Compressed sensing, sparse approximation, and low-rank matrix estimation

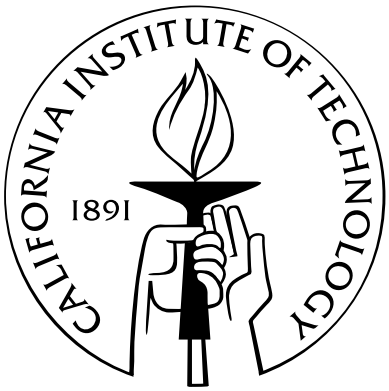
Thesis by

Yaniv Plan

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2011

(Defended February 1, 2011)

Dedicated to my wife, Jasmin

Abstract

The importance of sparse signal structures has been recognized in a plethora of applications ranging from medical imaging to group disease testing to radar technology. It has been shown in practice that various signals of interest may be (approximately) sparsely modeled, and that sparse modeling is often beneficial, or even indispensable to signal recovery. Alongside an increase in applications, a rich theory of sparse and compressible signal recovery has recently been developed under the names compressed sensing (CS) and sparse approximation (SA). This revolutionary research has demonstrated that many signals can be recovered from severely undersampled measurements by taking advantage of their inherent low-dimensional structure. More recently, an offshoot of CS and SA has been a focus of research on other low-dimensional signal structures such as matrices of low rank. Low-rank matrix recovery (LRMR) is demonstrating a rapidly growing array of important applications such as quantum state tomography, triangulation from incomplete distance measurements, recommender systems (e.g., the Netflix problem), and system identification and control.

In this dissertation, we examine CS, SA, and LRMR from a theoretical perspective. We consider a variety of different measurement and signal models, both random and deterministic, and mainly ask two questions.

How many measurements are necessary? How large is the recovery error?

We give theoretical lower bounds for both of these questions, including oracle and minimax lower bounds for the error. However, the main emphasis of the thesis is to demonstrate the efficacy of convex optimization—in particular ℓ_1 and nuclear-norm minimization based programs—in CS, SA, and LRMR. We derive upper bounds for the number of measurements required and the error derived by convex optimization, which in many cases match the lower bounds up to constant or logarithmic factors. The majority of these results do not require the restricted isometry property (RIP), a ubiquitous condition in the literature.

Acknowledgements

I am very grateful for the intelligent way in which my advisor, Emmanuel Candès, stoked my interest first in the classes that I took from him and later through years of research. His intuitive view of the material and his emphasis on tackling important (and difficult) problems was ideal for me. His ability to identify such research problems and ask substantial questions about them made it much easier for me to produce as a graduate student. I also appreciate the many mathematical concepts he has taught me, and exposed me to, and the resources he has shown me to help me learn material on my own.

Aside from my advisor, I would like to thank several fellow graduate students and postdocs for their help and support, both in working through math problems and also for giving me advice on my writing. In particular, thanks to Stephen Becker, Alex Gittens, Ewout Vandenberg, Deanna Needell, and Mark Davenport.

Throughout my stay at Caltech, Sheila Schull and Sydney Garstang did a wonderful job in providing information and help. It is amazing how much easier daily work is with them around.

I would like to thank Joel Tropp, Houman Owhadi, and Babak Hassibi for attending my thesis defense. Also, thanks to Joel and Houman for the material you have taught me in class—it is quite clear that you care about teaching well, and the benefits of this care are evident in the classroom. In fact, as I hope to lecture classes myself in the near future, I plan to incorporate some of the teaching techniques that I learned from watching you two (and from Emmanuel as well).

Finally, I would like to thank my family for the many implicit roles they have played. Thanks to my parents for never pushing me and always encouraging my mathematical development. I believe this is why I am self-motivated to continue work in applied math. Thanks to my brother for all of the useful advice. And of course, thanks to my wife Jasmin for always being there, supporting my work, moving up and down California with me, and energizing me.

Contents

Abstract **iv**

Acknowledgements **v**

List of Tables **x**

List of Figures **xi**

1 Introduction **1**

1.0.1 Compressed sensing 1
1.0.2 Sparse approximation 2
1.0.3 Low-rank matrix recovery 2
1.0.4 Peek at the results 3
1.0.5 The restricted isometry property 3
1.0.6 Organization 5

2 A general model for CS **7**

2.1 Introduction 7
2.1.1 A RIP-less theory? 7
2.1.2 A general theory 8
2.1.3 Examples of incoherent measurements 11
2.1.4 Matrix notation 13
2.1.5 Incoherent sampling theorem 13
2.1.6 Main results 14
2.1.7 Our contribution 16
2.1.8 Organization of the chapter 17
2.1.9 Notation 18
2.2 Background CS literature 18
2.2.1 Asymptotic results and phase transitions 18
2.2.2 Nonasymptotic results and the RIP 19

2.2.3 Null space conditions . . . 21
2.2.4 Other algorithms for CS . . . 23
2.3 Fundamental Estimates . . . 23
2.3.1 Local isometry . . . 23
2.3.2 Off-support incoherence . . . 25
2.3.3 Weak RIP . . . 27
2.3.4 Implications . . . 28
2.4 Noiseless and Sparse Recovery . . . 28
2.4.1 Dual certificates . . . 28
2.4.2 Proof of Lemma 2.4.3 . . . 30
2.5 General Signal Recovery from Noisy Data . . . 32
2.5.1 Proof of Theorem 2.1.2 . . . 33
2.5.2 Proof of Lemma 2.5.2 . . . 36
2.5.3 Proof of Theorem 2.1.3 . . . 39
2.6 Proof of Theorem 2.3.7 (the weak RIP) . . . 40
2.6.1 Proof of Lemma 2.6.3 . . . 43
2.6.2 Fine scale: $k \geq k_1$ . . . 45
2.6.3 Coarse scale: $k \leq 0$ . . . 46
2.6.4 Concentration around the mean . . . 49
2.7 Stochastic Incoherence . . . 49
2.8 Discussion . . . 52

3 Sparse approximation and model selection **53**
3.1 Introduction . . . 53
3.1.1 The coherence property . . . 54
3.1.2 Background literature . . . 55
3.1.3 Sparse model selection . . . 59
3.1.4 Exact model recovery . . . 62
3.1.5 General model selection . . . 63
3.1.6 Implications for signal estimation . . . 67
3.1.7 Organization of the chapter . . . 69
3.2 Optimality . . . 69
3.2.1 For almost all sparse models . . . 69
3.2.2 For sufficiently incoherent matrices . . . 72
3.3 Proofs . . . 74
3.3.1 Preliminaries . . . 74

3.3.2 Proof of Theorem 3.1.4 . . . 75
3.3.3 Norms of random submatrices . . . 78
3.3.4 Proof of Theorem 3.1.6 . . . 81
3.3.5 Proof of Theorem 3.1.5 . . . 83
3.3.6 Proof of (3.3.23) . . . 85
3.4 Discussion . . . 88
3.4.1 Comparison to related theoretical results . . . 88

4 Low-rank matrix estimation with the RIP **90**
4.1 Introduction . . . 90
4.1.1 A few applications . . . 91
4.1.2 Related literature . . . 92
4.1.3 Problem setup . . . 94
4.1.4 Algorithms . . . 95
4.1.5 Organization . . . 96
4.1.6 Notation . . . 97
4.2 Main Results . . . 97
4.2.1 Matrix RIP . . . 97
4.2.2 The matrix Dantzig selector and the matrix LASSO are nearly minimax . . . 99
4.2.3 Oracle inequalities . . . 101
4.2.4 Extension to full-rank matrices . . . 105
4.3 Proofs . . . 106
4.3.1 Proof of Lemma 4.1.1 . . . 107
4.3.2 Proof of Theorem 4.2.3 . . . 107
4.3.3 Proof of Theorem 4.2.4 . . . 110
4.3.4 Proof of Theorem 4.2.4 . . . 113
4.3.5 Proof of Theorem 4.2.7 . . . 113
4.3.6 Proof of Theorem 4.2.8 . . . 115
4.3.7 Extension of proofs to the solution to the LASSO (4.1.5) . . . 122
4.3.8 Proof of Theorem 4.2.5 . . . 123
4.3.9 Proof of Theorem 4.2.6 . . . 125
4.4 Discussion . . . 126

5 Matrix completion with noise **127**
5.1 Introduction . . . 127
5.2 Exact Matrix Completion . . . 129
5.2.1 Geometry and dual certificates . . . 134

5.3 Stable Matrix Completion 137
5.3.1 Proof of Theorem 5.3.1 139
5.3.2 Comparison with an oracle 141
5.4 Numerical Experiments 142
5.5 Discussion 146

6 Conclusion **147**

Bibliography **149**

List of Tables

5.1 RMS error ($\|\hat{M} - M\|_F/n$) as a function of n when subsampling 20% of an $n \times n$ matrix of rank two. Each RMS error is averaged over 20 experiments. 143

List of Figures

3.1 The vector $X\beta_0$ is the projection of $X\beta$ on an ideally selected subset of covariates. These covariates span a plane of optimal dimension which, among all planes spanned by subsets of the same dimension, is closest to $X\beta$. 65

3.2 Sparse signal recovery with the LASSO. (a) Values of the estimated coefficients. All the spike coefficients are obtained by soft-thresholding y and are nonzero. (b) LASSO signal estimate; $X\hat{\beta}$ is just a shifted version of the noisy signal. 71

5.1 Comparison between the recovery error, the oracle error times 1.68, and the estimated oracle error times 1.68. Each point on the plot corresponds to an average over 20 trials. Top left: in this experiment $n = 600, r = 2$, and p varies. The x-axis is the number of measurements per degree of freedom (df). Top right: n varies, whereas $r = 2$, $p = .2$. Bottom: $n = 600$, r varies, and $p = .2$. 145

Chapter 1

Introduction

An image of blood vessels in the body, which may be captured through MRI, has an abundance of structure; in particular, it is sparse in space, and its spatial finite differences are even sparser. Can this structure be utilized to improve MR imaging? Resoundingly, yes! In fact, by taking into account sparsity, MR imaging can be (and has been) sped up by a factor of 7, as demonstrated through a double blind study [159]. The improvement applies outside of just angiography, to MR images in general, which tend to be sparse in the appropriate basis. Moreover, the benefits can be life altering in cases when a slow MR scan is not feasible [71]. The research that has lead to these improvements is called compressed sensing and is closely related to sparse approximation. CS was born seven years ago, pioneered by the papers [39, 42, 55], and it is still being intensely researched today.

1.0.1 Compressed sensing

In CS, a signal $x \in \mathbb{C}^n$ is modeled as a superposition of a small number of elements from a given dictionary, $\Phi \in \mathbb{C}^{n \times k}$. In other words, $x = \Phi v$, where $v \in \mathbb{C}^k$ is sparse (it has few nonzero elements). The basic goal is to recover an approximation of x from linear measurements corrupted by noise

$$y = Ax + z \tag{1.0.1}$$

where $A \in \mathbb{C}^{m \times n}$ is a measurement matrix (often constructed by the scientist), and z is a noise term. For example, in MRI Φ may be a wavelet dictionary, and A can be modeled as a subsampling of the rows of a descrete fourier transform (DFT).

An interesting point about the theory of CS is that it generally requires random measurements (e.g., A could be a *random* subsampling of the DFT). Not only is this assumption crucial to the derivation of many strong theoretical results, but also random measurements seem to give better results in practice and are sought out in real applications.

Before continuing, for simplicity of this thesis we will absorb Φ into the definition of A (or take

it to be the identity), so that x itself should be sparse.[1] With this simplification CS is a special case of a more general class of problems called SA.

1.0.2 Sparse approximation

Once again in SA, we would like to recover a sparse vector x from linear measurements $y = Ax + z$, but there is still one subtle difference between CS and SA. Whereas in CS A is a measurement matrix which in most applications may be constructed (randomly) by the scientist (e.g., in MRI, one chooses which Fourier coefficients to sample), in SA in general there is no expectation that the structure of A may be affected by the scientist. A is often a deterministic, unalterable matrix, and this leads to different theoretical aspects of the problem and a quite different analysis. As an example of SA, in statistical model selection, A is the design matrix, filled with response variables (e.g., answers to a survey). Sparsity plays an important role because often only a few regressors (columns of A) are significant. (Statisticians will be more familiar with the notation $y = X\beta + z$, $X \in \mathbb{R}^{n \times p}$.)

While MRI and the statistical linear model are important examples, sparse signal structures are ubiquitous throughout science and engineering (and we will point to several more examples throughout the thesis). They often arise from the parsimonious nature of the signal—often the underlying structure depends on just a few parameters. In some cases this leads to sparse signals, or, in another vein, this may lead to matrix-valued signals with low rank; this is the second signal structure considered in this thesis.

1.0.3 Low-rank matrix recovery

As a quintessential example of LRMR, consider the semi-famous Netflix problem, in which one has available several entries of the Netflix movie-rating matrix. This matrix contains at position i, j the rating (or estimated rating) of user i for movie j. The goal is to fill in the missing entries—in particular, Netflix would like to be able to predict user ratings for unrated movies. Now, it turns out that the Netflix matrix is (approximately) low-rank; the theoretical justification for this phenomena is that there are only a few important factors which effect peoples' movie ratings. For example, a few ostensible factors would be the user's predilection towards drama, comedy, and violence. However, due to the general methods used for LRMR (see Chapter 4), it is completely unnecessary to know exactly which factors contribute—these are discovered along with the missing entries.

The Netflix problem is an example of a subclass of LRMR problems called *matrix completion*. However, LRMR, has many applications (see Chapters 4 and 5), which follow from a more general

[1] This is fairly innocuous for unitary transforms, since they are isometries in the ℓ_2 norm. For a treatment of CS with non-unitary dictionaries, see [31].

model:

$$y = \mathcal{A}(M) + z$$

where y is a vector of noisy measurements, $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to m$ is a linear measurement operator, z is a noise term, and $M \in \mathbb{R}^{n_1 \times n_2}$ is a matrix with low rank. The goal is to recover M.

1.0.4 Peek at the results

In this thesis we study the effectiveness of convex optimization to recover sparse vectors and low-rank matrices from noisy, linear measurements. In fact, beyond estimating x from equation (1.0.1), we also consider the accurate recovery of Ax and the support of x. The theory will show that one may take far fewer samples than the ambient dimension of the signal; instead the number of samples must match (approximately) the number of dimensions of the manifold that the signal resides in. For example, vectors of length n with sparsity level exactly s lie in an s dimensional manifold; in Chapters 2 and 3 we show that in many cases on the order of $s \log n$ measurements are sufficient for stable recovery by ℓ_1-minimization-based programs. Similarly, rank(r), $n \times n$ matrices lie in a manifold with dimension $2nr - r^2$; in Chapters 4 and 5 we show that they may be stably recovered by nuclear-norm-minimization-based programs from approximately nr, or $nr \log^2 n$, measurements (depending on the measurement model).

While the manifolds that our signals lie in are quite nonlinear, they contain many linear subspaces with (approximately) the same dimension as the original manifolds. In Chapters 3 and 4, this is used to develop lower bounds on the error achievable by any recovery method. We consider an oracle which gives away the smaller linear subspace that the signal resides in; from this point it is easy to analyze the error achieved by least squares regression (which is minimax). As is well known, in the case of Gaussian noise, this leads to an error proportional to the dimension of the linear subspace, and thus proportional to the dimension of the underlying manifold. Interestingly, we give upper bounds for the error achieved by convex optimization, which nearly match these lower bounds. In other words, by taking into account the parsimony of the model, the error in estimation is not proportional to the entire noise vector, but rather to the norm of the noise vector projected onto a much smaller subspace.

1.0.5 The restricted isometry property

A quite prevalent way to prove results about the efficacy of ℓ_1 minimization, nuclear-norm minimization, and a large array of other recovery techniques is the use of the RIP. From here on, we call a vector s-sparse if it has at most s nonzero entries.

Definition 1.0.1 (Restricted isometry property) *We say that an $m \times n$ matrix A obeys the*

RIP with parameters s and δ if

$$(1-\delta)\|v\|_{\ell_2}^2 \leq \|Av\|_{\ell_2}^2 \leq (1+\delta)\|v\|_{\ell_2}^2 \tag{1.0.2}$$

for all s-sparse vectors v.

In other words, A should be well conditioned when acting on signals of interest. When the RIP holds with parameters $2s$ and $\delta < \sqrt{2}-1$ [28] or even $\delta \leq 0.453\ldots$ [78], it is known that certain convex optimization programs are stable. In particular the LASSO and Dantzig selector, both ℓ_1-minimization-based convex programs introduced in Chapter 2, accurately recover all signals x with at most s nonzero elements.

An analogous version of the RIP holds for LRMR (see Chapter 4); in this case one asks the measurement operator $\mathcal{A}$ to be well conditioned when acting on low-rank matrices. As shown in Chapter 4, once again this demonstrates stability of convex optimization.

However, there are a number of limitations to RIP-based theory: 1) testing for the RIP is generally an intractable, combinatorial problem; 2) the only deterministic measurement ensembles which are known to satisfy the RIP do so only under extremely strong conditions; 3) the random measurement ensembles that satisfy the RIP (with high probability) are lacking in many applications; 4) the RIP provides uniform guarantees over *all* low-dimensional signals of interest, and thus theory with the RIP is *necessarily* limited by worst-case signals (and similarly for other conditions which provide uniform guarantees such as the RIP-1 [81] and restricted strong convexity [116]). However, numerical experiments [58] and theoretical results (some of which are contained in this thesis) demonstrate that in many cases *typical* signals may be accurately recovered far past the point when worst-case signals are unrecoverable.

In fact, proving 'RIP-less' results is a delicate matter. To see the difficulty, note that to have universal results, i.e., results that hold for all sparse x (or low-rank M), simultaneously, the lower bound of the RIP would be a necessary condition for stability. To illustrate the point, take the case of SA, and suppose that an oracle gives the exact support, T, of the signal x. Then, one would need the pseudo inverse of A_T (A restricted to the columns in T) to be bounded. In other words, the minimum singular value of A_T should be away from zero.

Thus, the RIP-less results in Chapters 2, 3, 5 are not universal, but rather they must take into account the structure of a typical signal. We do this in a variety of ways.

- **Fix the signal independent of the measurement ensemble (see Chapters 2 and 5):** This method is incontrovertible in many CS setups where the measurement ensemble, A, may often be constructed randomly by the scientist. One expects that the measurement ensemble has no dependence on the signal x, and so it is quite innocuous to fix x, rather than proving results about worst-case signals that may be chosen dependent on the random matrix A.

- **Adopt a statistical model for the signal (see Chapter 3):** Adopting a statistical model is a straightforward way to avoid worst-case signals, and can be used to prove results (with high probability) about general signals. This may be interpreted as proving results for *most* signals of interest.

- **Assume extra signal structure (see Chapter 5):** Here, we assume that the signal x belongs to a certain (large) subset of its inherent low-dimensional space, and prove results given this assumption. (These assumptions are also called *incoherence* assumptions in matrix completion.)

1.0.6 Organization

Each of the chapters, as described below, is self-contained, including notation. Also, they are all based upon research conducted jointly with my advisor, Emmanuel Candès.

Chapter 2: In this chapter, we introduce a simple and very general theory of CS. In this theory, the sensing mechanism simply selects sensing vectors independently at random from a probability distribution F; it includes all models—e.g. Gaussian, frequency measurements—discussed in the literature, but also provides a framework for new measurement strategies as well. We prove that if the probability distribution F obeys a simple incoherence property and an isotropy property, one can faithfully recover approximately sparse signals from a minimal number of noisy measurements. The novelty is that these recovery results do not require the restricted isometry property (RIP)—they make use of a much weaker notion—or a random model for the signal. As an example, in this chapter we show that a signal with s nonzero entries can be faithfully recovered from about $s \log n$ Fourier coefficients that are contaminated with noise.

Chapter 3: In this chapter, we turn to the sparse approximation problem which applies in particular to model selection; thus we switch to the standard statistics notation. We first consider the fundamental problem of estimating the mean vector, $X\beta$, from the data $y = X\beta + z$. X is an $n \times p$ design matrix in which one can have far more variables than observations and z is z is a mean-zero, stochastic error term—the so-called '$p > n$' setup. When β is sparse, or more generally, when there is a sparse subset of covariates providing a close approximation to the unknown mean vector, we ask whether or not it is possible to accurately estimate $X\beta$ using convex optimization.

We show that in a surprisingly wide range of situations, the LASSO happens to nearly select the best subset of variables. In fact, if all of the nonzero entries of β stand above the noise, we show that the support of β is recovered exactly. Quantitatively speaking, we prove that solving a simple, ℓ_1-minimization-based, quadratic program achieves a squared error within a

logarithmic factor of the ideal mean squared error one would achieve with an oracle supplying perfect information about which variables should be included in the model and which variables should not. Interestingly, our results describe the average performance of the LASSO; that is, the performance one can expect in an vast majority of cases where $X\beta$ is a sparse or nearly sparse superposition of variables, but not in all cases.

These results are widely applicable since they simply require that pairs of predictor variables are not too collinear.

Chapter 4: This chapter presents several novel theoretical results regarding the recovery of a low-rank matrix from just a few measurements consisting of linear combinations of the matrix entries. We show that properly constrained nuclear-norm minimization stably recovers a low-rank matrix from a constant number of noisy measurements per degree of freedom. Further, with high probability the recovery error from noisy data is within a constant of three targets: 1) the minimax risk, 2) an oracle error that would be available if the column space of the matrix were known, and 3) a more adaptive oracle error which would be available with the knowledge of the column space corresponding to the part of the matrix that stands above the noise. Lastly, the error bounds regarding low-rank matrices are extended to provide an error bound when the matrix has full rank with decaying singular values. The analysis in this chapter is based on the restricted isometry property.

Chapter 5: This chapter turns to the RIP-less matrix completion problem. We first survey the novel literature on matrix completion, which shows that under some suitable conditions, one can recover an unknown low-rank matrix from a nearly minimal set of entries by nuclear-norm minimization subject to data constraints. Further, this chapter introduces novel results showing that matrix completion is provably accurate even when the few observed entries are corrupted with a small amount of noise. A typical result is that one can recover an unknown $n \times n$ matrix of low rank r from just about $nr \log^2 n$ noisy samples with an error which is proportional to the noise level. We present numerical results which complement our quantitative analysis and show that, in practice, nuclear-norm minimization accurately fills in the many missing entries of large low-rank matrices from just a few noisy samples. Some analogies between matrix completion and compressed sensing are discussed throughout.

Chapter 6: In this chapter, we give a brief summary of the results discussed in the earlier chapters, and discuss the related open problems still left to be researched.

Chapter 2

A general model for CS

2.1 Introduction

This chapter develops a novel, simple, general, and 'RIP-less' theory of CS [39, 42, 55]. We begin by motivating and stating the results, and in turn give a discussion of related literature including a discussion of the *restricted isometry property* (RIP) in Section 2.1.7.

2.1.1 A RIP-less theory?

The early paper [39] triggered a massive amount of research by showing that it is possible to sample signals at a rate proportional to their information content rather than their bandwidth. For instance, in a discrete setting, this theory asserts that a digital signal $x \in \mathbb{R}^n$ (which can be viewed as Nyquist samples of a continuous-time signal over a time window of interest) can be recovered from a small random sample of its Fourier coefficients provided that x is sufficiently sparse. Formally, suppose that our signal x has at most s nonzero amplitudes at completely unknown locations and that we are given the value of its discrete Fourier transform (DFT) at m frequencies selected uniformly at random (we think of m as being much smaller than n). Then [39] showed that one can recover x by solving an optimization problem which simply finds, among all candidate signals, that with the minimum ℓ_1 norm; the number of samples we need must be on the order of $s \log n$. In other words, if we think of s as a measure of the information content, we can sample *nonadaptively* nearly at the information rate without information loss. By swapping time and frequency, this also says that signals occupying a very large bandwidth but with a sparse spectrum can be sampled (at random time locations) at a rate far below the Shannon-Nyquist rate.

Despite considerable progress in the field, some important questions have still been left open. We discuss two that have both a theoretical and practical appeal.

> *Is it possible to faithfully recover a nearly sparse signal $x \in \mathbb{R}^n$, one which is well approximated by its s largest entries, from about $s \log n$ of its Fourier coefficients? Is it still*

possible when these coefficients are further corrupted by noise?

These issues are paramount since in real-world applications, signals are never exactly sparse, and measurements are never perfect either. Now the traditional way of addressing these types of problems in the field is by means of the restricted isometry property (RIP) [41]. The trouble here is that it is unknown whether or not this property holds when the sample size m is on the order of $s \log n$. In fact, answering this one way or the other is generally regarded as extremely difficult, and so the restricted isometry machinery does not directly apply in this setting.

In this chapter, we prove that the two questions formulated above have positive answers. In fact, we introduce recovery results which are—up to a logarithmic factor—as good as those one would get if the restricted isometry property were known to be true. To fix ideas, suppose we observe m noisy discrete Fourier coefficients about an s-sparse signal x,

$$\tilde{y}_k = \sum_{t=0}^{n-1} e^{-\imath 2\pi \omega_k t} x[t] + \sigma z_k, \quad k = 1, \dots, m. \tag{2.1.1}$$

Here, the frequencies ω_k are chosen uniformly at random in $\{0, 1/n, 2/n, \dots, (n-1)/n\}$ and z_k is white noise with unit variance. Then if the number of samples m is on the order of $s \log n$, it is possible to get an estimate $\hat{x}$ obeying

$$\|\hat{x} - x\|_{\ell_2}^2 = \text{polylog}(n) \, \frac{s}{m} \sigma^2 \tag{2.1.2}$$

by solving a convex ℓ_1-minimization program. (Note that when the noise vanishes, the recovery is exact.) Up to the logarithmic factor, which may sometimes be on the order of $\log n$ and at most a small power of this quantity, this is optimal. Now if the RIP held, one would get a squared error bounded by $O(\log n) \frac{s}{m} \sigma^2$ [17, 43] and, therefore, the 'RIP-less' theory developed in this chapter roughly enjoys the same performance guarantees.

2.1.2 A general theory

The estimate we have just seen is not isolated and the real purpose of this chapter is to develop a theory of compressive sensing which is both as simple and as general as possible.

At the heart of compressive sensing is the idea that randomness can be used as an effective sensing mechanism. We note that random measurements are not only crucial in the derivation of many theoretical results, but also generally seem to give better empirical results as well. Therefore, we propose a mechanism whereby sensing vectors are independently sampled from a population F. Mathematically, we observe

$$\tilde{y}_k = \langle a_k, x \rangle + \sigma z_k, \quad k = 1, \dots, m, \tag{2.1.3}$$

where $x \in \mathbb{R}^n$, $\{z_k\}$ is a noise sequence, and the sensing vectors $a_k \overset{\text{iid}}{\sim} F$. For example, if F is the

family of complex sinusoids, this is the Fourier sampling model introduced earlier. All we require from F is an isotropy property and an incoherence property.

