On Lasso parameter sensitivity

by

Aaron Berk

B.Sc. Mathematics & Statistics, McMaster University, 2013

M.Sc. Applied Mathematics, University of Toronto, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Mathematics)

The University of British Columbia

(Vancouver)

April 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

On Lasso parameter sensitivity

submitted by **Aaron Berk** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Mathematics**.

Examining Committee:

Yaniv Plan, Mathematics, UBC
Supervisor

Özgür Yilmaz, Mathematics, UBC
Co-Supervisor

Brian Wetton, Mathematics, UBC
Supervisory Committee Member

Bruce Shepherd, Computer Science, UBC
University Examiner

Philip Loewen, Mathematics, UBC
University Examiner

Jared Tanner, Mathematics, University of Oxford
External Examiner

Additional Supervisory Committee Members:

Michael Friedlander, Mathematics and Computer Science, UBC
Supervisory Committee Member

Abstract

Compressed sensing (CS) is a paradigm in which a structured high-dimensional signal may be recovered from random, under-determined, corrupted linear measurements. LASSO programs are effective for solving CS problems due to their proven ability to leverage underlying signal structure. Three popular LASSO programs are equivalent in a sense and sometimes used interchangeably. Tuned by a governing parameter, each admits an optimal parameter choice yielding minimax order-optimal error. CS is well-studied, though theory for LASSO programs typically concerns this optimally tuned setting. However, the optimal parameter value for a LASSO program depends on properties of the data, and is typically unknown in practical settings. Performance in empirical problems thus hinges on a program's *parameter sensitivity*: it is desirable that small variation about the optimal parameter choice begets small variation about the optimal risk.

We examine the risk of three LASSO programs as a function of their governing parameters and further demonstrate that their parameter sensitivity can differ for the same data, thereby informing the selection of a LASSO program in practice. We prove a *gauge-constrained* program admits asymptotic cusp-like behaviour of its risk in the limiting low-noise regime; a *residual-constrained* program has asymptotically suboptimal risk for very sparse vectors (i.e., for *any* fixed parameter choice, the ratio of the risk to the optimally achievable risk grows unbounded). These results contrast observations about an unconstrained program with sufficiently large parameter. Our theory is supported with extensive numerical simulations, demonstrating the parameter sensitivity phenomenon for even modest dimensional parameters.

We first analyze these risks for proximal denoising (PD), in which one directly observes signal plus noise (i.e., the measurement matrix is identity). There, we further reveal a data regime in which the unconstrained PD risk can be asymptotically suboptimal. We also show how our theory extends to analyze generalized Lasso programs for generalized CS. Finally, we extend a keystone of our theoretical analysis, *the projection lemma*. We generalize the result to an arbitrary Hilbert space, and extend it from scaled projections to proximal mappings of a dilated gauge. We discuss applications and possible directions for these results.

Lay Summary

CS is a paradigm to recover high-dimensional structured data from a small number of random measurements, with important applications in medical imaging, geophysics and data science. LASSO programs are popular in CS for implementing this recovery in practice. There are many LASSO variants and each is typically tuned by a governing parameter. Existing theory has thoroughly characterized LASSO performance when the parameter is chosen optimally. However, the optimal parameter value is typically unknown in practice. Therefore, it is important to characterize LASSO program performance for sub-optimal parameter values, too. Ideally, small deviations in the parameter value would still offer nearly optimal recovery performance. In this thesis, we prove the ideal is not upheld in general. The parameter choice for some LASSO programs can be highly sensitive to measurement noise; to underlying signal structure for others. Importantly, one program can be sensitive, even when others are not, which challenges the commonly held intuition that LASSO programs can be treated interchangeably.

Preface

A majority of this thesis is based on an article and two conference proceedings that have been published, as well as an article that has been submitted for publication.

A version of Chapter 3 has been published as *Sensitivity of ℓ_1 minimization to parameter choice*. Authored by myself, Plan Y and Yilmaz Ö (my advisors). It is published in *Information and Inference: A Journal of the IMA* (doi:10.1093/imaiai/iaaa014). A preliminary version of this work first appeared as a conference paper: *Parameter Instability Regimes in Sparse Proximal Denoising Programs*, in the proceedings for *2019 13th International conference on Sampling Theory and Applications (SampTA)* (pages 1–5), authored by myself, Plan Y and Yilmaz Ö. In both cases, I was the lead investigator for the work and was responsible for preparing the original manuscript and performing revisions.

The content of Chapter 4 is unpublished original work that has been submitted for publication. The manuscript, *On the best choice of* Lasso *program given data parameters* was co-authored with Plan Y and Yilmaz Ö. I was the lead investigator for this project, and was responsible for drafting the original manuscript. Additional credit goes to Li X, a fellow student of Plan Y and Yilmaz Ö, for helpful comments on the manuscript to improve correctness and clarity of presentation.

The content of Chapter 5 is unpublished original work excerpted from a manuscript that will be co-authored with Plan Y and Yilmaz Ö. I was the lead investigator for this project and was responsible for drafting the original manuscript. It is worth noting that this chapter begins with an original

presentation of a previously known result, which motivates the novel work comprising the chapter's remainder.

Table of Contents

Abstract . . . . . **iii**

Lay Summary . . . . . **v**

Preface . . . . . **vi**

Table of Contents . . . . . **viii**

List of Tables . . . . . **xii**

List of Figures . . . . . **xiii**

Glossary . . . . . **xv**

Acknowledgments . . . . . **xvi**

1 Introduction **1**

1.1 Related work 11

1.2 Roadmap 13

1.3 Notation 14

2 Mathematical Background **17**

2.1 Tools from convex analysis 17

2.1.1 Projection lemma 20

2.2 Tools from probability theory 20

2.2.1 Geometric tools from probability 24

2.3 Effective dimension of structured signals 26
2.3.1 Refined bounds on Gaussian mean width 28

3 Proximal denoising parameter sensitivity 30
3.1 Overview 30
3.2 On parameter sensitivity for (LS^*_τ) 33
3.3 On parameter sensitivity for (QP^*_λ) 34
3.3.1 Smoothness of the risk 35
3.3.2 Left-sided parameter sensitivity 36
3.3.3 Right-sided parameter stability 38
3.4 On parameter sensitivity for (BP^*_σ) 38
3.4.1 Underconstrained (BP^*_σ) 39
3.4.2 Overconstrained (BP^*_σ) 39
3.4.3 Minimax results 42
3.4.4 Maximin results 42
3.5 Numerical Results 44
3.5.1 (LS^*_τ) numerical simulations 45
3.5.2 (QP^*_λ) analytic plots 46
3.5.3 (BP^*_σ) numerical simulations 47
3.5.4 Parameter stability in sparse proximal denoising . . 51
3.5.5 Triptych comparison of program sensitivity 51
3.5.6 Realistic denoising examples 53
3.6 Proofs 62
3.6.1 Preliminary results 62
3.6.2 Proof for (LS^*_τ) parameter sensitivity 68
3.6.3 Proofs for (QP^*_λ) results 70
3.6.4 Proofs for (BP^*_σ) parameter sensitivity 76

4 Compressed sensing parameter sensitivity 89
4.1 Overview 89
4.2 On parameter sensitivity for (LS_τ) 92
4.3 A brief note regarding (QP_λ) 93
4.3.1 Right-sided parameter stability 93

4.4 On parameter sensitivity for (BP_σ) . . . 94
4.4.1 Underconstrained parameter sensitivity . . . 95
4.4.2 Minimax suboptimality . . . 96
4.5 Numerical results . . . 97
4.5.1 (LS_τ) numerics . . . 99
4.5.2 (QP_λ) numerics . . . 100
4.5.3 (BP_σ) numerics . . . 101
4.5.4 More synthetic examples . . . 104
4.5.5 Realistic Examples . . . 105
4.6 Proofs . . . 110
4.6.1 Risk equivalences . . . 110
4.6.2 $\hat{R}$ is nearly monotone . . . 111
4.6.3 Proofs for constrained LASSO sensitivity . . . 118
4.6.4 Proofs for basis pursuit suboptimality . . . 126

5 A well-ordering property for proximal operators . . . 155
5.1 Additional background . . . 155
5.1.1 Proof technique . . . 155
5.1.2 Proximal operators and projections . . . 158
5.2 A notion of ordering for proximal operators . . . 161
5.2.1 Derivation and notation . . . 161
5.2.2 Main result . . . 162
5.3 Quasi-ordering of generalized projections . . . 164
5.4 Applications . . . 170

6 Conclusion . . . 172
6.1 Proximal denoising . . . 172
6.2 Compressed sensing . . . 173
6.3 Projection lemma and extensions . . . 174
6.4 Future directions . . . 175

Bibliography . . . 177

A Supporting Materials 185
A.1 Auxiliary proofs (CS) 185
A.1.1 Proofs for refinements on bounds for Gaussian mean width (gmw) 185
A.1.2 Proofs for projection lemma 186
A.1.3 Parameter sensitivity of nuclear norm recovery . . . 186
A.2 RBF Approximation 188
A.2.1 Why RBF approximation? 188
A.2.2 RBF approximation of the average loss 189
A.2.3 Interpolation parameter settings 191

List of Tables

Table A.1 Average loss RBF interpolation parameter settings for (LS_τ) parameter sensitivity numerics 191
Table A.2 Average loss RBF interpolation parameter settings for (QP_λ) parameter sensitivity numerics 192
Table A.3 Average loss RBF interpolation parameter settings for (BP_σ) parameter sensitivity numerics 193
Table A.4 Average loss RBF interpolation parameter settings for 1D wavelet CS example 194
Table A.5 Average loss RBF interpolation parameter settings for 2D realistic examples 194

List of Figures

Figure 3.1 A visualization of the geometric lemma 41
Figure 3.2 (LS^*_τ) parameter sensitivity in the low-noise regime . . . 46
Figure 3.3 (QP^*_λ) parameter sensitivity in the low-noise regime . . 48
Figure 3.4 (BP^*_σ) parameter sensitivity in the very sparse regime . 50
Figure 3.5 Parameter stability of sparse PD programs 52
Figure 3.6 Dependence of each PD program on noise level η 53
Figure 3.7 Dependence of each PD program on sparsity level s . . . 54
Figure 3.8 Image-space denoising simulation 55
Figure 3.9 Haar wavelet denoising of 1D transform-sparse signal . . 58
Figure 3.10 Recovery correspondence for 1D Haar wavelet denoising 59
Figure 3.11 Haar wavelet denoising of transform-sparse image 62
Figure 3.12 Grid plots for transform-sparse image denoising ($\eta = 10^{-5}$) 87
Figure 3.13 Grid plots for transform-sparse image denoising ($\eta = 0.5$) 88

Figure 4.1 (LS_τ) parameter sensitivity in the low-noise regime . . . 100
Figure 4.2 (QP_λ) parameter sensitivity *vs.* aspect ratio 101
Figure 4.3 Realization variance in (QP_λ) parameter sensitivity . . . 102
Figure 4.4 Parameter sensitivity of (BP_σ) 145
Figure 4.5 Approximation quality and realization variance 146
Figure 4.6 Approximation quality and realization variance for (BP_σ) numerics 147
Figure 4.7 Parameter sensitivity in the low-noise, high-sparsity regime 148
Figure 4.8 Parameter sensitivity in the low-sparsity regime 148
Figure 4.9 Average loss behaviour for intermediate regime 149

Figure 4.10 CS parameter sensitivity for realistic 1D example 150
Figure 4.11 Realization variance for 1D CS example 151
Figure 4.12 Square Shepp-Logan phantom 151
Figure 4.13 Realization variance for realistic 2D CS example 152
Figure 4.14 Grid plots for realistic 2D CS example (low-noise) . . . 153
Figure 4.15 Grid plots for realistic 2D CS example (large noise) . . . 154

Figure 5.1 A visualization of the projection lemma 158

Glossary

CS	compressed sensing	1
GCS	generalized compressed sensing	5
gmw	Gaussian mean width	19
iid	independent identically distributed	15
lsc	lower semi-continuous	159
MRI	magnetic resonance imaging	1
mse	mean squared error	11
nnse	noise-normalized squared error	30
PD	proximal denoising	10
PS	parameter sensitivity	9
psnr	peak signal-to-noise ratio	98
RBF	radial basis function	54
RIP	restricted isometry property	22

Acknowledgments

Foremost, I express my gratitude to my advisors, Yaniv Plan and Özgür Yilmaz, for their encouragement, guidance and patience over the years. I feel fortunate to have benefitted from their mathematical and academic insight, and to have been regularly energized by our rich discussions. Thank you, both.

I thank Brian Wetton and Michael Friedlander for serving on my supervisory committee, Bruce Shepherd and Philip Loewen for serving as university examiners, and Jared Tanner for serving as the external examiner. I would moreover like to thank Brian and Philip for their careful attention and insightful comments which have surely improved the correctness and clarity of this dissertation.

I express my gratitude to Ipek Oruc, whose mentorship and research insight have deftly honed my applied research competencies.

I thank my friends and family for their boundless support over the years. A special thanks to Adela Gherga. Thank you to my parents, Norman and Barbara, and my brother Adam, who all are role models of hard work and encouragement.

Finally, I gratefully acknowledge the financial support I have received from NSERC, UBC, and the Province of BC.

Chapter 1

Introduction

$\{(x, y) : y \geq f(x)\}$ — Federico Poloni (2015)

A fundamental problem of signal processing concerns the development and analysis of efficacious methods for structured signal recovery that are widely applicable in practice. Frequently in applications, the signal is assumed to be structured according to some data model and measured by a particular acquisition method. For example, in image deblurring one might assume the objects of interest lie in the dual of a Besov space and their coefficients are obtained in the Radon transform domain [38, 51], while in magnetic resonance imaging (MRI) applications, one might assume the images are sparse in a wavelet domain, and measured by subsampling their Fourier coefficients [48, 49]. There is extensive literature concerned with those applications in which the goal is to recover the ground-truth signal from acquired measurements by a prescribed convex program that exploits the signal structure. For example, compressed sensing (CS) has demonstrated that a scale-invariant structure such as sparsity can be captured by convex optimization [35].

This latter point is a remarkable property of CS that may seem surprising at first. Indeed, the set of sparse vectors is non-convex, scale-invariant, *thin* (e.g., it has Lebesgue measure 0) and has infinite diameter; the feasible set of a CS convex program has full measure and (in some formulations) is bounded.

Nevertheless, in the coming pages, we shall describe how the convex setting may serve as a proxy for the true underlying structure. In turn, we will specialize our discussion to a particular group of convex programs for CS, and examine some issues that arise in constructing a convex program with the "wrong" proxy set.

The CS paradigm can be put into the following mathematical language. One aims to recover or approximate an unknown ground truth signal $x_0 \in \mathbb{R}^N$ from noisy measurements $y \in \mathbb{R}^m$ when $m \ll N$. In the simplest setting, CS prescribes that the signal x_0 be s-sparse (i.e., have no more than s non-zero entries), or sparse in a basis; and that the measurements be linear: $y = Ax_0 + \eta z$. Here, z is a random or deterministic corruption with noise scale $\eta > 0$. The *measurement matrix* $A \in \mathbb{R}^{m \times N}$ is typically *random*. We reserve a discussion on the kinds of admissible randomness for A until later.

The now classical CS result [20–22, 29, 30, 35] shows that when $x_0 \in \mathbb{R}^N$ is an s-sparse signal, $m \geq Cs \log(eN/s)$ measurements suffice to efficiently recover x_0 from (y, A) *with high probability* on the realization of A. In fact, so-called ℓ_0 minimization can achieve these optimal results with an error that is also proportional to $\eta^2 s \log(N/s)$. Remarkably, methods in convex optimization can also achieve these rates when their parameters are tuned correctly. Research has typically focused on this optimally tuned setting. However, the optimal parameter value depends on properties of the data, and is typically unknown in practical settings. In this thesis, we analyze three such convex optimization approaches for the entire range of their parameters. We show how two of the programs exhibit undesirable error rates in particular data regimes as a function of their parameter and show how our results inform a practitioner's selection of optimization program for solving CS problems.

The Lasso is a common and well-analyzed tool for effecting the recovery of x_0 [13, 24, 25, 35, 69, 73, 75]. Currently, "Lasso" is an umbrella term referring to three or more different convex optimization programs, though it originally referred to the ℓ_1-constrained program introduced by Tibshirani [73], which we introduce below as (LS_τ). Effective for its ability to perform simultaneous best-basis and subset selection [73], the Lasso is a convex

optimization approach that has several variants and cousins [13, 56, 72, 75].

The problem posed by CS is a type of under-determined linear inverse problem. Generally, under-determined inverse problems are of fundamental importance to modern mathematical and machine learning applications. In these problems, one aims to recover or approximate an unknown ground truth signal from a number of (noisy) measurements less than the dimension of the signal's ambient space. For example, in geophysics, one may wish to determine a region's bathymetry from a small number of radar measurements taken at the surface [46], or to obtain a subsurface image using a small number of geophones [42]. Recent investigations suggest how well-analyzed approaches for solving inverse problems may help to elucidate "mysterious" behaviours of high-dimensional non-linear function approximators [41, 50]. Moreover, compressed sensing theory may be used to prove recovery guarantees for certain neural network architectures and particular data regimes [39].

A recurring theme in the solution of under-determined linear inverse problems is a notion of the *effective dimension* of the underlying structure set. For example, solvability of CS hinges on the characterization of the effective dimension of the set of sparse signals. Several approaches exist for characterizing the effective dimension of sparse signals [2, 35, 47, 77]. Common to modern approaches is the notion of approximating the structure set by a convex structural proxy. For example, the convex proxy set for the set of sparse vectors is frequently the (scaled) ℓ_1-ball. Indeed, convexifying the structure set is the connection that permits the aforementioned efficient approximation of the ground truth via convex optimization. We discuss some approaches to quantifying the effective dimension of the convexified structure set in § 2.1.

To motivate this convexificiation process we provide a brief derivation of convex ℓ_1 minimization programs for compressed sensing. First, for $x \in \mathbb{R}^n$ define the so-called ℓ_0 norm of x by $\|x\|_0 := |\{j \in [n] : x_j \neq 0\}|$, where $[n] := \{1, \ldots, n\}$. The ℓ_0 norm gives the number of non-zero entries of x. An s-sparse signal x_0 thereby satisfies $\|x_0\|_0 \leq s$, and an exactly s-sparse signal $\|x_0\|_0 = s$. Importantly, $\|\cdot\|_0$ is not a norm, but is named suggestively,

because

$$\|x\|_0 = \lim_{p \to 0} \|x\|_p^p.$$

It is a well-known fact [35, Theorem 2.13] that if every subset of $2s$ columns of A is linearly independent (necessitating $m \geq 2s$), then one may uniquely recover an unknown s-sparse signal x_0 from noiseless (i.e., $\eta = 0$) linear measurements using ℓ_0 minimization:

$$\hat{x}_{\ell_0} := \arg\min \{\|x\|_0 : Ax = Ax_0\}.$$

Thus, without further considerations, ℓ_0 minimization is the natural choice for recovering sparse vectors. However, critically, Natarajan [54] established that any program for ℓ_0 minimization is NP-hard [3]. Moreover, implementations for ℓ_0 minimization are typically not *stable* in practice. Namely, they may be susceptible to error if the ground truth has *sparsity defect* [35] (i.e., if the true signal is not exactly s-sparse, but *close enough* to an exactly s-sparse vector in ℓ_2 norm).

Remarkably, the above optimization problem admits a convex relaxation with more favourable properties. Basis pursuit, referred to as (BP_σ) in this thesis, was introduced by Chen and Donoho [24], Chen et al. [25] and has been well studied [35]. Notably, (BP_σ) is stable with respect to sparsity defect, and robust to measurement noise ηz. Finally, the $Cs\log(N/s)$ sample complexity referred to previously is minimax order-optimal over the class of arbitrary estimators that yield stable recovery [17, 35]. In addition to (LS_τ) and (BP_σ), which are both constrained optimization programs, we introduce a third variant of the Lasso, known as the unconstrained Lasso, which is hereafter referred to as (QP_λ). These three programs and their solutions are

of particular interest to this thesis, and are defined as:

$$\hat{x}(\tau) \in \underset{x\in\mathbb{R}^N}{\arg\min} \left\{ \|y - Ax\|_2^2 : \|x\|_1 \leq \tau \right\} \tag{LS$_\tau$}$$

$$\tilde{x}(\sigma) \in \underset{x\in\mathbb{R}^N}{\arg\min} \left\{ \|x\|_1 : \|y - Ax\|_2^2 \leq \sigma^2 \right\} \tag{BP$_\sigma$}$$

$$x^\sharp(\lambda) \in \underset{x\in\mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1 \right\}. \tag{QP$_\lambda$}$$

The advent of suitable fast and scalable algorithms has made this collection of convex ℓ_1 minimization programs extremely useful in practice [36, 37, 57, 75]. Note that some naming ambiguity for these programs exists in the literature. Our notation and naming convention for these three programs is similar to that used in Van Den Berg and Friedlander [75]. A purpose of this thesis is to compare and contrast these three programs. Thus, we reserve until later further discussion on their interrelation.

Generalizations of these programs, commonly referred to as generalized Lasso, allow for the recovery of signals with other kinds of structure that are well modelled by convex proxy sets. For instance, generalized Lasso programs are favoured for solving generalized compressed sensing (GCS) problems. To introduce the three generalized Lasso programs of greatest relevance to this thesis, first let $\emptyset \neq \mathcal{K} \subseteq \mathbb{R}^N$ be a convex set and denote by $\|\cdot\|_\mathcal{K}$ the Minkowski functional of $\mathcal{K}$ (i.e., gauge, as defined in § 2.1). For $\sigma, \tau, \lambda > 0$, the following *generalized* Lasso programs, which are convex, are defined by:

$$\hat{x}(\tau; A, y, \mathcal{K}) \in \underset{x\in\mathbb{R}^N}{\arg\min} \left\{ \|y - Ax\|_2^2 : x \in \tau\mathcal{K} \right\} \tag{LS$_{\tau,\mathcal{K}}$}$$

$$\tilde{x}(\sigma; A, y, \mathcal{K}) \in \underset{x\in\mathbb{R}^N}{\arg\min} \left\{ \|x\|_\mathcal{K} : \|y - Ax\|_2^2 \leq \sigma^2 \right\} \tag{BP$_{\sigma,\mathcal{K}}$}$$

$$x^\sharp(\lambda; A, y, \mathcal{K}) \in \underset{x\in\mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_\mathcal{K} \right\}. \tag{QP$_{\lambda,\mathcal{K}}$}$$

For brevity of notation, when it is clear from context, we omit explicit dependence of $\hat{x}, \tilde{x}, x^\sharp$ on y, A and $\mathcal{K}$. Note that the argmin may not be unique. When A is under-determined and has a suitable randomness, it is straight-

forward to show the programs (BP_σ) and (QP_λ) admit unique solutions almost surely on the realization of A. A detailed exposition for (QP_λ) is given in Tibshirani [74]. For a sufficient condition on A giving uniqueness of (BP_σ), see Zhang et al. [80]. However, (LS_τ) does not always admit a unique solution. For instance, if τ is "too large", then there may be infinitely many solutions $x \in \tau B_1^N$ satisfying $\|y - Ax\|_2 = 0$. This fact is fundamental to one of our results in § 4.2. By mild abuse of notation, when the solution to a program is unique we will replace "$\in$" with "$=$" in the definitions of the solutions for each program. Otherwise, we define each of $\hat{x}(\tau), \tilde{x}(\sigma)$, and $x^\sharp(\lambda)$ as the solution yielding worst-case error, and which appears first when ordered lexicographically. For example, $\hat{x}(\tau)$ refers to the particular solution solving (LS_τ) such that $\|\hat{x}(\tau) - x_0\|_2 \geq \|\hat{x} - x_0\|_2$ for any other $\hat{x}$ solving (LS_τ). We make an analogous modification to the definitions of the solutions to the generalized Lasso programs.

In the standard CS setting, the gauge is the ℓ_1-norm, though x_0 is assumed to belong to the set of s-sparse vectors $\Sigma_s^N := \{x \in \mathbb{R}^N : |\operatorname{supp}(x)| \leq s\}$, where $\operatorname{supp}(x) := \{j \in [N] : x_j \neq 0\}$ denotes the support of the vector $x \in \mathbb{R}^N$. So, x_0 does not necessarily belong to the convex proxy set $\mathcal{K} = B_1^N$, where $B_1^N := \{x \in \mathbb{R}^N : \|x\|_1 \leq 1\}$ denotes the N-dimensional unit 1-norm ball. In particular, B_1^N itself serves as a convex proxy set for sparse vectors in the sense that if $x \in \mathbb{R}^N$ is s-sparse, then $\|x\|_1/\|x\|_2$ is small relative to non-sparse vectors. We include below several other examples of this general set-up, noting that 2 is a repetition of the CS programs already introduced:

1. To obtain total variation (TV) denoising for (continuous-valued discrete) images, define for $x \in \mathbb{R}^{N \times N}$,

$$\|x\|_{\mathrm{BV}} := \|x\|_1 + \sum_{\alpha \in [N]^2} \sum_{\beta \in \nu(\alpha)} |x_\alpha - x_\beta|,$$

   where $\nu : [N]^2 \to \mathcal{P}([N]^2)$ is the neighbour map that determines which "pixels" x_β of the image are the neighbours of the pixel x_α. If $\alpha = (i, j)$ and $2 \leq i, j \leq N - 1$ then one typically has $\nu(i, j) = \{(i - 1, j), (i, j - 1), (i + 1, j), (i, j + 1)\}$ with a variety of choices for

the remaining indices. So defined, $x^\sharp(\lambda; y, I, K)$ is a well-known denoising model for two-dimensional images when $A = I$ is the identity matrix and $\mathcal{K} := \{\|x\|_{\mathrm{BV}} \leq 1\}$ [66]. Instead defining $\|x\|_{\mathrm{BV}} := \|x\|_1 + \sum_{i=1}^{N-1} |x_{i+1} - x_i|$ for $x \in \mathbb{R}^N$, one obtains an equivalent denoising method for one-dimensional signals. With minor modification of $x^\sharp(\lambda)$ to allow for A to act as a bounded linear operator on $x \in \mathbb{R}^{N \times N}$ (e.g., convolution with a Gaussian kernel), one may extend the model for image deblurring [28].

2. Say that $x \in \mathbb{R}^N$ is s-sparse if $x \in \Sigma_s^N := \{x \in \mathbb{R}^N : \|x\|_0 \leq s\}$ where $\|x\|_0 = |\{j : x_j \neq 0\}|$. Define $\mathcal{K} := B_1^N$, where $B_1^N \subseteq \mathbb{R}^N$ is the unit ℓ_1 ball. Suppose $x_0 \in \mathbb{R}^N$ is s-sparse for some $s \geq 1$ and suppose that $A \in \mathbb{R}^{m \times N}$ is a Gaussian random matrix with $A_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, m^{-1/2})$. Then we obtain three common variants of the LASSO that solve the "vanilla" CS problem: the constrained LASSO yielding $\hat{x}(\tau; y, A, \mathcal{K})$, basis pursuit denoise yielding $\tilde{x}(\sigma; y, A, \mathcal{K})$, and the unconstrained LASSO yielding $x^\sharp(\lambda; y, A, \mathcal{K})$.

3. When $A = I$ is the identity matrix, $(\mathrm{LS}_{\tau,\mathcal{K}})$ yields the orthogonal projection onto $\tau\mathcal{K}$, which we denote by $\mathrm{Proj}_{\tau\mathcal{K}}(y) := \hat{x}(\tau; y, I, \mathcal{K})$. Similarly, $(\mathrm{QP}_{\lambda,\mathcal{K}})$ yields the proximal operator for the gauge induced by $\mathcal{K}$, which we denote by $\mathrm{prox}_{\lambda^{-1}\mathcal{K}}(y) := x^\sharp(\lambda; y, I, \mathcal{K})$. Proximal operators are the workhorses of proximal algorithms: projected gradient descent methods rely on $\mathrm{Proj}_{\tau\mathcal{K}}(y)$, while proximal gradient descent methods rely on $\mathrm{prox}_{\lambda^{-1}\mathcal{K}}(y)$.

4. Assume that $x' \in \mathbb{R}^N$ is s-sparse and let $x_0 = \psi^{-1}x'$ where ψ is the orthonormal DFT matrix. Given $y = x_0 + \eta z$, the vector $\hat{x}(\tau; y, \psi^{-1}, B_1^N)$ gives an analogue of running so-called constrained proximal denoising in Fourier space.

5. Consider a matrix $x \in \mathbb{R}^{N \times N}$, let $\|x\|_*$ denote its nuclear norm and define $\mathcal{K} := \{x \in \mathbb{R}^{N \times N} : \|x\|_* \leq 1\}$. Then $\tilde{x}(\sigma)$ gives the standard [34] optimization program for recovering a low-rank matrix $x_0 \in \mathbb{R}^{N \times N}$ from measurements $Ax := \langle A_i, x \rangle = \sum_{\alpha \in [N]^2} A_{i,\alpha} x_\alpha$.

To relate the recovery performance of each program, we compare their recovery errors. While there are several possibilities for measuring the recovery error, the expected squared error and noise-normalized expected squared error of the estimator are common when the noise, z, is random [56]. In this thesis, we'll define the loss for an estimator as the noise normalized squared error of that estimator (with respect to the ground truth signal x_0); and define the estimator's risk as the expectation of the loss with respect to z. Note that the risk and loss are functions of the random matrix A. Specifically, the loss is defined for (LS_τ), (BP_σ), (QP_λ) respectively by:

$$\hat{L}(\tau; x_0, A, \eta z) := \eta^{-2}\|\hat{x}(\tau) - x_0\|_2^2, \tag{1.1}$$

$$\tilde{L}(\sigma; x_0, A, \eta z) := \eta^{-2}\|\tilde{x}(\sigma) - x_0\|_2^2, \tag{1.2}$$

$$L^\sharp(\lambda; x_0, A, \eta z) := \eta^{-2}\|x^\sharp(\eta\lambda) - x_0\|_2^2, \tag{1.3}$$

and the risk by:

$$\hat{R}(\tau; x_0, A, \eta) := \mathop{\mathbb{E}}_z \hat{L}(\tau; x_0, A, \eta z) \tag{1.4}$$

$$\tilde{R}(\sigma; x_0, A, \eta) := \mathop{\mathbb{E}}_z \tilde{L}(\sigma; x_0, A, \eta z) \tag{1.5}$$

$$R^\sharp(\lambda; x_0, A, \eta) := \mathop{\mathbb{E}}_z L^\sharp(\lambda; x_0, A, \eta z). \tag{1.6}$$

Similar definitions may be given for the generalized Lasso variants.

Minimax order-optimal error rates are well-known for $\hat{x}(\tau; y, A, \mathcal{K})$ when τ is equal to the optimal parameter choice, A is a matrix whose rows are independent, isotropic subgaussian random vectors, and $\mathcal{K}$ is a symmetric, closed convex set containing the origin [35, 47, 56]. A kind of equivalence between the three estimators (*cf.* Proposition 4.1.4) allows, in kind, for the characterization of the error rates for $\tilde{x}(\sigma)$ and $x^\sharp(\lambda)$ when σ and λ are optimally tuned. However, the error of $\hat{x}(\tau; y, A, \mathcal{K})$ is not fully characterized in the setting where τ is not the optimal choice. Similarly, the programs (LS_τ), (BP_σ) and (QP_λ) are often referred to interchangeably, but a full comparison of the error of the three estimators $\hat{x}(\tau), \tilde{x}(\sigma)$, and $x^\sharp(\lambda)$, as a function of their governing parameters, is lacking. It is an open question if

there are settings in which one estimator is always preferable to another.

Understanding the sensitivity of a LASSO program to its parameter choice is crucial. While theoretical guarantees for recovery error are typically given for an oracular choice of the parameter, the optimal parameter setting is generally unknown in practice. Thus, the usefulness of theoretical recovery guarantees may hinge on the assumption that the recovery error is stable with respect to variation of the governing parameter. In particular, one may hope that small changes in the governing parameter beget no more than small changes in the risk or loss. In other words, if the optimal choice of parameter yields order-optimal recovery error, then one may hope that a "nearly" optimal choice of parameter admits "nearly" order-optimal recovery error, too, in the sense that the discrepancy in error is no greater than a multiplicative constant that depends smoothly on the discrepancy in parameter choice. For example, if $R(\alpha)$ is the mean-squared error of a convex program with parameter $\alpha > 0$, and $\alpha^* > 0$ is the value yielding minimal error, then one may hope for smooth dependence on α, such as

$$R(\alpha) \lesssim \mu(\alpha) R(\alpha^*),$$

where $\mu : \mathbb{R} \to \mathbb{R}^+$ is a nonnegative smooth function with $\mu(\alpha^*) = 1$. For example, the risk for (QP_λ) satisfies such an expression with $\mu(\lambda) \approx (\lambda/\lambda^*)^2$ when $\lambda \geq \lambda^*$ [13, 67, 71, 72].

In Chapter 4, we take a step toward characterizing the performance and sensitivity of the three programs introduced by examining particular asymptotic parameter regimes for each program. In the setting where A has rows that are independent, isotropic subgaussian random vectors, we prove the existence of regimes in which CS programs exhibit parameter sensitivity (PS). Namely, the programs exhibit sensitivity to their parameter choice, in that small changes in parameter values can lead to blow-up in risk. Despite the notion of equivalence captured by Proposition 4.1.4, we demonstrate regimes in which one program exhibits sensitivity, while the other two do not. For example, in the very sparse regime, our theory and simulations suggest not to use (BP_σ). In the low-noise regime, they suggest not to use (LS_τ). As-

suredly, we identify situations where CS programs perform well in theory and *in silico* alike. As an additional aim of this thesis, we hope that the asymptotic theory and fairly extensive numerical simulations aid practitioners in deciding which CS program to select.

A precursor to analyzing the more general CS setting, the simpler setting of proximal denoising (PD) is treated in Chapter 3. This treatment follows the original of that published in Berk et al. [8, 9]. PD is a simplification of its more general CS counterpart, in which the measurement matrix is identity. PD uses convex optimization as a means to recover a structured signal corrupted by additive noise. We define three convex programs for PD: constrained proximal denoising, basis pursuit proximal denoising, and unconstrained proximal denoising. To bear greatest relevance to CS, we assume that x_0 is s-sparse and that $y = x_0 + \eta z$, where $z \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\eta > 0$. For $\tau, \sigma, \lambda > 0$, respectively,

$$\hat{x}(\tau; y, I_N, B_1^N) := \underset{x \in \mathbb{R}^N}{\arg\min} \left\{ \|y - x\|_2^2 : \|x\|_1 \leq \tau \right\} \tag{LS$^*_\tau$}$$

$$\tilde{x}(\sigma; y, I_N B_1^N) := \underset{x \in \mathbb{R}^N}{\arg\min} \left\{ \|x\|_1 : \|y - x\|_2^2 \leq \sigma^2 \right\} \tag{BP$^*_\sigma$}$$

$$x^\sharp(\lambda; y, I_N, B_1^N) := \underset{x \in \mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2}\|y - x\|_2^2 + \lambda \|x\|_1 \right\}. \tag{QP$^*_\lambda$}$$

These are clear simplifications of (LS$_{\tau,\mathcal{K}}$), (QP$_{\lambda,\mathcal{K}}$) and (BP$_{\sigma,\mathcal{K}}$) introduced above, in which $\mathcal{K} = B_1^N$ is the ℓ_1 ball and where we use * to denote that the measurement matrix is identity, $A = I_N \in \mathbb{R}^{N \times N}$. Where clear from context, we may again overload notation and, for example, refer to $\hat{x}(\tau; y, I_N, B_1^N)$ simply as $\hat{x}$ or $\hat{x}(\tau)$. Since the focus of this thesis concerning PD is largely relegated to Chapter 3, and that for CS to Chapter 4, we are confident there shall be no ambiguity.

In both the PD setting of Chapter 3 and CS setting of Chapter 4, we explore PS numerically. In particular, we analyze sensitivity for the three programs in myriad data regimes, both for completely synthetic experiments and for more realistic experiments using a modification of the Shepp-Logan phantom [68], a standard test image for image reconstruction algorithms.

1.1 Related work

PD is a simple model that elucidates crucial properties of models in general [32]. As a central model for denoising, it lays the groundwork for CS, deconvolution and inpainting problems [33]. A fundamental signal recovery phase transition in CS is predicted by geometric properties of PD [2], because the minimax risk for PD is equal to the statistical dimension of the signal class [55]. This quantity is a generalized version of $R^*(s, N)$ introduced below.

Robustness of PD to inexact information is discussed briefly in Oymak and Hassibi [55], wherein sensitivity to constraint set perturbation is quantified, including an expression for right-sided stability of unconstrained PD. Essentially, PD programs are proximal operators, a powerful tool in convex and non-convex optimization [14, 27]. For a thorough treatment of proximal operators and proximal point algorithms, we refer the reader to Bertsekas et al. [11], Eckstein and Bertsekas [31], Rockafellar [64]. Thus, PD is interesting in its own right, as argued in Oymak and Hassibi [55].

Equivalence of the above programs is illuminated from several perspectives [11, 55, 75]. PD risk is considered with more general convex constraints in Chatterjee [23]. A connection has been made between the risk of Unconstrained Lasso and $R^\sharp(\lambda; x_0, N, \eta)$ in Bayati and Montanari [5, 6], and the risk of unconstrained Lasso, as a function of its governing parameter, has been analyzed in Thrampoulidis et al. [70, 72]. In addition, there are near-optimal error bounds for worst-case noise demonstrating that equality-constrained basis pursuit ($\sigma = 0$) performs well under the noisy CS model ($\eta \neq 0$) [79]. It should be noted that these results do not contradict those of this thesis, as random noise can be expected to perform better than worst-case noise in general. Recently, Miolane and Montanari [52, Theorem 3.2] proved a bound on the unconstrained Lasso mean squared error (mse), which is uniform in λ and uniform in $x_0 \in B_p^N$. Note that this also does not run contrary to the left-sided PS result mentioned below as the uniformity in λ is over a pre-specified interval chosen independently of the optimal parameter choice λ^*, and the assumption on signal structure is different.

Several versions of the Lasso program are well-studied in the context of solving CS problems [35]. The program (LS_τ) was first posed in Tibshirani [73]. An analysis of its risk when $\tau = \|x_0\|_1$ and the noise z is deterministic may be found in Foucart and Rauhut [35]. A sharp non-asymptotic analysis for the generalized constrained Lasso may be found in Oymak et al. [56]. There, the risk was shown to depend on specific geometric properties of the regularizer. When the measurement matrix has independent isotropic subgassian rows, Liaw et al. [47] demonstrated how a geometric quantity may unify the quantification of generalized constrained Lasso risk. Risk bounds for generalized constrained Lasso with nonlinear observations were characterized in Plan and Vershynin [62]. Recent work has shown how dimensional parameters governing signal recovery problems in ridgeless least squares regression affect the average out-of-sample risk in some settings [41, 50].

Non-asymptotic bounds for the unconstrained Lasso were developed in Bickel et al. [13], which also determines an order-optimal choice for the program's governing parameter. The asymptotic risk for the unconstrained Lasso is determined analytically in Bayati and Montanari [5, 6]. Sharp, non-asymptotic risk bounds for the generalized unconstrained Lasso are developed in Thrampoulidis et al. [70, 71]. In Thrampoulidis et al. [71], $R^\sharp(\lambda)$ is examined for λ about λ_{opt}, while Thrampoulidis et al. [72] examined the risk as a function of its governing parameter for other kinds of M-estimators. Both assume Gaussianity of the data, and neither considers sensitivity with respect to parameter choice.

Basis pursuit is a third popular phrasing of the Lasso program, first proposed in Chen and Donoho [24], Chen et al. [25]. For a theoretical treatment of basis pursuit, we refer to Foucart and Rauhut [35]. Analytic connections between basis pursuit and other Lasso programs are exploited for fast computation of solutions in Van Den Berg and Friedlander [75].

Other modifications of the standard Lasso have also been examined. For example, sharp non-asymptotic risk bounds for the so-called square-root Lasso were obtained in Oymak et al. [56]. Related to basis pursuit, instance optimality of an exact ℓ_1 decoder is analyzed in Wojtaszczyk [79].

As already noted, sensitivity to parameter choice was analyzed for three

proximal denoising (PD) programs that have CS analogues [8, 9]. There, Berk et al. [9] prove an asymptotic cusp-like behaviour for constrained PD risk in the low-noise regime, an asymptotic phase transition for unconstrained PD risk in the low-noise regime, and asymptotic suboptimality of the basis pursuit PD risk in the very sparse regime. Berk et al. [10] develop non-trivial generalizations of those results, proving asymptotic results about the sensitivity of ℓ_1 minimization for the generalized constrained LASSO and generalized basis pursuit. These works comprise two main chapters of this thesis.

1.2 Roadmap

The main content of this thesis is developed in Chapters 3, 4 and 5. In Chapter 3, we begin by establishing theory for PD parameter sensitivity. First, we show that (LS^*_τ) risk develops a cusp-like behaviour in the limiting low-noise regime, admitting a single optimal choice of its governing parameter. Next, in contrast, we highlight how (QP^*_λ) exhibits right-sided parameter stability, and prove the existence of a data regime in which (QP^*_λ) exhibits left-sided PS. Namely, (QP^*_λ) asymptotically exhibits PS if its governing parameter is below the optimal choice. Finally, we establish a data regime in which (BP^*_σ) risk is asymptotically sub-optimal for any choice of its governing parameter. Following the theory, we showcase extensive numerical simulations for PD supporting the theoretical results. In particular, our simulations suggest that parameter sensitivity manifests for realistic problem sizes. They further showcase how a PD loss may exhibit cusp-like behaviour, though this is not directly suggested by theory pertaining to the associated risk.

The content of Chapter 4 is structured similarly to Chapter 3, and develops theory for CS parameter sensitivity. First, we show that (LS_τ) risk develops a cusp-like behaviour in the limiting low-noise regime, admitting a single optimal choice of its governing parameter. In fact, components of this result are proved in the stronger setting of GCS for $\hat{L}$. Again, we contrast the PS result by highlighting how (QP_λ) exhibits parameter stability if its governing parameter is at least as large as the optimal choice. Finally, we prove

existence of a data regime in which (BP_σ) risk is asymptotically suboptimal for any choice of its governing parameter (with convergence in probability on A). Following the theory, we showcase extensive numerical simulations for CS supporting the theoretical results. We additionally note how some commonality is borne between the PD and CS numerical simulations. The numerical simulations again demonstrate that PS is readily observed for realistic problem sizes. Moreover, the simulations suggest a relationship between the sensitivity of the loss and that of its associated risk.

Some of the aforementioned results rely strongly on Lemma 2.1.2, presented in § 2.1.1. In particular, this key result concerns the ordering of the norm of points projected onto scaled sets. In Chapter 5, we show that this result may be generalized to an arbitrary Hilbert space, and extended from projection operators to proximal operators. Due to the difference in theme of this content, we reserve until § 5.1 a discussion of additional background. Though Lemma 2.1.2 is not a novel result, our original proof is presented in § 5.1.1 to build intuition for the fully general case. Helpful intermediate results are given in § 5.1.2. Finally, the main result of the chapter is presented in § 5.2.2, with applications and potential directions of interest described in § 5.4.

1.3 Notation

We use the standard notation for the p-norm, $\|\cdot\|_p$, for values $p \geq 1$. For $0 \leq p \leq \infty$, denote the unit ℓ_p-ball in $\mathbb{R}^n$ by $B_p^n := \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$. We use sgn to denote the operator that returns a vector whose elements are the "signs" of the argument. Namely, if $x \in \mathbb{R}^n$ then $\mathrm{sgn}(x) \in \mathbb{R}^n$ and $\mathrm{sgn}(\alpha) := 1$ if $\alpha > 0$, $\mathrm{sgn}(\alpha) = -1$ if $\alpha < 0$ and $\mathrm{sgn}(0) = 0$. We denote the set whose elements are the first n positive integers by $[n] := \{1, 2, \ldots, n\}$.

We occasionally make use of the notation $\|x\|_0 := |\{i \in [N] : x_i \neq 0\}|$ to denote the number of non-zero entries of a vector x; $\|\cdot\|_0$ is not a norm. Throughout this document, $m, n, N \in \mathbb{N}$ will be positive integers representing dimension. We use the notation $\mathcal{N}(\mu, \sigma^2)$ to refer to the univariate normal distribution with mean μ and variance σ^2; $\mathcal{N}(\vec{\mu}, \Sigma)$ the multi-

dimensional normal distribution with vector mean $\vec{\mu}$ and covariance matrix Σ. If a random vector $g \in \mathbb{R}^n$ has independent identically distributed (iid) standard normal entries we may write either $g_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ or equivalently $g \sim \mathcal{N}(0, I_n)$. In the latter expression, 0 denotes the 0 vector and I_n the $n \times n$ identity matrix. Denote the set of at most s-sparse vectors by $\Sigma_s^N := \{x \in \mathbb{R}^N : 0 \leq \|x\|_0 \leq s\}$ and define $\Sigma_{-1}^N := \emptyset$. Let $x_0 \in \Sigma_s^N \subseteq \mathbb{R}^N$ be an s-sparse signal with support set $T \subseteq [N] := \{1, 2, \ldots, N\}$, where $s \ll N$. Whereas x_0 shall refer to the "ground truth" signal for a given problem instance, we use x or x' to denote an arbitrary s-sparse signal, except where otherwise noted. For a vector $x \in \mathbb{R}^N$ with support set $S \subseteq [N]$, we denote $x_S \in \mathbb{R}^N$ to be the vector with entries $(x_S)_i = x_i$ for $i \in S$ and $(x_S)_i = 0$ otherwise. We shall use z to represent a normal random vector with covariance matrix equal to the identity: $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\eta > 0$ to represent the standard deviation of the noise (the "noise scale"), and y to represent a set of noisy linear measurements. For example, in Chapter 3, we shall typically have $z \in \mathbb{R}^N$ with $y = x_0 + \eta z \in \mathbb{R}^N$ and in Chapter 4 we shall typically have $z \in \mathbb{R}^m$ with $y = Ax_0 + \eta z \in \mathbb{R}^m$ where $A \in \mathbb{R}^{m \times N}$ is the *sensing matrix*. For convenience, we may use $Z \sim \mathcal{N}(0,1)$ to denote a univariate standard normal random variable. Let ϕ and Φ denote the standard univariate normal pdf and cdf, respectively. In particular, $\phi(t) := \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$ and $\Phi(t) = \mathbb{P}(Z \leq t) = \int_{-\infty}^{t} \phi(u)\,\mathrm{d}u$.

As discussed above, reference to $\hat{R}$, $\tilde{R}$ and $R^\sharp$; $\hat{L}$, $\tilde{L}$ and $L^\sharp$ could be in either the PD or CS setting. This will be connoted by the argument list when ambiguity may be present. Specifically, if the argument list includes the sensing matrix (e.g., $\hat{R}(\tau; x_0, A, \eta)$), then the function refers to the CS variant; otherwise, to the PD variant (e.g., $\hat{R}(\tau; x_0, N, \eta)$). The argument list may be omitted when the setting is clear from context. Finally, we clarify notation of constants and asymptotics. We will write $\tilde{K} := K\sqrt{\log K}$, and in each such instance it will be the case that K is a constant with $K > 1$. We will also use C to denote an absolute positive constant whose value may change from one appearance to the next. When we write something like C_ε, we likewise refer to a positive constant whose value depends only on ε. Throughout this thesis we may make use of standard Bachmann-Landau

notation (i.e., asymptotic notation), similar to how it was defined in Knuth [45]. In particular, say that $g(n) = \mathcal{O}(f(n))$ if there exist positive constants C and n_0 such that $|g(n)| \leq Cf(n)$ for all $n \geq n_0$. Say that $g(n) = \Omega(f(n))$ if there exist positive constants C and n_0 such that $|g(n)| \geq Cf(n)$ for all $n \geq n_0$. Say that $g(n) = \Theta(f(n))$ if $g(n) = \mathcal{O}(f(n))$ and $g(n) = \Omega(f(n))$.

All additional notation used in this thesis shall be introduced in context.

Chapter 2

Mathematical Background

In this chapter, we synthesize several known results from convex analysis and probability theory, some with proof sketches to provide intuition. We outline notation to refer to common objects from convex analysis. We introduce two well-known tools for characterizing the *effective dimension* of a set, and state a result that connects these tools with PD estimators [55]. In § 2.1.1, we re-state a lemma first appearing in Oymak and Hassibi [55]. This projection lemma gives a notion of ordering for projection operators. In § 2.3.1 we state two recent results giving refined bounds on the Gaussian mean width of convex polytopes intersected with Euclidean balls [7].

2.1 Tools from convex analysis

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a convex function and let $x \in \mathbb{R}^N$. Denote by $\partial f(x)$ the subdifferential of f at the point x,

$$\partial f(x) := \{v \in \mathbb{R}^N : \forall y, f(y) \geq f(x) + \langle v, y - x\rangle\}$$

Note that $\partial f(x)$ is a nonempty, convex and compact set. Given $\mathcal{C} \subseteq \mathbb{R}^n$ and $\lambda > 0$, denote by

$$\lambda\mathcal{C} := \{\lambda x : x \in \mathcal{C}\}, \qquad \text{cone}(\mathcal{C}) := \{\lambda x : x \in \mathcal{C}, \lambda \geq 0\}$$

the scaling of $\mathcal{C}$ by λ, and the cone of $\mathcal{C}$, respectively. Observe that $\operatorname{cone}\mathcal{C}$ is not necessarily convex, but is convex if $\mathcal{C}$ is convex. Let $\operatorname{cvx}(\mathcal{C})$ denote the convex hull of the set $\mathcal{C}$:

$$\begin{aligned}\operatorname{cvx}(\mathcal{C}) &:= \left\{ \sum_{j=1}^{J} \alpha_j x_j : x_j \in \mathcal{C}, \alpha_j \geq 0, \sum \alpha_j = 1, J < \infty \right\} \\ &= \bigcap_{\substack{\mathcal{C}' \supseteq \mathcal{C} \\ \mathcal{C}' \text{ is convex}}} \mathcal{C}'.\end{aligned}$$

For a non-empty convex set $\mathcal{C} \subseteq \mathbb{R}^n$, denote the gauge (i.e., Minkowski functional) of $\mathcal{C}$ by $\|\cdot\|_{\mathcal{C}} : \mathbb{R}^n \to [0, \infty]$ where

$$\|x\|_{\mathcal{C}} := \inf\{\lambda > 0 : x \in \lambda\mathcal{C}\}.$$

Define the descent cone of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ by

$$T_f(x) := \operatorname{cone}\left\{z - x : z \in \mathbb{R}^N, f(z) \leq f(x)\right\}.$$

When $f = \|\cdot\|_1$ we write $T(x) := T_{\|\cdot\|_1}(x)$. By abuse of notation, we write $T_{\mathcal{C}}(x) := T_{\|\cdot\|_{\mathcal{C}}}(x)$ to refer to the descent cone of a gauge $\|\cdot\|_{\mathcal{C}}$ at a point x when $\mathcal{C}$ is a non-empty convex set.

For a nonempty set $\mathcal{C}$ and $x \in \mathbb{R}^N$, denote the distance of x to $\mathcal{C}$ by $\mathrm{d}(x, \mathcal{C}) := \inf_{w \in \mathcal{C}} \|x - w\|_2$. If $\mathcal{C}$ is also closed and convex, then there exists a unique point in $\mathcal{C}$ attaining the minimum. This point is called the projection of x onto $\mathcal{C}$ and is denoted

$$\operatorname{Proj}_{\mathcal{C}}(x) := \operatorname*{arg\,min}_{w \in \mathcal{C}} \|x - w\|_2.$$

Denote by $C^\circ := \{v \mid \forall x \in \mathcal{C}, \langle v, x\rangle \leq 0\}$ the *polar cone* of $\mathcal{C}$; and define the statistical dimension [2] of $\mathcal{C}$ by

$$\mathbf{D}(\mathcal{C}) := \mathbb{E}[\mathrm{d}(g, \mathcal{C}^\circ)^2], \quad g \sim \mathcal{N}(0, I_N).$$

Given a non-empty convex set $\mathcal{C}$ and a point $x \in \mathbb{R}^N$, consider $F_{\mathcal{C}}(x) := \{h :$

$x + h \in \mathcal{C}\} = \mathcal{C} - x$. The tangent cone is given by $T_{\mathcal{C}}(x) := \mathrm{cl}(\mathrm{cone}(F_{\mathcal{C}}(x)))$ where cl denotes the closure operation; it is the smallest closed cone containing $F_{\mathcal{C}}(x)$. With these tools, we recall a result [55] in the PD context, giving a precise characterization of the risk for (LS_τ^*).

Theorem 2.1.1 ([55, Theorem 2.1])**.** *Let $\mathcal{C}$ be a non-empty closed and convex set, let $x \in \mathcal{C}$ be an arbitrary vector and assume that $z \sim \mathcal{N}(0, I_N)$. Then*

$$\sup_{\eta>0} \frac{1}{\eta^2} \mathbb{E} \,\|\, \mathrm{Proj}_{\mathcal{C}}(x + \eta z) - x\|_2^2 = \mathbf{D}(T_{\mathcal{C}}(x)). \tag{2.1}$$

Oymak and Hassibi [55] note that $\mathbf{D}(T_{\mathcal{C}}(x)) \approx \mathrm{w}^2\left(T_{\mathcal{C}}(x) \cap B_2^N\right)$, where $\mathrm{w}(\cdot)$ denotes the Gaussian mean width (gmw). Specifically, gmw gives a near-optimal characterization of the risk for (LS_τ^*). Thus, $\mathrm{w}^2(\cdot)$ serves as an *effective dimension* for a bounded set; $\mathrm{w}^2(\cdot \cap B_2^N)$ for a cone (or a set well represented by a cone) [59, 61, 62].

Definition 1 (Gaussian mean width)**.** The Gaussian mean width (gmw) of a set $\mathcal{K} \subseteq \mathbb{R}^N$ is given by

$$\mathrm{w}(\mathcal{K}) := \mathbb{E} \sup_{x \in \mathcal{K}} \langle x, g \rangle, \quad g \sim \mathcal{N}(0, I_N).$$

A closely related quantity that may also be used to capture the effective dimension of a set is the Gaussian complexity.

Definition 2 (Gaussian complexity)**.** Let $\mathcal{K} \subseteq \mathbb{R}^N$. Define the Gaussian complexity of $\mathcal{K}$ by

$$\gamma(\mathcal{K}) := \mathbb{E} \sup_{x \in \mathcal{K}} |\langle x, g \rangle|, \qquad g \sim \mathcal{N}(0, I_N).$$

Remark. If $\mathcal{K} \subseteq \mathbb{R}^N$ is symmetric, then

$$\mathrm{w}(\mathcal{K}) = \frac{1}{2} \mathbb{E} \sup_{x \in \mathcal{K} - \mathcal{K}} \langle x, g \rangle = \gamma(\mathcal{K}).$$

2.1.1 Projection lemma

The projection lemma and extensions thereof receive a full treatment in Chapter 5. However, the usefulness of this result throughout Chapters 3 and 4 makes prescient its introduction here. An alternate form of the result below was first proved in Oymak and Hassibi [55, Lemma 15.3], and a simpler proof given in Berk et al. [9]. Given a point $z \in \mathbb{R}^n$, the lemma provides an ordering for the one-parameter family of projections $z_t := \operatorname{Proj}_{t\mathcal{K}}(z)$ as a function of $t > 0$ when $\mathcal{K}$ is a closed convex set with $0 \in \mathcal{K}$. Namely, $\|\operatorname{Proj}_{t\mathcal{K}}(z)\|_2 \leq \|\operatorname{Proj}_{u\mathcal{K}}(z)\|_2$ for $0 < t \leq u < \infty$. The result is depicted graphically in Figure 5.1a.

Lemma 2.1.2 (Projection lemma)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^n$ be a non-empty closed and convex set with $0 \in \mathcal{K}$, and fix $\lambda \geq 1$. For $z \in \mathbb{R}^n$,*

$$\|\operatorname{Proj}_{\mathcal{K}}(z)\|_2 \leq \|\operatorname{Proj}_{\lambda\mathcal{K}}(z)\|_2.$$

The following is an alternative version of Lemma 2.1.2 which quickly follows. The corollary is useful in establishing Lemma 4.6.13. Its proof is deferred to § A.1.2.

