P_{IR} = P^*$) if and only there exists a scalar $\lambda \in \mathbb{R}_+$ such that*

$$|X_j^T M y| > \lambda \quad \text{for all } j \in S, \text{ and} \quad (6.21a)$$

$$|X_j^T M y| \leq \lambda \quad \text{for all } j \notin S, \quad (6.21b)$$

where $X_j \in \mathbb{R}^n$ denotes the j^{th} column of X .

Proof. From the proof of Corollary 14, recall the Boolean convex program (6.13). As shown in equation (6.12), its optimum is achieved at $\hat{v} = -(I_n + XD(u^*)X^T)y$, where u^* is a Boolean indicator for membership in S . Applying Proposition 5 with this choice of \hat{v} yields the necessary and sufficient conditions

$$\begin{aligned} |y^T (\rho I_n + XD(u^*)X^T)^{-1} X_j| &> \lambda && \text{for all } j \in S, \text{ and} \\ |y^T (\rho I_n + XD(u^*)X^T)^{-1} X_j| &\leq \lambda && \text{for all } j \in S^c, \end{aligned}$$

and completes the proof. □

In order to gain an understanding of the above corollary consider an example where the rows of X_S are orthonormal and $n = k$, hence $M = (I_n + \rho^{-1} I_n)^{-1} = \rho / (1 + \rho) I_n$. Then the conditions for integrality reduce to checking whether there exists $\lambda' \in \mathbb{R}_+$ such that

$$\begin{aligned} |X_j^T y| &> \lambda' && \text{for all } j \in S, \text{ and} \\ |X_j^T y| &\leq \lambda' && \text{for all } j \notin S. \end{aligned}$$

Intuitively the above condition basically checks if the columns in the correct support are more aligned to the response y compared to the columns outside the support.

Also note that by the matrix inversion formula, we have the alternative representation,

$$M = (I_n + \rho^{-1} X_S X_S^T)^{-1} = I_n - X_S (\rho I_d + X_S^T X_S)^{-1} X_S^T,$$

For random ensembles, Corollary 15 allows the use of a primal witness method to certify exactness of the IR method. In particular, if we can construct a scalar λ for which the two bounds (6.21a) and (6.21b) hold with high probability, then we can certify exactness of the relaxation. We illustrate this approach in the following subsection.

6.2.1 Sufficient conditions for random ensembles

In order to assess the performance of the interval relaxation (6.10), we performed some simple experiments for the least squares case, first generating a design matrix $X \in \mathbb{R}^{n \times d}$ with i.i.d. $N(0, 1)$ entries, and then forming the response vector $y = Xw^* + \epsilon$, where the noise vector $\epsilon \in \mathbb{R}^n$ has i.i.d. $N(0, \gamma)$ entries. The unknown regression vector w^* was k -sparse, with absolute entries of the order $1/\sqrt{k}$ on its support. Each such problem can be characterized by the triple (n, d, k) of sample size, dimension and sparsity, and the question of interest is to understand how large the sample size should be in order to ensure exactness of a method. For instance, for this random ensemble, the Lasso is known [143] to perform exact support recovery once $n \gtrsim k \log(d - k)$, and this scaling is information-theoretically optimal [142]. Does the interval relaxation also satisfy this same scaling?

In order to test the IR relaxation, we performed simulations with sample size $n = \alpha k \log d$ for a control parameter $\alpha \in [2, 8]$, for three different problem sizes $d \in \{64, 128, 256\}$ and sparsity $k = \lceil \sqrt{d} \rceil$. Figure 6.1 shows the probability of successful recovery versus the control parameter α for these different problem sizes, for both the Lasso and the IR method. Note that both methods undergo a phase transition once the sample size n is larger than some constant multiple of $k \log(d - k)$.

The following result provides theoretical justification for the phase transition behavior exhibited in Figure 6.1:

Theorem 12. *Suppose that we are given a sample size $n > c_0 \frac{\gamma^2 + \|w_S^*\|_2^2}{w_{\min}^2} \log d$, and that we solve the interval relaxation with $\rho = \sqrt{n}$. Then with probability at least $1 - 2e^{-c_1 n}$, the interval relaxation is integral, so that $P_{IR} = P^*$.*

For a typical k -sparse vector, we have $\frac{\|w_S^*\|_2^2}{w_{\min}^2} \asymp k$, so that Theorem 12 predicts that the interval relaxation should succeed with $n \gtrsim k \log(d - k)$ samples, as confirmed by the plots in Figure 6.1.

6.2.2 Analysis of randomized rounding

In this section, we describe a method to improve the interval relaxation scheme introduced earlier. The convex relaxation of the Boolean hypercube constraint $u \in$

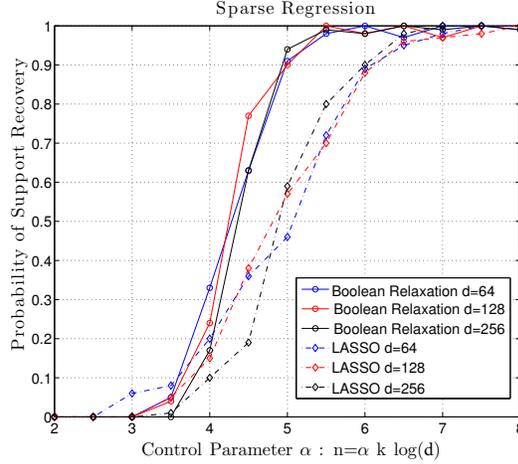


Figure 6.1: Problem of exact support recovery for the Lasso and the interval relaxation for different problem sizes $d \in \{64, 128, 256\}$. As predicted by theory, both methods undergo a phase transition from failure to success once the control parameter $\alpha := \frac{n}{k \log(d-k)}$ is sufficiently large. This behavior is confirmed for the interval relaxation in Theorem 12.

$\{0, 1\}^d$ to the standard hypercube constraint $u \in [0, 1]^d$ might produce an integral solution—in particular, when the conditions in Proposition 5 are not satisfied. In this case, it is natural to consider how to use the fractional solution $\hat{u} \in [0, 1]^d$ to produce a feasible Boolean solution $\tilde{u} \in \{0, 1\}^d$. By construction, the objective function values $(G(\hat{u}), G(\tilde{u}))$ defined by this pair will sandwich the optimal value—viz

$$G(\hat{u}) \leq P^* \leq G(\tilde{u}).$$

Here G is the objective function from the original Boolean problem (6.9).

Randomized rounding is a classical technique for converting fractional solutions into integer solutions with provable approximation guarantees [98]. Here we consider the simplest possible form of randomized rounding in application to our relaxation. Given the fractional solution $\hat{u} \in [0, 1]^d$, suppose that we generate a feasible Boolean solution $\tilde{u} \in \{0, 1\}^d$ as follows

$$\mathbb{P}[\tilde{u}_i = 1] = \hat{u}_i \quad \text{and} \quad \mathbb{P}[\tilde{u}_i = 0] = 1 - \hat{u}_i. \quad (6.22)$$

By construction, this random Boolean vector matches the fractional solution in expectation—that is, $\mathbb{E}[\tilde{u}] = \hat{u}$, and moreover its expected ℓ_0 -norm is given by

$$\mathbb{E}[\|\tilde{u}\|_0] = \sum_{i=1}^d \mathbb{P}[\tilde{u}_i = 1] = \sum_{i=1}^d \hat{u}_i \leq k,$$

where the final inequality uses the feasibility of the fractional solution \hat{u} . The random Boolean solution \tilde{u} can be used to define a randomized solution $\tilde{w} \in \mathbb{R}^d$ of the original problem via

$$\tilde{w} = \arg \min_{w \in \mathbb{R}^d} F(D(\tilde{u})w), \quad (6.23)$$

where the function F was defined in equation (6.1).

Without loss of generality, consider the least squares problem and assume the columns are normalized, i.e., $\|x_j\|_2 = 1$ for $j = 1, \dots, d$ and $\|y\|_2 = 1$, then we have the following result. Let $R \subset \{1, \dots, d\}$ be the subset of coordinates on which \hat{u} takes fractional values (i.e., $\hat{u}_j \in (0, 1)$ for all $j \in R$) and let $r = |R|$ be the cardinality of this set.

Theorem 13. *There are universal constants c_j such that for any $\delta \in (0, 1)$, with probability at least $1 - c_1 e^{-c_2 k \delta^2} - \frac{1}{\min\{r, n\}^{c_3}}$, the randomly rounded solution \tilde{w} has ℓ_0 -norm at most $(1 + \delta)k$, and has optimality gap at most*

$$F(\tilde{w}) - P^* \leq c_4 \frac{\sqrt{r \log \min\{r, n\}}}{\rho}. \quad (6.24)$$

Note that the optimality gap in the preceding bound is negligible when the number of fractional solutions are small enough, and vanishes when the solution is integral, i.e., $r = 0$. The optimality gap also decreases when ρ gets larger in which case the objective of the original problem is heavily regularized by $\frac{\rho}{2} \|w\|_2^2$. The bound in Theorem 13 uses concentration bounds from random matrix theory [3] which are known to be sharp estimates of the statistical deviation in random sampling.

In our simulations, in order to be sure that we compare with a feasible integral solution (i.e., with at most k entries), we generate T realizations—say $\{\tilde{u}^1, \dots, \tilde{u}^T\}$ of the rounding procedure—and then pick the one \tilde{u}^* that has smallest objective value $G(\tilde{u})$ among the feasible solutions. (Note that \tilde{u}^* will exist with high probability for reasonable choices of T .) Finally, we define $\tilde{w}^* = \arg \min_w F(D(\tilde{u}^*)w)$. Denoting this procedure as *randomized rounding of order T* , we study its empirical behavior in Section 6.4 in the sequel.

The computational complexity of the randomized rounding procedure is dominated by evaluating $F(D(\tilde{u})w)$ a total of T times. However since \tilde{u} are sparse vectors this procedure is very efficient. For the least squares problem with target cardinality k the complexity becomes $\mathcal{O}(Tk^2n)$ since evaluating $(D(\tilde{u})w)$ can be done in $\mathcal{O}(k^2n)$ time using QR decomposition.

We note that in some other applications there might be additional constraints imposed on the vector u such as block sparsity or graphical structure. In such cases the randomized rounding process needs to be altered accordingly, or variants of rejection sampling can be used to generate vectors until constraints are satisfied.

6.3 Penalized forms of cardinality

Up to this point, we have considered the cardinality-constrained versions of sparse learning problems. If we instead enforce sparsity by augmenting the objective with some multiple of the ℓ_0 -norm, this penalized objective can also be reformulated as a Boolean program with a convex objective.

6.3.1 Reformulation as Boolean program

More precisely, suppose that we begin with the cardinality-penalized program

$$P^*(\lambda) := \min_{w \in \mathbb{R}^d} \left\{ \sum_{i=1}^n f(\langle x_i, w \rangle; y_i) + \frac{1}{2} \rho \|w\|_2^2 + \lambda \|w\|_0 \right\}. \quad (6.25)$$

As before, we suppose that for each $y \in \mathcal{Y}$, the function $t \mapsto f(t; y)$ is closed and convex. Under this condition, the following result provides an equivalent formulation as a convex program in Boolean variables:

Theorem 14. *For any $\rho > 0$ and $\lambda > 0$, the cardinality-penalized program (6.25) can be represented exactly as the Boolean convex program*

$$P^*(\lambda) = \min_{u \in \{0,1\}^d} \max_{v \in \mathbb{R}^n} \left\{ -\frac{1}{2\rho} v^T X D(u) X^T v - \sum_{i=1}^n f^*(v_i; y_i) + \lambda \sum_{i=1}^d u_i \right\}, \quad (6.26)$$

where $D(u) := \text{diag}(u) \in \mathbb{R}^{d \times d}$ is a diagonal matrix.