Isotropy property: We say that F obeys the isotropy property if

$$\mathbb{E}\, aa^* = I, \quad a \sim F. \tag{2.1.4}$$

If F has mean zero (we do not require this), then $\mathbb{E}\, aa^*$ is the covariance matrix of F. In other words, the isotropy condition states that the components of $a \sim F$ have unit variance and are uncorrelated. This assumption may be weakened a little, as we shall see later.

Incoherence property: We may take the coherence parameter $\mu(F)$ to be the smallest number such that with $a = (a[1], \ldots, a[n]) \sim F$,

$$\max_{1 \le t \le n} |a[t]|^2 \le \mu(F) \tag{2.1.5}$$

holds either deterministically or stochastically in the sense discussed below. The smaller $\mu(F)$, i.e. the more incoherent the sensing vectors, the fewer samples we need for accurate recovery. When a simple deterministic bound is not available, one can take the smallest scalar μ obeying

$$\mathbb{E}[n^{-1}\|a\|_{\ell_2}^2 \mathbb{1}_{E^c}] \le \tfrac{1}{20} n^{-3/2} \qquad \text{and} \qquad \mathbb{P}(E^c) \le (nm)^{-1}, \tag{2.1.6}$$

where E is the event $\{\max_{1 \le t \le n} |a[t]|^2 > \mu\}$.

Suppose for instance that the components are i.i.d. $\mathcal{N}(0,1)$. Then a simple calculation we shall not detail shows that

$$\begin{aligned} \mathbb{E}[n^{-1}\|a\|_{\ell_2}^2 \mathbb{1}_{E^c}] &\le 2n\, \mathbb{P}(Z > \sqrt{\mu}) + 2\sqrt{\mu}\phi(\sqrt{\mu}), \\ \mathbb{P}(E^c) &\le 2n\, \mathbb{P}(Z \ge \sqrt{\mu}), \end{aligned} \tag{2.1.7}$$

where Z is standard normal and ϕ is its density function. The inequality $P(Z > t) \le \phi(t)/t$ shows that one can take $\mu(F) \le 6 \log n$ as long as $n \ge 16$ and $m \le n$. More generally, if the components of a are i.i.d. samples from a sub-Gaussian distribution, $\mu(F)$ is at most a constant times $\log n$. If they are i.i.d. from a sub-exponential distribution, $\mu(F)$ is at most a constant times $\log^2 n$. In what follows, however, it might be convenient for the reader to assume that the deterministic bound (2.1.5) holds.

It follows from the isotropy property that $\mathbb{E}\,|a[t]|^2 = 1$, and thus $\mu(F) \ge 1$. This lower bound is achievable by several distributions and one such example is obtained by sampling a row from the

DFT matrix as before, so that

$$a[t] = e^{i2\pi kt/n},$$

where k is chosen uniformly at random in $\{0, 1, \ldots, n-1\}$. Then another simple calculation shows that $\mathbb{E}\, aa^* = I$ and $\mu(F) = 1$ since $|a[t]|^2 = 1$ for all t. At the other extreme, suppose the measurement process reveals one entry of x selected uniformly at random so that $a = \sqrt{n}\, e_i$ where i is uniform in $\{1, \ldots, n\}$; the normalization ensures that $\mathbb{E}\, aa^* = I$. This is a lousy acquisition protocol because one would need to sample on the order of $n \log n$ times to recover even a 1-sparse vector (the logarithmic term comes from the coupon collector effect). Not surprisingly, this distribution is in fact highly coherent as $\mu(F) = n$.

We pause to note that when specializing to subsampled Fourier measurements, this is a slightly different model that what has been considered in most past works [42, 135]. In particular, our model samples rows from a DFT with replacement, allowing the possibility of duplicates, whereas older works have considered sampling without replacement. These models are in fact essentially the same. First, when significantly undersampling, very few rows will be duplicated. Second, in the noiseless case, the only relevant facet of A is its null space; sampling more rows decreases the null space and strictly aids in recovery. In particular, resampling a row provides no new information and does not decrease the null space. In other words, the probability that recovery fails when sampling m rows with replacement is strictly larger than than the probability that it fails when sampling m rows with replacement, i.e., our results extend to the other model. In the noisy case, the models appear to be quite similar, but neither is strictly weaker.

With the assumptions set, we now give a representative result of this chapter: suppose x is an arbitrary but fixed s-sparse vector and that one collects information about this signal by means of the random sensing mechanism (4.1.1), where z is white noise. Then if the number of samples is on the order $\mu(F) s \log n$, one can invoke ℓ_1 minimization to get an estimator $\hat{x}$ obeying

$$\|\hat{x} - x\|_{\ell_2}^2 \leq \text{polylog}(n)\, \frac{s}{m} \sigma^2.$$

This bound is sharp. It is not possible to substantially reduce the number of measurements and get a similar bound, no matter how intractable the recovery method might be. To be precise, as shown in Section 2.1.5 the number of measurements required is sharp modulo a constant. Further, with this many measurements, the upper bound is optimal up to logarithmic factors. Finally, we will see that when the signal is not exactly sparse, we just need to add an approximation error to the upper bound.

To summarize, this chapter proves that one can faithfully recover approximately s-sparse signals from about $s \log n$ random incoherent measurements for which $\mu(F) = O(1)$.

2.1.3 Examples of incoherent measurements

We have seen through examples that sensing vectors with low coherence are global or spread out. Incoherence alone, however, is not a sufficient condition: if F were a constant distribution (sampling from F would always return the same vector), one would not learn anything new about the signal by taking more samples regardless of the level of incoherence. However, as we will see, the incoherence and isotropy properties together guarantee that sparse vectors lie away from the nullspace of the sensing matrix whose rows are the a_k^*'s.

The role of the isotropy condition is to keep the measurement matrix from being rank deficient when sufficiently many measurements are taken (and similarly for subsets of columns of A). Specifically, one would hope to be able to recover *any* signal from an arbitrarily large number of measurements. However, if $\mathbb{E}\, aa^*$ were rank deficient, there would be signals $x \in \mathbb{R}^n$ that would not be recoverable from an arbitrary number of samples; just take $x \neq 0$ in the nullspace of $\mathbb{E}\, aa^*$. The nonnegative random variable $x^* aa^* x$ has vanishing expectation, which implies $a^* x = 0$ almost surely. (Put differently, all of the measurements would be zero almost surely.) In contrast, the isotropy condition implies that $\frac{1}{m}\sum_{k=1}^m a_k a_k^* \to I$ almost surely as $m \to \infty$ and, therefore, with enough measurements, the sensing matrix is well conditioned and has a left-inverse.[1]

We now provide examples of incoherent and isotropic measurements.

- **Sensing vectors with independent components.** Suppose the components of $a \sim F$ are independently distributed with mean zero and unit variance. Then F is isotropic. In addition, if the distribution of each component is light-tailed, then the measurements are clearly incoherent.

  A special case concerns the case where $a \sim N(0, I)$, also known in the field as the *Gaussian measurement ensemble*, which is perhaps the most commonly studied. Here, one can take $\mu(F) = 6 \log n$ as seen before.

  Another special case is the *binary measurement ensemble* where the entries of a are symmetric Bernoulli variables taking on the values ± 1. A shifted version of this distribution is the sensing mechanism underlying the single pixel camera [68].

- **Subsampled orthogonal transforms:** Suppose we have an orthogonal matrix obeying $U^* U = n\, I$. Then consider the sampling mechanism picking rows of U uniformly and independently at random. In the case where U is the DFT, this is the random frequency model introduced earlier. Clearly, this distribution is isotropic and $\mu(F) = \max_{ij} |U_{ij}|^2$. In the case where U is a Hadamard matrix, or a complex Fourier matrix, $\mu(F) = 1$.

[1] One could require 'near isotropy,' i.e., $\mathbb{E}\, aa^* \approx I$. If the approximation were tight enough, our theoretical results would still follow with minimal changes to the proof.

- **Random convolutions:** Consider the circular convolution model $y = Gx$ in which

$$G = \begin{bmatrix} g[0] & g[1] & g[2] & \dots & g[n-1] \\ g[n-1] & g[0] & g[1] & \dots & \\ & & & & \\ & & & & \\ g[1] & & \dots & g[n-1] & g[0] \end{bmatrix}.$$

  Because a convolution is diagonal in the Fourier domain (we just multiply the Fourier components of x with those of g), G is an isometry if the Fourier components of $g = (g[0], \dots, g[n-1])$ have the same magnitude. In this case, sampling a convolution product at randomly selected time locations is an isotropic and incoherent process provided g is spread out ($\mu(F) = \max_t |g(t)|^2$). This example extends to higher dimensions; e.g. to spatial 3D convolutions.

- **Subsampled tight or continuous frames:** We can generalize the example above by subsampling a tight frame or even a continuous frame. An important example might be the Fourier transform with a continuous frequency spectrum. Here,

$$a(t) = e^{\imath 2\pi\omega t},$$

  where ω is chosen uniformly at random in $[0, 1]$ (instead of being on an equispaced lattice as before). This distribution is isotropic and obeys $\mu(F) = 1$. A situation where this arises is in magnetic resonance imaging (MRI) as frequency samples rarely fall on an equispaced Nyquist grid. By swapping time and frequency, this is equivalent to sampling a nearly sparse trigonometric polynomial at randomly selected time points in the unit interval [125].

These examples could of course be multiplied, and we hope we have made clear that our framework is general and encompasses many of the measurement models discussed in compressive sensing—and perhaps many new ones as well.

In some specific cases our theory improves upon what is available in the literature (e.g., for Fourier measurements), but for certain other measurement models (e.g., Gaussian), our theory requires an increase in the number of measurements. In both cases the difference between our theory and the prior literature is the removal or inclusion of logarithmic factors. See Section 2.1.6 for a more detailed discussion.

2.1.4 Matrix notation

Before continuing, we pause to demonstrate exactly how we display this model in the matrix notation of the introduction. Divide both sides of (4.1.1) by $\sqrt{m}$, and rewrite our statistical model as

$$y = Ax + \sigma_m z; \tag{2.1.8}$$

the kth entry of y is $\tilde{y}_k$ divided by $\sqrt{m}$, the kth row of A is a_k^* divided by $\sqrt{m}$, and σ_m is σ divided by $\sqrt{m}$. This normalization implies that the columns of A are approximately unit-normed, and is most used in the compressive sensing literature.

2.1.5 Incoherent sampling theorem

To ease readability, we introduce our results by first presenting a recovery result from noiseless data. The recovered signal is obtained by the standard ℓ_1-minimization program

$$\min_{\bar{x}\in\mathbb{R}^n} \|\bar{x}\|_{\ell_1} \quad \text{subject to} \quad A\bar{x} = y. \tag{2.1.9}$$

(Recall that the rows of A are normalized independent samples from F.)

Theorem 2.1.1 (Noiseless incoherent sampling) *Let x be a fixed but otherwise arbitrary s-sparse vector in $\mathbb{R}^n$. Then with probability at least $1 - 5/n - e^{-\beta}$, x is the unique minimizer to* (2.1.9) *with $y = Ax$ provided that*

$$m \geq C_\beta \cdot \mu(F) \cdot s \cdot \log n.$$

More precisely, C_β may be chosen as $C_0(1+\beta)$ for some positive numerical constant C_0.

Among other things, this theorem states that one can perfectly recover an arbitrary sparse signal from about $s \log n$ convolution samples, or a signal that happens to be sparse in the wavelet domain from about $s \log n$ randomly selected noiselet coefficients. It extends an earlier result [38], which assumed a subsampled orthogonal model, and strengthens it since that reference could only prove the claim for randomly signed vectors x. Here, x is arbitrary, and we do not make any distributional assumption about its support or its sign pattern.

This theorem is also about a fundamental information theoretic limit: the number of samples for perfect recovery has to be on the order of $\mu(F) \cdot s \cdot \log n$, and cannot possibly be much below this number. More precisely, suppose we are given a distribution F with coherence parameter $\mu(F)$. Then there exist s-sparse vectors that cannot be recovered with probability at least $1 - 1/n$, say, from fewer than a constant times $\mu(F) \cdot s \cdot \log n$ samples. When $\mu(F) = 1$, this has been already established since [39] proves that some s sparse signals cannot be recovered from fewer than a constant times $s \cdot \log n$ random DFT samples. Our general claim follows from a modification of the argument in [39].

Assume, without loss of generality, that $\mu(F)$ is an integer and consider the isotropic process that samples rows from an $n \times n$ block diagonal matrix, each block being a DFT of a smaller size; that is, of size n/ℓ where $\mu(F) = \ell$. Then if $m \leq c_0 \cdot \mu(F) \cdot s \cdot \log n$, one can construct s-sparse signals just as in [39] for which $Ax = 0$ with probability at least $1/n$. We omit the details.

The important aspect, here, is the role played by the coherence parameter $\mu(F)$. In general, the minimal number of samples must be on the order of the coherence times the sparsity level s times a logarithmic factor. Put differently, *the coherence completely determines the minimal sampling rate.*

2.1.6 Main results

We assume for simplicity that we are undersampling so that $m \leq n$. Our general result deals with 1) arbitrary signals which are not necessarily sparse (images are never exactly sparse even in a transformed domain) and 2) noise. To recover x from the data y and the model (2.1.8), we consider the unconstrained LASSO [147] which solves the ℓ_1 regularized least-squares problem

$$\min_{\bar{x} \in \mathbb{R}^n} \ \tfrac{1}{2} \|A\bar{x} - y\|_{\ell_2}^2 + \lambda \sigma_m \|\bar{x}\|_{\ell_1}. \tag{2.1.10}$$

We assume that z is Gaussian $z \sim N(0, I)$. However, the theorem below may be adapted to any noise model that obeys $\|A^* z\|_{\ell_\infty} \leq C\sqrt{\log n}$ with high probability (for a fixed constant C). Thus many other noise models would work as well. In what follows, x_s is the best s-sparse approximation of x or, equivalently, a vector consisting of the s largest entries of x in magnitude. Ties may be resolved in any arbitrary way.

Theorem 2.1.2 *Let x be an arbitrary fixed vector in $\mathbb{R}^n$. Then with probability at least $1 - 6/n - 6e^{-\beta}$ the solution to (2.1.10) with $\lambda = 10\sqrt{\log n}$ obeys*

$$\|\hat{x} - x\|_{\ell_2} \leq \min_{1 \leq s \leq \bar{s}} C(1+\alpha) \left[\frac{\|x - x_s\|_{\ell_1}}{\sqrt{s}} + \sigma \sqrt{\frac{s \log n}{m}} \right] \tag{2.1.11}$$

provided that $m \geq C_\beta \cdot \mu(F) \cdot \bar{s} \cdot \log n$. If one measures the error in the ℓ_1 norm, then

$$\|\hat{x} - x\|_{\ell_1} \leq \min_{1 \leq s \leq \bar{s}} C(1+\alpha) \left[\|x - x_s\|_{\ell_1} + s\sigma \sqrt{\frac{\log n}{m}} \right]. \tag{2.1.12}$$

Above, C is a numerical constant, C_β can be chosen as before, and $\alpha = \sqrt{\frac{(1+\beta) s\mu \log n \log m \log^2(s\mu)}{m}}$ which is never greater than $\log^{3/2} n$ in this setup.

These robust error bounds do not require either (1) a random model on the signal or (2) the RIP nor one of a few closely related strong conditions such as the RIP-1 [81], the restricted eigenvalue assumption [17], or the compatibility condition [158]. The conditions are weak enough that they do

not necessarily imply uniform sparse-signal recovery, but instead they imply recovery of an arbitrary *fixed* sparse signal with high probability. Further, the error bound is within at most a $\log^{3/2} n$ factor of what has been established using the RIP since a variation on the arguments in [43] would give an error bound proportional to the quantity inside the square brackets in (2.1.11). As a consequence, the error bound is within a polylogarithmic factor of what is achievable with the help of an oracle that would reveal the locations of the significant coordinates of the unknown signal [43]. In other words, it cannot be substantially improved.

Because much of the compressive sensing literature works with restricted isometry conditions—we shall discuss exceptions such as [14,62] in Section 2.1.7—we pause here to discuss these conditions and to compare them to our own. As mentioned in Chapter 1, we say that an $m \times n$ matrix A obeys the RIP with parameters s and δ if

$$(1-\delta)\|v\|_{\ell_2}^2 \leq \|Av\|_{\ell_2}^2 \leq (1+\delta)\|v\|_{\ell_2}^2 \tag{2.1.13}$$

for all s-sparse vectors v. In other words, all the submatrices of A with at most s columns are well conditioned. When the RIP holds with parameters $2s$ and $\delta < 0.414\ldots$ [28] or even $\delta \leq 0.453\ldots$ [78], it is known that the error bound (2.1.11) holds (without the factor $(1+\alpha)$). This δ is sometimes referred to as the restricted isometry constant.

Bounds on the restricted isometry constant have been established in [42] and in [135] for partial DFT matrices, and by extension, for partial subsampled orthogonal transforms. For instance, [135] proves that if A is a properly normalized partial DFT matrix, then the RIP with $\delta = 1/4$ holds with high probability if $m \geq C \cdot s \log n \log m \log^2 s$ (C is some positive constant). We believe the proof extends with hardly any change to show that the measurement ensembles considered in this chapter obey the RIP with high probability when $m \geq C \cdot \mu(F) \cdot s \log n \log m \log^2(s\mu)$. Thus, our result bridges the gap between the region where the RIP holds and the region in which one has the minimum number of measurements needed to prove perfect recovery of exactly sparse signals from noisy data, which is on the order of $\mu(F) \cdot s \log n$. In doing so, we introduce an extra factor α into our error bounds, which does not exist in the RIP-based results. This factor is at most logarithmic ($\alpha < \log^{3/2} n$) and shrinks with the number of measurements; when the RIP is known to hold, the factor α disappears, i.e., $\alpha = O(1)$. With that said, we believe that in the region in which the RIP does not hold, and $\alpha > 1$, this extra factor is an artifact of the proof technique and could be removed by a different theoretical analysis. Last, we note that in certain regimes there are prior RIPless results that give stability guarantees when $m > Cs\mu \log m \log^5(\mu \log m)$ (see Section 3.1.2).

The careful reader will no doubt remark that for very specific models such as the Gaussian measurement ensemble, it is known that on the order $s \log(n/s)$ samples are sufficient for stable recovery while our result asserts that on the order of $s \log^2 n$ are sufficient (and $s \log n$ for the binary

measurement ensemble). This slight loss is a small price to pay for a very simple general theory, which accommodates a wide array of sensing strategies. Having said this, the reader will also verify that specializing our proofs below gives an optimal result for the Gaussian ensemble; i.e. establishes a near-optimal error bound from about $s \log(n/s)$ observations.

Finally, another frequently discussed algorithm for sparse regression is the Dantzig selector [43]. Here, the estimator is given by the solution to the linear program

$$\min_{\bar{x} \in \mathbb{R}^n} \|\bar{x}\|_{\ell_1} \quad \text{subject to} \quad \|A^*(A\bar{x} - y)\|_{\ell_\infty} \leq \lambda \sigma_m. \tag{2.1.14}$$

We show that the Dantzig selector obeys nearly the same error bound.

Theorem 2.1.3 *The Dantzig selector, with $\lambda = 10\sqrt{\log n}$ and everything else the same as in Theorem 2.1.2, obeys*

$$\|\hat{x} - x\|_{\ell_2} \leq \min_{s \leq \bar{s}} C(1 + \alpha^2) \left[\frac{\|x - x_s\|_{\ell_1}}{\sqrt{s}} + \sigma \sqrt{\frac{s \log n}{m}} \right] \tag{2.1.15}$$

$$\|\hat{x} - x\|_{\ell_1} \leq \min_{s \leq \bar{s}} C(1 + \alpha^2) \left[\|x - x_s\|_{\ell_1} + s\sigma \sqrt{\frac{\log n}{m}} \right] \tag{2.1.16}$$

with the same probabilities as before.

The only difference is α^2 instead of α in the right-hand sides.

2.1.7 Our contribution

Due to the plethora of background literature, we reverse the standard order and first describe our contribution before describing many of the important contributions that came before it, in Section 3.1.2.

From the perspective of an engineer or scientist with a problem that may fit in the CS framework, our main contribution is to provide a simple framework which applies to all the standard compressive sensing models and some new ones as well. With that said and as noted above, one could adapt the arguments of [135] to prove the RIP under our general framework, although this RIP-based theory would require about a factor of $\log m \log^2(s\mu)$ more measurements. From a theoretical standpoint, our main contribution is to reduce the minimal number of measurements required in some standard sensing models such as Fourier measurements, or, more generally, sensing matrices obtained by sampling a few rows from an orthogonal matrix. This is interesting theoretically, because in a sense (described above) the number of measurements required has been reduced to the absolute minimum, up to a constant. In fact, the theoretical developments necessary to idealize this number were quite involved, in particular using the majorizing measures theorem. Further, we establish

that the restricted isometry property is not necessarily needed to accurately recover nearly sparse vectors from noisy compressive samples. Thus our work is a significant departure from the majority of the literature, which establishes good noisy recovery properties via the RIP machinery. This literature is, of course, extremely large and we cannot cite all contributions but a partial list would include [9, 10, 17, 26, 40, 42, 43, 51, 56, 94, 124, 135, 166, 167].

The reason why one can get strong error bounds, which are within a polylogarithmic factor of what is available with the aid of an 'oracle', without the RIP is that our results do not imply universality. That is, we are not claiming that if A is randomly sampled and then fixed once for all, then the error bounds from Section 2.1.6 hold for all signals x. What we are saying is that if we are given an arbitrary x, and then collect data by applying our random scheme, then the recovery of *this* x will be accurate. As discussed in Chapter 1, if one wishes to establish universal results holding for *all* x simultaneously, then we would need the RIP or a property very close to it. As a consequence, we cannot possibly be in this setup and guarantee universality since we are not willing to assume that the RIP holds.

To the best of our knowledge, only a few papers have addressed non-universal stability (the literature grows so rapidly that an inadvertent omission is entirely possible). In Chapter 3 we also consider weak conditions that allow stable recovery; in this case the we assume that the signal is sampled according to a random model, but in return the measurement matrix A can be deterministic. In the asymptotic case, stable signal recovery has been demonstrated for the Gaussian measurement ensemble in a regime in which the RIP does not necessarily hold [14,62]; these papers will be discussed more below, and in the asymptotic limit they provide exact answers. This contrasts with our non-asymptotic theory which non-exact, but gives bounds that are tight to within logarithmic factors. Aside from these papers and the work in progress [35], it seems that that the literature regarding stable recovery with conditions weak enough that they do not imply universality is extremely sparse. Finally and to be complete, we would like to mention that earlier works have considered the recovery of perfectly sparse signals from subsampled orthogonal transforms [38], and of sparse trigonometric polynomials from random time samples [125].

2.1.8 Organization of the chapter

The chapter is organized as follows. In Section 3.1.2 we describe many of the important contributions in the literature of CS. In Section 2.3, we introduce several fundamental estimates which our arguments rely upon, but which also could be useful tools for other results in the field. In Section 3, we prove the noiseless recovery result, namely, Theorem 2.1.1. In Section 2.5, we prove our main results, Theorems 2.1.2 and 2.1.3. In Section 2.6, we give the proof of an important technical piece, which we call the weak RIP. Now all these sections assume for simplicity of exposition that the coherence bound holds deterministically (2.1.5). We extend the proof to distributions obeying the

coherence property in the stochastic sense (2.1.6) in Section 2.7. Finally, we conclude the main text with some final comments in Section 2.8.

2.1.9 Notation

We provide a brief summary of the notations used throughout the chapter. For an $m \times n$ matrix A and a subset $T \subset \{1, \ldots, n\}$, A_T denotes the $m \times |T|$ matrix with column indices in T. Also, $A_{\{i\}}$ is the i-th column of A. Likewise, for a vector $v \in \mathbb{R}^n$, v_T is the restriction of v to indices in T. Thus, if v is supported on T, $Av = A_T v_T$. In particular, $a_{k,T}$ is the vector a_k restricted to T. The operator norm of a matrix A is denoted $\|A\|$. The identity matrix, in any dimension, is denoted I. Further, e_i always refers to the i-th standard basis element, e.g., $e_1 = (1, 0, \ldots, 0)$. For a scalar t, $\mathrm{sgn}(t)$ is the sign of t if $t \neq 0$ and is zero otherwise. For a vector x, $\mathrm{sgn}(x)$ applies the sign function componentwise. We shall also use μ as a shorthand for $\mu(F)$ whenever convenient. Throughout, C is a constant whose value may change from instance to instance.

2.2 Background CS literature

While the theory and practice of ℓ_1 minimization to recover sparse signals had been around for quite some time (see Chapter 3, Section 3.1.2), CS emerged with the seminal works by Donoho [55] and Candès et al. [39]. In contrast to the prior theory of ℓ_1 minimization, these works extolled the value of taking random measurements, and showed that near-optimal results could be achieved with such measurements. However, two different paths of work grew from each of these results, focusing on different points of view and with a completely different theoretical analysis. Beyond these two paths, many researchers from various disciplines forged their own beautiful contributions to the theory of CS.

In this section, we discuss some of the various important results in the CS theory. This is by no means a comprehensive survey, as the number of papers on the subject is in the hundreds, and would be infeasible to review. We begin by discussing the pioneering CS papers and the subsequent line of theory described by the authors of those results. We then describe some of the keystone results in CS, focusing on those with relation to the theory in this chapter and we conclude by describing some of the techniques outside of ℓ_1 minimization used to recover sparse signals.

2.2.1 Asymptotic results and phase transitions

The early results addressed the noiseless problem $y = Ax$. Donoho et al. [57,64–66] focused on asymptotic results, and based the theory on polytope geometry and s-neighborliness (or k-neighborliness in the notation of these papers). They demonstrated sharp phase transitions theoretically for Gaussian measurements, and through numerical simulations demonstrated that these phase transitions

appeared to be universal to many measurement schemes (such as Fourier). To be a bit more precise, fix parameters (δ, ρ). Now suppose that we let $s, n, m \to \infty$ with $s/n \to \rho$ and $m/n \to \delta$. Then there is a curve, defined by a specific function $\rho_{CG}(\delta)$ dividing the region in which reconstruction succeeds and reconstruction fails. In particular, let $\hat{x}$ be the ℓ_1 minimization (2.1.9) solution; if $\rho < \rho_{CG}(\delta)$ then with probability converging to 1 in the asymptotic limit, $\hat{x} = x$. If $\rho > \rho_{CG}(\delta)$, the probability that $\hat{x} = x$ tends to zero. In particular, note that there is a linear relationship between the number of measurements needed and the sparsity level of x. Further, Donoho et al. [61] demonstrated that a certain algorithm based on message passing achieves the same phase transitions curve, while offering a large speed up in computational time. In fact, this message passing algorithm was shown to converge to the LASSO solution, a fact that led to important theoretical breakthroughs in the noisy problem.