Corollary 2.1.3. *Let $\mathcal{K} \subseteq \mathbb{R}^n$ be a non-empty closed and convex set with $0 \in \mathcal{K}$ and let $\|\cdot\|_{\mathcal{K}}$ be the gauge of $\mathcal{K}$. Given $y \in \mathbb{R}^n$ define*

$$x_\alpha := \arg\min\{\|x\|_{\mathcal{K}} : \|x - y\|_2 \leq \alpha\}$$

Then $\|x_\alpha\|_2$ is non-increasing in α.

2.2 Tools from probability theory

For a full treatment of the topics herein, we refer the reader to Adler and Taylor [1], Foucart and Rauhut [35], van Handel [76], Vershynin [77]. We start by introducing subgaussian random variables, which generalize Gaussian random variables, but retain certain desirable properties, such as Gaussian-like tail decay and moment bounds.

Definition 3 (Subgaussian random variable)**.** A random variable X is called subgaussian if there exists a constant $K > 0$ such that the moment generating function of X^2 satisfies, for all λ such that $|\lambda| \leq K^{-1}$,

$$\mathbb{E} \exp(\lambda^2 X^2) \leq \exp(K^2 \lambda^2).$$

The subgaussian norm of X is defined by

$$\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E} \exp(X^2/t^2) \leq 2\}.$$

We recall Hoeffding's inequality, which characterizes how subgaussian random variables concentrate in high dimensions.

Theorem 2.2.1 (General Hoeffding's inequality [77, Theorem 2.6.3])**.** *Let X_i, $i = 1, \dots n$, be independent mean-zero subgaussian random variables and let $a \in \mathbb{R}^n$. For $t > 0$,*

$$\mathbb{P}\big(\big|\sum_{i=1}^n a_i X_i\big| \geq t\big) \leq e \cdot \exp\Big(\frac{-t^2}{C \sum_{i=1}^n a_i^2 \|X_i\|_{\psi_2}^2}\Big)$$

One may similarly define subexponential random variables.

Definition 4 (Subexponential random variable)**.** A random variable X is called subexponential if there exists a constant $K > 0$ such that the moment generating function of $|X|$ satisfies, for all λ such that $0 \leq \lambda \leq K^{-1}$,

$$\mathbb{E} \exp(\lambda |X|) \leq \exp(K\lambda).$$

The subexponential norm of X is defined by

$$\|X\|_{\psi_1} := \inf\{t > 0 : \mathbb{E} \exp(|X|/t) \leq 2\}.$$

They, too, admit a concentration inequality.

Theorem 2.2.2 (Bernstein's inequality [77, Theorem 2.8.1])**.** *Let $X_1, \dots, X_n$ be independent mean-zero subexponential random variables. Then, for $k :=$*

$\max_i \|X_i\|_{\psi_1}$, *for all* $\{a_1, \ldots, a_n\} \in \mathbb{R}^n$ *and any* $t \geq 0$,

$$\mathbb{P}\big(|\sum_{i=1}^{n} a_i X_i| \geq t\big) \leq 2\exp\Big(-C\min\Big\{\frac{t^2}{k^2\|a\|_2^2}, \frac{t}{k\|a\|_\infty}\Big\}\Big).$$

Additionally, we call $X \in \mathbb{R}^n$ a K-subgaussian random vector if $\|X\|_{\psi_2} := \sup_{a\in\mathbb{R}^n} \|\langle a, X\rangle\|_{\psi_2} \leq K$; analogously so for K-subexponential random vectors. Where it is either clear or irrelevant, we may omit observing the norm parameter and refer to a K-subgaussian random vector simply as a subgaussian random vector; likewise with a subexponential random vector. For properties and equivalent definitions of subgaussian and subexponential random variables and vectors, see Vershynin [77, Chapter 2]. Say that a matrix $A \in \mathbb{R}^{m\times N}$ is *isotropic* if $\mathbb{E}\, A_i A_i^T = I_N$ where A_i^T is the ith row of A, $i \in [m]$. Next, we introduce a piece of jargon for the sake of concision.

Definition 5 (K-subgaussian matrix)**.** Given $m, N \in \mathbb{N}$, call $A \in \mathbb{R}^{m\times N}$ a K-subgaussian matrix if A has rows A_i^T that are independent, isotropic K-subgaussian random vectors:

$$\mathbb{E}\, A_i A_i^T = I, \qquad \|A_i\|_{\psi_2} \leq K, \qquad i \in [m].$$

Further, call $\frac{1}{\sqrt{m}}A$ a normalized K-subgaussian matrix.

A core element of CS is that the measurement matrix act as an approximate isometry for the signal class of interest. We will see that (normalized) subgaussian matrices indeed behave this way for s-sparse vectors. In essence, mandating isotropy of the rows is a mathematical expression of this notion; requiring that each row have subgaussian decay ensures the matrix *usually* behaves as such in high dimensions. The idea of a matrix behaving as an approximate isometry for a particular signal class is formalized as the now classical restricted isometry property (RIP), which can be stated in terms of restricted isometry constants, δ_s. Informally, A satisfies RIP if δ_s is small for reasonably large s.

Definition 6 ([35, Definition 6.1])**.** The sth restricted isometry constant $\delta_s = \delta_s(A)$ of a matrix $A \in \mathbb{R}^{m\times N}$ is the smallest $\delta \geq 0$ such that, for all $x \in \Sigma_s^N$,

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2.$$

In Chapter 4, we crucially leverage the fact that A satisfies a RIP. An exposition on RIP and restricted isometry constants may be found in Foucart and Rauhut [35]. As some results in this thesis concern K-subgaussian matrices, we state a classical version of RIP for such matrices restricted to the set of s-sparse vectors.

Theorem 2.2.3 (RIP for subgaussian matrices [35, Theorem 9.2])**.** *Let $A \in \mathbb{R}^{m\times N}$ be a normalized K-subgaussian matrix and fix $\varepsilon > 0$. There exists a constant $C = C_K > 0$ such that the restricted isometry constant of A, δ_s, satisfies $\delta_s \leq \delta$ with probability at least $1-\varepsilon$ provided*

$$m \geq C\delta^{-2}(s\ln(eN/s) + \ln(2\varepsilon^{-1})).$$

Remark 1. Setting $\varepsilon = 2\exp(-\delta^2 m/(2C))$ yields the condition

$$m \geq 2C\delta^{-2}s\ln(eN/s)$$

which guarantees that $\delta_s \leq \delta$ with probability at least $1-2\exp(-\delta^2 m/(2C))$.

Importantly, restricted isometry may be generalized. Indeed, large classes of random matrices satisfy deviation inequalities on sets (such as the set B_1^N). We formalize this idea in § 2.2.1 and connect it with RIP in § 2.3 and § 2.3.1.

Finally, we introduce a result that characterizes the variance and tail decay of the supremum of a Gaussian process. In particular, it establishes that the supremum of a Gaussian process defined over a topological space T behaves nearly like a normal random variable. This result is necessary in the development of the results of § 4.2 and § 4.4 (specifically Propositions 4.6.8, 4.6.16 and 4.6.19). We refer the reader to Adler and Taylor [1, Theorem 2.1.1] for a proof. More broadly, for an introduction to random processes, see Adler

and Taylor [1], Vershynin [77].

Theorem 2.2.4 (Borell-TIS inequality [15, 26]). *Let T be a topological space and let $\{f_t\}_{t\in T}$ be a centred (i.e., mean-zero) Gaussian process almost surely bounded on T with*

$$\|\!|f|\!\|_T := \sup_{t\in T} f_t, \qquad\qquad \sigma_T^2 := \sup_{t\in T} \mathbb{E}\left[f_t^2\right] \tag{2.2}$$

such that $\|\!|f|\!\|_T$ is almost surely finite. Then $\mathbb{E}\,\|\!|f|\!\|_T$ and σ_T are both finite and for each $u > 0$,

$$\mathbb{P}\big(\|\!|f|\!\|_T > \mathbb{E}\,\|\!|f|\!\|_T + u\big) \leq \exp\big(-\frac{u^2}{2\sigma_T^2}\big).$$

Observe that $\|\!|f|\!\|_T$ is notation; $\|\!|\cdot|\!\|_T$ is not a norm. By symmetry, one may derive an analogous lower-tail inequality. Consequently, under the assumptions of the theorem one also has for each $u > 0$,

$$\mathbb{P}\big(\,|\|\!|f|\!\|_T - \mathbb{E}\,\|\!|f|\!\|_T| > u\big) \leq 2\exp\big(-\frac{u^2}{2\sigma_T^2}\big).$$

Finally, we state the following comparison inequality for two centred Gaussian processes.

Theorem 2.2.5 (Sudakov-Fernique inequality [77, Theorem 7.2.11]). *Let $(X_t)_{t\in T}, (Y_t)_{t\in T}$ be mean-zero Gaussian processes. Assume, for all $s, t \in T$,*

$$\mathbb{E}(X_t - X_s)^2 \leq \mathbb{E}(Y_t - Y_s)^2.$$

Then,

$$\mathbb{E}\sup_{t\in T} X_t \leq \mathbb{E}\sup_{t\in T} Y_t.$$

2.2.1 Geometric tools from probability

To establish a connection between the notions of effective dimension introduced, and some of the probabilistic notions discussed above, we state two

results controlling the deviation of a K-subgaussian matrix on a bounded set, which generalize the idea of RIP introduced in Theorem 2.2.3. These results were first proved in Liaw et al. [47], and dependence on the constant K was later improved [43]. The results are stated using the improved constant $\tilde{K} := K\sqrt{\log K}$; we refer the reader to Jeong et al. [43, Theorem 2.1] for further details.

Theorem 2.2.6 ([47, Theorem 1.1])**.** *Let $A \in \mathbb{R}^{m\times N}$ be a K-subgaussian matrix and $T \subseteq \mathbb{R}^N$ bounded. Then*

$$\mathbb{E}\sup_{x\in T}\left|\|Ax\|_2 - \sqrt{m}\|x\|_2\right| \leq C\tilde{K}\gamma(T).$$

Another version of this result holds, where the deviation is instead controlled by the gmw and radius, rather than the Gaussian complexity.

Theorem 2.2.7 ([47, Theorem 1.4])**.** *Let $A \in \mathbb{R}^{m\times N}$ be a K-subgaussian matrix and $T \subseteq \mathbb{R}^N$ bounded. For any $u \geq 0$ the event*

$$\sup_{x\in T}\left|\|Ax\|_2 - \sqrt{m}\|x\|_2\right| \leq C\tilde{K}\left[\mathrm{w}(T) + u\cdot\mathrm{rad}(T)\right] \tag{2.3}$$

holds with probability at least $1 - 3\exp(-u^2)$. Here, $\mathrm{rad}(T) := \sup_{x\in T}\|x\|_2$ denotes the radius of T.

In particular, the deviation of the image of A restricted to a set T may be controlled using the gmw.

Remark 2. If $u \geq 1$ the bound in (2.3) can be loosened to the following simpler one:

$$\sup_{x\in T}\left|\|Ax\|_2 - \sqrt{m}\|x\|_2\right| \leq C\tilde{K}u\gamma(T).$$

Setting $T := \mathbb{S}^{N-1}$, and using the improved constant obtained in Jeong et al. [43, Theorem 2.1] gives the following corollary.

Corollary 2.2.8 (Largest singular value of K-subgaussian matrices)**.** *Let $A \in \mathbb{R}^{m\times N}$ be a K-subgaussian matrix. For all $t \geq 0$, with probability at*

least $1 - 3\exp(-t^2)$,

$$\left| \|A\| - \sqrt{m} \right| \leq C\tilde{K} \left[\sqrt{N} + t \right],$$

where $\|A\| := \sup_{x \in \mathbb{S}^{N-1}} \|Ax\|_2$ *denotes the operator norm of* A.

2.3 Effective dimension of structured signals

Because the deviation of a matrix can be controlled in terms of the gmw it is important to be able to express tight upper bounds on this quantity. In this section, we introduce tools primarily relevant to obtaining recovery bounds for compressed sensing in the classical setting where $\mathcal{K} = B_1^N$. Throughout this thesis, it will be useful to make reference to the following sets. For $r > 0$, define:

$$\mathcal{L}_s(r) := r \cdot \operatorname{cvx}\left(\Sigma_s^N \cap \mathbb{S}^{N-1}\right), \qquad \mathcal{L}_s := \mathcal{L}_s(2), \qquad \mathcal{L}_s^* := \mathcal{L}_{2s}(4). \tag{2.4}$$

Further, define the sets

$$\mathcal{J}_s^N := \{x \in \mathbb{R}^N : \|x\|_1 \leq \sqrt{s}\|x\|_2\}, \tag{2.5}$$

$$\mathcal{K}_s^N := \{x \in \mathbb{R}^N : \|x\|_2 \leq 1 \quad \& \quad \|x\|_1 \leq \sqrt{s}\}. \tag{2.6}$$

Observe that $\mathcal{J}_s^N$ is a cone and that $\mathcal{K}_s^N = B_2^N \cap \sqrt{s}B_1^N = \operatorname{cvx}\left(\mathcal{J}_s^N \cap \mathbb{S}^{N-1}\right)$.

We start by recalling that sparse vectors have low effective dimension, as does their difference set.

Lemma 2.3.1 (gmw of the sparse signal set [59, Lemma 2.3])**.** *There exist absolute constants* $c, C > 0$ *such that*

$$\begin{aligned} cs\log(2N/s) &\leq \mathrm{w}^2\left((\Sigma_s^N \cap B_2^N) - (\Sigma_s^N \cap B_2^N)\right) \\ &\leq Cs\log(2N/s) \end{aligned}$$

For possibly different absolute constants $c, C > 0$*, one also has*

$$cs\log(2N/s) \leq \mathrm{w}^2\left(\Sigma_s^N \cap B_2^N\right) \leq Cs\log(2N/s).$$

The descent cone of the ℓ_1 ball has comparable effective dimension to the set of sparse vectors. For example, Foucart and Rauhut [35, Proposition 9.24] present a result establishing $\mathrm{w}^2\left(T_{B_1^N}(x) \cap \mathbb{S}^{N-1}\right) \leq 2s\log(eN/s)$ when x is s-sparse. The following result further clarifies the connection between the gmw of B_1^N and that of Σ_s^N.

Lemma 2.3.2 (Convexification [60, Lemma 3.1])**.** *With $\mathcal{K}_s^N$ as defined in (2.6), one has*

$$\mathrm{cvx}(\Sigma_s^N \cap B_2^N) \subseteq \mathcal{K}_s^N \subseteq 2\,\mathrm{cvx}(\Sigma_s^N \cap B_2^N).$$

Next, it will be useful to leverage the following equivalent characterization for the ℓ_1 descent cone. These results are effectively simplifications of classical results [35].

Lemma 2.3.3 (Equivalent ℓ_1 descent cone characterization)**.** *Let $x \in \Sigma_s^N$ with non-empty support set $T \subseteq [N]$ and define $\mathcal{C} := \|x\|_1 B_1^N$. Then $T_{\mathcal{C}}(x) = K(x)$, where*

$$K(x) := \{h \in \mathbb{R}^N : \|h_{T^C}\|_1 \leq -\langle \mathrm{sgn}(x), h\rangle\}.$$

Finally, recall the following well-known descent cone condition.

Lemma 2.3.4 (ℓ_1 descent cone condition)**.** *Let $x \in \Sigma_s^N$ have non-empty support set $T \subseteq N$ and suppose $\hat{x} \in \mathbb{R}^N$ satisfies $\|\hat{x}\|_1 \leq \|x\|_1$. Then $\|h\|_1 \leq 2\sqrt{s}\|h\|_2$, where $h = \hat{x} - x$.*

Proof of Lemma 2.3.4. We use Lemma 2.3.3 above before applying Cauchy-Schwarz:

$$\begin{aligned}
\|\hat{x} - x\|_1 &= \|h_T\|_1 + \|h_{T^C}\|_1 \\
&\leq \langle \mathrm{sgn}(\hat{x}_T - x), h_T\rangle - \langle \mathrm{sgn}(x), h\rangle \\
&\leq \|\mathrm{sgn}(\hat{x}_T - x) - \mathrm{sgn}(x)\|_2 \|h\|_2 \\
&\leq 2\sqrt{s}\|h\|_2.
\end{aligned}$$

□

Remark. Lemma 2.3.4 specializes to LASSO solutions in the following way.

Let $x \in \Sigma_s^N$ with $s > 0$. Suppose $y = Ax + \eta z$ for $\eta > 0$, $z \in \mathbb{R}^m$, and $A \in \mathbb{R}^{m \times N}$. Let $\hat{x}$ solve (LS_τ) with $\tau = \|x\|_1$. Then $\|h\|_1 \leq 2\sqrt{s}\|h\|_2$, where $h = \hat{x} - x$.

Remark 3. An analogous result holds in the PD setting. For example, we present the following version for (BP_σ^*), used in the proof of Lemma 3.4.1.

For $s \geq 0$, let $x \in \Sigma_s^N$ and suppose $y = x + \eta z$ where $\eta > 0$ and $z \in \mathbb{R}^N$ with $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Condition z on the event $\mathcal{E} := \{\|z\|_2^2 \leq N - 2\sqrt{N}\}$. Let $\tilde{x}$ solve (BP_σ^*) with $\sigma \geq \eta\sqrt{N}$. Then $\|\tilde{x}\|_1 \leq \|x\|_1$ and $\|h\|_1 \leq 2\sqrt{s}\|h\|_2$ where $h := \tilde{x} - x$.

2.3.1 Refined bounds on Gaussian mean width

Two recent results of Bellec [7] yield improved upper- and lower-bounds on the gmw of convex polytopes intersected with Euclidean balls. Each is integral to establishing (BP_σ^*) parameter sensitivity in § 3.4 and (BP_σ) parameter sensitivity in § 4.4. These results are a fine-tuning of standard gmw results for bounded convex polytopes that describe how local effective dimension of a convex hull scales with neighbourhood size.

Proposition 2.3.5 ([7, Proposition 1])**.** *Let $m \geq 1$ and $N \geq 2$. Let $\mathcal{K}$ be the convex hull of $2N$ points in $\mathbb{R}^m$ and assume $\mathcal{K} \subseteq B_2^m$. Then for $\gamma \in (0,1)$,*

$$\mathrm{w}(\mathcal{K} \cap \gamma B_2^m) \leq \min\left\{4\sqrt{\max\left\{1, \log(8eN\gamma^2)\right\}}, \gamma\sqrt{\min\{m, 2N\}}\right\}.$$

The second result shows that Proposition 2.3.5 is tight up to multiplicative constants.

Proposition 2.3.6 ([7, Proposition 2])**.** *Let $m \geq 1$ and $N \geq 2$. Let $\gamma \in (0,1]$ and assume for simplicity that $s = 1/\gamma^2$ is a positive integer such that $s \leq N/5$. Let $M \in \mathbb{R}^{m \times N}$ have columns $\{M_i\}_{i \in [N]} \subseteq \mathbb{S}^{m-1}$. Let $\mathcal{K}$ be the convex hull of $\{\pm M_i\}_{i \in [N]}$. Assume that for some real number $\kappa \in (0,1)$ we have*

$$\kappa\|\theta\|_2 \leq \|M\theta\|_2 \qquad \text{for all } \theta \in \mathbb{R}^N \text{ such that } \|\theta\|_0 \leq 2s.$$

Then,

$$\mathrm{w}(\mathcal{K} \cap \gamma B_2^m) \geq (\sqrt{2}/4)\kappa\sqrt{\log(N\gamma^2/5)}.$$

In particular, the above two results may be combined to obtain a bound on a random polytope, obtained by considering the image of a (non-random) polytope under a normalized K-subgaussian matrix; proved in § A.1.1. This corollary is central to establishing the main result of § 4.4.

Corollary 2.3.7 (Controlling random hulls)**.** *Fix $\delta, \varepsilon > 0$, $\gamma \in (0, 1]$ and let $A \in \mathbb{R}^{m \times N}$ be a normalized K-subgaussian matrix. Assume for simplicity that $s = 1/\gamma^2 \in \mathbb{N}$ with $s < N/5$ and let T denote the convex hull of the $2N$ points $\{\pm A^j : j \in [N]\}$. Assume $m > C_\varepsilon \delta^{-2} \tilde{K}^2 s \log(2N/s)$. With probability at least $1 - \varepsilon$, for any $\alpha \in (0, (1-\delta))$,*

$$\begin{aligned}
&(\sqrt{2}/4)(1-\delta)^2\sqrt{\log \frac{N\alpha^2}{5(1-\delta)^2}} \\
&\leq \mathrm{w}(T \cap \alpha B_2^m) \\
&\leq \min\left\{4(1+\delta)\sqrt{\max\left\{1, \log \frac{8eN\alpha^2}{(1+\delta)^2}\right\}}, \alpha\sqrt{\min\{m, 2N\}}\right\}.
\end{aligned}$$

Chapter 3

Proximal denoising parameter sensitivity

3.1 Overview

We provide an overview of this chapter by presenting three sibling results that are simplifications of the main results presented in Chapter 3. In particular, these simplified results are asymptotic versions of the main results to follow. Throughout this chapter, "loss" refers to the noise-normalized squared error (nnse) of a PD estimator, and "risk" to the expected nnse of a PD estimator, unless otherwise noted. The losses and risks for CS estimators were given in (1.1)–(1.3) and (1.4)–(1.6), respectively. In the PD setting, we re-state the definition of the risks to incorporate one significant change in the definition of the risk associated to (QP^*_λ), and a minor notational change for all three:

$$\hat{R}(\tau; x_0, N, \eta) := \eta^{-2}\,\mathbb{E}\,\|\hat{x}(\tau; y) - x_0\|_2^2,$$

$$R^\sharp(\lambda; x_0, N, \eta) := \eta^{-2}\,\mathbb{E}\,\|x^\sharp(\eta\lambda; y) - x_0\|_2^2,$$

$$\tilde{R}(\sigma; x_0, N, \eta) := \eta^{-2}\,\mathbb{E}\,\|\tilde{x}(\sigma; y) - x_0\|_2^2.$$

The choice to use $x^\sharp(\eta\lambda)$ in the definition of $R^\sharp$, rather than $x^\sharp(\lambda)$, clarifies the presentation of the results in § 3.3. One could have similarly defined $\tilde{R}$

using $\tilde{x}(\eta\sigma)$, which would result in an appropriate change of normalization for the results of § 3.4.

We denote the optimally tuned worst-case risk for (LS_τ^*) by $R^*(s,N)$. This quantity is given by:

$$R^*(s,N) := \sup_{x_0 \in \Sigma_s^N} \hat{R}(\|x_0\|_1; x_0, N, \eta) = \max_{\substack{x_0 \in \Sigma_s^N \\ \|x_0\|_1 = 1}} \lim_{\eta \to 0} \hat{R}(1; x_0, N, \eta).$$

A proof of the second equality appears in Proposition 3.6.2. We use $R^*(s,N)$ as a benchmark, noting it is order-optimal in Proposition 3.1.5.

In § 3.2, we show that (LS_τ^*) exhibits an asymptotic phase transition in the low-noise regime. There is exactly one value τ^* of the governing parameter yielding minimax order-optimal error, with any choice $\tau \neq \tau^*$ yielding markedly worse behaviour. The intuition for this result is that (LS_τ^*) is extremely sensitive to the value of τ in the low-noise regime, making empirical use of (LS_τ^*) woefully unstable in this regime.

Theorem 3.1.1.

$$\lim_{N \to \infty} \max_{\substack{x_0 \in \Sigma_s^N \\ \|x_0\|_1 = 1}} \lim_{\eta \to 0} \frac{\hat{R}(\tau; x_0, N, \eta)}{R^*(s,N)} = \begin{cases} \infty & \tau < \tau^* \\ 1 & \tau = \tau^* = 1 \\ \infty & \tau > \tau^* \end{cases}$$

Next, in § 3.3, we show that (QP_λ^*) exhibits an asymptotic phase transition. The worst-case risk over $x_0 \in \Sigma_s^N$ is minimized for parameter choice $\lambda^* = O(\sqrt{\log(N/s)})$ [55]. While λ^* has no closed form expression, it satisfies $\lambda^*/\sqrt{2\log(N)} \xrightarrow{N \to \infty} 1$ for s fixed (Proposition 3.3.5). Thus, we consider the normalized parameter $\mu = \lambda/\sqrt{2\log(N)}$. The risk $R^\sharp(\lambda; x_0, N, \eta)$ is minimax order-optimal when $\mu > 1$ and suboptimal for $\mu < 1$.

Theorem 3.1.2. *Let $\lambda(\mu, N) := \mu\sqrt{2\log N}$ for $\mu > 0$. Then,*

$$\lim_{N \to \infty} \sup_{x_0 \in \Sigma_s^N} \frac{R^\sharp(\lambda(\mu,N); x_0, N, \eta)}{R^*(s,N)} = \begin{cases} \mathcal{O}(\mu^2) & \mu \geq 1 \\ \infty & \mu < 1 \end{cases}$$

Lastly, we show in § 3.4 that (BP_σ^*) is poorly behaved for all $\sigma > 0$ when x_0 is very sparse. Namely, $\tilde{R}(\sigma; x_0, N, \eta)$ is asymptotically suboptimal for *any* $\sigma > 0$ when s/N is sufficiently small.

Theorem 3.1.3.

$$\lim_{N\to\infty} \sup_{x_0\in\Sigma_s^N} \inf_{\sigma>0} \frac{\tilde{R}(\sigma; x_0, N, \eta)}{R^*(s, N)} = \infty$$

All numerical results are discussed in § 3.5, and proofs of most theoretical results are deferred to § 3.6. Next, we add two clarifications. First, the three PD programs are equivalent in a sense.

Proposition 3.1.4. *Let* $0 \neq x_0 \in \mathbb{R}^N$ *and* $\lambda > 0$. *Where* $x^\sharp(\lambda)$ *solves* (QP_λ^*), *define* $\tau := \|x^\sharp(\lambda)\|_1$ *and* $\sigma := \|y - x^\sharp(\lambda)\|_2$. *Then* $x^\sharp(\lambda)$ *solves* (LS_τ^*) *and* (BP_σ^*).

However, τ and σ have *stochastic* dependence on z, and this mapping may not be smooth. Thus, parameter stability of one program is not implied by that of another. Second, $R^*(s, N)$ has the desirable property that it is computable up to multiplicative constants. The proof follows by Oymak and Hassibi [55] and standard bounds in Foucart and Rauhut [35]. We don't claim novelty for this result, and defer its full proof to § 3.6.1.

Proposition 3.1.5. *Let* $s \geq 1, N \geq 2$ *be integers, let* $\eta > 0$ *and suppose* $y = x_0 + \eta z$ *for* $z \in \mathbb{R}^N$ *with* $z_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. *Let*

$$M^*(s, N) := \inf_{x_*} \sup_{x_0\in\Sigma_s^N} \eta^{-2}\|x_* - x_0\|_2^2$$

be the minimax risk over arbitrary estimators $x_* = x_*(y)$. *There is* $c, C_1, C_2 > 0$ *such that for* $N \geq N_0 = N_0(s)$, *with* $N_0 \geq 2$ *sufficiently large,*

$$cs\log(N/s) \leq M^*(s, N) \leq \inf_{\lambda>0} \sup_{x_0\in\Sigma_s^N} R^\sharp(\lambda; x_0, N, \eta)$$
$$\leq C_1 R^*(s, N) \leq C_2 s\log(N/s).$$

Thus, in the simplified theorems above, we could have normalized by any of the quantities in the conclusion of Proposition 3.1.5 instead of $R^*(s, N)$, because these expressions are asymptotically equivalent up to constants. A consequence of Theorem 3.1.3 offers a stark contrast to Proposition 3.1.5:

$$\inf_{\sigma>0} \sup_{x_0 \in \Sigma_s^N} \tilde{R}(\sigma; x_0, N, \eta) \geq \sup_{x_0 \in \Sigma_s^N} \inf_{\sigma>0} \tilde{R}(\sigma; x_0, N, \eta) \gg R^*(s, N).$$

In particular, removing the parameters' noise dependence destroys the equivalence attained in Proposition 3.1.4.

3.2 On parameter sensitivity for (LS_τ^*)

We describe a PS regime for (LS_τ^*), revealing a regime in which there is exactly one choice of parameter $\tau^* > 0$ such that $\hat{R}(\tau^*; x_0, N, \eta)$ is minimax order-optimal. Specifically, Theorem 3.2.1 shows that $\hat{R}(\tau; x_0, N, \eta)$ exhibits an asymptotic singularity in the limiting low-noise regime (by low-noise regime, we mean hereafter the regime in which $\eta \to 0$).

In § 3.5.1 we complement this asymptotic result with numerical simulations that contrast how the three risks behave in a simplified experimental context. The numerics support that Theorem 3.2.1 provides accurate intuition to guide how (LS_τ^*) can be expected to perform in practice when the noise level is small relative to the magnitude of the signal's entries.

The analogue of the classical CS result is included in our result as the special case $\tau = \tau^* = \|x_0\|_1$ (*cf.* Proposition 3.1.5). The cases for $\tau \neq \tau^*$ may seem surprising initially, but can be understood with the following key intuition: the approximation error is controlled by the effective dimension of the constraint set.

First, one should generally not expect good recovery when the signal lies outside the constraint set. When $\tau < \tau^*$, y lies outside of the constraint set with high probability in the limiting low-noise regime. Accordingly, there is a positive distance between the true signal and the recovered signal which may be lower-bounded by a dimension-independent constant. Hence, the risk is determined by the reciprocal of the noise variance, growing unboundedly as

$\eta \to 0$.

On the other hand, when $\tau > \tau^*$, y lies within the constraint set with high probability in the limiting low-noise regime. Thus, the problem is essentially unconstrained in this setting, so the effective dimension of the constraint set for the problem should be considered equal to that of the ambient dimension. In particular, one should expect that the error be proportional to N.

Theorem 3.2.1 ((LS_τ^*) parameter sensitivity)**.** *Let $s \geq 1, \eta > 0$ and let $x_0 \in \Sigma_s^N \setminus \Sigma_{s-1}^N$. Given $\tau > 0$,*

$$\lim_{\eta \to 0} \hat{R}(\tau; x_0, N, \eta) = \begin{cases} \infty & \tau < \|x_0\|_1 \\ R^*(s, N) & \tau = \|x_0\|_1 \\ N & \tau > \|x_0\|_1 \end{cases}$$

In summary, the surprising part of this result is that there is a sharp phase transition between two "unstable" regimes, with the optimal regime lying on the boundary of the two phases. We argue this suggests that there is only one reasonable choice for τ in the low-noise regime. Observe, that Theorem 3.2.1 connects with Theorem 3.1.1 by taking the limit of the problem as $N \to \infty$ after first restricting to signals of a finite norm (arbitrarily, 1) so that the essence of the result is preserved.

3.3 On parameter sensitivity for (QP_λ^*)

We show that $R^\sharp(\lambda; x_0, N, \eta)$ is smooth in the low-noise regime. This result becomes evident from the closed-form expression for $R^\sharp(\lambda; s, N)$ that emerges for this special case. At first, this smoothness result seems to stand in contrast to the "cusp-like" behaviour that we observe analytically and numerically for $\lim_{\eta \to 0} \hat{R}(\tau; x_0, N, \eta)$ (*cf.* Figure 3.2). However, $R^\sharp(\lambda; s, N)$ possesses unfavourable dependence on N that is elucidated in Theorem 3.3.4.

Briefly, if the governing parameter λ is too small, then the risk grows unboundedly as a power law of N in high dimensions. This rate of growth implies that the risk is minimax suboptimal for such λ. To our knowledge, the establishment of this result in Berk et al. [9] was novel. In contrast,

for all suitably large λ, $R^\sharp(\lambda; s, N)$ admits the desirable property suggested in Chapter 1: $R^\sharp(\lambda; s, N) \lesssim (\lambda/\lambda^*)^2 R^*(s, N)$. The result, stated in Theorem 3.3.6, essentially follows from known LASSO bounds for RIP matrices: $R(\lambda) \leq \lambda^2 s$. Thus, in the low-noise regime, $R^\sharp(\lambda; x_0, N, \eta)$ exhibits a phase transition between order-optimal and suboptimal regimes.

The numerics of § 3.5.2 suggest a viable constant for the growth rate of the risk when λ is too small, and support Theorem 3.3.6 in the case where λ is sufficiently large. These numerics also clarify the role that the dimension-dependent growth rate serves in the stability of (QP^*_λ) about λ^*.

3.3.1 Smoothness of the risk

The (QP^*_λ) estimator for a problem with noise level $\eta > 0$ and with parameter $\lambda > 0$ is given by soft-thresholding by $\eta\lambda$. In particular, $x^\sharp(\eta\lambda)$ is a smooth function with respect to the problem parameters, hence so is $R^\sharp(\lambda; x_0, N, \eta)$ (being a composition of smooth functions). However, the closed form expression for $R^\sharp(\lambda; x_0, N, \eta)$ is unavailable, because the expectations involved are untractable in general. When the noise-level vanishes this is no longer true and we may compute an exact expression in terms of λ, s and N for the risk. Specifically, we note that the smoothness result below is not special to the case where $\eta \to 0$, but is notable because of the closed form expression for the risk that is obtained.

Moreover, the result is notable, because the closed form expression is equivalent (in some precisely definable sense) to $R^\sharp(\lambda; x_0, N, \eta)$ when $\eta > 0$ and the magnitudes of the entries of x_0 are all large (i.e., "the signal is well-separated from the noise"). We make this connection after the main results discussed below. In turn, this connects Theorem 3.3.4 and Theorem 3.3.6 to Theorem 3.1.2, where the analytic expression is used to derive the so-called left-sided PS and right-sided parameter stability results.

Proposition 3.3.1 ($R^\sharp(\lambda; x_0, N, \eta)$ smoothness)**.** *Let $s \geq 0, N \geq 1, x_0 \in \Sigma_s^N$ and $\lambda > 0$. Then, where ϕ and Φ are the normal pdf and cdf,*

$$\lim_{\eta \to 0} R^\sharp(\lambda; x_0, N, \eta) = s(1 + \lambda^2) + 2(N - s)\big[(1 + \lambda^2)\Phi(-\lambda) - \lambda\phi(\lambda)\big] \quad (3.1)$$

Remark 4. Here and beyond, we denote the limiting low-noise risk by

$$R^\sharp(\lambda; s, N) := \lim_{\eta \to 0} R^\sharp(\lambda; x_0, N, \eta),$$

and define the function $G(\lambda) := (1+\lambda^2)\Phi(-\lambda) - \lambda\phi(\lambda)$ for notational brevity.

An equivalence in behaviour is seen between the low-noise regime $\eta \to 0$ and the large-entry regime $|x_{0,j}| \to \infty$ for $j \in \mathrm{supp}(x_0)$ with $\eta > 0$. For both programs, the noise level is "effectively" zero by comparison to the size of the entries of x_0. This type of scale invariance allows us to re-state the previous result as a max formulation.

Corollary 3.3.2 (max-formulation)**.** *Let $s \geq 0, N \geq 1, x_0 \in \Sigma_s^N$ and $\eta > 0$. For $\lambda > 0$,*

$$\sup_{x_0 \in \Sigma_s^N} R^\sharp(\lambda; x_0, N, \eta) = R^\sharp(\lambda; s, N)$$

3.3.2 Left-sided parameter sensitivity

We reveal an asymptotic regime in which $R^\sharp(\lambda; s, N)$ is minimax suboptimal for all λ sufficiently small. The result follows from showing the risk derivative is large for all $\lambda < \bar{\lambda}$ when s is sufficiently small relative to N. Here, $\bar{\lambda} := \sqrt{2 \log N}$ is an Ansatz estimate of λ^* used to make the proof proceed cleanly. Finally, we show in what sense $\bar{\lambda}$ is asymptotically equivalent to λ^* in Proposition 3.3.5.

The proof for the bound on the risk derivative follows by calculus and a standard estimate of $\Phi(-\lambda)$ in terms of $\phi(\lambda)$. Its scaling with respect to the ambient dimension destroys the optimal behaviour of $R^\sharp(\lambda; x_0, N)$ for all $\lambda < \bar{\lambda}$. The proof of this result, stated in Theorem 3.3.4, follows immediately from Lemma 3.3.3 by the fundamental theorem of calculus.

Lemma 3.3.3 (risk derivative growth)**.** *Fix $s \geq 1$. For any $\varepsilon \in (0, 1)$, there*

exists $C > 0$ *and an integer* $N_0 = N_0(s) \geq s$ *so that for all* $N \geq N_0$

$$-\frac{\mathrm{d}}{\mathrm{d}u}\bigg|_{u=1-\varepsilon} R^\sharp(u\bar{\lambda}; s, N) \geq CN^\varepsilon$$

where $\bar{\lambda} = \sqrt{2\log(N)}$ *is an estimate of the optimal parameter choice for* (QP^*_λ).

Theorem 3.3.4 ((QP^*_λ) parameter sensitivity)**.** *Under the conditions of the previous lemma, for* $\varepsilon \in (0,1)$ *there exists a constant* $C > 0$ *and integer* $N_0 \geq 1$ *such that for all* $N \geq N_0$*,*

$$R^\sharp((1-\varepsilon)\bar{\lambda}; s, N) \geq C\frac{N^\varepsilon}{\log N}.$$

Though these results may initially seem surprising, we claim they are sensible when viewed in comparison to unregularized proximal denoising (i.e., $\lambda = 0$). In this case, sparsity of the signal x_0 is unused and so one expects the error to be proportional to the ambient dimension, as in § 3.2. In the low-noise regime, the sensitivity of the program to λ is apparently amplified, and for $\lambda > 0$ one may still expect (QP^*_λ) to behave similarly to unregularized proximal denoising, begetting risk that behaves like a power law of N.

Proposition 3.3.5 (Asymptotic equivalence)**.** *Let* $N \in \mathbb{N}$ *with* $N \geq 2$*,* $s \in [N]$ *and* $\bar{\lambda} = \sqrt{2\log N}$*. For given problem data, suppose* $x^\sharp(\lambda)$ *solves* (QP^*_λ)*, and let* λ^* *be the optimal parameter choice for* $R^\sharp(\lambda; s, N)$*. Then*

$$\lim_{N\to\infty} \frac{\bar{\lambda}}{\lambda^*} = 1$$

Remark 5. The value $\bar{\lambda}$ estimates the optimal parameter choice for (QP^*_λ) in the following sense as $N \to \infty$ with $s = o(N)$ [55].

$$\lambda^* = \mathcal{O}(\sqrt{\log(N/s)}) \approx \sqrt{2\log N} =: \bar{\lambda}$$

3.3.3 Right-sided parameter stability

In the low-noise regime, $R^\sharp$ may still be order-optimal if λ is chosen large enough. Specifically, if $\lambda = L\lambda^*$ for some $L > 1$, then $R^\sharp(\lambda; x_0, N)$ is still minimax order-optimal. We claim no novelty for the result of this section (for example, this result is a consequence of [55, Theorem 2.2]), but use it as a contrast to elucidate the previous theorem. Whereas for $\lambda < \bar{\lambda}$ we are penalized for under-regularizing in the low-noise regime in high dimensions, the theorem below implies that we are not penalized for over-regularizing.

Theorem 3.3.6. (QP^*_λ) *is parameter stable in the sense that for any* $\lambda > 0$ *satisfying* $L = \lambda/\lambda^* > 1$*, there is* $N_0 = N_0(s, \lambda) \geq 2$ *so that for all* $N \geq N_0$*,*

$$\frac{R^\sharp(\lambda; s, N)}{R^*(s, N)} \leq CL^2.$$

Observe that the theorem still holds in the event that λ^* is replaced by $\bar{\lambda}$. Thus, one may obtain the exact point of the phase transition, $\bar{\lambda}$, observed in Theorem 3.1.2. In fact, with this note, Theorem 3.1.2 follows as a direct consequence of the results of this section by letting $N \to \infty$.

3.4 On parameter sensitivity for (BP^*_σ)

The program (BP^*_σ) is maximin suboptimal for very sparse vectors x_0. We show that $\tilde{R}(\sigma; x_0, N, \eta)$ scales as a power law of N for all $\sigma > 0$. This rate is significantly worse than $R^*(s, N)$. When x_0 is very sparse and (BP^*_σ) is underconstrained, then $\sigma \geq \eta N$ and § 3.4.1 proves that $\tilde{R}(\sigma; x_0, N, \eta) = \Omega(\sqrt{N})$. When (BP^*_σ) is overconstrained, then $\sigma \leq \eta\sqrt{N}$ and § 3.4.2 proves that $\tilde{R}(\sigma; x_0, N, \eta) = \Omega(N^q)$ for some $q > 0$ when x_0 is very sparse.

Intuitively, (BP^*_σ) kills not only the noise, but also eliminates too much of the signal content when underconstrained and s is small compared to N. Because the signal is very sparse, destroying the signal content is disastrous to the risk. When overconstrained, the remaining noise overwhelms the risk, because the off-support has size approximately equal to the ambient dimension.

The above two steps are combined in Theorem 3.4.6 as a minimax formulation over all $\sigma > 0$ and $x_0 \in \Sigma_s^N$. In Theorem 3.4.9, this result is strengthened to a maximin statement over $x_0 \in \Sigma_s^N$ and all $\sigma > 0$.

Although these results may seem to run contrary to the apparent efficacy of the CS analogue of (BP_σ^*) in empirical settings, we assure the reader that they are consistent. The type of PS described in this section occurs at very large dimensions, in the setting where $s \geq 1$ is fixed. Thus, although these results bode poorly for the ability of (BP_σ^*) to recover even the 0 vector (arguably a desirable property of a denoising program), many structured high-dimensional signals observed in practice are not so sparse [in a basis] as to belong to the present regime. Nevertheless, this result serves as a caveat for the limits of a popular ℓ_1 convex program.

3.4.1 Underconstrained (BP_σ^*)

The proof of this result uses standard methods from CS and may be found in § 3.6.4.

Lemma 3.4.1. *Let $s \geq 1$ and let $x_0 \in \Sigma_s^N \setminus \Sigma_{s-1}^N$ be an exactly s-sparse signal with $|x_j| \gtrsim N$ for all $j \in \mathrm{supp}(x_0)$. If $\sigma > \eta\sqrt{N}$, then there exists a constant $C > 0$ and integer $N_0 = N_0(s) \geq 2$ such that if $N \geq N_0$ then*

$$\tilde{R}(\sigma; x_0, N, \eta) \geq C\sqrt{N}.$$

3.4.2 Overconstrained (BP_σ^*)

The proof that $\tilde{R}(\sigma; x_0, N, \lambda)$ scales as a power law of N when $\sigma \leq \eta\sqrt{N}$ proceeds by an involved argument, hinging on two major steps. The first step is to find an event whose probability is lower-bounded by a universal constant, on which (BP_σ^*) fails to recover the 0 vector when $\sigma = \eta\sqrt{N}$. Then, Lemma 2.1.2 extends this result to all $\sigma \leq \eta\sqrt{N}$. At this point, one may obtain the minimax result of Theorem 3.4.6, as well as a partial maximin result for all $x_0 \in \Sigma_s^N$ on the restriction to $\sigma \leq \eta\sqrt{N}$. Then, to strengthen these claims to a maximin result over all $\sigma > 0$, we prove a lemma that

leverages elementary properties from convex analysis to show how the error of an estimator may be controlled by that of a lower dimensional estimator from the same class.

In this section, we state key results for building intuition and defer technical results and proofs to § 3.6.4.

Theorem 3.4.2 (Overconstrained Maximin)**.** *There exist universal constants $C > 0, q \in (0, \frac{1}{2})$ and integer $N_0 \geq 2$ such that for all $N \geq N_0, s \geq 0$ and $\eta > 0$,*

$$\sup_{x_0 \in \Sigma_s^N} \inf_{\sigma \leq \eta\sqrt{N}} \tilde{R}(\sigma; x_0, N, \eta) \geq CN^q.$$

By scaling, it is sufficient to prove this result in the case where $\eta = 1$. The discussion below thus assumes $y = x_0 + z$, while results are stated in full generality. The main result relies on proving

$$\inf_{\sigma \leq \sqrt{N}} \tilde{R}(\sigma; x_0, N, 1) \geq CN^q$$

when $x_0 \equiv 0$, trivially implying the equation before it. Thus, the problem now becomes that of recovering the 0 vector from standard normally distributed noise:

$$\tilde{x}(\sigma) = \arg\min\{\|x\|_1 : \|x - z\|_2^2 \leq \sigma^2\}.$$

Here and below, we denote the feasible set in (BP_σ^*) by $F(z; \sigma) = B_2^N(z; \sigma)$ and use the notation $F := F(z; \sqrt{N})$. For $\lambda > 0$ and $0 < \alpha_2 \leq \alpha_1 < \infty$, define $K_i = \lambda B_1^N \cap \alpha_i B_2^N$ to be the intersection of the ℓ_1-ball scaled by λ with the ℓ_2-ball scaled by α_i for $i = 1, 2$.

With $\sigma = \sqrt{N}$, we prove a geometric lemma. A pictorial representation of this lemma appears in Figure 3.1, in which we have represented λB_1^N using Milman's 2D representation of high-dimensional ℓ_1 balls to facilitate the intuition for how they behave in the present context. The key to the proof of Theorem 3.4.2 is the geometric lemma below, Lemma 3.4.3. It proves there exists an ℓ_1 ball of radius λ that intersects the feasible set,

hence a solution $\tilde{x}(\sigma)$ must satisfy $\|\tilde{x}(\sigma)\|_1 \leq \lambda$. Further, it shows that any vector in the ball λB_1^N which has small Euclidean norm does not intersect the feasible set. Thus, the solution must have large Euclidean norm.

Finally, this geometric lemma verifies that the previous three conditions hold simultaneously with probability at least $k_3 > 0$. As an immediate consequence, this lemma yields a lower risk bound, Corollary 3.4.4.

Lemma 3.4.3 (Geometric lemma)**.** *Let K_1, K_2, F be defined as above. For a particular choice of α_1, there is a positive integer $N_0^{(3.4.3)}$, strictly positive universal constants $k_3 = k_3(N_0), C_3, q$, and an event*

$$\mathcal{E} := \{K_1 \cap F \neq \emptyset\} \cap \{K_2 \cap F = \emptyset\},$$

such that for any $N \geq N_0^{(3.4.3)}$, it holds that $\mathbb{P}(\mathcal{E}) > k_3$ and $\alpha_2 > C_3 N^q$.

Corollary 3.4.4. *Fix $\eta > 0$. There are universal constants $C, q > 0$ such that for all $N \geq N_0^{(3.4.3)}$,*

$$\tilde{R}(\eta\sqrt{N}; 0, N, \eta) \geq CN^q.$$

Note that $N_0^{(3.4.3)}$ may be selected as the maximum of the N_0 values determined by the technical results Proposition 3.6.8 and Proposition 3.6.11 of § 3.6.4.

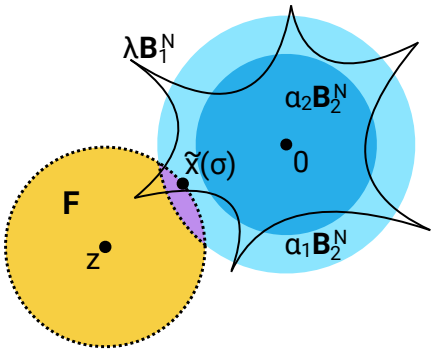


Figure 3.1: A visualization of the geometric lemma. We use Milman's 2D representation of high-dimensional ℓ_1 balls to facilitate intuition. In this setting, $\tilde{x}(\sigma)$ lies within λB_1^N. On the lemma's event $\mathcal{E}$, one has simultaneously $K_1 \cap F \neq \emptyset$ and $K_2 \cap F = \emptyset$.

Next we extend Corollary 3.4.4 from the case where $\sigma = \sqrt{N}$ to any positive $\sigma \leq \sqrt{N}$. The proof of this result follows near immediately from the projection lemma in Lemma 2.1.2. Thus, one finds $\tilde{x}(\sigma)$ has Euclidean norm at least as large as $\tilde{x}(\sqrt{N})$ when $\tilde{x}(\sigma)$ is an estimator of the 0 vector.

Lemma 3.4.5. *Let $0 < \sigma_1 < \sigma_0 = \sqrt{N}$ and $x_0 \equiv 0$. Define $\tilde{x}(\sigma_0), \tilde{x}(\sigma_1)$ as in (BP_σ^*) for $\sigma = \sigma_0, \sigma_1$, respectively. Then $\|\tilde{x}(\sigma_1)\|_2^2 \geq \|\tilde{x}(\sigma_0)\|_2^2$. Moreover, for $N \geq 2$,*

$$\mathbb{E}\,\|\tilde{x}(\sigma_1)\|_2^2 \geq \mathbb{E}\,\|\tilde{x}(\sigma_0)\|_2^2.$$

3.4.3 Minimax results

We now have the tools to state a minimax sensitivity result for (BP_σ^*). Informally, the best worst-case risk scales as a power law of N in the very sparse regime. In particular, for s fixed and N sufficiently large, there is no choice of $\sigma > 0$ yielding order-optimal risk for its corresponding worst-case signal.

Theorem 3.4.6 (Minimax Suboptimality)**.** *There are universal constants $C > 0, q \in (0, \frac{1}{2}], N_0 \geq 2$ such that for all $N \geq N_0, \eta \geq 0$ and $s \geq 1$,*

$$\inf_{\sigma > 0} \sup_{x \in \Sigma_s^N} \tilde{R}(\sigma; x, N, \eta) \geq C N^q$$

3.4.4 Maximin results

The final result of this section establishes maximin PS for all $x_0 \in \Sigma_s^N$ and $\sigma > 0$. To do this, we must show there exists a choice of signal $x_0 \in \mathbb{R}^N$ admitting no choice of $\sigma > 0$ bestowing order optimal recovery error. To this end, we will demonstrate that the previous overconstrained sensitivity results extend to s-sparse signals with $s \geq 1$. This will be enough to yield a choice of x_0 whose recovery is suboptimal over the whole parameter range.

Lemma 3.4.7 (Overconstrained (BP_σ^*), $s \geq 1$)**.** *Let $x_0 \in \Sigma_s^N$ with $\mathrm{supp}(x_0) \subseteq T \subseteq [N]$, let $y = x_0 + \xi$ for some $\xi \in \mathbb{R}^N$, let $x_1 := 0 \in \mathbb{R}^{N-s}$, and fix $\sigma > 0$. Let $\tilde{x} = \tilde{x}(\sigma) \in \mathbb{R}^N$ be the solution of (BP_σ^*) where x_0 is the ground truth,*

and let $\tilde{x}' = \tilde{x}'(\sigma) \in \mathbb{R}^{N-s}$ *be the solution of* (BP_σ^*) *where* x_1 *is the ground truth. Then*

$$\|\tilde{x}_{T^C}\|_2 \geq \|\tilde{x}'\|_2.$$

An immediate consequence of this result is the following inequality between the Euclidean norms of the error vectors.

Corollary 3.4.8. *Let* $h := \tilde{x} - x_0$ *and* $h' := \tilde{x}' - x_1$*, where* $x_0, x_1, \tilde{x}, \tilde{x}'$ *are defined as above. Then,*

$$\|h\|_2 \geq \|h'\|_2.$$

Remark 6. The above corollary is not yet sufficient to imply the desired maximin result below. As per the lemma, if $N - s \geq N_0^{(3.4.3)}$ then $\tilde{x}'$ is parameter unstable for $\sigma \leq \sqrt{N-s}$ and hence so is $\tilde{x}$. The fix for this slight mismatch is trivial, but technical. The result can be extended to the range $\sigma \leq \sqrt{N}$ by adjusting the constants in the proof of Lemma 3.4.3 and its constituents, leveraging the fact that $(N-s)/N \to 1$ as $N \to \infty$ and reselecting $N_0^{(3.4.3)}$ if necessary. We omit the details of this technical exercise.

We proceed under the assumption that the constants have been tuned to allow for $\tilde{x}'$ PS to imply $\tilde{x}$ PS for all $\sigma \leq \sqrt{N}$. Thus equipped, we state the following maximin PS result for (BP_σ^*). The proof of this result proceeds by finding a signal $x_0 \in \Sigma_s^N$ such that $\tilde{R}(\sigma; x_0, N, \eta)$ is suboptimal for all $\sigma > 0$. Since Lemma 3.4.1 applies only to signals x_0 with at least one non-zero entry, one shows there exists such a signal which simultaneously admits poor risk for $\sigma \leq \eta\sqrt{N}$ and $\sigma \geq \eta\sqrt{N}$. For example, it is enough to take $x_0 := Ne_1$ where $e_1 \in \mathbb{R}^N$ is the first standard basis vector.

Theorem 3.4.9 ((BP_σ^*) maximin suboptimality)**.** *There are universal constants* $C > 0, q \in (0, \frac{1}{2}]$ *and* $N_0 \geq 1$ *such that for all* $N \geq N_0$

$$\sup_{x_0 \in \Sigma_s^N} \inf_{\sigma > 0} \tilde{R}(\sigma; x_0, N, \eta) \geq CN^q.$$

Remark 7. The current result is given in a maximin framework. This framework is stronger than the minimax one in which these types of results are typically framed. In essence, the maximin framework assumes that the minimizer has knowledge about the ground truth signal x_0. Even still, it is not possible to choose σ to achieve order-optimal risk.

3.5 Numerical Results

Let $\mathfrak{P} \in \{(\mathrm{LS}^*_\tau), (\mathrm{QP}^*_\lambda), (\mathrm{BP}^*_\sigma)\}$ be a PD program with solution $x^*(\upsilon)$, where $\upsilon \in \{\tau, \lambda, \sigma\}$ is the associated parameter. Given a signal $x_0 \in \mathbb{R}^N$ and noise $\eta z \in \mathbb{R}^N$, denote by $\mathscr{L}(\upsilon; x_0, N, \eta z)$ the loss associated to $\mathfrak{P}$. For instance, if $\mathfrak{P} = (\mathrm{LS}^*_\tau)$, then $\mathscr{L} = \hat{L}$. In most cases, the signal x_0 for our numerical simulations will be s-sparse and s will be "small". For simplicity, and to ensure adequate separation of the "signal" from the "noise", each non-zero entry of x_0 will be equal to N, except where otherwise noted.

Define $\upsilon^* := \upsilon^*(x_0, \eta) > 0$ to be the population minimizer of $\mathscr{L}(\cdot; x_0, N, \eta z)$ — that is, the value υ where $\mathbb{E}_z \mathscr{L}(\cdot; x_0, N, \eta z)$ is minimal. Let the normalized parameter ρ for the problem $\mathfrak{P}$ be given by $\rho := \upsilon/\upsilon^*$ and let $L(\rho) := \mathscr{L}(\rho\upsilon^*)$ denote the loss for $\mathfrak{P}$ as a function of the normalized parameter. Note that $\rho = 1$ corresponds to a population minimizer of $\mathscr{L}$ *via* the re-parametrization. For $\{\rho_i\}_{i=1}^n$ being a sequence of points in the normalized parameter space, define the average loss for $\mathfrak{P}$ at any ρ_i by

$$\bar{L}(\rho_i; x_0, N, \eta, k) := k^{-1} \sum_{j=1}^{k} L(\rho_i; x_0, N, \eta \hat{z}_{ij}), \tag{3.2}$$

where $\hat{z}_{ij}$ is the (i,j)-th realization of noise; $\hat{z}_{ij} \sim \mathcal{N}(0, I_N)$ for all $(i,j) \in [n] \times [k]$. We may refer to L as the nnse. We may also refer to $\bar{L}$ as the empirical risk, average noise-normalized squared error (nnse) or simply the average loss. Note that $\bar{L}$ depends on $(\hat{z}_{ij} : i \in [n], j \in [k])$ and that notating this dependence is omitted for simplicity. Below, $\hat{z}_{ij}$ are not necessarily sampled independently. In fact, to obtain tractable computational simulations, we will frequently have $\hat{z}_{ij} = \hat{z}_{i'j}$ for $i, i' \in [n]$. Where necessary, we disambiguate the average losses with a subscript: $\bar{L}_{(\mathrm{LS}^*_\tau)}$, $\bar{L}_{(\mathrm{QP}^*_\lambda)}$, and $\bar{L}_{(\mathrm{BP}^*_\sigma)}$

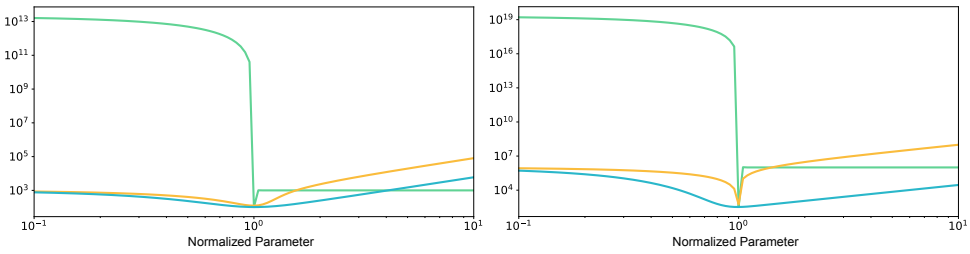
for the programs (LS_τ^*), (QP_λ^*) and (BP_σ^*), respectively.