The proof is very similar to that of Theorem 11, and so we omit it.

As a consequence of the equivalent Boolean form (6.26), we can also obtain various convex relaxations of the cardinality-penalized program. For instance, the first-order relaxation takes the form

$$P_{\text{IR}}(\lambda) = \min_{u \in [0,1]^d} \max_{v \in \mathbb{R}^n} \left\{ -\frac{1}{2\rho} v^T X D(u) X^T v - \sum_{i=1}^n f^*(v_i; y_i) + \lambda \sum_{i=1}^d u_i \right\}, \quad (6.27)$$

which is the analogue of our first-order relaxation (6.13) for the constrained version of sparse learning.

As with our previous analysis, it is possible to eliminate the minimization over u from this saddle point expression. Strong duality holds, so that the maximum and minimum may be exchanged. In order to evaluate the minimum over u , we observe that $\frac{1}{2\rho} v^T X D(u) X^T v = \sum_{i=1}^d u_i \left(\frac{1}{2\rho} (x_i^T v)^2 \right)$, and moreover that

$$\min_{u \in [0,1]^d} \left\{ -\sum_{i=1}^d u_i \left(\frac{1}{2\rho} (x_i^T v)^2 - \lambda \right) \right\} = -\sum_{i=1}^d \left(\frac{1}{2\rho} (x_i^T v)^2 - \lambda \right)_+,$$

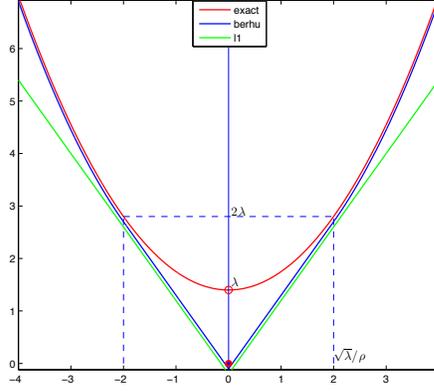


Figure 6.2: Plots of three different penalty functions as a function of $t \in \mathbb{R}$: reverse Huber (berhu) function $t \mapsto B(\frac{\sqrt{\rho t}}{\lambda})$ ℓ_1 -norm $t \mapsto \lambda|t|$ and the ℓ_0 -based penalty $t \mapsto \frac{t^2}{2} + \lambda\|t\|_0$.

Putting together the pieces, we can write the interval relaxation in the penalized case as the following convex (but non-differentiable) program

$$P_{\text{IR}}(\lambda) = \max_{v \in \mathbb{R}^n} \left\{ - \sum_{i=1}^d \left(\frac{1}{2\rho} (x_i^T v)^2 - \lambda \right)_+ - \sum_{i=1}^n f^*(v_i; y_i) \right\}. \quad (6.28)$$

6.3.2 Least-squares regression

As before, the relaxation (6.28) takes an especially simple form for the special but important case of least-squares regression. In particular, in the least-squares case, we have $f(t, y) = \frac{1}{2}(t - y)^2$, along with the corresponding conjugate dual function $f^*(s; y) = \frac{s^2}{2} + s y$. Consequently, the general relaxation (6.28) reduces to

$$P_{\text{IR}}(\lambda) = \max_{v \in \mathbb{R}^n} \left\{ - \sum_{i=1}^d \left(\frac{1}{2\rho} (x_i^T v)^2 - \lambda \right)_+ - v^T y - \frac{1}{2} \|v\|_2^2 \right\}, \quad (6.29)$$

As we now show, this convex program is equivalent to minimizing the least-squares objective using a form of regularization that combines the ℓ_1 and ℓ_2 -norms. In particular, let us define

$$B(t) = \frac{1}{2} \min_{z \in [0,1]} \left\{ z + \frac{t^2}{z} \right\} = \begin{cases} |t| & \text{if } |t| \leq 1 \\ \frac{t^2+1}{2} & \text{otherwise} \end{cases}. \quad (6.30)$$

This function combines the ℓ_1 and ℓ_2 norms in the way that is the opposite Huber's robust penalty; consequently, we call it the *reverse Huber penalty*.

Corollary 16. *The interval relaxation (6.29) for the cardinality-penalized least-squares problem has the equivalent form*

$$P_{\text{IR}}(\lambda) = \min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2} \|Xw - y\|_2^2 + 2\lambda \sum_{i=1}^d B\left(\frac{\sqrt{\rho}w_i}{\sqrt{\lambda}}\right) \right\}, \quad (6.31)$$

where B denotes the reverse Huber penalty.

A plot of the reverse Huber penalty is displayed in Figure 6.2 and compared with the ℓ_1 -norm $t \mapsto \lambda|t|$, as well as the ℓ_0 -based penalty $t \mapsto \lambda\|t\|_0 + \frac{1}{2}t^2$.

Proof. Consider the representation (6.29) for the least-squares case. We can represent the coordinatewise functions $(\cdot)_+$ function using a vector $p \in \mathbb{R}^d$ of auxiliary variables as follows

$$P_{\text{IR}}(\lambda) = \max_{v,p} \left\{ -1^T p - \frac{1}{2} \|v\|_2^2 - \langle v, y \rangle \right\} \quad \text{subject to } p \geq 0, \text{ and } p_i \geq \frac{1}{2\rho} (\langle x_i, v \rangle)^2 - \lambda \text{ for } i = 1, \dots, d.$$

Making use of rotated second order cone constraints, we have the equivalence

$$p_i \geq \frac{1}{2\rho} (\langle x_i, v \rangle)^2 - \lambda \iff \left\| \begin{pmatrix} \langle x_i, v \rangle \\ p_i + \lambda - 1 \end{pmatrix} \right\| \leq p_i + \lambda + 1, \quad \text{for } i = 1, \dots, d.$$

Thus, the relaxation (6.29) has the equivalent representation

$$P_{\text{IR}}(\lambda) = \max_{\substack{v \in \mathbb{R}^n \\ p \in \mathbb{R}^d}} \left\{ -\langle 1, p \rangle - \frac{1}{2} \|v\|_2^2 - \langle v, y \rangle \right\} \quad \text{subject to } p \geq 0, \left\| \begin{pmatrix} \sqrt{\rho^{-1}} \langle x_i, v \rangle \\ p_i + \lambda - 1 \end{pmatrix} \right\| \leq p_i + \lambda + 1, \quad i = 1, \dots, d.$$

which is a second order cone program (SOCP) in variables $(v, p) \in \mathbb{R}^n \times \mathbb{R}^d$.

Introducing Lagrange vectors for the constraints, we have

$$P_{\text{IR}}(\lambda) = \max_{v,p} \min_{\alpha, \beta, \gamma} \left\{ -\langle 1, p \rangle - \frac{1}{2} \|v\|_2^2 - \langle v, y \rangle + \sum_{i=1}^d \left(\gamma_i (p_i + \lambda - 1) - \sqrt{\rho^{-1}} \alpha_i \langle x_i, v \rangle - \beta_i (p_i + \lambda + 1) \right) \right\} \\ \text{subject to } p \geq 0, \left\| \begin{pmatrix} \alpha_i \\ \beta_i + \lambda - 1 \end{pmatrix} \right\| \leq \gamma_i, \quad i = 1, \dots, d.$$

Since $\lambda > 0$, strong duality holds by primal strict feasibility (see e.g., [28]), we may exchange the order of the minimum and the maximum. Making the substitutions $w = \alpha/\rho, u = \gamma + \beta, z = \gamma - \beta$, and then eliminating $v = y - Xw$ yields the equivalent

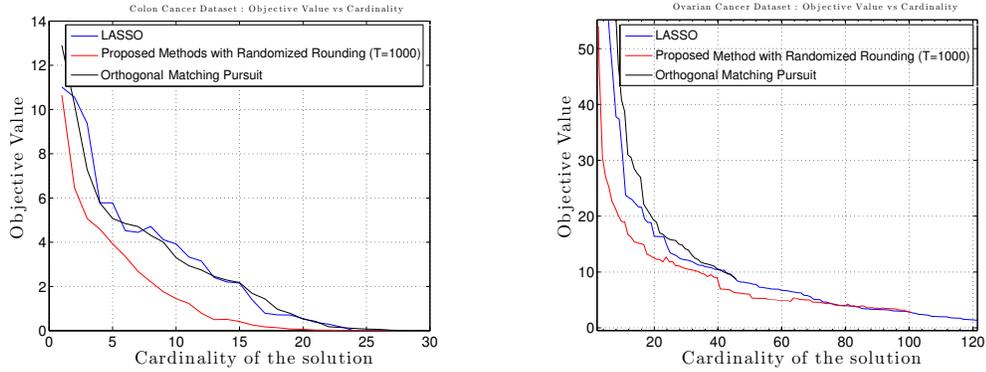


Figure 6.3: Objective value versus cardinality trade-off in a real dataset from cancer research. The proposed randomized rounding method considerably outperforms other methods by achieving lower objective value with smaller cardinality.

expression

$$\begin{aligned}
P_{\text{IR}}(\lambda) &= \min_{w,u,z} \max_{p \geq 0} \left\{ \frac{1}{2} \|Xw - y\|_2^2 + \langle p, z - 1 \rangle + \langle 1, \lambda z + y \rangle \right\} \quad \text{subject to} \quad \left\| \begin{pmatrix} \sqrt{\rho} x_i \\ y_i - z_i \end{pmatrix} \right\| \leq y_i + z_i \\
&= \min_{w,u,z} \left\{ \frac{1}{2} \|Xw - y\|_2^2 + \langle p, \lambda z + y \rangle \right\} \quad \text{subject to} \quad 0 \leq z_i \leq 1, y_i \geq 0, \rho w_i^2 \leq y_i z_i, \quad i = 1, \dots, n, \\
&= \min_{w,z} \left\{ \frac{1}{2} \|Xw - y\|_2^2 + \sum_{i=1}^d \left(\lambda z_i + \frac{\rho w_i^2}{z_i} \right) \right\}, \quad 0 \leq z_i \leq 1, \quad i = 1, \dots, n, \\
&= \min_w \left\{ \frac{1}{2} \|Xw - y\|_2^2 + 2\lambda \sum_{i=1}^d B\left(\frac{\sqrt{\rho} w_i}{\sqrt{\lambda}}\right) \right\},
\end{aligned}$$

which completes the proof. \square

We note that the alternative reverse Huber representation of the least squares problem can potentially be used to apply convex optimization toolboxes (e.g., [41, 64]) where the reverse Huber function is readily available.

6.4 Numerical Results

In this section, we discuss some numerical aspects of solving the relaxations that we have introduced, and illustrate their behavior on some real-world problems of sparse learning.

6.4.1 Optimization techniques

Although efficient polynomial-time methods exist for solving semi-definite programs, solving large-scale problems remains challenging using current computers and algorithms. For the SDP problems of interest here, one attractive alternative is to instead develop algorithms to solve the saddle-point problem in equation (6.10). For instance, in the least-squares case, the gradients of the relaxed objective in equation (6.19) are given by

$$\partial_i G(u) = -\left(x_i^T (I + XD(u)X^T/\rho)^{-1}y\right)^2.$$

Computing such a gradient requires the solution of a rank- $\|u\|_0$ linear system of size n , which can be done exactly in time $\mathcal{O}(\|u\|_0^3) + \mathcal{O}(nd)$ via the QR decomposition. Therefore, the overall complexity of using first-order and quasi-Newton methods is comparable to the Lasso when the sparsity level k is relatively small. We then employ a projected quasi-Newton method [125] to numerically optimize the convex objective. The randomized rounding procedure requires T evaluations of function value, which takes additional $\mathcal{O}(T\|\tilde{u}\|_0^3)$ time.