By analyzing the message passing algorithm, and lifting the results to the LASSO case, Donoho et al. [62] demonstrated an asymptotic phase transition for the noisy problem (once again under the assumption of Gaussian measurements). In fact, the curve, $\rho_{CG}(\delta)$ is exactly as in the noiseless case. Above the curve, the worst-case error is unbounded and below the curve there is an exact expression for the mean squared error. We emphasize that while this is also a noisy RIP-less result, it is clearly of a different nature than the results given in this chapter. In particular, it considers the asymptotic case and restricts to Gaussian measurements, but in return a very precise theory is offered.

A result of a similar nature, which once again used the message passing algorithm as a key part of the analysis and considered Gaussian measurements, was proven by Bayati–Montanari [14]. Here, the authors considered a sequence of problem instances $y^j = A^j x^j + z^j$, and assumed that in the asymptotic limit the empirical distribution of x^j converged weakly to a probability measure (thereby avoiding worst-case signals). Under this assumption, the authors gave an explicit form for the asymptotic error under a family of norms, but with no assumption on the sparsity level. Once again, this is a RIP-less result complementary to the theory described in this chapter.

We also note that a number of other researchers have considered the asymptotics, especially in the case of Gaussian measurements. For example, Wainwright [163] addresses the asymptotics for a family of Gaussian measurement ensembles, not restricting to the i.i.d. case. See also the work of Fletcher et al. [76].

2.2.2 Nonasymptotic results and the RIP

We begin by reviewing some of the relevant works of Candès and co-workers. This nonasymptotic theory began with the paper [39], which, as noted in the introduction, demonstrated that m random noiseless Fourier measurements were sufficient to recover an s-sparse vector by ℓ_1 minimization (2.1.9) as long as $m \gtrsim s \log n$. In [38], Candès and Romberg extended these results to general sub-

sampled orthogonal matrices with small entries, but in this case they required a random model on the signs of x. In [42] Candès and Tao introduced the uniform uncertainty principle, now called the restricted isometry property (RIP), and used it to demonstrate that ℓ_1 minimization could recover *approximately* sparse vectors from subsampled measurements. They considered Fourier measurements, Gaussian measurements, and binary measurements and gave non-asymptotic error bounds under the assumption that the entries of x decay following a power law. Under this assumption, and using the theory of Gelfand widths, they showed that their results were near optimal, up to constant or logarithmic factors. (See [51] for an explanation of Gelfand widths tailored to CS and see [82, 83] for the relevant theory on Gelfand widths.) Along the way, they proved the RIP for these ensembles, although their requirements were refined in later papers. Numerical results supporting the theory were described in [27], demonstrating that in practice, $3s - 5s$ measurements were necessary for accurate signal recovery by ℓ_1 minimization. In [40] the authors considered the noisy problem, and demonstrated that the ℓ_2 norm of the error in recovery when solving the constrained LASSO was within a constant of size of ℓ_2 norm of the noise—this required the RIP. In [43] the authors demonstrated that a different convex program, called the Danzig selector, in fact achieved a stronger error bound in the case of Gaussian noise: they showed that the error in recovery is proportional to the sparsity of the signal. In other words it was nearly as if one were able to project onto the low-dimensional space spanned by the non-zero coefficients of x (this is the type of error bound given in this chapter).

Beginning with this line of work, much of the theory of CS concentrated on RIP conditions. We pause to note that although the RIP was introduced to the CS community by Candès–Tao [42], similar constructions had already been considered in the approximation theory literature [95]. Now, a number of papers proved that different random measurement ensembles satisfy the RIP with high probability, as long as m is large enough. The case of a subsampled Fourier transform was first considered in [42] and then refined in [135] and [126], giving the sufficient condition $m \gtrsim s \log n \log m \log^2 s$. These results also extend to subsampled orthogonal matrices and were proved using subtle arguments—in each case chaining techniques—in particular, the latter two papers carefully applied Dudley's inequality. In contrast, matrices with independent sub-Gaussian entries can be handled with more straightforward techniques. For example, as shown in [10], a simple covering argument, similar to the proof of the Johnson-Lindenstrauss lemma, gives the RIP while only requiring $m \gtrsim s \log(n/s)$. In fact, this requirement is optimal up to a constant, which can be proven with the theory of Gelfand widths.

The RIP has been considered in a number of other measurement setups as well. Tropp et al. [157] demonstrated that random demodulators satisfy the RIP under weak conditions; this has clear applications in analog to digital conversion. Rauhut et al. [127] showed that random circulant matrices satisfy the RIP, but under the somewhat strong condition $m \gtrsim (s \log n)^{3/2}$. To be clear,

they considered a measurement matrix A that acts as a sampling of fixed (non-random) entries of the convolution of x with a random vector. To prove the RIP in this setup under the weaker condition $m \gtrsim s\text{polylog}(n)$ appears to be difficult and is still an open problem.

2.2.3 Null space conditions

An important direction of the theory of CS (and sparse recovery in general), was the consideration of null space conditions, i.e., conditions on the null space of A that can be used to imply that ℓ_1 minimization is exact in the noiseless, exactly sparse case and robust in the noisy, approximately sparse case. The quintessential null space condition is as follows. Below, $N(A)$ is the null space of A.

Definition 2.2.1 (Null space property) *A matrix $A \in \mathbb{C}^{m\times n}$ satisfies the null space property of order s if for all subsets $T \in \{1, 2, \cdots, n\}$ with $|T| = s$ it holds that*

$$\|v_T\|_{\ell_1} < \|v_{T^c}\|_{\ell_1} \quad \textit{for all} \quad v \in N(A)\backslash\{0\}.$$

It is straightforward to prove that this is a necessary and sufficient condition for ℓ_1 minimization (2.1.9) to exactly recover all s-sparse signals in the noiseless problem. This result appeared explicitly in [87] and was implicit in the earlier works [60, 70].

Zhang [168] used a stronger null space property to prove stability to noise. In particular, he introduced the requirement

$$\|x\|_0 = \frac{\nu}{4}\left(\frac{\|u\|_{\ell_1}}{\|u\|_{\ell_2}}\right)^2 \quad \text{for some} \quad \nu \in (0,1). \tag{2.2.1}$$

In this condition, we may take u to be any vector in the null space of A, or we can be more specific, as shown below. As noted by Zhang, this is a variation on a well-known sufficient condition considered in the noiseless case (see [168] for details). We pause to give the intuition for this condition: $\|u\|_{\ell_1}/\|u\|_{\ell_2}$ is in some sense an approximation of the sparsity level of u (in particular this ratio is bounded by $\sqrt{\|u\|_0}$). Thus, intuitively, the condition requires the vectors u to be spread, i.e., not overly sparse.

We now describe Zhang's main result; it is RIP-less and quite pertinent in comparison to the results given in this chapter. To best compare, we specialize Zhang's main theorem to the case when A is a subsampled orthogonal matrix. In particular, to simplify, we assume that it is sampled with replacement so that no rows are repeated. We also take A to have unit normed rows (rather than norm $\sqrt{n/m}$ as in our chapter) so that $AA^* = I$. Zhang considered the solution to the constrained LASSO

$$\min \|\bar{x}\|_{\ell_1} \qquad \text{subject to} \quad \|A\bar{x} - y\|_{\ell_2} \le \gamma \tag{2.2.2}$$

where γ should be chosen so that $\|z\|_{\ell_2} \leq \gamma$ (with high probability). He proved the following result.

Theorem 2.2.2 *Let $\gamma \geq \|z\|_{\ell_2}$ and let $\hat{x}$ be the solution to* (2.2.2)*. Assume that $\|x\|_0$ satisfies* (2.2.1) *for $u = (\mathrm{I} - AA^*)(\hat{x} - x)$ whenever $u = (\mathrm{I} - AA^*)(\hat{x} - x) \neq 0$. Then, for either $p = 1$ or $p = 2$*

$$\|\hat{x} - x\|_{\ell_p} \leq \lambda_p (C_\nu + 1)(\|z\|_{\ell_2} + \gamma)$$

where $\gamma_1 = \sqrt{n}, \gamma_2 = 1$ and

$$C\nu = \frac{1 + \nu\sqrt{2 - \nu^2}}{1 - \nu^2}.$$

Note that $I - AA^*$ is the projection onto the null space of A, and thus $u \in N(A)$. To be clear, the theorem states that $\|\hat{x} - x\|_p$ follows the bound above for either $p = 1$ or $p = 2$, not necessarily for both norms simultaneously, and it is not known which norm satisfies the bound. Nevertheless this is an important RIP-less stability result, and combined with a result on Kashin splittings by Guedon et al. [91], it applies to the subsampled Fourier problem and sometimes gives weaker requirements than RIP-based results.

We state the result of Guedon et al. [91, Theorem 3], written in the language of our chapter, except that we once again take the rows of our matrix to have unit norm.

Theorem 2.2.3 *Let $U \in \mathbb{C}^{n \times n}$ be an orthonormal matrix ($UU^* = \mathrm{I}$), whose rows, u_i, satisfy $\|u_i\|_{\ell_\infty}^2 \leq \mu/n$. Then there exists a matrix $A \in \mathbb{C}^{m \times n}$, created as a subsampling of m distinct rows of U, such that for any x in the null space of A, we have*

$$\|x\|_{\ell_1} \geq C \sqrt{\frac{m}{\mu \log m \log^5(\mu \cdot (n/m) \cdot \log m)}} \|x\|_{\ell_2}$$

where C is a fixed constant.

Now, combine this theorem with Zhang's requirement (2.2.1) to demonstrate the existence of subsampled orthogonal matrices that can be used to stably compress s-sparse signals when

$$m \geq Cs\mu \log m \log^5(\mu \cdot (n/m) \cdot \log m).$$

Now, specialize to the case of Fourier measurements which are not drastically undersampled, so that $m \geq n/\log m$. The requirement becomes

$$m \geq Cs \log m \log^5(\log m)$$

in this situation, which compares quite favorably with the best known RIP-requirement [135]

$$m \geq Cs \log m \log n \log^2 s.$$

This also comes quite close to the number of samples required in this chapter, $m \gtrsim s \log n$.

2.2.4 Other algorithms for CS

While there has been quite a bit of work on ℓ_1-minimization-based programs (e.g., the LASSO, basis pursuit, and the Dantzig selector), algorithms based on different approaches offer distinct advantages in certain areas. In particular, *greedy algorithms* tend to be much faster, but in return they tend to require more measurements for successful signal recovery.

There is a strong base of theoretical results on such greedy algorithms and we state a few such results. The simplest greedy algorithm, orthogonal matching pursuit (OMP) [54, 123], selects one coefficient at a time to include in the support of β. In particular, at each step it creates a residual by taking the projection of y onto the complement of the space spanned by the columns already included in the model, and adds to the model the column which has the highest inner product with this residual (i.e., forward selection). In [156] Tropp–Gilbert demonstrated that with high probability $O(s \log n)$ Gaussian measurements are sufficient to recover an s-sparse signal by OMP. Variations on this algorithm have also been developed, e.g., stagewise OMP [63] and regularized OMP [114,115]. In fact, Needell–Vershynin [114, 115] proved that under RIP conditions, regularized OMP is stable to noise, although in comparison to analogous results in convex optimization, the error bound is suboptimal by a logarithmic factor and so is the requirement of the RIP constant δ. More recently, Needell–Tropp [113] introduced a greedy-type algorithm called compressive sampling matching pursuit (CoSaMP) and proved stability to noise via the RIP, but without any extra logarithmic factors. Dai–Milenkovic [53] gave similar RIP-based guarantees for the greedy algorithm termed subspace pursuit.

2.3 Fundamental Estimates

Our proofs rely on several estimates, and we provide an interpretation of each whenever possible. The first estimates **E1**–**E4** are used to prove the noiseless recovery result; when combined with the weak RIP, they imply stability and robustness. Lemmas 2.3.1, 2.3.3, and 2.3.4 below are involved in the construction of an approximation of a dual vector (see Section 2.4), inspired by a similar construction in [88]. Thus, these lemmas are adaptations of similar results from [88]. Throughout this section, δ is a parameter left to be fixed in later sections; it is always less than or equal to one.

2.3.1 Local isometry

Let T of cardinality s be the support of x in Theorem 2.1.1, or the support of the best s-sparse approximation of x in Theorem 2.1.2. We shall need that with high probability,

$$\|A_T^* A_T - I\| \leq \delta \tag{2.3.1}$$

with $\delta \le 1/2$ in the proof of Theorem 2.1.1 and $\delta \le 1/4$ in that of Theorem 2.1.2. Put differently, the singular values of A_T must lie away from zero. This condition essentially prevents A_T from being singular as, otherwise, there would be no hope of recovering our sparse signal x. Indeed, letting h be any vector supported on T and in the null space of A, we would have $Ax = A(x+h)$ and thus, recovery would be impossible even if one knew the support of x. The condition (2.3.1) is much weaker than the restricted isometry property because it does not need to hold uniformly over all sparse subsets—only on the support set.

Lemma 2.3.1 (E1: local isometry) *Let T be a fixed set of cardinality s. Then for $\delta > 0$,*

$$\mathbb{P}\left(\|A_T^* A_T - I\| \ge \delta\right) \le 2s \exp\left(-\frac{m}{\mu(F)s} \cdot \frac{\delta^2}{2(1+\delta/3)}\right). \tag{2.3.2}$$

In particular, if $m \ge \frac{56}{3}\mu(F) \cdot s \cdot \log n$, then

$$\mathbb{P}\left(\|A_T^* A_T - I\| \ge 1/2\right) \le 2/n.$$

Note that $\|A_T^* A_T - I\| \le \delta$ implies that $\left\|(A_T^* A_T)^{-1}\right\| \le 1/(1-\delta)$, a fact that we will use several times.

In compressive sensing, the standard way of proving such estimates is via Rudelson's selection theorem [133]. Here, we use a more modern technique based on the matrix Bernstein inequality of Ahlswede and Winter [3], developed for this setting by Gross [88], and tightened in [155] by Tropp and in [120] by Oliveira. We present the version in [155].

Theorem 2.3.2 (Matrix Bernstein inequality) *Let $\{X_k\} \in \mathbb{R}^{d\times d}$ be a finite sequence of independent random self-adjoint matrices. Suppose that $\mathbb{E}\, X_k = 0$ and $\|X_k\| \le B$ a.s. and put*

$$\sigma^2 := \left\|\sum_k \mathbb{E}\, X_k^2\right\|.$$

Then for all $t \ge 0$,

$$\mathbb{P}\left(\left\|\sum_k X_k\right\| \ge t\right) \le 2d \exp\left(\frac{-t^2/2}{\sigma^2 + Bt/3}\right). \tag{2.3.3}$$

Proof Decompose $A_T^* A_T - I$ as

$$A_T^* A_T - I = m^{-1}\sum_{k=1}^m (a_{k,T} a_{k,T}^* - I) = m^{-1}\sum_{k=1}^m X_k, \qquad X_k := a_{k,T} a_{k,T}^* - I.$$

The isotropy condition implies $\mathbb{E}\, X_k = 0$, and since $\|a_T\|_{\ell_2}^2 \le \mu(F)\cdot s$, we have $\|X_k\| = \max(\|a_{i,T}\|_{\ell_2}^2 - 1, 1) \le \mu(F)\cdot s$. Last, $0 \le \mathbb{E}\, X_k^2 = \mathbb{E}(a_{k,T} a_{k,T}^*)^2 - I \le \mathbb{E}(a_{k,T} a_{k,T}^*)^2 = \mathbb{E}\,\|a_{k,T}\|^2 a_{k,T} a_{k,T}^*$. However,

$$\mathbb{E}\,\|a_{k,T}\|^2 a_{k,T} a_{k,T}^* \le \mu(F)\cdot s \cdot \mathbb{E}\, a_{k,T} a_{k,T}^* = \mu(F)\cdot s \cdot I$$

and, therefore, $\sum_k \mathbb{E}\, X_k^2 \le m\cdot\mu(F)\cdot s\cdot I$ so that σ^2 is bounded above by $m\cdot\mu(F)\cdot s$. Plugging $t=\delta m$ into (2.3.3) gives the lemma. ∎

Instead of having A act as a near isometry on all vectors supported on T, we could ask that it preserves the norm of an arbitrary fixed vector (with high probability), i.e. $\|Av\|_{\ell_2}\approx\|v\|_{\ell_2}$ for a fixed v supported on T. Not surprisingly, this can be proved with generally (slightly) weaker requirements.

Lemma 2.3.3 (E2: low-distortion) *Let v be a fixed vector supported on a set T of cardinality at most s. Then for each $t\le 1/2$,*

$$\mathbb{P}(\|(A_T^*A_T-I)v_T\|_{\ell_2}\ge t\|v\|_{\ell_2})\le\exp\Big(-\frac{1}{4}\Big(t\sqrt{\frac{m}{\mu(F)s}}-1\Big)^2\Big).$$

The proof is an application of the vector Bernstein inequality described in the fourth estimate **E4**. It is analogous to the proof shown there and is not repeated.

2.3.2 Off-support incoherence

Lemma 2.3.4 (E3: off-support incoherence) *Let v be supported on T with $|T|=s$. Then for each $t>0$,*

$$\mathbb{P}(\|A_{T^c}^*Av\|_{\ell_\infty}\ge t\|v\|_{\ell_2})\le 2n\exp\left(-\frac{m}{2\mu(F)}\cdot\frac{t^2}{1+\frac{1}{3}\sqrt{s}t}\right). \tag{2.3.4}$$

This lemma says that if $v=x$, then $\max_{i\in T^c}|\langle A_{\{i\}},Ax\rangle|$ cannot be too large so that the off-support columns do not correlate too well with Ax. The proof of **E3** is an application of Bernstein's inequality—the matrix Bernstein inequality with $d=1$—together with the union bound.

Proof We have

$$\|A_{T^c}^*Av\|_{\ell_\infty}=\max_{i\in T^c}|\langle e_i,A^*Av\rangle|.$$

Assume without loss of generality that $\|v\|_{\ell_2}=1$, fix $i\in T^c$ and write

$$\langle e_i,A^*Av\rangle=\frac{1}{m}\sum_k g_k,\qquad g_k:=\langle e_i,a_ka_k^*v\rangle.$$

Since $i\in T^c$, $\mathbb{E}\,g_k=0$ by the isotropy property. Next, the Cauchy-Schwartz inequality gives $|g_k|=|\langle e_i,a_k\rangle\cdot\langle a_k,v\rangle|\le|\langle e_i,a_k\rangle|\|a_{k,T}\|_{\ell_2}$. Since $|\langle e_i,a_k\rangle|\le\sqrt{\mu(F)}$ and $\|a_{k,T}\|_{\ell_2}\le\sqrt{\mu(F)s}$, we have $|g_k|\le\mu(F)\sqrt{s}$. Last, for the total variance, we have

$$\mathbb{E}\,g_k^2\le\mu(F)\,\mathbb{E}\langle a_{k,T},v\rangle^2=\mu(F)$$

where the equality follows from the isotropy property. Hence, $\sigma^2\le m\mu(F)$, and Bernstein's inequality gives

$$\mathbb{P}(|\langle e_i,A^*Av\rangle|\ge t)\le 2\exp\left(-\frac{m}{2\mu(F)}\cdot\frac{t^2}{1+\frac{1}{3}\sqrt{s}t}\right).$$

Combine this with the union bound over all $i \in T^c$ to give the desired result. ∎

We also require the following related bound:

$$\max_{i \in T^c} \|A_T^* A_{\{i\}}\|_{\ell_2} \le \delta.$$

In other words, none of the column vectors of A outside of the support of x should be well approximated by *any* vector sharing the support of x.

Lemma 2.3.5 (E4: uniform off-support incoherence) *Let T be a fixed set of cardinality s. For any $0 \le t \le \sqrt{s}$,*

$$\mathbb{P}\left\{\max_{i \in T^c} \|A_T^* A_{\{i\}}\|_{\ell_2} \ge t\right\} \le n \exp\left(-\frac{mt^2}{8\mu(F)s} + \frac{1}{4}\right).$$

In particular, if $m \ge 8\mu(F) \cdot s \cdot (2 \log n + 1/4)$, then

$$\mathbb{P}\left\{\max_{i \in T^c} \|A_T^* A_{\{i\}}\|_{\ell_2} \ge 1\right\} \le 1/n.$$

The estimate follows from the vector Bernstein inequality, which essentially follows from Chapter 6 of [100], and was proved by Gross [88, Theorem 11]. We use a slightly weaker version, which we find slightly more convenient.

Theorem 2.3.6 (Vector Bernstein inequality) *Let $\{v_k\} \in \mathbb{R}^d$ be a finite sequence of independent random vectors. Suppose that $\mathbb{E}\, v_k = 0$ and $\|v_k\|_{\ell_2} \le B$ a.s. and put $\sigma^2 \ge \sum_k \mathbb{E}\, \|v_k\|_{\ell_2}^2$. Then for all $0 \le t \le \sigma^2/B$,*

$$\mathbb{P}\left\{\|\sum_k v_k\|_{\ell_2} \ge t\right\} \le \exp\left(-\frac{(t/\sigma - 1)^2}{4}\right) \le \exp\left(-\frac{t^2}{8\sigma^2} + \frac{1}{4}\right). \tag{2.3.5}$$

Note that the bound does not depend on the dimension d.

Proof Fix $i \in T^c$ and write

$$A_T^* A_{\{i\}} = \frac{1}{m} \sum_{j=1}^m a_{k,T}^* \langle a_k, e_i \rangle := \frac{1}{m} \sum_{k=1}^m v_k.$$

As before, $\mathbb{E}\, v_k = \mathbb{E}\, a_{k,T}^* \langle a_k, e_i \rangle = 0$ since $i \in T^c$. Also, $\|v_k\|_{\ell_2} = \|a_{k,T}\|_{\ell_2} |\langle a_k, e_i \rangle| \le \mu(F)\sqrt{s}$. Last, we calculate the sum of expected squared norms,

$$\sum_{k=1}^m \mathbb{E}\, \|v_k\|_{\ell_2}^2 = m\, \mathbb{E}\, \|v_1\|_{\ell_2}^2 \le m\, \mathbb{E}[\|a_{1,T}\|_{\ell_2}^2 \langle e_i, a_1 \rangle^2] \le m\mu(F)s \cdot \mathbb{E}\langle e_i, a_1 \rangle^2 = m\mu(F)s.$$

As before, the last equality follows from the isotropy property. Bernstein's inequality together with the union bound give the lemma. ∎

2.3.3 Weak RIP

In the nonsparse and noisy setting, we shall make use of a variation on the restricted isometry property to control the size of the reconstruction error. This variation is as follows:

Theorem 2.3.7 (E5: weak RIP) *Let T be a fixed set of cardinality s and fix $\delta > 0$. Then for all v supported on $T \cup R$, where R is any set of cardinality $|R| \leq r$, we have*

$$(1-\delta)\|v\|_{\ell_2}^2 \leq \|Av\|_{\ell_2}^2 \leq (1+\delta)\|v\|_{\ell_2}^2 \tag{2.3.6}$$

with probability at least $1 - 5e^{-\beta}$ provided that

$$m \geq C_\delta \cdot \beta \cdot \mu(F) \cdot \max(s\log(s\mu), r\log n \log^2(r\mu)\log(r\mu\log n)).$$

Here C_δ is a fixed numerical constant which only depends upon δ.

This theorem is proved in Section 2.6 using Talagrand's generic chaining construction, and combines the framework and results of Rudelson and Vershynin in [135] and [133]. In the proof of Theorem 2.1.2, we take $\delta = 1/4$. Out of our estimates, the weak RIP is the one that is truly novel, while the others have clear antecedents.

The condition says that the column space of A_T should not be too close to that spanned by another small disjoint set R of columns. To see why a condition of this nature is necessary for any recovery algorithm, suppose that x has fixed support T and that there is a single column $A_{\{i\}}$ which is a linear combination of columns in T, i.e., $A_{T\cup\{i\}}$ is singular. Let $h \neq 0$ be supported on $T \cup \{i\}$ and in the null space of A. Then $Ax = A(x+th)$ for any scalar t. Clearly, there are some values of t such that $x+th$ is at least as sparse as x, and thus one should not expect to be able to recover x by any method. In general, if there were a vector v as above obeying $\|Av\|_{\ell_2} \ll \|v\|_{\ell_2}$ then one would have $A_T v_T \approx -A_R v_R$. Thus, if the signal x were the restriction of v to T, it would be very difficult to distinguish it from that of $-v$ to R under the presence of noise.

The weak RIP is a combination of the RIP and the local conditioning estimate **E1**. When $r = 0$, this is **E1** whereas this is the restricted isometry property when $s = 0$. The point is that we do not need the RIP to hold for sparsity levels on the order of $m/[\mu(F)\log n]$. Instead we need the following property: consider an arbitrary submatrix formed by concatenating columns in T with r other columns from A selected in any way you like; then we would like this submatrix to be well conditioned. Because T is fixed, one can prove good conditioning when s is significantly larger than the maximum sparsity level considered in the standard RIP.

2.3.4 Implications

The careful reader may ask why we bothered to state estimates **E1**–**E4** since they are all implied by the weak RIP! Our motivation is threefold: (1) some of these estimates, e.g. **E2** hold with better constants and weaker requirements than those implied by the weak RIP machinery; (2) the weak RIP requires an in-depth proof whereas the other estimates are simple applications of well-known theorems, and we believe that these theorems and the estimates should be independently useful tools to other researchers in the field; (3) the noiseless theorem does not require the weak RIP.

2.4 Noiseless and Sparse Recovery

This section proves the noiseless recovery theorem, namely, Theorem 2.1.1. Our proof essentially adapts the arguments of David Gross [88] from the low-rank matrix recovery problem.

2.4.1 Dual certificates

A standard method for establishing exact recovery is to exhibit a *dual certificate*; that is to say, a vector v obeying the two properties below.

Lemma 2.4.1 (Exact duality) *Set $T = supp(x)$ with x feasible for (2.1.9), and assume A_T has full column rank. Suppose there exists $v \in \mathbb{R}^n$ in the row space of A obeying*

$$v_T = sgn(x_T) \qquad and \qquad \|v_{T^c}\|_{\ell_\infty} < 1. \tag{2.4.1}$$

Then x is the unique ℓ_1 minimizer to (2.1.9).