For example visualizations of the average loss for each program, we refer the reader to Figures 3.2a, 3.2b, 3.3c, 3.4a and 3.5. Typically the average loss was evaluated on a grid $\{\rho_i\}_{i=1}^n$ of size n and plotted on a log-log scale. Each of the nk realizations of the noise is distributed according to $\hat{z}_{ij} \sim \mathcal{N}(0,1)$ unless otherwise noted. The noise level is denoted by $\eta > 0$ and the signal is given by x_0 where $x_0 = N \sum_{i=1}^s e_i$ with e_i being the ith standard basis vector. The grid $\{\rho_i\}_{i=1}^n$ was logarithmically spaced and centered about $\rho_{(n+1)/2} = 1$ with n always odd. The solutions to each PD problem were obtained using standard available methods in Python: `sklearn`'s `minimize_scalar` function from the `optimize` module was used for solving (LS_τ^*) and (BP_σ^*) [58], while the solution to (QP_λ^*) was obtained *via* soft-thresholding. Finally, the optimal values τ^*, λ^* and σ^* were either determined analytically (e.g., $\tau^* = \|x_0\|_1$), or estimated on a dense grid about an approximately optimal value for that parameter. Initial guesses for σ^* and λ^* were $\eta\sqrt{N}$ and $\eta\sqrt{2\log(N/s)}$ respectively.

3.5.1 (LS_τ^*) numerical simulations

This section presents numerical simulations demonstrating PS of (LS_τ^*) in the low-noise regime for two different ambient dimensions $N = 10^3, 10^6$. This repetition has the benefit of showcasing the behaviour of (LS_τ^*) at two different sparsity levels, as well as contrasting the behaviour of (LS_τ^*) with (QP_λ^*) and (BP_σ^*) at relatively low and high dimensions. Using the notation above, $n = 501$ points and $s = 20$; $(k, N) = (50, 10^3)$ for Figure 3.2a, while $(k, N) = (25, 10^6)$ for Figure 3.2b. In both regimes, x_0 is quite sparse $(s/N \sim 10^{-2}, 10^{-5})$ with entries that are well separated from the noise $(N/\eta \sim 10^6, 10^9)$.

We may glean several pieces of information from these two plots. Most notably, the (LS_τ^*) PS manifests in very low dimensions, relative to practical problem sizes. Moreover, the curve for (LS_τ^*) average loss seems to approach something resembling the sharp asymptotic phase transition described by Theorem 3.2.1. One may also notice the behaviour of the other two programs



(a) **(b)**

Figure 3.2: (LS^*_τ) PS in the low-noise regime. The loss for (LS^*_τ), in green, exhibits greater parameter sensitivity than either (BP^*_σ) (in orange) or (QP^*_λ) (in blue). Plotted is the average loss, as per (3.2), for each program. The data are shown on a log-log scale and plotted with respect to the normalized parameter ρ. The data parameters for (a) are $(s, N, \eta, k, n) = (20, 10^3, 10^{-3}, 50, 501)$ and those for (b) are $(s, N, \eta, k, n) = (20, 10^6, 10^{-3}, 25, 501)$.

in the low-noise regime. It is apparent that the magnitude of the derivative for the (QP^*_λ) risk increases markedly on the left-hand side of the optimal normalized parameter value (i.e., below 1) between the $N = 10^3$ and $N = 10^6$ plots. This behaviour is consistent with the result in Theorem 3.3.4 that the left-sided risk scales as a power law of N.

Finally, we observe that (BP^*_σ) develops a shape resembling the sensitivity of (LS^*_τ) when $N = 10^6$. We defer commentary on this behaviour to § 3.5.3.

3.5.2 (QP^*_λ) analytic plots

In this section we plot $R^\sharp(\lambda; s, N)$ using (3.1). In addition we plot the magnitude of its derivative with respect to the normalized parameter u, where $\lambda = u\lambda^*$. For fixed sparsity, we visualize $R^\sharp$ as a function of u and the ambient dimension N.

In Figure 3.3a we plot $R^\sharp(u\lambda^*; 1, N)$ as a function of N for $u \in \{1/2, 1 + 10^{-6}, 2\}$. As reference, we also plot the curves $y = CN^\kappa$ for $\kappa \in \{1/20, 7/10\}$. It is evident from the reference lines that $R^\sharp(u\lambda^*; 1, N)$ scales like a power law

of N for $u = 0.5$, while $R^\sharp(u\lambda^*; 1, N)$ appears to have approximately order-optimal growth for $u = 1 + 10^{-6}, 2$. The derivatives of these three functions are visualized in Figure 3.3b, which includes reference lines $y = CN^\kappa, \kappa \in \{1/20, 7/10\}$. Again, it is evident that the derivative scales as a power law of N for those risks with $u < 1$; apparently order-optimal otherwise. As a technical point, we note that we selected $u = 1 + 10^{-6}$ instead of $u = 1$, because in the latter case the derivative is 0, which is poorly represented on a log-log scale.

In Figure 3.3c we plot $R^\sharp(u\lambda^*; 1, 10^{10})$ as a function of the normalized parameter u. One may observe PS for $u < 1$ (i.e., $\lambda < \lambda^*$), for example by comparison to the plotted reference line $y \sim u^{-25}$. Similarly, one may observe right-sided parameter stability of $R^\sharp(u\lambda^*; 1, 10^{10})$ by comparing with the second plotted reference line, $y \sim u^2$. From these simulations one may observe that choosing $\lambda = .5\lambda^*$ accrues at least a 10^6 fold magnification of the risk. Observe that the plot of the analytic expression for $R^\sharp(\lambda; s, N)$ agrees well with the simulations approximating $R^\sharp(\lambda; x_0, N, \eta)$ in Figure 3.2 and Figure 3.4. In Figure 3.3d we plot $\left|\frac{\mathrm{d}}{\mathrm{d}u} R^\sharp(u\lambda^*; 1, 10^{10})\right|$ as a function of the normalized parameter u. Again, the right-sided stability and left-sided sensitivity are apparent. The cusp about $u = 1$ is due to the fact that the derivative vanishes for $u = 1$ (this is by definition the value of u yielding minimal risk); this cusp was cut-off at 10^{-5} for visualization purposes.

Finally, we would like to clarify a potentially confusing issue. Though our theory for (QP^*_λ) refers to λ^* only through its connection with $\bar{\lambda}$, we were able to approximate λ^* empirically in our numerical simulations. Accordingly, we have made reference to it when discussing parameter stability regimes.

3.5.3 (BP^*_σ) numerical simulations

This section presents numerical simulations demonstrating PS of (BP^*_σ) in the regime where x_0 is very sparse. Figure 3.4a is generated as described by the above procedure in § 3.5, with parameters $(s, N, \eta, k, n) = (1, 10^6, 1, 31, 501)$, while Figure 3.4b was generated in a way that mirrors the proof of Theorem 3.4.9, with parameters $(s, \eta, k, n) = (1, 1, 31, 51)$.

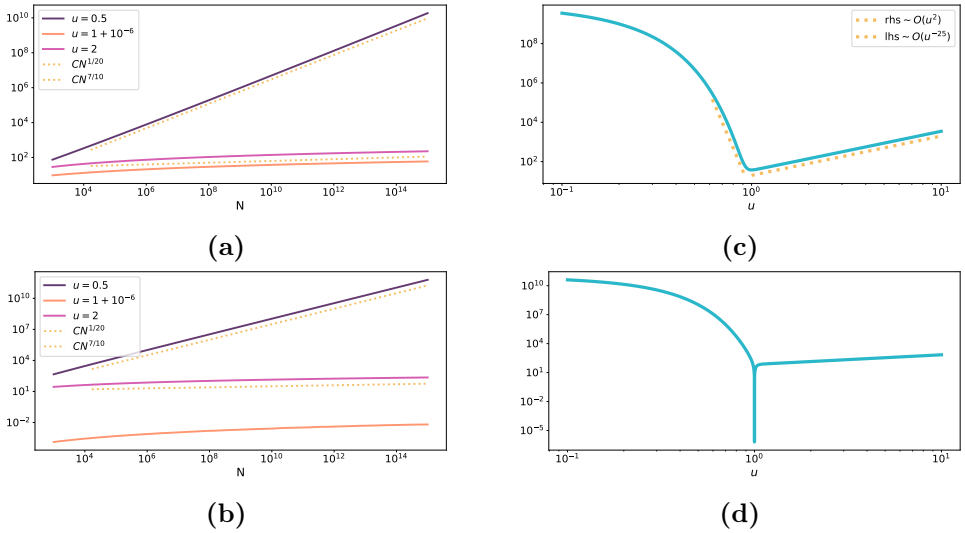


Figure 3.3: (QP^*_λ) PS in the low-noise regime. All curves are generated analytically using the expressions obtained in § 3.3 and plotted on a log-log scale. (a) A plot of $R^\sharp(u\lambda^*; 1, N)$ as a function of N for $u \in \{1/2, 1 + 10^{-6}, 2\}$. The lines $y = CN^\kappa$ for $\kappa \in \{1/20, 7/10\}$ are plotted for reference. (b) A plot of the magnitude of $\frac{\mathrm{d}}{\mathrm{d}u} R^\sharp(u\lambda^*; 1, N)$ as a function of N for $u \in \{1/2, 1 + 10^{-6}, 2\}$. The lines $y = CN^\kappa$ for $\kappa \in \{1/20, 7/10\}$ are plotted as reference. Note that $\frac{\mathrm{d}}{\mathrm{d}u} R^\sharp(\lambda^*; 1, N) = 0$ and so we have omitted plotting this quantity on a log-log scale. (c) A plot of $R^\sharp(u\lambda^*; 1, 10^{10})$ as a function of the normalized parameter u. Two lines are plotted as reference for risk growth rate with respect to u. (d) A plot of $\left|\frac{\mathrm{d}}{\mathrm{d}u} R^\sharp(u\lambda^*; 1, 10^{10})\right|$ as a function of the normalized parameter u. We exclude $u = 1$ from the plot; the function's value at this point is undefined on a log-log scale, since $\frac{\mathrm{d}}{\mathrm{d}u}\Big|_{u=0} R^\sharp(u\lambda^*; 1, 10^{10}) = 0$.

The purpose of Figure 3.4a is to resolve PS of (BP^*_σ) about the optimal parameter choice. Because the theory suggests that $\tilde{R}(\sigma; x_0, N, \eta)$ is surely resolved when the ambient dimension is sufficiently large, we set $N = 10^6$; this value was expected to resolve the sensitivity, as per the discussion later in this section. We limited the number of realizations, because the problem size was computationally prohibitive. The minimal average loss observed on

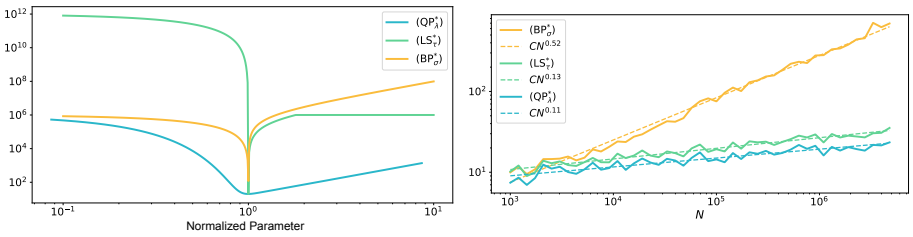
the plot was significantly larger than the respective minimal average losses of (LS^*_τ) and (QP^*_λ) by a factor of 5.9, supporting the theory. We expect this discrepancy to increase as dimension increases; however, this investigation was limited by the computational demands of the problem. We also noticed a cusp-like behaviour, which would be an interesting object of further study.

Figure 3.4b was generated so as to mirror the theory backing Theorem 3.4.9. Specifically, noise realizations were constrained to the constant probability event $\{\|z\|_2^2 - N \in (.5\sqrt{N}, 5\sqrt{N})\}$. Plotted in the figure is the best average loss as a function of N, for each program,

$$\bar{L}_{\text{best}}(N; x_0, \eta, k, n) := \frac{1}{k} \min_{i \in [n]} \sum_{j=1}^{k} L(\rho_i(N); x_0, N, \eta \hat{z}_j).$$

The domain for N ranges from 10^3 to $4.786 \cdot 10^6$, computed on a logarithmically spaced grid composed of 47 points. For each program and each value N in the grid, the average loss was computed for $n = 51$ values of the normalized parameter, each using $k = 31$ realizations of noise. The grid of n values were centered at the optimal value of the normalized parameter, and were densest about this value, so as to best resolve the average loss curves about their optimal values. Included for reference is a simple linear regression fitted on the log-scaled data (i.e., a linear regression fitted to the log-best average loss as a function of the log-normalized parameter, for each program). The regression coefficients were computed using standard tools in Python [58]. The slope of the linear regression corresponds to the power law of the data, as depicted in the legend. Observe that the value for (BP^*_σ) approximately matches the lower bound derived for $\tilde{R}(\eta\sqrt{N}; x_0, N, \eta)$ in Lemma 3.4.1.

Finally, observe that in both Figures 3.4a and 3.2b, the average loss for (BP^*_σ) has a cusp-like shape resembling the sensitivity of (LS^*_τ) in Figure 3.2b when $N = 10^6$. We offer the plausible explanation that the relative sparsity of the signal is small $(s/N = 2/10^5)$ and thus this regime coincides with the regime in which (BP^*_σ) develops PS. We observed similar behaviour in Figure 3.4, which depicted cusp-like behaviour about the optimal choice of the normalized parameter. It is interesting that parameter sensitivity



(a) **(b)**

Figure 3.4: (BP^*_σ) PS in the very sparse regime. (a) Data parameters: $(s, N, \eta, k, n) = (1, 10^6, 1, 31, 501)$. Average losses plotted on a log-log scale with respect to the normalized parameter. (b) Best average loss for (BP^*_σ) as a function of N. Data parameters: $(s, \eta, k, n) = (1, 1, 31, 51)$. The function $\sigma_{\mathrm{opt}}(N)$ was obtained as the value of σ bestowing minimal average loss of the program for each N. A simple linear regression was fitted the log-scaled data and plotted for reference; each model's slope is equal to the power appearing in its respective legend entry. .

developed by (BP^*_σ) seems to manifest in a way similar to that of (LS^*_τ), because Theorem 3.4.9 shows that there is no good choice of parameter σ, though Figure 3.4 supports that there is a single best choice, albeit minimax suboptimal, when N is moderately large.

Simulating theorem parameters

Here we clarify the relationship between some of the constants appearing in the proofs of Theorem 3.4.6 and Theorem 3.4.9. We provide two examples of minimal N_0 values guaranteeing PS behaviour of (BP^*_σ) for given parameter choices. The theory does not claim these values to be optimal, nor do we claim that the constants are tuned. In particular, these demonstrations seem rather pessimistic, especially by comparison with the numerical simulations in Figure 3.4.

The following values were determined by computing

$$N_0 := \max\{N_0^{(3.6.6)}(a_1, C_1, L), N_0^{(3.6.9)}(C_2, L)\}$$

for particular choices of a_1, C_1, C_2 and L, using their definitions in the technical results of § 3.6.4. Thus, the theory of § 3.4 guarantees PS for all $N \geq N_0$ when

$$\begin{aligned} N_0 &\approx 1.5\text{e}6 \quad \text{and} \quad (a_1, C_1, C_2, L) \approx (1.45, 5, 4, 3.78) \qquad \text{or} \\ N_0 &\approx 4.9\text{e}5 \quad \text{and} \quad (a_1, C_1, C_2, L) \approx (1.58, 4.04, 4, 3.62). \end{aligned}$$

These numbers appear pessimistic, given that N_0 is large, while $(C_2, C_1) \approx (4, 5)$ implies the sensitivity arises on the event $\{\|z\|_2^2 - N \in (4\sqrt{N}, 5\sqrt{N})\}$, which occurs with relatively minute (but constant) probability. Thus, it may not be all that surprising that (BP_σ^*) suboptimality is difficult to ascertain empirically from a small number of realizations in only moderately large dimension when $\sigma \approx \sigma^*$.

3.5.4 Parameter stability in sparse proximal denoising

In this section we show numerical simulations in which the three programs appear to exhibit better parameter stability. For these simulations, $\eta \approx 233.0$, $s = 2500$ and $N = 10^4$. Average loss was computed from $k = 31$ realizations for $n = 501$ grid points. As the noise is large, this setting lies (mostly) outside the regime in which (LS_τ^*) and (QP_λ^*) exhibit PS. Moreover, the signal is not very sparse, since $s/N = .25$. Thus, this setting also lies outside the regime in which (BP_σ^*) exhibits PS. Accordingly, smooth risk curves are seen for (BP_σ^*) and (QP_λ^*). While (QP_λ^*) and (BP_σ^*) appear relatively gradual, (LS_τ^*) appears to avoid a pronounced cusp-like point about $\tau/\tau^* = 1$. These data are visualized in Figure 3.5.

3.5.5 Triptych comparison of program sensitivity

Here, we include two sets of visualizations. In each, a key parameter is varied to explicitly show the way in which it affects the PS of a program without affecting the PS of the other(s).

In Figure 3.6, the parameters for the synthetic experiment were $(s, N, k, n) = (1, 10^5, 50, 501)$ with $\eta \in \{10^{2j-1}\}_{j=-2}^{2}$. As usual, the plots visualize the

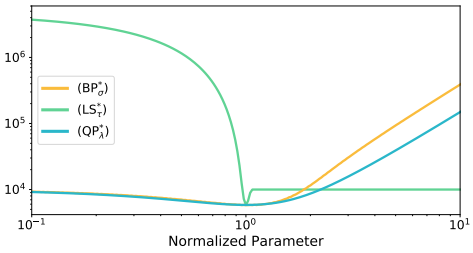


Figure 3.5: Parameter stability of sparse PD programs when $(s, N, \eta, k, n) = (2500, 10^4, 233, 31, 501)$. Plotted curves represent average loss, plotted on a log-log scale.

average loss as a function of the normalized parameter, with the plots corresponding to (BP^*_σ), (LS^*_τ) and (QP^*_λ), respectively, from left to right. One readily observes that the average loss curves for (LS^*_τ) become sharper as the noise-level decreases (centre plot). In contrast, there is little observable difference in the average loss for both (BP^*_σ) (left plot) and (QP^*_λ) (right plot) as η varies. We relegate other comparisons drawn between the three programs to the relevant subsections of this thesis.

The parameters for the synthetic experiment shown in Figure 3.7 were $(N, \eta, k, n) = (10^6, 100, 50, 501)$ with $s \in \{1, 5, 25, 125\}$. As above, the plots visualize average loss as a function of the normalized parameter for the programs (BP^*_σ), (LS^*_τ), and (QP^*_λ), respectively, from left to right. In each plot, which is plotted on a log-log scale, s is varied to showcase this behaviour. The plots serve to exemplify PS of (BP^*_σ) in the very sparse regime. The average curves for (BP^*_σ) become more parameter unstable as s decreases (left-most plot). Indeed, away from the optimal choice of the normalized parameter, the curves are nearly identical; however, at the optimal parameter, the minimum average loss increases as s does (the minimum average loss values for (BP^*_σ) are highlighted by arrows of matching colour). Thus, when the sparsity level is small, small perturbations about the optimal parameter choice lead to larger changes in average loss relative to larger values of s. In particular, when s is small, small variations in the normalized parameter essentially destroy the advantage imbued by sparsity. In contrast, the average loss curves for (LS^*_τ) become sharper about $\tau/\tau^* = 1$ as sparsity increases.

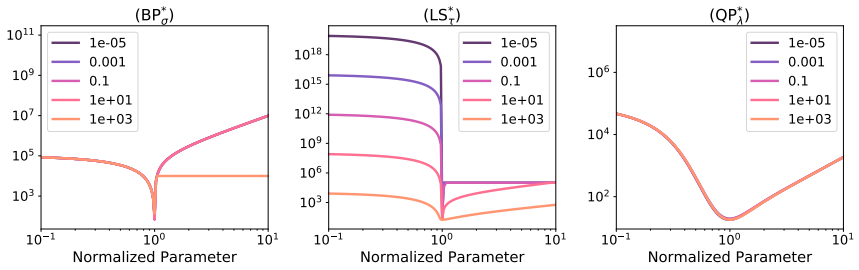


Figure 3.6: Dependence of each PD program on noise level η; $(s, N, k, n) = (1, 10^5, 50, 501)$. Each plot corresponds to a single program: (BP^*_σ), (LS^*_τ), (QP^*_λ), from left to right. In each plot, that program's average loss is plotted with respect to its normalized parameter on a log-log scale for each $\eta \in \{10^{-5}, 10^{-3}, 10^{-1}, 10^1, 10^3\}$. These plots are best observed on a computer (in colour).

As s increases, so too does the norm of x_0; we have seen that this is in a sense equivalent to decreasing η. Therefore, as s increases, one expects the average loss curve to become sharper (i.e., more parameter unstable). This is indeed what happens. Similarly, one sees that small variations about the optimal choice of the normalized parameter lead to larger variation in (QP^*_λ) average loss when s is smaller. Still, in the present regime, this program appears to be the most stable of the three with respect to variation of the normalized parameter about its optimal value.

3.5.6 Realistic denoising examples

Image-space denoising

We visualize how proximal denoising behaves for a realistic denoising problem. The ground truth signal is the standard $512 \times 512 \times 3$ colour image of a mandrill face, ravelled to a vector $x_0 \in [0,1]^N \subseteq \mathbb{R}^N$, $N = 786\,432$. The denoising is performed in image space. Specifically, the signal x_0 is not sparse: 99.98% of its coefficients are nonzero. We set $y_j = x_{0,j} + \eta z_j \in \mathbb{R}^N$ where $\eta = 10^{-5}, 1$ and $z_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. The results of this example are displayed in Figure 3.8: the ground truth and noisy images in the top row, and

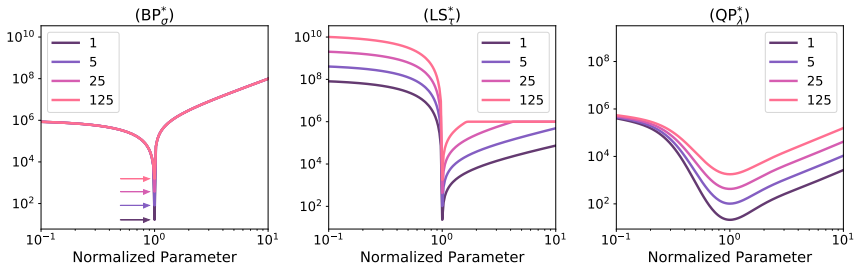


Figure 3.7: Dependence of each PD program on sparsity level s; $(N, \eta, k, n) = (10^6, 100, 50, 501)$. Each plot corresponds to a single program: (LS^*_τ), (BP^*_σ), (QP^*_λ), from left to right. In each plot, that program's average loss is plotted with respect to its normalized parameter on a log-log scale for each of $s \in \{1, 5, 25, 125\}$. For (BP^*_σ), the minimum average loss for each value of s is annotated with an arrow of matching colour. These plots are best observed on a computer (in colour).

quantitative results captured by plots of the average loss (3.2) in the bottom row.

The plots of average loss were generated from $k = 25$ realizations of noise z, with a logarithmically spaced grid of $n = 501$ points centered about the optimal parameter value for each of the three proximal denoising programs. The optimal parameter value for each program was determined analytically where possible, or numerically using standard solvers [78]. A smooth approximating curve of the non-uniformly spaced point cloud of loss realizations was computed using radial basis function (RBF) approximation (see § A.2.2). The RBF approximation used multiquadric kernels with parameters $(\varepsilon_{\mathrm{rbf}}, \mu_{\mathrm{rbf}}, n_{\mathrm{rbf}}) = (10^{-3}, 10^{-2}, 301)$. Here, $\varepsilon_{\mathrm{rbf}}$ is the associated RBF scale parameter, μ_{rbf} is a smoothing parameter and n_{rbf} is the number of grid points at which to approximate [78]. The RBF parameters for the approximation were selected so as to generate a smooth line that best represents the path about which the individual (noisy) data points concentrate.

About the optimal average loss (where the normalized parameter is 1), an average difference of 1% in the value of τ results in a nearly 5×10^5 fold difference in nnse on average when $\eta = 10^{-5}$. In contrast, that error

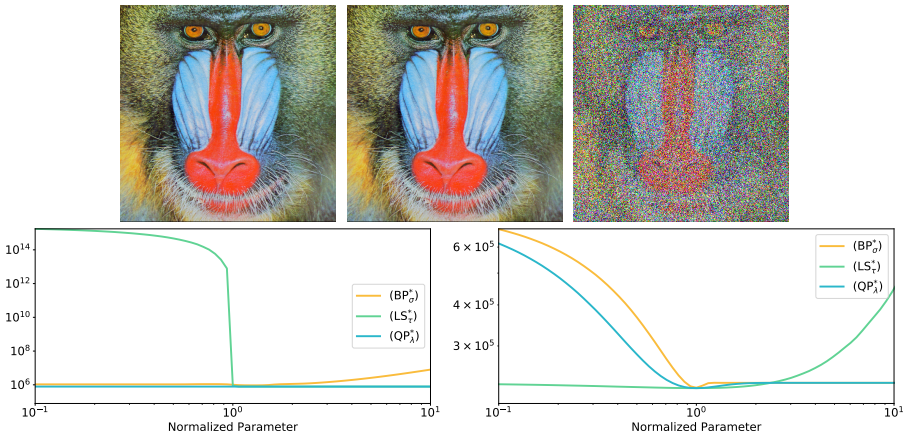


Figure 3.8: Top (left-to-right): The underlying signal is the $512 \times 512 \times 3$ mandrill image; the middle image is corrupted by iid normally distributed noise ($\eta = 10^{-5}$); the right-most image is corrupted by iid normally distributed noise ($\eta = 1$). The pixel values of the original image lie in $[0, 1]^3$; those of the noisy images are scaled to this range for plotting. **Bottom:** Average loss is plotted with respect to the normalized parameter for (LS^*_τ), (QP^*_λ) and (BP^*_σ) respectively when $\eta = 10^{-5}$ **(left)** and $\eta = 1$ **(right)**. The associated parameters are $(N, k, n) = (786\,432, 25, 501)$. Plotted lines are smoothed approximations of loss realization data using multiquadric RBFs. In the left-hand plot, the nnse curve for (BP^*_σ) is obscured by that of (QP^*_λ).

varies by no more than a factor of three in the large noise regime ($\eta = 1$). Moreover, we observe that the average losses computed for $\eta = 10^{-5}$ upper bound those computed for $\eta = 1$. These results suggest not to use (LS^*_τ) for proximal denoising when η is small, even when the underlying data are not sparse.

1D denoising example

In this section, we demonstrate PS regimes for a realistic example of a 1D signal using wavelet domain denoising. Specifically, an s-sparse 1D signal $x_0 \in \mathbb{R}^N$ was generated in the Haar wavelet domain, where $(s, N) = (10, 4096)$.

In the signal domain, iid normal random noise was added to the signal to generate $\mathcal{W}^{-1}y := \mathcal{W}^1 x_0 + \eta z$ where $\eta = \frac{N}{100}, \frac{N}{10}$. The denoising problem was solved in the wavelet domain on a grid of size 501 centered about the optimal normalized parameter and logarithmically spaced. Namely, the input to each program was y. The loss was computed in the signal domain after applying the inverse transform to the estimated solution:

$$L(\rho; x_0, N, \eta\hat{z}) := \eta^{-2}\|\mathcal{W}^{-1}(x^*(\rho) - x_0)\|_2^2.$$

A smooth approximation to the average loss $\bar{L}(\rho_i; x_0, N, \eta, k)$ was computed from $k = 50$ realizations of the noise using linear radial basis function approximation with parameters $(\mu_{\text{rbf}}, n_{\text{rbf}}) = (2, 501)$.

In Figure 3.9, we visualize how the three programs behave for denoising a 1D signal, sparse in the Haar wavelet domain, which has been corrupted by one of two different noise levels in the signal domain. The top row visualizes the ground truth signal with a realization of the corrupted signal for $\eta = N/100$ (top-left) and $\eta = N/10$ (top-right). The bottom row visualizes the average loss with respect to the normalized parameter of each program. In the high-noise regime (bottom-right), it is clearly seen that (BP^*_σ) is the most parameter unstable about the optimal parameter choice. Moreover, the best average loss for (BP^*_σ) is greater than that for (QP^*_λ) or (LS^*_τ), as suggested by the supporting theory. We note that (QP^*_λ) also has an average loss greater than the minimal one, and suggest — noting the local variability in the curve — that this is an artifact of the RBF approximation through the optimality region. In the moderate-noise regime, we see a situation in which (QP^*_λ) appears to be the most parameter stable — again consistent with our reasoning that unconstrained programs should exhibit better stability. In contrast, (LS^*_τ) is most parameter unstable below the optimal parameter, while it is (BP^*_σ) that is most parameter unstable above the optimal parameter. This behaviour may be indicative of a regime intermediate to those we have previously discussed (i.e., lying between strictly low-noise and strictly very sparse).

With the grid in Figure 3.10, we intend to elucidate how PS manifests

for each program as a function of the normalized parameter, by visualizing the recovered signal for different values of the normalized parameter. The top plot shows the same average losses that are plotted in the bottom-left of Figure 3.9. The dotted lines at $\rho = .5, .75, 1, 4/3, 2$, and the markers located approximately at the intersection of these lines with the loss curves, mark sections of the loss for which the solution to the program will be visualized. Indeed, for each value of ρ and each program, there is a corresponding plot in the grid that depicts the solution to the program for that normalized parameter value ρ, along with the original signal x_0, which is depicted as a black dotted line in each of the 15 plots. When ρ is too small for (BP^*_σ) and (QP^*_λ), it is clear that the noise fails to be thresholded away. In contrast, this occurs for (LS^*_τ) when ρ is too large. On the other hand, the signal content is thresholded away by (BP^*_σ) when ρ too large, and by (LS^*_τ) when ρ is too small. Notice that this behaviour does not seem to occur with (QP^*_λ), further supporting that (QP^*_λ) admits right-sided parameter stability.

Wavelet-space denoising

In this section, we demonstrate PS regimes for a realistic example using proximal denoising of an image signal in a wavelet domain. Namely, noise is added in the image domain, the data denoised in Haar wavelet space, and performance of the back-transformed estimator is evaluated in the image domain. The image was designed to resemble a Shepp-Logan phantom, but to admit a very sparse expansion in Haar wavelets. This modified phantom, which we coin the "Square Shepp-Logan phantom", was created so as to be sparse enough to allow for better visualization of (BP^*_σ) PS. Specifically, if one were to generate the same figures for the Shepp-Logan phantom, one would see that (BP^*_σ) is less parameter stable than (QP^*_λ), but that the behaviour is markedly less pronounced than the behaviour we visualize in Figure 3.11 or Figure 3.13. Indeed this discrepancy results from the standard Shepp-Logan phantom being less sparse (having more non-zero entries) in its Haar wavelet transform than our modification. An alternative demonstration using the standard Shepp-Logan phantom might proceed using a different

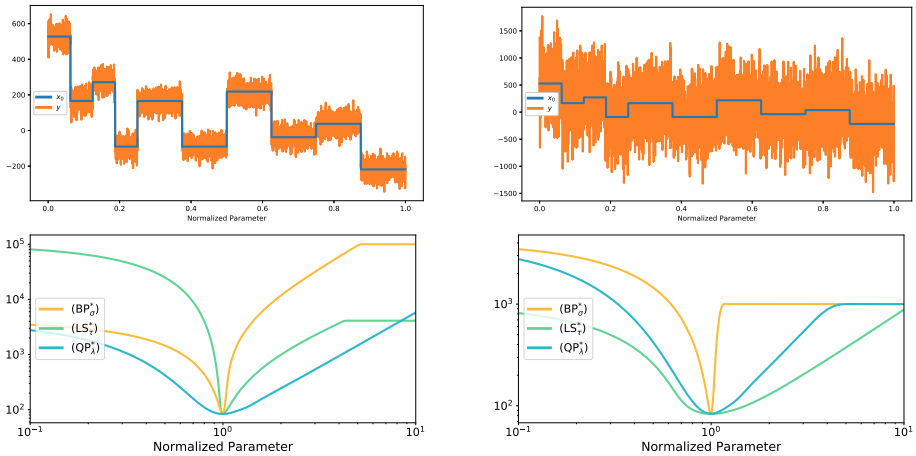


Figure 3.9: Haar wavelet space denoising of a 1D signal that is sparse in the Haar wavelet domain for two different noise levels, $\eta = N/100$ (left column) and $\eta = N/10$ (right column). **Top:** each plot contains a realization of the noisy 1D signal plotted in orange with the ground truth signal in blue. **Bottom:** plots of the average loss as a function of the normalized parameter for each program, computed using RBF approximation. The parameters for this example are: $(s, N, k, n) = (10, 4096, 50, 501)$.

transform domain in which its representation is sparser.

A corrupted Square Shepp-Logan phantom was obtained by adding iid noise $z_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ to the image pixels $I = (I_{i,j})_{i,j}$, yielding y where $y_{i,j} = I_{i,j} + \eta z_{i,j}$ with $\eta = 10^{-5}, 0.5$ and where $I_{i,j} \in [0,1]$ is the (i,j)th pixel of the uncorrupted Square Shepp-Logan phantom. The input signal to each recovery program was the vectorized 2D Haar wavelet transform of $y_{i,j}$: $w = \mathcal{W}(y_{i,j})_{i,j}$ where $\mathcal{W}$ is the operator connoting a Haar wavelet transform to (vectorized) Haar wavelet coefficients. Loss was computed in the image domain, using the nnse of the inverse-transformed proximal denoising estimator. For example, the loss for (BP^*_σ) is given by $\eta^{-2}\|\mathcal{W}^{-1}(\tilde{x}(\sigma)) - I\|_2^2$. Average loss (3.2) was thus computed by averaging the loss over $k = 25$ realizations of the noise z.

The associated parameters of the problem are $(s, N, k, n) = (5188, 409618, 25, 501)$, implying a relative sparsity of 1.27%. To create ef-

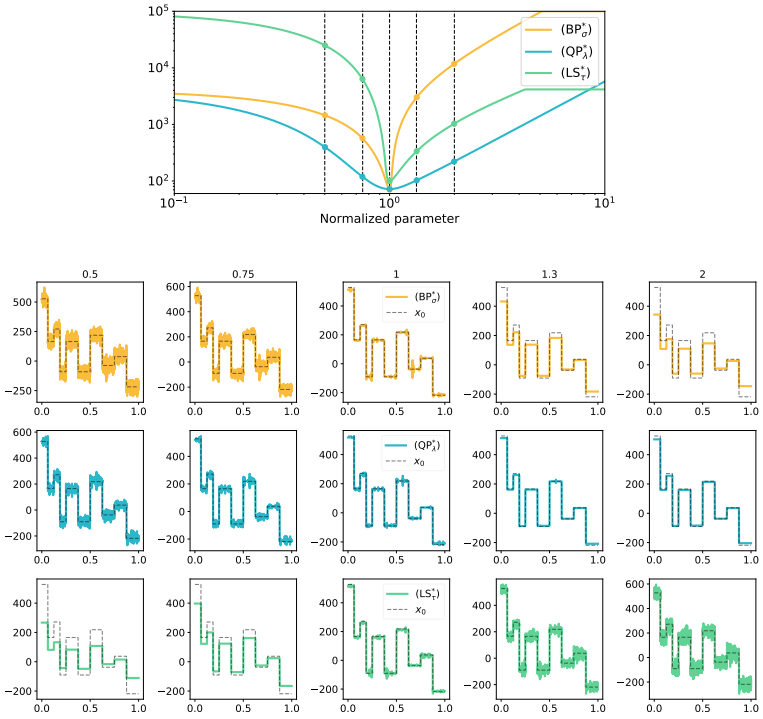


Figure 3.10: Wavelet space denoising of a 1D signal for different values of the normalized parameter when $\eta \approx 41$. **Top:** The sections of the average-loss surface for which estimator recovery will be visualized are depicted by the dots which lie nearly on the black dotted lines, themselves located at $\rho = 0.5, 0.75, 1, 4/3, 2$. **Bottom:** This group of fifteen plots represents a program's solution for a particular value of the normalized parameter, arranged in a grid. Each row of the 15 plot grouping represents a program, as denoted by the legend label; each column a value of the normalized parameter, as determined by the heading above the top row.

fective visualizations of the PS behaviour, the noisy images seen in the top row of Figure 3.11 are scaled to the interval $[0, 1]$. Subsequent visualizations do not perform this rescaling so that a perceptual evaluation of the recovery is better facilitated.

The plots in the bottom row of Figure 3.11 depict the average loss as a

function of the normalized parameter ρ of each program. For each of the k realizations, the loss was computed on a logarithmically spaced grid of $n = 501$ points about the optimal parameter. As in § 3.5.6, a smooth approximating curve to the non-uniformly spaced point cloud of loss realizations was computed using RBF approximation. The RBF approximation used multiquadric kernels with parameters $(\varepsilon_{\text{rbf}}, \mu_{\text{rbf}}, n_{\text{rbf}}) = (10^{-3}, 10^{-2}, 301)$ [78]. The RBF parameters for the approximation were selected so as to generate a smooth line that best represents the path about which the individual (noisy) data points concentrate, especially so as to resolve the behaviour of the loss about $\rho = 1$.

About the optimal average loss, an approximate 10^8 fold difference in average loss results from a less than 2% perturbation of τ with $\tau < \tau^*$ in the low-noise and very sparse regime ($\eta = 10^{-5}, s/N \approx 1.27\%$); an approximate 10^2 fold difference in average loss results from a less than 2% perturbation of τ with $\tau > \tau^*$, which is consistent with the theory. In this regime, we observe that (BP^*_σ) is less stable than (QP^*_λ) for values of the normalized parameter greater than 1, as suggested by our theory. In the very sparse regime with large noise ($\eta = 0.5$), (BP^*_σ) is markedly more parameter unstable than (LS^*_τ) or (QP^*_λ), especially for values of the normalized parameter exceeding 1.

In Figure 3.12 and Figure 3.13 we depict estimator performance by visualizing the solution to each program at specific values of the normalized parameter. The description of each figure is identical, but the noise levels η differ between them. Specifically, for each program we show the recovered image and its pixel-wise nnse for values of the normalized parameter $\rho = 0.5, 0.75, 1, 4/3, 2$. The plot in the top row of the figure depicts the loss as a function of the normalized parameter for each program and for a realization of the noise z. Also depicted are reference lines for the corresponding values of the normalized parameter whose recovered image are visualized. The middle row contains a grid of 15 images; each column corresponds to a value of the normalized parameter as denoted by the title heading, while each row corresponds to a proximal denoising program as denoted by the labels along the left-most y-axis. The bottom grouping of 15 images depicts the pixel-wise nnse, arranged identically to the middle row. Because

the average loss curves were computed on a grid of n logarithmically spaced points centered about the optimal parameter value, we do not visualize the recovered image for the exact values of ρ given above, but for those values represented by the coloured points seen in the plot of the top row. These points are sufficiently close to the quoted values of ρ so as to visualize the program behaviour all the same.

The numerics of Figure 3.12 occur in the low-noise regime ($\eta = 10^{-5}$), and so, as expected, demonstrate PS of (LS^*_τ). We note that pixel-wise nnse for (BP^*_σ) is approximately 20 times worse than (QP^*_λ) when $\rho \approx 2$. Moreover, the pathologies (in the sense of pixel-wise nnse) of these latter two programs appear similar. We also observe that the pixel-wise nnse varies more greatly for (BP^*_σ) than for (QP^*_λ) as ρ varies from 0.75 to 4/3. This is consistent with our theory for the behaviour of (BP^*_σ) in the very sparse regime. The numerics of Figure 3.13 occur in the high-noise regime ($\eta = 0.5$). Failure of (BP^*_σ) in the very sparse regime is seen from examining the solution itself. For example, when $\rho < 1$, pixel values of the solution to (BP^*_σ) may reach more than 2 or even be negative. This pathology manifests as large-magnitude pixelation in the corresponding plots of pixel-wise nnse. Catastrophic failure of (BP^*_σ) is observed for $\rho > 1$, in which the program fails to recover any semblance of the original image. Specifically, large σ shrinks the wavelet coefficients to near the origin, enforcing few non-zero components that are small in magnitude. This yields the rectangular pattern observed in the solutions for (BP^*_σ) (top-right of the middle row). In contrast, moderate deformation of the image is observed for $\rho \neq 1$ for both (QP^*_λ) and (LS^*_τ).

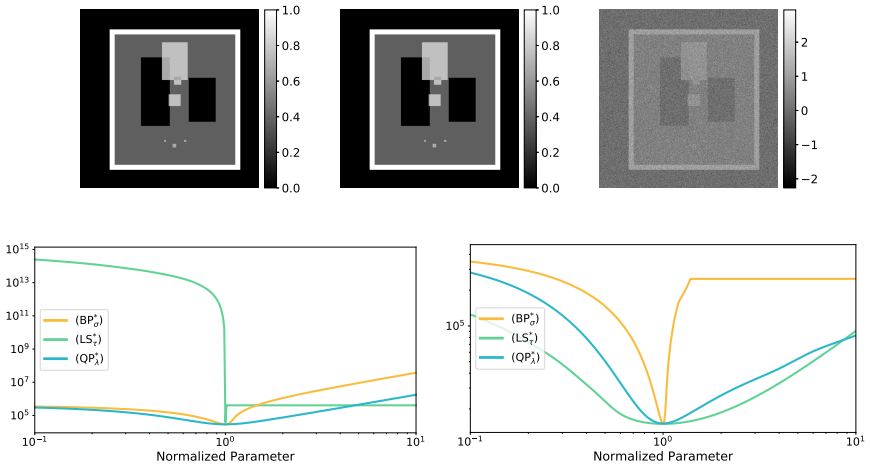


Figure 3.11: Top (left-to-right): The underlying signal is the 640×640 Square Shepp-Logan phantom image; the middle image is corrupted by iid normally distributed noise ($\eta = 10^{-5}$); the right-most image is corrupted by iid normally distributed noise ($\eta = 0.5$). The pixel values of the original image lie in $[0, 1]$; those of the noisy images are scaled to $[0, 1]$. **Bottom:** Average loss is plotted with respect to the normalized parameter for (LS^*_τ), (QP^*_λ) and (BP^*_σ) respectively when $\eta = 10^{-5}$ (left) and $\eta = 0.5$ (right). The associated parameters are $(s, N, k, n) = (5188, 409618, 50, 501)$, implying relative sparsity of 1.27%. Plotted lines are smoothed approximations of loss realization data using multiquadric RBFs.

3.6 Proofs

3.6.1 Preliminary results

Proof of ℓ_1 tangent cone equivalence

Proof of Lemma 2.3.3. First observe that the definition of $F_{\mathcal{C}}(x)$ is equivalent to

$$F_{\mathcal{C}}(x) = \{h \in \mathbb{R}^N : h = z - x,\ \|z\|_1 \leq \|x\|_1\}.$$

Next, observe that $K(x)$ is a cone. So, for left containment, it suffices to show $F_{\mathcal{C}}(x) \subseteq K(x)$ since the cone generated by a set is no larger than any cone containing that set. By definition of $h = z - x \in F_{\mathcal{C}}(x)$:

$$\begin{aligned}\langle \operatorname{sgn}(x), x\rangle = \|x\|_1 \geq \|z\|_1 = \|z_T\|_1 + \|h_{T^C}\|_1;\\ \|z_T\|_1 = \langle \operatorname{sgn}(z), z_T\rangle \geq \langle \operatorname{sgn}(x), z_T\rangle.\end{aligned}$$

The above expressions combine to yield left containment:

$$\|h_{T^C}\|_1 \leq -\langle \operatorname{sgn}(x), z_T - x\rangle = -\langle \operatorname{sgn}(x), h_T\rangle = -\langle \operatorname{sgn}(x), h\rangle.$$

To show right containment, it suffices to show, for any $w \in K(x)$, that there exists $\alpha > 0$ so that $\alpha w \in F_{\mathcal{C}}(x)$. Fix $w \in K(x)$ and select $\alpha > 0$ sufficiently small so that $z := x + \alpha w$ admits $z_j x_j \geq 0$ for all $j \in T$. Using $\alpha\|w_{T^C}\|_1 \leq -\alpha\langle \operatorname{sgn}(x), w_T\rangle$, we show $\|z\|_1 \leq \|x\|_1$ implying that $\alpha w \in F_{\mathcal{C}}(x)$, whence $w \in T_{\mathcal{C}}(x)$. Defining $h := \alpha w = z - x$, we have

$$\begin{aligned}\|z\|_1 &= \|z_T\|_1 + \|z_{T^C}\|_1 = \langle \operatorname{sgn}(z_T), z_T\rangle + \|h_{T^C}\|_1\\ &\leq \langle \operatorname{sgn}(z_T), z_T\rangle - \langle \operatorname{sgn}(x), h_T\rangle\\ &= \langle \operatorname{sgn}(z_T), z_T\rangle - \langle \operatorname{sgn}(x), h_T\rangle + \langle \operatorname{sgn}(x), x\rangle - \langle \operatorname{sgn}(x), x\rangle\\ &= \langle \operatorname{sgn}(x), x\rangle + \langle \operatorname{sgn}(z_T), z_T\rangle - \langle \operatorname{sgn}(x), z_T\rangle\\ &= \|x\|_1 + \langle \operatorname{sgn}(z_T) - \operatorname{sgn}(x), z_T\rangle = \|x\|_1.\end{aligned}$$

The latter equality follows from the fact that $\langle \operatorname{sgn}(z_T) - \operatorname{sgn}(x), z_T\rangle = 0$ by choice of α. Thus, $w \in T_{\mathcal{C}}(x)$ and $T_{\mathcal{C}}(x) = K(x)$ as desired. □

Proof of worst-case risk equivalence

Lemma 3.6.1 (Increasing risk)**.** *Fix $x_0 \in \Sigma_s^N$, $\|x_0\|_1 = 1$. Then $\hat{R}(\tau; \tau x_0, N, \eta)$ is an increasing function of $\tau \geq 0$.*

Proof of Lemma 3.6.1. Given $y(\tau) := \tau x_0 + \eta z$ for $\eta > 0$ and $z \in \mathbb{R}^N, z_i \overset{\text{iid}}{\sim}$

$\mathcal{N}(0,1)$, let $\hat{x}(\tau) := \hat{x}(\tau; y(\tau))$ solve

$$\hat{x}(\tau; y(\tau)) := \arg\min_x \{\|y(\tau) - x\|_2 : \|x\|_1 \leq \tau\}$$

Let $K := B_1^N - x_0$, a convex set containing the origin. Using a standard scaling property of orthogonal projections,

$$\begin{aligned}\hat{x}(\tau) - \tau x_0 &= \arg\min_w \{\|\eta z - w\|_2 : \|w + \tau x_0\|_1 \leq \tau\} \\ &= \mathrm{Proj}_{\tau K}(\eta z).\end{aligned}$$

Hence, it follows by Lemma 2.1.2 that $\|\hat{x}(\tau) - \tau x_0\|_2$ is an increasing function of τ. □

Proposition 3.6.2 (Risk equivalence)**.** *Let $\eta, \tau > 0$ and fix $N \geq 2$. Then*

$$\sup_{x \in \Sigma_s^N} \hat{R}(\|x\|_1; x, N, \eta) = \max_{\substack{x \in \Sigma_s^N \\ \|x\|_1 = 1}} \lim_{\tau \to \infty} \hat{R}(\tau; \tau x, N, \eta) = \max_{\substack{x \in \Sigma_s^N \\ \|x\|_1 = 1}} \lim_{\eta \to 0} \hat{R}(1; x, N, \eta).$$

Proof of Proposition 3.6.2. The first equality is an immediate consequence of Lemma 3.6.1:

$$\sup_{x \in \Sigma_s^N} \hat{R}(\|x\|_1; x, N, \eta) = \max_{\substack{x \in \Sigma_s^N \\ \|x\|_1 = 1}} \sup_{\tau > 0} \hat{R}(\tau; \tau x, N, \eta) = \max_{\substack{x \in \Sigma_s^N \\ \|x\|_1 = 1}} \lim_{\tau \to \infty} \hat{R}(\tau; \tau x, N, \eta).$$

The second equality follows from a standard property of orthogonal projections, and the risk expression derived in Lemma 3.6.1. For $K := B_1^N - x$,

$$\begin{aligned}&\max_{\substack{x \in \Sigma_s^N \\ \|x\|_1 = 1}} \lim_{\tau \to \infty} \hat{R}(\tau; \tau x, N, \eta) = \max_{\substack{x \in \Sigma_s^N \\ \|x\|_1 = 1}} \lim_{\tau \to \infty} \eta^{-2} \, \mathbb{E} \, \| \mathrm{Proj}_{\tau K}(\eta z) \|_2^2 \\ &= \max_{\substack{x \in \Sigma_s^N \\ \|x\|_1 = 1}} \lim_{\tau \to \infty} \frac{\tau^2}{\eta^2} \, \mathbb{E} \, \| \mathrm{Proj}_K(\tau^{-1} \eta z) \|_2^2 = \max_{\substack{x \in \Sigma_s^N \\ \|x\|_1 = 1}} \lim_{\tilde{\eta} \to 0} \tilde{\eta}^{-2} \, \mathbb{E} \, \| \mathrm{Proj}_K(\tilde{\eta} z) \|_2^2 \\ &= \max_{\substack{x \in \Sigma_s^N \\ \|x\|_1 = 1}} \lim_{\eta \to 0} \hat{R}(1; x, N, \eta).\end{aligned}$$

□

Proof of (LS^*_τ) optimal risk

Proof of Proposition 3.1.5. Directly from Theorem 2.1.1,

$$R^*(s,N) = \max_{\substack{x_0 \in \Sigma_s^N \\ \|x_0\|_1 = 1}} \mathbf{D}(T_{B_1^N}(x_0)),$$

where $\mathbf{D}(T_{B_1^N}(x_0))$ is the statistical dimension of the ℓ_1 descent cone (i.e., the mean-squared distance to the polar of the ℓ_1 descent cone. The operator $\mathbf{D}$ has the following desirable relation to the Gaussian mean width: if $\mathcal{C} \neq \{0\}$ is a non-empty convex cone [2, Prop 10.2],

$$\mathrm{w}^2(\mathcal{C} \cap \mathbb{S}^{N-1}) \leq \mathbf{D}(C) \leq \mathrm{w}^2(\mathcal{C} \cap \mathbb{S}^{N-1}) + 1.$$

Thus, it suffices to lower- and upper-bound $\mathrm{w}^2\left(T_{B_1^N} \cap \mathbb{S}^{N-1}\right)$. The desired upper bound is an elementary but technical exercise using Hölder's inequality, Stirling's approximation and a bit of calculus. The lower bound may be computed using Sudakov's inequality and [35, Lemma 10.12]. It thereby follows that

$$cs\log(N/s) \leq \mathbf{D}(T_{B_1^N}(x_0)) \leq Cs\log(N/s),$$

where $c, C > 0$ are universal constants. Accordingly, $cs\log(N/s) \leq R^*(s,N) \leq Cs\log(N/s)$.

From Theorem 3.3.6, $R^\sharp(\lambda^*; s, N) \leq Cs\log N$ for any $N \geq N_0(s)$ with $N_0(s)$ sufficiently large. Using the above equation gives, for $c, C_1 > 0$, $Cs\log N \leq C_1 cs\log(N/s) \leq C_1 R^*(s,N)$. Finally, $R^\sharp(\lambda^*; s, N)$ is lower-bounded trivially by $M^*(s,N)$ and $M^*(s,N) = \Theta(s\log(N/s))$ [17]. □

Elementary results from probability

We briefly recall two aspects of how normal random vectors concentrate in high dimensions.

Proposition 3.6.3. *Let $z \in \mathbb{R}^N$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, fix constants $C_2 < C_1 < \infty$ and define the event $\mathcal{Z}_\pm$ by $\mathcal{Z}_\pm := \{C_2\sqrt{2N} \leq \|z\|_2^2 - N \leq C_1\sqrt{2N}\}$. There exists a constant $p = p(C_1, C_2) > 0$ and integer $N_0 \geq 1$ such that for all $N \geq N_0$,*

$$\mathbb{P}\big(\mathcal{Z}_\pm\big) \geq p$$

Proof of Proposition 3.6.3. Define the following random variable, which is an affine transformation of a χ^2_N-distributed random variable:

$$X_N := \frac{\|z\|_2^2 - N}{\sqrt{2N}}.$$

By applying the central limit theorem for the collection $\{Z_i^2\}_{i=1}^N$ one has that X_N converges in distribution to the standard normal distribution: for every $t \in \mathbb{R}$,

$$\lim_{N\to\infty} \mathbb{P}\left(X_N \leq t\right) = \Phi(t),$$

where Φ is the standard normal cdf. In particular, for any $\varepsilon, t > 0$ there is $N_0 \in \mathbb{N}$ such that

$$\big|\mathbb{P}\big(X_N \leq t\big) - \Phi(t)\big| \leq \varepsilon$$

for all $N \geq N_0$. As such, one need merely choose $\varepsilon > 0$ so that

$$\mathbb{P}\big(\mathcal{Z}_\pm\big) \geq \Phi(C_1) - \Phi(C_2) - 2\varepsilon =: p(C_1, C_2) > 0$$

and choose the first N_0 for which the chain of inequalities is valid for all $N \geq N_0$. □

Corollary 3.6.4. *Fix $N, N_0 \in \mathbb{N}$ with $N \geq N_0 \geq 2$. Let $z \in \mathbb{R}^N$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and define the event*

$$A_N := \big\{\|z\|_2^2 \leq N - 2\sqrt{N} \quad \& \quad \|z\|_\infty \leq \sqrt{3\log N}\big\}$$

There exists a real constant $C = C(N_0) > 0$ such that $\mathbb{P}(A_N) \geq C$.

Proof of Corollary 3.6.4. Given N, define the events $E_N := \{\|z\|_2^2 \leq N - 2\sqrt{N}\}$ and $F_N := \{\|z\|_\infty \leq \sqrt{3 \log N}\}$. Using the standard identity $\Phi(-x) \leq \phi(x)/x$, we note that

$$\mathbb{P}(F_N) \geq 1 - 2N\mathbb{P}(|Z| > \sqrt{3 \log N}) \geq 1 - \frac{2}{\sqrt{\frac{3}{2}\pi N_0 \log N_0}} > 0$$

With this, and noting that $\mathbb{P}(E_N \mid F_N) \geq \mathbb{P}(E_N)$, one has

$$\mathbb{P}(A_N) = \mathbb{P}(E_N \cap F_N) = \mathbb{P}(E_N \mid F_N)\mathbb{P}(F_N) \geq \mathbb{P}(E_N)\mathbb{P}(F_N) \geq C \geq 0.$$

□

We also recall that an event holding with high probability, intersected with an event occurring with constant probability, still occurs with constant probability.

Proposition 3.6.5. *Let $N \geq 1$ be an integer and suppose that $\mathcal{E} = \mathcal{E}(N)$ is an event that holds with high probability in the sense that*

$$\mathbb{P}(\mathcal{E}_N) \geq 1 - p(N)$$

for some function $p(N) > 0$ with $\lim_{N\to\infty} p(N) = 0$. Suppose also that for an event $\mathcal{F} = \mathcal{F}(N)$ there exists $q > 0$ such that $\inf_{N\geq 1} \mathbb{P}(\mathcal{F}(N)) \geq q$. Then there exists a constant $q' > 0$ and integer $N_0 \geq 1$ such that $\mathbb{P}\big(\mathcal{E}(N) \cap \mathcal{F}\big) \geq q'$ for all $N \geq N_0$.

Proof of Proposition 3.6.5. The proof is very similar to that of Proposition 3.6.3. Simply choose a threshold $\varepsilon > 0$ and select the first $N_0 \geq 1$ for which

$$\mathbb{P}\big(\mathcal{E}(N) \cap \mathcal{F}\big) \geq q - p(N) \geq q - \varepsilon =: q' > 0$$

for all $N \geq N_0$. □

Remark 8. An example of such a $p(N)$ as in Proposition 3.6.5 is $p(N) \sim O(e^{-N})$ when $\mathcal{E}_N := \{|X - \mu| \leq t\}$ for X a subgaussian random variable, $\mathbb{E}\, X = \mu$ and $t > 0$.