6.4.2 Experiments on real datasets

We consider two well known high-dimensional datasets studied in cancer research, the 62×2000 *Colon cancer* dataset¹ and 216×4000 *Ovarian cancer* dataset² which contain ion intensity levels corresponding to related proteins and corresponding *cancer* or *normal* output labels.. We consider classical ℓ_2^2 -regularized least squares classification using the mapping -1 for *cancer* label and $+1$ for *normal* label. We numerically implemented the proposed randomized rounding procedure of $T = 1000$ trials based on the relaxed solution. For other methods we identify their support and predict using regularized least squares solution constrained to that support where regularization parameter is optimized for each method on the training set. Figure 6.3 depicts optimization error (training error) as a function of the cardinality of the solution for both of the datasets. It is observed that the randomized rounding approach provides a considerable improvement in the optimal value for any fixed cardinality. In order to assess the learning and generalization performance of the trained model, we then split the dataset into two halves for training and testing. We present the plots of the test error as a function of cardinality over 1000 realizations of data splits and show the corresponding error-bars calculated for 1.5σ in Figure 6.4. The proposed algorithm also shows a considerable improvement in both training and test error compared

¹Taken from the Princeton University Gene Expression Project; for original source and further details please see the references therein.

²Taken from FDA-NCI Clinical Proteomics Program Databank; for original source and further details please see the references therein.

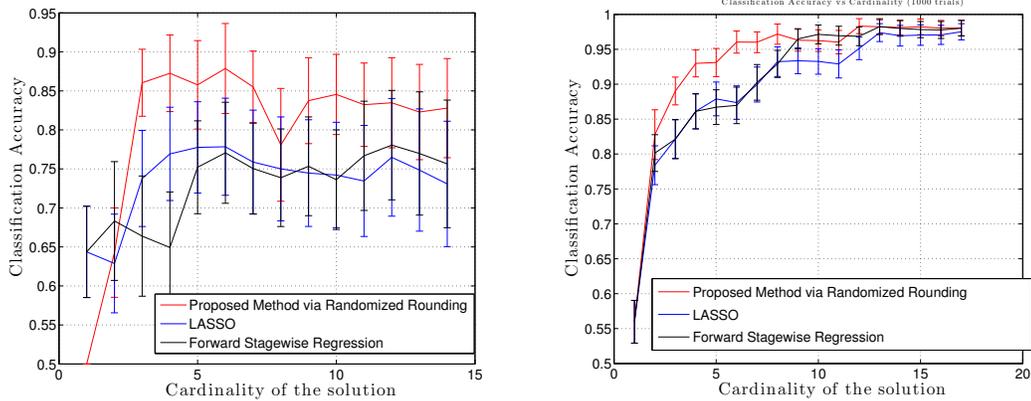


Figure 6.4: Classification accuracy versus cardinality in a real dataset from cancer research. The proposed method has considerably higher classification accuracy for a fixed cardinality.

to the other methods, as can be seen from the figures. We observed that choosing $T \in [100, 1000]$ gave satisfactory results however T can be chosen larger for higher dimensional problems without any computational difficulty.

We also note that in many applications choosing a target cardinality k with good predictive accuracy is an important problem. For a range of cardinality values the proposed approach can be combined with cross-validation and other model selection methodologies such as the Bayesian information criterion (BIC) or Akaike information criterion (AIC) [6, 146]. However there are also machine learning applications where the target cardinality is specified due to computational complexity requirements at runtime (see e.g. [46]). In these applications the cardinality directly effects the number of features that needs to be checked for classifying a new sample.

6.5 Simplex Constrained Problems

In this section we consider optimization problems of the following form,

$$p^* = \min_{x \in C} f(x) + \lambda \text{card}(x)$$

where f is a convex function, C is a convex set, $\text{card}(x)$ denotes the number of nonzero elements of x and $\lambda \geq 0$ is a given tradeoff parameter for adjusting desired sparsity. Since the cardinality penalty is inherently of combinatorial nature, these problems are in general not solvable in polynomial-time. In recent years ℓ_1 norm penalization as a proxy for penalizing cardinality has attracted a great deal of attention in machine learning, statistics, engineering and applied mathematics [34], [36],

[29], [35]. However the aforementioned types of sparse probability optimization problems are not amenable to the ℓ_1 heuristic since $\|x\|_1 = 1^T x = 1$ is constant on the probability simplex. Numerous problems in machine learning, statistics, finance and signal processing fall into this category however to the authors' knowledge there is no known general convex optimization strategy for such problems constrained on the probability simplex. We claim that the reciprocal of the infinity-norm, i.e., $\frac{1}{\max_i x_i}$ is the correct convex heuristic for penalizing cardinality on the probability simplex and the resulting relaxations can be solved via convex optimization. Figure 6.5 depicts an example of a sparse probability measure which also has maximal infinity norm. In the following sections we expand our discussion by exploring two specific problems: recovering a measure from given moments where $f = 0$ and C is affine, and convex clustering where f is a log-likelihood and $C = \mathbb{R}$. For the former case we give a sufficient condition for this convex relaxation to exactly recover the minimal cardinality solution of p^* . We then present numerical simulations for the both problems which suggest that the proposed scheme offers a very efficient convex relaxation for penalizing cardinality on the probability simplex.

6.6 Optimizing over sparse probability measures

We begin the discussion by first taking an alternative approach to the cardinality penalized optimization by directly lower-bounding the original hard problem using the following relation

$$\|x\|_1 = \sum_{i=1}^n |x_i| \leq \mathbf{card}(x) \max_i |x_i| \leq \mathbf{card}(x) \|x\|_\infty$$

which is essentially one of the core motivations of using ℓ_1 penalty as a proxy for cardinality. When constrained to the probability simplex, the lower-bound for the cardinality simply becomes $\frac{1}{\max_i x_i} \leq \mathbf{card}(x)$. Using this bound on the cardinality, we immediately have a lower-bound on our original NP-hard problem which we denote by p_∞^* :

$$p^* \geq p_\infty^* := \min_{x \in C, 1^T x = 1, x \geq 0} f(x) + \lambda \frac{1}{\max_i x_i} \quad (6.32)$$

The function $\frac{1}{\max_i x_i}$ is concave and hence the above lower-bounding problem is not a convex optimization problem. However below we show that the above problem can be exactly solved using convex programming.

Proposition 1. The lower-bounding problem defined by p_∞^* can be globally solved using the following n convex programs in $n + 1$ dimensions:

$$p^* \geq p_\infty^* = \min_{i=1, \dots, n} \left\{ \min_{x \in C, 1^T x = 1, x \geq 0, t \geq 0} f(x) + t \quad : \quad x_i \geq \lambda/t \right\}. \quad (6.33)$$

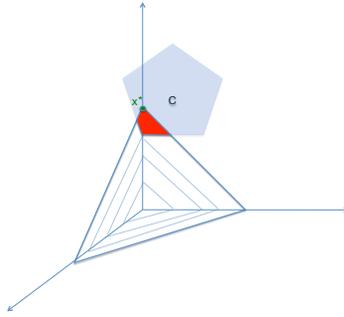


Figure 6.5: Probability simplex and the reciprocal of the infinity norm . The sparsest probability distribution on the set C is x^* (green) which also minimizes $\frac{1}{\max_i x_i}$ on the intersection (red)

Note that the constraint $x_i \geq \lambda/t$ is jointly convex since $1/t$ is convex in $t \in \mathbb{R}^+$, and they can be handled in most of the general purpose convex optimizers, e.g. `cvx`, using either the positive inverse function or rotated cone constraints.

Proof.

$$p_\infty^* = \min_{x \in C, 1^T x = 1, x \geq 0} f(x) + \min_i \frac{\lambda}{x_i} \quad (6.34)$$

$$= \min_i \min_{x \in C, 1^T x = 1, x \geq 0} f(x) + \frac{\lambda}{x_i} \quad (6.35)$$

$$= \min_i \min_{x \in C, 1^T x = 1, x \geq 0, t \geq 0} f(x) + t \quad s.t. \quad \frac{\lambda}{x_i} \leq t \quad (6.36)$$

□

The above formulation can be used to efficiently approximate the original cardinality constrained problem by lower-bounding for arbitrary convex f and C . In the next section we show how to compute the quality of approximation.

6.6.1 Computing a bound on the quality of approximation

By the virtue of being a relaxation to the original cardinality problem, we have the following remarkable property. Let \hat{x} be an optimal solution to the convex program p_∞^* , then we have the following relation

$$f(\hat{x}) + \lambda \text{card}(\hat{x}) \geq p^* \geq p_\infty^* \quad (6.37)$$

Since the left-hand side and right-hand side of the above bound are readily available when p_∞^* defined in (6.33) is solved, we immediately have a bound on the quality of relaxation. More specifically the relaxation is exact, i.e., we find a solution for the original cardinality penalized problem, if the following holds:

$$f(\hat{x}) + \lambda \mathbf{card}(\hat{x}) = p_\infty^*$$

It should be noted that for general cardinality penalized problems, using ℓ_1 heuristic does not yield such a quality bound, since it is not a lower or upper bound in general. Moreover most of the known equivalence conditions for ℓ_1 heuristics such as Restricted Isometry Property and variants are NP-hard to check. Therefore a remarkable property of the proposed scheme is that it comes with a simple computable bound on the quality of approximation.

6.7 Recovering a Sparse Measure

Suppose that μ is a discrete probability measure and we would like to know the sparsest measure satisfying some arbitrary moment constraints:

$$p^* = \min_{\mu} \mathbf{card}(\mu) \quad : \quad \mathbb{E}_{\mu}[X_i] = b_i, \quad i = 1, \dots, m$$

where X_i 's are random variables and E_{μ} denotes expectation with respect to the measure μ . One motivation for the above problem is the fact that it upper-bounds the minimum entropy power problem:

$$p^* \geq \min_{\mu} \exp H(\mu) \quad : \quad \mathbb{E}_{\mu}[X_i] = b_i, \quad i = 1, \dots, m$$

where $H(\mu) := -\sum_i \mu_i \log \mu_i$ is the Shannon entropy. Both of the above problems are non-convex and in general very hard to solve.

When viewed as a finite dimensional optimization problem the minimum cardinality problem can be cast as a linear sparse recovery problem:

$$p^* = \min_{1^T x=1, x \geq 0} \mathbf{card}(x) \quad : \quad Ax = b \tag{6.38}$$

As noted previously, applying the ℓ_1 heuristic doesn't work and it does not even yield a unique solution when the problem is underdetermined since it simply solves a feasibility problem:

$$p_1^* = \min_{1^T x=1, x \geq 0} \|x\|_1 \quad : \quad Ax = b \tag{6.39}$$

$$= \min_{1^T x=1, x \geq 0} 1 \quad : \quad Ax = b \tag{6.40}$$

and recovers the true minimum cardinality solution if and only if the set $1^T x = 1, x \geq 0, Ax = b$ is a singleton. This condition may hold in some cases, i.e. when the first $2k - 1$ moments are available, i.e., A is a Vandermonde matrix where $k = \mathbf{card}(x)$ [38]. However in general this set is a polyhedron containing dense vectors. Below we show how the proposed scheme applies to this problem.