The proof is now standard, see [42, 80, 150]. Roughly, the existence of a dual vector implies that there is a subgradient of the ℓ_1 norm at x that is perpendicular to the feasible set. This geometric property shows that x is solution. Following Gross, we slightly modify this definition as to make use of an 'inexact dual vector'.

Lemma 2.4.2 (Inexact duality) *Set $T = supp(x)$ where x is feasible, and assume that*

$$\left\|(A_T^* A_T)^{-1}\right\| \le 2 \qquad and \qquad \max_{i \in T^c} \|A_T^* A_{\{i\}}\|_{\ell_2} \le 1. \tag{2.4.2}$$

Suppose there exists $v \in \mathbb{R}^n$ in the row space of A obeying

$$\|v_T - sgn(x_T)\|_{\ell_2} \le 1/4 \qquad and \qquad \|v_{T^c}\|_{\ell_\infty} \le 1/4. \tag{2.4.3}$$

Then x is the unique ℓ_1 minimizer to (2.1.9).

Proof Let $\hat{x} = x + h$ be a solution to (2.1.9) and note that $Ah = 0$ since both x and $\hat{x}$ are feasible. To prove the claim, it suffices to show that $h = 0$. We begin by observing that

$$\|\hat{x}\|_{\ell_1} = \|x_T + h_T\|_{\ell_1} + \|h_{T^c}\|_{\ell_1} \geq \|x_T\|_{\ell_1} + \langle \mathrm{sgn}(x_T), h_T \rangle + \|h_{T^c}\|_{\ell_1}.$$

Letting $v = A^* w$ be our (inexact) dual vector, we have

$$\langle \mathrm{sgn}(x_T), h_T \rangle = \langle \mathrm{sgn}(x_T) - v_T, h_T \rangle + \langle v_T, h_T \rangle = \langle \mathrm{sgn}(x_T) - v_T, h_T \rangle - \langle v_{T^c}, h_{T^c} \rangle,$$

where we used $\langle v_T, h_T \rangle = \langle v, h \rangle - \langle v_{T^c}, h_{T^c} \rangle = -\langle v_{T^c}, h_{T^c} \rangle$ since $\langle v, h \rangle = \langle w, Ah \rangle = 0$. The Cauchy-Schwartz inequality combined with Hölder's inequality and the properties of v yield

$$\begin{aligned}
|\langle \mathrm{sgn}(x_T), h_T \rangle| &\leq |\langle \mathrm{sgn}(x_T) - v_T, h_T \rangle| + |\langle v_T, h_T \rangle| \\
&= |\langle \mathrm{sgn}(x_T) - v_T, h_T \rangle| + |\langle v_{T^c}, h_{T^c} \rangle| \\
&\leq \|\mathrm{sgn}(x_T) - v_T\|_{\ell_2} \cdot \|h_T\|_{\ell_2} + \|v_{T^c}\|_{\ell_\infty} \cdot \|h_{T^c}\|_{\ell_1} \\
&\leq \frac{1}{4}(\|h_T\|_{\ell_2} + \|h_{T^c}\|_{\ell_1}).
\end{aligned}$$

Therefore,

$$\|\hat{x}\|_{\ell_1} \geq \|x\|_{\ell_1} - \frac{1}{4}\|h_T\|_{\ell_2} + \frac{3}{4}\|h_{T^c}\|_{\ell_1}.$$

We now bound $\|h_T\|_{\ell_2}$. First, it follows from

$$h_T = (A_T^* A_T)^{-1} A_T^* A_T h = -(A_T^* A_T)^{-1} A_T^* A_{T^c} h_{T^c}$$

that $\|h_T\|_{\ell_2} \leq 2\|A_T^* A_{T^c} h_{T^c}\|_{\ell_2}$. Second,

$$\|A_T^* A_{T^c} h_{T^c}\|_{\ell_2} \leq 2 \sum_{i \in T^c} \|A_T^* A_{\{i\}}\|_{\ell_2} |h_i| \leq \max_{i \in T^c} \|A_T^* A_{\{i\}}\|_{\ell_2} \|h_{T^c}\|_{\ell_1} \leq \|h_{T^c}\|_{\ell_1}.$$

In conclusion, $\|h_T\|_2 \leq 2\|h_{T^c}\|_1$ and thus,

$$\|\hat{x}\|_{\ell_1} \geq \|x\|_{\ell_1} + \frac{1}{4}\|h_{T^c}\|_{\ell_1}.$$

This implies $h_{T^c} = 0$, which in turn implies $h_T = 0$ since we must have $A_T h_T = Ah = 0$ (and A_T has full rank). ■

Lemma 2.4.3 (Existence of a dual certificate) *Under the hypotheses of Theorem 2.1.1, one can find $v \in \mathbb{R}^n$ obeying the conditions of Lemma 2.4.2 with probability at least $1 - e^{-\beta} - 1/n$.*

This lemma, which is proved next, implies Theorem 2.1.1. The reason is that we just need to verify conditions (2.4.2). However, by Lemmas 2.3.1 and 2.3.5, they jointly hold with probability at least $1 - 3/n$ provided that $m \geq \mu \cdot s \cdot (19 \log n + 2)$ (recall that μ is a shorthand for $\mu(F)$).

2.4.2 Proof of Lemma 2.4.3

The proof uses the clever *golfing scheme* introduced in [88]. Partition A into row blocks so that from now on, A_1 are the first m_1 rows of the matrix A, A_2 the next m_2 rows, and so on. The ℓ matrices $\{A_i\}_{i=1}^{\ell}$ are independently distributed, and we have $m_1 + m_2 + \ldots + m_\ell = m$. As before, $A_{i,T}$ is the restriction of A_i to the columns in T.

The golfing scheme then starts with $v_0 = 0$, inductively defines

$$v_i = \frac{m}{m_i} A_i^* A_{i,T}(\operatorname{sgn}(x_T) - v_{i-1,T}) + v_{i-1}$$

for $i = 1, \ldots, \ell$, and sets $v = v_\ell$. Clearly v is in the row space of A. To simplify notation, let $q_i = \operatorname{sgn}(x_T) - v_{i,T}$, and observe the two identities

$$q_i = \left(I - \frac{m}{m_i} A_{i,T}^* A_{i,T}\right) q_{i-1} = \prod_{j=1}^{i} \left(I - \frac{m}{m_j} A_{j,T}^* A_{j,T}\right) \operatorname{sgn}(x_T) \tag{2.4.4}$$

and

$$v = \sum_{i=1}^{\ell} \frac{m}{m_i} A_i^* A_{i,T}\, q_{i-1}, \tag{2.4.5}$$

which shall be used frequently. From (2.4.4) and the fact that $I - \frac{m}{m_i} A_{i,T}^* A_{i,T}$ should be a contraction (local isometry **E1**), we see that the norm of q_i decreases geometrically fast—the terminology comes from this fact since each iteration brings us closer to the target just as each golf shot would bring us closer to the hole—so that v_T should be close to $\operatorname{sgn}(x_T)$. Hopefully, the process keeps the size of v_{T^c} under control as well.

To control the size of v_{T^c} and that of $\operatorname{sgn}(x_T) - v_T$, we claim that the following inequalities hold for each i with high probability: first,

$$\|q_i\|_{\ell_2} \leq c_i \|q_{i-1}\|_{\ell_2} \tag{2.4.6}$$

and, second,

$$\|\frac{m}{m_i} A_{i,T^c}^* A_{i,T}\, q_{i-1}\|_{\ell_\infty} \leq t_i \|q_{i-1}\|_{\ell_2} \tag{2.4.7}$$

(the values of the parameters t_i and c_i will be specified later). Let $p_1(i)$ (resp. $p_2(i)$) be the

probability that the bound (2.4.6) (resp. (2.4.7)) does not hold. Lemma 2.3.3 gives

$$p_1(i) \le \exp\left(-\frac{1}{4}(c_i\sqrt{m_i/(s\mu)} - 1)^2\right). \tag{2.4.8}$$

Thus, if

$$m_i \ge \frac{2 + 8(\beta + \log\alpha)}{c_i^2} s\mu, \tag{2.4.9}$$

then $p_1(i) \le \frac{1}{\alpha}e^{-\beta}$. Next, Lemma 2.3.4 gives

$$p_2(i) \le 2n\exp\left(-\frac{3t_i^2 m_i}{6\mu + 2\mu\sqrt{s}t_i}\right). \tag{2.4.10}$$

Thus, if

$$m_i \ge \left(\frac{2}{t_i^2 s} + \frac{2}{3t_i\sqrt{s}}\right)(\beta + \log(2\alpha) + \log n)s\mu, \tag{2.4.11}$$

then $p_2(i) \le \frac{1}{\alpha}e^{-\beta}$.

It is now time to set the number of blocks ℓ, the block sizes m_i and the values of the parameters c_i and t_i. These are as follows:

- $\ell = \lceil(\log_2 s)/2\rceil + 2$;
- $c_1 = c_2 = 1/[2\sqrt{\log n}]$ and $c_i = 1/2$ for $3 \le i \le \ell$;
- $t_1 = t_2 = 1/[8\sqrt{s}]$ and $t_i = \log n/[8\sqrt{s}]$ for $3 \le i \le \ell$;
- $m_1, m_2 \ge 35(1 + \log 4 + \beta)s\mu c_i^{-2}$ and $m_i \ge 35(1 + \log 6 + \beta)s\mu c_i^{-2}$ for $3 \le i \le \ell$.

It is not hard to see that the total number of samples $m = \sum_i m_i$ obeys the assumptions of the lemma. To see why v is a valid certificate, suppose first that for each i, (2.4.6) and (2.4.7) hold. Then (2.4.4) gives

$$\|\mathrm{sgn}(x_T) - v_T\|_{\ell_2} = \|q_\ell\|_{\ell_2} \le \|\mathrm{sgn}(x_T)\|_{\ell_2} \prod_{i=1}^{\ell} c_i \le \frac{\sqrt{s}}{2^\ell} \le \frac{1}{4}$$

as desired. Further, (2.4.5) yields

$$\|v_{T^c}\|_{\ell_\infty} \le \sum_{i=1}^{\ell} \|\frac{m}{m_i} A^*_{i,T^c} A_{i,T} q_{i-1}\|_{\ell_\infty} \le \sum_{i=1}^{\ell} t_i \|q_{i-1}\|_{\ell_2} \le \sqrt{s}\sum_{i=1}^{\ell} t_i \prod_{j=1}^{i-1} c_i.$$

Now with our choice of parameters, the right-hand side is bounded above by

$$\frac{1}{8}\left(1 + \frac{1}{2\sqrt{\log n}} + \frac{\log n}{4\log n} + \cdots\right) < \frac{1}{4},$$

which is the desired conclusion.

Now we must show that the bounds (2.4.6), (2.4.7) hold with probability at least $1 - e^{-\beta} - 1/n$. It follows from (2.4.9) and (2.4.11) that $p_1(i), p_2(i) \le \frac{1}{4}e^{-\beta}$ for $i = 1, 2$ and $p_1(i), p_2(i) \le \frac{1}{6}e^{-\beta} \le 1/6$ for $i \ge 3$. Thus, $p_1(1) + p_1(2) + p_2(1) + p_2(2) \le e^{-\beta}$ and $p_1(i) + p_2(i) \le 1/3$ for $i \ge 3$. Now the union bound would never show that (2.4.6) and (2.4.7) hold with probability at least $1 - 1/n$ for all $i \ge 3$ because of the weak bound on $p_1(i) + p_2(i)$. However, using a clever idea in [88], it is not necessary for each subset of rows to 'succeed' and give the desired bounds. Instead, one can sample a 'few' extra batches of rows, and throw out those that fail our requirements. We only need $\ell - 2$ working batches, after the first 2. In particular, pick $\ell' + 2 > \ell$ batches of rows, so that we require $m \ge 2 \cdot \lceil 140(1 + \log 4 + \beta) \cdot \mu \cdot s \cdot \log n \rceil + \ell \cdot \lceil 140(1 + \log 6 + \beta) s\mu \rceil$ (note that we have made no attempt to optimize constants). Now as in [88], let N be the the number of batches—after the first 2—obeying (2.4.6) and (2.4.7); this N is larger (probabilistically) than a binomial$(\ell', 2/3)$ random variable. Then a standard concentration bound [107, Theorem 2.3a]

$$\mathbb{P}(N < \ell - 2) \le \exp\left(-2\,\frac{(\frac{2}{3}\ell' - \ell + 2)^2}{\ell'}\right)$$

tells us that if we were to pick $\ell' = 3\lceil \log n \rceil + 1$, we would have

$$\mathbb{P}(N < \ell - 2) \le 1/n.$$

In summary, from $p_1(1)+p_2(1)+p_1(2)+p_2(2) \le e^{-\beta}$ and the calculation above, the dual certificate v obeys the required properties with probability at least $1 - 1/n - e^{-\beta}$, provided that $m \ge C(1 + \beta) \cdot \mu \cdot s \cdot \log n$.

2.5 General Signal Recovery from Noisy Data

We prove the general recovery theorems from Section 2.1.6 under the assumption of Gaussian white noise but would like to emphasize that the same result would hold for other noise distributions. Specifically, suppose we have the noisy model

$$y = Ax + z, \quad \text{where} \quad \|A^* z\|_{\ell_\infty} \le \lambda_n \tag{2.5.1}$$

holds with high probability. Then the conclusions of Theorem 2.1.3 remain valid. In details, the Dantzig selector with constraint $\|A^*(y - A\bar{x})\|_{\ell_\infty} \le 4\lambda_n$ obeys

$$\|\hat{x} - x\|_{\ell_2} \le C_1(1 + \alpha^2)\left[\frac{\|x - x_s\|_{\ell_1}}{\sqrt{s}} + \lambda_n\sqrt{s}\right] \tag{2.5.2}$$

with high probability. Hence, (2.1.15) is a special case corresponding to $\lambda_n = 2.5\sigma_m\sqrt{\log n} = 2.5\sigma\sqrt{\frac{\log n}{m}}$. Likewise, the bound on the ℓ_1 loss (2.1.16) with λ_n in place of $\sigma\sqrt{\frac{\log n}{m}}$ holds as well. A similar generality applies to the LASSO as well, although in this case we need a second noise correlation bound, namely,

$$\|A_{T^c}^*(I-P)z\|_{\ell_\infty} \le \lambda_n$$

where $P := A_T(A_T^*A_T)^{-1}A_T^*$ is the projection onto the column space of A_T.

Now when $z \sim \mathcal{N}(0, I)$ and A is a fixed matrix, we have

$$\|A^*z\|_{\ell_\infty} \le 2\|A\|_{1,2}\sqrt{\log n} \tag{2.5.3}$$

with probability at least $1 - 1/2n$; here, $\|A\|_{1,2}$ is the maximum column norm of A. Indeed, the ith component of A^*z is distributed as $\mathcal{N}(0, \|A_{\{i\}}\|_{\ell_2}^2)$ and, therefore, the union bound gives

$$\mathbb{P}(\|A^*z\|_{\ell_\infty} > 2\|A\|_{1,2}\sqrt{\log n}) \le n\,\mathbb{P}(|\mathcal{N}(0,1)| > 2\sqrt{\log n}).$$

The conclusion follows for $n \ge 2$ from the well-known tail bound $\mathbb{P}(|\mathcal{N}(0,1)| > t) \le 2\phi(t)/t$, where ϕ is the density of the standard normal distribution. The same steps demonstrate that

$$\|A^*(I-P)z\|_{\ell_\infty} \le 2\|(I-P)A\|_{1,2}\sqrt{\log n} \le 2\|A\|_{1,2}\sqrt{\log n} \tag{2.5.4}$$

with probability at least $1 - 1/2n$.

2.5.1 Proof of Theorem 2.1.2

We assume $\sigma_m = 1$ since the general result follows from a simple rescaling.

Fix s obeying $s \le \bar{s}$, and let $T = \mathrm{supp}(x_s)$. We prove the error bounds of Theorem 2.1.2 with s fixed, and the final result follows by considering that s which minimizes either the ℓ_2 (2.1.11) or ℓ_1 (2.1.12) error bound. This is proper since the minimizing s has a deterministic value. With T as above, we assume in the rest of the proof that

(i) all of the requirements for noiseless recovery in Lemma 2.4.2 are met,

(ii) and that the inexact dual vector v of Section 2.4 is successfully constructed.

All of this occurs with probability at least $1 - 4/n - e^{-\beta}$. Further, we assume that

(iii) the weak RIP holds with $\delta = 1/4$, $r = \frac{m}{C(1+\beta)\cdot\mu\cdot\log n\log m\log^2(s\mu)} \vee s$ and T is as above.

This occurs with probability at least $1 - 5e^{-\beta}$, and implies the RIP at sparsity level r and restricted isometry constant $\delta = 1/4$. Last, we assume

(iv) the noise correlation bound

$$\|A^* z\|_{\ell_\infty} \le 2.5\sqrt{\log n}. \tag{2.5.5}$$

Assuming the weak RIP above, which implies $\|A\|_{1,2} \le 5/4$, the conditional probability that this occurs is at least $1 - 1/2n$ because of (2.5.3). Because the weak RIP implies the local isometry condition **E1** with $\delta = 1/4$, all of these conditions together hold with probability at least $1 - 4/n - 6e^{-\beta}$. All of the steps in the proof are now deterministic consequences of (i)–(iv); from now on, we will assume they hold.

With $h = \hat{x} - x$, our goal is to bound both the ℓ_2 and ℓ_1 norms of h. We will do this with a pair of lemmas. The first is frequently used (recall that λ is set to $10\sqrt{\log n}$).

Lemma 2.5.1 (Tube constraint) *The error h obeys*

$$\|A^* A h\|_{\ell_\infty} \le \frac{5\lambda}{4}.$$

Proof As shown in [33, Lemma 3.1], writing that the zero vector is a subgradient of the LASSO functional $\frac{1}{2}\|y - A\bar{x}\|_{\ell_2}^2 + \lambda\|\bar{x}\|_{\ell_1}$ at $\bar{x} = \hat{x}$ gives

$$\|A^*(y - A\hat{x})\|_{\ell_\infty} \le \lambda.$$

Then it follows from the triangle inequality that

$$\|A^* A h\|_{\ell_\infty} \le \|A^*(y - A\hat{x})\|_{\ell_\infty} + \|A^* z\|_{\ell_\infty} \le \lambda + \|A^* z\|_{\ell_\infty},$$

where z is our noise term. The claim is a consequence of (2.5.5). ∎

Lemma 2.5.2 *The error h obeys*

$$\|h_{T^c}\|_{\ell_1} \le C_0(s\lambda + \|x_{T^c}\|_{\ell_1}) \tag{2.5.6}$$

for some numerical constant C_0.

Before proving this lemma, we show that it gives Theorem 2.1.2. Some of the steps are taken from the proof of Theorem 1.1 in [43].

Proof [Theorem 2.1.2] Set r as in (iii) above. We begin by partitioning T^c and let T_1 be the indices of the r largest entries of h_{T^c}, T_2 be those of the next r largest, and so on. We first bound $\|h_{T\cup T_1}\|_{\ell_2}$ and set $\bar{T}_1 = T \cup T_1$ for short. The weak RIP assumption (iii) gives

$$\frac{3}{4}\|h_{\bar{T}_1}\|_{\ell_2}^2 \le \|A_{\bar{T}_1} h_{\bar{T}_1}\|_{\ell_2}^2 = \langle A_{\bar{T}_1} h_{\bar{T}_1}, Ah\rangle - \langle A_{\bar{T}_1} h_{\bar{T}_1}, A_{\bar{T}_1^c} h_{\bar{T}_1^c}\rangle. \tag{2.5.7}$$

From Lemma 2.5.1, we have

$$\langle A_{\bar{T}_1} h_{\bar{T}_1}, Ah \rangle = \langle h_{\bar{T}_1}, A^*_{\bar{T}_1} Ah \rangle \leq \|h_{\bar{T}_1}\|_{\ell_1} \|A^*_{\bar{T}_1} Ah\|_{\ell_\infty} \leq \frac{5}{4}\lambda \|h_{\bar{T}_1}\|_{\ell_1}.$$

Since $\bar{T}_1$ has cardinality at most $2s$, the Cauchy-Schwarz inequality gives

$$\langle A_{\bar{T}_1} h_{\bar{T}_1}, Ah \rangle \leq \frac{5}{4}\lambda\sqrt{2s} \|h_{\bar{T}_1}\|_{\ell_2}. \tag{2.5.8}$$

Next, we bound $|\langle A_{\bar{T}_1} h_{\bar{T}_1}, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle| \leq |\langle A_T h_T, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle| + |\langle A_{T_1} h_{T_1}, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle|$. We have

$$\langle A_T h_T, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle \leq \sum_{j \geq 2} |\langle A_T h_T, A_{T_j} h_{T_j} \rangle|. \tag{2.5.9}$$

As shown in [41, Lemma 1.2], the parallelogram identity together with the weak RIP imply that

$$|\langle A_T h_T, A_{T_j} h_{T_j} \rangle| \leq \frac{1}{4} \|h_T\|_{\ell_2} \|h_{T_j}\|_{\ell_2}$$

and, therefore,

$$\langle A_T h_T, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle \leq \frac{1}{4} \|h_T\|_{\ell_2} \sum_{j \geq 2} \|h_{T_j}\|_{\ell_2}. \tag{2.5.10}$$

To bound the summation, we use a standard result [43, (3.10)] [2]

$$\sum_{j \geq 2} \|h_{T_j}\|_{\ell_2} \leq r^{-1/2} \|h_{T^c}\|_{\ell_1}, \tag{2.5.11}$$

which gives

$$|\langle A_T h_T, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle| \leq \frac{1}{4} r^{-1/2} \|h_T\|_{\ell_2} \|h_{T^c}\|_{\ell_1}.$$

The same analysis yields $|\langle A_{T_1} h_{T_1}, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle| \leq \frac{1}{4} r^{-1/2} \|h_{T_1}\|_{\ell_2} \|h_{T^c}\|_{\ell_1}$ and thus,

$$|\langle A_{\bar{T}_1} h_{\bar{T}_1}, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle| \leq \frac{1}{2} r^{-1/2} \|h_{\bar{T}_1}\|_{\ell_2} \|h_{T^c}\|_{\ell_1}.$$

Plugging these estimates into (2.5.7) gives

$$\|h_{\bar{T}_1}\|_{\ell_2} \leq \frac{1}{2}\Big(\frac{5}{2}\sqrt{2s}\lambda + r^{-1/2} \|h_{T^c}\|_{\ell_1}\Big). \tag{2.5.12}$$

[2] We note that this method to compare ℓ_1 and ℓ_2 has older roots in approximation theory.

The conclusion is now one step away. Obviously,

$$\begin{aligned}\|h\|_{\ell_2} \le \|h_{\bar{T}_1}\|_{\ell_2} + \sum_{j\ge 2} \|h_{T_j}\|_{\ell_2} &\le \|h_{\bar{T}_1}\|_{\ell_2} + r^{-1/2}\|h_{T^c}\|_{\ell_1} \\ &\le \frac{1}{2}\Big(\frac{5}{2}\sqrt{2s}\lambda + 3r^{-1/2}\|h_{T^c}\|_{\ell_1}\Big),\end{aligned}$$

where the second line follows from (2.5.12). Lemma 2.5.2 completes the proof for the ℓ_2 error. For the ℓ_1 error, note that by the Cauchy-Schwarz inequality

$$\|h\|_{\ell_1} = \|h_T\|_{\ell_1} + \|h_{T^c}\|_{\ell_1} \le \sqrt{s}\|h_T\|_{\ell_2} + \|h_{T^c}\|_{\ell_1} \le \sqrt{s}\|h_{\bar{T}_1}\|_{\ell_2} + \|h_{T^c}\|_{\ell_1}.$$

Combine this with (2.5.12) and Lemma 2.5.2. ■

2.5.2 Proof of Lemma 2.5.2

Since $\hat{x}$ is the minimizer to (2.1.10),

$$\frac{1}{2}\|A\hat{x} - y\|_{\ell_2}^2 + \lambda\|\hat{x}\|_{\ell_1} \le \frac{1}{2}\|Ax - y\|_{\ell_2}^2 + \lambda\|x\|_{\ell_1},$$

which can be massaged into the more convenient form

$$\frac{1}{2}\|Ah\|_{\ell_2}^2 + \lambda\|\hat{x}\|_{\ell_1} \le \langle Ah, z\rangle + \lambda\|x\|_{\ell_1}.$$

Lemma 2.5.3

$$\|\hat{x}\|_{\ell_1} \ge \|x\|_{\ell_1} + \langle h_T, \operatorname{sgn}(x_T)\rangle + \|h_{T^c}\|_{\ell_1} - 2\|x_{T^c}\|_{\ell_1}.$$

Proof We have $\|\hat{x}\|_{\ell_1} = \langle \hat{x}, \operatorname{sgn}(\hat{x})\rangle \ge \langle x_T + h_T, \operatorname{sgn}(x_T)\rangle + \|x_{T^c} + h_{T^c}\|_{\ell_1}$ and the claim follows from the triangle inequality. ■

It follows from this that

$$\frac{1}{2}\|Ah\|_{\ell_2}^2 + \lambda\|h_{T^c}\|_{\ell_1} \le \langle Ah, z\rangle - \lambda\langle h_T, \operatorname{sgn}(x_T)\rangle + 2\lambda\|x_{T^c}\|_{\ell_1}, \tag{2.5.13}$$

and the proof is now a consequence of the two short lemmas below.