3.6.2 Proof for (LS^*_τ) parameter sensitivity

Proof of Theorem 3.2.1. Let $x_0 \in \Sigma^N_s$ with non-empty support and let $\tau > 0$ be the governing parameter of (LS^*_τ). First suppose the parameter is chosen smaller than the optimal value, i.e., $\tau < \|x_0\|_1$. The discrepancy of the guess, $\rho := |\|x_0\|_1 - \tau| = \|x_0\|_1 - \tau > 0$, induces the sensitivity.

The solution $\hat{x}(\tau)$ to (LS^*_τ) satisfies $0 \leq \|\hat{x}(\tau)\|_1 \leq \tau$ by construction. Therefore, by the Cauchy-Schwarz inequality and an application of the triangle inequality,

$$\|\hat{x}(\tau) - x_0\|_2^2 \geq N^{-1}\|\hat{x}(\tau) - x_0\|_1^2 \geq \frac{\rho^2}{N} > 0.$$

Accordingly,

$$\lim_{\eta \to 0} \frac{1}{\eta^2}\|\hat{x}(\tau) - x_0\|_2^2 \geq \lim_{\eta \to 0} \frac{\rho^2}{N\eta^2} = \infty.$$

Next assume τ is chosen too large, with discrepancy between the correct and actual guesses for the parameter again being denoted $\rho = \tau - \|x_0\|_1 > 0$. Two key pieces of intuition guide this result. The first is that the error of approximation should be controlled by the effective dimension of the constraint set. The second suggests that y continues to lie within the constraint set for sufficiently small noise level, meaning recovery behaves as though it were unconstrained. Hence, the effective dimension of the problem is that of the ambient dimension, and so one should expect the error to be proportional to N.

First, we show that for η sufficiently small, $y \in \tau B_1^N$ with high probability. Fix a sequence $\eta_j \xrightarrow{j \to \infty} 0$ and define $y_j := x_0 + \eta_j z$. Since $\|z\|_1$ is subgaussian, Theorem 2.2.1 implies there is a constant $C > 0$ such that for

any $t > 0$,

$$\mathbb{P}\big(\|z\|_1 \geq t + N\sqrt{\frac{2}{\pi}}\big) \leq \mathbb{P}\big(\big|\|z\|_1 - N\sqrt{\frac{2}{\pi}}\big| \geq t\big) \leq e \cdot \exp\big(\frac{-t^2}{CN}\big).$$

In order to satisfy $x_0 + \eta z \in \tau B_1^N$, we need $\|x_0 + \eta z\|_1 < \tau$, for which $\eta\|z\|_1 < \rho$ is sufficient. We select a $t > 0$ so that the probability this event does not occur may be upper bounded as

$$\mathbb{P}\big(\|z\|_1 \geq \frac{\rho}{\eta}\big) \leq \mathbb{P}\big(\|z\|_1 \geq t + N\sqrt{\frac{2}{\pi}}\big) \leq e \cdot \exp\big(-\frac{t^2}{CN}\big).$$

For $t = \rho/\eta - N\sqrt{\frac{2}{\pi}}$, and $\tilde{C} > 0$ a new constant,

$$\begin{aligned}\mathbb{P}\left(\|z\|_1 \geq \frac{\rho}{\eta}\right) &\leq e \cdot \exp\left(-\frac{\big(\rho/\eta - N\sqrt{2/\pi}\big)^2}{CN}\right) \\ &\leq \tilde{C}\exp\left(-\frac{\rho^2}{CN\eta^2}\right) \xrightarrow{\eta\to 0} 0.\end{aligned}$$

Let $E_j := \{\|z\|_1 < \frac{\rho}{\eta_j}\}$ for $j \geq 1$; their respective probabilities lower-bounded by $p_j := 1 - \tilde{C}\exp(-\rho^2/N\eta_j^2)$. Given $0 < \varepsilon \ll 1$, denote by j_0 the first integer such that $p_j \geq 1-\varepsilon$ for all $j \geq j_0$. On E_j with $j \geq j_0$, $y_j \in \tau B_1^N$ so y_j is the unique minimizer of (LS_τ^*), meaning:

$$\frac{1}{\eta^2}\|\hat{x}(\tau) - x_0\|_2^2 = \|z\|_2^2.$$

The result follows by bounding the following expectations:

$$\begin{aligned}&\lim_{\eta\to 0}\frac{1}{\eta^2}\,\mathbb{E}\,\|\hat{x}(\tau) - x_0\|_2^2 \\ &= \lim_{j\to\infty}\mathbb{E}\left[\eta_j^{-2}\|\hat{x}(\tau) - x_0\|_2^2\,\chi(E_j)\right] + \mathbb{E}\left[\eta_j^{-2}\|\hat{x}(\tau) - x_0\|_2^2\,\chi(E_j^C)\right]. \qquad (\star)\end{aligned}$$

The first term converges by dominated convergence theorem:

$$\lim_{j\to\infty}\mathbb{E}\left[\eta_j^{-2}\|\hat{x}(\tau) - x_0\|_2^2\,\chi(E_j)\right] = \lim_{j\to\infty}\mathbb{E}\left[\,\chi(E_j)\|z\|_2^2\right] = \mathbb{E}\left[\|z\|_2^2\right] = N.$$

On E_j^C, $\|\hat{x}(\tau) - x_0\|_2^2 \leq \|\hat{x}(\tau) - x_0\|_1^2 \leq \rho^2\eta^2$, so by dominated convergence theorem,

$$\lim_{j\to\infty} \mathbb{E}\left[\eta_j^{-2}\|\hat{x}(\tau) - x_0\|_2^2\,\chi(E_j^C)\right] \leq \lim_{j\to\infty} \mathbb{E}\left[\|z\|_1^2\,\chi(E_j^C)\right] = 0.$$

This immediately yields the desired result,

$$\lim_{\eta\to 0} \frac{1}{\eta^2}\,\mathbb{E}\,\|\hat{x}(\tau) - x_0\|_2^2 = N.$$

To prove the final case where $\tau = \|x_0\|_1$, set $\mathcal{C} = B_1^N$ in (2.1) of Theorem 2.1.1. Then,

$$\lim_{\eta\to 0} \eta^{-2}\,\mathbb{E}\,\|\hat{x}(\tau) - x_0\|_2^2 = \mathbf{D}(T_{\mathbb{B}_1^N}(x_0)) = \Theta(s\log(N/s)) \ll N.$$

□

3.6.3 Proofs for (QP_λ^*) results

Proof of Proposition 3.3.1. Because z is isotropic and iid, one can split the signal $x_0 = x_0^+ - x_0^-$ into "positive" and "negative" components, and so it suffices to consider the case where $x_{0,j} \geq 0$ for all $j \in [N]$. The heart of this proposition again relies on the fact that the noise limits to 0. In general, $\lambda > 0$ is finite and typically small ($\sim \mathcal{O}(\eta\sqrt{\log N})$, so we require only that $|x_{0,j}| = \mathcal{O}(1)$ for $j \in T = \mathrm{supp}(x_0)$. This requirement can be written $x_{0,j} \geq a > 0$ for all $j \in T$ and some real number $a > 0$. Recall that the minimizer of (QP_λ^*) is given by the soft-thresholding operator which we denote by

$$x^\sharp(\eta\lambda) = S_{\eta\lambda}(x_0 + \eta z).$$

Where $k \in T, \ell \in T^C$ so that $x_{0,k} \geq a$, $x_{0,\ell} = 0$, one has

$$S_{\eta\lambda}(x_{0,k} + \eta z_k) - x_{0,k} = \begin{cases} \eta(z_k - \lambda) & x_{0,k} > \eta(\lambda - z_k) \\ -x_{0,k} & |x_{0,k} + \eta z_k| \leq \eta\lambda \\ \eta(z_k + \lambda) & x_{0,k} < -\eta(\lambda + z_k) \end{cases}$$

$$S_{\eta\lambda}(\eta z_\ell) = \begin{cases} \eta(z_\ell - \lambda) & z_\ell > \lambda \\ 0 & |z_\ell| \leq \lambda \\ \eta(z_\ell + \lambda) & z_\ell < -\lambda \end{cases}.$$

Independence of z_j yields

$$\begin{aligned} &\lim_{\eta\to 0} \frac{1}{\eta^2} \mathbb{E} \|x^\sharp(\eta\lambda) - x_0\|_2^2 \\ &= \lim_{\eta\to 0} \frac{s}{\eta^2} \mathbb{E}\left[(S_{\eta\lambda}(x_{0,k} + \eta z_k) - x_{0,k})^2\right] + \lim_{\eta\to 0} \frac{N-s}{\eta^2} \mathbb{E}\left[S_{\eta\lambda}(z_\ell)^2\right]. \end{aligned}$$

Passing to a sequence $\eta_j \to 0$, there exists $J \in \mathbb{N}$ such that for all $j \geq J$,

$$S_{\eta_j\lambda}(x_{0,k} + \eta_j z_k) - x_{0,k} = \eta_j(z_k - \lambda) \qquad \text{with high probability.} \qquad (\star)$$

If this equality were true almost surely, it would follow that for $k \in T$,

$$\eta^{-2} \mathbb{E} \|(x^\sharp(\eta\lambda) - x_0)_T\|_2^2 = s \mathbb{E}\left[(z_k - \lambda)^2\right] = s(1 + \lambda^2).$$

Indeed this is still true in the case of $(\star)$ with $\eta \to 0$. In particular, using independence of z_k for $k \in T$ and denoting by E_j the high probability event $(\star)$, we obtain by similar means as in the proof of Theorem 3.2.1,

$$\begin{aligned} &\lim_{\eta\to 0} \eta^{-2} \mathbb{E} \|(x^\sharp(\eta\lambda) - x_0)_T\|_2^2 \\ &= \lim_{j\to\infty} \frac{s}{\eta_j^2} \mathbb{E}\left[\eta_j^2 (z_k - \lambda)^2 \chi(E_j)\right] + \frac{s}{\eta_j^2} \mathbb{E}\left[(x_k^\sharp(\eta_j\lambda) - x_{0,k})^2 \chi(E_j^C)\right] \\ &= s(1 + \lambda^2) \end{aligned}$$

Next, define $G(\lambda) := (1+\lambda^2)\Phi(-\lambda) - \lambda\phi(\lambda)$. By independence of the entries of z_{T^C}, with any $\ell \in T^C$, the second quantity is exactly computable as

$$\lim_{\eta\to 0} \eta^{-2}\,\mathbb{E}\,\|(x^\sharp(\eta\lambda) - x)_{T^C}\|_2^2 = (N-s)\,\mathbb{E}\left[S_\lambda(z_\ell)^2\right] = 2(N-s)G(\lambda),$$

where the final equality is by definition of S_λ and elementary calculations (*cf.* Remark 10). Therefore, as desired,

$$\lim_{\eta\to 0} \eta^{-2}\,\mathbb{E}\,\|x^\sharp(\eta\lambda) - x\|_2^2 = s(1+\lambda^2) + 2(N-s)G(\lambda).$$

□

Proof of Corollary 3.3.2. For $0 \le t \le s$, where we define for simplicity of notation $\Sigma_{-1}^N := \emptyset$, observe that

$$\sup_{x_0 \in \Sigma_t^N \setminus \Sigma_{t-1}^N} R^\sharp(\lambda; x_0, N, \eta) = R^\sharp(\lambda; t, N)$$

because the regime $\eta \to 0$ is equivalent, by a rescaling argument, to the regime in which $\eta > 0$ and $|x_{0,j}| \to \infty$ for $j \in \mathrm{supp}(x_0)$ (as shown explicitly in the proof of Proposition 3.6.2). Therefore,

$$\begin{aligned}
\sup_{x_0 \in \Sigma_s^N} R^\sharp(\lambda; x_0, N, \eta) &= \max_{0\le t\le s} \sup_{x_0 \in \Sigma_t^N \setminus \Sigma_{t-1}^N} R^\sharp(\lambda; x_0, N, \eta) \\
&= \max\{R^\sharp(\lambda; 0, N), R^\sharp(\lambda; s, N)\} \\
&= R^\sharp(\lambda; s, N)
\end{aligned}$$

by linearity of the max argument and the fact that $1 + \lambda^2 \ge G(\lambda)$ for $\lambda > 0$. □

Proof of (QP_λ^*) parameter sensitivity

Proof of Lemma 3.3.3. By Proposition 3.3.1, $R^\sharp(\lambda; s, N) = s(1+\lambda^2) + 2(N-s)G(\lambda)$. We prove the result by controlling $G'(\lambda)$ using integration by parts.

Thus,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}G(\lambda) &= 2\lambda\Phi(-\lambda) - 2\phi(\lambda) \\ &\leq 2\lambda(\frac{1}{\lambda} - \frac{1}{\lambda^3} + \frac{3}{\lambda^5})\phi(\lambda) - 2\phi(\lambda) = -2\frac{\lambda^2-3}{\lambda^4}\phi(\lambda)\end{aligned}$$

A simple substitution yields, for all $N > \exp\left(\frac{3}{2}(1-\varepsilon)^{-2}\right)$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u}\bigg|_{u=1-\varepsilon} G(u\bar{\lambda}) &\leq \left[-2\frac{(u\bar{\lambda})^2-3}{u^4\bar{\lambda}^3}\phi(u\bar{\lambda})\right|_{u=1-\varepsilon} \\ &= -\frac{2(1-\varepsilon)^2\log(N)-3}{(1-\varepsilon)^4\sqrt{\pi\log^3(N)}}N^{-(1-\varepsilon)^2} =: -\frac{1}{2}\gamma(N,\varepsilon)N^{-(1-\varepsilon)^2}.\end{aligned}$$

Multiplying $G((1-\varepsilon)\bar{\lambda})$ by $N-s$ yields

$$\begin{aligned}&\left|\frac{\mathrm{d}}{\mathrm{d}u}R^\sharp(u\bar{\lambda};s,N)\right|_{u=1-\varepsilon} \geq (N-s)\gamma(N,\varepsilon)N^{-(1-\varepsilon)^2} - 2s(1-\varepsilon)\sqrt{2\log N} \\ &= \gamma(N,\varepsilon)N^{2\varepsilon-\varepsilon^2} - s\gamma(N,\varepsilon)N^{-(1-\varepsilon)^2} - 2s(1-\varepsilon)\sqrt{2\log N} \\ &\geq CN^{\varepsilon}\end{aligned}$$

for some constant $C > 0$ under the condition that $N \geq N_0$, where $N_0 > \exp\left(\frac{3}{2}(1-\varepsilon)^{-2}\right)$ is chosen so that for all $N \geq N_0$ the following two conditions are satisfied:

$$\begin{cases}(N-s)\gamma(N,\varepsilon)N^{-(1-\varepsilon)^2} \geq 2s(1-\varepsilon)\sqrt{2\log N} \\ \gamma(N,\varepsilon)\big(1-\frac{s}{N}\big) \geq 2s(1-\varepsilon)N^{-2\varepsilon+\varepsilon^2}\sqrt{2\log N} + CN^{-\varepsilon+\varepsilon^2}\end{cases}$$

In this regime, one achieves unbounded growth of the risk as a power law of the ambient dimension. □

Remark 9. Using integration by parts, one has for $x > 0$,

$$\begin{aligned}\Phi(-x) = \int_x^\infty \phi(t)\,\mathrm{d}t &= \Big(\frac{1}{x} - \frac{1}{x^3}\Big)\phi(x) + 3\int_x^\infty \frac{t\phi(t)}{t^5}\,\mathrm{d}t \\ &\le \Big(\frac{1}{x} - \frac{1}{x^3}\Big)\phi(x) + 3x^{-5}\int_x^\infty t\phi(t)\,\mathrm{d}t = \Big(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5}\Big)\phi(x)\end{aligned}$$

Proof of Theorem 3.3.4. Define $f(u) := \frac{\mathrm{d}}{\mathrm{d}u} R^\sharp(u\bar{\lambda}; s, N)$ and write $F(u) := R^\sharp(u\bar{\lambda}; s, N)$ for its anti-derivative. The proof is an application of the fundamental theorem of calculus:

$$F(1) - F(1-\varepsilon) = \int_0^\varepsilon f(1-t)\,\mathrm{d}t \le -C\int_0^\varepsilon N^t\,\mathrm{d}t = C\frac{1-N^\varepsilon}{\log N}.$$

The result follows by substituting:

$$R^\sharp((1-\varepsilon)\bar{\lambda}; s, N) \ge C\frac{N^\varepsilon - 1}{\log N} + R^\sharp(\bar{\lambda}; s, N) \ge C\frac{N^\varepsilon}{\log N}$$

where the latter inequality holds after taking N sufficiently large, and $C > 0$ is a universal constant that has changed values in the final expression. □

Proof of Proposition 3.3.5. By Proposition 3.3.1, $R^\sharp(\lambda; s, N) = s(1+\lambda^2) + 2(N-s)G(\lambda)$. We prove the result by controlling $G'(\lambda)$. One may lower bound $G'(\lambda)$ as

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}G(\lambda) = 2\lambda\Phi(-\lambda) - 2\phi(\lambda) \ge 2\lambda\Big(\frac{\lambda}{\lambda^2+1}\Big)\phi(\lambda) - 2\phi(\lambda) = -2\frac{\phi(\lambda)}{\lambda^2+1}.$$

This gives the following lower bound for $\frac{\mathrm{d}}{\mathrm{d}\lambda}R^\sharp(\lambda; s, N)$:

$$\begin{aligned}\frac{\mathrm{d}R^\sharp}{\mathrm{d}\lambda}(\lambda; s, N) &\ge 2s\lambda - 4(N-s)\frac{\phi(\lambda)}{\lambda^2+1} \ge 2\lambda - 4N\frac{\phi(\lambda)}{\lambda^2+1} \\ &= \frac{2}{\lambda^2+1}\big(\lambda(\lambda^2+1) - 2N\phi(\lambda)\big).\end{aligned}$$

Substituting $\bar{\lambda}$ gives a positive quantity, since $N \ge 2$:

$$\frac{2}{2\log N + 1}\big(\sqrt{2\log N}(2\log N + 1) - \frac{2}{\sqrt{2\pi}}\big) > 0.$$

Consequently, $\lambda^* < \bar{\lambda}$ because λ^* is the value giving optimal risk and $\frac{\mathrm{d}}{\mathrm{d}\lambda} R^\sharp(\lambda; s, N)$ is increasing for all $\lambda \geq \bar{\lambda}$. Then it must be that $|\lambda^* - \bar{\lambda}| < \varepsilon$ for any $\varepsilon > 0$ when N is sufficiently large. Indeed, fix $\varepsilon > 0$. By Lemma 3.3.3 there exists $N_0 \geq 1$ so that for all $N \geq N_0$ (QP^*_λ) is parameter unstable for $\lambda < \bar{\lambda}$, yielding $R^\sharp(\lambda; s, N) \gtrsim N^\varepsilon$. But $R^\sharp(\lambda^*; s, N) \leq C R^*(s, N)$ for $N \geq N_0$ by Proposition 3.1.5, where we re-choose $N_0 = N_0(s)$ if necessary. Thus, it must be that $|\lambda^* - \bar{\lambda}| < \varepsilon$ for all $N \geq N_0$. In particular, $\lim_{N\to\infty} \bar{\lambda}/\lambda^* = 1$. □

Remark 10. One may derive the following lower bound using integration by parts:

$$\Phi(-\lambda) \geq \frac{\lambda}{\lambda^2 + 1} \phi(\lambda).$$

Let $Z \sim \mathcal{N}(0, 1)$ be a standard normal random variable and let $S_\lambda(\cdot)$ denote soft-thresholding by $\lambda > 0$. Then,

$$0 \leq \mathbb{E}\left[S_\lambda(Z)^2\right] = 2 \int_\lambda^\infty (z - \lambda)^2 \phi(\lambda)\, \mathrm{d}z = 2(1 + \lambda^2)\Phi(-\lambda) - 2\lambda\phi(\lambda).$$

Thus, $(1 + \lambda^2)\Phi(-\lambda) \geq \lambda\phi(\lambda)$, giving the desired lower bound.

Proof of (QP^*_λ) right-sided stability

A more general version of Theorem 3.3.6 for general convex regularizers appears in Oymak and Hassibi [55]; and a related version for CS appears in Bickel et al. [13]. The present case allows for a simpler, direct proof to be given, proceeding by (3.1).

Proof of Theorem 3.3.6. Given $L = \lambda/\lambda^* > 1$, define $\bar{L} = \bar{L}(s, N) > 0$ by $\lambda = \bar{L}\bar{\lambda} = \bar{L}\sqrt{2 \log N}$. Note $\lim_{N\to\infty} \bar{L}(s, N) = L$, because $\bar{\lambda}$ is asymptotically equivalent to λ^* up to constants. A direct substitution of $\lambda = \bar{L}\bar{\lambda} = \bar{L}\sqrt{2 \log N}$ in the analytic formula for $R^\sharp(\lambda; s, N)$ yields the desired bound, noting that $R^\sharp(\lambda^*; s, N)$ equals $R^*(s, N)$ up to constants. Thus, there is

$C > 0$ and $N_0 = N_0(s) \geq 2$ so that for all $N \geq N_0$

$$R^\sharp(\lambda; s, N) \leq s(1 + 2\bar{L}^2 \log N) + \frac{N - s}{\bar{L} N^{\bar{L}^2} \sqrt{\pi \log N}} \leq C L^2 R^*(s, N).$$

□

3.6.4 Proofs for (BP^*_σ) parameter sensitivity

Proof of underconstrained (BP^*_σ) suboptimality

Proof of Lemma 3.4.1. By scaling properties of the risk, it suffices to consider the case where $\eta = 1$. Define the event

$$A_N := \big\{ \|z\|_2^2 \leq N - 2\sqrt{N} \quad \& \quad \|z\|_\infty \leq \sqrt{3 \log N} \big\}.$$

On A_N, it follows from the KKT conditions, where $h = \tilde{x}(\sigma) - x_0$, that

$$N \leq \sigma^2 = \|h\|_2^2 - 2\langle h, z \rangle + \|z\|_2^2 \leq \|h\|_2^2 - 2\langle h, z \rangle + N - 2\sqrt{N}.$$

By Cauchy-Schwartz and definition of A_N,

$$\frac{1}{2}\|h\|_2^2 \geq \sqrt{N} + \langle h, z \rangle \geq \sqrt{N} - \|h\|_1 \|z\|_\infty \geq \sqrt{N} - \|h\|_1 \sqrt{3 \log N}.$$

Applying Remark 3 and the binomial inequality $2ab \leq a^2 + b^2$ gives

$$\begin{aligned} \sqrt{N} - \|h\|_1 \sqrt{3 \log N} &\geq \sqrt{N} - 2\sqrt{s} \|h\|_2 \sqrt{3 \log N} \\ &\geq \sqrt{N} - \frac{1}{2}\|h\|_2^2 - 6s \log N. \end{aligned}$$

Combining these two groups of inequalities gives $\|h\|_2^2 \geq \sqrt{N} - 6s \log N$. Hence, by Bayes' rule and Corollary 3.6.4 there exist dimension independent constants $C, C' > 0$ such that

$$\begin{aligned} \mathbb{E}\, \|\tilde{x}(\sigma) - x_0\|_2^2 &\geq \mathbb{P}(A_N) \cdot \mathbb{E}\left[\|\tilde{x}(\sigma) - x_0\|_2^2 \mid A_N \right] \\ &\geq C' \big(\sqrt{N} - 6s \log N \big) \geq C \sqrt{N}. \end{aligned}$$

The final inequality follows by the assumption that $N \geq N_0(s)$. $\square$

Supporting propositions for the geometric lemma

This section is dedicated to several results necessary for the proof of Lemma 3.4.3, a main lemma in the proofs of Theorem 3.4.6 and Theorem 3.4.9. We state and prove these propositions in line.

Proposition 3.6.6. *Fix $C_1 > 0$. Let $K_1 := \lambda B_1^N \cap \alpha_1 B_2^N$ with $\alpha_1 := a_1 N^{1/4}$ and $\lambda := L\sqrt{\frac{N}{\log N}}$. There exists a choice of universal constants $a_1 > 0$, $L > 1$ and $N_0 = N_0^{(3.6.6)}(a_1, C_1, L) \geq 1$ satisfying*

$$N_0^{(3.6.6)}(a_1, C_1, L) := D_1^{2/(2D_2-1)}, \quad \text{where}$$
$$D_1 := \frac{a_1^2}{5L^2} < 1, \qquad D_2 := 2\Big(\frac{C_1 + a_1^2}{L^2}\Big)^2 < \frac{1}{2},$$

so that for all $N > N_0$,

$$\mathrm{w}(K_1) \geq (\frac{a_1^2 + C_1}{2})\sqrt{N}.$$

Proof of Proposition 3.6.6. Since $\mathrm{w}(K_1) = \mathbb{E}_z \sup_{q \in K_1} \langle q, z \rangle$ is the Gaussian mean width of K_1, we may invoke Proposition 2.3.6 to obtain a sufficient chain of inequalities:

$$\mathbb{E} \sup_{K_1} \langle q, z \rangle = gmw(K_1) \overset{(2.3.6)}{\geq} \frac{\sqrt{2}}{4}\kappa\lambda\sqrt{\log\big(\frac{N\alpha_1^2}{5\lambda^2}\big)} \overset{(*)}{\geq} \Big(\frac{\alpha_1^2 + C_1}{2}\Big)\sqrt{N}.$$

In particular, Proposition 2.3.6 holds with $\kappa = 1$, since κ is the lower-RIP constant of the sensing matrix for (BP_σ^*), which is the identity. We thus turn our attention to $(*)$, which is equivalent to

$$\log\big(D_1\sqrt{N}\log N\big) \geq D_2 \log N, \qquad D_1 := \frac{a_1^2}{5L^2}, D_2 := 2\Big(\frac{C_1 + a_1^2}{L^2}\Big)^2.$$

Rearranging gives

$$\frac{1}{2} + \frac{\log D_1 + \log\log N}{\log N} \geq D_2,$$

and for $D_1, 2D_2 \leq 1$, this is certainly satisfied for $N \geq D_1^{2/(2D_2-1)}$ (e.g., $L = 11$ imposes $N \gtrsim 10^5$ when $a_1 = 1, C_1 = 2$). Accordingly, it suffices to choose $N_0 = N_0(a_1, C_1, L)$ as in the proposition statement so that for all $N \geq N_0$, as desired,

$$\mathrm{w}(K_1) \geq \Big(\frac{a_1^2 + C_1}{2}\Big)\sqrt{N}.$$

□

Proposition 3.6.7. *Fix $\delta > 0$, $c \in (0,1)$. Let $K_1 = \lambda B_1^N \cap \alpha_1 B_2^N$ be as defined above. There are universal constants $\tilde{D}_1 > 0, N_0 \geq 2$ such that for*

$$N > N_0 := N_0^{(3.6.7)}(c, \tilde{D}_1, \delta, L) := \Big(\frac{1}{\tilde{D}_1 L^2 (1-c)^2} \log\big(\frac{1}{\delta}\big)\Big)^2$$

there exists $q \in K_1$ such that $\langle q, z\rangle \geq c\,\mathrm{w}(K_1)$ with probability at least $1 - \delta$, where $z \in \mathbb{R}^N$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$.

Proof of Proposition 3.6.7. Note that $K_1 \subseteq \mathbb{R}^N$ is a topological space and define the centered Gaussian process $f_x := \langle x, g\rangle$ for $g_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Observe that $\|f\|_{K_1} := \sup_{x \in K_1} |f_x|$ is almost surely finite. For any $u > 0$,

$$\mathbb{P}\big(\sup_{x \in K_1} |\langle x, g\rangle| < \mathrm{w}(K_1) - u\big) \leq \exp\big(-\frac{u^2}{2\sigma_{K_1}^2}\big).$$

by Theorem 2.2.4. Therefore, for $c \in (0,1)$,

$$\begin{aligned}\mathbb{P}\big(\sup_{x \in K_1} |\langle x, g\rangle| < c\,\mathrm{w}(K_1)\big) &\leq \exp\big(-\frac{(1-c)^2\,\mathrm{w}^2(K_1)}{2\sigma_{K_1}^2}\big) \\ &\leq \exp\big(-\frac{(1-c)^2 L^2 \sqrt{N}\log(D_1\sqrt{N}\log N)}{16\log N}\big) \leq \delta\end{aligned}$$

because

$$\sigma_{K_1}^2 = \sup_{x\in K_1} \mathbb{E}\,|\langle x, g\rangle|^2 = \sup_{x\in K_1} \sum_{i=1}^N x_i^2\, \mathbb{E}\,|g_i|^2 = \sup_{x\in K_1} \|x\|_2^2 = \alpha_1^2 = \sqrt{N}.$$

A specific choice of $q \in K_1$ follows by choosing the $q \in K_1$ that realizes the supremum, since K_1 is closed.

□

Proposition 3.6.8. *Fix $C_1, \delta > 0$ and define the event $\mathcal{Z}_- := \{\|z\|_2^2 \le N + C_1\sqrt{N}\}$ for $z \in \mathbb{R}^N$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$. There is a universal constant $N_0 = N_0^{(3.6.8)} \ge 1$ satisfying*

$$N_0^{(3.6.8)} \ge \max\{N_0^{(3.6.6)}(a_1, C_1, L), N_0^{(3.6.7)}(c, \tilde{D}_1, \delta, L)\},$$

and a universal constant $k_1 = k_1(N_0^{(3.6.8)}, \delta) > 0$ so that for all $N \ge N_0$ there is an event $\mathcal{E} \subseteq \mathcal{Z}_-$ satisfying

$$K_1 \cap F \ne \emptyset \text{ on } \mathcal{E} \qquad \text{and} \qquad \mathbb{P}(\mathcal{E}) \ge \mathbb{P}(\mathcal{Z}_-) - \delta.$$

Proof of Proposition 3.6.8. By Proposition 3.6.7, for any $c_1 \in (0,1)$ there is an event $\mathcal{E}_1$ that holds with high probability such that $\sup_{q\in K_1}\langle q, z\rangle \ge c_1\, \mathrm{w}(K_1)$ on $\mathcal{E}_1$. Subsequent statements are made on the restriction to $\mathcal{E}_1$.

As K_1 is closed, there is $q \in K_1$ realizing the supremum, whence $\langle q, z\rangle \ge c_1\, \mathrm{w}(K_1)$. Now, choose $C_1' > 0$ such that $C_1 \ge c_1^{-1}\big(a_1^2 + C_1\big) - a_1^2$. Then $q \in K_1$ satisfies

$$\langle q, z\rangle \ge c_1\, \mathrm{w}(K_1) \ge c_1\Big(\frac{a_1^2 + C_1'}{2}\Big)\sqrt{N} \ge \Big(\frac{a_1^2 + C_1}{2}\Big)\sqrt{N}.$$

Now, because $\|q\|_2 \le \alpha_1$ and $q \in K_1$, it holds on the event $\mathcal{E}_1 \cap \mathcal{Z}_-$ that

$$\Big(\frac{a_1^2 + C_1}{2}\Big)\sqrt{N} \ge \frac{1}{2}\|q\|_2^2 + \frac{1}{2}\big(\|z\|_2^2 - N\big).$$

Combining the two previous chains of inequalities implies

$$\|q - z\|_2^2 \leq N.$$

Namely, there exists an event $\mathcal{Z}_- \cap \mathcal{E}_1$, such that $q \in K_1 \cap F$, so long as $N \geq N_0^{(3.6.8)}$. Because $\mathcal{E}_1$ holds with high probability and the probability of $\mathcal{Z}_-$ is lower-bounded by a universal constant, Proposition 3.6.5 implies $\mathbb{P}(\mathcal{Z}_- \cap \mathcal{E}_1) \geq k_1(N_0^{(3.6.8)}, \delta)$ for $N \geq N_0^{(3.6.8)}$, where

$$N_0^{(3.6.8)} \geq \max\{N_0^{(3.6.6)}(a_1, C_1, L), N_0^{(3.6.7)}(c, \tilde{D}_1, \delta, L)\}.$$

□

Proposition 3.6.9. *Fix $C_2 > 0$ and let $L \geq 1$. Set $K_2 := \lambda B_1^N \cap \alpha_2 B_2^N$, where $\lambda = L\sqrt{\frac{N}{\log N}}$. There is a maximal choice of $\alpha_2 = \alpha_2(N) > 0$ so that for all $N \geq 1$,*

$$\mathrm{w}(K_2) \leq \frac{C_2}{2}\sqrt{N}.$$

Proof of Proposition 3.6.9. Since $\mathrm{w}(K_2) = \mathbb{E}_z \sup_{q \in K_2} \langle q, z \rangle$ is the Gaussian mean width of K_2, we may invoke Proposition 2.3.5 to obtain a sufficient chain of inequalities:

$$\mathrm{w}(K_2) \overset{(2.3.5)}{\leq} 4\lambda\sqrt{\log(\frac{4eN\alpha_2^2}{\lambda^2})} \overset{(**)}{\leq} \frac{C_2}{2}\sqrt{N}.$$

The first inequality follows by Proposition 2.3.5 immediately. Rearranging and substituting for λ, $(**)$ is equivalent to

$$D_3 \log N \geq \log\left(D_4 \alpha_2^2 \log N\right), \qquad D_3 := \left(\frac{C_2}{8L}\right)^2, D_4 := \frac{4e}{L^2}.$$

This inequality is satisfied for any α_2 with

$$\alpha_2^2 \leq \frac{N^{D_3}}{D_4 \log N} =: A^2(C_2, N).$$

For example, one may choose

$$\alpha_2 = \frac{LN^{D_5}}{2\sqrt{e\log N}}, \qquad D_5 := \frac{C_2^2}{32L^2}.$$

For such $0 < \alpha_2 \le A(N; C_2, L)$, it holds as desired that $\mathrm{w}(K_2) \le \frac{C_2}{2}\sqrt{N}$.

□

Remark 11. Notice that we want to choose N_0 so that $A(C_2, N)$ is increasing for all $N \ge N_0$. A quick calculation reveals that $N_0 = N_0^{(3.6.9)}(C_2, L) := \exp\big((2D_5)^{-1}\big)$ is sufficient.

Proposition 3.6.10. *Fix $\delta > 0$, and $C > 1$. Let $K_2 = \lambda B_1^N \cap \alpha_2 B_2^N$ as above. There are universal constants $\tilde{D}_2 > 0, N_0 \ge 1$ such that for*

$$N > N_0 := N_0^{(3.6.10)}(C, \tilde{D}_2, \delta, L) := \Big(\frac{1}{\tilde{D}_2 L^2 (C-1)^2}\log\big(\frac{1}{\delta}\big)\Big)^2$$

one has $\sup_{q\in K_2}\langle q, z\rangle \le C\,\mathrm{w}(K_2)$ with probability at least $1-\delta$, where $z \in \mathbb{R}^N$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$.

Proof of Proposition 3.6.10. Define the centered Gaussian process $f_x := \langle x, g\rangle$ for $x \in K_2 \subseteq \mathbb{R}^N$, a topological space, and where $g_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Observe $\|f\|_{K_2} = \sup_{x\in K_2}|f_x| < \infty$ almost surely. For any $u > 0$,

$$\mathbb{P}\Big(\sup_{x\in K_2}|\langle x, g\rangle| > \mathrm{w}(K_2) + u\Big) \le \exp\Big(-\frac{u^2}{2\sigma_{K_2}^2}\Big)$$

by Theorem 2.2.4. Hence, for $C > 1$,

$$\begin{aligned}\mathbb{P}\Big(\sup_{x\in K_2}|\langle x, g\rangle| > C\,\mathrm{w}(K_2)\Big) &\le \exp\Big(-\frac{(C-1)^2\,\mathrm{w}^2(K_2)}{2\sigma_{K_2}^2}\Big)\\ &\le \exp\Big(-\frac{(C-1)^2 L^2 N\log(D_1\sqrt{N}\log N)}{16\alpha_2^2\log N}\Big) \le \delta\end{aligned}$$

because

$$\sigma_{K_2}^2 = \sup_{x\in K_2} \mathbb{E}\,|\langle x, g\rangle|^2 = \sup_{x\in K_2} \sum_{i=1}^N x_i\, \mathbb{E}\,|g_i|^2 = \sup_{x\in K_2} \|x\|_2^2 = \alpha_2^2 \le \alpha_1^2 = \sqrt{N}.$$

Finally, for $\delta > 0$ and $C > 1$, $\sup_{x\in K_2} |\langle x, g\rangle| \le C\,\mathrm{w}(K_2)$ with probability at least $1-\delta$ for any $N \ge N_0^{(3.6.10)}$.

□

Proposition 3.6.11. *Fix $C_2, \delta > 0$ and define the event $\mathcal{Z}_+ := \{\|z\|_2^2 \ge N + C_2\sqrt{N}\}$ where $z \in \mathbb{R}^N$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$. There is a universal constant $N_0 := N_0^{(3.6.11)} \ge 1$ satisfying*

$$N_0^{(3.6.11)} \ge \max\left\{N_0^{(3.6.9)}, N_0^{(3.6.10)}\right\}.$$

and a universal constant $k_2 = k_2(N_0, \delta) > 0$ so that for all $N \ge N_0$ there is an event $\mathcal{E} \subseteq \mathcal{Z}_+$ satisfying

$$K_2 \cap F = \emptyset \text{ on } \mathcal{E} \qquad \text{and} \qquad \mathbb{P}(\mathcal{E}) \ge k_2 := \mathbb{P}(\mathcal{Z}_+) - \delta.$$

Proof of Proposition 3.6.11. By Proposition 3.6.10, for any $0 < c_2 < 1$ there is an event $\mathcal{E}_2$ that holds with high probability such that $\sup_{q\in K_2}\langle q, z\rangle \le c_2\,\mathrm{w}(K_2)$ on $\mathcal{E}_2$. Because K_2 is closed, there is $q \in K_2$ realizing the supremum when restricted to $\mathcal{E}_2$, whence

$$\langle q, z\rangle \le \sup_{q'\in K_2} \langle q', z\rangle \le c_2\,\mathrm{w}(K_2).$$

Now, choose $C_2' > 0$ such that $0 \le C_2 \le c_2 C_2'$. Then $q \in K_2$ satisfies

$$\langle q, z\rangle \le c_2\,\mathrm{w}(K_2) \le c_2\frac{C_2'}{2}\sqrt{N} \le \frac{C_2}{2}\sqrt{N}.$$

On the other hand, for any $q' \in F$ on the event $\mathcal{Z}_+$,

$$C_2\sqrt{N} \le \|q'\|_2^2 + \|z\|_2^2 - N \le 2\langle q', z\rangle,$$

whence $K_2 \cap F = \emptyset$ on the event $\mathcal{Z}_+ \cap \mathcal{E}_2$. Because $\mathcal{E}_2$ holds with high probability and the probability of $\mathcal{Z}_+$ is lower-bounded by a universal constant, Proposition 3.6.5 implies $\mathbb{P}\big(\mathcal{Z}_+ \cap \mathcal{E}_2\big) > k_2(N_0^{(3.6.11)}, \delta)$ for $N \geq N_0^{(3.6.8)}$ where

$$N_0^{(3.6.11)} \geq \max\big\{N_0^{(3.6.9)}, N_0^{(3.6.10)}\big\}.$$

□

Proof of the geometric lemma

We now have the tools required for Lemma 3.4.3. A graphical depiction of the result appears in Figure 3.1 of § 3.4.2.

Proof of Lemma 3.4.3. We obtain the event $\mathcal{E}$ trivially from Proposition 3.6.8 and Proposition 3.6.11, choosing $\alpha_1 := a_1 N^{1/4}$ as in Proposition 3.6.6. Namely, define the event

$$\mathcal{E} := \mathcal{Z}_- \cap \mathcal{E}_1 \cap \mathcal{Z}_+ \cap \mathcal{E}_2.$$

Observe that $\mathbb{P}\big(\mathcal{E}\big) \geq \mathbb{P}\big(\mathcal{Z}_- \cap \mathcal{Z}_+\big) - 2\delta \geq k_3$ for all sufficiently large N. This is a direct consequence of Proposition 3.6.3 and Proposition 3.6.5. Proving that $\alpha_2 > C_3 N^q$ follows from a note in Proposition 3.6.9. Specifically, the result holds for any choice of α_2 satisfying

$$0 < \alpha_2 \leq A(N; C_2; L) = \frac{L N^{D_5}}{2\sqrt{e \log N}}, \quad D_5 := \frac{C_2^2}{32 L^2}.$$

Hence, choose $C_3, q > 0$ so that $\alpha_2 > C_3 N^q$ for all $N \geq N_0^{(3.4.3)} \geq N_0^{(3.6.9)}$.

□

Proofs for overconstrained suboptimality

We start by proving a key ingredient of the main results for parameter sensitivity of $\tilde{R}(\sigma; x_0 N, \eta)$. Then, we prove the lemma that extends (BP_σ^*) parameter sensitivity from $\sigma = \sqrt{N}$ and $x_0 \equiv 0$ to $\sigma \leq \sqrt{N}$ and $x_0 \equiv 0$. Fi-

nally, we prove the restricted maximin result, yielding PS for overconstrained (BP^*_σ).

Proof of Corollary 3.4.4. Restrict to the event $\mathcal{E}$ as given in the lemma and assume that $N \geq N_0^{(3.4.3)}$. $K_1 \cap F$ is non-empty, so $\tilde{x}(\sigma) \in K_1 \cap F$ by definition. $K_2 \cap F = \emptyset$ thereby implies

$$\tilde{x}(\sigma) \in \lambda B_1^N \cap \left(\alpha_1 B_2^N \setminus \alpha_2 B_2^N\right) \cap F = (K_1 \setminus K_2) \cap F.$$

Whence follows $\|\tilde{x}(\sigma)\|_1 \leq \lambda$ and $\alpha_2 \leq \|\tilde{x}(\sigma)\|_2 \leq \alpha_1$. Applying Bayes' rule to the noise-normalized risk yields:

$$\tilde{R}(\sigma; 0, N, \eta) \geq \frac{\mathbb{P}(\mathcal{E})}{\eta^2} \mathbb{E}\left[\|\tilde{x}(\sigma)\|_2^2 \mid \mathcal{E}\right] \geq k_3 C_3 N^q =: C N^q.$$

□

Proof of Lemma 3.4.5. This result is an immediate consequence of Corollary 2.1.3.

□

Proof of Theorem 3.4.2. Without loss of generality, assume $\eta = 1$. We may trivially lower-bound the minimax expression by considering only the case where $x_0 \equiv 0$,

$$\sup_{x_0 \in \Sigma_s^N} \inf_{\sigma \leq \sqrt{N}} \tilde{R}(\sigma; x_0, N, 1) \geq \inf_{\sigma \leq \sqrt{N}} \tilde{R}(\sigma; 0, N, 1).$$

Then, Lemma 3.4.5 and Corollary 3.4.4 imply in turn,

$$\inf_{\sigma \leq \sqrt{N}} \tilde{R}(\sigma; 0, N, \eta) \geq \tilde{R}(\sqrt{N}; 0, N, \eta) \geq C N^q$$

for all $N \geq N_0$, where $N_0 \geq N_0^{(3.4.3)}$ and $C, q > 0$ are chosen according to Lemma 3.4.3. □

Proof of (BP^*_σ) minimax suboptimality

In this section we prove Theorem 3.4.6, establishing a data regime in which (BP^*_σ) is asymptotically minimax suboptimal.

Proof of Theorem 3.4.6. Without loss of generality, take $\eta = 1$. Observe,

$$\inf_{\sigma>0} \sup_{x_0 \in \Sigma_s^N} \tilde{R}(\sigma; x_0, N, 1) = \min\left\{ \inf_{\sigma \le \sqrt{N}} S(\sigma), \inf_{\sigma > \sqrt{N}} S(\sigma) \right\}$$

where $S(\sigma) := \sup_{x_0 \in \Sigma_s^N} \tilde{R}(\sigma; x_0, N, 1)$. Next, assume $N \ge N_0^{(3.4.3)}$. Then one has $\inf_{\sigma > \sqrt{N}} S(\sigma) \ge C_1 \sqrt{N}$ by Lemma 3.4.1. Next, successively applying a trivial lower bound, Lemma 3.4.5, and Corollary 3.4.4 gives

$$\inf_{\sigma \le \sqrt{N}} S(\sigma) \ge \inf_{\sigma \le \sqrt{N}} \tilde{R}(\sigma; 0, N, 1) \ge \tilde{R}(\sqrt{N}; 0, N, 1) \ge C_2 N^q.$$

In particular, there is a universal constant $C > 0$ so that

$$\inf_{\sigma>0} \sup_{x_0 \in \Sigma_s^N} \tilde{R}(\sigma; x_0, N, 1) \ge \min\{C_2 N^q, C_1 \sqrt{N}\} \ge C N^q.$$

□

Proof of (BP^*_σ) maximin suboptimality

In this section we prove Theorem 3.4.9, establishing a data regime in which (BP^*_σ) is asymptotically maximin suboptimal. We begin with a proof of the key ingredient, Lemma 3.4.7.

Proof of Lemma 3.4.7. The proof is completed by the following chain of inequalities. The first and last equalities are by definition of the (BP^*_σ) estimator. The first inequality follows by relaxing the objective; the second

inequality follows by relaxing the constraint condition.

$$\begin{aligned}\|\tilde{x}_{T^C}\|_2 &= \left\| \arg\min\{\|x\|_1 : \|y-x\|_2^2 \le \sigma^2\}_{T^C} \right\|_2 \\ &\ge \left\| \arg\min\{\|x_{T^C}\|_1 : \|y-x\|_2^2 \le \sigma^2\}_{T^C} \right\|_2 \\ &\ge \left\| \arg\min\{\|x_{T^C}\|_1 : \|(y-x)_{T^C}\|_2^2 \le \sigma^2\}_{T^C} \right\|_2 \\ &\equiv \|\tilde{x}'\|_2\end{aligned}$$

□

Proof of Theorem 3.4.9. We may trivially lower-bound the maximin expression by considering the case where $x_0 := Ne_1$ where e_1 is the first standard basis vector. Without loss of generality, we may assume that this entry is in the first coordinate, and is at least N. Again without loss of generality, it suffices to consider the case where $\eta = 1$. We write the lower bound as

$$\sup_{x \in \Sigma_s^N} \inf_{\sigma > 0} \tilde{R}(\sigma; x, N, 1) \ge \inf_{\sigma > 0} \tilde{R}(\sigma; x_0, N, 1).$$

If $\sigma \ge \sqrt{N}$, then the result follows by Lemma 3.4.1. Otherwise, it must be that $\sigma \le \sqrt{N}$, in which case the result follows immediately by Lemma 3.4.7. In this latter case, we have implicitly assumed that if $\sigma \in (\sqrt{N-1}, \sqrt{N})$, then the omitted technical exercise of adjusting constants in Corollary 3.4.4 and its constituents has been carried out. For further detail on this caveat, see the remark immediately succeeding Corollary 3.4.8. □

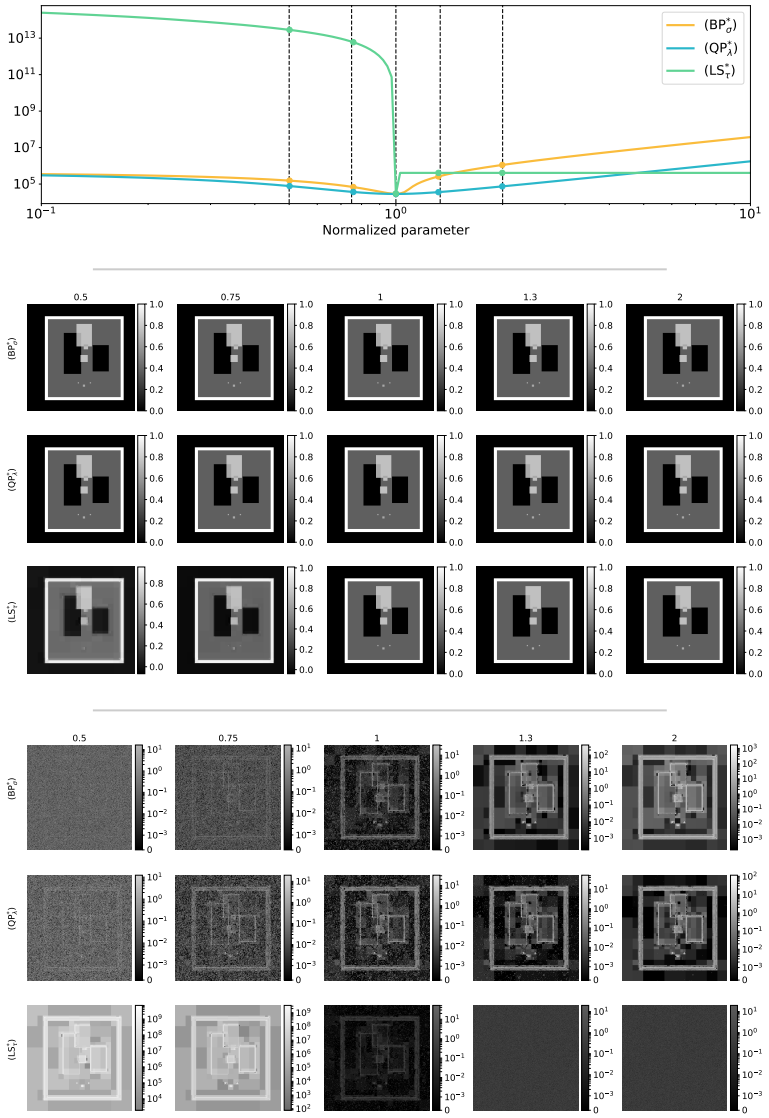


Figure 3.12: Wavelet space denoising of the square Shepp-Logan phantom for different values of the normalized parameter when $\eta = 10^{-5}$. **Top:** The sections of the loss for which estimator recovery will be visualized are depicted by the dots which lie nearly on the blacked dotted lines, themselves located at $\rho = 0.5, 0.75, 1, 4/3, 2$. **Middle:** This group of fifteen plots represents a program's solution for a particular value of the normalized parameter, arranged in a grid. Image pixel values are not scaled to $[0, 1]$; their range is given by the associated colour bar. **Bottom:** This group of fifteen plots depicts pixel-wise nnse for each (program, normalized parameter) pairing. In both the middle and bottom groups, the program is denoted along the left-hand side, while the normalized parameter value is denoted along the top row of each group.

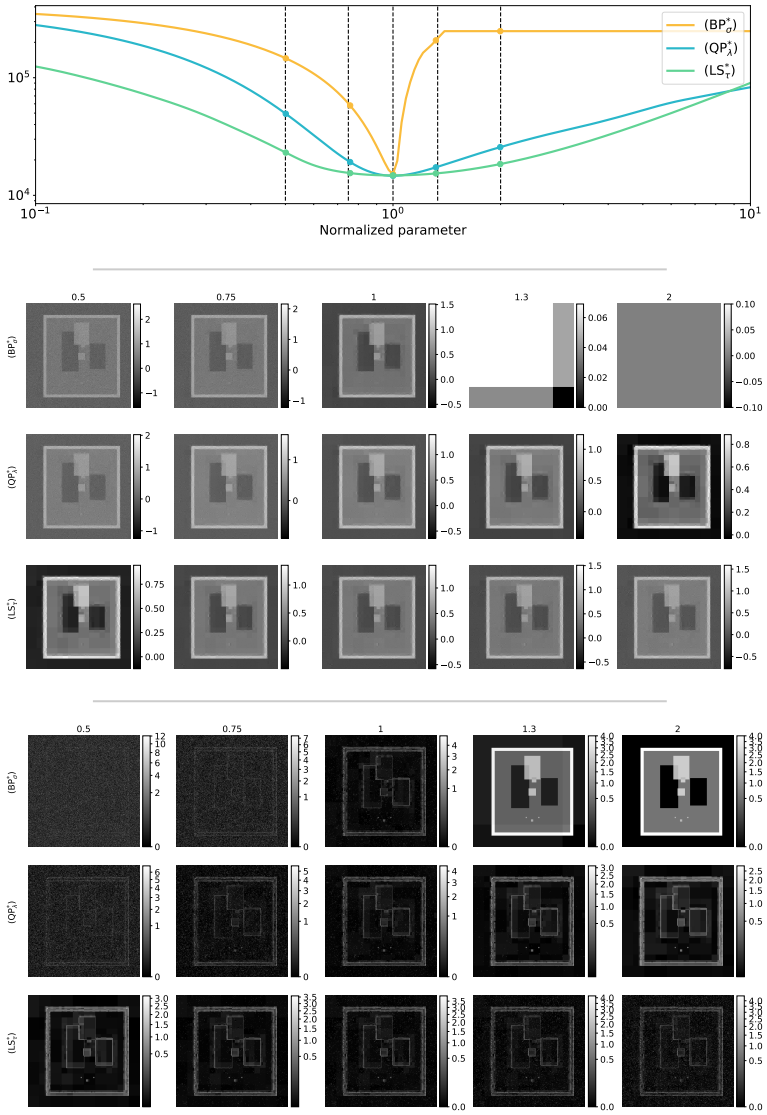


Figure 3.13: Wavelet space denoising of the square Shepp-Logan phantom for different values of the normalized parameter when $\eta = 0.5$. **Top:** The sections of the loss for which estimator recovery will be visualized are depicted by the dots which lie nearly on the blacked dotted lines, themselves located at $\rho = 0.5, 0.75, 1, 4/3, 2$. **Middle:** This group of fifteen plots represents a program's solution for a particular value of the normalized parameter, arranged in a grid. Image pixel values are not scaled to $[0, 1]$; their range is given by the associated colour bar. **Bottom:** This group of fifteen plots depicts pixel-wise nnse for each (program, normalized parameter) pairing. In both the middle and bottom groups, the program is denoted along the left-hand side, while the normalized parameter value is denoted along the top row of each group.

Chapter 4

Compressed sensing parameter sensitivity

4.1 Overview

In this chapter, we examine parameter sensitivity in the setting of CS. We start by presenting three sibling results as an allusion to the main results of this chapter. We intend for each result to encapsulate the behaviour of the three ℓ_1 programs that are the main focus of this chapter. Define the worst-case risk for (LS_τ) in the low-noise regime by:

$$R^*(s, A) := \lim_{\eta \to 0} \sup_{x \in \Sigma_s^N \cap \partial B_1^N} \hat{R}(1; x, A, \eta).$$

Importantly, under mild assumptions, $R^*(s, A)$ is nearly equivalent to the optimally tuned worst-case risk for (LS_τ). Namely, for all $\eta > 0$,

$$R^*(s, A) \leq \sup_{x \in \Sigma_s^N} \hat{R}(\|x\|_1; x, A, \eta) \leq C R^*(s, A).$$

This result is treated formally in § 4.6.2. In each case discussed below, the performance of the estimators will be compared to $R^*(s, A)$ as a benchmark, noting that this quantity is minimax order optimal in the sense of Proposition 4.1.5.

In § 4.2, we show that (LS_τ) exhibits an asymptotic PS in the low-noise regime. There is exactly one value τ^* of the governing parameter yielding minimax order-optimal error, with any choice $\tau \neq \tau^*$ yielding markedly worse behaviour. This result holds for normalized K-subgaussian matrices A, which are defined in Definition 5. The intuition provided by this result is that (LS_τ) is extremely sensitive to the value of τ in the low-noise regime, making empirical use of (LS_τ) woefully unstable in this regime.

Theorem 4.1.1 ((LS_τ) sensitivity simplified)**.** *Fix $\varepsilon > 0$ and let $1 \leq s \leq m < N < \infty$ be integers. If $A \in \mathbb{R}^{m\times N}$ is a normalized K-subgaussian matrix with $m > C_\varepsilon \tilde{K}^2 s \log(N/s)$, then with probability at least $1-\varepsilon$ on the realization of A,*

$$\lim_{\eta\to 0} \sup_{x\in\Sigma_s^N \cap B_1^N} \frac{\hat{R}(\tau; x, A, \eta)}{R^*(s, A)} = \begin{cases} 1 & \tau = 1 \\ \infty & \textit{otherwise} \end{cases}$$

Next, in § 4.3 we state a rephrasing of Shen et al. [67, Theorem 3]. The result shows there is a parameter λ^* such that (QP_λ) is not sensitive to its parameter choice for $\lambda \geq \lambda^*$. Right-sided parameter stability of (QP_λ) was first established in Bickel et al. [13, Theorem 7.2]. This well-known result is contrasted in § 4.5 with numerical results demonstrating a left-sided PS for (QP_λ) in the regime of high sparsity, low noise, and large dimension.