Using general form in (6.33), the proposed relaxation is given by the following,

$$(p^*)^{-1} \leq (p_\infty^*)^{-1} = \max_{i=1, \dots, n} \left\{ \max_{1^T x=1, x \geq 0} x_i : Ax = b \right\}. \quad (6.41)$$

which can be solved very efficiently by solving n linear programs in n variables. The total complexity is at most $O(n^4)$ using a primal-dual LP solver.

It's easy to check that strong duality holds and the dual problems are given by the following:

$$(p_\infty^*)^{-1} = \max_{i=1, \dots, n} \left\{ \min_{w, \lambda} w^T b + \lambda : A^T w + \lambda 1 \geq e_i \right\}. \quad (6.42)$$

where 1 is the all ones vector and e_i is all zeros with a one in only i 'th coordinate.

6.7.1 An alternative minimal cardinality selection scheme

When the desired criteria is to find a minimum cardinality probability vector satisfying $Ax = b$, the following alternative selection scheme offers a further refinement, by picking the lowest cardinality solution among the n linear programming solutions. Define

$$\hat{x}_i : = \arg \max_{1^T x=1, x \geq 0} x_i : Ax = b \quad (6.43)$$

$$\hat{x}_{min} : = \arg \min_{i=1, \dots, n} \mathbf{card}(\hat{x}_i) \quad (6.44)$$

The following theorem gives a sufficient condition for the recovery of a sparse measure using the above method.

Theorem 2. Assume that the solution to p^* in (6.38) is unique and given by x^* . If the following condition holds

$$\min_{1^T x=1, y \geq 0, 1^T y=1} x_i \text{ s.t. } A_S x = A_{S^c} y > 0$$

where $b = Ax^*$ and A_S is the submatrix containing columns of A corresponding to non-zero elements of x^* and A_{S^c} is the submatrix of remaining columns, then the convex linear program

$$\max_{1^T x=1, x \geq 0} x_i : Ax = b$$

has a unique solution given by x^* .

Let $\text{Conv}(a_1, \dots, a_m)$ denote the convex hull of the m vectors $\{a_1, \dots, a_m\}$. The following corollary depicts a geometric condition for recovery.

Corollary 3. If $\text{Conv}(A_{S^c})$ does not intersect an extreme point of $\text{Conv}(A_S)$ then $\hat{x}_{min} = x^*$, i.e. we recover the minimum cardinality solution using n linear programs.

Proof. Consider k 'th inner linear program defined in the problem p_∞^* . Using the optimality conditions of the primal-dual linear program pairs in (6.41) and (6.42), it can be shown that the existence of a pair (w, λ) satisfying

$$A_S^T w + \lambda 1 = e_k \quad (6.45)$$

$$A_{S^c}^T w + \lambda 1 > 0 \quad (6.46)$$

implies that the support of solution of the linear program is exactly equal to the support of x^* , and in particular they have the same cardinality. Since the solution of p^* is unique and has minimum cardinality, we conclude that x^* is indeed the unique solution to the k 'th linear program. Applying Farkas' lemma and duality theory we arrive at the conditions defined in Theorem 2. The corollary follows by first observing that the condition of Theorem 2 is satisfied if $\text{Conv}(A_{S^c})$ does not intersect an extreme point of $\text{Conv}(A_S)$. Finally observe that if any of the n linear programs recover the minimal cardinality solution then $\hat{x}_{min} = x^*$, since $\text{card}(\hat{x}_{min}) \leq \text{card}(\hat{x}_k), \forall k$.

6.7.2 Noisy measure recovery

When the data contains noise and inaccuracies, such as the case when using empirical moments instead of exact moments, we propose the following noise-aware robust version, which follows from the general recipe given in the first section:

$$\min_{i=1, \dots, n} \left\{ \min_{1^T x=1, x \geq 0, t \geq 0} \|Ax - b\|_2^2 + t \quad : \quad x_i \geq \lambda/t \right\}. \quad (6.47)$$

where $\lambda \geq 0$ is a penalty parameter for encouraging sparsity. The above problem can be solved using n second-order cone programs in $n + 1$ variables, hence has $O(n^4)$ worst case complexity.

The proposed measure recovery algorithms are investigated and compared with a known suboptimal heuristic in Section 6.10.

6.8 Convex Clustering

In this section we base our discussion on the exemplar based convex clustering framework of [81]. Given a set of data points $\{z_1, \dots, z_n\}$ of d -dimensional vectors, the

task of clustering is to fit a mixture probability model to maximize the log likelihood function

$$L := \frac{1}{n} \sum_{i=1}^n \log \left[\sum_{j=1}^k x_j f(z_i; m_j) \right]$$

where $f(z; m)$ is an exponential family distribution on Z with parameter m , and x is a k -dimensional vector on the probability simplex denoting the mixture weights. For the standard multivariate Normal distribution we have $f(z_i; m_j) = e^{-\beta \|z_i - m_j\|_2^2}$ for some parameter $\beta > 0$. As in [81] we'll further assume that the mean parameter m_j is one of the examples z_i which is unknown a-priori. This assumption helps to simplify the log-likelihood whose data dependence is now only through a kernel matrix $K_{ij} := e^{-\beta \|z_i - z_j\|_2^2}$ as follows

$$L = \frac{1}{n} \sum_{i=1}^n \log \left[\sum_{j=1}^k x_j e^{-\beta \|z_i - z_j\|_2^2} \right] \quad (6.48)$$

$$= \frac{1}{n} \sum_{i=1}^n \log \left[\sum_{j=1}^k x_j K_{ij} \right] \quad (6.49)$$

Partitioning the data $\{z_1, \dots, z_n\}$ into few clusters is equivalent to have a sparse mixture x , i.e., each example is assigned to few centers (which are some other examples). Therefore to cluster the data we propose to approximate the following cardinality penalized problem,

$$p_c^* := \max_{1^T x = 1, x \geq 0} \sum_{i=1}^n \log \left[\sum_{j=1}^k x_j K_{ij} \right] - \lambda \mathbf{card} x \quad (6.50)$$

As hinted previously, the above problem can be seen as a lower-bound for the entropy penalized problem

$$p_c^* \leq \max_{1^T x = 1, x \geq 0} \sum_{i=1}^n \log \left[\sum_{j=1}^k x_j K_{ij} \right] - \lambda \exp H(x) \quad (6.51)$$

where $H(x)$ is the Shannon entropy of the mixture probability vector.

Applying our convexification strategy, we arrive at another upper-bound which can be computed via convex optimization

$$p_c^* \leq p_\infty^* := \max_{1^T x = 1, x \geq 0} \sum_{i=1}^n \log \left[\sum_{j=1}^k x_j K_{ij} \right] - \frac{\lambda}{\max_i x_i} \quad (6.52)$$

We investigate the above approach in a numerical example in Section 6.10 and compare with the well-known soft k-means algorithm.

6.9 Algorithms

6.9.1 Exponentiated Gradient

Exponentiated gradient [77] is a proximal algorithm to optimize over the probability simplex which uses the Kullback-Leibler divergence $D(x, y) = \sum_i x_i \log \frac{x_i}{y_i}$ between two probability distributions as a proximal map. For minimizing a convex function ψ the exponentiated gradient updates are given by the following:

$$x^{k+1} = \arg \min_x \psi(x^k) + \nabla \psi(x^k)^T (x - x^k) + \frac{1}{\alpha} D(x, x^k)$$

When applied to the general form of 6.33 it yields the following updates to solve the i 'th problem of p_∞^*

$$x_i^{k+1} = r_i^k x_i^k / \left(\sum_j r_j^k x_j^k \right)$$

where the weights r_i are exponentiated gradients:

$$r_i^k = \exp \left(\alpha (\nabla_i f(x^k) - \lambda / x_i^2) \right)$$

We also note that the above updates can be done in parallel for the n convex programs, and they are guaranteed to converge to the optimum.

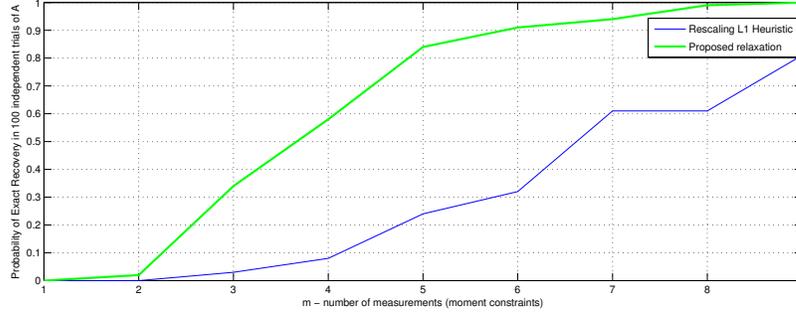
6.10 Numerical Results

6.10.1 Recovering a Measure from Gaussian Measurements

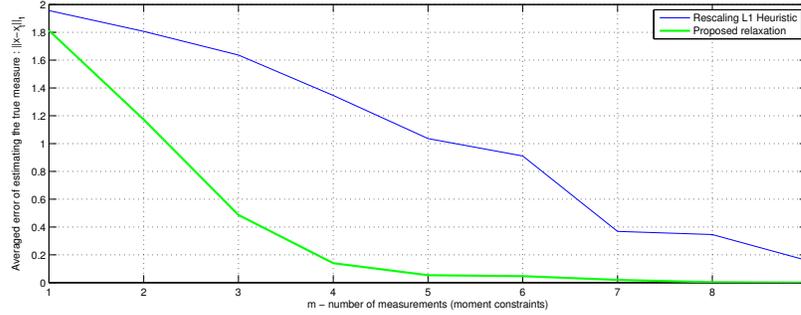
Here we show that the proposed recovery scheme is able to recover a sparse measure exactly with overwhelming probability, when the matrix $A \in \mathbb{R}^{m \times n}$ is chosen from the independent Gaussian ensemble, i.e, $A_{i,j} \sim \mathcal{N}(0, 1)$ i.i.d.

As an alternative method we consider a commonly employed simple heuristic to optimize over a probability measure which first drops the constraint $1^T x = 1$ and solves the corresponding ℓ_1 penalized problem. And finally rescales the optimal x such that $1^T x = 1$. This procedure is clearly suboptimal and we will refer it as the *rescaling heuristic*. We set $n = 50$ and randomly pick a 2-sparse probability vector x^* which is k sparse, let $b = Ax^*$ be m noiseless measurements, then check the probability of recovery, i.e. $\hat{x} = x^*$ where \hat{x} is the solution to,

$$\max_{i=1, \dots, n} \left\{ \max_{1^T x=1, x \geq 0} x_i \quad : \quad Ax = b \right\}. \quad (6.53)$$



(a) Probability of exact recovery as a function of m



(b) Average error for noisy recovery as a function of m

Figure 6.6: A comparison of the exact recovery probability in the noiseless setting (top) and estimation error in the noisy setting (bottom) of the proposed approach and the rescaled ℓ_1 heuristic

Figure 6.6(a) shows the probability of *exact* recovery as a function of m , the number of measurements, in 100 independent realizations of A for the proposed LP formulation and the rescaling heuristic. As it can be seen in Figure 6.6(a), the proposed method recovers the correct measure with probability almost 1 when $m \geq 5$. Quite interestingly the rescaling heuristic doesn't succeed to recover the true measure with high probability even for a cardinality 2 vector.