Lemma 2.5.4

$$\langle Ah, z\rangle \le \frac{5}{12}s\lambda^2 + \frac{\lambda}{4}\|h_{T^c}\|_{\ell_1}. \tag{2.5.14}$$

Proof The proof is similar to an argument in [33]. Let $P = A_T(A_T^*A_T)^{-1}A_T^*$ be the orthogonal projection onto the range of A_T. Then

$$\begin{aligned}
\langle Ah, z\rangle &= \langle PAh, z\rangle + \langle (I-P)A_{T^c}h_{T^c}, z\rangle \\
&= \langle A_T^*Ah, (A_T^*A_T)^{-1}A_T^*z\rangle + \langle h_{T^c}, A_{T^c}^*(I-P)z\rangle \\
&\le \|A_T^*Ah\|_{\ell_\infty}\|(A_T^*A_T)^{-1}A_T^*z\|_{\ell_1} + \|h_{T^c}\|_{\ell_1}\|A_{T^c}^*(I-P)z\|_{\ell_\infty} \\
&\le \frac{5}{4}\lambda\|(A_T^*A_T)^{-1}A_T^*z\|_{\ell_1} + 2.5\sqrt{\log n}\|h_{T^c}\|_{\ell_1}. \qquad (2.5.15)
\end{aligned}$$

The last line follows from Lemma 2.5.1 and (2.5.4). We now bound the first term, and write

$$\begin{aligned}
\|(A_T^*A_T)^{-1}A_T^*z\|_{\ell_1} &\le \sqrt{s}\|(A_T^*A_T)^{-1}A_T^*z\|_{\ell_2} \\
&\le \frac{4}{3}\sqrt{s}\|A_T^*z\|_{\ell_2} \\
&\le \frac{4}{3}s\|A_T^*z\|_{\ell_\infty} \le \frac{1}{3}s\,\lambda. \qquad (2.5.16)
\end{aligned}$$

The first inequality follows from Cauchy-Schwarz, the second from $\|A_T^*A_T\| \le 4/3$, and the fourth from $\|A^*z\|_{\ell_\infty} \le \lambda/4$. Inequality (2.5.15) establishes the claim. ■

Lemma 2.5.5

$$|\langle h_T, \mathrm{sgn}(x_T)\rangle| \le Cs\lambda + \frac{7}{12}\|h_{T^c}\|_{\ell_1} + \frac{1}{2\lambda}\|Ah\|_{\ell_2}^2. \qquad (2.5.17)$$

Proof Let v be the inexact dual vector, and decompose $\langle h_T, \mathrm{sgn}(x_T)\rangle$ as

$$\begin{aligned}
|\langle h_T, \mathrm{sgn}(x_T)\rangle| &\le |\langle h_T, \mathrm{sgn}(x_T) - v_T\rangle| + |\langle h_T, v_T\rangle| \\
&\le |\langle h_T, \mathrm{sgn}(x_T) - v_T\rangle| + |\langle h, v\rangle| + |\langle h_{T^c}, v_{T^c}\rangle|. \qquad (2.5.18)
\end{aligned}$$

First,

$$|\langle h_T, \mathrm{sgn}(x_T) - v_T\rangle| \le \|h_T\|_{\ell_2}\|\mathrm{sgn}(x_T) - v_T\|_{\ell_2} \le \tfrac{1}{4}\|h_T\|_{\ell_2}.$$

Now

$$\begin{aligned}
\|h_T\|_{\ell_2} \le \left\|(A_T^*A_T)^{-1}\right\|\|A_T^*A_Th_T\|_{\ell_2} &\le \frac{4}{3}\|A_T^*A_Th_T\|_{\ell_2} \\
&\le \frac{4}{3}\|A_T^*Ah\|_{\ell_2} + \frac{4}{3}\|A_T^*A_{T^c}h_{T^c}\|_{\ell_2} \\
&\le \frac{4}{3}\sqrt{s}\|A_T^*Ah\|_{\ell_\infty} + \frac{4}{3}\|h_{T^c}\|_{\ell_1}\max_{j\in T^c}\|A_T^*A_{\{j\}}\|_{\ell_2} \\
&\le \frac{5}{3}\sqrt{s}\lambda + \frac{4}{3}\|h_{T^c}\|_{\ell_1}, \qquad (2.5.19)
\end{aligned}$$

where the last line follows from Lemma 2.5.1 and (2.4.2). Second, it follows from the definition of v that

$$|\langle h_{T^c}, v_{T^c}\rangle| \le \|h_{T^c}\|_{\ell_1}\|v_{T^c}\|_{\ell_\infty} \le \tfrac{1}{4}\|h_{T^c}\|_{\ell_1}.$$

Hence, we established

$$|\langle h_T, \mathrm{sgn}(x_T)\rangle| \le \frac{5}{12}\sqrt{s}\lambda + \frac{7}{12}\|h_{T^c}\|_{\ell_1} + |\langle h, v\rangle|. \tag{2.5.20}$$

Third, we bound $|\langle h, v\rangle|$ by Lemma 2.5.6 below. With the notation of this lemma,

$$|\langle h, v\rangle| = |\langle h, A^*w\rangle| = |\langle Ah, w\rangle| \le \|Ah\|_{\ell_2}\|w\|_{\ell_2} \le C_0\sqrt{s}\|Ah\|_{\ell_2}$$

for some $C_0 > 0$. Since

$$\|Ah\|_{\ell_2}\sqrt{s} \le \frac{\|Ah\|_{\ell_2}^2}{2C_0\lambda} + \frac{C_0 s\lambda}{2},$$

it follows that

$$|\langle h, v\rangle| \le \frac{C_0^2}{2}s\lambda + \frac{1}{2\lambda}\|Ah\|_{\ell_2}^2. \tag{2.5.21}$$

Plugging this into (2.5.20) finishes the proof. ∎

Lemma 2.5.6 *The inexact dual certificate from Section 2.4 is of the form $v = A^*w$ where $\|w\|_{\ell_2} \le C_0\sqrt{s}$ for some positive numerical constant C_0.*

Proof For notational simplicity, assume without loss of generality that the first ℓ batches of rows were those used in constructing the dual vector v (none were thrown out) so that

$$v = \sum_{i=1}^{\ell} \frac{m}{m_i} A_i^* A_{i,T} q_{i-1}.$$

Hence, $v = A^*w$ with $w^* = (w_1^*, \dots, w_\ell^*, 0, \dots, 0)$ and $w_i := \frac{m}{m_i}A_{i,T}q_{i-1}$ so that $\|w\|_{\ell_2}^2 = \sum_{i=1}^{\ell}\|w_i\|_{\ell_2}^2$. We have

$$\begin{aligned}
\tfrac{m}{m_i}\|A_{i,T}\,q_{i-1}\|_{\ell_2}^2 &= \langle \tfrac{m}{m_i}A_{i,T}^*A_{i,T}q_{i-1}, q_{i-1}\rangle \\
&= \langle (\tfrac{m}{m_i}A_{i,T}^*A_{i,T} - I)q_{i-1}, q_{i-1}\rangle + \|q_{i-1}\|_{\ell_2}^2 \\
&\le \|q_i\|_{\ell_2}\|q_{i-1}\|_{\ell_2} + \|q_{i-1}\|_{\ell_2}^2 \\
&\le 2\|q_{i-1}\|_{\ell_2}^2 \\
&\le 2s\prod_{j=1}^{i-1} c_j^2.
\end{aligned} \tag{2.5.22}$$

It follows that

$$\|w\|_{\ell_2}^2 \le 2s \cdot \sum_{i=1}^{\ell}\frac{m}{m_i}\prod_{j=1}^{i-1}c_j^2.$$

Assume that $m \le C(1+\beta)\mu s \log n$ so that m is just large enough to satisfy the requirements of Theorem 2.1.2 (up to a constant). Then recall that $m_i \ge C(1+\beta)\mu s c_i^{-2} \Rightarrow \frac{m}{m_i} \le C c_i^2 \log n$. (If m is much larger, rescale each m_i proportionally to achieve the same ratio.) This gives

$$\|w\|_{\ell_2}^2 \le Cs \log n \sum_{i=1}^{\ell} \prod_{j=1}^{i} c_j^2 \le Cs \sum_{i=1}^{\ell} \prod_{j=2}^{i} c_j^2$$

since $c_1 = (2\sqrt{\log n})^{-1}$. For $i \ge 1$, $\prod_{j=2}^{i} 4^{-(i-1)}$ and the conclusion follows. ∎

2.5.3 Proof of Theorem 2.1.3

Proof Fix s and T as in Section 2.5.1 and assume that (i)–(iv) hold. The proof parallels that for the LASSO; this is why we only sketch the important points and reuse the earlier techniques with minimal extra explanation. We shall repeatedly use the inequality

$$ab \le ca^2/2 + b^2/(2c), \tag{2.5.23}$$

which holds for positive scalars a, b, c. Our first intermediate result is analogous to Lemma 2.5.2.

Lemma 2.5.7 *The error $h = \hat{x} - x$ obeys*

$$\|h_{T^c}\|_{\ell_1} \le C(s\lambda + \|x_{T^c}\|_{\ell_1} + \sqrt{s}\|Ah\|_{\ell_2}).$$

Proof Since x is feasible, $\|\hat{x}\|_{\ell_1} \le \|x\|_{\ell_1}$ and it follows from Lemma 2.5.3 that

$$\|h_{T^c}\|_{\ell_1} \le -\langle h_T, \mathrm{sgn}(x_T)\rangle + 2\|x_{T^c}\|_{\ell_1}. \tag{2.5.24}$$

We bound $|\langle h_T, \mathrm{sgn}(x_T)\rangle|$ in exactly the same way as before, but omitting the last step, and obtain

$$|\langle h_T, \mathrm{sgn}(x_T)\rangle| \le Cs\lambda + \frac{7}{12}\|h_{T^c}\|_{\ell_1} + C\sqrt{s}\|Ah\|_{\ell_2}.$$

This concludes the proof. ∎

The remainder of this section proves Theorem 2.1.3. Observe that $\|A^*Ah\|_{\ell_\infty} \le \frac{5}{4}\lambda$ (Lemma 2.5.1) since the proof is identical (we do not even need to consider subgradients). Partitioning the indices as before, one can repeat the earlier argument leading to (2.5.12). Then combining (2.5.12) with Lemma 2.5.7 gives

$$\|h_{\bar{T}_1}\|_{\ell_2} \le C\sqrt{s}\lambda + Cr^{-1/2}(s\lambda + \|x_{T^c}\|_{\ell_1} + \sqrt{s}\|Ah\|_{\ell_2}). \tag{2.5.25}$$

The term proportional to $\sqrt{s/r}\|Ah\|_{\ell_2}$ in the right-hand side was not present before, and we must develop an upper bound for it. Write

$$\|Ah\|_{\ell_2}^2 = \langle A^*Ah, h\rangle \le \|A^*Ah\|_{\ell_\infty}\|h\|_{\ell_1} \le \frac{5}{4}\lambda(\|h_T\|_{\ell_1} + \|h_{T^c}\|_{\ell_1})$$

and note that (2.5.24) gives

$$\|h_{T^c}\|_{\ell_1} \le \|h_T\|_{\ell_1} + 2\|x_{T^c}\|_{\ell_1}.$$

These last two inequalities yield $\|Ah\|_{\ell_2}^2 \le \frac{5}{2}\lambda(\|h_T\|_{\ell_1} + \|x_{T^c}\|_{\ell_1})$, and since $\sqrt{\lambda\|x_{T^c}\|_{\ell_1}} \le \frac{1}{2}\lambda\sqrt{s} + \frac{1}{2\sqrt{s}}\|x_{T^c}\|_{\ell_1}$ because of (2.5.23), we have

$$\|Ah\|_{\ell_2} \le \sqrt{\tfrac{5}{2}\lambda}(\sqrt{\|h_T\|_{\ell_1}} + \sqrt{\|x_{T^c}\|_{\ell_1}}) \le \sqrt{\tfrac{5}{2}}\Big(\sqrt{\lambda\|h_T\|_{\ell_1}} + \tfrac{1}{2}\lambda\sqrt{s} + \tfrac{1}{2\sqrt{s}}\|x_{T^c}\|_{\ell_1}\Big).$$

In short,

$$\|h_{\bar{T}_1}\|_{\ell_2} \le C\Big(\sqrt{s}\lambda + r^{-1/2}\Big(s\lambda + \|x_{T^c}\|_{\ell_1} + \sqrt{s\lambda\|h_T\|_{\ell_1}}\Big)\Big).$$

The extra term on the right-hand side has been transmuted into $C\sqrt{\frac{s}{r}\lambda\|h_T\|_{\ell_1}}$, which may be bounded via (2.5.23) as

$$C\sqrt{\frac{s}{r}\lambda\|h_T\|_{\ell_1}} \le C^2\frac{s}{r}\sqrt{s}\lambda + \frac{1}{2\sqrt{s}}\|h_T\|_{\ell_1} \le C^2\frac{s}{r}\sqrt{s}\lambda + \frac{1}{2}\|h_T\|_{\ell_2}.$$

Since $\|h_T\|_{\ell_2} \le \|h_{\bar{T}_1}\|_{\ell_2}$, we have

$$\|h_{\bar{T}_1}\|_{\ell_2} \le C\left(1 + \sqrt{\frac{s}{r}} + \frac{s}{r}\right)\sqrt{s}\lambda + C\,\frac{\|x_{T^c}\|_{\ell_1}}{\sqrt{r}}.$$

The remaining steps are the same as those in the proof for the LASSO. ∎

2.6 Proof of Theorem 2.3.7 (the weak RIP)

Our proof uses some the results and techniques of [133] and [135]. Recall that A is a matrix with rows drawn independently from a probability distribution F obeying the isotropy and incoherence conditions, and that we wish to show that for any fixed $0 \le \delta < 1$,

$$(1-\delta)\|v\|_{\ell_2}^2 \le \|Av\|_{\ell_2}^2 \le (1+\delta)\|v\|_{\ell_2}^2.$$

These inequalities should hold with high probability, uniformly over all vectors v obeying $\mathrm{supp}(v) \subset T \cup R$ where T is fixed, R may vary, and

$$|T| \le c\frac{m}{\mu \log m}, \qquad |R| \le c\frac{m}{\mu \log n \log m \log^2(|R|\mu)}$$
$$\Leftrightarrow \qquad m \ge C|T|\mu \log(|T|\mu), \qquad m \ge C|R| \log n \log(r\mu \log n) \log^2(|R|\mu).$$

To express this in another way, set

$$X := \sup_{v \in V} |\|Av\|_{\ell_2}^2 - \|v\|_{\ell_2}^2|,$$

where

$$V = \{v : \|v\|_{\ell_2} = 1, \mathrm{supp}(v) \subset T \cup R, |R| \le r, T \cap R = \varnothing\}. \tag{2.6.1}$$

In words, v is a unit-normed vector supported on $T \cup R$ where T is fixed of cardinality $s \le cm/(\mu \log m)$, and R is any set disjoint from T of cardinality at most $r \le cm/(\mu \log n \log m \log^2(r\mu))$. We wish to show that $X \le \delta$ with high probability. We will first bound this random variable in expectation and then show that it is unlikely to be much larger than its expectation. The bound in expectation is contained in the following lemma.

Lemma 2.6.1 *Fix $\epsilon > 0$. Suppose $m \ge C\,\mu[s \log m \vee r \log n \log m \log^2(r\mu)]$, where C is a constant only depending on ϵ. Then*

$$\mathbb{E}\, X \le \epsilon.$$

To begin the proof, note that for any v with $\mathrm{supp}(v) \subset T \cup R$, we have

$$\|Av\|_{\ell_2}^2 = \|A_T v_T\|_{\ell_2}^2 + \|A_R v_R\|_{\ell_2}^2 + 2\langle v_T, A_T^* A_R v_R\rangle.$$

The first two terms are easily dealt with using prior results. To be sure, under the conditions of Lemma 2.6.1, a slight modification of the proof of Theorem 3.4 in [135] gives[3]

$$\mathbb{E} \sup_{v_R : |R| \le r} |\|A_R v_R\|_{\ell_2}^2 - \|v_R\|_{\ell_2}^2| \le \frac{\epsilon}{4}\|v_R\|_{\ell_2}^2. \tag{2.6.2}$$

Next, it follows from [134], or the matrix Bernstein inequality in Estimate 1, that

$$\mathbb{E} \sup_{v_T} |\|A_T v_T\|_{\ell_2}^2 - \|v_T\|_{\ell_2}^2| \le \frac{\epsilon}{4}\|v_T\|_{\ell_2}^2. \tag{2.6.3}$$

[3]Rudelson and Vershynin consider a slightly different model but the proof in [135] extends to our model with hardly any adjustments.

Thus, to prove Lemma 2.6.1, it suffices to prove that

$$\mathbb{E} \max_R \|A_R^* A_T\| \le \epsilon/4.$$

This is the content of the following theorem.

Theorem 2.6.2 *Under the assumptions of Lemma 2.6.1, we have*

$$\mathbb{E} \max_R \|A_R^* A_T\| \le C \left(\sqrt{\frac{s\mu \log m}{m}} + \sqrt{\frac{r\mu \log n \log m \log^2(r\mu)}{m}} \right). \tag{2.6.4}$$

Put differently, the theorem develops a bound on

$$\mathbb{E} \max_{(x,y)\in B_T \times D} \frac{1}{m} \sum_{i=1}^m \langle a_i, x\rangle \langle a_i, y\rangle \tag{2.6.5}$$

in which

$$B_T := \{x : \|x\|_{\ell_2} \le 1, \operatorname{supp}(x) \subset T\},$$
$$D := \{y : \|y\|_{\ell_2} \le 1, \operatorname{supp}(y) \cap T = \varnothing, |\operatorname{supp}(y)| \le r\}.$$

By symmetrization followed by a comparison principle—both of which follow by Jensen's inequality (see [100, Lemma 6.3] followed by [100, inequality (4.8)])—(2.6.5) is less or equal to a numerical constant times

$$\mathbb{E} \max_{(x,y)\in B_T \times D} \frac{1}{m} \sum_{i=1}^m g_i \langle a_i, x\rangle \langle a_i, y\rangle,$$

where the g_i's are independent $\mathcal{N}(0,1)$ random variables. The main estimate is a bound on the conditional expectation of the right-hand side; that is, holding the vectors a_i fixed.

Lemma 2.6.3 (Main lemma) *Fix vectors $\{a_i\}_{i=1}^m$ and let*

$$R_1 := \max_{x\in B_T} \frac{1}{m} \sum_{i=1}^m \langle a_i, x\rangle^2, \qquad R_2 := \max_{y\in D} \frac{1}{m} \sum_{i=1}^m \langle a_i, y\rangle^2.$$

Suppose $m \ge C\,\mu\,[s \log m \vee r \log n \log m \log^2(r\mu)]$.

Then

$$\mathbb{E} \max_{(x,y)\in B_T \times D} \frac{1}{m} \sum_{i=1}^m g_i \langle a_i, x\rangle \langle a_i, y\rangle \le C \left(\sqrt{\frac{(1+R_2)(1+R_1)s\mu \log m}{m}} + \sqrt{\frac{(1+R_1)r\mu \log n \log m \log^2(r\mu)}{m}} \right).$$

Proof [Theorem 2.6.2] We have $\sqrt{(1+R_2)(1+R_1)} \le \frac{1}{2}(2 + R_1 + R_2)$. Now, under the assumptions of the theorem, it follows from the results in [135] that $\mathbb{E}\, R_2 \le C$. Likewise, the results in [134] and give $\mathbb{E}\, R_1 \le C$, and thus $\mathbb{E}\sqrt{1+R_1} \le C$. (These inequalities were also noted, in a different form, in

(2.6.3) and (2.6.2)). Hence, Lemma 2.6.3 implies

$$\mathbb{E}\max_R \|A_R^* A_T\| \le C\left(\sqrt{\frac{s\mu\log m}{m}} + \sqrt{\frac{r\mu\log n\log m\log^2(r\mu)}{m}}\right)$$

where the expectation is taken over the randomly selected rows $\{a_i^*\}$. ■

2.6.1 Proof of Lemma 2.6.3

We need to develop a bound about the expected maximum of a Gaussian process, namely,

$$\mathbb{E}\max_{(x,y)\in B_T\times D} F(x,y),$$

where

$$F(x,y) \coloneqq \sum_{i=1}^m g_i\langle a_i,x\rangle\langle a_i,y\rangle.$$

We shall do this by means of a certain version of the majorizing measure theorem, which may be found in [133] and is attributed to Talagrand. It is derived as a combination of a majorizing measure construction of Talagrand (combine Theorem 4.1 with Propositions 2.3 and 4.4 in [145]) and the majorizing measure theorem of Fernique [100].

From now on, (M,d) is a metric space and $B(t,\epsilon)$ is the ball of center t and radius ϵ under the metric d.

Theorem 2.6.4 (Majorizing measure theorem) *Let $(X_t)_{t\in M}$ be a collection of zero-mean random variables obeying the subgaussian tail estimate*

$$\mathbb{P}(|X_t - X_{t'}| > u) \le \exp\left(-c\frac{u^2}{d^2(t,t')}\right), \tag{2.6.6}$$

for all $u > 0$. Fix $\rho > 1$ and let k_0 be an integer so that the diameter of M is less than ρ^{-k_0}. Suppose there exist $\sigma > 0$ and a sequence of functions $\{\varphi_k\}_{k=k_0}^\infty$, $\varphi_k : M \to \mathbb{R}^+$, with the following two properties: 1) the sequence is uniformly bounded by a constant depending only on ρ; 2) for each k and for any $t \in M$ and any points $t_1,\cdots,t_{\tilde{N}} \in B(t,\rho^{-k})$ with mutual distances at least ρ^{-k-1}, we have

$$\max_{j=1,\cdots,\tilde{N}} \varphi_{k+2}(t_j) \ge \varphi_k(s) + \sigma\rho^{-k}\sqrt{\log\tilde{N}}. \tag{2.6.7}$$

Then

$$\mathbb{E}\sup_{t\in M} X_t \le C(\rho)\cdot\sigma^{-1}. \tag{2.6.8}$$

To apply this theorem, we begin by bounding the variance between increments in order to

ascertain the metric we need to use (the induced metric). We compute

$$\begin{aligned} d((x,y),(x',y')) &:= \sqrt{\mathrm{Var}(F(x,y)-F(x',y'))} \\ &= \sqrt{\sum_{i=1}^{m}\Big(\langle a_i,x\rangle\langle a_i,y\rangle - \langle a_i,x'\rangle\langle a_i,y'\rangle\Big)^2} \end{aligned}$$

Before continuing, we record two useful lemmas for bounding $\tilde{N}$. Here and below, $N(M,d,\epsilon)$ is the covering number of M in the metric d with balls of radius ϵ.

Lemma 2.6.5 (Packing number bound) *Let $t_1, t_2, \cdots, t_{\tilde{N}} \in M$ be points with mutual distances at least 2ϵ under the metric d. Then*

$$\tilde{N} \le N(M,d,\epsilon).$$

This is a standard result proved by creating an injective mapping from the points $\{t_j\}$ to those in the cover set (map each t_j to the nearest point in the cover).

The next lemma is a standard tool used to obtain bounds on covering numbers, see [122] and [20] for a more general statement.

Lemma 2.6.6 (Dual Sudakov minorization) *Let B_{ℓ_2} be the unit ℓ_2 ball in $\mathbb{R}^d$, and let $\|\cdot\|$ be a norm. Let $z \in \mathbb{R}^d$ be a Gaussian vector with independent $\mathcal{N}(0,1)$ entries. Then there is a numerical constant $C > 0$ such that*

$$\sqrt{\log N(B_{\ell_2}, \|\cdot\|, \epsilon)} \le \frac{C}{\epsilon}\sqrt{\mathbb{E}\,\|z\|^2}.$$

We now invoke the majorizing measure theorem to prove Lemma 2.6.3. We start by bounding the diameter of $B_T \times D$ under the metric d. By Cauchy-Schwarz, for any $y \in D$, we have $|\langle a_i, y\rangle| \le \sqrt{r\mu}$. This may be used to derive the following bound on the diameter.

$$d((x,y),(x',y')) \le 2\sqrt{2r\mu m R_1}.$$

Thus set k_0 to be the largest integer such that

$$\rho^{-k_0} \ge 2\sqrt{2r\mu m R_1}.$$

We also set k_1—whose meaning will become apparent in a moment—to be the largest integer such that

$$\rho^{-k_1} \ge 2\sqrt{2mR_1}$$

We now define φ_k on coarse and fine scales. In what follows, we may take $\rho = 8$ so that $C(\rho)$

(2.6.8) is an absolute constant.

Coarse scales: for $k = k_0, k_0+1, \dots, k_1 - 1$,

$$\varphi_k(x,y) := \min\{\|u\|_{\ell_2}^2 : (u,v) \in B((x,y), 2\rho^{-k})\} + \frac{k-k_0}{\log(r\mu)}.$$

Fine scales: for $k \geq k_1$, φ_k is a constant function given by

$$\varphi_k(x) := 3\rho\sigma \int_{\rho^{-k}}^{\rho^{-k_1}} \sqrt{\log N(B_T \times D, d, \epsilon)}d\epsilon + 2.$$

Last, set

$$\sigma^{-1} := C\sqrt{m}\left(\sqrt{(1+R_2)(1+R_1)s\mu \log m} + \sqrt{(1+R_1)r\mu \log n \log m \log^2 r}\right).$$

Our definition of φ_k is closely related to—and inspired by—the functions defined in [133]. We need to show that these functions are uniformly bounded and obey (2.6.7) for all k. We begin by verifying these properties for fine scale elements as this is the less subtle calculation.

2.6.2 Fine scale: $k \geq k_1$

To show that (2.6.7) holds, observe that,

$$\begin{aligned}
\varphi_{k+2} - \varphi_k &= 3\sigma\rho \int_{\rho^{-(k+2)}}^{\rho^{-k}} \sqrt{\log N(B_T \times D, d, \epsilon)}d\epsilon \\
&\geq 3\sigma\rho \int_{\rho^{-(k+2)}}^{\frac{1}{2}\rho^{-(k+1)}} \sqrt{\log N(B_T \times D, d, \epsilon)}d\epsilon \\
&\geq 3\sigma\rho\Big(\frac{1}{2}\rho^{-(k+1)} - \rho^{-(k+2)}\Big)\sqrt{\log N\Big(B_T \times D, d, \frac{1}{2}\rho^{-(k+1)}\Big)} \\
&\geq \sigma\rho^{-k}\sqrt{\log \tilde{N}}.
\end{aligned}$$

The last line follows from $\rho \geq 6$ and the packing number bound (Lemma 2.6.5). Note that this same calculation holds when $k = k_1 - 1, k_1 - 2$ because for $k \leq k_1 - 1$, $\varphi_k \leq 3$ (see Section 2.6.3).