Theorem 4.1.2 ((QP_λ) right-sided stability)**.** *Fix $\varepsilon > 0$, let $1 \leq s \leq m < N < \infty$ be integers, $x_0 \in \Sigma_s^N$ and $A \in \mathbb{R}^{m\times N}$ a normalized K-subgaussian matrix. There is an absolute constant $C > 0$ such that if $\lambda \geq C\sqrt{\log N}$ and $m \geq C_\varepsilon \tilde{K}^2 s \log(N/s)$, then with probability at least $1-\varepsilon$ on the realization of A,*

$$R^\sharp(\lambda; x_0, A, \eta) \leq C\lambda^2 s.$$

In the initial work, $\lambda^* = 2\sqrt{2\log N}$ [13, Theorem 7.2]. In the present phrasing, we write only that $\lambda^* = C\sqrt{\log N}$ for an absolute constant $C > 0$. Note that when the data are Gaussian, right-sided stability of (QP_λ) and ($\mathrm{QP}_{\lambda,\mathcal{K}}$) has been examined by Thrampoulidis et al. [71, 72].

Finally, in § 4.4 we show that (BP_σ) is poorly behaved for all $\sigma > 0$ when x_0 is very sparse. In particular, under mild restrictions on the aspect ratio of the measurement matrix, we show that $\tilde{R}(\sigma; x_0, N, \eta)$ is asymptotically suboptimal for *any* $\sigma > 0$ when s/N is sufficiently small. Below, this theorem shows that the minimax risk for (BP_σ), relative to the benchmark risk R^*, converges in probability to ∞.

Theorem 4.1.3 ((BP_σ) sensitivity simplified)**.** *Fix $\eta > 0$, an integer $s \geq 1$, and suppose for $m : \mathbb{N} \to \mathbb{N}$ that $m(N)/N \to \gamma \in (0, 1)$. For each N, suppose $A = A(N) \in \mathbb{R}^{m(N)\times N}$ is a normalized K-subgaussian matrix. Then, for all $M > 0$,*

$$\lim_{N\to\infty} \mathbb{P}\left(\inf_{\sigma>0} \sup_{x\in\Sigma_s^N} \frac{\tilde{R}(\sigma; x, A, \eta)}{R^*(s, A)} > M \right) = 1.$$

Numerical results supporting § 4.2 and § 4.4 are discussed in § 4.5. Proofs of most of the theoretical results are deferred to § 4.6. Next, we add two clarifications. First, the three programs are equivalent in a sense.

Proposition 4.1.4 (Program equivalence [35, Proposition 3.2])**.** *Let $0 \neq x_0 \in \mathbb{R}^N$ and $\lambda > 0$. Where $x^\sharp(\lambda)$ solves (QP_λ), define $\tau := \|x^\sharp(\lambda)\|_1$ and $\sigma := \|y - Ax^\sharp(\lambda)\|_2$. Then $x^\sharp(\lambda)$ solves (LS_τ) and (BP_σ).*

However, τ and σ are functions of z, a random variable, and this mapping may not be smooth. Thus, parameter stability of one program is not implied by that of another. Second, $R^*(s, A)$ has the desirable property that it is computable up to multiplicative constants [47].

Proposition 4.1.5 (Risk equivalences)**.** *Fix $\varepsilon > 0$, let $1 \leq s \leq m < \infty, N \geq 2$ be integers, let $\eta > 0$. Suppose $A \in \mathbb{R}^{m\times N}$ is a normalized K-subgaussian matrix satisfying $m > C_\varepsilon K^2 \log Ks \log(N/s)$, and suppose that $y = Ax_0 + \eta z$ for $z \in \mathbb{R}^m$ with $z_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. Let*

$$M^*(s, N) := \inf_{x^*} \sup_{x_0\in\Sigma_s^N} \eta^{-2}\|x_* - x_0\|_2^2$$

be the minimax risk over arbitrary estimators $x^* = x^*(y)$. There are $c, C > 0$ such that with probability at least $1 - \varepsilon$ on the realization of A,

$$cs\log(N/s) \leq M^*(s,N) \leq \inf_{\lambda>0} \sup_{x_0 \in \Sigma_s^N} R^\sharp(\lambda; x_0, A, \eta)$$
$$\leq CR^*(s,A) \leq Cs\log(N/s).$$

In this thesis, we focus primarily on the versions of LASSO for which $\|\cdot\|_1$ is the structural proxy, and Σ_s^N the structure set for the data x_0. In addition, we discuss the pertinence of our results to the Generalized LASSO setting. For instance, in Lemma A.1.1, we show how Theorem 4.2.1 adapts to the setting of PS for low-rank matrix recovery using the nuclear norm. Further, we connect our discussion on (QP_λ) in § 4.3 to results for more general gauges [71, 72], which works have developed tools suitable for analyzing PS of $(\mathrm{QP}_{\lambda,\mathcal{K}})$ when the data are Gaussian. While it remains an open question to determine how our results in § 4.4 may be extended to analyze PS of $(\mathrm{BP}_{\sigma,\mathcal{K}})$, we conjecture that a suboptimality result like that exemplified in Theorem 4.1.3 exists under analogous assumptions for $(\mathrm{BP}_{\sigma,\mathcal{K}})$.

4.2 On parameter sensitivity for (LS_τ)

The main result of this section is proved in the case of standard CS, where $x_0 \in \Sigma_s^N$ and where tight bounds on the effective dimension of the structure set are known (e.g., bounds on the Gaussian complexities $\gamma(\mathcal{L}_s)$ or $\gamma(\mathcal{K}_s^N)$). Define $\tau^* := \|x_0\|_1$. The following result states that $\hat{L}$ is almost surely suboptimal in the limiting low-noise regime when $\tau \neq \tau^*$, while $\hat{R}(\tau^*)$ is order-optimal. A proof of the result may be found in § 4.6.3, with supporting lemmata in § 4.6.3.

Theorem 4.2.1 (Asymptotic singularity)**.** *Fix $\delta, \varepsilon > 0$ and let $1 \leq s \leq m < N < \infty$ be integers. Let $x_0 \in \Sigma_s^N \setminus \Sigma_{s-1}^N$ with $\tau^* := \|x_0\|_1$ and $\tau > 0$ such that $\tau \neq \tau^*$. Let $\eta > 0$ and let $z \in \mathbb{R}^m$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$. Suppose*

$A \in \mathbb{R}^{m\times N}$ is a normalized K-subgaussian matrix, and assume m satisfies

$$m > C_\varepsilon \tilde{K}^2 \delta^{-2} s \log \frac{eN}{s}.$$

Almost surely on the realization of (A, z),

$$\lim_{\eta\to 0} \hat{L}(\tau; x_0, A, \eta z) = \infty.$$

With probability at least $1-\varepsilon$ on the realization of A, there exist constants $0 < c_\delta < C_\delta < \infty$ such that

$$c_\delta s \log \frac{N}{s} \leq \lim_{\eta\to 0} \sup_{x\in\Sigma_s^N} \hat{R}(\|x\|_1; x, A, \eta) \leq C_\delta s \log \frac{N}{2s}.$$

For clarity, observe the similarity to the definition of τ^* of the precise parameter, $\|x\|_1$, appearing in the lower bound. Importantly, the spirit of this result extends to the situation where x_0 belongs, more generally, to some convex proxy set $\mathcal{K} \subseteq \mathbb{R}^N$. In particular, the blow-up of $\hat{L}$ in the limiting low-noise regime holds independent of the assumptions on $\mathcal{K}$, except that $\mathcal{K}$ be bounded. For instance, Lemma A.1.1 addresses an analogous result in the case where the signal is a $d \times d$ matrix and $\mathcal{K}$ is the nuclear norm ball. Further, worst-case bounds on $\hat{R}$ are well-known in the case where $\mathcal{K}$ is a convex polytope, and are useful when $\mathcal{K}$ has small gmw [7, 47].

4.3 A brief note regarding (QP_λ)

4.3.1 Right-sided parameter stability

In this section we present a contrast to the type of sensitivity observed in § 4.2. Specifically, we observe via Theorem 4.1.2 that (QP_λ) is not sensitive to its parameter choice if the chosen parameter is too large. This so-called right-sided parameter stability is important in practical settings, as it suggests that recovery will not be penalized "too heavily" if the parameter is chosen incorrectly to be too large. Having such a leniency is reassuring, since knowing the exact choice of λ in an experimental setting is unlikely at

best.

The right-sided parameter stability for (QP_λ) was first proved in [13, Theorem 7.2]. When the data are Gaussian, right-sided stability of (QP_λ) has been examined for (QP_λ) and ($\mathrm{QP}_{\lambda,\mathcal{K}}$) by Thrampoulidis et al. [71, 72]. Here, Theorem 4.1.2 is a specialized rephrasing of a known result [67, Theorem 3] that is more suitably adapted to the present work.

In particular, over-guessing λ results in no more than a quadratic penalty on the bound for the recovery error. Consequently, (QP_λ) is right-sided parameter stable — it is not sensitive to variation of its governing parameter when the parameter is sufficiently large. We note, thereby, that there exist regimes in which LASSO programs are not sensitive to their parameter choice. Perhaps more importantly: there are regimes in which one program may be sensitive to its parameter choice, and another program is not. Namely, we see that (LS_τ) can be sensitive to its parameter choice in the low-noise regime (so the correct choice of τ is an imperative), while recovery for the very same data using (QP_λ) is not sensitive to λ (if λ is sufficiently large).

4.4 On parameter sensitivity for (BP_σ)

The final program that we subject to scrutiny is (BP_σ). It is well-known under standard assumptions that an optimal choice of σ yields order-optimal risk $\tilde{R}(\sigma^*; x_0, A, \eta)$ with high probability on the realization of A. In this section, we demonstrate the existence of a regime in which any choice of σ fails to yield order-optimal recovery for (BP_σ). A key message of this section is that (BP_σ) performs poorly if the signal is too sparse and the number of measurements is too large. We demonstrate this behaviour for two regimes: the underconstrained setting, where σ is "too large" and the overconstrained setting where σ is "too small". Each of these settings covers the case where σ is chosen "just right"; we will see how (BP_σ) risk fails to achieve order optimality in this case, as well.

For the duration of this section, we will consider $x_0 \in \Sigma_s^N$ where s may or may not be allowed to be 0. We will clarify this explicitly in each instance. The main result of the section will be stated in the case where $A \in \mathbb{R}^{m\times N}$ is

a normalized K-subgaussian matrix with $m = m(N)$ satisfying a particular growth condition. Thus, the measurement vector will be given by $y = Ax_0 + \eta z$ where $\eta > 0$ is the noise scale and $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ as before. For the sake of analytical and notational simplicity, we assume that η is independent of N. However, we eventually make clear how σ may be allowed to depend on the ambient dimension, and that our result holds irrespective of this dependence.

4.4.1 Underconstrained parameter sensitivity

As a "warm-up" for the main result, we start by demonstrating that there is a regime in which $\tilde{R}(\sigma; x_0, A, \eta)$ fails to achieve minimax order-optimality when restricted to $\sigma \geq \eta\sqrt{m}$. Specifically, if m is too large, then there is a (sufficiently sparse) vector $x_0 \in \Sigma_s^N$ such that, with high probability on the realization of A, the risk $\tilde{R}(\sigma; x_0, A, \eta)$ is large regardless of the choice of $\sigma \in [\eta\sqrt{m}, \infty)$. We defer the proof of this result to § 4.6.4.

Lemma 4.4.1 (Underconstrained maximin (BP_σ))**.** *Fix $\delta, \varepsilon, \eta > 0$, let $1 \leq s < m \leq N$ be integers, and suppose $A \in \mathbb{R}^{m \times N}$ is a normalized K-subgaussian matrix. If*

$$m > C_\varepsilon \delta^{-2} \tilde{K}^2 s^2 \log^2\left(\frac{N}{s}\right),$$

then with probability at least $1 - \varepsilon$ on the realization of A,

$$\inf_{\sigma \geq \eta\sqrt{m}} \sup_{x \in \Sigma_s^N} \tilde{R}(\sigma; x, A, \eta) \geq C\sqrt{m}.$$

Remark 12. In some settings, it may be not be appropriate for m to depend logarithmically on N. If m and N satisfy the power law relation $m = N^k$ for some $k \in (0,1)$, then under the assumptions of Lemma 4.4.1,

$$\mathop{\mathbb{E}}_{z} \|\tilde{x}(\sigma) - x_0\|_2^2 = \Omega(N^{k/2}).$$

4.4.2 Minimax suboptimality

Just as observed in § 4.4.1, the results of this section hold in the regime where the aspect ratio approaches a constant: $m/N \to \gamma \in (0,1)$. Our simulations in § 4.5.3 support suboptimality of $\tilde{R}$, and sensitivity of (BP_σ) to its parameter choice for aspect ratios ranging from $\gamma = 0.1$ to $\gamma = 0.45$.

Our result is of an asymptotic nature in one additional sense. We have stated that $\tilde{R}$ may be suboptimal for "very sparse" signals x_0. This is specified in the sense that, while m and N may be allowed to grow, s remains fixed. Our numeric simulations demonstrate how this assumption may be interpreted as the inability of (BP_σ) to effectively recover the off-support of the signal x_0 (i.e., the all 0 sub-vector x_{T^C} where $T \subseteq [N]$ denotes the support of x_0). Thus, it is in this setting, where the number of measurements is sufficiently large, and the sparsity sufficiently small, that we show $\tilde{R}(\sigma; x_0, A, \eta)$ is asymptotically suboptimal, regardless of the choice of σ.

Theorem 4.4.2 ((BP_σ) minimax suboptimality)**.** *Fix $\varepsilon, \eta > 0$ and $m : \mathbb{N} \to \mathbb{N}$ with $m(N)/N \to \gamma \in (0,1)$. There is $N_0 \geq 2$ and $p > 0$ so that for any $N \geq N_0$ and any $1 \leq s < m(N_0)$, if $A \in \mathbb{R}^{m \times N}$ is a normalized K-subgaussian matrix, then with probability at least $1 - \varepsilon$ on the realization of A,*

$$\inf_{\sigma > 0} \sup_{x \in \Sigma_s^N} \tilde{R}(\sigma; x, A, \eta) \geq C_{\gamma,K} N^p. \tag{4.1}$$

A minor modification of the above result allows one to show that the minimax risk for (BP_σ), relative to the benchmark risk $R^*(s, A)$, converges in probability to ∞.

Corollary 4.4.3 ((BP_σ) suboptimal in probability)**.** *Fix $\eta > 0$, an integer $s \geq 1$, and suppose for $m : \mathbb{N} \to \mathbb{N}$ that $m(N)/N \to \gamma \in (0,1)$. For each N, suppose $A = A(N) \in \mathbb{R}^{m(N) \times N}$ is a normalized K-subgaussian matrix. Then, for all $M > 0$,*

$$\lim_{N \to \infty} \mathbb{P}\left(\inf_{\sigma > 0} \sup_{x \in \Sigma_s^N} \frac{\tilde{R}(\sigma; x, A, \eta)}{R^*(s, A)} > M \right) = 1.$$

4.5 Numerical results

Let $\mathfrak{P} \in \{(\mathrm{LS}_\tau), (\mathrm{QP}_\lambda), (\mathrm{BP}_\sigma)\}$ be a CS program with solution $x^*(\upsilon)$, where $\upsilon \in \{\tau, \lambda, \sigma\}$ is the associated parameter. Given a signal $x_0 \in \mathbb{R}^N$, matrix $A \in \mathbb{R}^{m \times N}$, and noise $\eta z \in \mathbb{R}^m$, denote by $\mathscr{L}(\upsilon; x_0, A, \eta z)$ the loss associated to $\mathfrak{P}$. For instance, if $\mathfrak{P} = (\mathrm{LS}_\tau)$, then $\mathscr{L} = \hat{L}$. In most cases, the signal x_0 for our numerical simulations will be s-sparse, and s will be "small". For simplicity, and to ensure adequate separation of the "signal" from the "noise", each non-zero entry of x_0 will be equal to N, except where otherwise noted. Unless otherwise noted the measurement matrix A will have entries $A_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, m^{-1})$.

Define $\upsilon^* := \upsilon^*(x_0, A, \eta) > 0$ to be the value of υ yielding best risk (i.e., where $\mathbb{E}_z \mathscr{L}(\cdot; x_0, A, \eta z)$ is minimal) and let the normalized parameter ρ for the problem $\mathfrak{P}$ be given by $\rho := \upsilon/\upsilon^*$. Note that $\rho = 1$ is a population estimate of the argmin of $\mathscr{L}(\rho\upsilon^*; x_0, A, \eta z)$; by the law of large numbers, this risk estimates well an average of such losses over many realizations $\hat{z}$. Let $L(\rho) := \mathscr{L}(\rho\upsilon^*)$ denote the loss for $\mathfrak{P}$ as a function of the normalized parameter, let $\{\rho_i\}_{i=1}^n$ denote a sequence of points in the normalized parameter space, and define the average loss for $\mathfrak{P}$ at any point ρ_i by

$$\bar{L}(\rho_i; x_0, A, \eta, k) := k^{-1} \sum_{j=1}^{k} L(\rho_i; x_0, A, \eta \hat{z}_{ij}), \tag{4.2}$$

where $\hat{z}_{ij}$ is the (i,j)-th realization of noise; $\hat{z}_{ij} \sim \mathcal{N}(0, I_m)$ for all $(i,j) \in [n] \times [k]$. We may also refer to $\bar{L}$ as the empirical risk or average noise-normalized squared error (nnse). Note that $\bar{L}$ depends on $(\hat{z}_{ij} : i \in [n], j \in [k])$ and that notating this dependence is omitted for simplicity. Below, $\hat{z}_{ij}$ are not necessarily sampled independently. In fact, to obtain tractable computational simulations, we will frequently have $\hat{z}_{ij} = \hat{z}_{i'j}$ for $i, i' \in [n]$. Where necessary, we disambiguate the average losses with a subscript: $\bar{L}_{(\mathrm{LS}_\tau)}$, $\bar{L}_{(\mathrm{QP}_\lambda)}$ and $\bar{L}_{(\mathrm{BP}_\sigma)}$ for the programs (LS_τ), (QP_λ) and (BP_σ), respectively.

Given two signals x^*, x define the peak signal-to-noise ratio (psnr) by

$$\operatorname{psnr}(x^*, x) := 10 \log_{10} \Big(\frac{\max_{i \in [N]} x_i^2}{\operatorname{mse}(x^*, x)} \Big), \tag{4.3}$$

$$\operatorname{mse}(x^*, x) := \frac{1}{N} \sum_{i=1}^{N} (x_i^* - x_i)^2. \tag{4.4}$$

As with defining loss $L(\rho; x_0, N, \eta \hat{z}_{ij})$, by abuse of notation we define psnr as a function of the normalized parameter ρ:

$$\operatorname{psnr}(\rho) := \operatorname{psnr}(\rho; x_0, N, \eta \hat{z}_{ij}) := \operatorname{psnr}(x^*(\rho \upsilon^*), x_0).$$

In this section, we include plots representing average loss of a program with respect to that program's normalized governing parameter. Both the optimal parameter υ^* and the average loss $\bar{L}$ are approximated by $\upsilon^\dagger$ and $L^\dagger$, respectively, using RBF interpolation as described in § A.2.2. Parameter settings for the interpolations are provided in § A.2.3. To this end, the average loss $\mathbb{E}_z \, L(\rho; x_0, A, \eta z)$ is approximated from k realizations of the true loss on a logarithmically spaced grid of n points centered about $\rho = 1$. The approximation is computed using multiquadric RBF interpolation and the values of the parameters of the interpolation, $(k, n, \varepsilon_{\text{rbf}}, \mu_{\text{rbf}}, n_{\text{rbf}})$, are stated in each instance where the computation was performed. In every case, due to concentration effects, for a given program and given parameter value the realizations cluster very closely about the average loss. Therefore, RBF interpolation is very close to the true approximated average loss curve computed from the loss realizations; and has the added advantage of facility to account for nonuniformly spaced data points. In the main graphics, we omit the original data point cloud in favour of presenting clean, interpretable plots. However, we include auxiliary plots of the average loss approximant and the point cloud to visualize goodness of fit. In addition, these latter visualizations serve to support how a program's order-optimality for a single realization may be impacted by averaging over noise.

There is one final caveat to note in how the plots were generated. Computational methods available to the authors for computing solutions to (BP_σ)

and (LS_τ) were much slower than those available for (QP_λ). Consequently, ensuring computational tractability of our numerical simulations required solving (QP_λ) and obtaining corresponding parameter values from those problem instances. Namely, given $\{\rho_i\}_{i=1}^n$ and solutions $x^\sharp(\lambda_i)$ where $\lambda_i = \rho_i\lambda^*, i \in [n]$, we use Proposition 4.1.4 to obtain $\tau_i := \|x^\sharp(\lambda_i)\|_1$ and $\sigma_i := \|y - Ax^\sharp(\lambda_i)\|_2$. Thus, we obtain loss curves $\hat{L}(\tau_i; x_0, A, \eta z)$ and $\tilde{L}(\sigma_i; x_0, A, \eta z)$ by solving (QP_λ) on a sufficiently fine grid, yielding $(\lambda_i, x^\sharp(\lambda_i))$. This allows us to approximate the optimal parameter choices for (LS_τ) and (BP_σ) and therefore determine within some numerical tolerance all of $\bar{L}_{(\mathrm{LS}_\tau)}$, $\bar{L}_{(\mathrm{QP}_\lambda)}$ and $\bar{L}_{(\mathrm{BP}_\sigma)}$. Further details for approximating the average loss of a program are given in § A.2.2 where we describe (RBF) interpolation [16, 78].

4.5.1 (LS_τ) numerics

The data generating process for the numerics in this section is as follows. Fix $A \in \mathbb{R}^{m\times N}$, $\eta > 0$ and $x_0 \in \Sigma_s^N$. Fix a logarithmically spaced grid of n points for the normalized parameter, $\{\rho_i\}_{i=1}^n$, centered about 1. Generate k realizations $\{z_j\}_{j=1}^k$ of the noise. Obtain $\lambda_i := \rho_i\lambda^*(A, x_0, \eta)$ and obtain $\ell_{ij} := L^\sharp(\lambda_i; x_0, A, \eta z_j)$ after computing $x^\sharp(\lambda_i; z_j)$. Observe that $\hat{L}(\tau_{ij}; x_0, A, \eta z_j) = \ell_{ij}$ where $\tau_{ij} := \|x^\sharp(\lambda_i; z_j)\|_1$. Similarly, for $\sigma_{ij} := \|y_j - Ax^\sharp(\lambda_i; z_j)\|_2$, one has $\tilde{L}(\sigma_{ij}; x_0, A, \eta z_j) = \ell_{ij}$. Finally, for a sufficiently fine and wide numerical grid, one may approximate the normalized parameter grids $\{\tau_{ij}\}$ and $\{\sigma_{ij}\}$ using the values $\{\ell_{ij}\}$. Consequently, for each program we are able to approximate the average loss $\bar{L}$ by obtaining a clever approximate interpolant of $\{(\tau_{ij}, \ell_{ij}) : (i,j) \in [n]\times[k]\}$, $\{(\sigma_{ij}, \ell_{ij}) : (i,j) \in [n]\times[k]\}$ or $\cup_{j\in[k]}\{(\rho_i, \ell_{ij}) : i \in [n]\}$, respectively, while only having to solve (QP_λ). This particular bit of good fortune is guaranteed to us by Proposition 4.1.4. As stated, the clever approximant is obtained using multiquadric RBF interpolation [16, 78]. For more background on kernel methods for function approximation and radial basis functions in particular, we refer the reader to Buhmann [16], Hastie et al. [40], Murphy [53]. Some additional detail to this end is provided in § A.2.2.

The numerics for (LS_τ), appearing in Figure 4.1, concern the case in

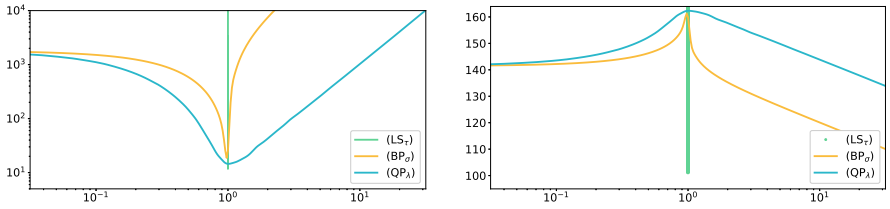


Figure 4.1: (LS_τ) PS in the low-noise regime. Average loss (left) plotted on a log-log scale with respect to the normalized parameter; average psnr (right) plotted on a log-linear scale with respect to the normalized parameter. The data parameters are $(s, m, N, \eta, k, n) = (1, 2500, 10^4, 2 \cdot 10^{-3}, 15, 301)$.

which η is small. When the ambient dimension is modest ($N = 10^4$) and the noise scale only moderately small ($\eta = 2 \cdot 10^{-3}$), PS of (LS_τ) is readily observed. Minute changes in τ lead to blow-up in the nnse and the psnr (left and right plot, respectively). Indeed, for the range plotted, it is difficult to visually segment the left half of the (LS_τ) average loss curve from the right half. These observations support the asymptotic theory of § 4.2. Moreover, the simulations suggest that the other two programs, (BP_σ) and (QP_λ) are relatively much less sensitive to the choice of their governing parameter.

4.5.2 (QP_λ) numerics

In this section we visualize the average loss of (QP_λ) as a function of its normalized parameter $\rho = \lambda/\lambda^*$. In Figure 4.2, the average loss for (QP_λ) is plotted with respect to the ρ for an aspect ratio δ ranging between 0.25 and 4. As suggested by Theorem 4.1.2, the average loss appears to scale quadratically with respect to the normalized parameter for values $\rho > 1$. For $\rho \in (0.5, 0.9)$, the average loss appears to scale super-quadratically with respect to the normalized parameter, with the rate of growth increasing as a function of δ. This behaviour suggests that (QP_λ) can be sensitive to its parameter choice if ρ is too small. The intuition for the observed behaviour is that (QP_λ) increasingly behaves like ordinary least squares when $\lambda \to 0$. Each average loss was approximated from 15 realizations of the loss using multiquadric RBF interpolation. Due to concentration effects, the realiza-

tions for each parameter value clustered very closely to the approximated average loss. The left-hand plot Figure 4.2 is plotted on a log-log scale, while the right-hand plot is plotted on a linear-linear scale. The linear-linear plot readily demonstrates how over-guessing λ by a factor of 2 is more robust to error than under-guessing λ by a factor of 2.

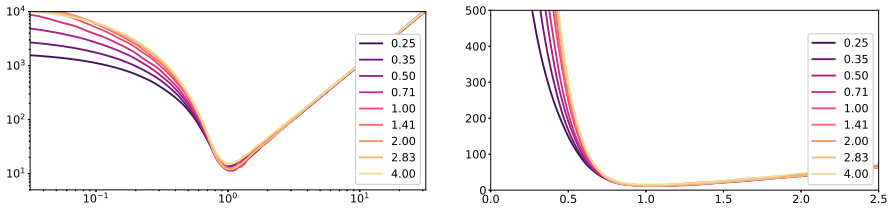


Figure 4.2: Average loss of (QP_λ) plotted with respect its normalized parameter in the low-noise, high sparsity regime. Parameters for the simulation are $(s, N, \eta, k, n) = (1, 10^4, 10^{-5}, 15, 301)$. The aspect ratio of the matrix $A \in \mathbb{R}^{m \times N}$ with $A_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, m^{-1})$ takes values $\delta \in \{4^{-1}, \dots, 4\}$, as shown in the legend. The data are visualized on a log-log scale (left) and linear scale (right).

In Figure 4.3, we visualize (QP_λ) average loss with respect to its normalized parameter. In the top row, we include two plots similar to Figure 4.2, but for $\delta = 0.25, 0.45$ only. Again, the left-hand plot is on a log-log scale while the right-hand plot is on a linear-linear scale. The bottom row depicts the goodness of fit of the RBF approximation for the average loss.

4.5.3 (BP_σ) numerics

This section includes numerical simulations depicting the sensitivity of (BP_σ) to its parameter choice. These numerics serve to support the asymptotic theory developed in § 4.4.

The graphics in Figure 4.4 serve as an initial depiction of (BP_σ) PS, depicting the average loss for each program and for $N \in \{4000, 7000\}$, $\delta \in \{0.1, 0.25, 0.45\}$. Each plot depicts the average loss as a function of the normalized parameter for (LS_τ) (green), (BP_σ) (orange) and (QP_λ) (blue). The domain of the normalized parameter in each plot is $(0.2, 5)$. A single realization of A was fixed and the average loss was computed from $k = 50$

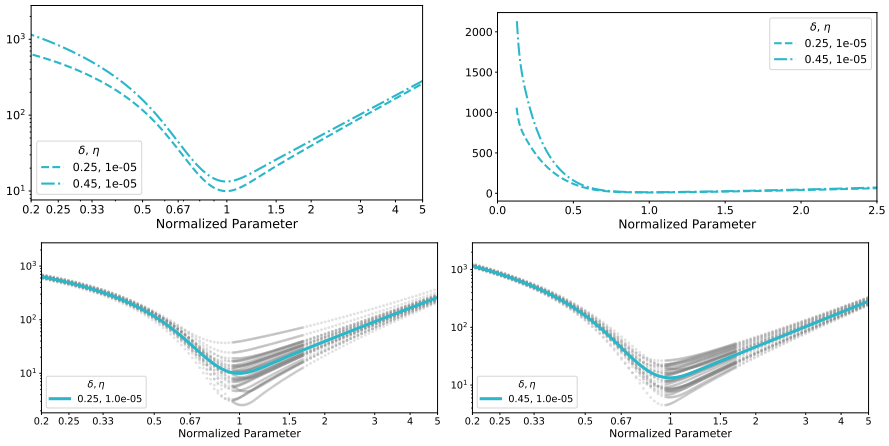


Figure 4.3: Average loss of (QP_λ) plotted with respect to its normalized parameter in the low-noise, high-sparsity regime. Parameters for the simulation are $(s, N, \eta, k, n) = (1, 10^4, 10^{-5}, 25, 201)$. The aspect ratio of the matrix $A \in \mathbb{R}^{m \times N}$ with $A_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, m^{-1})$ takes values $\delta = 0.25, 0.45$, as shown in the legend. The data in the top row are visualized on a log-log scale (left) and linear scale (right). The bottom row depicts the quality of fit for the RBF approximation of the average loss for $\delta = 0.25$ (left) and $\delta = 0.45$ (right).

realizations of the noise by constructing a function approximator using radial basis function approximation with a multiquadric kernel. The RBF approximator was evaluated on a logarithmically spaced grid of $n_{\mathrm{rbf}} = 301$ points centered about 1. The loss values for (LS_τ) and (BP_σ) were computed by using the program equivalence described by Proposition 4.1.4. In particular, for computational expediency, once A and z were fixed, the Lasso program was solved only using (QP_λ), for all λ in a specified range. For each $x^\sharp(\lambda)$, we obtained $\hat{x}(\tau)$ and $\tilde{x}(\sigma)$ using that $\hat{x}(\tau) = \tilde{x}(\sigma) = x^\sharp(\lambda)$ for $\tau := \|x^\sharp(\lambda)\|_1$ and $\sigma := \|y - Ax^\sharp(\lambda)\|_2$.

For convenience, we refer to each plot in Figure 4.4 *via* its (row, column) position in the figure. The collection of plots serves to depict how the average loss changes as a function of N and $\delta = m/N$ when $\eta = 1$. Namely, as N increases, the average loss for (BP_σ) becomes sharper about the op-

timal parameter choice. In addition, as δ increases, we observe the same phenomenon. In Figure 4.6, similar content is depicted, but for $\eta = 100$. In this case, $n_{\text{rbf}} = 501$ was used.

Specific parameter settings for the RBF approximation for each set of problem parameters and program are detailed in Table A.3. Because (BP_σ) PS is not easily visualized in small dimensions (e.g., for $N < 10^6$), we supply several plots visualizing the quality of the RBF approximation. Namely, approximation quality plots corresponding with Figure 4.4 may be found in Figure 4.5; and approximation quality plots are included in Figure 4.6. Each row of these plots is a triptych; each column corresponds to a program: (LS_τ) for the left-most, (BP_σ) in the centre; and (QP_λ) on the right. These plots depict a single line and a collection of points. The points correspond to individual loss values for each realization of the noise and each normalized parameter value computed. The line corresponds to the RBF approximation of the average loss for that program. The domain for the (LS_τ) plots is $(0.95, 0.95^{-1})$ in the normalized parameter space. For (BP_σ) it is $(0.9, 0.9^{-1})$, and for (QP_λ) $(0.75, 0.75^{-1})$.

In Figure 4.5, one may observe by inspection that the loss realizations for (LS_τ) and (QP_λ) typically are achieved very close to 1. In contrast, there is a relatively wider range in the domain for where the (BP_σ) loss achieves its optimum. This is integral to how (BP_σ) risk is sensitive to the choice of σ.

There appear to be two competing factors that impact sensitivity to parameter choice and program optimality. The first is sensitivity of the parameter with respect to variation due to the noise realization. This has already been described: because $\sigma(\lambda^*, z)$ varies greatly as a function of z, program optimality is destroyed by suboptimal loss values near $\sigma = \sigma^*$. On the other hand, this also tends to have somewhat of a smooth effect about the optimal normalized parameter. Thus, as $\tau(\lambda^*, z)$ does not vary as greatly in this manner, its sensitivity to parameter choice is not smoothed due to local averaging effects.

4.5.4 More synthetic examples

In this section, we display three synthetic examples where only s and η were changed. Thus, the effect of sparsity and noise scale is readily observed. In each of these figures, the aspect ratio of the measurement matrix was $\delta = 0.25, 0.45$ for the left- and right-hand plots respectively (except Figure 4.8 where $\delta = 0.25$ was too small to achieve recovery). The average loss curves for each program were computed from $k = 25$ realizations of loss curves that were, themselves, generated on a logarithmically spaced grid of $n = 201$ points centered about the optimal choice of the normalized parameter, $\rho = 1$. The loss realizations were again computed by solving (QP_λ) and using the correspondence between Lasso programs to compute the loss curves for (LS_τ) and (BP_σ).

The bottom row of Figure 4.7 and Figure 4.9, and the right half of Figure 4.8 depict the quality of the approximation of the average loss curve for each program. Specifically, each program appears with its own facet, in which are displayed the individual loss realizations $L(\rho_i; x_0, A, \eta z_j), i \in [n], j \in [k]$ as grey points, and the average loss $\bar{L}(\rho; x_0, A, \eta)$ as a coloured line. The top row of Figure 4.7 and Figure 4.9, and the left half of Figure 4.8 compare the the average loss curves for each program, where the average losses are plotted on a log-log scale with respect to the normalized parameter.

The first figure, Figure 4.7, displays a setting similar to Figure 4.1. The noise scale was $\eta = 10^{-5}$ and $s = 1$. Thus, the setting depicts the low-noise high-sparsity regime. The second figure, Figure 4.8, depicts a moderately low-noise regime, with a large value of s (so large that $\delta = 0.25$ did not yield adequate recovery). Thus, this figure depicts the regime in which x_0 is near the limit of acceptable sparsity for the CS regime. Finally, the parameter settings for Figure 4.9 were $s = 100$ and $\eta = 100$. In particular, sparsity is modest, and the noise scale is large (the variance equals the ambient dimension, $\eta^2 = N = 10^4$).

It is readily observed that (LS_τ) is highly sensitive to its parameter choice in both low-noise regimes. We observe that (BP_σ) becomes more sensitive to its parameter choice as sparsity decreases from 750 to 100 to 1. Finally, we

observe that (QP_λ) is most sensitive to its parameter choice in the low-noise high-sparsity regime. This left-sided sensitivity is consistent with the theory and numerical simulations for the corresponding proximal denoising setting in Chapter 3 [8, 9].

4.5.5 Realistic Examples

We next include two realistic examples in addition to the synethetic ones of the previous sections. We show how CS programs may exhibit sensitivity as a function of their governing parameter for a 1D and 2D wavelet problem. In each example, we will include plots similar to those appearing above; however there will be some key differences. As above, the average loss is computed from several realizations of the loss, which depend in turn on realizations of the noise. However, we will plot the loss corresponding to a single realization as a function of the normalized parameter. Computing the normalized parameter is what requires computing the average loss. As before, we approximate the average loss and the normalized parameter using RBF interpolation, described in § A.2.2. The figures of this section contain three main pieces. We will plot psnr, as a function of the normalized parameter; loss, equal to the nnse, as a function of the normalized parameter; and we will include a grid of plots that allows for comparison of CS recovery by visualizing the recovery in the signal domain. The latter grid of plots shall be referred to as "grid plots" while the psnr and nnse plots shall be referred to as "reference plots", as they contain annotations that relate them to the grid plots.

We now include a brief description of the so-called grid plots and associated reference plots that appear in this section. Other than plotting loss, rather than average loss, a key difference of the reference plots to the plots of § 4.5.1–4.5.3 is that they have been annotated with vertical black dashed lines, and coloured dots. Where the loss for a program intersects the black dashed line, we show a representative solution for that program where the normalized parameter for the problem is given by the intercept of the vertical line (approximately). Because the programs were solved on a grid, the

true value of the normalized program is given by the coloured dot appearing nearest the black dashed line. The horizontal axis for the reference plots is the normalized parameter (plotted on a log scale). The vertical axis for the reference plots is either the psnr (plotted on a linear scale) or the nnse (plotted on a log scale). The representatives for each chosen normalized parameter value and each program are plotted as a faceted grid below the reference plot. The chosen normalized parameter value is given at the top of each column, while the program used to recover the noisy ground truth signal is described in the legend.

1D wavelet compressed sensing

The signal $\xi_0 \in \mathbb{R}^N$, $N = 4096$, was constructed in the Haar-wavelet domain. In particular, $x_0 \in \mathbb{R}^N$ has 10 non-zero coefficients, each equal to N. Let $\mathcal{W}_1$ denote the 1D Haar wavelet transform. Thus, $\xi_0 = \mathcal{W}_1 x_0$ where $\|x_0\|_0 = 10$. Next, for $A \in \mathbb{R}^{m \times N}$ where $m = 1843$, define $y = Ax_0 + \eta z$ where $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\eta = 50$. The signal's wavelet coefficients were recovered using (LS_τ), (QP_λ) and (BP_σ) for several realizations of the noise z and over a grid of normalized parameter values. For example $\hat{x}(\tau_i; x_0, A, z^{(j)})$ is the (LS_τ) recovery of the wavelet coefficients x_0 from $(y^{(j)}, A)$ with $\tau = \tau_i$, where $y^{(j)} := Ax_0 + \eta z^{(j)}$. The recovered signal is thus given by $\hat{\xi}(\tau_i) := \mathcal{W}^{-1}\hat{x}(\tau_i)$ and the loss given by $\eta^{-2}\|\hat{\xi}(\tau_i) - \xi_0\|_2^2$. The loss is modified similarly for the other programs. Specifically, the loss is measured in the signal domain and not the wavelet domain. The average loss was approximated from $k = 50$ loss realizations using RBF interpolation, as described in § A.2.2, on a grid of $n = 501$ points logarithmically spaced and centered about $\rho = 1$.

Results of this simulation are depicted in Figure 4.10 with RBF interpolation parameter settings given in Table A.4. The results shown in Figure 4.10 depict data from only a single noise realization: the top-most graphic shows psnr as a function of the normalized parameter; the middle graphic plots the loss as a function of the normalized parameter; and the bottom group compares the ground-truth signal and recovered signals in the signal domain. While psnr and loss are plotted instead of average psnr and average

loss, the normalized parameter was computed from the average loss, as usual (*cf.* § A.2.2). Correspondingly, observe that the optimal parameter choice for each program may not appear at $\rho = 1$, since the optimal normalized parameter for a particular loss realization is not necessarily equal to the optimal normalized parameter for the expected loss. In the bottom group of 15 plots, each row corresponds with a particular program — (LS_τ), (QP_λ) and (BP_σ), from top to bottom — while each column corresponds with a particular vaue of the normalized parameter — 0.5, 0.75, 1, 1.3 and 2, from left to right. The recovered signal for that program and normalized parameter value is shown as a coloured line, while the ground truth signal is shown as a black line.

As $\eta = 50$, the problem lies outside of the small-noise regime. As such, (BP_σ) is more sensitive to its parameter choice than (QP_λ), and more sensitive than (LS_τ) for $\rho > 1$ due to the relatively high sparsity of the signal. Since suboptimality of (BP_σ) is observed for risk or average loss rather than for individual loss realizations, we do not observe suboptimality of (BP_σ) loss in these graphics. As expected, (LS_τ) is sensitive to its parameter choice for $\rho < 1$, as the ground-truth solution lies outside the feasible set in this setting. It appears that the loss is mildly more sensitive to under-guessing τ in this regime, than it is to over-guessing σ. This is readily observed from all of the plots in the figure, especially by comparing those in the bottom group of 15.

For comparison with the middle plot of Figure 4.10, we include a plot of the average loss for each program as a function of the normalized parameter in Figure 4.11 (left plot). Beside it is a triptych visualizing the RBF approximation quality for the average loss.

2D Wavelet Compressed Sensing

In this section, we describe numerical simulations for a 2D wavelet compressed sensing problem. The signal, ξ_0, is an 80×80 image of the so-called square Shepp-Logan phantom (sslp), visualized in Figure 4.12. The sslp was first used in Berk et al. [9]. Let $\mathcal{W}$ denote the Haar wavelet transform and define $x_0 := \mathcal{W}\xi_0 \in \mathbb{R}^{6400}$ to be the vector of Haar wavelet coefficients for

the signal ξ_0. The linear measurements are taken as

$$y = Ax_0 + \eta z, \qquad A_{ij} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, m^{-1}\right), z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \eta > 0.$$

The signal's wavelet coefficients were recovered using (QP_λ) to obtain $x^\sharp(\lambda_i)$, where $i \in [n]$ enumerates the grid of parameter values. The recovered image is then given by $\xi^\sharp(\lambda_i) := \mathcal{W}^{-1}(x^\sharp(\lambda_i))$. By using the method described previously at the beginning of § 4.5, the corresponding solutions for (LS_τ) and (BP_σ) were computed, obtaining $\hat{x}(\tau_i)$ and $\tilde{x}(\sigma_i)$, respectively, in addition to the corresponding images $\hat{\xi}(\tau_i) := \mathcal{W}^{-1}(\hat{x}(\tau_i))$ and $\tilde{\xi}(\sigma_i) := \mathcal{W}^{-1}(\tilde{x}(\sigma_i))$, $i \in [n]$. As in § 4.5.5, the loss has been modified to measure the nnse in the image domain. For example, the (LS_τ) loss is given as $\eta^{-2}\|\hat{\xi}(\tau_i) - \xi_0\|_2^2$; similarly for the other two programs.

Average loss as a function of the normalized parameter ρ is shown in Figure 4.13 for $\eta = 10^{-2}, 1/2$ with $m = 2888$ (i.e., $m/N \approx 0.45$). The average loss was approximated using RBF interpolation from $k = 50$ realizations along a logarithmically spaced grid of 501 points centered about $\rho = 1$ using the method described in § A.2.2. The parameter settings for the RBF interpolation are provided in Table A.5. Plots showing the approximation quality of the RBF interpolation are given in the bottom row of Figure 4.13. In these plots, individual realizations of the nnse for the recovery are shown as grey points. The RBF interpolant is given by the coloured line in each plot. The approximation quality is only visualized for a narrow region about $\rho = 1$. Indeed, the approximation quality of the RBF interpolant was observed, in every case, to be better away from $\rho = 1$ than about $\rho = 1$: ensuring good interpolation of the loss realizations about $\rho = 1$ was observed to be sufficient for ensuring good interpolation of the average loss over the region of interest, $\rho \in [10^{-1}, 10^1]$.

In the left column of the figure, where $\eta = 10^{-2}$, we observe that (LS_τ) is relatively more sensitive to its parameter choice than either (BP_σ) or (QP_λ). In particular, for this problem, we observe that $\eta = 10^{-2}$ is sufficient to lie within the low-noise regime. Due to how the solutions for (LS_τ) were computed from those for (QP_λ), the average loss curve for (LS_τ) is not

resolved over the full domain for the normalized parameter. This reinforces how small changes in the normalized parameter value for (LS_τ) correspond to relatively much larger changes in the normalized parameter value for (QP_λ).

In the right column of the figure, where $\eta = 1/2$, we observe that (BP_σ) is relatively more sensitive to its parameter choice than either (LS_τ) or (QP_λ). We expect this is due to the relatively high sparsity of the signal. Again, the average loss curve for (BP_σ) is not resolved over the full plotted domain of the normalized parameter. This underscores how changes in the governing parameter for (QP_λ) correspond with relatively smaller changes in the governing parameter for (BP_σ). In particular (BP_σ) is more sensitive to its governing parameter than (QP_λ) in the present problem. This observation is supported by the theory of § 4.4.2.

As in previous numerical simulations, the bottom row of Figure 4.13 includes triptyches depicting the average loss approximation quality for the RBF interpolation of the loss realizations (*cf.* § A.2.2).

In both Figure 4.14 and Figure 4.15, we use four main elements to depict the results of a 2D wavelet CS problem, each for a single realization of the noise. In each figure, the top row depicts the psnr curves for each program (left) and loss curves for each program (right). The bottom row of the figure contains two groupings of the 15 plots each. Each grid of 15 plots is faceted by program ((LS_τ), (BP_σ) and (QP_λ), top-to-bottom) and normalized parameter value (0.5, 0.75, 1, 1.3, 2, left-to-right). Each (program, normalized parameter) tuple on the left-hand side of the figure corresponds with its partner on the right-hand side. Specifically, the left-hand grid of images depicts the recovered image for a given (program, normalized parameter) tuple, while the corresponding right-hand image depicts the pixel-wise nnse in the signal domain. The details of these images are best examined on a computer.

In Figure 4.14, the parameter settings are
$(s, N, m, \eta, k, n) = (416, 6418, 2888, 10^{-2}, 50, 501)$. In particular, η lies within the low-noise regime, as observed by the relative sensitivity of (LS_τ) to its parameter choice. In Figure 4.15, the parameter settings are $(s, N, m, \eta, k, n) = (416, 6418, 2888, 1/2, 50, 501)$. In particular, the noise scale is relatively larger.

The relatively high sparsity of the signal causes (BP_σ) to be relatively more sensitive to its parameter choice than either (LS_τ) or (QP_λ). These observations are supported by the theory of § 4.2 and § 4.4.2.

4.6 Proofs

4.6.1 Risk equivalences

Proof of Proposition 4.1.5. The left-most inequality,

$$cs\log(N/s) \leq M^*(s,N),$$

is a consequence of [17, Theorem 1], and the second inequality,

$$M^*(s,N) \leq \inf_{\lambda>0} \sup_{x_0\in\Sigma_s^N} R^\sharp(\lambda; x_0, A, \eta)$$

is trivial. The third inequality,

$$\inf_{\lambda>0} \sup_{x_0\in\Sigma_s^N} R^\sharp(\lambda; x_0, A, \eta) \leq CR^*(s,A)$$

is a consequence of [13, Theorem 6] and Corollary 4.6.3. Indeed, $R^*(s,A)$ may be lower-bounded by the optimally tuned worst-case risk, given by $\sup_{x\in\Sigma_s^N} \hat{R}(\|x\|_1; x, A, \eta)$, which is in turn lower-bounded by $cs\log(N/s)$ due to [17, Theorem 1]. In particular, selecting constants appropriately gives

$$\begin{aligned}
&\inf_{\lambda>0} \sup_{x_0\in\Sigma_s^N} R^\sharp(\lambda; x_0, A, \eta) \\
&\qquad \leq Cs\log(N/s) \leq CM^*(s,N) \\
&\qquad \leq C \sup_{x\in\Sigma_s^N} \hat{R}(\|x\|_1; x, A, \eta) \leq CR^*(s,A).
\end{aligned}$$

The final inequality, a variant of which may be found in [47] or [56], easily follows from Lemma 4.6.9.

□

4.6.2 $\hat{R}$ is nearly monotone

We first quote a specialized version of a result introduced in [47], which gives a kind of local characterization of the deviation inequality presented in Theorem 2.2.7.

Theorem 4.6.1 ([47, Theorem 1.7])**.** *Let A be a normalized K-subgaussian matrix and $T \subseteq \mathbb{R}^N$ a convex set. For any $t \geq 1$, it holds with probability at least $1 - \exp(-t^2)$ that*

$$\big| \|Ax\|_2 - \sqrt{m}\|x\|_2 \big| \leq t \cdot C\tilde{K}\gamma(T \cap \|x\|_2 B_2^N), \qquad \text{for all } x \in T.$$

We now present the main result of this section.

Proposition 4.6.2 ($\hat{R}$ is nearly monotone)**.** *Let A be a normalized K-subgaussian matrix and $\mathcal{K} \subseteq \mathbb{R}^N$ a non-empty closed convex set. Fix $\delta, \eta > 0$, $0 < \tau_1 \leq \tau_2 < \infty$ and $x_0 \in \mathcal{K}$ with $\|x_0\|_{\mathcal{K}} = 1$. For any $t \geq 1$, if m satisfies $m > Ct^2\tilde{K}^2\delta^{-2}\gamma^2(T_{\mathcal{K}}(x_0) \cap \mathbb{S}^{N-1})$, then with probability at least $1 - \exp(-t^2)$ on the realization of A,*

$$\hat{R}(\tau_1; \tau_1 x_0, A, \eta) \leq \frac{1+\delta}{1-\delta} \hat{R}(\tau_2; \tau_2 x_0, A, \eta).$$

Proof of Proposition 4.6.2. Given $0 < \tau_1 \leq \tau_2 < \infty$, let $\tau \in \{\tau_1, \tau_2\}$, define $y(\tau) = A\tau x_0 + \eta z$, and define

$$q(\tau) := A\hat{w}(\tau), \quad \text{where} \quad \hat{w}(\tau) := \hat{x}(\tau; A, y) - \tau x_0, \quad \tau \in \{\tau_1, \tau_2\}.$$

Observe that $q(\tau)$ may be written as

$$q(\tau) \in \arg\min\{\|q - \eta z\|_2 : q \in \tau\mathcal{K}'\},$$
$$\mathcal{K}' := \{A(x - x_0) : x \in \mathcal{K}\}.$$

The set $\mathcal{K}' \subseteq \mathbb{R}^m$ is non-empty, closed and convex, with $0 \in \mathcal{K}'$. In particular,

Lemma 2.1.2 implies

$$\|q(\tau_1)\|_2 \le \|q(\tau_2)\|_2.$$

By [47, Theorem 1.7], for any $t \ge 1$ it holds with probability at least $1 - \exp(-t^2)$ on A that for all $w \in T_{\mathcal{K}}(x_0)$,

$$\sqrt{m} \cdot |\|Aw\|_2 - \|w\|_2| \le Ct\tilde{K}\gamma(T_{\mathcal{K}}(x_0) \cap \mathbb{S}^{N-1}).$$

Accordingly, since $\hat{w}(\tau) \in T_{\mathcal{K}}(x_0)$ for $\tau = \tau_1, \tau_2$, under the assumption on m it holds with probability at least $1 - \exp(-t^2)$

$$\begin{aligned}(1-\delta)\|\hat{w}(\tau_1)\|_2 &\le \|q(\tau_1)\|_2 \\ &\le \|q(\tau_2)\|_2 \le (1+\delta)\|\hat{w}(\tau_2)\|_2.\end{aligned}$$

In particular, $\|\hat{w}(\tau_1)\|_2 \le \frac{1+\delta}{1-\delta}\|\hat{w}(\tau_2)\|_2$. As z was arbitrary, the result follows:

$$\hat{R}(\tau_1, \tau_1 x_0, A, \eta) \le \frac{1+\delta}{1-\delta}\hat{R}(\tau_2; \tau_2 x_0, A, \eta).$$

□

Corollary 4.6.3. *Under the assumptions of Proposition 4.6.2, the optimally tuned worst-case risk for* (LS_τ) *is nearly equivalent to* $R^*(s, A)$*, in the sense that*

$$R^*(s, A) \le \sup_{x \in \Sigma_s^N} \hat{R}(\|x\|_1; x, A, \eta) \le CR^*(s, A).$$

Proof of Corollary 4.6.3. The sup defining the optimally tuned worst-case risk may be decoupled as

$$\sup_{x' \in \Sigma_s^N} \hat{R}(\|x'\|_1; x', A, \eta) = \sup_{\tau > 0} \sup_{x \in \Sigma_s^N \cap \mathbb{S}^{N-1}} \hat{R}(\tau; \tau x, A, \eta). \tag{4.5}$$

Applying a standard scaling property gives the relation:

$$\begin{aligned}\hat{R}(\tau;\tau x,A,\eta) &= \left(\frac{\tau}{\eta}\right)^2 \mathbb{E}\,\|\hat{x}(1;y/\tau,A)-x\|_2^2\\ &= \hat{R}(1;x,A,\eta/\tau).\end{aligned}$$

The lower bound follows trivially from these two observations. To prove the upper bound, we start by observing two facts. First, $\Sigma_s^N \cap \mathbb{S}^{N-1}$ is compact, so there is $x^*(\tau)$ achieving the supremum over the set $\Sigma_s^N \cap \mathbb{S}^{N-1}$ in (4.5). Next, if the supremum over $\tau > 0$ is achieved for $\tau \to \infty$, there is nothing to show, since

$$\begin{aligned}&\sup_{\tau>0}\sup_{x\in\Sigma_s^N\cap\mathbb{S}^{N-1}} \hat{R}(\tau;\tau x,A,\eta)\\ &\quad= \lim_{\tau\to\infty}\sup_{x\in\Sigma_s^N\cap\mathbb{S}^{N-1}} \hat{R}(\tau;\tau x,A,\eta)\\ &\quad= \lim_{\tau\to\infty}\sup_{x\in\Sigma_s^N\cap\partial B_1^N} \hat{R}(1;x,A,\eta/\tau)\\ &\quad= \lim_{\eta\to 0}\sup_{x\in\Sigma_s^N\cap\partial B_1^N} \hat{R}(1;x,A,\eta).\end{aligned}$$

Otherwise, the supremum is achieved for some $0 \leq \tau^* < \infty$. Let $(\tau_i)_{i\in\mathbb{Z}}$ be an arbitrary bi-infinite monotone sequence with $\tau_i \xrightarrow{i\to-\infty} \tau^*$ and $\tau_i \xrightarrow{i\to\infty} \infty$. For any $i \leq j$, Proposition 4.6.2 and properties of the supremum give

$$\begin{aligned}&\sup_{x\in\Sigma_s^N\cap\mathbb{S}^{N-1}} \hat{R}(\tau_i;\tau_i x,A,\eta)\\ &\qquad= \hat{R}(\tau_i;\tau_i x^*(\tau_i),A,\eta)\\ &\qquad\leq C\hat{R}(\tau_j;\tau_j x^*(\tau_i),A,\eta)\\ &\qquad\leq C\hat{R}(\tau_j;\tau_j x^*(\tau_j),A,\eta)\\ &\qquad= C\sup_{x\in\Sigma_s^N\cap\mathbb{S}^{N-1}} \hat{R}(\tau_j;\tau_j x,A,\eta)\end{aligned}$$

As the above chain of inequalities holds for any pair $i < 0$ and $j > 0$, taking

$i \to -\infty$ and $j \to \infty$ gives,

$$\sup_{\tau>0} \sup_{x \in \Sigma_s^N \cap \mathbb{S}^{N-1}} \hat{R}(\tau; \tau x, A, \eta) \leq C \sup_{x \in \Sigma_s^N \cap \mathbb{S}^{N-1}} \hat{R}(\tau_j; \tau_j x, A, \eta)$$
$$\xrightarrow{j \to \infty} C \liminf_{\tau \to \infty} \sup_{x \in \Sigma_s^N \cap \mathbb{S}^{N-1}} \hat{R}(\tau; \tau x, A, \eta).$$

Finally, combining the above with an application of the standard scaling property yields

$$\begin{aligned}
&\sup_{\tau>0} \sup_{x \in \Sigma_s^N \cap \mathbb{S}^{N-1}} \hat{R}(\tau; \tau x, A, \eta) \\
&\quad \leq C \liminf_{\tau \to \infty} \sup_{x \in \Sigma_s^N \cap \mathbb{S}^{N-1}} \hat{R}(\tau; \tau x, A, \eta) \\
&\quad = C \liminf_{\tau \to \infty} \sup_{x \in \Sigma_s^N \cap \partial B_1^N} \hat{R}(1; x, A, \eta/\tau) \\
&\quad = C \liminf_{\eta \to 0} \sup_{x \in \Sigma_s^N \cap \partial B_1^N} \hat{R}(1; x, A, \eta)
\end{aligned}$$

□

Controlling a conditionally Gaussian process

Here we ready two technical results that are used to control the error of the tuned approximation ($\tau = \tau^*$) uniformly with respect to the noise scale $\eta > 0$. First, we specialize a result of [47]. Next, with high probability we control in expectation the extreme values of a conditionally Gaussian process.