We then add normal distributed noise with standard deviation 0.1 on the observations and solve,

$$\min_{i=1,\dots,n} \left\{ \min_{1^T x=1, x \geq 0, t \geq 0} \|Ax - b\|_2^2 + t \quad : \quad x_i \geq \lambda/t \right\}. \quad (6.54)$$

We compare the above approach by the corresponding rescaling heuristic, which first solves a nonnegative Lasso,

$$\min_{x \geq 0} \|Ax - b\|_2^2 + \lambda \|x\|_1 \quad (6.55)$$

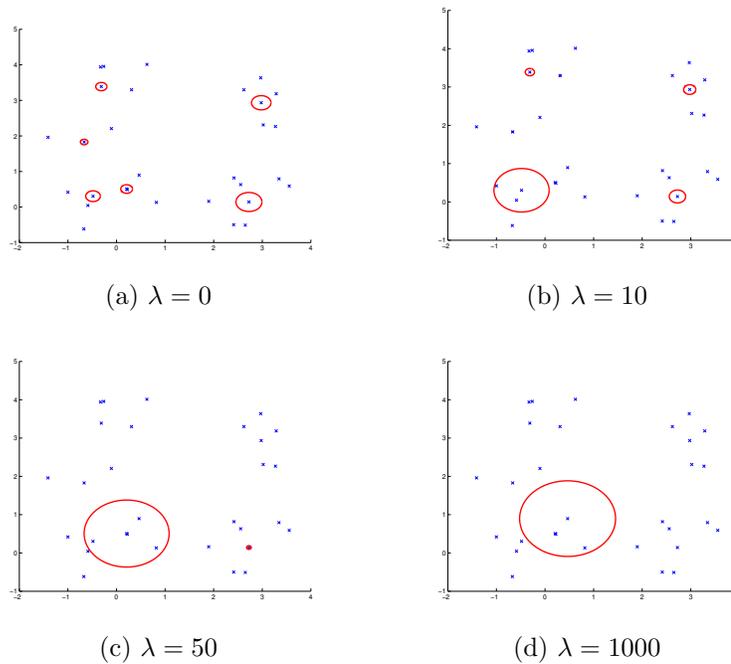


Figure 6.7: Proposed convex clustering scheme

then rescales x such that $1^T x = 1$. For each realization of A and measurement noise we run both methods using a primal-dual interior point solver for 30 equally spaced values of $\lambda \in [0, 10]$ and record the minimum error $\|\hat{x} - x^*\|_1$. The average error over 100 realizations are shown in Figure 6.6(b). As it can be seen in the figure the proposed scheme clearly outperforms the rescaling heuristic since it can utilize the fact that x is on the probability simplex, without trivializing its complexity regularizer.

6.10.2 Convex Clustering

We generate synthetic data using a Gaussian mixture of 4 components with identity covariances and cluster the data using the proposed method, the resulting clusters given by the mixture density is presented in Figure 6.7. The centers of the circles represent the means of the mixture components and the radii are proportional to the respective mixture weights. We then repeat the clustering procedure using the well known soft k-means algorithm and present the results in Figure 6.8.

As it can be seen from the figures the proposed convex relaxation is able to penalize the cardinality on the mixture probability vector and produce clusters close to the soft k-means algorithm. Note that soft k-means is a non-convex procedure whose performance depends heavily on the initialization. The proposed approach is convex hence insensitive to the initializations. Note that in [81] the number of clusters are

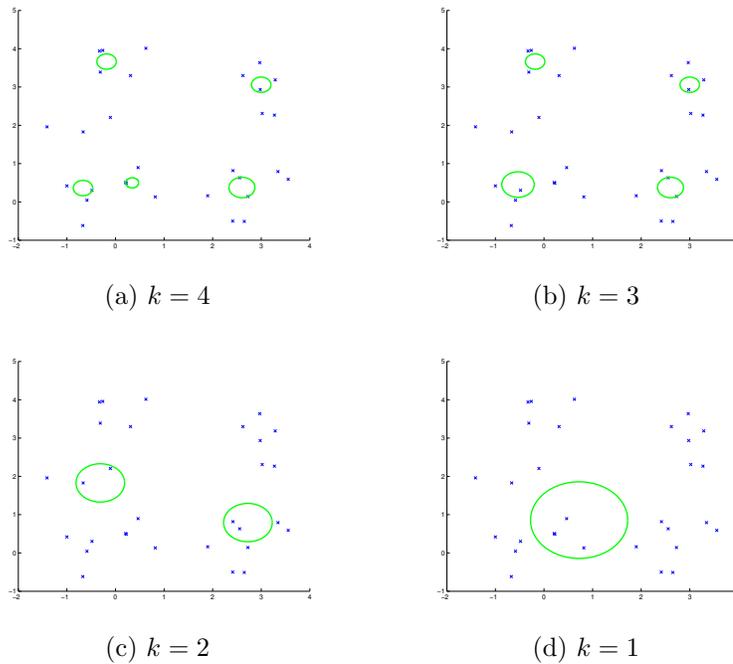


Figure 6.8: Soft k-means algorithm

adjusted indirectly by varying the β parameter of the distribution. In contrast our approach tries to implicitly optimize the likelihood/cardinality tradeoff by varying λ .

6.11 Discussion

We first showed how a broad class of cardinality-constrained (or penalized) sparse learning problems can be reformulated exactly as Boolean programs involving convex objective functions. The utility of this reformulation is in permitting the application of various types of relaxation hierarchies, such as the Sherali-Adams and Lasserre hierarchies for Boolean programs. The simplest such relaxation is the first-order interval relaxation, and we analyzed the conditions for its exactness in detail. In contrast to the classical ℓ_1 heuristic, the presented method provides a lower bound on the solution value, and moreover a certificate of optimality when the solution is integral. We provided sufficient conditions for the solution to be integral for linear regression problems with random Gaussian design matrices. For problems in which the solution is not integral, we proposed an efficient randomized rounding procedure, and showed that its approximation accuracy can be controlled in terms of the number of fractional entries, and a regularization parameter in the algorithm. In our experiments with real

data sets, the output of this randomized rounding procedure provided considerably better solutions than standard competitors such as the Lasso or orthogonal matching pursuit.

We also presented a convex cardinality penalization scheme for problems constrained on the probability simplex. We then derived a sufficient condition for recovering the sparsest probability measure in an affine space using the proposed method. The geometric interpretation suggests that it holds for a large class of matrices. An interesting direction is to extend the recovery analysis to the noisy setting and arbitrary functions such as the log-likelihood in the clustering example. There might also be other problems where proposed approach could be practically useful such as portfolio optimization, or sparse multiple kernel learning where a sparse convex combination of assets is sought.

There are a range of interesting open problem suggested by our developments. In particular, we have studied only the most naive first-order relaxation for the problem: it would be interesting to see whether one quantify how quickly the performance improves (relative to the exact cardinality-constrained solution) as the level of relaxation—say in one of the standard hierarchies for Boolean problems [128, 88, 82, 83, 145]—is increased. This question is particularly interesting in light of recent work [156] showing that, under a standard conjecture in computational complexity, there are fundamental gaps between the performance of cardinality-constrained estimators and polynomial-time methods for the prediction error in sparse regression.

6.12 Proofs of technical results

In this section, we provide the proofs of Theorems 12 and Theorem 13.

6.12.1 Proof of Theorem 12

Recalling the definition (6.20) of the matrix M , for each $j \in \{1, \dots, d\}$, define the rescaled random variable $U_j := \frac{X_j^T M y}{\rho n}$. In terms of this notation, it suffices to find a scalar λ such that

$$\min_{j \in S} |U_j| > \lambda \quad \text{and} \quad \max_{j \in S^c} |U_j| < \lambda. \quad (6.56)$$

By definition, we have $y = X_S w_S^* + \varepsilon$, whence

$$U_j = \underbrace{\frac{X_j^T M X_S w_S^*}{\rho n}}_{A_j} + \underbrace{\frac{X_j^T M \varepsilon}{\rho n}}_{B_j}.$$

Based on this decomposition, we then make the following claims:

Lemma 31. *There are numerical constants c_1, c_2 such that*

$$\mathbb{P}\left[\max_{j=1,\dots,d} |B_j| \geq t\right] \leq c_1 e^{-c_2 \frac{nt^2}{\gamma^2} + \log d}. \quad (6.57)$$

Lemma 32. *There are numerical constants c_1, c_2 such that*

$$\mathbb{P}\left[\min_{j \in S} |A_j| < \frac{w_{\min}}{4}\right] \leq c_1 e^{-c_2 n \frac{w_{\min}^2}{\|w_S^*\|_2^2} + \log(2k)} \quad \text{and} \quad (6.58a)$$

$$\mathbb{P}\left[\max_{j \in S^c} |A_j| \geq \frac{w_{\min}}{16}\right] \leq c_3 e^{-c_4 n \frac{w_{\min}^2}{\|w_S^*\|_2^2} + \log(d-k)}, \quad (6.58b)$$

Using these two lemmas, we can now complete the proof. Recall that Theorem 12 assumes a lower bound of the form $n > c_0 \frac{\gamma^2 + \|w_S^*\|_2^2}{w_{\min}^2} \log d$, where c_0 is a sufficiently large constant. Thus, setting $t = \frac{w_{\min}}{16}$ in Lemma 31 ensures that $\max_{j=1,\dots,d} |B_j| \leq \frac{w_{\min}}{16}$ with high probability. Combined with the bound (6.58a) from Lemma 32, we are guaranteed that

$$\min_{j \in S} |U_j| \geq \frac{w_{\min}}{4} - \frac{w_{\min}}{16} = \frac{3w_{\min}}{16} \quad \text{with high probability.}$$

Similarly, the bound (6.58b) guarantees that

$$\max_{j \in S^c} |U_j| \leq \frac{w_{\min}}{16} + \frac{w_{\min}}{16} = \frac{2w_{\min}}{16} \quad \text{also with high probability.}$$

Thus, setting $\lambda = \frac{5w_{\min}}{32}$ ensures that the condition (6.56) holds.

The only remaining detail is to prove the two lemmas.

6.12.1.0.3 Proof of Lemma 31: Define the event $\mathcal{E}_j = \{\|X_j\|_2 / \sqrt{n} \leq 2\}$, and observe that

$$\mathbb{P}[|B_j| > t] \leq \mathbb{P}[|B_j| > t \mid \mathcal{E}] + \mathbb{P}[\mathcal{E}^c].$$

Since the variable $\|X_j\|_2^2$ follows a χ^2 -distribution with n degrees of freedom, we have $\mathbb{P}[\mathcal{E}^c] \leq 2e^{-c_2 n}$. Recalling the definition (6.20) of the matrix M , note that $\lambda_{\max}(M) \leq \rho^{-1}$, whence conditioned on \mathcal{E} , we have $\|MX_j\|_2 \leq \|X_j\|_2 \leq 2\sqrt{n}$. Consequently, conditioned on \mathcal{E} , the variable $\frac{X_j^T M \varepsilon}{\rho}$ is a Gaussian random vector with variance at most $4\gamma^2 / \rho^2$, and hence $\mathbb{P}[|B_j| > t \mid \mathcal{E}] \leq 2e^{-\frac{\rho^2 t^2}{32\gamma^2}}$.

Finally, by union bound, we have

$$\mathbb{P}\left[\max_{j=1,\dots,d} |B_j| > t\right] \leq d \mathbb{P}[|B_j| > t] \leq d \left\{ 2e^{-\frac{\rho^2 t^2}{32\gamma^2}} + 2e^{-c_2 \rho n} \right\} \leq c_1 e^{-c_2 \frac{\rho^2 t^2}{\gamma^2} + \log d},$$

as claimed.