We now show that φ_k is bounded. Since

$$\varphi_k \leq 3\rho\sigma \int_0^{\rho^{-k_1}} \sqrt{\log N(B_T \times D, d, \epsilon)}d\epsilon + 3 \leq 3\rho\sigma \int_0^{\rho\sqrt{8mR_1}} \sqrt{\log N(B_T \times D, d, \epsilon)}d\epsilon + 3 \qquad (2.6.9)$$

it suffices to show that the right-hand side is bounded. This follows from crude upper bounds on

the covering number. Indeed, observe that

$$\begin{aligned} d((x,y),(x',y')) &\le \sqrt{2\sum_{i=1}^{m}\langle a_i, x-x'\rangle^2\langle a_i, y\rangle^2} + \sqrt{2\sum_{i=1}^{m}\langle a_i, x'\rangle^2\langle a_i, y-y'\rangle^2} \\ &\le \sqrt{2mR_2}\max_{1\le i\le m}|\langle a_i, x-x'\rangle| + \sqrt{2mR_1}\max_{1\le i\le m}|\langle a_i, y-y'\rangle| \\ &\le \sqrt{2mR_2 s\mu}\|x-x'\|_{\ell_2} + \sqrt{2mR_1 r\mu}\|y-y'\|_{\ell_2} \\ &\le m^{3/2}\|x-x'\|_{\ell_2} + m^{3/2}\|y-y'\|_{\ell_2}. \end{aligned}$$

Thus,

$$\begin{aligned} N(B_T\times D, d, \epsilon) &\le N\left(B, \|\cdot\|_{\ell_2}, \frac{\epsilon}{m^{3/2}}\right)\cdot N\left(D, \|\cdot\|_{\ell_2}, \frac{\epsilon}{m^{3/2}}\right) \\ &\le \left(\frac{2m^{3/2}}{\epsilon}\right)^s \cdot \binom{n}{r}\left(\frac{2m^{3/2}}{\epsilon}\right)^r. \end{aligned}$$

The second line comes from the standard volumetric estimate $N(B, \|\cdot\|_{\ell_2}, \epsilon) \le \left(\frac{3}{\epsilon}\right)^s$ for $\epsilon \le 1$. The factor $\binom{n}{r}$ arises from decomposing D as the union of $\binom{n-s}{r} \le \binom{n}{r}$ sets of the same form as B_T, but with support size bounded by r. Now, in order to bound the last inequality, we further write $\binom{n}{r} \le n^r$. Plugging this in, we obtain

$$\sqrt{\log N(B_T\times D, d, \epsilon)} \le C(\sqrt{s\log(m/\epsilon)} + \sqrt{r\log(mn/\epsilon)}).$$

To conclude, bounding the integration gives

$$\int_0^{\sqrt{8R_1 m}} \sqrt{s\log(m/\epsilon)} + \sqrt{r\log(mn/\epsilon)}d\epsilon \le C\sqrt{8R_1 ms}(\sqrt{\log(m)}+1) + \sqrt{8R_1 mr}(\sqrt{\log(mn)}+1),$$

which establishes the claim since the right-hand side is dominated by σ^{-1}.

2.6.3 Coarse scale: $k \le 0$

This section contains the crucial estimates, which must be developed very carefully. To show that φ_k is bounded, observe that by definition, $\rho^{k_1-1-k_0} \le \sqrt{r\mu}$, and thus $(k_1 - 1 - k_0) \le \frac{1}{2}\log(r\mu)$. It follows that $\varphi_k \le 2$.

Next, we show that the more subtle bound (2.6.7) holds. Let $\{(x_i, y_i)\}$ be the points in the definition of the majorizing measure theorem with mutual distances at least ρ^{-k-1}, so that $\tilde{N} = |\{(x_i, y_i)\}|$. Let (z_x, v_x) be the point of $B((x,y), \rho^{-k})$ which minimizes the value of $\|z_x\|_{\ell_2}^2$. Let (z_j, v_j) be the similar points of $B((x_i, y_i), \rho^{-k-2})$. Finally, introduce the pivotal quantity

$$\theta := \max_{1\le j\le \tilde{N}} \|z_j\|_{\ell_2}^2 - \|z_x\|_{\ell_2}^2.$$

We must show that

$$\rho^{-k}\sigma\sqrt{\log \tilde{N}} \le \max_{1\le j\le \tilde{N}} \varphi_{k+2}(x_j, y_j) - \varphi_k(x, y) = \theta + 2/\log(r\mu).$$

In order to bound $\tilde{N}$, we consider the points $\{z_j, v_j\}$ and note that $\tilde{N} = |\{z_j, v_j\}|$.

We shall need two key properties of the points $\{z_j, v_j\}$. First, these points are well separated. Indeed, the triangle inequality, gives for $i \neq j$

$$\begin{aligned} d((z_i, v_i), (z_j, v_j)) &\ge d((x_i, y_i), (x_j, y_j)) - d((x_i, y_i), (z_i, v_i)) - d((x_j, y_j), (z_j, v_j)) \\ &\ge \rho^{-k-1} - 4\rho^{-k-2} \\ &\ge \frac{1}{2}\rho^{-k-1} \end{aligned}$$

provided that $\rho \ge 8$. Second, each point (z_j, v_j) is close to (x, y) in the sense that

$$d((x, y), (z_j, v_j)) \le d((x, y), (x_j, y_j)) + d((x_j, y_j), (z_j, v_j)) \le \rho^{-k} + 2\rho^{-k-2} \le 2\rho^{-k}$$

provided that $\rho \ge 2$. In other words $(z_j, v_j) \in B((x, y), 2\rho^{-k})$, and thus $\|z_j\|_{\ell_2} \ge \|z_x\|_{\ell_2}$.

Now, the benefit of the special construction of φ_k on the coarse scale is that the size of θ restricts the space that $\{z_j\}$ can inhabit. To demonstrate this, since $B((x, y), 2\rho^{-k})$ is convex, $(\frac{z_x+z_j}{2}, \frac{v_x+v_j}{2}) \in B((x, y), 2\rho^{-k})$. Now combine $\|\frac{z_x+z_j}{2}\|_{\ell_2} \ge \|z_x\|_{\ell_2}$ with $\|z_j\|_{\ell_2} \ge \|z_x\|_{\ell_2}$ to give

$$\|\frac{z_j - z_x}{2}\|_{\ell_2}^2 = \frac{1}{2}\|z_j\|_{\ell_2}^2 + \frac{1}{2}\|z_x\|_{\ell_2}^2 - \|\frac{z_j + z_x}{2}\|_{\ell_2}^2 \le \|z_j\|_{\ell_2}^2 - \|z_x\|_{\ell_2}^2 \le \theta.$$

Hence,

$$\|z_j - z_x\|_{\ell_2} \le 2\sqrt{\theta}.$$

Combined with Lemma 2.6.5, we obtain

$$\tilde{N} \le N(B_T(z_x, 2\sqrt{\theta}) \times D, d, \rho^{-k-1}/4), \tag{2.6.10}$$

where $B_T(z_x, 2\sqrt{\theta}) := \{x : \mathrm{supp}(x) \subset T, \|x\|_{\ell_2} \le 1, \|x - z_x\|_{\ell_2} \le 2\sqrt{\theta}\}$.

We now bound the metric, d, but we will do so more carefully than in the fine scale. We have

$$\begin{aligned} d((x, y), (x', y')) &\le d((x, y), (x', y)) + d((x', y), (x', y')) \\ &\le \|x - x'\|_y + \sqrt{mR_1}\|y - y'\|_X \end{aligned}$$

where

$$\bullet \ \|y\|_X \coloneqq \max_{1\le i\le m} |\langle a_i, x\rangle|$$
$$\bullet \ \|x\|_y \coloneqq \sqrt{\sum_{i=1}^{m} \langle a_i, x\rangle^2 \langle a_i, y\rangle^2}$$

Note that the norm $\|\cdot\|_y$ varies with y. Also note that they may be pseudonorms, but this makes no difference to the proof. All of the utilized lemmas and theorems generalize to pseudonorms.

We cover $B_T(u, 2\sqrt{\theta}) \times D_S$ to precision ϵ by covering D_S to precision $\epsilon/2$ under the norm $\sqrt{mR_1}\|\cdot\|_X$, and then, for each y contained in this cover, we cover $B_T(u, 2\sqrt{\theta})$ to precision $\epsilon/2$ under the norm $\|\cdot\|_y$.

Thus,

$$\begin{aligned}\sqrt{\log \tilde{N}} &\le \sqrt{\log N(D_S, \sqrt{mR_1}, \|\cdot\|_X, \epsilon/2)} + \max_{y'\epsilon D_S} \sqrt{\log N(B_T(u, 2\sqrt{\theta}), \|\cdot\|_y, \epsilon/2)} \\ &\coloneqq K_1 + \max_{y\epsilon D} K_2(y) \end{aligned} \tag{2.6.11}$$

Using Lemma 3.7 from [135] (which follows from an argument of [45]) gives

$$K_1 \le C\sqrt{mR_1 r\mu \log n \log m}\rho^k \tag{2.6.12}$$

(We do not reproduce the proof of this lemma here, but encourage the interested reader to explore the extremely clever, short, arguments used).

Now we bound $K_2(y)$. For y fixed, using dual Sudakov minorization (Lemma 2.6.6) and Jensen's inequality, we have

$$K_2(y) \le C\frac{\sqrt{\theta}}{\epsilon}\sqrt{\mathbb{E}\sum_{i=1}^{m} \langle a_i, z_T\rangle^2 \langle a_i, y\rangle^2} \tag{2.6.13}$$

where z is a guassian vector with standard normal entries. Note that $\mathbb{E}\langle a_i, z_T\rangle^2 = \|a_{i,T}\|_{\ell_2}^2 \le s\mu$, and thus

$$K_2(y) \le C\frac{\sqrt{\theta}}{\epsilon}\sqrt{\mu s m R_2} \le C\sqrt{\mu s m R_2}\rho^k\sqrt{\theta}. \tag{2.6.14}$$

Plug in the covering number bounds (i.e., plug in (2.6.12) and (2.6.14) into (2.6.11)) to give

$$\sqrt{\log \tilde{N}} \le C(\sqrt{mR_1 ru \log n \log m}\rho^k + \sqrt{\mu s m R_2}\rho^k\sqrt{\theta}).$$

Now, plug in $\sqrt{\theta} \le \theta\sqrt{\log(r\mu)} + 1/\sqrt{\log(r\mu)}$, along with the definition of σ, to give

$$\rho^{-k}\sigma\sqrt{\log \tilde{N}} \le \frac{2}{\log(r\mu)} + \theta$$

as desired, thus proving Theorem 2.6.2.

2.6.4 Concentration around the mean

We have now proved that $\mathbb{E}\, X \leq \epsilon$ for any $\epsilon > 0$ provided that $m \geq C_\epsilon\, \mu\, [s \log m \vee r \log n \log m \log^2(r\mu)]$. This already shows that for any fixed $\delta > 0$,

$$\mathbb{P}(X > \delta) \leq \frac{\epsilon}{\delta}$$

and so taking ϵ to be a small fraction of δ gives a first crude bound. However, we wish to show that if $m \geq C\mu\,\beta\, [s \log m \vee r \log n \log m \log^2(r\mu)]$ then the probability of 'failure' decreases as $e^{-\beta}$. This can be proved using a theorem of [135] which in turn is a combination of Theorem 6.17 and inequality (6.19) of [100]. We restate this theorem below.

Theorem 2.6.7 *Let $Y_1, \cdots, Y_m$ be independent symmetric random variables taking values in some Banach space. Assume that $\|Y_j\| \leq R$ for all j and for some norm $\|\cdot\|$. Then for any integer $\ell \geq q$, and any $t > 0$, the random variable*

$$Z := \left\| \sum_{j=1}^{m} Y_j \right\|$$

obeys

$$\mathbb{P}(Z \geq 8q\, \mathbb{E}\, Z + 2R\ell + t) \leq \left(\frac{C}{q}\right)^{\ell} + 2\exp\left(-\frac{t^2}{256q(\mathbb{E}\, Z)^2}\right).$$

In our setup, we work with a norm on positive semidefinite matrices given by

$$\|M\| := \sup_{v \in V}\ v^* M v,$$

where V is given by (2.6.1). The rest of the details of the proof of concentration around the mean follows exactly as in the steps of [135, pages 11–12] and so we do not repeat them, but encourage the interested reader to check [135]. This is the final step in proving Theorem 2.3.7.

2.7 Stochastic Incoherence

In Sections 2–4, we have assumed that the coherence bound holds deterministically, and it is now time to prove our more general statement; that is to say, we need to extend the proof to the case where it holds stochastically. We propose a simple strategy: condition on the (likely) event that each row has 'small' entries, as to recreate the case of deterministic coherence (on this event). Outside of this event, we give no guarantees, but this is of little consequence because we will require the event to hold with probability at least $1 - 1/n$. A difficulty arises because the conditional distribution of the rows no longer obeys the isotropy condition (although the rows are still independent). Fortunately,

this conditional distribution obeys a *near isotropy condition*, and all of our results can be reproved using this condition instead. In particular, all of our theorems follow (with adjustments to the absolute constants involved) from the following two conditions on the distribution of the rows:

$$\begin{aligned} \|\mathbb{E}\, aa^* - I\| &\le 1/(8\sqrt{n}) && \text{(near isotropy)} \\ \max_{1\le t\le n} \|a[t]\|_{\ell_2}^2 &\le \mu && \text{(deterministic coherence).} \end{aligned} \tag{2.7.1}$$

We first illustrate how to use near isotropy to prove our results. There are several results that need to be reproved, but they are all adjusted using the same principle, so to save space we just prove that a slight variation on Lemma 2.3.1 still holds when requiring near isotropy, and leave the rest of the analogous calculations to the interested reader.

Set $W := \mathbb{E}\, aa^*$ and let $W_{T,T}$ be the restriction of W to rows and columns in T. We first show that

$$\mathbb{P}(\|A_T^* A_T - W_{T,T}\| \ge \delta) \le 2s \exp\left(-\frac{m}{\mu(s+1)} \frac{\delta^2}{2+2\delta/3}\right). \tag{2.7.2}$$

To prove this bound, we use the matrix Bernstein inequality of Section 2.3.1, and also follow the framework of the calculations of Section 2.3.1. Thus, we skim the steps. To begin, decompose $A_T^* A_T - W_{T,T}$ as follows:

$$m(A_T^* A_T - W_{T,T}) = \sum_{k=1}^m (a_{k,T} a_{k,T}^* - W_{T,T}) := \sum_{k=1}^m X_k.$$

We have $\mathbb{E}\, X_k = 0$ and $\|X_k\| \le \left\|a_{k,T} a_{k,T}^* - I\right\| + \|I - W_{T,T}\| \le s\mu + \frac{1}{8\sqrt{n}} \le (s+1)\mu := B$. Also, the total variance obeys

$$\|\mathbb{E}\, X_k\|^2 \le \left\|\mathbb{E}(a_{k,T} a_{k,T}^*)^2\right\| \le s\mu \left\|\mathbb{E}\, a_{k,T} a_{k,T}^*\right\| = s\mu \|W_{T,T}\| \le s\mu(1 + \tfrac{1}{8\sqrt{n}}) \le (s+1)\mu.$$

Thus, $\sigma^2 \le m(s+1)\mu$, and (2.7.2) follows from the matrix Bernstein inequality.

Now, it follows from $\|W_{T,T} - I\| \le \|W - I\| \le \frac{1}{8\sqrt{n}}$ that

$$\mathbb{P}\Big(\|A_T^* A_T - I\| \ge \tfrac{1}{8\sqrt{n}} + \delta\Big) \le 2s \exp\left(-\frac{m}{\mu(s+1)} \frac{\delta^2}{2+2\delta/3}\right).$$

In the course of the proofs of Theorems 2.1.1 and 2.1.2 we require $\|A_T^* A_T - I\| \le 1/2$ for noiseless results and $\|A_T^* A_T - I\| \le 1/4$ for noisy results. This can be achieved under the near isotropy condition by increasing the required number of measurements by a tiny bit. In fact, when proving the analogous version of Lemma 2.3.1, one could weaken the near isotropy condition and instead require $\|\mathbb{E}\, aa^* - I\| \le 1/8$, for example. However, in extending some of the other calculations to work with the near isometry condition—such as (2.4.10)—the factor of $\sqrt{n}$ (or at least $\sqrt{s}$) in the denominator appears necessary; this seems to be an artifact of the method of proof, namely, the

golfing scheme. It is our conjecture that all of our results could be established with the weaker requirement $\|\mathbb{E}\, aa^* - I\| \le \epsilon$ for some fixed positive constant ϵ.

We now describe the details concerning the conditioning on rows having small entries. Fix the coherence bound μ and let

$$E_k = \left\{ \max_{1\le t\le n} |a_k[t]|^2 \le \mu \right\} \qquad \text{and} \qquad G = \cap_{1\le k\le m}\, E_k.$$

Thus G is the 'good' event (G is for good) on which $\max_{1\le t\le n} |a_k[t]|^2 \le \mu$ for all k. By the union bound, $\mathbb{P}(G^c) \le m\,\mathbb{P}(E_1^c)$. We wish for $\mathbb{P}(G^c)$ to be bounded by $1/n$, and so we require μ to be large enough so that $\mathbb{P}(E_1^c) \le (mn)^{-1}$.

Next we describe how conditioning on the event G induces the near isometry condition. Because of the independence of the rows of A, we may just consider the conditional distribution of a_1 given E_1. Drop the subindex for simplicity and write

$$I = \mathbb{E}[aa^*] = \mathbb{E}[aa^*\mathbb{1}_E] + \mathbb{E}[aa^*\mathbb{1}_{E^c}] = \mathbb{E}[aa^*|E]\,\mathbb{P}(E) + \mathbb{E}[aa^*\mathbb{1}_{E^c}].$$

Thus,

$$\|\mathbb{E}[aa^*|E] - I\| \cdot \mathbb{P}(E) = \|(1-\mathbb{P}(E))I - \mathbb{E}[aa^*\mathbb{1}_{E^c}]\| \le \mathbb{P}(E^c) + \|\mathbb{E}[aa^*\mathbb{1}_{E^c}]\|\,. \tag{2.7.3}$$

We now bound $\|\mathbb{E}[aa^*\mathbb{1}_{E^c}]\|$. By Jensen's inequality (which is a crude, but still fruitful, bound here),

$$\|\mathbb{E}[aa^*\mathbb{1}_{E^c}]\| \le \mathbb{E}[\|aa^*\mathbb{1}_{E^c}\|] = \mathbb{E}[\|a\|_{\ell_2}^2\mathbb{1}_{E^c}], \tag{2.7.4}$$

and, therefore,

$$\|\mathbb{E}[aa^*|E] - I\| \le \frac{1}{1-\mathbb{P}(E^c)}\left(\mathbb{P}(E^c) + \mathbb{E}[\|a\|_{\ell_2}^2\mathbb{1}_{E^c}]\right).$$

Combine this with the requirement that $\mathbb{P}(E^c) \le (mn)^{-1}$ to give

$$\|\mathbb{E}[aa^*|E] - I\| \le \frac{19}{20}\left(\frac{1}{20\sqrt{n}} + \mathbb{E}[\|a\|_{\ell_2}^2\mathbb{1}_{E^c}]\right)$$

as long as $m\sqrt{n} \ge 20$. It now follows that in order to ensure near isotropy, it is sufficient that

$$\mathbb{E}[\|a\|_{\ell_2}^2\mathbb{1}_{E^c}] \le \frac{1}{20\sqrt{n}}.$$

It may be helpful to note a simple way to bound the left-hand side above. If $f(t)$ is such that

$$\mathbb{P}\left(\max_{1\le t\le n} |a[t]|^2 \ge t\right) \le f(t),$$

then a straightforward calculation shows that

$$\mathbb{E}[\|a\|_{\ell_2}^2 \mathbb{1}_{E^c}] \leq n\mu f(\mu) + n \int_{\mu}^{\infty} f(t)dt.$$

2.8 Discussion

This chapter developed a general and accessible theory of compressive sensing, in which sensing vectors are drawn independently at random from a probability distribution. In addition to establishing a general framework, we showed that nearly sparse signals could be accurately recovered from a small number of noisy compressive samples by means of tractable convex optimization. For example, s-sparse signals can be recovered accurately from about $s \log n$ DFT coefficients corrupted by noise. Our analysis shows that stable recovery is possible from a minimal number of samples, and improves on previously known results. This improvement comes from novel stability arguments, which do not require the restricted isometry property to hold.

We have seen that the isotropy condition is not really necessary, and it would be interesting to know the extent in which it can be relaxed. In particular, for which values of α and β obeying $\alpha I \preceq \mathbb{E}\, aa^* \preceq \beta I$ would our results continue to hold? Also, we have assumed that the sensing vectors are sampled independently at random, and although the main idea in compressive sensing is to use randomness as a sensing mechanism, it would be interesting to know how the results would change if one were to introduce some correlations.

Chapter 3

Sparse approximation and model selection

3.1 Introduction

One of the most common problems in statistics is to estimate a mean response $X\beta$ from the data $y = (y_1, y_2, \ldots, y_n)$ and the linear model

$$y = X\beta + \sigma z, \tag{3.1.1}$$

where X is an $n \times p$ matrix of explanatory variables, β is a p-dimensional parameter of interest, and $z \sim \mathcal{N}(0, I)$ is a Gaussian error term. (Gaussian errors are chosen for simplicity, but our results and methods can easily accommodate other types of distributions.) We measure the performance of any estimator $X\hat{\beta}$ with the usual squared Euclidean distance $\|X\beta - X\hat{\beta}\|_{\ell_2}^2$, or with the mean-squared error which is simply the expected value of this quantity.

In this chapter and although this is not a restriction, we are primarily interested in situations in which there are as many or more explanatory variables than observations—the so-called and now widely popular '$p > n$' setup. In such circumstances, however, it is often the case that a relatively small number of variables have substantial explanatory power so that to achieve accurate estimation, one needs to select the 'right' variables and determine which components β_i are not equal to zero. A standard approach is to find $\hat{\beta}$ by solving

$$\min_{b \in \mathbb{R}^p} \frac{1}{2} \|y - Xb\|_{\ell_2}^2 + \lambda_0 \sigma^2 \|b\|_{\ell_0}, \tag{3.1.2}$$

where $\|b\|_{\ell_0}$ is the number of nonzero components in b. In other words, the estimator (3.1.2) achieves the best trade-off between the goodness of fit and the complexity of the model—here the number of variables included in the model. Popular selection procedures such as AIC, C_p, BIC, and RIC are all of this form with different values of the parameter: $\lambda_0 = 1$ in AIC [4, 106], $\lambda_0 = \frac{1}{2} \log n$ in BIC [139],

and $\lambda_0 = \log p$ in RIC [77]. It is known that these methods perform well both empirically and theoretically, see [77] and [12, 18] and the many references therein. Having said this, the problem of course is that these 'canonical selection procedures' are highly impractical. Solving (3.1.2) is in general NP-hard [112] and to the best of our knowledge, requires exhaustive searches over all subsets of columns of X, a procedure which clearly is combinatorial in nature and has exponential complexity, since for p of size about n there are about 2^p such subsets.

In recent years, several methods based on ℓ_1 minimization have been proposed to overcome this problem. The most well-known is probably the LASSO [147] (as introduced in Chapter 2), which replaces the nonconvex ℓ_0 norm in (3.1.2) with the convex ℓ_1 norm $\|b\|_{\ell_1} = \sum_{i=1}^{p} |b_i|$. The LASSO estimate $\hat{\beta}$ is defined as the solution to

$$\min_{b \in \mathbb{R}^p} \frac{1}{2} \|y - Xb\|_{\ell_2}^2 + \lambda \sigma \|b\|_{\ell_1}, \tag{3.1.3}$$

where λ is a regularization parameter essentially controlling the sparsity (or the complexity) of the estimated coefficients; see also [136] and [48] for exactly the same proposal. In contrast to (3.1.2), the optimization problem (3.1.3) is a quadratic program which can be solved efficiently. It is known that the LASSO performs well in some circumstances. Further, there is also an emerging literature on its theoretical properties [17,23,24,85,86,108,109,162,167,169] showing that in some special cases, the LASSO is effective. These important results, as well as many related results on ℓ_1 minimization, are discussed below in Section 3.1.2.

In this chapter, we will show that the LASSO provably works well in a surprisingly broad range of situations. We establish that under minimal assumptions guaranteeing that the predictor variables are not highly correlated, the LASSO achieves a squared error which is nearly as good as that one would obtain if one had an oracle supplying perfect information about which β_i's were nonzero. Continuing in this direction, we also establish that the LASSO correctly identifies the true model with very large probability provided that the amplitudes of the nonzero β_i are sufficiently large.

3.1.1 The coherence property

Throughout the chapter, we will assume without loss of generality that the matrix X has unit-normed columns as one can otherwise always rescale the columns. We denote by X_i the ith column of X ($\|X_i\|_{\ell_2} = 1$) and introduce the notion of coherence, which essentially measures the maximum correlation between unit-normed predictor variables and is defined by

$$\mu(X) = \sup_{1 \le i < j \le p} |\langle X_i, X_j \rangle|. \tag{3.1.4}$$

In words, the coherence is the maximum inner product between any two distinct columns of X. It follows that if the columns have zero mean, the coherence is just the maximum correlation between pairs of predictor variables.

We will be interested in problems in which the variables are not highly collinear or redundant.

Definition 3.1.1 (Coherence property) *A matrix X is said to obey the coherence property if*

$$\mu(X) \leq A_0 \cdot (\log p)^{-1}, \tag{3.1.5}$$

where A_0 is some positive numerical constant.

We will sometimes shorten notation by writing μ for the coherence and leaving the dependence on X implicit.

A matrix obeying the coherence property is a matrix in which the predictors are not highly collinear. This is a mild assumption. Suppose X is a Gaussian matrix with i.i.d. entries whose columns are subsequently normalized. The coherence of X is about $\sqrt{(2\log p)/n}$ so that such matrices trivially obey the coherence property unless n is ridiculously small, i.e., of the order of $(\log p)^3$. We will give other examples of matrices obeying this property later in the chapter, and will soon contrast this assumption with what is traditionally assumed in the literature.

In order to prove results with this relatively weak requirement on the correlations, we make use of random matrix theory derived by Tropp in [153]. Tropp also used this theory to prove noiseless results [152] with the same condition on the correlations.

3.1.2 Background literature

In the last few years, there have been many beautiful works attempting to understand the properties of the LASSO and other minimum ℓ_1 algorithms such as the Dantzig selector when the number of variables may be larger than the sample size [17,23,24,43,59,85,86,108,109,151,162,167,169]. Some papers focus on the estimation of the parameter β and on recovering its support, others focus on estimating $X\beta$.

Preceding the recent surge in attention, the efficacy of ℓ_1 minimization as a method for sparse approximation and model selection had been treated in the literature for quite some time (see, e.g., the 1938 work [16]); we could not hope to cover all of the relevant literature here. Instead, we detail several of the keystone achievements, and the results that set the background for the material in this chapter. In particular, we mainly restrict our discusssion to theoretical papers that use coherence-related conditions. This removes from the discussion the many RIP-based results discussed in chapter 2, whose applicability rests mainly with CS matrices, i.e., matrices that can be randomly chosen by the scientist. Further, when reviewing noisy results, we concentrate on papers

that consider the estimation of $X\beta$ and the support of β since this is what is considered in this chapter.

We begin with an early example in the geophysics literature, which demonstrates that ℓ_1 minimization was known to be useful in practice for quite some time. In 1979 Taylor et al. [146] applied ℓ_1 minimization to the problem (3.1.1), in the case when $X\beta$ is the convolution of β with a given wavelet w. They showed numerically that this recovered sparse signals quite well, while ridge regression (regularization with an ℓ_2 penalty), was shown to smooth out sparse signals, thus losing their inherent structure. A similar numerical result by Levy–Fullagar [103] demonstrated that a sparse signal could be reconstructed from subsampled Fourier measurements. Similar numerical results have been accumulating over the last few decades (see e.g., [47, 48]). In 1986, Santo-Synes [137] provided some theoretical justification for these empirical results. However, starting around the year 2000, their has been a spring of new and rigorous theory about ℓ_1 minimization.