Lemma 4.6.4 (Corollary of Theorem 2.2.7)**.** *Fix $\delta, \varepsilon, r > 0$ and let $A \in \mathbb{R}^{m \times N}$ be a normalized K-subgaussian matrix. For a constant $C_\varepsilon > 0$, if*

$$m > C_\varepsilon \delta^{-2} \tilde{K}^2 r^2 s \log\left(\frac{2N}{s}\right), \tag{4.6}$$

it holds with probability at least $1-\varepsilon$ *on the realization of* A *that*

$$\sup_{x\in\mathcal{L}_s(r)} |\|Ax\|_2 - \|x\|_2| < \delta. \tag{4.7}$$

Proof of Lemma 4.6.4. If $s = 0$ the result holds trivially. For $s \geq 1$, this lemma is a straightforward consequence of Theorem 2.2.7. Set $u := \sqrt{\log(2\varepsilon^{-1})}$. Indeed, by that result, it holds with probability at least $1-\varepsilon$ on the realization of A that

$$\begin{aligned}\sup_{x\in\mathcal{L}_s(r)} &|\|Ax\|_2 - \|x\|_2| \\ &\leq C_1 m^{-1/2}\tilde{K}r\left[\mathrm{w}(\Sigma_s^N\cap\mathbb{S}^{N-1}) + u\right],\end{aligned}$$

where C_1 is an absolute constant. By Lemma 2.3.1, there is an absolute constant $C_2 > 0$ so that

$$\mathrm{w}^2(\Sigma_s^N\cap\mathbb{S}^{N-1}) \leq C_2^2 s\log\left(\frac{2N}{s}\right).$$

In particular, (4.7) holds if

$$C_1\tilde{K}m^{-1/2}r\left[C_2\sqrt{s\log\left(\frac{2N}{s}\right)} + u\right] < \delta.$$

Observe that this condition is satisfied if (4.6) holds:

$$\begin{aligned}m &> C_\varepsilon\delta^{-2}\tilde{K}^2r^2 s\log\left(\frac{2N}{s}\right),\\ C_\varepsilon &:= 4C_1^2\cdot\max\left\{\log\left(2\varepsilon^{-1}\right), C_2^2\right\}.\end{aligned}$$

□

Lemma 4.6.5 (Conditionally Gaussian process)**.** *Let* $\mathcal{K}\subseteq\mathcal{K}_s^N\cap\mathbb{S}^{N-1}$ *and suppose* $A\in\mathbb{R}^{m\times N}$ *is a normalized* K*-subgaussian matrix. Let* $z\in\mathbb{R}^m$ *with*

$z_i \overset{iid}{\sim} \mathcal{N}(0,1)$ *and define*

$$f(A,z) := \sup_{x \in \mathcal{K}} \langle Ax, z \rangle.$$

Let $\delta, \varepsilon > 0$ *and* $s \in \mathbb{N}$ *with* $s \geq 1$. *There is an absolute constant* $C_\varepsilon > 0$, *depending only on* ε, *so that if*

$$m > C_\varepsilon \delta^{-2} \tilde{K}^2 s \log(N/s),$$

then with probability at least $1 - \varepsilon$ *on the realization of* A,

$$\mathbb{E}\left[f(A,z) \mid A\right] \leq C_\delta \sqrt{s \log(2N/s)}$$

where $C_\delta > 0$ *is an absolute constant depending only on* δ.

Proof of Lemma 4.6.5. By Lemma 2.3.2,

$$\mathcal{K} \subseteq \mathcal{K}_s^N \cap \mathbb{S}^{N-1} \subseteq \mathcal{L}_s.$$

Therefore, $f(A,z) \leq \sup_{x \in \mathcal{L}_s} \langle Ax, z \rangle$. Furthermore,

$$\mathcal{L}_s - \mathcal{L}_s \subseteq \mathcal{L}_s^*.$$

By Lemma 4.6.4,

$$\max_{j \in [N]} \left| \|A^j\|_2 - 1 \right| \leq \sup_{x \in \mathcal{L}_s^*} \left| \|Ax\|_2 - \|x\|_2 \right| < \delta \tag{4.8}$$

with probability at least $1 - \varepsilon$ if m satisfies

$$m > 32 C_\varepsilon \delta^{-2} \tilde{K}^2 s \log(N/s). \tag{4.9}$$

Next, where $x \in \mathcal{L}_s$, define the random processes

$$X_x := \langle Ax, z \rangle, \qquad z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1);$$
$$Y_x := (1+\delta)\langle x, g \rangle, \qquad g_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

Assume (4.9) holds and condition on the event $\mathcal{A}$ described by (4.8). Then $x - y \in \mathcal{L}_s - \mathcal{L}_s \subseteq \mathcal{L}_s^*$, so

$$\begin{aligned}\mathbb{E}(X_y - X_x)^2 &= \|A(x-y)\|_2^2 \\ &\leq (1+\delta)^2 \|x-y\|_2^2 = \mathbb{E}(Y_y - Y_x)^2.\end{aligned}$$

Namely, conditioned on $\mathcal{A}$, the Sudakov-Fernique inequality (Theorem 2.2.5) gives

$$\begin{aligned}\mathbb{E}\left[f(A,z) \mid A\right] &\leq \mathbb{E}\left[\sup_{x \in \mathcal{L}_s} X_x\right] \leq \mathbb{E}\left[\sup_{x \in \mathcal{L}_s} Y_x\right] \\ &= (1+\delta)\,\mathrm{w}(\mathcal{L}_s) \\ &\leq 2C(1+\delta)\sqrt{s\log(2N/s)},\end{aligned}$$

where $C > 0$ is an absolute constant.

□

Remark 13 (Subgaussianity of $f(A,z) \mid A$). Conditioned on A, Borell-TIS (Theorem 2.2.4) gives subgaussian concentration of $f(A,z)$ about $\mathbb{E}[f(A,z) \mid A]$. In particular,

$$\left\|f(A,z) - \mathbb{E}\left[f(A,z) \mid A\right]\right\|_{\psi_2} \lesssim \sigma_{\mathcal{K}}$$

where, on the event $\mathcal{A}$ as defined in the proof of Lemma 4.6.5,

$$\sigma_{\mathcal{K}}^2 = \sup_{x \in \mathcal{K}} \mathbb{E}\left[|\langle Ax, z\rangle|^2\right] = \sup_{x \in \mathcal{K}} \|Ax\|_2^2 \leq (1+\delta)^2.$$

Note that subgaussianity of $f(A,z)$ about $\mathbb{E}[f(A,z) \mid A]$ can also be established using concentration of Lipschitz functions of Gaussians. Indeed, since $\mathcal{K} \subseteq \mathbb{S}^{N-1}$, for each A it holds that $f(A,z)$ is Lipschitz in z. In fact, one can show that "for most" A, $f(A,z)$ is "nearly" 1-Lipschitz.

4.6.3 Proofs for constrained Lasso sensitivity

Suboptimal choice of τ

The first result required to prove Theorem 4.2.1 concerns the case where (LS_τ) is controlled by a parameter that is too large. Under mild regularity assumptions on the mapping A, we show that this underconstrained problem cannot recover even the least-squares proximal denoising error rate in the limiting low-noise regime. The second result of this section concerns the situation where τ is too small, $\tau < \tau^*$. In this overconstrained problem, the ground truth does not lie in the feasible set and one expects this to be detrimental to recovery performance. We confirm this intuition irrespective of the assumptions on the measurement matrix A.

Lemma 4.6.6 (Underconstrained (LS_τ))**.** *Let $A \in \mathbb{R}^{m\times N}$ and assume that $\dim(\mathrm{null}(A)) > 0$. Given $x_0 \in \mathbb{R}^N, \eta > 0$ and $z \in \mathbb{R}^m$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, let $y := Ax_0 + \eta z$. Suppose $\tau > \|x_0\|_1$. Almost surely on the realization of z,*

$$\lim_{\eta\to 0} \hat{L}(\tau; x_0, A, \eta z) = \infty.$$

Proof of Lemma 4.6.6. Define $\rho := \tau - \tau^*$, where $\tau^* := \|x_0\|_1$. For simplicity, first assume $\mathrm{span}(A) = \mathbb{R}^m$. There exists $\zeta \in \mathbb{R}^N$ such that $A\zeta = z$, and so $A(x_0+\eta\zeta) = Ax_0+\eta z = y$. Moreover, if $\eta < \rho\|\zeta\|_1^{-1}$ then $x_0+\eta\zeta \in \tau B_1^N$. In particular, $\xi := x_0 + \eta\zeta$ solves (LS_τ), because it is feasible and achieves the lowest possible objective value for (LS_τ). Notice $\|\zeta\|_1 < \infty$ almost surely, so for any realization of z, $\eta < \rho\|\zeta\|_1^{-1}$ holds for all η sufficiently small. Specifically, we have constructed ξ solving (LS_τ), and lying on the interior of τB_1^N. Consequently, almost surely there is $\nu \in \mathrm{null}(A)$ so that $\xi + \nu \in \tau B_1^N$ and still $A(\xi + \nu) = y$. Scale ν if necessary so that $\|\xi + \nu\|_1 \in [\frac{1}{2}(\tau + \tau_*), \tau]$.

Then, almost surely on the realization of z,

$$\begin{aligned}\hat{L}(\tau; x_0, A, \eta z) &\geq \eta^{-2}\|\xi + \nu - x_0\|_2^2 \\ &\geq \frac{1}{N\eta^2}\|\xi + \nu - x_0\|_1^2 \\ &\geq \frac{\rho^2}{4N\eta^2} \xrightarrow{\eta \to 0} \infty.\end{aligned}$$

The case $\text{span}(A) \neq \mathbb{R}^m$ is similar. This case is interesting only when $z \in \mathbb{R}^m \setminus \text{span}(A) \neq \emptyset$, otherwise we argue as above. In this setting, define P to be the projection onto the range of A with $P^\perp = (I - P)$ being its orthogonal component. We may re-write the objective of (LS_τ) as

$$\begin{aligned}\|y - Ax\|_2^2 &= \|(P + P^\perp)(y - Ax)\|_2^2 \\ &= \|P(y - Ax) + P^\perp y\|_2^2 \\ &= \|\tilde{y} - Ax\|_2^2 + \|P^\perp y\|_2^2\end{aligned}$$

where $\tilde{y} := Py$. Therefore, when $z \notin \text{span}(A)$, solving (LS_τ) is equivalent to solving

$$\arg\min \{\|Py - Ax\|_2 : \|x\|_1 \leq \tau\}. \tag{$\star$}$$

By construction, $\tilde{y} = Py \in \text{range}(A)$, so we may apply the same argument as above to the program ($\star$), implying

$$\lim_{\eta \to 0} \hat{L}(\tau; x_0, A, \eta z) = \infty.$$

□

Lemma 4.6.7 (Overconstrained (LS_τ))**.** *Fix $\tau < \tau^*$. Almost surely on the realization z,*

$$\lim_{\eta \to 0} \hat{L}\left(\tau; x_0, A, \eta z\right) = \infty.$$

Proof of Lemma 4.6.7. Let $\rho := \tau^* - \tau > 0$. For any solution ξ to (LS_τ),

one has

$$\eta^{-2}\|\xi - x_0\|_2^2 \geq \frac{\rho^2}{N\eta^2}.$$

By definition, the desired result follows immediately:

$$\lim_{\eta\to 0} \hat{L}(\tau; x_0, A, \eta z) \geq \eta^{-2}\|\xi - x_0\|_2^2 \geq \frac{\rho^2}{N\eta^2} \xrightarrow{\eta\to 0} \infty.$$

□

Uniform control over noise scales

In this section, we control (LS_τ) in the optimally tuned setting, uniform over the noise scale η. Specifically, for any $x_0 \in \Sigma_s^N$ we control the expected error of recovery for (LS_τ) uniformly over the noise scale $\eta > 0$. The results of § 4.6.2 are crucial for this purpose.

Proposition 4.6.8 (Uniform over noise scale)**.** *Let $0 \leq s < N < \infty$ be integers and let $m \in \mathbb{N}$. Let $A \in \mathbb{R}^{m\times N}$ be a normalized K-subgaussian matrix, and fix $\delta, \varepsilon > 0$. Suppose that $y = Ax_0 + \eta z$ for $\eta > 0$ and $z \in \mathbb{R}^m$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$. With probability at least $1 - \varepsilon$ on the realization of A, there exist constants $C_\delta, C_\varepsilon > 0$ so that if*

$$m > C_\varepsilon \delta^{-2} \tilde{K}^2 s \log\left(\frac{N}{2s}\right),$$

then for all $\eta > 0$:

$$\mathbb{E}\left[\|\hat{x} - x_0\|_2^2 \mid A\right] \leq C_\delta \eta^2 s \log\left(\frac{N}{2s}\right).$$

where $\hat{x} = \hat{x}(\tau^*)$ solves (LS_τ) with $\tau = \tau^* := \|x_0\|_1$.

Proof of Proposition 4.6.8. If $s = 0$, the result holds trivially as, by construction, $\|\hat{x} - x_0\|_2 = 0$ almost surely. Suppose $s \geq 1$. By definition of $\hat{x}$,

where $h := \hat{x} - x_0$,

$$\|A\hat{x} - y\|_2^2 \leq \|Ax_0 - y\|_2^2 \quad \Longrightarrow \quad \|Ah\|_2^2 \leq 2\eta\langle Ah, z\rangle.$$

Step 1: Lower bound $\|Ah\|_2$ with high probability. Note that $\|Aw\|_2 = \|w\|_2\|A\hat{w}\|_2$ for $w \neq 0$ where $\hat{w} := w/\|w\|_2$. By Lemma 4.6.4, there is an event $\mathcal{A}_1$ with $\mathbb{P}(\mathcal{A}_1) \geq 1 - \varepsilon/2$ on which

$$\sup_{x \in \mathcal{K}_{4s}^N} |\|Ax\|_2 - \|x\|_2| \leq \sup_{x \in \mathcal{L}_{4s}} |\|Ax\|_2 - \|x\|_2| < \delta_1$$

if m satisfies

$$m > 16C'_\varepsilon \delta_1^{-2} \tilde{K}^2 s \log\left(\frac{N}{2s}\right). \tag{4.10}$$

Specifically, $h \in \mathcal{J}_{4s}^N$ by Lemma 2.3.4, meaning $\hat{h} \in \mathcal{J}_{4s}^N \cap \mathbb{S}^{N-1} \subseteq \mathcal{K}_{4s}^N$ if $h \neq 0$. So, conditioning on $\mathcal{A}_1$ and enforcing (4.10), one has

$$\|Ah\|_2^2 = \|h\|_2^2\|A\hat{h}\|_2^2 \geq \|h\|_2^2 \left(\|\hat{h}\|_2 - \delta_1\right)^2 \geq (1 - \delta_1)^2\|h\|_2^2. \tag{4.11}$$

The inequality $(1 - \delta_1)^2\|h\|_2^2 \leq \|Ah\|_2^2$ holds also for $h = 0$.
Step 2a: Upper bound $\langle Ah, z\rangle$. Again using that $h \in \mathcal{J}_{4s}^N$,

$$2\eta\langle Ah, z\rangle \leq 2\eta\|h\|_2 \sup_{\hat{h} \in \mathcal{K}_{4s}^N \cap \mathbb{S}^{N-1}} \langle A\hat{h}, z\rangle. \tag{4.12}$$

Step 2b: Control the latter quantity in expectation. By Lemma 4.6.5, there is $C''_\varepsilon > 0$ so that for

$$m > 4C''_\varepsilon \delta^{-2} \tilde{K}^2 s \log\left(\frac{N}{4s}\right), \tag{4.13}$$

there is an event $\mathcal{A}_2$ holding with probability at least $1 - \varepsilon/2$, on which there

is a constant $C_\delta > 0$ such that

$$\mathbb{E}\left[\sup_{\hat{h}\in\mathcal{K}_{4s}^N} \langle A\hat{h}, z\rangle \mid A\right] \leq 2C_\delta\sqrt{s\log\left(\frac{N}{2s}\right)}.$$

Step 3: Now combine steps 1 and 2a. Assume m simultaneously satisfies (4.10) and (4.13), and condition on $\mathcal{A}_1 \cap \mathcal{A}_2$, which holds with probability at least $1-\varepsilon$. Combining (4.11) and (4.12), and letting $\delta_1 := 1 - 2^{-1/2}$ gives

$$\|h\|_2 \leq 4\eta \sup_{\hat{h}\in\mathcal{K}_{4s}^N} \langle A\hat{h}, z\rangle.$$

Take expectation of both sides and bound the quantity by applying step 2b. This yields,

$$\mathbb{E}\left[\|h\|_2 \mid A\right] \leq 8C_\delta\eta\sqrt{s\log\left(\frac{N}{2s}\right)}.$$

Note that by setting $C_\varepsilon := \max\{\frac{32C'_\varepsilon}{3-2\sqrt{2}}, 4C''_\varepsilon\}$, it suffices to require

$$m > C_\varepsilon\delta^{-2}\tilde{K}^2 s\log\left(\frac{N}{2s}\right).$$

Alternatively, one may also apply a standard fact for subgaussian random variables. Recall as in (2.2), $|||X_w|||_{K_{4s}^N} := \sup_{w\in K_{4s}^N} X_w$. Then

$$\left\| |||X_w|||_{K_{4s}^N} - \mathbb{E}\,|||X_w|||_{K_{4s}^N} \right\|_{\psi_2} \leq \sigma^2_{K_{4s}^N}$$

by Theorem 2.2.4. So, there is an absolute constant $C > 0$ such that on $\mathcal{A}_1$,

$$\begin{aligned}
&\left\| |||X_w|||_{K_{4s}^N} - \mathop{\mathbb{E}}_z |||X_w|||_{K_{4s}^N} \right\|^2_{L^2} \\
&= \mathop{\mathbb{E}}_z |||X_w|||^2_{K_{4s}^N} - \left(\mathop{\mathbb{E}}_z |||X_w|||_{K_{4s}^N}\right)^2 \\
&\leq C\sigma^2_{K_{4s}^N} \leq C(1+\delta_1)^2.
\end{aligned}$$

where $X_w := \langle Aw, z\rangle$ conditioned on A. In particular, choosing instead

$\delta_1 := 1 - 2^{-1/4}$,

$$\mathbb{E}\left[\|h\|_2^2 \mid A\right] \leq 8\eta^2 \left(4C_\delta^2 s \log \frac{N}{2s} + C\sqrt{2}\right).$$

Rearranging, and observing that the right hand term in parentheses is small relative to the left hand term, we may obtain a new absolute constant $C_\delta > 0$ depending only on δ such that

$$\eta^{-2}\,\mathbb{E}\left[\|\hat{x} - x_0\|_2^2 \mid A\right] \leq C_\delta s \log \frac{N}{2s}.$$

□

Remark 14. In the proof above, no attempt was made to optimize constants. In fact, several simplifications were made for clarity of presentation, which in turn resulted in larger than necessary constants.

Remark 15 (Uniform control over noise scale and signal class). Observe that the result above is uniform over noise scale $\eta > 0$ and signal $x_0 \in \Sigma_s^N$. In particular, we could have written (conditioning on A),

$$\sup_{\eta>0} \sup_{x_0 \in \Sigma_s^N} \hat{R}(\tau^*; x_0, A, \eta) \leq C_\delta s \log \frac{N}{2s}.$$

Optimal choice of τ and phase transition

Here, we synthesize the technical results of § 4.6.3 to show that, with high probability on the realization of A, (LS_τ) achieves order-optimal risk in the limiting low-noise regime when m is sufficiently large and $\tau = \tau^*$.

Lemma 4.6.9 (Tuned (LS_τ))**.** *Fix $\delta, \varepsilon > 0$ and let $A \in \mathbb{R}^{m\times N}$ be a normalized K-subgaussian matrix. For $s \in \mathbb{N}$ fixed with $0 \leq s \leq m$, suppose $x_0 \in \Sigma_s^N$ and $\eta > 0$. If m satisfies*

$$m > C'_\varepsilon \delta^{-2} \tilde{K}^2 s \log \frac{N}{2s},$$

then, with probability at least $1 - \varepsilon$ on the realization A, there exist constants

$0 < c_\delta < C_\delta < \infty$ *such that*

$$c_\delta \cdot s \log\left(\frac{N}{s}\right) \leq \lim_{\eta \to 0} \sup_{x_0 \in \Sigma_s^N} \hat{R}(\tau^*; x_0, N, \eta) \leq C_\delta \cdot s \log\left(\frac{N}{2s}\right).$$

Proof of Lemma 4.6.9. For simplicity of the proof, we assume $\tilde{K}^2 = 1$.
Upper bound: Given $\delta, \varepsilon_1 > 0$, assume

$$m > C_{\varepsilon_1} \delta^{-2} s \log \frac{N}{2s}.$$

With probability at least $1 - \varepsilon_1$ on the realization of A, by Proposition 4.6.8, for any $x_0 \in \Sigma_s^N$ and $\eta > 0$,

$$\hat{R}(\tau^*; x_0, A, \eta) \leq C_\delta \cdot s \log \frac{N}{2s}.$$

In particular,

$$\lim_{\eta \to 0} \sup_{x_0 \in \Sigma_s^N} \hat{R}(\tau^*; x_0, A, \eta) \leq C_\delta \cdot s \log \frac{N}{2s}.$$

Lower bound: From Corollary 4.6.3 and [17, Theorem 1],

$$\sup_{x_0 \in \Sigma_s^N} \hat{R}(\tau^*; x_0, A, \eta) \geq \inf_{x_*} \sup_{x_0 \in \Sigma_s^N} \eta^{-2} \, \mathbb{E} \, \|x_* - x_0\|_2^2 \geq \frac{C_1 N}{\|A\|_F^2} s \log\left(\frac{N}{s}\right).$$

In particular,

$$\lim_{\eta \to 0} \sup_{x_0 \in \Sigma_s^N} \hat{R}(\tau^*; x_0, A, \eta) \geq \frac{C_1 N}{\|A\|_F^2} s \log\left(\frac{N}{s}\right).$$

Now, $\mathbb{E} \, \|A\|_F^2 = N$, and $\|A\|_F^2$ admits subexponential concentration around its expectation by Bernstein's inequality [77, Corollary 2.8.3]. Therefore, with probability at least $1 - \varepsilon_2$ on the realization of A, there is a constant

$c_\delta > 0$ depending only on C_1 and δ such that

$$\lim_{\eta \to 0} \sup_{x_0 \in \Sigma_s^N} \hat{R}(\tau^*; x_0, A, \eta) \geq c_\delta \cdot s \log\left(\frac{N}{s}\right),$$

under the condition that

$$m \geq C\delta^{-2} N^{-1} \log \frac{2}{\varepsilon_2}.$$

Combine: Finally, set $\varepsilon_1 = \varepsilon_2 = \varepsilon/2$. Under the assumptions on m, with probability at least $1 - \varepsilon$ on the realization of A it holds that

$$\begin{aligned} c_\delta \cdot s \log\left(\frac{N}{s}\right) &\leq \lim_{\eta \to 0} \sup_{x_0 \in \Sigma_s^N} \hat{R}(\tau^*; x_0, A, \eta) \\ &\leq C_\delta \cdot s \log \frac{N}{2s}. \end{aligned}$$

□

We conclude this section with the proof of Theorem 4.2.1 which combines Lemma 4.6.9 and the results of § 4.6.3. Namely, even when m is sufficiently large, (LS_τ) admits order-optimal risk in the limiting low-noise regime only when the governing parameter is chosen optimally.

Proof of Theorem 4.2.1. This result follows immediately from the lemmata of this section. Indeed, a direct application of Lemma 4.6.9 gives

$$\begin{aligned} c_\delta \cdot s \log\left(\frac{N}{s}\right) &\leq \lim_{\eta \to 0} \sup_{x \in \Sigma_s^N} \hat{R}(\tau^*; x, A, \eta) \\ &\leq C_\delta \cdot s \log \frac{N}{2s}. \end{aligned}$$

Otherwise, $\tau \neq \tau^*$. First, if $\tau < \tau^*$, then Lemma 4.6.7 immediately implies

$$\lim_{\eta \to 0} \hat{L}(\tau; x_0, A, \eta z) = \infty.$$

Otherwise, assume $\tau > \tau^*$. In order to apply Lemma 4.6.6, A must satisfy $\dim(\mathrm{null}(A)) > 0$, which holds trivially, as $m < N$. In particular,

Lemma 4.6.6 implies almost surely on (A, z),

$$\lim_{\eta \to 0} \hat{L}(\tau; x_0, A, \eta z) = \infty.$$

□

Remark 16. The proof for Theorem 4.2.1 proceeds whether z be deterministic (say with fixed norm $\|z\|_2 = \sqrt{m}$) or have entries $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. We have presented it this way so that the assumption is consistent with the implicit assumption on the noise for the result concerning $\hat{R}(\tau^*; x_0, A, \eta)$.

4.6.4 Proofs for basis pursuit suboptimality

Suboptimal regime for underconstrained basis pursuit

This section contains the proof for Lemma 4.4.1 in § 4.4.1.

Proof of Lemma 4.4.1. It suffices to prove this result for the best choice of σ and any $x \in \Sigma_s^N$. In particular, choose $x_0 \in \Sigma_s^N$ having at least one non-zero entry, and for which the non-zero entries have magnitude satisfying $|x_{0,j}| \geq C\eta\sqrt{m}$, $j \in \text{supp}(x_0) \subseteq [N]$. For this choice of x_0, let $y = Ax_0 + \eta z$ and define the event $\mathcal{F} := \{\|y\|_2 \leq \sigma\}$.

For any $\sigma \geq \eta\sqrt{m}$, re-choose $x_0 \in \Sigma_s^N$ if necessary so that moreover $\mathbb{P}(\mathcal{F}^C) \geq 0.99$. Restricting to $\mathcal{F}^C$, the solution to (BP_σ) satisfies, by the KKT conditions [11],

$$\eta^2 m \leq \sigma^2 = \|Ah\|_2^2 - 2\eta\langle Ah, z\rangle + \eta^2\|z\|_2^2.$$

By Lemma 4.6.4, it holds with probability at least $1 - \varepsilon$ on the realization of A that

$$(1+\delta)^2\|h\|_2^2 \geq \|Ah\|_2^2 \geq \eta^2(m - \|z\|_2^2) + 2\eta\langle Ah, z\rangle$$

Define the event $\mathcal{Z}_{\leq} := \{\|z\|_2^2 \leq m - 2\sqrt{m}\}$ and observe that further re-

stricting to $\mathcal{F}^C \cap \mathcal{Z}_{\leq}$ thereby gives

$$\begin{aligned}(1+\delta)^2\|h\|_2^2 &\geq 2\eta^2\sqrt{m} - 2\eta\|h\|_2 f(A,z) \\ &\geq 2\eta^2\sqrt{m} - \frac{1}{2}\|h\|_2^2 - 2\eta^2 f^2(A,z),\end{aligned}$$

where $f(A,z)$ is defined as in Lemma 4.6.5 with $\mathcal{K} = \mathcal{K}_{2s}^N \cap \mathbb{S}^{N-1}$. Indeed, where $\hat{h} = h/\|h\|_2$, one has $\langle A\hat{h}, z\rangle \leq f(A,z)$ since $\hat{h} \in \mathcal{K}_{2s}^N \cap \mathbb{S}^{N-1}$ with high probability on the realization of A. This yields the following bound on the risk:

$$\begin{aligned}&\tilde{R}(\sigma; x_0, A, \eta) \\ &\quad\geq \eta^{-2}\mathop{\mathbb{E}}_{z}\left[\|h\|_2^2 \cdot \chi\left(\mathcal{F}^C \cap \mathcal{Z}_{\leq}\right)\right] \\ &\quad\geq C_\delta \mathop{\mathbb{E}}_{z}\left[\left(\sqrt{m} - f^2(A,z)\right) \cdot \chi\left(\mathcal{F}^C \cap \mathcal{Z}_{\leq}\right)\right] \\ &\quad= C_\delta\sqrt{m}\mathbb{P}\left(\mathcal{F}^C \cap \mathcal{Z}_{\leq}\right) - C_\delta \mathop{\mathbb{E}}_{z}\left[f^2(A,z) \cdot \chi\left(\mathcal{F}^C \cap \mathcal{Z}_{\leq}\right)\right] \\ &\quad\geq C_\delta\sqrt{m}\mathbb{P}\left(\mathcal{F}^C \cap \mathcal{Z}_{\leq}\right) - C_\delta \mathop{\mathbb{E}}_{z} f^2(A,z) \qquad (4.14)\end{aligned}$$

Finally, we bound $\mathbb{E}_z\, f^2(A,z) = \mathbb{E}[f^2(A,z) \mid A]$. With high probability on the realization of A:

$$\mathbb{E}[f^2(A,z) \mid A] \leq C\,\mathbb{E}[f(A,z) \mid A]^2 \leq C_\delta s\log(N/s).$$

Above, we have first used [77, Exercise 7.6.1] followed by an application of Lemma 4.6.5. Another way to see this would be through the successive application of Remark 13 and Lemma 4.6.5, noting that $f(A,z) - \mathbb{E}[f(A,z) \mid A]$ is a centered subgaussian random variable.

Consequently, using that $\mathbb{P}(\mathcal{F}^C \cap \mathcal{Z}_{\leq}) \geq C$, (4.14) becomes

$$\tilde{R}(\sigma; x_0, A, \eta) \geq C_\delta\left(\sqrt{m} - s\log(N/s)\right). \qquad (4.15)$$

The result follows trivially from the definition of sup and by the initial assumption on m.

□

Suboptimal regime for overconstrained basis pursuit

In this section, we show that $\tilde{R}(\sigma; x_0, A, \eta)$ is suboptimal for $\sigma \leq \eta\sqrt{m}$. To the chagrin of the beleaguered reader, the proofs in this section require several technical lemmata, some assumptions and notation. In particular, we state and prove the technical lemmata required for the main lemma, Lemma 4.6.11 in § 4.6.4.

The flow of this section will proceed as follows. After establishing required preliminary details, we state and prove results concerning the ability of (BP_σ) to recover the 0 vector from noisy random measurements. The results exhibit a regime in which $\tilde{R}(\sigma; x_0, A, \eta)$ may be lower-bounded in the case where $\sigma = \eta\sqrt{m}$. Then, we proceed by showing that (BP_σ) performs no better if σ is allowed to be smaller. In particular, we obtain lower bounds on $\tilde{R}(\sigma; x_0, A, \eta)$ for $\sigma \leq \eta\sqrt{m}$. Motivation for this latter result is readily observed by a re-phrasing of the projection lemma in Proposition 4.6.10.

Preliminaries. For $z \in \mathbb{R}^m$ and $\sigma > 0$ define the set of feasible points by $F(z;\sigma) := \{q \in \mathbb{R}^m : \|q - z\|_2^2 \leq \sigma^2\}$ and denote $F := F(z;\sqrt{m})$. For a matrix $A \in \mathbb{R}^{m\times N}$, denote $B_{1,A} := \{Ax \in \mathbb{R}^m : x \in B_1^N\}$, and define the gauge of $B_{1,A}$ by

$$\begin{aligned}\|q\|_{1,A} &:= \inf\{\|x\|_1 : Ax = q, x \in \mathbb{R}^N\} \\ &= \inf\{\lambda > 0 : q \in \lambda B_{1,A}\}. \end{aligned} \tag{4.16}$$

Recall a gauge is nonnegative, positively homogeneous, convex and vanishes at the origin. Moreover, note that $B_{1,A}$ is a random set, and so $\|\cdot\|_{1,A}$ is random. Now, for a matrix $A \in \mathbb{R}^{m\times N}$, $z \in \mathbb{R}^m$ and $\sigma > 0$, define the program

$$\tilde{q}(\sigma; A, z) := \arg\min\big\{\|q\|_{1,A} : q \in F(z;\sigma)\big\}, \tag{BQ_σ}$$

where $\|\cdot\|_{1,A}$ is defined as in (4.16). Where clear, we omit notating the dependence of $\tilde{q}(\sigma; A, z)$ on A and z, writing simply $\tilde{q}(\sigma)$.

With the above notation, we define an admissible ensemble. The el-

ements of an admissible ensemble will be used to state the main lemma, Lemma 4.6.11. The technical arguments characterizing an admissible ensemble appear in § 4.6.4.

Definition 7 (Admissibile ensemble)**.** Let $0 \leq s < N$ be integers, and let $m : \mathbb{N} \to \mathbb{N}$ be an integer-valued function mapping $N \mapsto m(N)$ such that $\lim_{N\to\infty} m(N)/N = \gamma \in (0,1)$. For any $0 < \theta < \min\{1-\gamma, \gamma\}$, define $N_\theta \geq 1$ to be the least integer such that for all $N \geq N_\theta$,

$$\left| \frac{m(N)}{N} - \gamma \right| < \theta.$$

Where $N \geq 2$, let $A(N)$ be a family of normalized K-subgaussian matrices $A = A(N) \in \mathbb{R}^{m(N)\times N}$. Define $N_* := \max\{N_\theta, N_{\mathrm{RIP}}\}$ where $N_{\mathrm{RIP}} \geq 1$ is the least positive integer such that for all $N \geq N_{\mathrm{RIP}}$,

$$m(N) \geq C_\varepsilon \delta^{-2} \tilde{K}^2 s \log \frac{2N}{s}.$$

where $\delta, \varepsilon > 0$ are fixed in advance.

Let $z = z(N) \in \mathbb{R}^{m(N)}$ with $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Define $F = F(z; \sqrt{m(N)}) = \{q \in \mathbb{R}^{m(N)} : \|q - z\|_2^2 \leq m(N)\}$ and omit writing explicitly its dependence on N, unless necessary. Define $\alpha_1 = \alpha_1(N) := a_1 m(N)^{1/4}$ for some dimension-independent constant $a_1 > 0$; $\lambda = \lambda(N) := L\sqrt{\frac{m(N)}{\log N}}$ for some dimension-independent constant $L > 1$; and

$$K_1 = K_1(N) := \lambda(N) B_{1,A} \cap \alpha_1(N) B_2^{m(N)},$$
$$K_2 = K_2(N) := \lambda(N) B_{1,A} \cap \alpha_2(N) B_2^{m(N)},$$

where $0 < \alpha_2 = \alpha_2(N) \leq \alpha_1$ will be quantified in Proposition 4.6.18. Lastly, define the following random processes. For $g \in \mathbb{R}^m$ with $g_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, let

$$X_1 := \sup_{x \in K_1} |\langle x, g \rangle|, \qquad X_2 := \sup_{x \in K_2} |\langle x, g \rangle|.$$

Thus we define an $(s, m(N), N, \delta, \varepsilon, \theta)$-admissible ensemble as the collection $(A(N), z(N), K_1(N), K_2(N), X_1, X_2)$ satisfying the conditions just de-

scribed, defined for all $N \geq N_*$. This collection will generally be abbreviated to $(A, z, K_1, K_2, X_1, X_2)$ where clear.

Where possible, we simplify notation by omitting explicit dependence on arguments. For example, if N is fixed, then we may refer to $m(N)$ simply as m. Note, however, that for any $N \geq 2$, K_1 and K_2 always depend on $\alpha_1 = \alpha_1(N)$ and $\alpha_2 = \alpha_2(N)$, respectively. Further observe that K_1 and K_2 are random, as they depend on the matrix A. Observe that N_θ depends on θ, γ and $m(\cdot)$, and omit writing explicitly its dependence on the latter two; we assume $m(\cdot)$ and γ are fixed in advance. Requiring $N \geq N_{\text{RIP}}$ is the key condition on $m(N)$ so that Lemma 4.6.4 holds. Clearly, N_{RIP} depends on the parameters $\delta, \varepsilon, K, s$ and the function $m(\cdot)$; for simplicity of presentation we omit writing explicitly its dependence on these parameters. Finally, note that the parameters on which N_* depends are exactly those for which N_θ and N_{RIP} depend.

Proposition 4.6.10. *Let $z \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times N}$ be a normalized K-subgaussian matrix with $1 \leq m < N$. If $0 < \sigma_1 < \sigma_2 < \infty$ and $\tilde{q}(\sigma)$ solves (BQ_σ) then almost surely on (A, z),*

$$\|\tilde{q}(\sigma_1)\|_2 \geq \|\tilde{q}(\sigma_2)\|_2.$$

Proof of Proposition 4.6.10. The result follows by Corollary 2.1.3, because $\|\cdot\|_{1,A}$ is a gauge. □

The geometric lemma. Next, we state a lemma with a geometric flavour, Lemma 4.6.11, which is the main workhorse for proving suboptimality of $\tilde{R}$ in the overconstrained setting. It is a generalization of [9, Lemma 6.2].

Lemma 4.6.11 (Geometric Lemma)**.** *Fix $\delta, \varepsilon_1, \varepsilon_2 > 0$ and $\theta \in (0, \gamma)$. Given an $(s, m, N, \delta, \varepsilon_1, \theta)$-admissible ensemble, there is a choice of $a_1 > 0$ defining $\alpha_1(N)$; $L > 1$ defining $\lambda(N)$; an integer $N_0 \geq N_*$; and absolute constants $p, k > 0$, so that the following occurs. For all $N \geq N_0$, with probability at least $1 - \varepsilon_1$ on the realization of A, there is an event $\mathcal{E} := \mathcal{E}(\varepsilon_1, \varepsilon_2)$ for z on*

which

$$\text{1. } K_1 \cap F \neq \emptyset, \qquad \text{2. } K_2 \cap F = \emptyset,$$
$$\text{3. } \alpha_2 > CN^p, \qquad \text{4. } \mathbb{P}(\mathcal{E}) > k.$$

Above, k depends on N_0 and ε_2 only; p on δ, γ and θ only.

Proof of Lemma 4.6.11. For constants $0 < C_2 < C_1 < \infty$, define the events

$$\mathcal{Z}_< := \{\|z\|_2^2 \leq m + C_1\sqrt{m}\}$$
$$\mathcal{Z}_> := \{\|z\|_2^2 \geq m + C_2\sqrt{m}\}.$$

By Propositions 4.6.17 and 4.6.20, there is an integer $N_0 \geq N_*$ (select the larger of the two bestowed by each result), and respective events, $\mathcal{E}_1, \mathcal{E}_2$, so that with probability at least $1 - \varepsilon_1$ on the realization of A,

$$\mathbb{P}(\mathcal{E}_1) \geq \mathbb{P}(\mathcal{Z}_<) - \varepsilon_2 \qquad \mathbb{P}(\mathcal{E}_2) \geq \mathbb{P}(\mathcal{Z}_>) - \varepsilon_2.$$

In particular, for $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2$, choose a largest such absolute constant $k := k(N_0, C_1, C_2, \varepsilon_2) > 0$ so that

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq \mathbb{P}(\mathcal{Z}_< \cap \mathcal{Z}_>) - 2\varepsilon_2 \geq k.$$

As per Proposition 4.6.17 and Proposition 4.6.20, conditioning on $\mathcal{E}$ and letting $N \geq N_0$ gives $K_1 \cap F \neq \emptyset$ and $K_2 \cap F = \emptyset$ with probability at least $1 - \varepsilon_1$ on the realization of A, as desired. In this regime, that there exists $p > 0$ satisfying $\alpha_2 = \alpha_2(N) \geq CN^p$ is a consequence of Proposition 4.6.19. One need simply select the largest p satisfying for all $N \geq N_0$:

$$CN^p\sqrt{\log N} \leq C_{\delta,\gamma,L,\theta} N^{d/2}.$$

Thus, for all $N \geq N_0$, with probability at least $1 - \varepsilon$ on the realization of A there exists an event $\mathcal{E}$ for z on which all four of the desired criteria hold.

□

Implications for overconstrained basis pursuit. Finally, we state the main results of this section. The first result, Lemma 4.6.12, uses the geometric lemma to show that there exists a regime in which $\tilde{R}$ is suboptimal in the setting where $x_0 = 0$ and $\sigma = \eta\sqrt{m}$. From there, we show in Lemma 4.6.13 that $\tilde{R}$ is no better if σ is any larger. This is enough to state a maximin suboptimality result for (BP_σ), with σ restricted to $(0, \eta\sqrt{m}]$, in Theorem 4.6.14. Notably, this result is stronger than the analogous minimax statement, which necessarily follows from the maximin result.

Lemma 4.6.12 (Lower bound $\tilde{R}(\eta\sqrt{m}; 0, A, \eta)$)**.** *Fix $\delta, \varepsilon, \eta > 0$ and suppose $m : \mathbb{N} \to \mathbb{N}$ satisfies $m(N)/N \to \gamma \in (0, 1)$. There is $N_0 \in \mathbb{N}$ and an absolute constant $p > 0$ so that for all $N \geq N_0$, if $A \in \mathbb{R}^{m(N)\times N}$ is a normalized K-subgaussian matrix, then with probability at least $1 - \varepsilon$ on the realization of A,*

$$\tilde{R}(\eta\sqrt{m}; 0, A, \eta) \geq C_{\delta,\gamma,K} N^p.$$

Proof of Lemma 4.6.12. By a simple scaling argument, it suffices to assume $\eta = 1$. Consider an $(s, m, N, \delta, \varepsilon, \theta)$-admissible ensemble. By Lemma 4.6.11, there is a choice of $a_1 > 0$ for $\alpha_1(N)$ and $L > 1$ for $\lambda(N)$, an integer $N_0 \geq N_*$ and absolute constants $k, p > 0$ so that with probability at least $1-\varepsilon/2$ on the realization of A, there is an event $\mathcal{E}$ for z on which $K_1 \cap F \neq \emptyset$, $K_2 \cap F = \emptyset$, and for which $\mathbb{P}(\mathcal{E}) \geq k_3$. Where $\tilde{q}$ solves (BQ_σ), observe that $\tilde{q} = A\tilde{x}(\sqrt{m})$ and moreover, by construction, $\tilde{q} \in (K_1 \setminus K_2) \cap F$. In particular,

$$\|\tilde{q}\|_{1,A} \leq \lambda, \qquad \alpha_2 \leq \|\tilde{q}\|_2 \leq \alpha_1.$$

By Corollary 2.2.8 and our initial assumptions,

$$\begin{aligned}\|A\| &\leq 1 + C\tilde{K}\left(1 + \sqrt{\frac{N}{m}}\right) \\ &\leq 1 + C\tilde{K}\left(1 + (\gamma - \theta)^{-1/2}\right) = C_{\gamma,K,\theta},\end{aligned}$$

with probability at least $1-C\exp(-m)$. Note, by re-choosing N_0 if necessary,

$$\begin{aligned}1 - C\exp(-m) &\geq 1 - C\exp(-N(\gamma-\theta)) \\ &\geq 1 - C_{\gamma,\theta}\exp(-N_0) \geq 1 - \varepsilon/2.\end{aligned}$$

In particular, for $N \geq N_0$, with probability at least $1-\varepsilon$ on the realization A, it holds with probability at least k on z that

$$\alpha_2 \leq \|\tilde{q}\|_2 \leq \|A\|\|\tilde{x}(\sqrt{m})\|_2 \leq C_{\gamma,K,\theta}\|\tilde{x}(\sqrt{m})\|_2.$$

On the same event, by item 3 of Lemma 4.6.11, there is an absolute constant $p > 0$ so that $\alpha_2 \geq C_{\delta,\gamma,L,\theta}N^p$, whence

$$\|\tilde{x}(\sqrt{m})\|_2 \geq C_{\delta,\gamma,K,L,\theta}N^p.$$

Finally, this immediately implies that for $N \geq N_0$, with probability at least $1-\varepsilon_1$ on the realization of A,

$$\tilde{R}(\sqrt{m}; 0, N, 1) \geq \mathbb{E}\left[\|\tilde{x}(\sqrt{m})\|_2^2 \mid \mathcal{E}\right]\mathbb{P}(\mathcal{E}) \geq C_{\delta,\gamma,K,L,\theta}kN^p.$$

□

Lemma 4.6.13 (Lower bound $\tilde{R}(\sigma; 0, A, \eta)$, $\sigma < \eta\sqrt{m}$)**.** *Fix $\delta, \varepsilon, \eta > 0$ and suppose $m : \mathbb{N} \to \mathbb{N}$ satisfies $m(N)/N \to \gamma \in (0,1)$. There is $N_0 \in \mathbb{N}$ and absolute constant $p > 0$ so that for all $N \geq N_0$, if $A \in \mathbb{R}^{m(N)\times N}$ is a normalized K-subgaussian matrix, it holds with probability at least $1-\varepsilon$ on the realization of A that for any $0 < \sigma \leq \eta\sqrt{m}$,*

$$\tilde{R}(\sigma; 0, A, \eta) \geq C_{\delta,\gamma,K}N^p.$$

Proof of Lemma 4.6.13. The proof of this result is nearly identical to that of Lemma 4.6.12. The crucial difference is its use of Proposition 4.6.10, using which one argues

$$\alpha_2 \leq \|\tilde{q}(\sqrt{m})\|_2 \leq \|\tilde{q}(\sigma)\|_2 \leq C_{\gamma,K,\theta}\|\tilde{x}(\sigma)\|_2$$

to show, in the appropriate regime, that $\|\tilde{x}(\sigma)\|_2 \geq C_{\delta,\gamma,K,L,\theta} N^p$. □

Theorem 4.6.14 (Overconstrained maximin)**.** *Fix $\delta, \varepsilon, \eta > 0$ and suppose $m : \mathbb{N} \to \mathbb{N}$ satisfies $m(N)/N \to \gamma \in (0,1)$. For any $s \geq 0$, there is an integer $N_0 \in \mathbb{N}$ and an absolute constant $p > 0$ so that for all $N \geq N_0$, if $A \in \mathbb{R}^{m(N) \times N}$ is a normalized K-subgaussian matrix, then it holds with probability at least $1 - \varepsilon$ on the realization of A that*

$$\sup_{x \in \Sigma_s^N} \inf_{\sigma \leq \eta\sqrt{m}} \tilde{R}(\sigma; x, A, \eta) \geq C_{\delta,\gamma,K,\theta} N^p.$$

Proof of Theorem 4.6.14. By a scaling argument, it suffices to consider the case $\eta = 1$. Establishing an admissible ensemble and using Lemma 4.6.13, there is $N_0 \geq N_*$ such that for any $N \geq N_0$, with probability at least $1 - \varepsilon$ on A,

$$\sup_{x \in \Sigma_s^N} \inf_{\sigma \leq \sqrt{m}} \tilde{R}(\sigma; x, A, 1) \geq \inf_{\sigma \leq \sqrt{m}} \tilde{R}(\sigma; 0, A, 1) \geq C_{\delta,\gamma,K,\theta} N^p.$$

□

Technical lemmata for overconstrained basis pursuit

The six lemmata of this section are the elements establishing Lemma 4.6.11. These lemmata are strict generalizations of their analogues in § 3.6.4.

Proposition 4.6.15 (Lower bound $\mathrm{w}(K_1)$)**.** *Fix $C_1, \delta, \varepsilon > 0$ and $0 < \theta < \min\{1-\gamma, \gamma\}$. Given an admissible ensemble, there exists a choice of absolute constants $a_1 > 0$ and $L > 1$, as well as an integer $N_0^{(4.6.15)} \geq N_*$ so that, for each $N \geq N_0^{(4.6.15)}$, it holds with probability at least $1 - \varepsilon$ on the realization of A that*

$$\mathrm{w}(K_1) \geq \left(\frac{a_1^2 + C_1}{2} \right) \sqrt{m}.$$

Proof of Proposition 4.6.15. Since $\mathrm{w}(K_1) = \mathbb{E} \sup_{q \in K_1} \langle q, z \rangle$ is the gmw of

K_1 we may invoke Corollary 2.3.7 to obtain a sufficient chain of inequalities:

$$\mathrm{w}(K_1) \geq \frac{\sqrt{2}}{4}(1-\delta)^2\lambda\sqrt{\log\left(\frac{N\alpha_1^2}{5(1-\delta)^2\lambda^2}\right)} \tag{4.17}$$

$$\overset{(*)}{\geq} \left(\frac{a_1^2+C_1}{2}\right)\sqrt{m}. \tag{4.18}$$

The first inequality, by Corollary 2.3.7, holds with probability at least $1-\varepsilon$ on the realization of A. Therefore, it is enough to show $(*)$ holds. Rewriting $(*)$ gives the equivalent condition:

$$C_{\delta,L}\sqrt{\frac{m}{\log N}}\sqrt{\log\left(\frac{N\log N}{C_{\delta,L,a_1}\sqrt{m}}\right)} \geq C_{a_1,C_1}\sqrt{m}.$$

The latter term of the left-hand side may be simplified using that $m \leq N(\gamma+\theta)$, since $N \geq N_\theta$:

$$\log\left(\frac{N\log N}{C_{\delta,L,a_1}\sqrt{m}}\right) \geq \log\left(C_{\delta,L,a_1,\gamma,\theta}\sqrt{N}\log N\right).$$

In particular, $(*)$ is satisfied if

$$\log\left(C_{\delta,L,a_1,\gamma,\theta}\sqrt{N}\log N\right) \geq C_{\delta,L,a_1,C_1}\log N,$$

which holds when

$$\left(\frac{1}{2} - C_{\delta,L,a_1,C_1}\right)\log(N) \geq -\log\left(C_{\delta,L,a_1,\gamma,\theta}\log N\right).$$

This is eventually true so long as one chooses (a_1, L) abiding

$$C_{\delta,L,a_1,C_1} = 2\left(\frac{a_1^2+C_1}{(1-\delta)^2 L}\right)^2 < \frac{1}{2}.$$

□

Proposition 4.6.16 (Lower bound X_1)**.** *Fix $\delta, \varepsilon_1, \varepsilon_2 > 0$ and $\theta \in (0, \gamma)$. Given an admissible ensemble, there exists a choice of absolute constants*

$a_1 > 0$ and $L > 1$, as well as an integer $N_0^{(4.6.16)} \geq N_*$ so that, for each $N \geq N_0^{(4.6.16)}$, it holds with probability at least $1 - \varepsilon_1$ on the realization of A that with probability at least $1 - \varepsilon_2$ on the realization of z, for any $c \in (0,1)$ there exists $q \in K_1$ satisfying $\langle q, z \rangle \geq c\, \mathrm{w}(K_1)$.

Proof of Proposition 4.6.16. Observe that $K_1 \subseteq \mathbb{R}^m$ is a topological space and define the centered Gaussian process $T_x := \langle x, g \rangle$ for $g_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Observe that $X_1 = \sup_{x \in K_1} |T_x|$ is almost surely finite. So, for any $u > 0$,

$$\mathbb{P}\left(X_1 < \mathrm{w}(K_1) - u\right) \leq \exp\left(-\frac{u^2}{2\sigma_{K_1}^2}\right)$$

by Theorem 2.2.4, where

$$\begin{aligned}\sigma_{K_1}^2 &= \sup_{x \in K_1} \mathbb{E}\, T_x^2 = \sup_{x \in K_1} \sum_{i=1}^N x_i^2\, \mathbb{E}\, |g_i|^2 \\ &= \sup_{x \in K_1} \|x\|_2^2 = \alpha_1^2 = a_1^2 \sqrt{m}.\end{aligned}$$

Now, combine Theorem 2.2.4 and Corollary 2.3.7. For $N \geq N_*$, it holds with probability at least $1 - \varepsilon_1$ on the realization of A that for any $c \in (0,1)$,

$$\begin{aligned}\mathbb{P}\left(X_1 < c\, \mathrm{w}(K_1)\right) &\leq \exp\left(-\frac{(1-c)^2\, \mathrm{w}^2(K_1)}{2\sigma_{K_1}^2}\right) \\ &\leq \exp\left(-C_{a_1,c,\delta,L}\sqrt{m} \cdot \frac{\log\left(C_{a_1,\delta,\gamma,L,\theta}\sqrt{N}\log N\right)}{\log N}\right).\end{aligned}$$

Choose $N_1 \geq N_*$ so that the following chain of inequalities is satisfied:

$$\begin{aligned}\frac{\log\left(C_{a_1,\delta,\gamma,L,\theta}\sqrt{N}\log N\right)}{\log N} &= \frac{1}{2} + \frac{\log\left(C_{a_1,\delta,\gamma,L,\theta}\log N\right)}{\log N} \\ &\geq \frac{1}{4}.\end{aligned}$$

Further, select $N^{(4.6.16)} \geq N_1$ such that all $N \geq N^{(4.6.16)}$ satisfy

$$N \geq \frac{\log^2 \varepsilon_2^{-1}}{C_{a_1,c,\delta,\gamma,L,\theta}}.$$

Then, for all $N \geq N^{(4.6.16)}$, it holds with probability at least $1-\varepsilon_1$ on the realization of A that for any $c \in (0,1)$,

$$\mathbb{P}\left(X_1 < c\,\mathrm{w}(K_1)\right) \leq \exp\left(-C_{a_1,c,\delta,L}\sqrt{m}\right) < \varepsilon_2.$$

Thus, under the specified conditions, $X_1 \geq c\,\mathrm{w}(K_1)$ with probability at least $1-\varepsilon_1$ on the realization of A and probability at least $1-\varepsilon_2$ on the realization of z. In particular, since K_1 is closed, it holds with probability at least $(1-\varepsilon_1)(1-\varepsilon_2)$ that there exists $q \in K_1$ realizing the supremum, thereby admitting existence of a q as claimed. □

Proposition 4.6.17 (Control $K_1 \cap F$)**.** *Fix $C_1, \delta, \varepsilon_1, \varepsilon_2 > 0$ and $\theta \in (0,\gamma)$. Given an admissible ensemble, there is an integer $N_0^{(4.6.17)} \geq N_*$ and an absolute constant $k_1 = k_1(N_0^{(4.6.17)}, C_1, \varepsilon_2) > 0$ so that for all $N \geq N_0^{(4.6.17)}$, with probability at least $1-\varepsilon_1$ on the realization of A, there is an event $\mathcal{E}$ for z satisfying*

$$K_1 \cap F \neq \emptyset \quad \textit{on} \quad \mathcal{E}, \qquad \textit{and} \qquad \mathbb{P}(\mathcal{E}) \geq k_1.$$

Proof of Proposition 4.6.17. Fix $c_1 \in (0,1)$. By Proposition 4.6.16, there is a choice of $a_1 > 0$ and $L > 1$, and an integer $N_0 \geq N_*$ such that, with probability at least $1-\varepsilon_1/2$ on the realization of A, there is an event $\mathcal{E}_1$ for z, with $\mathbb{P}(\mathcal{E}_1) \geq 1-\varepsilon_2$, on which

$$\sup_{q \in K_1} \langle q, z \rangle \geq c_1\,\mathrm{w}(K_1).$$

Further, there exists $q \in K_1$ realizing that supremum because K_1 is closed. Selecting this q, we have $\langle q, z\rangle \geq c_1\,\mathrm{w}(K_1)$. Next, define $C_1' := c_1^{-1}(a_1^2 + C_1) - a_1^2$. By Proposition 4.6.15, increasing L if necessary, there is an integer

$N_1 \geq N_0$ so that with probability at least $1 - \varepsilon_1/2$ on the realization of A,

$$\mathrm{w}(K_1) \geq \left(\frac{a_1^2 + C_1'}{2} \right) \sqrt{m}.$$

In particular, with probability at least $1 - \varepsilon_1$ on the realization of A and probability at least $1 - \varepsilon_2$ on the realization of z, one has simultaneously:

$$\langle q, z \rangle \geq c_1 \, \mathrm{w}(K_1) \geq \left(\frac{a_1^2 + C_1}{2} \right) \sqrt{m}.$$

Define the event $\mathcal{Z}_< := \{\|z\|_2^2 \leq m + C_1\sqrt{m}\}$. Because $q \in K_1$, $\|q\|_2 \leq a_1 m^{-1/4}$, whence conditioning on $\mathcal{E}_1 \cap \mathcal{Z}_<$ gives $q \in F$. Indeed,

$$\begin{aligned} \|q - z\|_2^2 &= \|q\|_2^2 - 2\langle q, z \rangle + \|z\|_2^2 \\ &\leq a_1^2\sqrt{m} - (a_1^2 + C_1)\sqrt{m} + m + C_1\sqrt{m} \\ &= m. \end{aligned}$$

Choose $N_0^{(4.6.17)} := N_1$. Then, for each $N \geq N_0^{(4.6.17)}$, with probability at least $1 - \varepsilon_1$ on the realization of A, there is an event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{Z}_<$ on which $q \in K_1 \cap F$. Next, define

$$k_1 := k_1(N_0^{(4.6.17)}, C_1, \varepsilon_2) := \left[\inf_{N \geq N_0^{(4.6.17)}} \mathbb{P}(\mathcal{Z}_<) \right] - \varepsilon_2$$

and observe that $k_1 > 0$, because $\mathbb{P}(\mathcal{Z}_<)$ is bounded below by a dimension independent constant for $N \geq 2$. Finally, because $\mathcal{E}_1$ holds with probability at least $1 - \varepsilon_2$ on the realization of z one has

$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}(\mathcal{Z}_<) - \varepsilon_2 \geq k_1.$$

□

Proposition 4.6.18 (Upper bound $\mathrm{w}(K_2)$)**.** *Fix $C_2, \delta, \varepsilon > 0$ and $\theta \in (0, \gamma)$. Given an admissible ensemble, there exists a choice of absolute constants $a_1 > 0$ and $L > 1$, an integer $N_0^{(4.6.18)} \geq N_*$, and a maximal choice of*

$\alpha_2 = \alpha_2(N)$ *so that, for each* $N \geq N_0^{(4.6.18)}$*, it holds with probability at least* $1-\varepsilon$ *on the realization of* A *that*

$$\mathrm{w}(K_2) \leq \frac{C_2}{2}\sqrt{m}.$$

Proof of Proposition 4.6.18. First, invoke Corollary 2.3.7 in obtaining a sufficient chain of inequalities on $\mathrm{w}(K_2)$:

$$\mathrm{w}(K_2) \overset{(2.3.7)}{\leq} 4(1+\delta)\lambda\sqrt{\log\left(\frac{8eN\alpha_2^2}{(1+\delta)^2\lambda^2}\right)} \overset{(**)}{\leq} \frac{C_2}{2}\sqrt{m}.$$

The first inequality holds with probability at least $1-\varepsilon$ on the realization of A. Therefore, showing $(**)$ implies the desired result. Rewriting $(**)$ gives the equivalent condition

$$\log\left(C_{\delta,L}\alpha_2^2\frac{N\log N}{m}\right) \leq C_{C_2,\delta,L}\log N.$$

The left-hand side may be simplified using that $m \geq N(\gamma-\theta)$, since $N \geq N_\theta$, yielding a new sufficient condition:

$$\log\left(C_{\delta,\gamma,L,\theta}\alpha_2^2\log N\right) \leq C_{C_2,\delta,L}\log N. \tag{4.19}$$

Thus, (4.19) is valid for any α_2 satisfying $\alpha_2 \leq \alpha_2(N)$, where

$$\begin{aligned}
\alpha_2^2(N) &:= C_{\delta,\gamma,L,\theta}\frac{N^d}{\log N}, \quad \text{where} \\
d &:= \left(\frac{C_2}{8(1+\delta)L}\right)^2, \\
C_{\delta,\gamma,L,\theta} &:= \frac{(1+\delta)^2L^2(\gamma-\theta)}{8e}.
\end{aligned}$$

Finally, set $N_0^{(4.6.18)} := N_*$ and observe that for any $N \geq N_0^{(4.6.18)}$, with $\alpha_2 := \alpha_2(N)$, it holds with probability at least $1-\varepsilon$ on the realization of A that $\mathrm{w}(K_2) \leq \frac{C_2}{2}\sqrt{m}$, as desired. □

Remark 17. It will be convenient to reselect $N_0^{(4.6.18)} \geq N_*$ in Proposi-

tion 4.6.18 so that $\alpha_2(N)$ is increasing for all $N \geq N_0^{(4.6.18)}$. A quick calculation verifies that $N_0^{(4.6.18)} \geq \exp(d^{-1})$ suffices.