6.12.1.0.4 Proof of Lemma 32: We split the proof into two parts.

6.12.1.0.5 (1) Proof of the bound (6.58a): Note that

$$\frac{1}{\rho} X_S^T M X_S = X_S^T (\rho I_n + X_S X_S^T)^{-1} X_S$$

We now write $X_S = U D V^T$ for singular value decomposition of $\frac{1}{\sqrt{n}} X_S$ in compact form. We thus have

$$\frac{1}{\rho} X_S^T M X_S = V (\rho I_n + n D^2)^{-1} D^2 V^T.$$

We will prove that for a fixed vector z , the following holds with high probability

$$\frac{\| \left(\frac{1}{\rho} X_S^T M X_S - I \right) z \|_\infty}{\|z\|_\infty} \leq \epsilon. \quad (6.59)$$

Applying the above bound to w_s^* , which is a fixed vector we obtain

$$\| \left(\frac{1}{\rho} X_S^T M X_S - I \right) w_s^* \|_\infty \leq \epsilon \|w_s^*\|_\infty \quad (6.60)$$

Then by triangle inequality the above statement implies that

$$\min_{i \in S} \left| \frac{1}{\rho} X_S^T M X_S w_i^* \right| > (1 - \epsilon) \min_{i \in S} |w_i^*|.$$

and setting $\epsilon = 3/4$ yields the claim.

Next we let $\frac{1}{\rho} X_S^T M X_S - I = V \tilde{D} V$ where we defined $\tilde{D} := ((\rho I_n + D^2)^{-1} D^2 - I)$. By standard results on operator norm of Gaussian random matrices (e.g., see Davidson and Szarek [44]), the minimum singular value

$$\sigma_{\min} \left(\frac{1}{\sqrt{n}} X_S \right) = \min_{i=1, \dots, k} D_{ii}$$

of the matrix X_S/\sqrt{n} can be bounded as

$$\mathbb{P} \left[\frac{1}{\sqrt{n}} \min_{i=1, \dots, k} |D_{ii}| \leq 1 - \sqrt{\frac{k}{n}} - t \right] \leq 2e^{-c_1 n t^2}, \quad (6.61)$$

where c_1 is a numerical constant (independent of (n, k)).

Now define $Y_i := e_i^T V \tilde{D} V^T z = z_i v_i \tilde{D} v_i + v_i^T \tilde{D} \sum_{l \neq i} z_l v_l$. Then note that,

$$\begin{aligned} |Y_1| &\leq \|\tilde{D}\|_2 |z_1| + v_1^T \tilde{D} \sum_{l \neq i} z_l v_l \\ &= \frac{\rho}{\rho + \min_{i=1, \dots, k} |D_{ii}|^2} |z_1| + F(v_1) \end{aligned}$$

where we defined $F(v_1) := v_1^T \tilde{D} \sum_{l \neq i} z_l v_l$ and v_1 is uniformly distributed over a sphere in $k - 1$ dimensions and hence $\mathbb{E}F(v_1) = 0$. Observe that F is a Lipschitz map satisfying

$$\begin{aligned} |F(v_1) - F(v'_1)| &\leq \|\tilde{D}\|_\infty \sqrt{\sum_{l \neq i} |z_l^2|} \|v_1 - v'_1\|_2 \\ &= \frac{\rho}{\rho + \min_i |D_{ii}|^2} |\sqrt{k-1}|z||_\infty \|v_1 - v'_1\|_2 \end{aligned}$$

Applying concentration of measure for Lipschitz functions on the sphere (e.g., see [84]) the function $F(v_1)$ we get that for all $t > 0$ we have,

$$\mathbb{P}[F(v_1) > t|z||_\infty] \leq 2e^{-c_4(k-1) \frac{t^2}{\left(\frac{\rho}{\rho + \min_i |D_{ii}|^2}\right)^2 (k-1)}}. \quad (6.62)$$

Conditioning on the high probability event $\{\min_i |D_{ii}|^2 \leq \frac{n}{2}\}$ and then applying the tail bound (6.61) yields

$$\begin{aligned} \mathbb{P}[F(v_1) > t|z||_\infty] &\leq 2 \exp\left(-c_4 \frac{n^2 t^2}{\rho^2}\right) + 2e^{-c_2 \frac{nt^2}{\rho^2}} \\ &\leq 4e^{-c_5 \frac{n^2 t^2}{\rho^2}}. \end{aligned} \quad (6.63)$$

Combining the pieces in (6.63) and (6.62), we take a union bound over $2k$ coordinates,

$$\begin{aligned} \mathbb{P}\left[\min_{j \in S} |Y_j| > t|z||_\infty\right] &\leq 2k \cdot 3 \exp(-c_5 n^2 t^2 / \rho^2) \\ &\leq 2k \cdot 3 \exp(-c_5 n t^2). \end{aligned}$$

where the final line follows from our choice $\rho = \sqrt{n}$. Finally setting $t = \epsilon$ we obtain the statement in (6.59) and hence complete the proof.

6.12.1.0.6 Proof of the bound (6.58b): A similar calculation yields

$$A_j = \frac{1}{\rho} X_{S^c}^T M X_S w_S^* = X_{S^c}^T (\rho I_n + X_S X_S^T)^{-1} X_S w_S^*,$$

for each $j \in S^c$. Defining the event $\mathcal{E} = \{\lambda_{\max}(X_S)/\rho \leq 2\sqrt{n}\}$, standard bounds in random matrix theory [44] imply that $\mathbb{P}[\mathcal{E}^c] \leq 2e^{-c_2 n}$. Conditioned on \mathcal{E} , we have

$$\|(\rho I_n + X_S X_S^T)^{-1} X_S w_S^*\|_2 \leq \frac{2}{\rho} \|w_S^*\|_2,$$

so that the variable A_j is conditionally Gaussian with variance at most $\frac{4}{\rho^2} \|w_S^*\|_2^2$. Consequently, we have

$$\mathbb{P}[|A_j| \geq t] \leq \mathbb{P}[|A_j| \geq t \mid \mathcal{E}] + \mathbb{P}[\mathcal{E}^c] = 2e^{-\frac{\rho^2 t^2}{32 \|w_S^*\|_2^2}} + 2e^{-c_2} \leq c_1 e^{-c_2 \frac{\rho^2 t^2}{\|w_S^*\|_2^2}},$$

Setting $t = \frac{w_{\min}}{8}$, $\rho = \sqrt{n}$ and taking union bound over all $d - k$ indices in S^c yields the claim (6.58b).

6.12.2 Proof of Theorem 13

The vector $\tilde{u} \in \{0, 1\}^d$ consists of independent Bernoulli trials, and we have $\mathbb{E}[\sum_{j=1}^d \tilde{u}_j] \leq k$. Consequently, by the Chernoff bound for Bernoulli sums, we have

$$\mathbb{P}\left[\sum_{j=1}^d \tilde{u}_j \geq (1 + \delta)k\right] \leq c_1 e^{-c_2 k \delta^2}.$$

as claimed.

It remains to establish the high-probability bound on the optimal value. As shown previously, the Boolean problem admits the saddle point representation

$$P^* = \min_{u \in \{0,1\}^d, \sum_{i=1}^d u_i \leq k} \underbrace{\left\{ \max_{\alpha \in \mathbb{R}^n} -\frac{1}{\rho} \alpha^T X D(u) X^T \alpha - \|\alpha\|_2^2 - 2\alpha^T y \right\}}_{G(u)}. \quad (6.64)$$

Since the optimal value is non-negative, the optimal dual parameter $\alpha \in \mathbb{R}^n$ must have its ℓ_2 -norm bounded as $\|\alpha\|_2 \leq 2\|y\|_2 \leq 2$. Using this fact, we have

$$\begin{aligned} G(\hat{u}) - G(\tilde{u}) &= \max_{\|\alpha\|_2 \leq 2} \left\{ -\frac{1}{\rho} \alpha^T X D(\hat{u}) X^T \alpha - \|\alpha\|_2^2 - 2\alpha^T y \right\} - \max_{\|\alpha\|_2 \leq 2} \left\{ -\frac{1}{\rho} \alpha^T X D(\tilde{u}) X^T \alpha - \|\alpha\|_2^2 - 2\alpha^T y \right\} \\ &\leq \max_{\|\alpha\|_2 \leq 2} \left\{ -\frac{1}{\rho} \alpha^T X (D(\hat{u}) - D(\tilde{u})) X^T \alpha \right\} \\ &\leq \frac{2}{\rho} \lambda_{\max}(X (D(\hat{u}) - D(\tilde{u})) X^T), \end{aligned}$$

where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of a symmetric matrix.

It remains to establish a high probability bound on this maximum eigenvalue. Recall that R is the subset of indices associated with fractional elements of \widehat{u} , and moreover that $\mathbb{E}[\tilde{u}_j] = \widehat{u}_j$. Using these facts, we can write

$$X(D(\tilde{u}) - D(\widehat{u}))X^T = \sum_{j \in R} \underbrace{(\tilde{u}_j - \mathbb{E}[\tilde{u}_j])X_j X_j^T}_{A_j}$$

where $X_j \in \mathbb{R}^n$ denotes the j^{th} column of X . Since $\|X_j\|_2 \leq 1$ by assumption and \tilde{u}_j is Bernoulli, the matrix A_j has operator norm at most 1, and is zero mean. Consequently, by the Ahlswede-Winter matrix bound [3, 109], we have

$$\mathbb{P}\left[\lambda_{\max}\left(\sum_{j \in R} A_j\right) \geq \sqrt{rt}\right] \leq 2 \min\{n, r\}e^{-t^2/16},$$

where $r = |R|$ is the number of fractional components. Setting $t^2 = c \log \min\{n, r\}$ for a sufficiently large constant c yields the claim.