The main thrust of the theory on ℓ_1 minimization began in the noiseless setting, $y = X\beta$; here the goal was to recover the sparsest vector β fitting the data, or a good approximation of it. Chen–Donoho [48] introduced to the signal processing community the now standard ℓ_1 recovery procedure

$$\min_b \|b\|_{\ell_1} \qquad \text{subject to } Xb = y. \tag{3.1.6}$$

This program is called basis pursuit, and may be interpreted as the LASSO in the limit as $\lambda \to 0$ (assuming that X has full row rank). Its introduction caused a chain reaction of theoretical publications on model selection and sparse approximation by ℓ_1 minimization, leading within a few years to the invention of CS.

An early paper by Donoho–Xuo [60] considered the case where X is the concatenation of two orthonormal bases. To state their result, we let s be an upper bound on $\|\beta\|_0$. The authors demonstrated that ℓ_1 minimization deterministically succeeds in recovering β whenever $s < \frac{1}{2}(1 + 1/\mu(X))$. They also demonstrated that under this coherence condition β is the unique sparsest solution to $y = Xb$. Elad–Bruckstein [69, 70] fine tuned this result by softening the condition on μ and $\|\beta\|_0$, as in the following theorem.

Theorem 3.1.2 *Suppose that $y = X\beta$ with $\|\beta\|_0 \le s$, where X is the concatenation of two orthonormal bases. Let $\hat{\beta}_0$ be the solution to*

$$\min_b \|b\|_0 \qquad \textit{subject to } Xb = y$$

and let $\hat{\beta}_1$ be the basis pursuit (3.1.6) *solution. Then, if $s < 1/\mu(X)$, $\hat{\beta}_0 = \beta$. Further, if $s < (\sqrt{2} - .5)/\mu(X)$, then $\hat{\beta}_1 = \beta$.*

Not surprisingly, there is a gap between where ℓ_0 minimization provably succeeds, and where its

convex relaxation, ℓ_1 minimization, succeeds. It may be surprising that this gap is so small. In fact, both bounds are tight (see [75]), and thus the gap cannot be removed.

These results were extended to concatenations of arbitrary numbers of orthogonal bases by Gribonval–Nielsen [87]. Shortly afterwards, Donoho–Elad [67] and Fuchs [80] further extended these results to arbitrary incoherent dictionaries. In this case the requirement given for exact recovery by ℓ_1 minimization was $s \leq \frac{1}{2}(1 + 1/\mu(X))$. Tropp [149] then introduced a general condition which ensures the efficacy of ℓ_1 minimization. Specifically, let $T \subset \{1, 2, \cdots, p\}$ be the support of β, and let X_T be the matrix X restricted to just the columns indexed by T. Tropp proved the following theorem.

Theorem 3.1.3 *Suppose that $y = X\beta$ and $T = \mathrm{supp}(\beta)$. Let $\hat{\beta}$ be the ℓ_1 minimizer of* (3.1.6)*. If*

$$\max_{i \notin T} \|(X_T^* X_T)^{-1} X_T^* X_i\|_{\ell_1} < 1 \tag{3.1.7}$$

then $\hat{\beta} = \beta$.

Further, Tropp showed that the condition $s = |T| < (1+1/\mu(X))/2$ implies that (3.1.7) holds, thereby recovering the incoherence-based results above. In the same paper, Tropp also considered somewhat weaker conditions; instead of a requirement on the maximum inner product between columns, he gave conditions for recovery based on sums of multiple inner products; we do not detail this result. We pause to note that Tropp's analysis also demonstrated that under the condition (3.1.7) a completely different sparsity-promoting algorithm also provably recovers β. This algorithm is called orthogonal matching pursuit [123], and is known as a *greedy algorithm*, which selects one coefficient at a time to include in the support set. See [149] for details.

We now move to the noisy model considered in this paper, $y = X\beta + z$. Basis pursuit (3.1.6) is somewhat ill-suited for this problem because the constraint $y = Xb$ would generally mean that β is infeasible. (Nevertheless, in certain cases, strong theoretical results on the ability of the basis pursuit to approximate a sparse vector from a noisy model do exist [166].) A program that makes more sense, and has better stability in practice, is the LASSO, introduced by Tibshirani [147] in 1995 (this is also equivalent to *basis pursuit denoising*, as introduced by Chen–Donoho [48]). In fact, Tibshirani credits Breiman [21] for motivating the definition of the LASSO with his own sparsity-inducing program: the *non-negative garotte*. Breiman's program solves

$$\min_c \|y - X(\hat{\beta}^0 \cdot c)\|_{\ell_2} \qquad \text{subject to } c_j \geq 0, \|c\|_{\ell_1} \leq t$$

where $\hat{\beta}^0$ is the least-squares estimate and $(\hat{\beta}^0 \cdot c)_j := \hat{\beta}^0_j \cdot c_j$, i.e., componentwise multiplication. In other words, the garotte shrinks the least-squares estimate by factors whose sum is constrained by t, a parameter which must be chosen by the scientist.

Returning to the LASSO, Tibshirani [147] demonstrated through numerical simulations, and analysis of the simple case where X is an orthogonal matrix, that the LASSO tends to promote sparse solutions. His numerical simulations further showed that the LASSO outperforms both ridge regression and subset selection when there are a small to moderate number of effects (i.e., when $\|\beta\|_0$ is small to moderate). Tibshirani's inspiring work on the LASSO led to years of study of this seemingly simple program, with the main goal of demonstrating its theoretical efficacy.

In the last decade, Bunea et al. [22–24] addressed the predictive error of the LASSO in a series of papers. There are several results here with different assumptions in this string of papers, one of which matches the coherence-based, $p > n$ discussion. Specifically, [23, Theorem 4.3] demonstrates similar error bounds to the ones given in this chapter (see Theorem 3.1.6), but under the coherence condition $s < \frac{1}{32\mu}$. Bickel et al. [17] also consider the prediction error of the LASSO, but under the assumption that the RIP holds. This is a somewhat different model than the one considered in this chapter. Nevertheless, we note that the RIP is satisfied at sparsity level s as long as $s < O(1)/\mu$, a condition on the coherence which is now perhaps looking familiar. We also note that this is a tight bound in the sense that there are matrices with s-sparse vectors in their null space with $s\mu = O(1)$.

Next, we would like to point to a work by Tropp [151], which, under the coherence condition $s < \mu/2$ demonstrated three key properties of the LASSO solution: 1) the support of the large entries of β are recovered; 2) the support of $\hat{\beta}$ is contained in the support of β; and 3) the error in recovery of β is comparable to the noise level in infinity norm. Further, Tropp noted that the coherence conditions may be weakened by techniques in Banach space geometry—indeed such techniques led to the weaker coherence conditions available in this Chapter.

Having a sparsity level substantially smaller than the inverse of the coherence is a common assumption in the modern literature on the subject, although in some circumstances a few papers have developed some weaker assumptions. To be a little more specific, [169] reports an asymptotic result in which the LASSO recovers the exact support of β provided that the strong irrepresentable condition of Section 3.3.5 holds. The references [108, 162] develop very similar results and use very similar requirements. The recent paper [93] develops similar results, but requires either a good initial estimator, or a level of coherence on the order of $n^{-1/2}$. In [43, 109] the singular values of X restricted to any subset of size proportional to the sparsity of β must be bounded away from zero while [17] introduces an extension of this condition. In all these works, a sufficient condition is that the sparsity be much smaller the inverse of the coherence.

This is a quite strong condition. In fact, as shown in [132] or [144, Theorem 2.3]

$$\mu \geq \sqrt{\frac{p-n}{n(p-1)}}.$$

It follows that for $p \geq 2n$, $\mu \geq 1/\sqrt{2n}$. Thus, in the significantly underdetermined case, these universal

results based on coherence all require $s \lesssim \sqrt{n}$ (the '$\sqrt{n}$ bottleneck').

In contrast, as noted in the previous section Tropp drastically weakened the coherence assumption in the paper [152]. He did this by considering a statistical model for β, as we do in this chapter. The main contribution of this chapter is to extend these results to the estimation of $X\beta$ in the noisy problem.

3.1.3 Sparse model selection

We begin by discussing the intuitive case where the vector β is sparse before extending our results to a completely general case. The basic question we would like to address here is how well can one estimate the response $X\beta$ when β happens to have only s nonzero components? Recall that we call such vectors *s-sparse*.

First and foremost, we would like to emphasize that in this chapter, we are interested in quantifying the performance one can expect from the LASSO in an overwhelming majority of cases. This viewpoint needs to be contrasted with an analysis concentrating on the worst case performance; when the focus is on the worst case scenario, one would study very particular values of the parameter β for which the LASSO does not work well. Our non-universal approach will enable us to show that one can reliably estimate the mean response $X\beta$ under much weaker conditions than those required for worst-case analysis.

Our point of view emphasizes the average performance (or the performance one could expect in a large majority of cases) and we thus need a statistical description of sparse models. To this end, we consider the *generic s-sparse model*, defined as follows:

1. The support $T \subset \{1, \ldots, p\}$ of the s nonzero coefficients of β is selected uniformly at random.

2. Conditional on T, the signs of the nonzero entries of β are independent and equally likely to be -1 or 1.

The consideration of random signs is not new (see, e.g., [38]); nor is the consideration of uniformly random support along with random, mean-zero signs, (see [37, 153]).

We make no assumption on the amplitudes. In some sense, this is the simplest statistical model one could think of; it simply says that that all subsets of a given cardinality are equally likely, and that the signs of the coefficients are equally likely. In other words, one is not biased towards certain variables nor do we have any reason to believe a priori whether a given coefficient is positive or negative.

Our first result is that for most s-sparse vectors β, the LASSO is provably accurate. Throughout, $\|X\|$ refers to the operator norm of the matrix A (the largest singular value).

Theorem 3.1.4 *Suppose that X obeys the coherence property and assume that β is taken from the generic s-sparse model. Suppose that $s \le c_0 p/[\|X\|^2 \log p]$ for some positive numerical constant c_0. Then the LASSO estimate* (3.1.3) *computed with $\lambda = 2\sqrt{2\log p}$ obeys*

$$\|X\beta - X\hat{\beta}\|_{\ell_2}^2 \le C_0 \cdot (2\log p) \cdot s \cdot \sigma^2 \tag{3.1.8}$$

with probability at least $1 - 6p^{-2\log 2} - p^{-1}(2\pi \log p)^{-1/2}$. The constant C_0 may be taken as $8(1+\sqrt{2})^2$.

For simplicity, we have chosen $\lambda = 2\sqrt{2\log p}$ but one could take any λ of the form $\lambda = (1+a)\sqrt{2\log p}$ with $a > 0$. Our proof indicates that as a decreases, the probability with which (3.1.8) holds decreases but the constant C_0 also decreases. Conversely, as a increases, the probability with which (3.1.8) holds increases but the constant C_0 also increases.

Theorem 3.1.4 asserts that one can estimate $X\beta$ with nearly the same accuracy as if one knew ahead of time which β_i's were nonzero. To see why this is true, suppose that the support T of the true β was known. In this ideal situation, we would presumably estimate β by regressing y onto the columns of X with indices in T, and construct

$$\beta^\star = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \|y - Xb\|_{\ell^2}^2 \quad \text{subject to} \quad b_i = 0 \text{ for all } i \notin T. \tag{3.1.9}$$

It is a simple to calculation to show that this ideal estimator (it is ideal because we would not know the set of nonzero coordinates) achieves[1]

$$\mathbb{E}\,\|X\beta - X\beta^\star\|_{\ell_2}^2 = s \cdot \sigma^2. \tag{3.1.10}$$

Hence, one can see that (3.1.8) is optimal up to a factor proportional to $\log p$. It is also known that one cannot in general hope for a better result; the log factor is the price we need to pay for not knowing ahead of time which of the predictors are actually included in the model.

The assumptions of our theorem are pretty mild. Roughly speaking, if the predictors are not too collinear and if s is not too large, then the LASSO works most of the time. An important point here is that the restriction on the sparsity can be very mild. We give two examples to illustrate our purpose.

- *Random design.* Imagine as before that the entries of X are i.i.d. $\mathcal{N}(0,1)$ and then normalized. Then the operator norm of X is sharply concentrated around $\sqrt{p/n}$ so that our assumption essentially reads $s \le c_0 n/\log p$. Expressed in a different way, β does not have to be sparse at all. It has to be smaller than the number of observations of course, but not by a very large margin.

[1] One could also develop a similar estimate with high probability but we find it simpler and more elegant to derive the performance in terms of expectation.

Similar conclusions would apply to many other types of random matrices.

- *Signal estimation.* A problem that has attracted quite a bit of attention in the signal processing community is that of recovering a signal which has a sparse expansion as a superposition of spikes and sinusoids (see, e.g., [37, 154]). Here, we have noisy data y

$$y(t) = f(t) + \sigma z(t), \quad t = 1, \ldots, n, \tag{3.1.11}$$

about a digital signal f of interest, which is expressed as the the 'time-frequency' superposition

$$f(t) = \sum_{k=1}^{n} \alpha_k^{(0)} \delta(t-k) + \sum_{k=1}^{n} \alpha_k^{(1)} \varphi_k(t); \tag{3.1.12}$$

δ is a Dirac or spike obeying $\delta(t) = 1$ if $t = 0$ and 0 otherwise, and $(\varphi_k(t))_{1 \le k \le n}$ is an orthonormal basis of sinusoids. The problem (3.1.11) is of the general form (3.1.1) with $X = [I_n \; F_n]$ in which I_n is the identity matrix, F_n is the basis of sinusoids (a discrete cosine transform), and β is the concatenation of $\alpha^{(0)}$ and $\alpha^{(1)}$. Here, $p = 2n$ and $\|X\| = \sqrt{2}$. Also, X obeys the coherence property if n or p is not too small since $\mu(X) = \sqrt{2/n} = 2/\sqrt{p}$.

Hence, if the signal has a sparse expansion with fewer than on the order of $n/\log n$ coefficients, then the LASSO achieves a quality of reconstruction which is essentially as good as what could be achieved if we knew in advance the precise location of the spikes and the exact frequencies of the sinusoids.

This fact extends to other pairs of orthobases and to general overcomplete expansions as we will explain later.

In our two examples, the condition of Theorem 3.1.4 is satisfied for s as large as on the order of $n/\log p$; that is, β may have a large number of nonzero components. The novelty here is that the assumptions on the sparsity level s and on the correlation between predictors are very realistic. This is different from the available literature, which typically requires a much lower bound on the coherence or a much lower sparsity level, see Section 3.4 for a comprehensive discussion. In addition, many published results assume that the entries of the design matrix X are sampled from a probability distribution—e.g., are i.i.d. samples from the standard normal distribution—which we are not assuming here (one could of course specialize our results to random designs as discussed above). Hence, we do not simply prove that in some idealized setting the LASSO would do well, but that it has a very concrete edge in practical situations—as shown empirically in a great number of works.

An interesting fact is that one cannot expect (3.1.8) to hold for all models as one can construct simple examples of incoherent matrices and special β for which the LASSO does not select a good

model, see Section 3.2. In this sense, (3.1.8) can be achieved on the average—or better, in an overwhelming majority of cases—but not in all cases.

3.1.4 Exact model recovery

In this section we consider the problem of recovering the support of β. Before beginning, we note that Tropp [151] has already considered this problem in the deterministic setting (without a random model for the signal), and has shown under some strong coherence conditions that the support of the entries of β which stand above the noise is recovered. In fact, our results are derived as a combination of the ideas of [151] and the theory of [153] together with the development of a few concentration inequalities.

Now, to be concrete, suppose that we are interested in estimating the set $T = \{i : \beta_i \neq 0\}$. Then we show that if the values of the nonvanishing β_i's are not too small, then the LASSO correctly identifies the 'right' model.

Theorem 3.1.5 *Let T be the support of β and suppose that*

$$\min_{i \in I} |\beta_i| > 8\,\sigma\,\sqrt{2\log p}.$$

Then under the assumptions of Theorem 3.1.4, the LASSO estimate with $\lambda = 2\sqrt{2\log p}$ obeys

$$\mathrm{supp}(\hat{\beta}) = \mathrm{supp}(\beta), \quad \textit{and} \tag{3.1.13}$$

$$\mathrm{sgn}(\hat{\beta}_i) = \mathrm{sgn}(\beta_i), \quad \textit{for all } i \in T, \tag{3.1.14}$$

with probability at least $1 - 2p^{-1}((2\pi\log p)^{-1/2} + |T|p^{-1}) - O(p^{-2\log 2})$.

In words, if the nonzero coefficients are significant in the sense that they stand above the noise, then the LASSO identifies all the variables of interest and only these. Further, the LASSO also correctly estimates the signs of the corresponding coefficients. Again, this does not hold for all β's as shown in the example of Section 3.2 but for a wide majority.

Our condition says that the amplitudes must be larger than a constant times the noise level times $\sqrt{2\log p}$ which is sharp modulo a small multiplicative constant. Our statement is nonasymptotic, and relies upon [169] and [24]. In particular, [169] requires X and β to satisfy the *Irrepresentable Condition*, which is sufficient to guarantee the exact recovery of the support of β in some asymptotic regime; Section 3.3.3 connects with this line of work by showing that the Irrepresentable Condition holds with high probability under the stated assumptions. We would also note in passing that [79] considers the Irrepresentable Condition as well (without calling it by this name) and in fact our proof of Theorem 3.1.5 has some ideas in common with that paper.

As before, we have decided to state the theorem for a concrete value of λ, namely, $2\sqrt{2\log p}$ but we could have used any value of the form $(1+a)\sqrt{2\log p}$ with $a > 0$. When a decreases, our proof indicates that one can lower the threshold on the minimum nonzero value of β but that at the same time, the probability of success is lowered as well. When a increases, the converse applies. Finally our proof shows that by setting λ close to $\sqrt{2\log p}$ and by imposing slightly stronger conditions on the coherence and the sparsity s, one can substantially lower the threshold on the minimum nonzero value of β and bring it close to $\sigma\sqrt{2\log p}$.

We would also like to remark that under the hypotheses of Theorem 3.1.5, one can improve the estimate (3.1.8) a little by using a two-step procedure similar to that proposed in [43].

1. Use the LASSO to find $\hat{T} \equiv \{i : \hat{\beta}_i \neq 0\}$.
2. Find $\tilde{\beta}$ by regressing y onto the columns (X_i), $i \in \hat{T}$.

Since $\hat{T} = T$ with high probability, we have that

$$\|X\tilde{\beta} - X\beta\|_{\ell_2}^2 = \sigma^2 \|P[T]z\|_{\ell_2}^2$$

with high probability, where $P[T]$ is the projection onto the space spanned by the variables (X_i). Because $\|P[T]z\|_{\ell_2}^2$ is concentrated around $|T| = s$, it follows that with high probability,

$$\|X\tilde{\beta} - X\beta\|_{\ell_2}^2 \leq C \cdot s \cdot \sigma^2,$$

where C is a some small numerical constant. In other words, when the values of the nonzero entries of β are sufficiently large, one does not have to pay the logarithmic factor.

3.1.5 General model selection

In many applications, β is not sparse or does not have a real meaning so that it does not make much sense to talk about the values of this vector. Consider an example to make this precise. Suppose we have noisy data y (3.1.11) about an n-pixel digital image f, where σz is white noise. We wish to remove the noise, i.e. estimate the mean of the vector y. A majority of modern methods express the unknown signal as a superposition of fixed waveforms $(\varphi_i(t))_{1 \leq i \leq p}$,

$$f(t) = \sum_{i=1}^{p} \beta_i \varphi_i(t), \tag{3.1.15}$$

and construct an estimate

$$\hat{f}(t) = \sum_{i=1}^{p} \hat{\beta}_i \varphi_i(t).$$

That is, one introduces a model $f = X\beta$ in which the columns of X are the sampled waveforms $\varphi_i(t)$. It is now extremely popular to consider overcomplete representations with many more waveforms than samples, i.e., $p > n$. The reason is that overcomplete systems offer a wider range of generating elements which may be well suited to represent contributions from different phenomena; potentially, this wider range allows more flexibility in signal representation and enhances statistical estimation.

In this setup, two comments are in order. First, there is no ground truth associated with each coefficient β_i; there is no real wavelet or curvelet coefficient. And second, signals of general interest are never really exactly sparse; they are only approximately sparse meaning that they may be well approximated by sparse expansions. These considerations emphasize the need to formulate results to cover those situations in which the precise values of β_i are either ill-defined or meaningless.

In general, one can understand model selection as follows. Select a model—a subset T of the columns of X—and construct an estimate of $X\beta$ by projecting y onto the subspace generated by the variables in the model. Mathematically, this is formulated as

$$X\hat{\beta}[T] = P[T]y = P[T]X\beta + \sigma P[T]z,$$

where $P[T]$ denotes the projection onto the space spanned by the variables (X_i), $i \in T$. What is the accuracy of $X\hat{\beta}[T]$? Note that

$$X\beta - X\hat{\beta}[T] = (\mathrm{Id} - P[T])X\beta - \sigma P[T]z$$

and, therefore, the mean-squared error (MSE) obeys[2]

$$\mathbb{E}\,\|X\beta - X\hat{\beta}[T]\|^2 = \|(\mathrm{Id} - P[T])X\beta\|^2 + |T|\,\sigma^2. \tag{3.1.16}$$

This is the classical bias variance decomposition; the first term is the squared bias one gets by using only a subset of columns of X to approximate the true vector $X\beta$. The second term is the variance of the estimator and is proportional to the size of the model T.

Hence, one can now define the *ideal model* achieving the minimum MSE over all models

$$\min_{I \subset \{1,\ldots,p\}} \|(\mathrm{Id} - P[T])X\beta\|^2 + |T|\,\sigma^2. \tag{3.1.17}$$

We will refer to this as the ideal risk. This is ideal in the sense that one could achieve this performance if we had available an oracle which—knowing $X\beta$— would select for us the best model to use, i.e. the best subset of explanatory variables.

To connect this with our earlier discussion, one sees that if there is a representation of $f = X\beta$

[2]It is again simpler to state the performance in terms of expectation.

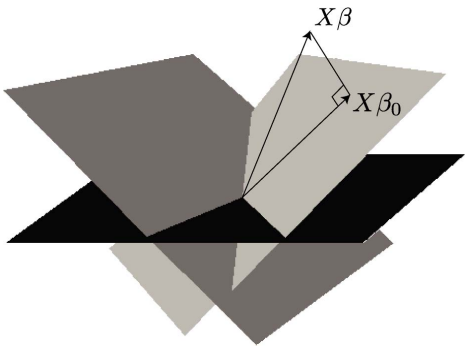


Figure 3.1: The vector $X\beta_0$ is the projection of $X\beta$ on an ideally selected subset of covariates. These covariates span a plane of optimal dimension which, among all planes spanned by subsets of the same dimension, is closest to $X\beta$.

in which β has s nonzero terms, then the ideal risk is bounded by the variance term, namely, $s \cdot \sigma^2$ (just pick T to be the support of β in (3.1.17)). The point we would like to make is that whereas we did not search for an optimal bias-variance trade off in the previous section, we will here. The reason is that even in the case where the model is interpretable, the projection estimate on the model corresponding to the nonzero values of β_i may very well be inaccurate and have a mean-squared error which is far larger than (3.1.17). In particular, this typically happens if out of the s nonzero β_i's, only a small fraction are really significant while the others are not (e.g., in the sense that any individual test of significance would not reject the hypothesis that they vanish). In this sense, the main result of this section, Theorem 3.1.6 generalizes but also strengthens Theorem 3.1.4.

An important question is of course whether one can get close to the ideal risk (3.1.17) without the help of an oracle. It is known that solving the combinatorial optimization problem (3.1.2) with a value of λ_0 being a sufficiently large multiple of $\log p$ would provide an MSE within a multiplicative factor of order $\log p$ of the ideal risk. That real estimators with such properties exist is inspiring. Yet solving (3.1.2) is computationally intractable. Our next result shows that in a wide range problems, the LASSO also nearly achieves the ideal risk.

We are naturally interested in quantifying the performance one can expect from the LASSO in nearly all cases and just as before, we now introduce a useful statistical description of these cases. Consider the best model T_0 achieving the minimum in (3.1.17). In case of ties, pick one uniformly at random. Suppose T_0 is of cardinality s. Then we introduce the *best s-dimensional subset model* defined as follows:

1. The subset $T_0 \subset \{1, \ldots, p\}$ of cardinality s is distributed uniformly at random.

2. Define β_0 with support T_0 via

$$X\beta_0 = P[T_0]X\beta. \tag{3.1.18}$$

In other words, β_0 is the vector one would get by regressing the true mean vector $X\beta$ onto

the variables in T_0; we call β_0 the ideal approximation. Conditional on T_0, the signs of the nonzero entries of β_0 are independent and equally likely to be -1 or 1.

We make no assumption on the amplitudes. Our intent is just the same as before. All models are equally likely (there is no bias towards special variables) and one has no a priori information about the sign of the coefficients associated with each significant variable.

Theorem 3.1.6 *Suppose that X obeys the coherence property and assume that the ideal approximation β_0 is taken from the best s-dimensional subset model. Suppose that $s \le c_0 p/[\|X\|^2 \log p]$ for some positive numerical constant c_0. Then the LASSO estimate* (3.1.3) *computed with $\lambda = 2\sqrt{2\log p}$ obeys*

$$\|X\beta - X\hat{\beta}\|_{\ell_2}^2 \le (1+\sqrt{2})\left[\min_{I\subset\{1,\ldots,p\}} \|X\beta - P[T]X\beta\|_{\ell_2}^2 + C_0'\,(2\log p)\cdot|T|\cdot\sigma^2\right] \tag{3.1.19}$$

with probability at least $1 - 6p^{-2\log 2} - p^{-1}(2\pi\log p)^{-1/2}$. The constant C_0' may be taken as $12+10\sqrt{2}$.

In words, the LASSO nearly selects the best model in a very large majority of cases. As argued earlier, this also strengthens our earlier result since the right-hand side in (3.1.19) is always less or equal to $O(\log p)\,s\,\sigma^2$ whenever there is an s-sparse representation.[3]

Theorem 3.1.6 is guaranteeing excellent performance in a broad range of problems. That is, whenever we have a design matrix X whose columns are not too correlated, then for most responses $X\beta$, the LASSO will find a statistical model with low mean-squared error; simple extensions would also claim that the LASSO finds a statistical model with very good predictive power but we will not consider these here. As an illustrative example, we can consider predicting the clinical outcomes from different tumors on the basis of gene expression values for each of the tumors. In typical problems, one considers hundreds of tumors and tens of thousands of genes. While some of the gene expressions (the columns of X) are correlated, one can always eliminate redundant predictors, e.g., via clustering techniques. Once the statistician has designed an X with low coherence, then in most cases, the LASSO is guaranteed to find a subset of genes with near-optimal predictive power.