Proposition 4.6.19 (Upper bound X_2)**.** *Fix $\delta, \varepsilon_1, \varepsilon_2 > 0$ and $0 < \theta < \min\{1-\gamma, \gamma\}$. Given an admissible ensemble, there exists a choice of absolute constants $a_1 > 0$ and $L > 1$, as well as an integer $N_0^{(4.6.19)} \geq N_*$ so that, for each $N \geq N_0^{(4.6.19)}$, it holds with probability at least $1 - \varepsilon_1$ on the realization of A and with probability at least $1 - \varepsilon_2$ on the realization of z that for any $C > 1$,*

$$\sup_{q \in K_2} \langle q, z \rangle \leq C \, \mathrm{w}(K_2).$$

Proof of Proposition 4.6.19. Define the centered Gaussian process $T_x := \langle x, g \rangle$ for $x \in K_2 \subseteq \mathbb{R}^m$, a topological space, where $g_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Observe that $X_2 = \sup_{x \in K_2} |T_x| < \infty$ almost surely. So, for any $u > 0$,

$$\mathbb{P}\left(X_2 > \mathrm{w}(K_2) + u\right) \leq \exp\left(-\frac{u^2}{2\sigma_{K_2}^2}\right)$$

by Theorem 2.2.4, where

$$\begin{aligned}\sigma_{K_2}^2 &= \sup_{x \in K_2} \mathop{\mathbb{E}}_{g} |\langle x, g \rangle|^2 = \sup_{x \in K_2} \sum_{i=1}^m x_i \mathop{\mathbb{E}}_{i} |g_i|^2 \\ &= \sup_{x \in K_2} \|x\|_2^2 = \alpha_2^2 \leq \alpha_1^2 = a_1^2 \sqrt{m}.\end{aligned}$$

Now, invoke Corollary 2.3.7. For $N \geq N_*$, it holds with probability at least

$1-\varepsilon_1$ on the realization of A that for any $C>1$,

$$\begin{aligned}\mathbb{P}\left(X_2 > C\,\mathrm{w}(K_2)\right) &\le \exp\left(-\frac{(C-1)^2\,\mathrm{w}^2(K_2)}{2\sigma_{K_2}^2}\right)\\ &\le \exp\left(-\frac{(C-1)^2 C_{\delta,L}\sqrt{m}\log(C_{\delta,\gamma,L,\theta}\alpha_2^2\log N)}{2a_1^2\log N}\right)\\ &= \exp\left(-C_{a_1,C,\delta,L}\frac{\sqrt{m}\log(C_{\delta,\gamma,L,\theta}N^d)}{\log N}\right)\\ &= \exp\left(-C_{a_1,C,\delta,L}\sqrt{m}\left(d+\frac{C_{\delta,\gamma,L,\theta}}{\log N}\right)\right)\\ &\le \exp\left(-C_{a_1,C,\delta,L}d\sqrt{m}\right) < \varepsilon_2.\end{aligned}$$

The latter line follows by taking $N_1 \ge N_*$ sufficiently large so that for all $N \ge N_1$,

$$\frac{C_{\delta,\gamma,L,\theta}}{\log N} > -\frac{d}{2}, \qquad \text{and} \qquad N \ge \frac{\log^2 \varepsilon_2^{-1}}{C_{a_1,C,\delta,L}d^2(\gamma-\theta)}.$$

Finally, set $N_0^{(4.6.19)} := N_1$. Then, for all $N \ge N_0^{(4.6.19)}$, it holds with probability at least $1-\varepsilon_1$ on the realization of A that, for any $C>1$, $X_2 \le C\,\mathrm{w}(K_2)$ with probability at least $1-\varepsilon_2$ on the realization of z, as desired. □

Proposition 4.6.20 (Control $\mathrm{w}(K_2)\cap F$)**.** *Fix $C_2,\delta,\epsilon_1,\varepsilon_2>0$ and $\theta\in(0,\gamma)$. Given an admissible ensemble, there is an integer $N_0^{(4.6.20)} \ge N_*$ and an absolute constant $k_2 = k_2(N_0^{(4.6.20)}, C_2, \varepsilon_2) > 0$ so that for all $N \ge N_0^{(4.6.20)}$, with probability at least $1-\varepsilon_1$ on the realization of A, there is an event $\mathcal{E}$ for z satisfying*

$$K_2 \cap F = \emptyset \quad \text{on} \quad \mathcal{E} \qquad \text{and} \qquad \mathbb{P}(\mathcal{E}) \ge k_2.$$

Proof of Proposition 4.6.20. Fix $c_2 > 1$. By Proposition 4.6.19, there is a choice of $a_1 > 0$ and $L > 1$, and an integer $N_0 \ge N_*$ such that, with probability at least $1-\varepsilon_1/2$ on the realization of A, there is an event $\mathcal{E}_2$ for

z, with $\mathbb{P}(\mathcal{E}_2) \geq 1 - \varepsilon_2$, on which

$$\sup_{q \in K_2} \langle q, z \rangle \leq c_2 \, \mathrm{w}(K_2).$$

Now, select $C_2' > 0$ so that $0 < c_2 C_2' < C_2$. By Proposition 4.6.18, increasing L if necessary, there is an integer $N_1 \geq N_0$ so that with probability at least $1 - \varepsilon_1/2$ on the realization of A,

$$\mathrm{w}(K_2) \leq \frac{C_2'}{2} \sqrt{m}.$$

In particular, with probability at least $1 - \varepsilon_1$ on the realization of A and probability at least $1 - \varepsilon_2$ on the realization of z, one has simultaneously:

$$\sup_{q \in K_2} \langle q, z \rangle \leq c_2 \, \mathrm{w}(K_2) < c_2 \frac{C_2'}{2} \sqrt{m} < \frac{C_2}{2} \sqrt{m}.$$

Define the event $\mathcal{Z}_> := \{\|z\|_2^2 \geq m + C_2 \sqrt{m}\}$. With probability at least $1 - \varepsilon_1$ on the realization of A, conditioning on $\mathcal{E}_2 \cap \mathcal{Z}_>$ gives, for any $q \in K_2$,

$$\begin{aligned} \|q - z\|_2^2 &= \|q\|_2^2 - 2\langle q, z \rangle + \|z\|_2^2 \\ &> \|q\|_2^2 - C_2 \sqrt{m} + m + C_2 \sqrt{m} \\ &\geq m. \end{aligned}$$

In particular, $\|q - z\|_2^2 > m$ for all $q \in K_2$, and so $K_2 \cap F = \emptyset$. Choose $N_0^{(4.6.20)} := N_1$. Then, for each $N \geq N_0^{(4.6.20)}$, with probability at least $1 - \varepsilon_1$ on the realization of A, there is an event $\mathcal{E} := \mathcal{E}_2 \cap \mathcal{Z}_>$ on which $K_2 \cap F = \emptyset$. Next, define

$$k_2 := k_2(N_0^{(4.6.20)}, C_2, \varepsilon_2) := \left[\inf_{N \geq N_0^{(4.6.20)}} \mathbb{P}(\mathcal{Z}_>) \right] - \varepsilon_2$$

and observe that $k_2 > 0$ because $\mathbb{P}(\mathcal{Z}_>)$ is bounded below by a dimension independent constant for $N \geq 2$. Finally, because $\mathcal{E}_2$ holds with probability

at least $1 - \varepsilon_2$ on the realization of z, one has

$$\mathbb{P}(\mathcal{E}) \geq \mathcal{P}(\mathcal{Z}_>) - \varepsilon_2 \geq k_2.$$

□

Remark 18 (Dimension-independent bound for k_1, k_2). In Proposition 4.6.17 and Proposition 4.6.20 above, it is not necessary to have k_1 depend on $N_0^{(4.6.17)}$ or k_2 on $N_0^{(4.6.20)}$. For example, one could bound $\mathbb{P}(\mathcal{Z}_<)$ for all $N \geq 2$. However, the dependence of k_1 on $N_0^{(4.6.17)}$ is simply to note that the lower bound on $\mathbb{P}(\mathcal{Z}_<)$ is improved by considering $N \geq N_0^{(4.6.17)}$ as opposed to merely $N \geq 2$. Analogously so for k_2.

Suboptimal regime for basis pursuit

This section contains the proof for Theorem 4.4.2, the main result of § 4.4 establishing a regime in which (BP_σ) is minimax suboptimal. In essence, it combines Lemma 4.4.1 and Lemma 4.6.13.

Proof of Theorem 4.4.2. By a scaling argument, it suffices to consider the case $\eta = 1$. Re-write the minimax expression (4.1) as

$$\inf_{\sigma>0} \sup_{x \in \Sigma_s^N} \tilde{R}(\sigma; x, A, 1) = \min \left\{ \inf_{\sigma \leq \sqrt{m}} S(\sigma), \inf_{\sigma > \sqrt{m}} S(\sigma) \right\},$$
$$S(\sigma) := \sup_{x \in \Sigma_s^N} \tilde{R}(\sigma; x, A, 1).$$

For any $N \geq N_*$, by Lemma 4.4.1, it holds with probability at least $1 - \varepsilon/2$ on the realization of A that

$$\inf_{\sigma > \sqrt{m}} S(\sigma) \geq C_{\delta,\gamma,\theta} \sqrt{N}.$$

Next observe that the trivial lower bound $S(\sigma) \geq \tilde{R}(\sigma; 0, A, 1)$ holds for any $\sigma > 0$, because $0 \in \Sigma_s^N$. In particular, Lemma 4.6.13 yields an $N_0 \geq N_*$ and absolute constant $p > 0$ such that, with probability at least $1 - \varepsilon/2$ on the

realization of A,

$$\inf_{\sigma \le \sqrt{m}} S(\sigma) \ge \inf_{\sigma \le \sqrt{m}} \tilde{R}(\sigma; 0, A, 1) \ge C_{\delta,\gamma,K,\theta} N^p.$$

Consequently, there is an absolute constant $p > 0$ so that, for all $N \ge N_0$, it holds with probability at least $1 - \varepsilon$ on the realization of A that

$$\begin{aligned} \inf_{\sigma > 0} \sup_{x \in \Sigma_s^N} \tilde{R}(\sigma; x, A, 1) &\ge \min \left\{ C_{\delta,\gamma,\theta} \sqrt{N}, C_{\delta,\gamma,K,\theta} N^p \right\} \\ &\ge C_{\delta,\gamma,K,\theta} N^p. \end{aligned}$$

□

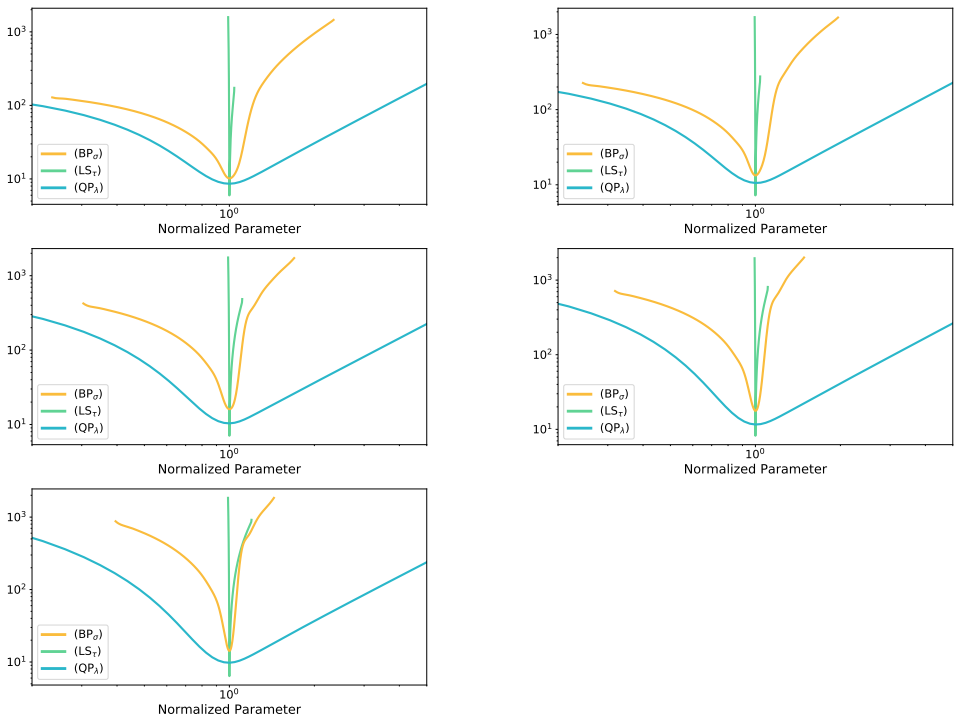


Figure 4.4: Each plot depicts the average loss as a function of the normalized parameter for each of the three programs under consideration. The collection of plots depicts how the average loss changes as a function of N and $\delta = m/N$. Details for each plot will be given by referencing the (row, column) position of the plot in this figure. The domain of the normalized parameter in each plot is $(0.2, 5)$. A single realization of A was fixed and the average loss was computed from $k = 50$ realizations of the noise by constructing a function approximator using radial basis function approximation with a multiquadric kernel. The RBF approximator was evaluated on a logarithmically spaced grid of $n = 301$ points centered about 1. **(1,1):** $(s, N, \delta, \eta) = (1, 4000, 0.1, 1)$; **(1,2):** $(s, N, \delta, \eta) = (1, 7000, 0.1, 1)$; **(2,1):** $(s, N, \delta, \eta) = (1, 4000, 0.25, 1)$; **(2,2):** $(s, N, \delta, \eta) = (1, 7000, 0.25, 1)$; **(3,1):** $(s, N, \delta, \eta) = (1, 4000, 0.45, 1)$.

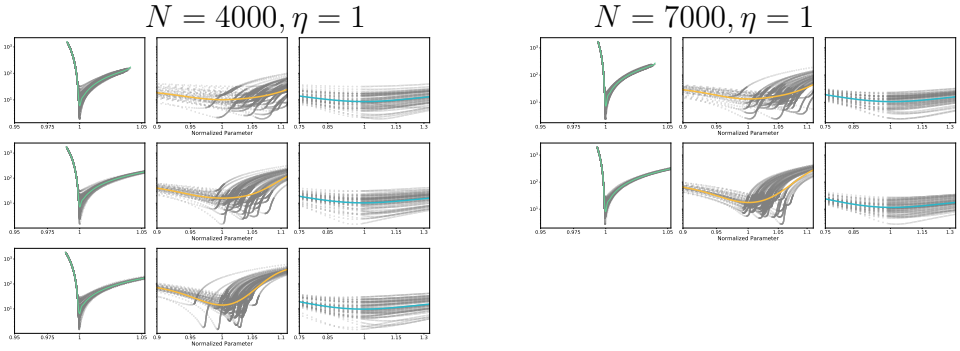


Figure 4.5: Each plot depicts the quality of the RBF approximation about the optimal normalized parameter. The left-most plot is in every case depicting the loss and (approximate) average loss of (LS_τ); the middle that for (BP_σ); and the right that for (QP_λ). Top-to-bottom:$(s, N, \delta, \eta) = (1, 4000, 0.1, 1)$; $(s, N, \delta, \eta) = (1, 4000, 0.25, 1)$; $(s, N, \delta, \eta) = (1, 4000, 0.45, 1)$; $(s, N, \delta, \eta) = (1, 7000, 0.1, 1)$; $(s, N, \delta, \eta) = (1, 7000, 0.25, 1)$.

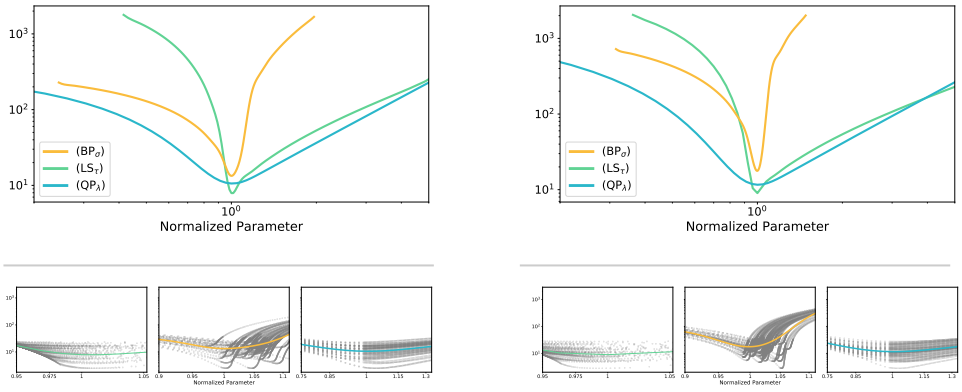


Figure 4.6: Top row: Each plot depicts the average loss as a function of the normalized parameter for each of the three programs under consideration. The collection of plots depicts how the average loss changes as a function of $\delta = m/N$. The domain of the normalized parameter in each plot is $(0.2, 5)$. A single realization of A was fixed and the average loss was computed from $k = 50$ realizations of the noise by constructing a function approximator using radial basis function approximation with a multiquadric kernel. The RBF approximator was evaluated on a logarithmically spaced grid of $n_{\text{rbf}} = 501$ points centered about 1. **Bottom row:** Each plot depicts the quality of the RBF approximation about the optimal normalized parameter. In each triptych, the left plot depicts the loss and (approximate) average loss of (LS_τ); the middle that for (BP_σ); and the right that for (QP_λ). **Left column:** $(s, N, \delta, \eta) = (1, 7000, 0.1, 100)$; **Right column:** $(s, N, \delta, \eta) = (1, 7000, 0.25, 100)$.

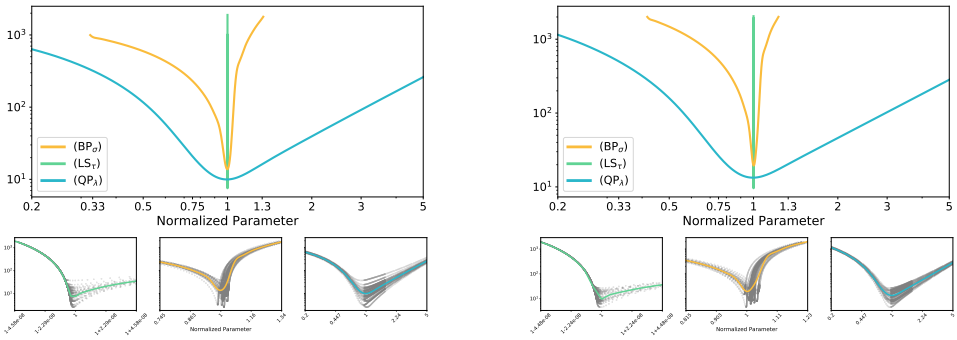


Figure 4.7: PS numerics in the low-noise, high-sparsity regime. **Top row:** Average loss is plotted with respect to the normalized parameter for each program. **Bottom row:** Visualizations of RBF approximation quality for average loss (best seen on a computer). **Left:** $(s, N, m, \eta, k, n) = (1, 10^4, 2500, 10^{-5}, 25, 201)$; **Right:** $(s, N, m, \eta, k, n) = (1, 10^4, 4500, 10^{-5}, 25, 201)$.

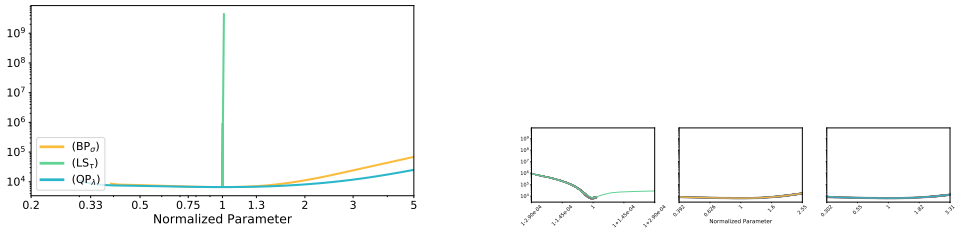


Figure 4.8: PS numerics in the low-sparsity regime with parameters $(s, N, m, \eta, k, n) = (750, 10^4, 4500, 10^{-1}, 25, 201)$. **Left:** Average loss is plotted with respect to the normalized parameter for each program. **Right:** Visualizations of the RBF approximation quality for average loss.

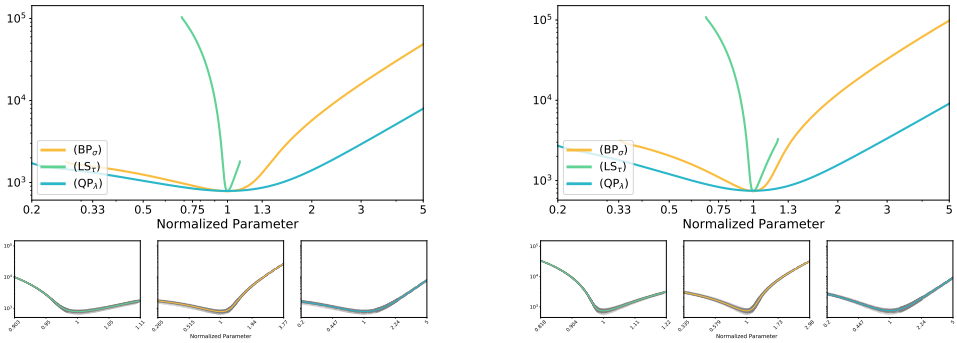


Figure 4.9: PS numerics for intermediate parameter values: $(s, N, \eta, k, n) = (10^2, 10^4, 10^{-1}, 25, 201)$. **Left:** $m = 2500$; **Right:** $m = 4500$. **Top:** Average loss is plotted with respect to the normalized parameter for each program. **Bottom:** Visualizations of average loss approximation quality.

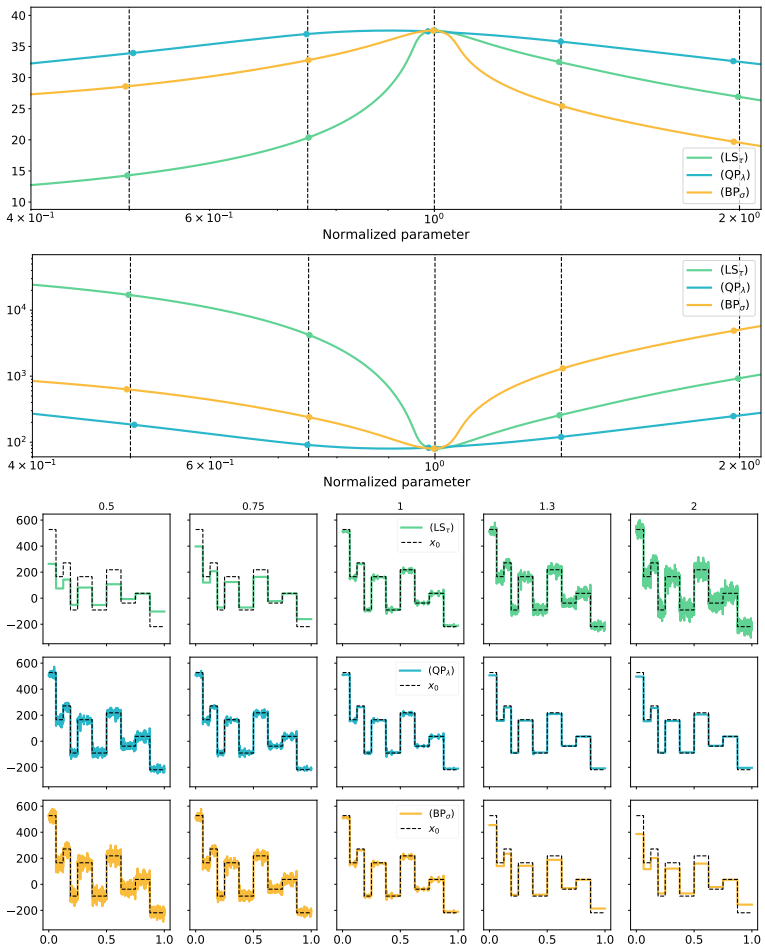


Figure 4.10: Realistic example in 1D for $(s, N, m, \eta, k, n) = (10, 4096, 1843, 50, 50, 501)$. Ground truth signal x_0 defined in the Haar wavelet domain with first s coefficients equal to N. Noise added in the Haar wavelet domain; recovery error measured in the signal domain. **Top:** Average psnr as a function of the normalized parameter for each parameter. **Middle:** Average nnse as a function of the normalized parameter for each parameter. **Bottom:** The ground truth and recovered signal for a single realization of the noise, faceted by the approximate normalized parameter value (given in the title) and by program (as depicted in the legend).

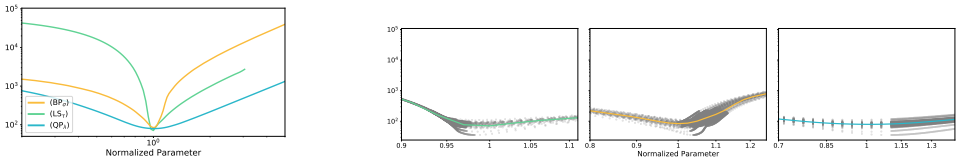


Figure 4.11: PS numerics for 1D wavelet CS example with parameter settings $(s, N, m, \eta, kn) = (10, 4096, 1843, 50, 50, 501)$. **Left:** Average loss for each program plotted with respect to the normalized parameter. **Right:** A visualization of approximation quality for the average loss: (LS_τ), (BP_σ) and (QP_λ), from left to right.

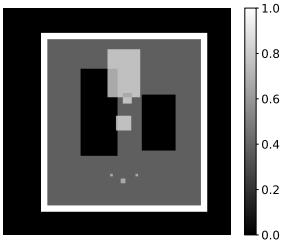


Figure 4.12: The square Shepp-Logan phantom (sslp).

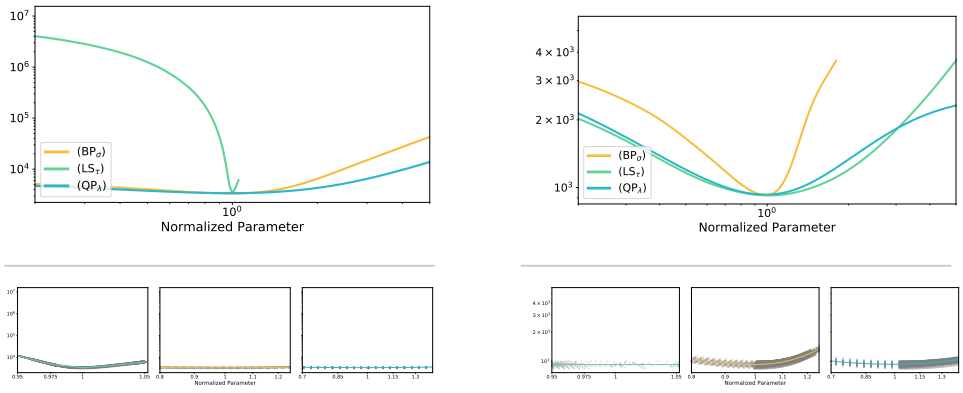


Figure 4.13: Average loss for a 2D wavelet compressed sensing problem, plotted as a function of ρ; $(s, N, m) = (416, 6418, 2888)$ with $(k, n) = (50, 501)$. **Left:** $\eta = 10^{-2}$. **Right:** $\eta = 1/2$. **Top:** The average loss (i.e., nnse) for each program as a function of ρ. The average loss was approximated using RBF interpolation with parameters given in Table A.5. **Bottom:** Plots to evaluate the quality of the RBF interpolation. In each polot, individual realizations of the loss are visible as grey points; the approximation to the average loss is visible as the coloured line through those points.

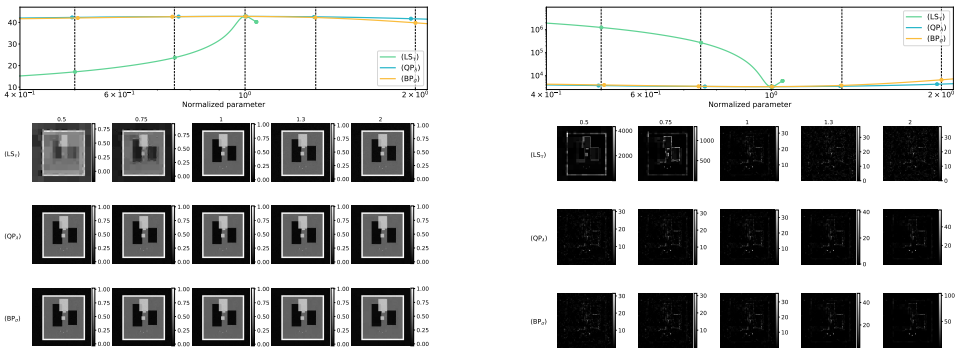


Figure 4.14: A 2D wavelet compressed sensing problem using the square Shepp-Logan phantom; $(s, N, m, \eta) = (416, 6418, 2888, 10^{-2})$ with $(k, n) = (50, 501)$. **Top row:** psnr (left) and nnse (right), plotted as a function of ρ. The plotted curves were generated from the single realization of the measurements that correspond to the grids depicted below them. **Bottom grids:** The left grid of 15 images shows the recovered image for each of five values of ρ: $\rho \in \{\frac{1}{2}, \frac{3}{4}, 1, \frac{4}{3}, 2\}$; and for each program: (LS_τ), (QP_λ), (BP_σ). The right grid of 15 images shows the pixel-wise nnse of the recovered image for the same values of ρ, and for the three programs. Colour bars provide scale, and are best observed on a computer. The stated values of ρ are approximate; the values of ρ for which the images are depicted are marked by points in the nnse and psnr plots of the same colour as the loss curve on top of which they're plotted.

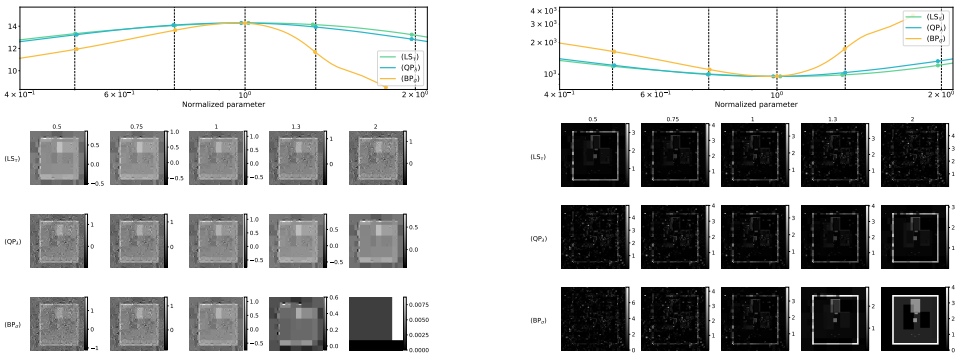


Figure 4.15: A 2D wavelet compressed sensing problem using the square Shepp-Logan phantom; $(s, N, m, \eta) = (416, 6418, 2888, 1/2)$ with $(k, n) = (50, 501)$. **Top row:** psnr (left) and nnse (right), plotted as a function of ρ for each program. The plotted curves were generated from the single realization of the measurements that correspond to the grids depicted below them. **Bottom grids:** The left grid of 15 images shows the recovered image for each of five values of ρ: $\rho \in \{\frac{1}{2}, \frac{3}{4}, 1, \frac{4}{3}, 2\}$; and for each program: (LS_τ), (QP_λ), (BP_σ). The right grid of 15 images shows the pixel-wise nnse of the recovered image for the same values of the normalized parameter, and for the three programs. Colour bars provide scale, and are best observed on a computer. The stated values of ρ are approximate; the values of ρ for which the images are depicted are marked by points in the nnse and psnr plots of the same colour as the loss curve on top of which they're plotted.

Chapter 5

A well-ordering property for proximal operators

This chapter is devoted to the development of a theoretical property satisfied by the proximal mapping of certain functions. We begin by supplying the relevant background material and notation, including the definition of a proximal operator. Then, we include a more detailed discussion of the projection lemma, introduced in § 2.1.1 (*cf.* Lemma 2.1.2). A re-statement of Lemma 2.1.2 was first proved in Oymak and Hassibi [55, Lemma 15.3]. However, Berk et al. [9] provided an alternative proof that allows the lemma to be generalized to a larger class of operators.

5.1 Additional background

5.1.1 Proof technique

We begin the additional background with a proof of Lemma 2.1.2, as originally given in Berk et al. [9]. This proof will provide a framework for generalizing the projection lemma. An element of the proof is the following well-known projection theorem [11, Proposition 2.2.1].

Proposition 5.1.1 (Projection Theorem)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be nonempty closed and convex.*

1. *For every $x \in \mathbb{R}^n$, there is a unique vector that minimizes $\|w-x\|$ over all $w \in \mathcal{C}$, which we refer to as* $\operatorname{Proj}_{\mathcal{C}}(x)$.

2. *For every $x \in \mathbb{R}^n$, a vector $z \in \mathcal{C}$ is equal to* $\operatorname{Proj}_{\mathcal{C}}(x)$ *if and only if*

$$\langle w - z, x - z \rangle \leq 0, \qquad \forall w \in \mathcal{C}.$$

   If $\mathcal{C}$ is affine, this is equivalent to $x - z \in S^{\perp}$ where S is the subspace parallel to $\mathcal{C}$.

3. *The function* $\operatorname{Proj}_{\mathcal{C}} : \mathbb{R}^n \to \mathcal{C}$ *is continuous and nonexpansive:*

$$\|\operatorname{Proj}_{\mathcal{C}}(y) - \operatorname{Proj}_{\mathcal{C}}(x)\| \leq \|y - x\|, \qquad \forall x, y \in \mathbb{R}^n.$$

4. *The distance function* $\mathrm{d} : \mathbb{R}^n \to \mathbb{R}$, *defined by*

$$\mathrm{d}(x, \mathcal{C}) := \min_{z \in \mathcal{C}} \|z - x\|_2$$

   is convex.

We may now present the proof.

Proof of Lemma 2.1.2. Define $z_\alpha := \operatorname{Proj}_{\alpha K}(z)$ for $\alpha = 1, \lambda$ and define $f(t) := \|u_t\|_2^2$, where $u_t := tz_\lambda + (1-t)z_1$ for $t \in [0,1]$. Our goal is to show $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} f(t) \geq 0$; this implies $\|z_\lambda\|_2 \geq \|z_1\|_2$, because f is convex. Expanding $f(t)$ gives

$$f(t) = t^2\big(\|z_\lambda\|_2^2 - 2\langle z_1, z_\lambda \rangle + \|z_1\|_2^2\big) + 2t\big(\langle z_1, z_\lambda \rangle - \|z_1\|_2^2\big) + \|z_1\|_2^2.$$

Thus, it is required to verify the condition $(\star)$:

$$\frac{\mathrm{d}}{\mathrm{d}t} f(t)\bigg|_{t=0} = \big[2t\|z_\lambda - z_1\|_2^2 + 2\langle z_1, z_\lambda - z_1 \rangle\big|_{t=0} = 2\langle z_1, z_\lambda - z_1 \rangle \overset{(\star)}{\geq} 0$$

The projection condition says that if $\operatorname{Proj}_C(x)$ is the projection of x onto a convex set C then for any $y \in C$, $\langle y - \operatorname{Proj}_C(x), x - \operatorname{Proj}_C(x) \rangle \leq 0$. From the projection condition [11], we have

- $\langle \lambda^{-1} z_\lambda - z_1, z - z_1 \rangle \leq 0$
- $\langle \lambda z_1 - z_\lambda, z - z_\lambda \rangle \leq 0$.

Accordingly,

$$\begin{aligned} 0 &\geq \langle z_\lambda - \lambda z_1, z - z_1 \rangle + \langle \lambda z_1 - z_\lambda, z - z_\lambda \rangle \\ &= \langle \lambda z_1 - z_\lambda, z_1 - z_\lambda \rangle = \langle (\lambda - 1) z_1, z_1 - z_\lambda \rangle + \|z_1 - z_\lambda\|_2^2 \\ &\geq (\lambda - 1) \langle z_1, z_1 - z_\lambda \rangle \end{aligned}$$

which is equivalent to $\langle z_1, z_\lambda - z_1 \rangle \geq 0$. Therefore, f is a convex function increasing on the interval $t \in [0, 1]$, whence $\|z_1\|_2 \leq \|z_\lambda\|_2$ as desired. □

Lemma 2.1.2 has immediate consequences for the ability of PD algorithms to recover the 0 vector from corrupted measurements, as revealed in Chapter 3. More generally, this lemma has consequences on the sensitivity of recovery error for both PD and CS Lasso programs. Its role is seen in § 3.6 and § 4.6. Note that the set $\mathcal{K}$ need be neither symmetric nor origin-centered. However, it must be convex and contain the origin, in general; we have included a pictorial counterexample in Figure 5.1b to depict why.

Our proof of Lemma 2.1.2 examines the derivative of the function $f(t) := \|u_t\|_2^2$, where $u_t := t \operatorname{Proj}_{\lambda K}(z) + (1 - t) \operatorname{Proj}_K(z)$. From the latter lines of the proof, observe that a growth condition on the derivative at $t = 0$ can be specified:

$$\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} f(t) \bigg|_{t=0} = \langle z_1, z_\lambda - z_1 \rangle \geq \frac{\|z_\lambda - z_1\|_2^2}{\lambda - 1}. \tag{5.1}$$

Remark 19. There is a simpler way to begin the proof of the projection lemma. To show

$$\|z_1\|_2 \leq \|z_\lambda\|_2 \quad \iff \quad \|z_1\|_2^2 \leq \|z_1\|_2 \|z_\lambda\|_2,$$

one may instead prove the following chain,

$$\|z_1\|_2^2 \leq \langle z_1, z_\lambda \rangle \leq \|z_1\|_2 \|z_\lambda\|_2.$$

The latter inequality is true by Cauchy-Schwarz, so it remains only to prove the former:

$$\langle z_1, z_\lambda \rangle - \|z_1\|_2^2 \geq 0 \quad \iff \quad \langle z_1, z_\lambda - z_1 \rangle \geq 0.$$

Rearranging shows this inequality is equivalent to $(\star)$, and the remainder of the proof proceeds as is. This remark is included for intuition, but this approach is less generalizable. For example, it does not yield the aforementioned rate of growth on the derivative.

5.1.2 Proximal operators and projections

Let $\mathcal{H}$ denote a Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle$. The associated norm is given as $\|x\| := \sqrt{\langle x, x \rangle}$. Observe that the previous notion of projection still applies in the Hilbert space setting. Namely, for a closed,

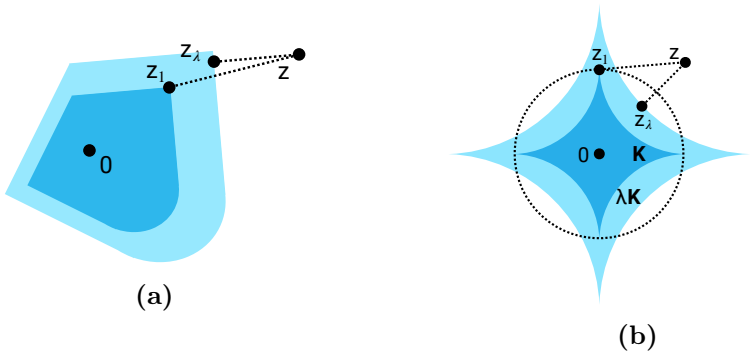


Figure 5.1: (a) A visualization of the lemma. Projecting z onto the outer and inner sets gives z_λ and z_1, respectively; evidently, $\|z_1\|_2 \leq \|z_\lambda\|_2$. (b) A counterexample using scaled ℓ_p balls for some $0 < p < 1$, suggesting why K must be convex in general. Here, z is projected inwards onto λK, but towards a distal vertex when projected onto K. This figure has been reproduced from Berk et al. [9].

convex and nonempty set $\mathcal{K} \subseteq \mathcal{H}$ and point $x \in \mathcal{H}$ one has

$$\mathrm{Proj}_{\mathcal{K}}(x) = \arg\min_{w \in \mathcal{H}} \|x - w\|.$$

We define a proper function in the fashion of Bertsekas et al. [11, §1.2].

Definition 8 (Proper)**.** An extended real-valued convex function $f : \mathcal{H} \to (-\infty, \infty]$ is *proper* if there exists $x \in \mathcal{H}$ such that $f(x) < \infty$.

Next, we recall the definition of lower semi-continuous.

Definition 9 (lower semi-continuous (lsc))**.** Say that a function f is lower semi-continuous at a point x if for every net $(x_a)_{a \in A}$, $x_a \to x$ implies that $f(x) \leq \liminf f(x_a)$. If f is lower semi-continuous for all x in its domain then we say that f is lower semi-continuous (lsc).

Denote the set of proper lsc convex functions on $\mathcal{H}$ by $\Gamma_0(\mathcal{H})$.

Definition 10 (Proximal mapping)**.** Let $f \in \Gamma_0(\mathcal{H})$. For a parameter $\lambda > 0$ define the envelope function $\mathrm{env}_\lambda f$ and the proximal mapping $\mathrm{prox}_\lambda f$ by

$$\mathrm{env}_\lambda f(x) := \inf_{w \in \mathcal{H}} \left\{ f(w) + \frac{1}{2\lambda}\|w - x\|^2 \right\},$$
$$\mathrm{prox}_\lambda f(x) := \arg\min_{w \in \mathcal{H}} \left\{ f(w) + \frac{1}{2\lambda}\|w - x\|^2 \right\}.$$

If $\lambda = 1$ we write simply $\mathrm{prox}\, f := \mathrm{prox}_1 f$. We refer the reader to Bertsekas et al. [11, Ex. 2.13–14] for common properties of env and prox; and to Chapter 1 and § 1.1 for a discussion about motivations for and applications of proximal operators. Observe that $\mathrm{Proj}_{\mathcal{K}}$ is the special case of prox in which $f := \chi_{\mathcal{K}}$ is the indicator function of the set $\mathcal{K}$. Specifically, the indicator function $\chi_{\mathcal{K}} : \mathcal{H} \to \{0, \infty\}$ of the set K is defined as $\chi_{\mathcal{K}}(x) = 0$ if $x \in \mathcal{K}$ and $\chi_{\mathcal{K}}(x) = \infty$ otherwise.

We next state a well-known result [65, Theorem 2.26] that describes important behaviour of the prox and env operators when $f \in \Gamma_0(\mathcal{H})$.

Theorem 5.1.2. *Suppose $f \in \Gamma_0(\mathcal{H})$. Then, for every $\lambda > 0$.*

1. *The proximal mapping* $\operatorname{prox}_\lambda f$ *is single-valued and continuous. In fact,* $\operatorname{prox}_\lambda f \to \operatorname{prox}_{\bar\lambda} f(\bar x)$ *whenever* $(\lambda, x) \to (\bar\lambda, \bar x)$ *with* $\bar\lambda > 0$.

2. *The envelope* $\operatorname{env}_\lambda f$ *is convex and continuously differentiable, with gradient*

$$\nabla \operatorname{env}_\lambda f(x) = \frac{1}{\lambda}\left(x - \operatorname{prox}_\lambda f(x)\right).$$

Proximal operators and envelopes admit the following scaling property.

Proposition 5.1.3. *Let* $\phi : \mathcal{H} \to \mathbb{R}$ *and* $\tau > 0$. *Define* $f_\tau(x) := \phi(x/\tau)$. *Then*

$$\operatorname{prox} f_\tau(x) = \tau \operatorname{prox}\left(\frac{\phi}{\tau^2}\right)\left(\frac{x}{\tau}\right),$$
$$\operatorname{env} f_\tau(x) = \tau^2 \operatorname{env}\left(\frac{\phi}{\tau^2}\right)\left(\frac{x}{\tau}\right).$$

Proof of Proposition 5.1.3. Establishing the first identity is a direct computation:

$$\begin{aligned}
\operatorname{prox} f_\tau(x) &= \operatorname*{arg\,min}_{w\in\mathcal{H}} \left\{\phi(w/\tau) + \frac{1}{2}\|x - w\|_2^2\right\} \\
&= \tau \operatorname*{arg\,min}_{w'\in\mathcal{H}} \left\{\phi(w') + \frac{1}{2}\|x - \tau w'\|_2^2\right\} \\
&= \tau \operatorname*{arg\,min}_{w'\in\mathcal{H}} \left\{\frac{\phi(w')}{\tau^2} + \frac{1}{2}\|\frac{x}{\tau} - w'\|_2^2\right\} \\
&= \tau \operatorname{prox}\left(\frac{\phi}{\tau^2}\right)\left(\frac{x}{\tau}\right).
\end{aligned}$$

Similarly for the envelope of f_τ,

$$\begin{aligned}
\operatorname{env} f_\tau(x) &= \min_{w\in\mathcal{H}} \left\{\phi(w/\tau) + \frac{1}{2}\|x - w\|_2^2\right\} \\
&= \min_{w'\in\mathcal{H}} \left\{\phi(w') + \frac{\tau^2}{2}\|\frac{x}{\tau} - w'\|_2^2\right\} \\
&= \tau^2 \operatorname{env}\left(\frac{\phi}{\tau^2}\right)\left(\frac{x}{\tau}\right)
\end{aligned}$$

□

Proximal operators also admit nice subgradients in the sense of the following proposition [4, Proposition 12.26].

Proposition 5.1.4 (prox-subgradient connection)**.** *Let* $f \in \Gamma_0(\mathcal{H})$ *and let* $x, p \in \mathcal{H}$*. Then* $p = \operatorname{prox} f(x)$ *iff*

$$\forall y \in \mathcal{H}, \quad f(y) \geq f(p) + \langle y - p, x - p \rangle. \tag{5.2}$$

5.2 A notion of ordering for proximal operators

5.2.1 Derivation and notation

The purpose of this chapter is to establish an analogue of Lemma 2.1.2 for proximal operators. To this end, we use Proposition 5.1.4 above and the scaffold established in the proof of Lemma 2.1.2. First, we derive an expression that suggests what ordering proximal operators should satisfy.

As noted, projections onto a set $\mathcal{K}$ satisfy $\operatorname{Proj}_{\mathcal{K}}(x) = \operatorname{prox} \chi_{\mathcal{K}}(x)$. Projections onto scalings of $\mathcal{K}$ satisfy, in turn,

$$\chi_{\tau\mathcal{K}}(x) = 0 \iff x \in \tau\mathcal{K} \iff \tau^{-1}x \in \mathcal{K} \iff \chi_{\mathcal{K}}(\tau^{-1}x) = 0.$$

Define the map $\mu_\lambda(x) := \lambda x$ and observe that the latter equivalence may be re-written as $(\chi_{\mathcal{K}} \circ \mu_{\tau^{-1}})(x) = 0$. Thus, $\operatorname{Proj}_{\tau\mathcal{K}}(x) = \operatorname{prox}(\chi_{\mathcal{K}} \circ \mu_{\tau^{-1}})(x)$. In this special case, Lemma 2.1.2 establishes that if $\tau > 1$, then

$$\|\operatorname{prox}(\chi_{\mathcal{K}} \circ \mu_{\tau^{-1}})(x)\| \geq \|\operatorname{prox} \chi_{\mathcal{K}}(x)\|.$$

For the remainder of this chapter, we assume $\phi : \mathcal{H} \to [0, \infty)$ to be a gauge and define $\phi_\tau(x) := (\phi \circ \mu_{\tau^{-1}})(x) = \phi(x/\tau)$. Further, denote $p_\tau := \operatorname{prox} \phi_\tau(x)$.

Proposition 5.2.1 (Gauge-ordering of solutions)**.** *Let* $\phi : \mathcal{H} \to (-\infty, \infty]$ *be*

a gauge and define $p_\tau := \operatorname{prox} \phi_\tau(x)$. *Then,*

$$\phi(p_\tau) \geq \phi(p_1).$$

Proof of Proposition 5.2.1. This result is a straightforward application of Proposition 5.1.4. Indeed, one has

$$\begin{aligned} \phi(p_\tau) - \phi(p_1) &\geq \langle p_\tau - p_1, x - p_1 \rangle; \\ \phi_\tau(p_1) - \phi_\tau(p_\tau) &\geq \langle p_1 - p_\tau, x - p_\tau \rangle. \end{aligned}$$

Adding the two inequalities produces

$$\begin{aligned} &\langle p_\tau - p_1, x - p_1 \rangle + \langle p_1 - p_\tau, x - p_\tau \rangle \\ &= \|p_\tau - p_1\|^2 \leq \phi_1(p_\tau) - \phi_1(p_1) + \phi_\tau(p_1) - \phi_\tau(p_\tau) \\ &= \phi(p_\tau) - \phi(p_1) + \phi\left(\frac{p_1}{\tau}\right) - \phi\left(\frac{p_\tau}{\tau}\right). \end{aligned}$$

Since ϕ is a gauge, it is positively homogeneous, whence

$$0 \leq \frac{\|p_\tau - p_1\|^2}{\tau - 1} \leq \phi_\tau(p_\tau) - \phi_\tau(p_1).$$

In particular, one has $\phi(p_\tau) \geq \phi(p_1)$ with equality if and only if $p_\tau = p_1$. □

5.2.2 Main result

We now have the tools to state and prove the first main result of this chapter.

Theorem 5.2.2 (Proximal lemma)**.** *Let* $\mathcal{H}$ *be a Hilbert space with norm* $\|\cdot\|$ *and inner product* $\langle \cdot, \cdot \rangle$. *Suppose* $x \in \mathcal{H}$ *and that* $\phi : \mathcal{H} \to [0, \infty)$ *is a gauge. For* $\tau \in [0, \infty)$, *define* $\phi_\tau(x) := \phi(x/\tau)$ *and let*

$$p_\tau := \operatorname{prox} \phi_\tau(x) := \operatorname*{arg\,min}_{w \in \mathcal{H}} \phi_\tau(w) + \frac{1}{2}\|x - w\|^2.$$

If $\tau > 1$ *then* $\|p_\tau\| \geq \|p_1\|$.

Proof of Theorem 5.2.2. The proof proceeds in the same fashion as that for

Lemma 2.1.2. Define $u_\alpha := (1-\alpha)p_1 + \alpha p_\tau$ and $F(\alpha) := \|u_\alpha\|^2$. We see that F is smooth and convex on $\alpha \in [0,1]$, whence it suffices to establish that $F(\alpha)$ is non-decreasing on $\alpha \in [0,1]$. By convexity of F, it suffices to establish that $F'(0) \geq 0$. A calculation verifies that

$$\frac{1}{2}F'(0) = \langle p_1, p_\tau \rangle - \|p_1\|^2 = \langle p_1, p_\tau - p_1 \rangle.$$

Hence, it remains to establish that $\langle p_1, p_\tau - p_1 \rangle \geq 0$. Before proceeding, we note that if $\tau = 1$ there is nothing to show. Similarly, if $\tau > 1$, but $p_1 = p_\tau$, then there is still nothing to show. Thus, we consider only the more interesting case where $\tau > 1$ and $p_1 \neq p_\tau$.

First, assume that $\phi(p_1) > 0$. Recall from Proposition 5.2.1 that $\phi(p_\tau) - \phi(p_1) \geq 0$ achieves equality iff $p_1 = p_\tau$. Since we assume $p_1 \neq p_\tau$, it must be that $\phi(p_\tau) > \phi(p_1)$. Select the value $\lambda > 1$ such that $\phi_\tau(\lambda p_1) = \phi_\tau(p_\tau)$ and observe that $\phi(\lambda^{-1}p_\tau) = \phi(p_1)$ by positive homogeneity. Using Proposition 5.1.4 establishes:

$$0 = \phi(\lambda^{-1}p_\tau) - \phi(p_1) \geq \langle \lambda^{-1}p_\tau - p_1, x - p_1 \rangle; \tag{5.3}$$

$$0 = \phi(\lambda p_1) - \phi(p_\tau) \geq \langle \lambda p_1 - p_\tau, x - p_\tau \rangle. \tag{5.4}$$

Taking a linear combination of these expressions yields

$$\begin{aligned} 0 &\geq \lambda \langle \lambda^{-1}p_\tau - p_1, x - p_1 \rangle + \langle \lambda p_1 - p_\tau, x - p_\tau \rangle \\ &= \langle p_\tau - \lambda p_1, x - p_1 \rangle + \langle p_\tau - \lambda p_1, p_\tau - x \rangle \\ &= \langle p_\tau - \lambda p_1, p_\tau - p_1 \rangle \\ &= \|p_\tau - p_1\|^2 + (1-\lambda)\langle p_1, p_\tau - p_1 \rangle, \end{aligned}$$

In particular, as desired, we have

$$\langle p_1, p_\tau - p_1 \rangle \geq \frac{\|p_\tau - p_1\|^2}{\lambda - 1} > 0.$$

Consequently, $\|p_\tau\| > \|p_1\|$. Now, if $\phi(p_1) = 0$, then $p_1 \in \mathrm{null}(\phi)$. It is straightforward to show that $\mathrm{null}(\phi) = \mathrm{null}(\phi_\tau)$ is a linear subspace because

ϕ is a gauge. So, if $p_\tau \in \text{null}(\phi)$, then

$$p_\tau = \text{Proj}_{\text{null}(\phi_\tau)}(x) = \text{Proj}_{\text{null}(\phi)}(x) = p_1.$$

Otherwise there is $q_\tau \in \text{null}(\phi)^\perp$ such that $p_\tau = p_1 + q_\tau$, whence $\|p_\tau\|^2 = \|p_1\|^2 + \|q_\tau\|^2$. Chiefly, $\|p_\tau\| \geq \|p_1\|$. □

Remark 20 (Projection lemma on Hilbert space). Observe that the same elements in the proof of Theorem 5.2.2 may be used to prove a version of Lemma 2.1.2 in the same general setting. Namely, it is straightforward to show Lemma 2.1.2 holds in the case where $0 \in \mathcal{K} \in \mathcal{H}$ with $\mathcal{H}$ being an arbitrary Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \| := \sqrt{\langle \cdot, \cdot \rangle}$.