Bibliography

- [1] D. Achlioptas, “Database-friendly random projections: Johnson-lindenstrauss with binary coins,” *Journal of computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.
- [2] S. Aeron, V. Saligrama, and M. Zhao, “Information theoretic bounds for compressed sensing,” *IEEE Trans. Info. Theory*, vol. 56, no. 10, pp. 5111–5130, 2010.
- [3] R. Ahlswede and A. Winter, “Strong converse for identification via quantum channels,” *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 569–579, March 2002.
- [4] N. Ailon and B. Chazelle, “Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform,” in *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*. ACM, 2006, pp. 557–563.
- [5] N. Ailon and E. Liberty, “Fast dimension reduction using Rademacher series on dual BCH codes,” *Discrete Comput. Geom*, vol. 42, no. 4, pp. 615–630, 2009.
- [6] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” in *Proceedings of the 2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, September 1971.
- [7] A. E. Alaoui and M. W. Mahoney, “Fast randomized kernel methods with statistical guarantees,” UC Berkeley, Tech. Rep. arXiv:1411.0306, 2014.
- [8] N. Alon, Y. Matias, and M. Szegedy, “The space complexity of approximating the frequency moments,” in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. ACM, 1996, pp. 20–29.
- [9] Y. Amit, M. Fink, N. Srebro, and S. Ullman, “Uncovering shared structures in multiclass classification,” in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 17–24. [Online]. Available: <http://doi.acm.org/10.1145/1273496.1273499>

- [10] J. Antony and A. R. Barron, “Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity,” *IEEE Trans. Info. Theory*, vol. 58, no. 5, pp. 2541–2557, 2012.
- [11] A. Argyriou, T. Evgeniou, and M. Pontil, “Convex multi-task feature learning,” *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008. [Online]. Available: <http://dx.doi.org/10.1007/s10994-007-5040-8>
- [12] E. Arias-Castro and Y. Eldar, “Noise folding in compressed sensing,” *IEEE Signal Proc. Letters.*, vol. 18, no. 8, pp. 478–481, 2011.
- [13] N. Aronszajn, “Theory of reproducing kernels,” *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.
- [14] H. Avron, P. Maymounkov, and S. Toledo, “Blendenpik: Supercharging lapack’s least-squares solver,” *SIAM Journal on Scientific Computing*, vol. 32, no. 3, pp. 1217–1236, 2010.
- [15] F. Bach, “Consistency of trace norm minimization,” *Journal of Machine Learning Research*, vol. 9, pp. 1019–1048, June 2008.
- [16] —, “Sharp analysis of low-rank kernel matrix approximations,” in *International Conference on Learning Theory (COLT)*, December 2012.
- [17] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, “Structured sparsity through convex optimization,” *Statistical Science*, vol. 27, no. 4, pp. 450–468, 2012.
- [18] F. Bach, R. Jenatton, J. Mairal, G. Obozinski *et al.*, “Convex optimization with sparsity-inducing norms,” *Optimization for Machine Learning*, pp. 19–53, 2011.
- [19] P. L. Bartlett, O. Bousquet, and S. Mendelson, “Local Rademacher complexities,” *Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.
- [20] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [21] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Norwell, MA: Kluwer Academic, 2004.
- [22] P. J. Bickel, Y. Ritov, and A. Tsybakov, “Simultaneous analysis of Lasso and Dantzig selector,” *Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.
- [23] L. Birgé, “Estimating a density under order restrictions: Non-asymptotic min-max risk,” *Annals of Statistics*, vol. 15, no. 3, pp. 995–1012, March 1987.

- [24] A. Bordes, L. Bottou, and P. Gallinari, “Sgd-qn: Careful quasi-Newton stochastic gradient descent,” *Journal of Machine Learning Research*, vol. 10, pp. 1737–1754, 2009.
- [25] J. Bourgain, S. Dirksen, and J. Nelson, “Toward a unified theory of sparse dimensionality reduction in euclidean space,” *Geometric and Functional Analysis*, vol. 25, no. 4, 2015.
- [26] C. Boutsidis and P. Drineas, “Random projections for the nonnegative least-squares problem,” *Linear Algebra and its Applications*, vol. 431, no. 5–7, pp. 760–771, 2009.
- [27] C. Boutsidis, P. Drineas, and M. Mahdon-Ismail, “Near-optimal coresets for least-squares regression,” *IEEE Trans. Info. Theory*, vol. 59, no. 10, pp. 6880–6892, 2013.
- [28] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK: Cambridge University Press, 2004.
- [29] A. Bruckstein, D. Donoho, and M. Elad, “From sparse solutions of systems of equations to spares modeling of signals and images,” *SIAM Review*, 2007.
- [30] P. Bühlmann and S. van de Geer, *Statistics for high-dimensional data*, ser. Springer Series in Statistics. Springer, 2011.
- [31] F. Bunea, Y. She, and M. Wegkamp, “Optimal selection of reduced rank estimators of high-dimensional matrices,” vol. 39, no. 2, pp. 1282–1309, 2011.
- [32] R. H. Byrd, G. M. Chin, M. Gillian, W. Neveitt, and J. Nocedal, “On the use of stochastic Hessian information in optimization methods for machine learning,” *SIAM Journal on Optimization*, vol. 21, no. 3, pp. 977–995, 2011.
- [33] R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer, “A stochastic quasi-Newton method for large-scale optimization,” *arXiv preprint arXiv:1401.7020*, 2014.
- [34] E. J. Candes and T. Tao, “Decoding by linear programming,” *IEEE Trans. Info Theory*, vol. 51, no. 12, pp. 4203–4215, December 2005.
- [35] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, “The convex geometry of linear inverse problems,” *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.
- [36] S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. Sci. Computing*, vol. 20, no. 1, pp. 33–61, 1998.

- [37] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Majoney, X. Meng, and D. P. Woodruff, “The fast cauchy transform and faster robust linear regression,” in *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2013, pp. 466–477.
- [38] A. Cohen and A. Yeredor, “On the use of sparsity for recovering discrete probability distributions from their moments,” in *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, 2011.
- [39] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.
- [40] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines (and other kernel based learning methods)*. Cambridge University Press, 2000.
- [41] I. CVX Research, “CVX: Matlab software for disciplined convex programming, version 2.0,” Aug. 2012.
- [42] A. Dasgupta, R. Kumar, and T. Sarlós, “A sparse Johnson-Lindenstrauss transform,” in *Proceedings of the forty-second ACM symposium on Theory of computing*. ACM, 2010, pp. 341–350.
- [43] A. d’Aspremont and L. E. Ghaoui, “Testing the nullspace property using semidefinite programming,” Princeton, Tech. Rep., 2009.
- [44] K. R. Davidson and S. J. Szarek, “Local operator theory, random matrices, and Banach spaces,” in *Handbook of Banach Spaces*. Amsterdam, NL: Elsevier, 2001, vol. 1, pp. 317–336.
- [45] V. de La Pena and E. Giné, *Decoupling: From dependence to independence*. New York: Springer, 1999.
- [46] O. Dekel and Y. Singer, “Support vector machines on a budget,” *Advances in neural information processing systems*, vol. 19, p. 345, 2007.
- [47] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, “Inexact Newton methods,” *SIAM Journal on Numerical analysis*, vol. 19, no. 2, pp. 400–408, 1982.
- [48] R. S. Dembo and T. Steihaug, “Truncated Newton algorithms for large-scale unconstrained optimization,” *Mathematical Programming*, vol. 26, no. 2, pp. 190–212, 1983.
- [49] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

- [50] D. L. Donoho, M. Elad, and V. M. Temlyakov, “Stable recovery of sparse over-complete representations in the presence of noise,” *IEEE Trans. Info Theory*, vol. 52, no. 1, pp. 6–18, January 2006.
- [51] D. Donoho, I. Johnstone, and A. Montanari, “Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising,” *IEEE Trans. Info. Theory*, vol. 59, no. 6, pp. 3396 – 3433, 2013.
- [52] P. Drineas, M. Magdon-Ismail, M. Mahoney, and D. Woodruff, “Fast approximation of matrix coherence and statistical leverage,” *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3475–3506, 2012.
- [53] P. Drineas and M. W. Mahoney, “On the Nystrom method for approximating a Gram matrix for improved kernel-based learning,” *Journal of Machine Learning Research*, vol. 6, pp. 2153–2175, 2005.
- [54] —, “Effective resistances, statistical leverage, and applications to linear equation solving,” *arXiv preprint arXiv:1005.3097*, 2010.
- [55] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlos, “Faster least squares approximation,” *Numer. Math*, vol. 117, no. 2, pp. 219–249, 2011.
- [56] C. Dwork, “Differential privacy,” in *Encyclopedia of Cryptography and Security*. Springer, 2011, pp. 338–340.
- [57] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.
- [58] L. El Ghaoui, V. Viallon, and T. Rabbani, “Safe feature elimination in sparse supervised learning,” EECS Dept., University of California at Berkeley, Tech. Rep. UC/EECS-2010-126, September 2010.
- [59] M. Fazel, “Matrix rank minimization with applications,” Ph.D. dissertation, Stanford, 2002, available online: <http://faculty.washington.edu/mfazel/thesis-final.pdf>.
- [60] J. J. Fuchs, “Recovery of exact sparse representations in the presence of noise,” in *ICASSP*, vol. 2, 2004, pp. 533–536.
- [61] A. Gittens and M. W. Mahoney, “Revisiting the nystrom method for improved large-scale machine learning,” *arXiv preprint arXiv:1303.1849*, 2013.
- [62] G. Golub and C. V. Loan, *Matrix Computations*. Baltimore: Johns Hopkins University Press, 1996.

- [63] Y. Gordon, A. E. Litvak, S. Mendelson, and A. Pajor, “Gaussian averages of interpolated bodies and applications to approximate reconstruction,” *Journal of Approximation Theory*, vol. 149, pp. 59–73, 2007.
- [64] M. Grant and S. Boyd, “Graph implementations for nonsmooth convex programs,” in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.
- [65] C. Gu, *Smoothing spline ANOVA models*, ser. Springer Series in Statistics. New York, NY: Springer, 2002.
- [66] T. Hastie and B. Efron, “LARS: Least angle regression, Lasso and forward stagewise,” *R package version 0.9-7*, 2007.
- [67] T. Hastie, R. Tibshirani, and M. J. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall, New York: CRC Press, 2015.
- [68] T. R. Hastie, T. and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag, 2001.
- [69] P. Huber, “Robust regression: Asymptotics, conjectures and Monte Carlo,” *Annals of Statistics*, vol. 1, pp. 799–821, 2001.
- [70] W. B. Johnson and J. Lindenstrauss, “Extensions of Lipschitz mappings into a Hilbert space,” *Contemp. Math.*, vol. 26, pp. 189–206, 1984.
- [71] I. M. Johnstone, *Gaussian estimation: Sequence and wavelet models*. New York: Springer, To appear.
- [72] A. Joseph and A. R. Barron, “Fast sparse superposition codes have near exponential error probability for $R < C$,” *IEEE Trans. Info. Theory*, vol. 60, no. 2, pp. 919–942, Feb 2014.
- [73] D. M. Kane and J. Nelson, “Sparsier Johnson-Lindenstrauss transforms,” *Journal of the ACM*, vol. 61, no. 1, 2014.
- [74] D. Kane and J. Nelson, “Sparsier Johnson-Lindenstrauss transforms,” *Journal of the ACM*, vol. 61, no. 1, p. 4, 2014.
- [75] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, “An interior-point method for large-scale ℓ_1 -regularized least squares,” *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.
- [76] G. Kimeldorf and G. Wahba, “Some results on Tchebycheffian spline functions,” *Jour. Math. Anal. Appl.*, vol. 33, pp. 82–95, 1971.

- [77] J. Kivinen and M. K. Warmuth, “Exponentiated gradient versus gradient descent for linear predictors,” *Information and Computation*, vol. 132, no. 1, pp. 1–63, 1997.
- [78] V. Koltchinski and D. Panchenko, “Rademacher processes and bounding the risk of function learning,” in *High-dimensional probability II*. Springer-Verlag, 2000, pp. 443–459.
- [79] V. Koltchinskii, “Local Rademacher complexities and oracle inequalities in risk minimization,” *Annals of Statistics*, vol. 34, no. 6, pp. 2593–2656, 2006.
- [80] F. Krahmer and R. Ward, “New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property,” *SIAM Journal on Mathematical Analysis*, vol. 43, no. 3, pp. 1269–1281, 2011.
- [81] D. Lashkari and P. Golland, “Convex clustering with exemplar-based models,” *Advances in neural information processing systems*, vol. 20, 2007.
- [82] L. B. Lasserre, “An explicit exact SDP relaxation for nonlinear 0–1 programs,” *K. Aardal and A.M.H. Gerards, eds., Lecture Notes in Computer Science*, vol. 2081, pp. 293–303, 2001.
- [83] M. Laurent, “A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for 0-1 programming,” *Mathematics of Operations Research*, vol. 28, pp. 470–496, 2003.
- [84] M. Ledoux, *The Concentration of Measure Phenomenon*, ser. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society, 2001.
- [85] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. New York, NY: Springer-Verlag, 1991.
- [86] Y. Li, I. Tsang, J. T. Kwok, and Z. Zhou, “Tighter and convex maximum margin clustering,” in *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 2009, pp. 344–351.
- [87] P. Loh and M. J. Wainwright, “High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity,” *Annals of Statistics*, vol. 40, no. 3, pp. 1637–1664, September 2012.
- [88] L. Lovász and A. Schrijver, “Cones of matrices and set-functions and 0 – 1 optimization,” *SIAM Journal of Optimization*, vol. 1, pp. 166–190, 1991.
- [89] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in *Proc. Int’l Conf. Automatic Face and Gesture Recognition*, 1998, pp. 200–205.