Remark 21. We seek the most general assumptions on ϕ under which a result like Theorem 5.2.2 or Lemma 2.1.2 remains true. Certainly, the indicator function of a set is not a gauge; it is not positively homogeneous, for one. As such, Lemma 2.1.2 falls outside the realm of consideration of Theorem 5.2.2. It is then natural to ask if one may state a unified version of these two results, where ϕ belongs to a class of functions containing both indicator functions of origin-containing convex sets, and proximal operators of gauges. We leave this as an open question.

5.3 Quasi-ordering of generalized projections

This section is motivated by the study of solvability of LASSO programs, in the sense seen in Chapters 3 and 4. We start by introducing a generalization of projections onto convex sets, which bears strong resemblance to $(\text{LS}_{\tau,\mathcal{K}})$. Given $\mathcal{K} \subseteq \mathbb{R}^N$, $A \in \mathbb{R}^{m \times N}$ and $y \in \mathbb{R}^m$ define $x_\alpha(y, A)$ by

$$x_\alpha(y, A) := \arg\max \{\|x_*\|_2 : x_* \in \mathcal{S}(y, A, \alpha\mathcal{K})\}, \tag{5.5}$$

$$\mathcal{S}(y, A, \alpha\mathcal{K}) := \arg\min \{\|y - Ax\|_2 : x \in \alpha\mathcal{K}\}. \tag{5.6}$$

For the duration of this chapter we assume that $\mathcal{K} \subseteq \mathbb{R}^N$ is a non-empty closed convex set containing the origin, and that $\alpha \geq 0$. We may refer to

$\mathcal{S}(y, A, \alpha\mathcal{K})$ colloquially as the *solution set*. The solution set is non-empty because $0 \in \alpha\mathcal{K}$. If the solution set contains more than one element, then $x_\alpha(y, A)$ corresponds to a member of the solution set with maximal ℓ_2 norm. In the event that the solution set has more than one element with largest ℓ_2 norm, simply order those members lexicographically by their coordinate values and define $x_\alpha(y, A)$ to be the first. In this way we write equality in (5.5).

Theorem 5.3.1 (Generalized projection lemma)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^N$ be a non-empty closed and convex set with $0 \in \mathcal{K}$ and $\operatorname{rad}(\mathcal{K}) \leq 2$. Fix $\lambda \geq 1$, suppose that $A \in \mathbb{R}^{m\times N}$ is a normalized K-subgaussian matrix and let $y \in \mathbb{R}^m$. There exists $\delta > 0$ sufficiently small so that if*

$$m > C\delta^{-2}\tilde{K}^2\, \mathrm{w}^2(\mathcal{K}),$$

then with probability at least $1 - 3\exp(-cm\delta^2/\tilde{K}^2)$ on the realization of A,

$$\|x_1(y, A)\|_2 \leq \|x_\lambda(y, A)\|_2.$$

Proof of Theorem 5.3.1. If $\lambda = 1$ the result follows trivially, so we assume $\lambda > 1$. We assume that $y \in \operatorname{range}(A)$. If A is from a continuous distribution, it has full rank almost surely, and so $y \in \operatorname{range}(A)$ almost surely. The case $y \notin \operatorname{range}(A)$ follows easily.

Define $\mathcal{Z}_y := \{x \in \mathbb{R}^N : y = Ax\}$. We first assume $|\lambda\mathcal{K} \cap \mathcal{Z}_y| \leq 1$. In this setting, $x_1(y, A)$ and $x_\lambda(y, A)$ are unique almost surely. For brevity, define $x_\alpha := x_\alpha(y, A)$ for $\alpha \in \{1, \lambda\}$. Set

$$f(t) := \|u(t)\|_2^2 \quad \text{where} \quad u(t) := tx_\lambda(y, A) + (1 - t)x_1(y, A).$$

Our goal is to show $f'(0) \geq 0$, which implies $\|x_\lambda\|_2 \geq \|x_1\|_2$, because f is convex. As before, this is equivalent to showing $\frac{1}{2}f'(0) = \langle x_1, x_\lambda - x_1\rangle \geq 0$. If $x_1 = x_\lambda$, the result holds trivially. So, we proceed under the assumption

that $x_\lambda \neq x_1$. For $x \in \mathbb{R}^N$, define $F(x) := \|y - Ax\|_2^2$ and observe that

$$F'(x) = A^T(Ax - y),$$

whence by the optimality conditions for convex programs, it holds for all $x \in \alpha\mathcal{K}$ that

$$\langle y - Ax_\alpha, A(x - x_\alpha)\rangle \leq 0.$$

Using again the same style of trick as in § 5.1.1 and § 5.2.2, we make an appropriate subsititution and take a linear combination of the following relationships:

$$\langle A(\lambda^{-1}x_\lambda - x_1), y - Ax_1\rangle \leq 0, \qquad \langle A(\lambda x_1 - x_\lambda), y - Ax_\lambda\rangle \leq 0.$$

The linear combination yields

$$\begin{aligned} 0 &\geq \langle A(x_\lambda - \lambda x_1), y - Ax_1\rangle + \langle A(\lambda x_1 - x_\lambda), y - Ax_\lambda\rangle \\ &= \langle A(\lambda x_1 - x_\lambda), A(x_1 - x_\lambda)\rangle \\ &= (\lambda - 1)\langle Ax_1, A(x_1 - x_\lambda)\rangle + \|A(x_1 - x_\lambda)\|_2^2. \end{aligned}$$

Rearranging the expression gives

$$0 \leq \frac{\|A(x_\lambda - x_1)\|_2^2}{\lambda - 1} \leq \langle Ax_1, A(x_\lambda - x_1)\rangle.$$

Effectively, it is enough to show $\langle Ax_1, A(x_\lambda - x_1)\rangle \approx \langle x_1, x_\lambda - x_1\rangle$. For simplicity, write:

$$u := x_1 \qquad v := x_\lambda \qquad w := v - u.$$

Thus, our aim is to show that $\langle Au, Aw\rangle$ behaves like $\langle u, w\rangle$; correctly con-

trolling $\langle u, w\rangle$ finishes the proof. To this end, set $B := A^T A - I$ and write

$$\begin{aligned}\sup_{\substack{u\in\mathcal{K}\\ v\in\lambda\mathcal{K}\\ w:=v-u}} |\langle Au, Aw\rangle - \langle u, w\rangle| &= \sup_{\substack{u\in\mathcal{K}\\ v\in\lambda\mathcal{K}\\ w:=v-u}} \left|u^T B w\right| \\ &= \frac{1}{2} \sup_{\substack{u\in\mathcal{K}\\ v\in\lambda\mathcal{K}\\ w:=v-u}} \left|v^T B v - u^T B u - w^T B w\right|.\end{aligned}$$

We bound this quantity with high probability on the realization of A. Observe that $0 \in \mathcal{K}$ implies $\mathcal{K} \subseteq (1+\lambda)\mathcal{K}'$, where $\mathcal{K}' := (1+\lambda)^{-1}(\lambda\mathcal{K} - \mathcal{K})$, so it suffices to control

$$\sup_{x\in(1+\lambda)\mathcal{K}'} \left|x^T B x\right| = \sup_{x\in(1+\lambda)\mathcal{K}'} \left|\|Ax\|_2^2 - \|x\|_2^2\right|.$$

For any $\delta \in (0,1)$, if $m \geq C\delta^{-2}\tilde{K}^2\,\mathrm{w}^2(\mathcal{K}')$, then Jeong et al. [44, Corollary 1.2] yields with probability at least $1 - 3\exp(-cm\delta^2/\tilde{K}^2)$ on the realization of A that

$$\sup_{x\in(1+\lambda)\mathcal{K}'} \left|\|Ax\|_2^2 - \|x\|_2^2\right| \leq (1+\lambda)^2\delta.$$

Consequently, on such an event,

$$\begin{aligned}(\lambda-1)^{-1}\left(\|w\|_2^2 - (1+\lambda)^2\delta\right) &\leq (\lambda-1)^{-1}\|Aw\|_2^2 \\ &\leq \langle Au, Aw\rangle \\ &\leq \langle u, w\rangle + \frac{3}{2}\delta(1+\lambda)^2.\end{aligned}$$

Rearranging gives

$$\langle u, w\rangle \geq \frac{\|w\|_2^2}{\lambda-1} - \frac{(1+\lambda)^2\delta}{\lambda-1} - \frac{3}{2}\delta(1+\lambda^2).$$

The right-hand expression is positive for any

$$\delta < \frac{2\|w\|_2^2}{(1+\lambda^2)(3\lambda-1)}.$$

Finally, $\mathrm{w}(\mathcal{K}') \leq \mathrm{w}(\mathcal{K})$, so it suffices to require $m \geq C\delta^{-2}K^2 \log K \mathrm{w}^2(\mathcal{K})$. Thus, with the stated probability on the realization of A, f is a convex function increasing on the interval $t \in [0,1]$, and so $\|x_1\|_2 \leq \|x_\lambda\|_2$ as desired.

Next, assume that $|\lambda\mathcal{K} \cap \mathcal{Z}_y| > 1$. One of two cases occurs: either $\mathcal{K} \cap \mathcal{Z}_y = \emptyset$ or $\mathcal{K} \cap \mathcal{Z}_y \neq \emptyset$. In the former case, choose $\tilde{\lambda} \in (1, \lambda)$ such that $\tilde{\lambda}\mathcal{K} \cap \mathcal{Z}_y \neq \emptyset$ and $\tilde{\lambda}\mathcal{K}^\circ \cap \mathcal{Z}_y = \emptyset$ (i.e., $\mathcal{Z}_y$ has non-empty intersection only with the boundary of $\tilde{\lambda}\mathcal{K}$). We may now apply the argument above to $\mathcal{K}$ and $\tilde{\lambda}\mathcal{K}$, immediately yielding $\|x_{\tilde{\lambda}}(y,A)\|_2 \geq \|x_1\|_2$ with high probability on the realization of A for appropriate choice of δ and m. By construction, it holds that $x_{\tilde{\lambda}}(y,A) \in \mathcal{S}(y,A,\lambda\mathcal{K})$, so $\|x_1\|_2 \leq \|x_{\tilde{\lambda}}(y,A)\|_2 \leq \|x_\lambda\|_2$ by definition of x_λ. Next, in the latter case where $\mathcal{K} \cap \mathcal{Z}_y \neq \emptyset$, it holds trivially that $\mathcal{S}(y,A,\mathcal{K}) \subseteq \mathcal{S}(y,A,\lambda\mathcal{K})$, whence $\|x_1\|_2 \leq \|x_\lambda\|_2$ as desired.

□

Remark 22 (Uniformity in $y \in \mathbb{R}^m$). Observe that the bound

$$\sup_{x \in (1+\lambda)\mathcal{K}} \left| \|Ax\|_2^2 - \|x\|_2^2 \right| \leq (1+\lambda)^2\delta$$

is uniform over all $x \in (1+\lambda)\mathcal{K}$, and moreover, for all $y \in \mathbb{R}^m$ one has $u, v, w \in (1+\lambda)\mathcal{K}$. Thus, suppose there is $c > 0$ so that $\inf_{y \in \mathbb{R}^m} \|x(\lambda; y, A) - x(1; y, A)\|_2^2 \geq c$ and define

$$c(\lambda) := \frac{2c}{(1+\lambda^2)(3\lambda - 1)}.$$

Then, for all $y \in \mathbb{R}^m$, for any $\delta \in (0, c(\lambda))$, if $m \geq C\delta^{-2}K^2 \log K \, \mathrm{w}(\mathcal{K})$, then it holds with probability at least $1 - 3\exp(-cm\delta^{-2}/(K^2 \log K))$ that $\|x(\lambda; y, A)\|_2 \geq \|x(1; y, A)\|_2$.

Corollary 5.3.2. *Let $\mathcal{K} \subseteq \mathbb{R}^N$ be a non-empty closed and convex set with $0 \in \mathcal{K}$ and $\mathrm{rad}(\mathcal{K}) \leq 2$. Suppose that $A \in \mathbb{R}^{m \times N}$ is a K-subgaussian matrix and let $y \in \mathbb{R}^m$. For $\beta > 0$, define*

$$\tilde{x}(\beta; y, A) := \arg\min\{\|x\|_{\mathcal{K}} : \|y - Ax\|_2 \leq \beta\}.$$

There exists $\delta > 0$ sufficiently small and a constant $\lambda > 0$, each depending on β_0, β_1, y and A, so that if

$$m > C\delta^{-2}K^2 \log(K) \mathrm{w}^2(\mathcal{K}),$$

then with probability at least $1 - 3\exp\left(-cm\delta^2/(K^2 \log K)\right)$ on the realization of A,

$$\|\tilde{x}(\beta_0; y, A)\|_2 \geq \|\tilde{x}(\beta_1; y, A)\|_2.$$

Proof of Corollary 5.3.2. Observe that the problem defining $\tilde{x}(\beta; y, A)$ is a generalized LASSO program. In particular, by Proposition 4.1.4, given $0 < \beta_0 < \beta_1 < \infty$, there exists $0 \leq \alpha_1 \leq \alpha_0 < \infty$ so that

$$x(\alpha_i; y, A) = \tilde{x}(\beta_i; y, A), \quad i = 0, 1.$$

If $x(\alpha_0; y, A) = x(\alpha_1; y, A)$, then there is nothing left to prove. Otherwise, define $\lambda := \alpha_0 \alpha_1^{-1}$. By Theorem 5.3.1 there exists $\delta > 0$ satisfying

$$\delta < \frac{2\|x(\lambda; \alpha_1^{-1} y, A) - x(1; \alpha_1^{-1} y, A)\|_2^2}{(1+\lambda)^2(3\lambda - 1)}$$

so that if $m > CK^2 \log K \mathrm{w}^2(\mathcal{K})$, then with probability at least $1 - 3\exp(-cm\delta^2/(K^2 \log K))$ on A,

$$\alpha_1 \|x(1; \alpha_1^{-1} y, A)\|_2 = \|x(\alpha_1; y, A)\|_2 \leq \|x(\alpha_0; y, A)\|_2 = \alpha_1 \|x(\lambda; \alpha_1^{-1} y, A)\|_2.$$

In particular, one has, as desired:

$$\|\tilde{x}(\beta_0; y, A)\|_2 \geq \|\tilde{x}(\beta_1; y, A)\|_2.$$

□

5.4 Applications

As already noted, a version of Lemma 2.1.2 was first developed [55, Lemma 15.3] to control the worst-case error for a constrained denoising problem. The present variant was used to control worst-case error for a constrained denoising problem for arbitrary parameter choices [9]. Additionally, it is used to obtain an ordering of the solutions for a basis pursuit PD problem, in order to obtain lower bounds on the maximin error of the problem [9]. The simpler variant has moreover been adapted to establish lower bounds on the minimax error for (BP_σ) problems in a particular data regime [10]. In Berk et al. [10], it is also used to compare worst-case error for (LS_τ) with arbitrary values of its governing parameter.

First, we present a "non-example". Observe that Theorem 5.2.2 may not be used to analyze $R^\sharp$ as in Lemma 3.6.1. There, we relied on $x_0 + B_1^N$ being a set that gives rise to a gauge, whereas $\|x_0 + \cdot\|_1$ is not itself a gauge.

Next, let us analyze (QP_λ^*) of Chapter 3 through the lens of Theorem 5.2.2. In particular, we analyze its ability to recover the 0 vector. As argued previously, efficacy of recovering 0 is a desirable property of a denoiser. Suppose, then that $\mathcal{H} = \mathbb{R}^N$ and $\|\cdot\| = \|\cdot\|_2$. By definition, for $z \in \mathcal{H}$,

$$x^\sharp(\lambda; z) := \operatorname*{arg\,min}_{x \in \mathcal{H}} \{\frac{1}{2}\|z - x\|_2^2 + \lambda \|x\|_1\} = \operatorname{prox}_\lambda \|\cdot\|_1(z).$$

Setting $\phi := \|\cdot\|_1$ and defining $\phi_\lambda(x) = \phi(x/\lambda)$ as above, observe that

$$x^\sharp(\lambda; z) = \operatorname{prox} \phi_{\lambda^{-1}}(z).$$

Indeed, we know that $x^\sharp(\lambda; z)$ is simply the soft-thresholding of z by λ. Separately applying Proposition 5.2.1 and Theorem 5.2.2 shows for $p = 1, 2$ that $\|x^\sharp(\lambda; z)\|_p \leq \|x^\sharp(\lambda'; z)\|_p$ when $\lambda \geq \lambda'$. Moreover, it is easy to use similar ideas from the proof of Theorem 5.2.2 to obtain the relationship for $\lambda > 0$,

$$\langle z, x^\sharp(\lambda; z) - z \rangle \geq \lambda^{-1} \|z - x^\sharp(\lambda; z)\|_2^2.$$

While this example was presented in the PD context discussed earlier, we observe that this relationship remains valid for arbitrary $\mathcal{H}$ and choice of $\|\cdot\|$.

It would be interesting to explore if this analysis could be adapted to analyze novel properties of common approaches to *dictionary learning*, or an infinite-dimensional de-noising problem. For example, the Beurling LASSO [12, 63] seeks to recover an unknown sparse measure

$$m_{a,\theta} := \sum_{i=1}^{n} a_i \delta_{\theta_i}, \qquad a_i \in \mathbb{R}, \theta_i \in \Theta \subseteq \mathbb{R}^N,$$

using measurements $y \in \mathcal{H}$ that are acquired via a continuous linear functional Ψ of the form

$$\Psi m := \int_{\Theta} \psi(\theta) \, \mathrm{d}m.$$

Here, $\psi : \Theta \to \mathcal{H}$ is a continuous function mapping into some Hilbert space $\mathcal{H}$; necessarily, $\Psi : \mathcal{M}(\Theta) \to \mathcal{H}$, where $\mathcal{M}(\Theta)$ is the space of compactly supported Radon measures. The unconstrained Beurling LASSO [12, 63] is then defined by

$$\min_{m \in \mathcal{M}(\Theta)} \frac{1}{2} \|\Psi m - y\|_2^2 + \lambda \|m\|_{\mathrm{TV}}$$

where $\|m\|_{\mathrm{TV}} := \sup_{\{\mathcal{A}_i\} \subseteq \Theta} \sum_i |m(\mathcal{A}_i)|$ denotes the total variation norm of the measure m. Similarly, Beurling basis pursuit is given by

$$\min_{m \in \mathcal{M}(\Theta)} \|m\|_{\mathrm{TV}} \quad \text{such that} \quad \Psi m = y.$$

Natural questions may thus concern the regimes in which each Beurling LASSO variant exhibits PS; it is natural to expect that answers to these questions may rely on the Hilbert space generalization of Lemma 2.1.2.

Chapter 6

Conclusion

We conclude by discussing the results of this thesis in a broader context. In addition, we discuss intriguing avenues for future work.

6.1 Proximal denoising

Chapter 3 of this thesis investigated the stability of three sibling convex programs. Specifically, it investigated the stability of their recovery error with respect to variation in their governing parameter. We have illustrated regimes in which each of the three PD programs is parameter sensitive. The theory of § 3.2–3.4 proves asymptotic results for each program, while the numerics of § 3.5 support using the asymptotic behaviour as a basis for practical intuition. Thus, we hope these results inform practitioners about which program to use.

In § 3.2 and § 3.5.1 we observe that (LS_τ^*) exhibits parameter sensitivity in the low-noise regime. The risk $\hat{R}(\tau; x_0, N, \eta)$ develops an asymptotic singularity as $\eta \to 0$, blowing up for any $\tau \neq \|x_0\|_1$, where $\hat{R}(\|x_0\|_1; x_0, N, \eta)$ attains minimax order-optimal error. Numerical simulations support that $\hat{R}(\tau; x_0, N, \eta)$ develops cusp-like behaviour in the low-noise regime, which agrees with the asymptotic singularity of Theorem 3.2.1. Notably, (LS_τ^*) parameter sensitivity manifests in very low dimensions relative to practical problem sizes. Outside of the low-noise regime, (LS_τ^*) appears to exhibit

better parameter stability, as exemplified in Figures 3.5, 3.9 and 3.13.

In § 3.3 and § 3.5.2 we observe that (QP^*_λ) exhibits left-sided parameter sensitivity in the low-noise regime. When $\lambda < \bar{\lambda}$ we prove that $R^\sharp(\lambda; s, N)$ scales asymptotically as a power law of N. The suboptimal scaling of the risk manifests in relatively higher dimensional problems, as suggested by Figure 3.3a. Minimax order-optimal scaling of the risk when $\lambda \geq \bar{\lambda}$ is clear from Figure 3.3c. The numerics of § 3.5 support that (QP^*_λ) is generally the most stable of the three PD programs considered.

In § 3.4 and § 3.5.3 we observe that (BP^*_σ) exhibits parameter sensitivity in the very sparse regime. Notably, $\tilde{R}(\sigma; x_0, N, \eta)$ is maximin suboptimal for *any* choice of $\sigma > 0$ for s/N sufficiently small. This behaviour is supported by Figure 3.4a, in which the best average loss of (BP^*_σ) is 82.2 times worse than that for (LS^*_τ) and (QP^*_λ). Further, the average loss for (BP^*_σ) exhibits a clear cusp-like behaviour in Figure 3.4a, like for that of (LS^*_τ), which would be an interesting object of further study. Outside of the very sparse regime, (BP^*_σ) appears to exhibit parameter stability, as exemplified in Figure 3.5.

Finally, in § 3.5.6 we portray how estimators behave as a function of the normalized parameter for each program. We show the kinds of pathologies from which these estimators suffer in regimes of sensitivity. These examples were performed in the setting of relatively more realistic data: a modification of the Shepp-Logan phantom.

6.2 Compressed sensing

Chapter 4 of this thesis examined the relative sensitivity of three Lasso programs to their governing parameters: (LS_τ), (BP_σ) and (QP_λ). We proved asymptotic cusp-like behaviour of $\hat{R}(\tau; x_0, A, \eta)$ in the limiting low-noise regime in § 4.2. Numerical simulations in § 4.5.1 support these observations for even modest dimensional parameters and noise scales, as do additional numerics later in § 4.5. We motivated how § 4.2 may be applicable to the $(\mathrm{LS}_{\tau,\mathcal{K}})$ setting by highlighting parameter instability of nuclear norm recovery in Lemma A.1.1. Interestingly, the theoretical and numerical results for (LS_τ) closely mirror those obtained for (LS^*_τ).

In § 4.3, we recall a result establishing right-sided stability of (QP_λ) for a class of matrices that satisfy a version of RIP. The result does not address sensitivity of (QP_λ) to its governing parameter when the governing parameter is less than its optimal value. In § 4.5, we demonstrate numerically that there are regimes in which (QP_λ) is sensitive to its governing parameter λ when $\lambda < \lambda^*$. This sensitivity is readily observed in the rightmost plot of Figure 4.2. This observation establishes a numerical connection to the numerics and theory of Chapter 3. Moreover, we observe that (QP_λ) is more sensitive to its choice of parameter when the aspect ratio is larger. We believe this is due to there being a smaller null-space, which has the effect of shrinking the space of possible solutions. This behaviour is visible in both plots of Figure 4.2: the error curves for larger δ are steeper for $\lambda < \lambda^*$ than those for smaller values of δ.

In § 4.4, we proved asymptotic suboptimality of $\tilde{R}(\sigma; x_0, A, \eta)$ in a certain dimensional regime that falls outside the typical CS regime where $m \approx Cs\log(N/s)$. In particular, for $m \approx \delta N$, $\delta \in (0,1)$, we show that (BP_σ) risk is asymptotically suboptimal for "very sparse" signals. We demonstrate that this theory is relevant to the CS regime in § 4.5, in which we show that the loss and average loss for (BP_σ) are sensitive to the value of the governing parameter if the ground-truth signal is very sparse. Furthermore, Figure 4.4 and Figure 4.6 depict suboptimality of the (BP_σ) risk for modest choices of dimesional parameters. Again, it is interesting to note that these results closely mirror those obtained for (BP^*_σ).

6.3 Projection lemma and extensions

We have demonstrated the usefulness of Lemma 2.1.2, which was an integral component of several results in this thesis. By this result, the size of $\tilde{x}(\eta\sqrt{N})$ controls the size of $\tilde{x}(\sigma)$ for $\sigma \leq \eta\sqrt{N}$ when $x_0 \equiv 0$. This was key to demonstrating risk suboptimality for underconstrained (BP^*_σ) and (BP_σ). Moreover, Lemma 2.1.2 was used to prove that $\hat{R}(\tau; \tau x_0, N, \eta)$ is an increasing function of τ when $\|x_0\|_1 = 1$ (Lemma 3.6.1) and similarly for $\hat{R}$ in the CS setting (Proposition 4.6.2). Thus, the projection lemma was particularly

effective for proving minimax order-optimality of $R^*(s, N)$ and $R^*(s, A)$.

In Chapter 5, we showed that the projection lemma may be generalized along two directions. In particular, Theorem 5.2.2 extends the projection lemma to the more general setting of arbitrary gauges defined on an arbitrary Hilbert space. The proof of Theorem 5.2.2 itself shows that Lemma 2.1.2 holds in the same general setting. Thus, as argued in § 5.4, equivalent notions of parameter sensitivity could be extended to the infinite dimensional setting. On the other hand, Theorem 5.3.1 establishes an ordering for projection-like operators, namely those resembling ($\mathrm{LS}_{\tau,\mathcal{K}}$) when the structure set $\mathcal{K}$ gives rise to a gauge and A is a subgaussian random matrix. This extension is imperfect in that it necessitates a sufficiently small choice of δ, meaning that it is not possible to use this result to obtain uniform statements, like those in § 4.6.2. Nevertheless, we believe this result reinforces the satisfying geometric link between the PD and CS problems. As argued in § 5.4, we hope that this result may motivate similar study of other variants of projection-like operators.

6.4 Future directions

Future works related to Chapter 3 include extending the main results in the PD setting to more general atomic norms. These results may also extend to ones under more general noise models. It would be interesting to see what role parameter instability might play in proximal point algorithms and those algorithms relying on proximal operators. Conversely, it would be useful to understand rigorously when a PD program exhibits parameter sensitivity, and to determine systematically the regime in which that sensitivity arises. A non-asymptotic theoretical analysis of parameter sensitivity remains an open problem.

Future works related to Chapter 4 include fully extending the main results to the generalized Lasso setting, using more general atomic norms. A rigorous examination of low-rank matrix recovery, particularly the (BP_σ) variant, could be interesting. Finally, it would be useful to understand when a convex program is expected to exhibit sensitivity to its governing param-

eter, and to determine systematically the regime in which that instability arises. Again, a non-asymptotic analysis for the CS setting remains open. Finally, it would be interesting to establish how parameter sensitivity relates to other popular LASSO variants, such as the so-called square-root LASSO, which is touted to be "noise-scale agnostic".

Future works related to Chapter 5 are relatively open. For instance, applications of Theorem 5.2.2 remain largely unexplored. Separately, note that indicator functions like $\chi_{\mathcal{K}}$ are not gauges. Thus, a notion of ordering has been established for both projection operators and for proximal operators of gauges. In particular, it is an open question to determine the largest class of functions that satisfy an inequality like that of Theorem 5.2.2. Moreover, the uniform ("coarse") analysis in the proof of Theorem 5.3.1 gave a result that requires a specific range for δ. It would be interesting to determine if a local analysis in the proof could relax the requirements on δ, improving the potential applicability of the theorem in applications.

Bibliography

[1] R. J. Adler and J. E. Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009. → pages 20, 23, 24

[2] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014. → pages 3, 11, 18, 65

[3] S. Arora and B. Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009. → page 4

[4] H. H. Bauschke, P. L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011. → page 161

[5] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011. → pages 11, 12

[6] M. Bayati and A. Montanari. The Lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011. → pages 11, 12

[7] P. C. Bellec. Localized gaussian width of m-convex hulls with applications to Lasso and convex aggregation. *Bernoulli*, 25(4A): 3016–3040, 2019. → pages 17, 28, 93

[8] A. Berk, Y. Plan, and Ö. Yilmaz. Parameter instability regimes in sparse proximal denoising programs. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–5. IEEE, 2019. → pages 10, 13, 105

[9] A. Berk, Y. Plan, and Ö. Yilmaz. Sensitivity of ℓ_1 minimization to parameter choice. *Information and Inference: A Journal of the IMA*, 2020. doi:10.1093/imaiai/iaaa014. URL https://doi.org/10.1093/imaiai/iaaa014. → pages 10, 13, 20, 34, 105, 107, 130, 155, 158, 170

[10] A. Berk, Y. Plan, and Ö. Yilmaz. On the best choice of Lasso program given data parameters. *IEEE Transactions on Information Theory*, Submitted October 2020. 48 pages. Submission ID: IT-20-0773. (arXiv preprint arXiv:2010.08884). → pages 13, 170

[11] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, Belmont, MA, USA, 2003. ISBN 1-886529-45-0. → pages 11, 126, 155, 156, 159

[12] A. Beurling. Sur les intégrales de fourier absolument convergentes et leur application à une transformation fonctionelle. In *Ninth Scandinavian Mathematical Congress*, pages 345–366, 1938. → page 171

[13] P. J. Bickel, Y. Ritov, A. B. Tsybakov, et al. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. → pages 2, 3, 9, 12, 75, 90, 94, 110

[14] J. Bolte, P. L. Combettes, and J.-C. Pesquet. Alternating proximal algorithm for blind image recovery. In *2010 IEEE International Conference on Image Processing*, pages 1673–1676. IEEE, 2010. → page 11

[15] C. Borell. The brunn-minkowski inequality in gauss space. *Inventiones mathematicae*, 30(2):207–216, 1975. → page 24

[16] M. D. Buhmann. *Radial basis functions: theory and implementations*, volume 12. Cambridge university press, 2003. → page 99

[17] E. J. Candes and M. A. Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2): 317–323, 2013. → pages 4, 65, 110, 124

[18] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010. → page 188

[19] E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4): 2342–2359, 2011. → page 188

[20] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006. → page 2

[21] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[22] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006. → page 2

[23] S. Chatterjee. A new perspective on least squares under convex constraint. *Annals of Statistics*, 42(6):2340–2381, 2014. → page 11

[24] S. Chen and D. Donoho. Basis pursuit. In *28th Asilomar Conf. Signals, Systems Computers*, 1994. → pages 2, 4, 12

[25] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001. → pages 2, 4, 12

[26] B. S. Cirel'son, I. A. Ibragimov, and V. Sudakov. Norms of gaussian sample functions. In *Proceedings of the Third JapanâĂŤUSSR Symposium on Probability Theory*, pages 20–41. Springer, 1976. → page 24

[27] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011. → page 11

[28] I. Daubechies and G. Teschke. Variational image restoration by means of wavelets: Simultaneous decomposition, deblurring, and denoising. *Applied and Computational Harmonic Analysis*, 19(1):1–16, 2005. → page 7

[29] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. *Introduction to Compressed Sensing*, page 1âĂŞ64. Cambridge University Press, 2012. doi:10.1017/CBO9780511794308.002. → page 2

[30] D. L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006. → page 2

[31] J. Eckstein and D. P. Bertsekas. On the douglasâĂŤrachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992. → page 11

[32] M. Elad. Sparse and redundant representation modelingâĂŤwhat next? *IEEE Signal Processing Letters*, 19(12):922–928, 2012. → page 11

[33] M. Elad, M. A. Figueiredo, and Y. Ma. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98(6):972–982, 2010. → page 11

[34] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002. → page 7

[35] S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing. *Bull. Am. Math*, 54:151–165, 2017. → pages 1, 2, 3, 4, 8, 12, 20, 23, 27, 32, 65, 91

[36] M. P. Friedlander, I. Macedo, and T. K. Pong. Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014. → page 5

[37] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010. → page 5

[38] J. B. Garnett, T. M. Le, Y. Meyer, and L. A. Vese. Image decompositions using bounded variation and generalized homogeneous besov spaces. *Applied and Computational Harmonic Analysis*, 23(1): 25–56, 2007. → page 1

[39] P. Hand and V. Voroninski. Global guarantees for enforcing deep generative priors by empirical risk. *IEEE Transactions on Information Theory*, 66(1):401–418, 2019. → page 3

[40] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning: prediction, inference and data mining. *Springer-Verlag, New York*, 2009. → page 99

[41] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019. → pages 3, 12

[42] F. J. Herrmann, M. P. Friedlander, and O. Yilmaz. Fighting the curse of dimensionality: Compressive sensing in exploration seismology. *IEEE Signal Processing Magazine*, 29(3):88–100, 2012. → page 3

[43] H. Jeong, X. Li, Y. Plan, and Ö. Yilmaz. Non-gaussian random matrices on sets: Optimal tail dependence and applications. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–5. IEEE, 2019. → page 25

[44] H. Jeong, X. Li, Y. Plan, and Ö. Yılmaz. Sub-gaussian matrices on sets: Optimal tail dependence and applications. *arXiv preprint arXiv:2001.10631*, 2020. → page 167

[45] D. E. Knuth. Big omicron and big omega and big theta. *ACM Sigact News*, 8(2):18–24, 1976. → page 16

[46] R. Kumar, H. Wason, and F. J. Herrmann. Source separation for simultaneous towed-streamer marine acquisitionâĂŤa compressed sensing approach. *Geophysics*, 80(6):WD73–WD88, 2015. → page 3

[47] C. Liaw, A. Mehrabian, Y. Plan, and R. Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric aspects of functional analysis*, pages 277–299. Springer, 2017. → pages 3, 8, 12, 25, 91, 93, 110, 111, 112, 114

[48] M. Lustig, D. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. → page 1

[49] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. Compressed sensing MRI. *IEEE signal processing magazine*, 25(2):72–82, 2008. → page 1

[50] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019. → pages 3, 12

[51] Y. Meyer. *Oscillating patterns in image processing and nonlinear evolution equations: the fifteenth Dean Jacqueline B. Lewis memorial lectures*, volume 22. American Mathematical Soc., 2001. → page 1

[52] L. Miolane and A. Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*, 2018. → page 11

[53] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. → page 99

[54] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995. → page 4

[55] S. Oymak and B. Hassibi. Sharp mse bounds for proximal denoising. *Foundations of Computational Mathematics*, 16(4):965–1029, 2016. → pages 11, 17, 19, 20, 31, 32, 37, 38, 75, 155, 170

[56] S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized Lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009. IEEE, 2013. → pages 3, 8, 12, 110

[57] M. Y. Park and T. Hastie. L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007. → page 5

[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. → pages 45, 49

[59] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012. → pages 19, 26

[60] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66 (8):1275–1297, 2013. doi:10.1002/cpa.21442. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21442. → page 27

[61] Y. Plan and R. Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, 51(2): 438–461, 2014. → page 19

[62] Y. Plan and R. Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3): 1528–1537, 2016. → pages 12, 19

[63] C. Poon, N. Keriven, and G. Peyré. The geometry of off-the-grid compressed sensing. *arXiv preprint arXiv:1802.08464*, 2018. → page 171

[64] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976. → page 11

[65] R. T. Rockafeller and R. J.-B. Wets. *Variational Analysis*. Springer-Verlag, 2009. → page 159

[66] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4): 259–268, 1992. → page 7

[67] Y. Shen, B. Han, and E. Braverman. Stable recovery of analysis based approaches. *Applied and Computational Harmonic Analysis*, 39(1): 161–172, 2015. → pages 9, 90, 94

[68] L. A. Shepp and B. F. Logan. The fourier reconstruction of a head section. *IEEE Transactions on nuclear science*, 21(3):21–43, 1974. → page 10

[69] M. Stojnic. A framework to characterize performance of Lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013. → page 2

[70] C. Thrampoulidis, S. Oymak, and B. Hassibi. Regularized linear regression: A precise analysis of the estimation error. *Proceedings of Machine Learning Research*, 40:1683–1709, 2015. → pages 11, 12

[71] C. Thrampoulidis, A. Panahi, and B. Hassibi. Asymptotically exact error analysis for the generalized ℓ_2^2-Lasso. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2021–2025. IEEE, 2015. → pages 9, 12, 90, 92, 94

[72] C. Thrampoulidis, E. Abbasi, and B. Hassibi. Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018. → pages 3, 9, 11, 12, 90, 92, 94

[73] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58 (1):267–288, 1996. → pages 2, 12

[74] R. J. Tibshirani. The Lasso problem and uniqueness. *Electronic Journal of statistics*, 7:1456–1490, 2013. → page 6

[75] E. Van Den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31 (2):890–912, 2008. → pages 2, 3, 5, 11, 12

[76] R. van Handel. Probability in high dimension. Technical report, Princeton University, NJ, 2014. → page 20

[77] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018. → pages 3, 20, 21, 22, 24, 124, 127

[78] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi:https://doi.org/10.1038/s41592-019-0686-2. → pages 54, 60, 99

[79] P. Wojtaszczyk. Stability and instance optimality for gaussian measurements in compressed sensing. *Foundations of Computational Mathematics*, 10(1):1–13, 2010. → pages 11, 12

[80] H. Zhang, W. Yin, and L. Cheng. Necessary and sufficient conditions of solution uniqueness in ℓ_1-norm minimization. *Journal of Optimization Theory and Applications*, 164(1):109–122, 2015. → page 6

Appendix A

Supporting Materials

A.1 Auxiliary proofs (CS)

A.1.1 Proofs for refinements on bounds for gmw

Proof of Corollary 2.3.7. Assuming

$$m > C_\varepsilon \delta^{-2} \tilde{K}^2 s \log \frac{2N}{s},$$

Lemma 4.6.4 gives

$$\sup_{x \in \mathcal{L}_s(1)} |\|Ax\|_2 - \|x\|_2| < \delta,$$

with probability at least $1 - \varepsilon$ on the realization of A. In particular, $(1 - \delta)^{-1}\|A^j\|_2 \geq 1$ and $(1 + \delta)^{-1}\|A^j\|_2 \leq 1$ for all $j \in [N]$. Define the sets

$$T_+ := \mathrm{cvx}\{\pm(1+\delta)^{-1} A^j : j \in [N]\},$$
$$T_- := \mathrm{cvx}\{\pm(1-\delta)^{-1} A^j : j \in [N]\}.$$

We will apply Proposition 2.3.5 and Proposition 2.3.6 to T_+ and T_-, respectively, then use that $T_+ = (1+\delta)^{-1}T$ and $T_- = (1-\delta)^{-1}T$. Indeed, observe

that

$$
\begin{aligned}
&(1+\delta)^{-1}\,\mathrm{w}(T\cap(1+\delta)\gamma B_2^m)\\
&\qquad=\mathrm{w}((1+\delta)^{-1}T\cap\gamma B_2^m)=\mathrm{w}(T_+\cap\gamma B_2^m)\\
&\qquad\le\min\left\{4\sqrt{\max\left\{1,\log(8eN\gamma^2)\right\}},\gamma\sqrt{\min\{m,2N\}}\right\}.
\end{aligned}
$$

Rearranging, with $\alpha=(1+\delta)\gamma$ gives

$$
\begin{aligned}
&\mathrm{w}(T\cap\alpha B_2^m)\\
&\le\min\left\{4(1+\delta)\sqrt{\max\left\{1,\log(8eN(1+\delta)^{-2}\alpha^2)\right\}},\alpha\sqrt{\min\{m,2N\}}\right\}.
\end{aligned}
$$

Similarly, one may derive the lower bound for $\alpha\in(0,(1-\delta))$, using that $\kappa=1-\delta$,

$$
\mathrm{w}(T\cap\alpha B_2^m)\ge(\sqrt{2}/4)(1-\delta)^2\sqrt{\log\frac{N\alpha^2}{5(1-\delta)^2}}.
$$

□

A.1.2 Proofs for projection lemma

Proof of Corollary 2.1.3. Define $\beta:=\|q_\alpha\|_{\mathcal{K}}$. Then $q_\alpha\in\beta\mathcal{K}$ and so $\|y-\mathrm{Proj}_{\beta\mathcal{K}}(y)\|_2\le\alpha$ by definition of $\mathrm{Proj}_{\beta\mathcal{K}}(\cdot)$. Again by definition of $\mathrm{Proj}_{\beta\mathcal{K}}(\cdot)$, it holds that $\|\mathrm{Proj}_{\beta\mathcal{K}}(y)\|_{\mathcal{K}}\le\beta$. In particular, $\mathrm{Proj}_{\beta\mathcal{K}}(y)$ is feasible and

$$
\|\mathrm{Proj}_{\beta\mathcal{K}}(y)\|_{\mathcal{K}}\le\|q_\alpha\|_{\mathcal{K}},
$$

whence optimality of q_α implies $q_\alpha=\mathrm{Proj}_{\beta\mathcal{K}}(y)$. Thus, by an elementary sequence of steps, the proof follows from the projection lemma. □

A.1.3 Parameter sensitivity of nuclear norm recovery

This section is a supplement to § 4.6.3, in support of the comment made at the end of § 4.2. We include a result that ports the two lemmata of § 4.6.3 from the setting of constrained Lasso to that of constrained nuclear norm

recovery. Before we state the lemma, define $B_* := \{X \in \mathbb{R}^{d\times d} : \|X\|_* = \sum_{i=1}^d \sigma_i(X) \leq r\}$, where $\sigma_i(X)$ is the ith largest singular value of the matrix X. Recall that $\|\cdot\|_*$ is dual to the operator norm $\|\cdot\|$ and so they admit the following inequality for any $X, Y \in \mathbb{R}^{d\times d}$

$$\langle X, Y\rangle \leq \|X\|\|Y\|_*.$$

Finally, since the space of matrices is finite, all matrix norms are equivalent. In particular, if X is a rank r matrix (i.e., $\sigma_i(X) = 0$ for $i = r+1, \ldots, d$), then

$$\|X\|_F \leq \|X\|_* \leq \sqrt{r}\|X\|_F.$$

Lemma A.1.1 (Nuclear Norm Recovery)**.** *Let $\mathcal{A} : \mathbb{R}^{d\times d} \to \mathbb{R}^m$ be an operator mapping $\mathcal{A}X = (\langle A_i, X\rangle)_{i=1}^m$ for $A_i \in \mathbb{R}^{d\times d}$. Given $X_0 \in \mathbb{R}^{d\times d}$, $\eta > 0$ and $z \in \mathbb{R}^m$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$, let $y = \mathcal{A}X_0 + \eta z$. Suppose that either $\tau > \|X_0\|_*$ and $\dim\ker\mathcal{A} > 0$ or $\tau < \|X_0\|_*$. Almost surely on the realization of z,*

$$\begin{aligned}\lim_{\eta\to 0} \hat{L}(\tau; X_0, \mathcal{A}, \eta z) &= \lim_{\eta\to 0} \eta^{-2}\|\hat{X}(\tau; y, \mathcal{A}, B_*) - X_0\|_F^2 \\ &= \infty.\end{aligned}$$

Proof of Lemma A.1.1. First, we examine the setting where $\tau > \|X_0\|_*$. In like manner as the proof for Lemma 4.6.6, define $\tau^* := \|X_0\|_*$ and set $\rho := \tau - \tau^*$. Assume $\operatorname{span}(\mathcal{A}) = \mathbb{R}^m$. There exists $\zeta \in \mathbb{R}^{d\times d}$ such that $\mathcal{A}\zeta = z$, and so $\mathcal{A}(X_0 + \eta\zeta) = y$. If η is sufficiently small then $X' := X_0 + \eta\zeta \in \tau B_*$ where $B_* = \{X \in \mathbb{R}^{d\times d} : \|X\|_* \leq 1\}$ is the nuclear norm ball. In particular, X' is a solution for $(\mathrm{LS}_{\tau,\mathcal{K}})$ where $\mathcal{K} = B_*$. Applying a similar argument as before, noting that

$$\|X' - X_0\|_F \geq \frac{\|X'\|_* - \|X_0\|_*}{d} \geq \rho > 0,$$

will complete the proof in the case where $\mathcal{A}$ spans $\mathbb{R}^m$. The case where $\operatorname{span}\mathcal{A} \subsetneq \mathbb{R}^m$ is similar.

Next, we examine the setting where $\tau < \|X_0\|_*$. For any solution Ξ to $(\mathrm{LS}_{\tau,\mathcal{K}})$, one has by norm equivalence and triangle inequality:

$$\hat{L}(\tau; X_0, \mathcal{A}, B_*) \geq \eta^{-2}\|X_0 - \Xi\|_F^2 \geq \frac{(\|X_0\|_* - \|\Xi\|_*)^2}{\eta^2 d^2} \xrightarrow{\eta \to 0} \infty.$$

□

Note that risk bounds are well-known for constrained nuclear norm recovery in the case where $\tau = \|X_0\|_*$ and $\mathcal{A} : X \mapsto (\langle A_i, X\rangle)_{i=1}^m$ with $A_i \in \mathbb{R}^{d\times d}$ independent and having independent subgaussian entries [18, 19]. In particular, combining such a result with the lemmata above gives an analogue to Theorem 4.2.1 in the constrained nuclear norm setting.

A.2 RBF Approximation

A.2.1 Why RBF approximation?

Before we describe the radial basis function (RBF) approximation method in detail, we include here a motivation for our choice to use RBF interpolation to obtain an approximation to the average loss. We describe the motivation in reference to the numerics of § 4.5, but the motivation equally well applies to those instances in § 3.5 where RBF approximation was used. As noted in those sections, the desideratum is the quantity $\bar{L}(\rho_i)$, where ρ_i is a normalized parameter value in a grid of such points $\{\rho_i\}_{i\in[n]}$. Recall from (4.2) ((3.2) for the PD setting) that for each i, $\bar{L}(\rho_i)$ is an average of k points. Unfortunately, computing this average directly is computationally intractable. In particular, this approach would involve solving each of the three programs directly for each ρ_i and each $\hat{z}_j$. Thus, taking the direct approach means solving $3nk$ problem instances (particularly burdensome in the CS setting where one is required to re-sample a realization $\hat{A}$ of $A \in \mathbb{R}^{m\times N}$). Exacerbating this issue is that (LS_τ) and (BP_σ) are significantly more computationally intensive than (QP_λ).

Instead, loss values for (LS_τ) and (BP_σ) were obtained by solving (QP_λ) for certain parameter values λ, and using Proposition 4.1.4 to obtain the corresponding parameter values for (LS_τ) and (BP_σ) that correspond to the obtained loss. This reduces the number of solved problem instances to nk, using only the fastest of the three programs, (QP_λ). Consequently, the collection of nk loss values for each of (LS_τ) and (BP_σ) were a non-uniformly sampled point cloud. In particular, there is no way to average k loss realizations for a fixed parameter ρ_i. Instead, one seeks a *localized average* that approximates the quantity.

Several approaches exist for computing a localized average. For example, one could average the losses of all points whose parameter value is within a certain radius r of the desired value ρ_i:

$$\bar{L}(\rho_i; x_0, A) \approx \sum_{\{\rho : |\rho - \rho_i| < r\}} L(\rho; x_0, A, \eta \hat{z}_j).$$

However, few options are entirely suitable for our purposes. In the example just given, there are two key issues. The first issue is how to select an appropriate neighbourhood size, especially as the resolution of the resulting curve may vary, particularly when PS is exhibited. The second difficulty is ensuring that each ρ_i-neighbourhood contains enough points. It is not enough to select a fixed neighbourhood size r, as that could result in averaging too few points in some regions, and yielding too low a resolution in others. Otherwise suitable localized averaging alternatives, such as Gaussian process regression, are computationally intractable due to the data size. Fortunately, RBF approximation is a standard approach for localized averaging that is both forgiving with hyperparameter selection, and computationally tractable in the present setting.

A.2.2 RBF approximation of the average loss

For a particular program $\mathfrak{P}$, $A \in \mathbb{R}^{m \times N}$, $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\eta > 0$ and $x_0 \in \Sigma_s^N$, define $y^{(j)} = Ax_0 + \eta \hat{z}^{(j)}, j \in [k]$ where $\hat{z}^{(j)}$ is an iid copy of z. For the program $\mathfrak{P}$ denote the solution to the program by $x^*(\upsilon; y^{(j)}, A)$, where $\upsilon > 0$

is the governing parameter. Here, $\upsilon \in \Upsilon^{(j)}$, where $\Upsilon^{(j)}$ is a logarithmically spaced grid of n points whose centre is approximately equal to the optimal parameter choice $\upsilon^*(x_0, A, \eta z^{(j)})$. Next, define the concatenated vectors

$$\mathbf{z} := (z^{(j)} : j \in [k]) \in \mathbb{R}^{kn}, \quad \mathbf{y} := (y^{(j)} : j \in [k]) \in \mathbb{R}^{kn},$$

and define the collection

$$\begin{aligned}\Gamma &:= \Gamma(x_0, A, \eta\mathbf{z}) \\ &:= \{(\upsilon_{ij}, \mathscr{L}(\upsilon_{ij}; x_0, A, \eta z^{(j)})) : \upsilon_{ij} \in \Upsilon^{(j)}, i \in [n], j \in [k]\},\end{aligned}$$

where $\mathscr{L}$ is the loss corresponding to $\mathfrak{P}$. From here, we describe how to approximate the average loss and the normalized parameter ρ. Specifically, we construct the RBF approximators $\mathscr{L}^\dagger, L^\dagger$ satisfying $\mathscr{L}^\dagger(\upsilon; \Gamma) = L^\dagger(\rho; \Gamma) \approx \bar{L}(\rho; x_0, A, \eta, k)$. Define the multiquadric RBF kernel by

$$\kappa(\upsilon, \upsilon') := \sqrt{1 + \left(\frac{|\upsilon - \upsilon'|}{\varepsilon_{\text{rbf}}}\right)^2}, \qquad \upsilon, \upsilon' > 0,$$

and define the matrix $X \in \mathbb{R}^{kn \times kn}$ by

$$X_{ij} = \kappa(\upsilon_i, \upsilon_j), \quad \upsilon_i, \upsilon_j \in \Gamma.$$

For $\mu_{\text{rbf}} \geq 0$, the coefficients of the RBF approximator are given by $\tilde{w} \in \mathbb{R}^{kn}$ where $\tilde{w}$ solves

$$\mathbf{y} = (X - \mu_{\text{rbf}} I_{kn})\tilde{w}$$

with $I_{kn} \in \mathbb{R}^{kn \times kn}$ being the identity matrix. To evaluate the approximant at a set of points $\xi \in \mathbb{R}^{n_{\text{rbf}}}$, one simply computes

$$\tilde{y} = \mathscr{L}^\dagger(\xi; \Gamma(\mathbf{y}, A)) := \tilde{X}\tilde{w} \quad \text{where} \quad \tilde{X}_{ij} := \kappa(\xi_i, \upsilon_j), i \in [n_{\text{rbf}}], j \in [kn].$$

The optimal parameter choice for the approximator, $\upsilon^\dagger > 0$, is given as

$$\upsilon^\dagger \in \arg\min_{\upsilon>0} \mathscr{L}^\dagger(\upsilon;\Gamma),$$

Finally, the normalized parameter ρ is approximated as $\rho \approx \upsilon/\upsilon^\dagger$ and the average loss thus approximated by

$$L^\dagger(\rho;\Gamma(x_0,A,\eta\mathbf{z})) := \mathscr{L}(\rho\upsilon^\dagger;\Gamma(x_0,A,\eta\mathbf{z})).$$

A.2.3 Interpolation parameter settings

This subsection includes the RBF parameter settings used throughout the numerics in § 3.5 and § 3.5. Note that any settings not presented in this section are either unavailable or already included in-line in the relevant section.

The RBF interpolation parameter settings for each of the approximations of the average loss pertaining to (LS_τ) PS numerics appearing in § 4.5.1 can be found in Table A.1. For those pertaining to (QP_λ), appearing in Figure 4.3 and 4.7–4.9, see Table A.2. For those pertaining to (BP_σ), appearing in § 4.5.3, see Table A.3. The RBF interpolation parameter settings for each of the average loss approximations in § 4.5.5 are given in Table A.4. For those pertaining to Figure 4.13–4.15 of § 4.5.5, see Table A.5.

program	ϵ_{rbf}	μ_{rbf}
(LS_τ)	10^{-5}	0.1
(QP_λ)	$3\cdot10^{-2}$	0.5
(BP_σ)	$3\cdot10^{-2}$	0.5

Table A.1: Average loss interpolation parameter settings for (LS_τ) PS numerics in § 4.5.1. $(s,N,m,\eta) = (1,10^5,2500,2\cdot10^{-3})$; $(n_{\mathrm{rbf}},\mathrm{function}) = (301,\mathrm{multiquadric})$.

s	N	m	η	program	ϵ_{rbf}	μ_{rbf}
1	10000	2500	10^{-5}	(LS_τ)	0.005	1
1	10000	2500	10^{-5}	(QP_λ)	0.05	1
1	10000	2500	10^{-5}	(BP_σ)	0.04	0.9
1	10000	4500	10^{-5}	(LS_τ)	0.005	1
1	10000	4500	10^{-5}	(QP_λ)	0.05	1
1	10000	4500	10^{-5}	(BP_σ)	0.04	0.9
750	10000	4500	0.1	(LS_τ)	0.005	1
750	10000	4500	0.1	(QP_λ)	0.05	1
750	10000	4500	0.1	(BP_σ)	0.05	0.5
100	10000	2500	100	(LS_τ)	0.005	1
100	10000	2500	100	(QP_λ)	0.05	1
100	10000	2500	100	(BP_σ)	0.05	0.5
100	10000	4500	100	(LS_τ)	0.005	1
100	10000	4500	100	(QP_λ)	0.05	1
100	10000	4500	100	(BP_σ)	0.05	0.5

Table A.2: Average loss interpolation parameter settings for (QP_λ) PS numerics in Figure 4.3 and Figure 4.7, 4.8 and 4.9. $n_{\text{rbf}} = 501$, function = multiquadric.

N	m	η	δ	program	ϵ_{rbf}	μ_{rbf}	n_{rbf}
4000	400	1	0.1	(BP_σ)	0.05	1	301
4000	400	1	0.1	(LS_τ)	0.001	1	301
4000	400	1	0.1	(QP_λ)	0.05	1	301
4000	1000	1	0.25	(BP_σ)	0.05	1	301
4000	1000	1	0.25	(LS_τ)	0.001	1	301
4000	1000	1	0.25	(QP_λ)	0.05	1	301
4000	1800	1	0.45	(BP_σ)	0.05	1	301
4000	1800	1	0.45	(LS_τ)	0.001	1	301
4000	1800	1	0.45	(QP_λ)	0.05	1	301
7000	700	1	0.1	(BP_σ)	0.05	1	301
7000	700	1	0.1	(LS_τ)	0.001	1	301
7000	700	1	0.1	(QP_λ)	0.05	1	301
7000	1750	1	0.25	(BP_σ)	0.05	1	301
7000	1750	1	0.25	(LS_τ)	0.001	1	301
7000	1750	1	0.25	(QP_λ)	0.05	1	301
7000	700	100	0.1	(LS_τ)	0.05	1	501
7000	700	100	0.1	(QP_λ)	0.05	1	501
7000	700	100	0.1	(BP_σ)	0.05	1	501
7000	1750	100	0.25	(LS_τ)	0.05	1	501
7000	1750	100	0.25	(QP_λ)	0.05	1	501
7000	1750	100	0.25	(BP_σ)	0.05	1	501

Table A.3: The RBF interpolation parameter settings for each of the approximations of the average loss in § 4.5.3, including for Figure 4.4. $s = 1$, function = multiquadric.

program	ε_{rbf}	μ_{rbf}	n_{rbf}
(LS_τ)	0.01	1	501
(QP_λ)	0.01	1	501
(BP_σ)	0.01	1	501

Table A.4: The RBF interpolation parameter settings for each of the average loss approximations in § 4.5.5. $(s, N, m, \eta) = (10, 4096, 1843, 50)$; function = multiquadric.

η	program	ε_{rbf}	μ_{rbf}	n_{rbf}
0.01	(LS_τ)	0.001	1	501
0.01	(QP_λ)	0.05	1	501
0.01	(BP_σ)	0.05	1	501
0.5	(LS_τ)	0.05	1	501
0.5	(QP_λ)	0.05	1	501
0.5	(BP_σ)	0.05	1	501

Table A.5: The RBF interpolation parameter settings for each of the RBF interpolations in Figure 4.13–4.15 of § 4.5.5. $(s, N, m) = (416, 6418, 2888)$; function = multiquadric.