- [90] M. W. Mahoney, “Randomized algorithms for matrices and data,” *Foundations and Trends in Machine Learning in Machine Learning*, vol. 3, no. 2, 2011.
- [91] H. M. Markowitz, *Portfolio Selection*. New York: Wiley, 1959.
- [92] P. Massart, “About the constants in Talagrand’s concentration inequalities for empirical processes,” *Annals of Probability*, vol. 28, no. 2, pp. 863–884, 2000.
- [93] J. Matousek, *Lectures on discrete geometry*. New York: Springer-Verlag, 2002.
- [94] P. McCullagh and J. Nelder, *Generalized linear models*, ser. Monographs on statistics and applied probability 37. New York: Chapman and Hall/CRC, 1989.
- [95] N. Meinshausen and P. Bühlmann, “High-dimensional graphs and variable selection with the Lasso,” *Annals of Statistics*, vol. 34, pp. 1436–1462, 2006.
- [96] S. Mendelson, “Geometric parameters of kernel machines,” in *Proceedings of COLT*, 2002, pp. 29–43.
- [97] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, “Reconstruction of sub-gaussian operators in asymptotic geometric analysis,” *Geometric and Functional Analysis*, vol. 17, no. 4, pp. 1248–1282, 2007.
- [98] R. Motwani and P. Raghavan, *Randomized Algorithms*. Cambridge, UK: Cambridge University Press, 1995.
- [99] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, “Restricted strong convexity and generalized linear models,” UC Berkeley, Department of Statistics, Tech. Rep., August 2011.
- [100] —, “A unified framework for high-dimensional analysis of M -estimators with decomposable regularizers,” *Statistical Science*, vol. 27, no. 4, pp. 538–557, December 2012.
- [101] S. Negahban and M. J. Wainwright, “Estimation of (near) low-rank matrices with noise and high-dimensional scaling,” *Annals of Statistics*, vol. 39, no. 2, pp. 1069–1097, 2011.
- [102] —, “Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise,” *Journal of Machine Learning Research*, vol. 13, pp. 1665–1697, May 2012.
- [103] J. Nelson and H. L. Nguyễn, “Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings,” in *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 2013, pp. 117–126.

- [104] Y. Nesterov, *Introductory Lectures on Convex Optimization*. New York: Kluwer Academic Publishers, 2004.
- [105] —, “Primal-dual subgradient methods for convex problems,” Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Tech. Rep., 2005.
- [106] —, “Gradient methods for minimizing composite objective function,” Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Tech. Rep. 76, 2007.
- [107] Y. Nesterov and A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied Mathematics, 1994.
- [108] V. Ojansivu and J. Heikkil, “Blur insensitive texture classification using local phase quantization,” in *Proc. Image and Signal Processing (ICISP 2008)*, 2008, pp. 236–243.
- [109] R. I. Oliveira, “Sums of random Hermitian matrices and an inequality by Rudelson,” *Elec. Comm. in Probability*, vol. 15, pp. 203–212, 2010.
- [110] M. R. Osborne, B. Presnell, and B. A. Turlach, “On the Lasso and its dual,” *Journal of Computational and Graphical Statistics*, vol. 2, no. 9, pp. 319–337, 2000b.
- [111] M. Pilanci, L. E. Ghaoui, and V. Chandrasekaran, “Recovery of sparse probability measures via convex programming,” in *Advances in Neural Information Processing Systems*, 2012, pp. 2420–2428.
- [112] M. Pilanci and M. J. Wainwright, “Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares,” UC Berkeley, Tech. Rep., 2014, full length version at arXiv:1411.0347.
- [113] —, “Newton sketch: A linear-time optimization algorithm with linear-quadratic convergence,” UC Berkeley, Tech. Rep., 2015. [Online]. Available: <http://arxiv.org/pdf/1505.02250.pdf>
- [114] —, “Randomized sketches of convex programs with sharp guarantees,” *IEEE Trans. Info. Theory*, vol. 9, no. 61, pp. 5096–5115, September 2015.
- [115] —, “Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares,” *Journal of Machine Learning Research*, pp. 1–33, 2015.
- [116] M. Pilanci, M. J. Wainwright, and L. El Ghaoui, “Sparse learning via boolean relaxations,” *Mathematical Programming*, vol. 151, no. 1, pp. 63–87, 2015.

- [117] G. Pisier, “Probabilistic methods in the geometry of Banach spaces,” in *Probability and Analysis*, ser. Lecture Notes in Mathematics. Springer, 1989, vol. 1206, pp. 167–241.
- [118] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984.
- [119] G. Raskutti, M. J. Wainwright, and B. Yu, “Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls,” *IEEE Trans. Information Theory*, vol. 57, no. 10, pp. 6976–6994, October 2011.
- [120] —, “Minimax-optimal rates for sparse additive models over kernel classes via convex programming,” *Journal of Machine Learning Research*, vol. 12, pp. 389–427, March 2012.
- [121] B. Recht, M. Fazel, and P. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.
- [122] V. Rokhlin and M. Tygert, “A fast randomized algorithm for overdetermined linear least-squares regression,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 36, pp. 13 212–13 217, 2008.
- [123] T. Sarlos, “Improved approximation algorithms for large matrices via random projections,” in *Foundations of Computer Science, 2006. FOCS’06. 47th Annual IEEE Symposium on*. IEEE, 2006, pp. 143–152.
- [124] C. Saunders, A. Gammerman, and V. Vovk, “Ridge regression learning algorithm in dual variables,” in *Proceedings of the Fifteenth International Conference on Machine Learning*, ser. ICML ’98, 1998, pp. 515–521.
- [125] M. Schmidt, E. van den Berg, M. Friedlander, and K. Murphy, “Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm,” *AISTATS, 2009*, vol. 5, 2009.
- [126] N. N. Schraudolph, J. Yu, and S. Günter, “A stochastic quasi-newton method for online convex optimization,” in *International Conference on Artificial Intelligence and Statistics*, 2007, pp. 436–443.
- [127] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press, 2004.
- [128] H. D. Sherali and W. P. Adams, “A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems,” *SIAM Journal on Discrete Mathematics*, vol. 3, pp. 411–430, 1990.

- [129] D. Spielman and N. Srivastava, “Graph sparsification by effective resistances,” *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1913–1926, 2011.
- [130] N. Srebro, N. Alon, and T. S. Jaakkola, “Generalization error bounds for collaborative prediction with low-rank matrices,” in *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2005.
- [131] I. Steinwart and A. Christmann, *Support vector machines*. New York: Springer, 2008.
- [132] G. W. Stewart and J. Sun, *Matrix perturbation theory*. New York: Academic Press, 1980.
- [133] C. J. Stone, “Optimal global rates of convergence for non-parametric regression,” *Annals of Statistics*, vol. 10, no. 4, pp. 1040–1053, 1982.
- [134] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.
- [135] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Foundations of Computational Mathematics*, vol. 12, pp. 389–434, 2012.
- [136] J. Tropp, “Just relax: Convex programming methods for subset selection and sparse approximation,” *ICES Report 04-04, UT-Austin, February.*, 2004.
- [137] J. A. Tropp, “Improved analysis of the subsampled randomized hadamard transform,” *Advances in Adaptive Data Analysis*, vol. 3, no. 01n02, pp. 115–126, 2011.
- [138] S. van de Geer, *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.
- [139] S. Vempala, *The Random Projection Method*, ser. Discrete Mathematics and Theoretical Computer Science. Providence, RI: American Mathematical Society, 2004.
- [140] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *Compressed Sensing: Theory and Applications*, 2012.
- [141] G. Wahba, *Spline models for observational data*, ser. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia, PN: SIAM, 1990.
- [142] M. J. Wainwright, “Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting,” *IEEE Trans. Info. Theory*, vol. 55, pp. 5728–5741, December 2009.

- [143] —, “Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (Lasso),” *IEEE Trans. Information Theory*, vol. 55, pp. 2183–2202, May 2009.
- [144] —, “Structured regularizers: Statistical and computational issues,” *Annual Review of Statistics and its Applications*, vol. 1, pp. 233–253, January 2014.
- [145] M. J. Wainwright and M. I. Jordan, “Treewidth-based conditions for exactness of the sherali-adams and lasserre relaxations,” UC Berkeley, Department of Statistics, No. 671, Tech. Rep., September 2004.
- [146] L. Wasserman, “Bayesian model selection and model averaging,” *Journal of mathematical psychology*, vol. 44, no. 1, pp. 92–107, 2000.
- [147] C. Williams and M. Seeger, “Using the Nyström method to speed up kernel machines,” in *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, 2001, pp. 682–688.
- [148] S. Wright and J. Nocedal, *Numerical optimization*. Springer New York, 1999, vol. 2.
- [149] T. T. Wu and K. Lange, “Coordinate descent algorithms for Lasso penalized regression,” *Annals of Applied Statistics*, vol. 2, no. 1, pp. 224–244, 2008.
- [150] N. Yamashita and M. Fukushima, “On the rate of convergence of the Levenberg-Marquardt method,” in *Topics in numerical analysis*. Springer, 2001, pp. 239–249.
- [151] Y. Yang, M. Pilanci, and M. J. Wainwright, “Randomized sketches for kernels: Fast and optimal non-parametric regression,” UC Berkeley, Tech. Rep., 2015. [Online]. Available: <http://arxiv.org/pdf/1501.06195.pdf>
- [152] B. Yu, “Assouad, Fano and Le Cam,” in *Festschrift for Lucien Le Cam*. Berlin: Springer-Verlag, 1997, pp. 423–435.
- [153] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, “Dimension reduction and coefficient estimation in multivariate linear regression,” *Journal Of The Royal Statistical Society Series B*, vol. 69, no. 3, pp. 329–346, 2007.
- [154] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society B*, vol. 1, no. 68, p. 49, 2006.
- [155] Y. Zhang, J. C. Duchi, and M. J. Wainwright, “Divide and conquer kernel ridge regression,” in *Computational Learning Theory (COLT) Conference*, Princeton, NJ, July 2013.

- [156] Y. Zhang, M. J. Wainwright, and M. I. Jordan, “Lower bounds on the performance of polynomial-time algorithms for sparse linear regression,” in *COLT conference*, Barcelona, Spain, June 2014, full length version at <http://arxiv.org/abs/1402.1918>.
- [157] P. Zhao and B. Yu, “On model selection consistency of Lasso,” *Journal of Machine Learning Research*, vol. 7, pp. 2541–2567, 2006.
- [158] S. Zhou, J. Lafferty, and L. Wasserman, “Compressed and privacy-sensitive sparse regression,” *IEEE Trans. Info. Theory*, vol. 55, pp. 846–866, 2009.
- [159] H. Zou and T. J. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society, Series B*, vol. 67, no. 2, pp. 301–320, 2005.