There is a slightly different formulation of this general result which may go as follows: let s_0 be the maximum sparsity level $s_0 = \lfloor c_0 p/[\|X\|^2\log p]\rfloor$ and for each $s \le s_0$, introduce $\mathcal{A}_s \subset \{-1,0,1\}^p$ as the set of all possible signs of vectors $\beta \in \mathbb{R}^p$ with $\mathrm{sgn}(\beta_i) = 0$ if $\beta_i = 0$ such that exactly s signs are nonzero. Then under the hypotheses of our theorem, for each $X\beta \in \mathbb{R}^n$,

$$\|X\beta - X\hat{\beta}\|_{\ell_2}^2 \le \min_{s\le s_0}\ \min_{b:\,\mathrm{sgn}(b)\in\mathcal{A}_{0,s}} (1+\sqrt{2})\left[\|X\beta - Xb\|_{\ell_2}^2 + C_0'\,(2\log p)\cdot s\cdot\sigma^2\right] \tag{3.1.20}$$

[3] We have assumed that the mean response f of interest is in the span of the columns of X (i.e. of the form $X\beta$) which always happens when $p \ge n$ and X has full column rank for example. If this is not the case, however, the error would obey $\|f - X\hat{\beta}\|_{\ell_2}^2 = \|Pf - X\hat{\beta}\|_{\ell_2}^2 + \|(\mathrm{Id}-P)f\|_{\ell_2}^2$ where P is the projection onto the range of X. The first term obeys the oracle inequality so that the LASSO estimates Pf in a near-optimal fashion. The second term is simply the size of the unmodelled part the mean response.

with probability at least $1 - O(p^{-1})$, where one can still take $C_0' = 12 + 10\sqrt{2}$ (the probability is with respect to the noise distribution). Above, $\mathcal{A}_{0,s}$ is a very large subset of $\mathcal{A}_s$ obeying

$$|\mathcal{A}_{0,s}|/|\mathcal{A}_s| \geq 1 - O(p^{-1}). \tag{3.1.21}$$

Hence, for most β, the sub-oracle inequality (3.1.20) is actually the true oracle inequality.

For completeness, $\mathcal{A}_{0,s}$ is defined as follows. Let $b \in \mathcal{A}_s$ be supported on T; b_T is the restriction of the vector b to the index set T, and X_T is the submatrix formed by selecting the columns of X with indices in T. Then we say that $b \in \mathcal{A}_{0,s}$ if and only if the following three conditions hold: 1) the submatrix $X_T^* X_T$ is invertible and obeys $\|(X_T^* X_T)^{-1}\| \leq 2$; 2) $\|X_{I^c}^* X_T (X_T^* X_T)^{-1} b_T\|_{\ell_\infty} \leq 1/4$ (recall that $b \in \{-1, 0, 1\}^p$ is a sign pattern); 3) $\max_{i \notin I} \|X_T (X_T^* X_T)^{-1} X_T^* X_i\| \leq c_0/\sqrt{\log p}$ for some numerical constant c_0. In Section 3.3, we will analyze these three conditions in detail and prove that $|\mathcal{A}_{0,s}|$ is large. The first condition is called the *invertibility condition* and the second and third conditions are needed to establish that a certain *complementary size condition* holds, see Section 3.3.

3.1.6 Implications for signal estimation

Our findings may be of interest to researchers interested in signal estimation and we now recast our main results in the language of signal processing. Suppose we are interested in estimating a signal $f(t)$ from observations

$$y(t) = f(t) + \sigma z(t), \qquad t = 0, \ldots, n-1,$$

where σz is white noise with variance σ^2. We are given a dictionary of waveforms $(\varphi_i(t))_{1 \leq i \leq p}$ which are normalized so that $\sum_{t=0}^{n-1} \varphi_i^2(t) = 1$, and are looking for an estimate of the form $\hat{f}(t) = \sum_{i=1}^{p} \hat{\alpha}_i \varphi_i(t)$. When we have an overcomplete representation in which $p > n$, there are infinitely many ways of representing f as a superposition of the dictionary elements.

Introduce now the best m-term approximation f_m defined via

$$\|f - f_m\|_{\ell_2} = \inf_{a:\, \#\{i,\, a_i \neq 0\} \leq m} \|f - \sum_i a_i \varphi_i\|_{\ell_2};$$

that is, it is that linear combination of at most m elements of the dictionary which comes closest to the object f of interest. [4] With these notations, if we could somehow guess the best model of dimension m, one would achieve a MSE equal to

$$\|f - f_m\|_{\ell_2}^2 + m\sigma^2.$$

[4] Note that again, finding f_m is in general a combinatorially hard problem.

Therefore, one can rewrite the ideal risk (which could be attained with the help of an oracle telling us exactly which subset of waveforms to use) as

$$\min_{0\le m\le p} \|f - f_m\|_{\ell_2}^2 + m\sigma^2, \tag{3.1.22}$$

which is exactly the trade-off between the approximation error and the number of terms in the partial expansion[5].

Consider now the estimate $\hat{f} = \sum_i \hat{\alpha}_i \varphi_i$ where $\hat{\alpha}$ is solution to

$$\min_{a\in\mathbb{R}^p} \frac{1}{2}\|y - \sum_i a_i\varphi_i\|_{\ell_2}^2 + \lambda\sigma\|a\|_{\ell_1} \tag{3.1.23}$$

with $\lambda = 2\sqrt{2\log p}$, say. Then provided that the dictionary is not too redundant in the sense that $\max_{1\le i<j\le p} |\langle\varphi_i, \varphi_j\rangle| \le c_0/\log p$, Theorem 3.1.6 asserts that for most signals f, the minimum-ℓ_1 estimator (3.1.23) obeys

$$\|\hat{f} - f\|_{\ell_2}^2 \le C_0 \left[\inf_m \|f - f_m\|_{\ell_2}^2 + \log p \cdot m\sigma^2\right], \tag{3.1.24}$$

with large probability and for some reasonably small numerical constant C_0. In other words, one obtains a squared error which is within a logarithmic factor of what can be achieved with information provided by a genie.

Overcomplete representations are now in widespread use as in the field of artificial neural networks for instance [49]. In computational harmonic analysis and image/signal processing, there is an emerging wisdom which says that 1) there is no universal representation for signals of interest and 2) different representations are best for different phenomena; 'best' is here understood as providing sparser representations. For instance:

- sinusoids are best for oscillatory phenomena;
- wavelets [105] are best for point-like singularities;
- curvelets [29, 30] are best for curve-like singularities (edges);
- local cosines are best for textures; and so on.

Thus, many efficient methods in modern signal estimation proceed by forming an overcomplete dictionary—a union of several distinct representations—and then by extracting a sparse superposition that fits the data well. The main result of this chapter says that if one solves the quadratic program (3.1.23), then one is provably guaranteed near-optimal performance for most signals of interest. This explains why these results might be of interest to people working in this field.

[5] It is also known that for many interesting classes of signals $\mathcal{F}$ and appropriately chosen dictionaries, taking the supremum over $f \in \mathcal{F}$ in (3.1.22) comes within a log factor of the minimax risk for $\mathcal{F}$.

The spikes and sines model has been studied extensively in the literature on information theory in the nineties and there, the assumption that the 'arrival times' of the spikes and the frequencies of the sinusoids are random is legitimate. In other situations, the model may be less adequate. For instance, in image processing, the large wavelet coefficients tend to appear early in the series, i.e., at low frequencies. Baraniuk et al. [11] address this situation, and show that taking into account prior information about the model can aid in the sparse approximation of the image. With this in mind, two comments are in order. First, it is likely that similar results would hold for other models (we just considered the simplest). And second, if we have a lot of a priori information about which coefficients are more likely to be significant, then we would probably not want to use the plain LASSO (3.1.3) but rather incorporate this side information.

3.1.7 Organization of the chapter

The chapter is organized as follows. In Section 3.2, we explain why our results are nearly optimal, and cannot be fundamentally improved. Section 3.3 introduces a recent result due to Joel Tropp regarding the norm of certain random submatrices which is essential to our proofs, and proves all of our results. We conclude with a discussion in Section 3.4 where for the most part, we relate our work with a series of other published results, and distinguish our main contributions.

3.2 Optimality

3.2.1 For almost all sparse models

A natural question is whether one can relax the condition about β being *generically* sparse or about $X\beta$ being well approximated by a *generically* sparse superposition of covariates. The emphasis is on 'generic' meaning that our results apply to nearly all objects taken from a statistical ensemble but perhaps not all. This begs a question: can one hope to establish versions of our results which would hold universally? The answer is negative. Even in the case when X has very low coherence, one can show that the LASSO does not provide an accurate estimation of certain mean vectors $X\beta$ with a sparse coefficient sequence. This section gives one such example.

Suppose as in Section 3.1.3 that we wish to estimate a signal assumed to be a sparse superposition of spikes and sinusoids. We assume that the length n of the signal $f(t)$, $t = 0, 1, \ldots, n-1$, is equal to $n = 2^{2j}$ for some integer j. The basis of spikes is as before and the orthobasis of sinusoids takes

the form

$$\begin{aligned}
\varphi_1(t) &= 1/\sqrt{n}, \\
\varphi_{2k}(t) &= \sqrt{2/n}\cos(2\pi kt/n), \quad k = 1, 2, \ldots, n/2-1, \\
\varphi_{2k+1}(t) &= \sqrt{2/n}\sin(2\pi kt/n), \quad k = 1, 2, \ldots, n/2-1, \\
\varphi_n(t) &= (-1)^t/\sqrt{n}.
\end{aligned}$$

Recall the discrete identity (a discrete analog of the Poisson summation formula)

$$\begin{aligned}
\sum_{k=0}^{2^j-1} \delta(t-k2^j) &= \sum_{k=0}^{2^j-1} \frac{1}{\sqrt{n}} e^{i2\pi\, k2^j t/n} \\
&= \frac{1}{\sqrt{n}}(1+(-1)^t) + \frac{2}{\sqrt{n}} \sum_{k=1}^{2^{j-1}-1} \cos(2\pi\, k2^j t/n) \\
&= \varphi_1(t) + \varphi_n(t) + \sqrt{2} \sum_{k=1}^{2^{j-1}-1} \varphi_{k2^{j+1}}(t). \qquad (3.2.1)
\end{aligned}$$

Then consider the model

$$y = \mathbf{1} + \sigma z = X\beta + \sigma z,$$

where $\mathbf{1}$ is the constant signal equal to 1 and X is the $n \times (2n-1)$ matrix

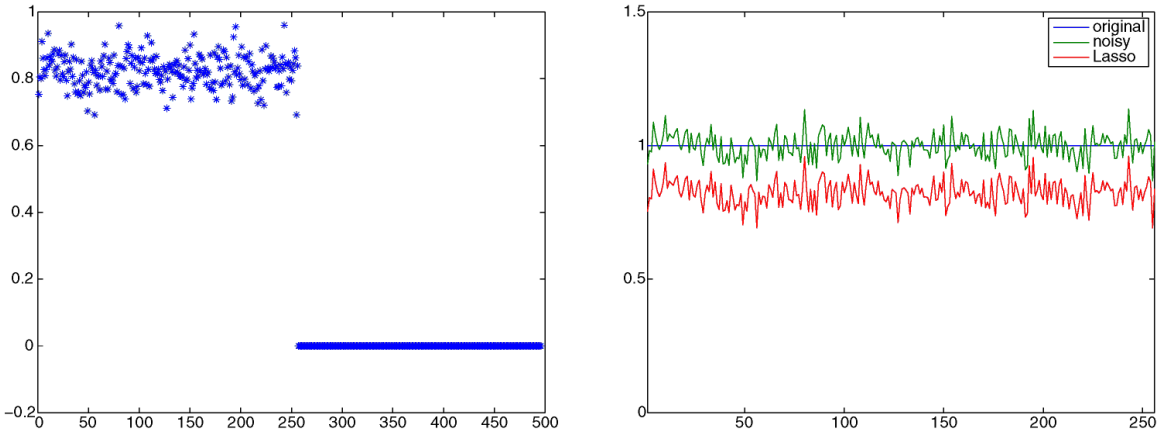
$$X = [I_n \; F_{n,2:n}]$$

in which I_n is the identity (the basis of spikes) and $F_{n,2:n}$ is the orthobasis of sinusoids minus the first basis vector φ_1. Note that this is a low-coherence matrix X since $\mu(X) = \sqrt{2/n}$. In plain English, we are simply trying to estimate a constant-mean vector. It follows from (3.2.1) that

$$1 = \sqrt{n}\left[\sum_{k=0}^{2^j-1} \delta(t-k2^j) - \varphi_n(t) - \sqrt{2}\sum_{k=1}^{2^{j-1}-1} \varphi_{k2^{j+1}}(t)\right],$$

so that $\mathbf{1}$ has a sparse expansion since it is a superposition of at most $\sqrt{n}$ spikes and $\sqrt{n}/2$ sinusoids (it can also be deduced from existing results that this is actually the sparsest expansion). In other words, if we knew which column vectors to use, one could obtain

$$\mathbb{E}\,\|X\beta^\star - X\beta\|_{\ell_2}^2 = \frac{3}{2}\sqrt{n}\sigma^2.$$



(a) (b) (a) (b)

Figure 3.2: Sparse signal recovery with the LASSO. (a) Values of the estimated coefficients. All the spike coefficients are obtained by soft-thresholding y and are nonzero. (b) LASSO signal estimate; $X\hat{\beta}$ is just a shifted version of the noisy signal.

How does the LASSO compare? We claim that with very high probability

$$\hat{\beta}_i = \begin{cases} y_i - \lambda\sigma, & i \in \{1, \dots, n\}, \\ 0, & i \in \{n+1, \dots, 2n-1\}, \end{cases} \tag{3.2.2}$$

so that

$$X\hat{\beta} = y - \lambda\sigma\,\mathbf{1} \tag{3.2.3}$$

provided that $\lambda\sigma \le 1/2$. In short, the LASSO does not find the sparsest model at all. As a matter of fact, it finds a model as dense as it can be, and the resulting mean-squared error is awful since

$$\mathbb{E}\,\|X\hat{\beta} - X\beta\|_{\ell_2}^2 \approx (1+\lambda^2)n\sigma^2.$$

Even if one could somehow remove the bias, this would still be a very bad performance.

An illustrative numerical example is displayed in Figure 3.2. In this example, $n = 256$ so that $p = 512 - 1 = 511$. The mean vector $X\beta$ is made up as above and there is a representation in which β has only 24 nonzero coefficients. Yet, the LASSO finds a model of dimension 256; i.e. select as many variables as there are observations.

We need to justify (3.2.2), as (3.2.3) would be an immediate consequence. It follows from taking the subgradient of the LASSO functional that $\hat{\beta}$ is a minimizer if and only if

$$\begin{aligned} X_i^*(y - X\hat{\beta}) &= \lambda\sigma\,\mathrm{sgn}(\hat{\beta}_i), & \hat{\beta}_i \neq 0, \\ |X_i^*(y - X\hat{\beta})| &\le \lambda\sigma, & \hat{\beta}_i = 0. \end{aligned} \tag{3.2.4}$$

One can further establish that $\hat{\beta}$ is the unique minimizer of (3.1.3) if

$$\begin{aligned} X_i^*(y - X\hat{\beta}) &= \lambda\sigma \operatorname{sgn}(\hat{\beta}_i), & \hat{\beta}_i \neq 0, \\ |X_i^*(y - X\hat{\beta})| &< \lambda\sigma, & \hat{\beta}_i = 0, \end{aligned} \tag{3.2.5}$$

and the columns indexed by the support of $\hat{\beta}$ are linearly independent (note the strict inequalities). We then simply need to show that $\hat{\beta}$ given by (3.2.2) obeys (3.2.5). Suppose that $\min_i y_i > \lambda\sigma$. A sufficient condition is that $\max_i \sigma \cdot |z_i| < 1 - \lambda\sigma$ which occurs with very large probability if $\lambda\sigma \le 1/2$ and $\lambda > \sqrt{2\log n}$ (see (3.3.4) with $X = I$). (One can always allow for larger noise by multiplying the signal by a factor greater than 1.) Note that $y - X\hat{\beta} = \lambda\sigma\,\mathbf{1}$ so that for $i \in \{1, \dots, n\}$ we have

$$X_i^*(y - X\hat{\beta}) = \lambda\sigma = \lambda\sigma \operatorname{sgn}(\hat{\beta}_i),$$

whereas for $i \in \{n+1, \dots, 2n-1\}$, we have

$$X_i^*(y - X\hat{\beta}) = \lambda\sigma\langle X_i, \mathbf{1}\rangle = 0,$$

which proves our claim.

To summarize, even when the coherence is low, i.e. of size about $1/\sqrt{n}$, there are sparse vectors β with sparsity level about equal to $\sqrt{n}$ for which the LASSO completely misbehaves (we presented an example but there are of course many others). It is therefore a fact that none of our theorems, namely, Theorems 3.1.4, 3.1.5, and 3.1.6 can hold for all β's. In this sense, they are sharp.

3.2.2 For sufficiently incoherent matrices

We now show that predictors cannot be too collinear, and begin by examining a small problem in which X is a 2×2 matrix, $X = [X_1, X_2]$. We violate the coherence property by choosing X_1 and X_2 so that $\langle X_1, X_2\rangle = 1 - \epsilon$, where we think of ϵ as being very small. Assume without loss of generality that $\sigma = 1$ to simplify. Consider now

$$\beta = \frac{a}{\epsilon}\begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

where a is some positive amplitude and observe that $X\beta = a\epsilon^{-1}(X_1 - X_2)$, and $X^*X\beta = a(1, -1)^*$. For example, we could set $a = 1$. It is well known that the LASSO estimate $\hat{\beta}$ vanishes if $\|X^*y\|_{\ell_\infty} \le \lambda$. Now

$$\|X^*y\|_{\ell_\infty} \le a + \sigma\|X^*z\|_{\ell_\infty}$$

so that if $a = 1$, say, and λ is not ridiculously small, then there is a positive probability π_0 that $\hat{\beta} = 0$ where $\pi_0 \geq \mathbb{P}(\sigma \cdot \|X^* z\|_\infty \leq \lambda - 1)$.[6] For example, if $\lambda > 1 + 3 = 4$, then $\hat{\beta} = 0$ as long as both entries of $X^* z$ are within 3 standard deviations of 0. When $\hat{\beta} = 0$, the squared error loss obeys

$$\|X\beta\|_{\ell_2}^2 = 2\frac{a^2}{\epsilon},$$

which can be made arbitrarily large if we allow ϵ to be arbitrarily small.

Of course, the culprit in our 2-by-2 example is hardly sparse and we now consider the $n \times n$ diagonal block matrix X_0 (n even)

$$X_0 = \begin{bmatrix} X & & & \\ & X & & \\ & & \ddots & \\ & & & X \end{bmatrix}$$

with blocks made out of $n/2$ copies of X. We now consider β from the s-sparse model with independent entries sampled from the distribution (we choose $a = 1$ for simplicity but we could consider other values as well)

$$\beta_i = \begin{cases} \epsilon^{-1} & \text{w. p. } n^{-1/2}, \\ -\epsilon^{-1} & \text{w. p. } n^{-1/2}, \\ 0 & \text{w. p. } 1 - 2n^{-1/2}. \end{cases}$$

Certainly, the support of β is random and the signs are random. One could argue that the size of the support is not fixed (the expected value is $2\sqrt{n}$ so that β is sparse with very large probability) but this is obviously unessential.[7]

Because X_0 is block diagonal, the LASSO functional becomes additive and the LASSO will minimize each individual term of the form $\frac{1}{2}\|Xb^{(i)} - y^{(i)}\|_{\ell_2}^2 + \lambda\|b^{(i)}\|_{\ell_1}$, where $b^{(i)} = (b_{2i-1}, b_{2i})$ and $y^{(i)} = (y_{2i-1}, y_{2i})$. If for any of these subproblems, $\beta^{(i)} = \pm\epsilon^{-1}(1, -1)$ as in our 2-by-2 example above, then the squared error will blow up (as ϵ gets smaller) with probability π_0. With i fixed, $\mathbb{P}\left(\beta^{(i)} = \pm\epsilon^{-1}(1, -1)\right) = 2/n$ and thus the probability that none of these sub-problems is poised to blow up is $\left(1 - \frac{2}{n}\right)^{\frac{n}{2}} \to \frac{1}{e}$ as $n \to \infty$. Formalizing matters, we have a squared loss of at least $2/\epsilon$ with probability at least $\pi_0\left(1 - \left(1 - \frac{2}{n}\right)^{\frac{n}{2}}\right)$. Note that when n is large, λ is large so that π_0 is close to 1, and the LASSO badly misbehaves with a probability greater or equal to a quantity approaching $1 - 1/e$.

In conclusion, the LASSO may perform badly—even with a random β—when all our assumptions

[6] π_0 can be calculated since $X^* z$ is a bivariate Gaussian variable.

[7] We could alternatively select the support at random and randomly assign the signs and this would not change our story in the least.

are met but the coherence property. To summarize, an upper bound on the coherence is also necessary.

3.3 Proofs

In this section, we prove all of our results. It is sufficient to establish our theorems with $\sigma = 1$ as the general case is treated by a simple rescaling. Therefore, we conveniently assume $\sigma = 1$ from now on. Here and in the remainder of this chapter, x_T is the restriction of the vector x to an index set T, and for a matrix X, X_T is the submatrix formed by selecting the columns of X with indices in T. In the following, it will also be convenient to denote by K the functional

$$K(y,b) = \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + 2\lambda_p \|b\|_{\ell_1} \tag{3.3.1}$$

in which $\lambda_p = \sqrt{2\log p}$.

3.3.1 Preliminaries

We will make frequent use of subgradients and we begin by briefly recalling what these are. We say that $u \in \mathbb{R}^p$ is a subgradient of a convex function $f : \mathbb{R}^p \to \mathbb{R}$ at x_0 if f obeys

$$f(x) \geq f(x_0) + \langle u, x - x_0 \rangle \tag{3.3.2}$$

for all x.

Further, our arguments will repeatedly use two general results that we now record. The first states that the LASSO estimate is feasible for the Dantzig selector optimization problem.

Lemma 3.3.1 *The LASSO estimate obeys*

$$\|X^*(y - X\hat{\beta})\|_{\ell_\infty} \leq 2\lambda_p. \tag{3.3.3}$$

Proof Since $\hat{\beta}$ minimizes $f(b) = K(y,b)$ over b, 0 must be a subgradient of f at $\hat{\beta}$. Now the subgradients of f at b are of the form

$$X^*(Xb - y) + 2\lambda_p \epsilon,$$

where ϵ is any p-dimensional vector obeying $\epsilon_i = \mathrm{sgn}(b_i)$ if $b_i \neq 0$ and $|\epsilon_i| \leq 1$ otherwise. Hence, since 0 is a subgradient at $\hat{\beta}$, there exists ϵ as above such that

$$X^*(X\hat{\beta} - y) = -2\lambda_p \epsilon.$$

The conclusion follows from $\|\epsilon\|_{\ell_\infty} \le 1$. ∎

The second general result states that $\|X^* z\|_{\ell_\infty}$ cannot be too large. With large probability, $z \sim \mathcal{N}(0, I)$ obeys

$$\|X^* z\|_{\ell_\infty} = \max_i |\langle X_i, z\rangle| \le \lambda_p. \tag{3.3.4}$$

This is standard and simply follows from the fact that $\langle X_i, z\rangle \sim \mathcal{N}(0,1)$. Hence for each $t > 0$,

$$\mathbb{P}(\|X^* z\|_{\ell_\infty} > t) \le 2p \cdot \phi(t)/t, \tag{3.3.5}$$

where $\phi(t) \equiv (2\pi)^{-1/2} e^{-t^2/2}$. Better bounds may be possible but we will not pursue these refinements here. Also note that $\|X^* z\|_{\ell_\infty} \le \sqrt{2}\lambda_p$ with probability at least $1 - p^{-1}(2\pi \log p)^{-1/2}$. These two general facts have an interesting consequence since it follows from the decomposition $y = X\beta + z$ and the triangle inequality that with high probability

$$\begin{aligned} \|X^* X(\beta - \hat{\beta})\|_{\ell_\infty} &\le \|X^*(X\beta - y)\|_{\ell_\infty} + \|X^*(y - X\hat{\beta})\|_{\ell_\infty} \\ &= \|X^* z\|_{\ell_\infty} + \|X^*(y - X\hat{\beta})\|_{\ell_\infty} \\ &\le (\sqrt{2} + 2)\lambda_p. \end{aligned} \tag{3.3.6}$$

3.3.2 Proof of Theorem 3.1.4

Put T for the support of β. To prove our claim, we first establish that (3.1.8) holds provided that the following three deterministic conditions are satisfied.

- *Invertibility condition.* The submatrix $X_T^* X_T$ is invertible and obeys

$$\|(X_T^* X_T)^{-1}\| \le 2. \tag{3.3.7}$$

  The number 2 is arbitrary; we just need the smallest eigenvalue of $X_T^* X_T$ to be bounded away from zero.

- *Orthogonality condition.* The vector z obeys $\|X^* z\|_{\ell_\infty} \le \sqrt{2}\lambda_p$.

- *Complementary size condition.* The following inequality holds

$$\|X_{I^c}^* X_T (X_T^* X_T)^{-1} X_T^* z\|_{\ell_\infty} + 2\lambda_p \|X_{I^c}^* X_T (X_T^* X_T)^{-1} \mathrm{sgn}(\beta_T)\|_{\ell_\infty} \le (2 - \sqrt{2})\lambda_p. \tag{3.3.8}$$

This section establishes the main estimate (3.1.8) assuming these three conditions hold whereas the next will show that all three conditions hold with large probability—hence proving Theorem 3.1.4. Note that when z is white noise, we already know that the orthogonality condition holds with probability at least $1 - p^{-1}(2\pi \log p)^{-1/2}$.

Assume then that all three conditions above hold. Since $\hat{\beta}$ minimizes $K(y,b)$, we have $K(y,\hat{\beta}) \leq K(y,\beta)$ or equivalently

$$\frac{1}{2}\|y - X\hat{\beta}\|_{\ell_2}^2 + 2\lambda_p\|\hat{\beta}\|_{\ell_1} \leq \frac{1}{2}\|y - X\beta\|_{\ell_2}^2 + 2\lambda_p\|\beta\|_{\ell_1}.$$

Set $h = \hat{\beta} - \beta$ and note that

$$\|y - X\hat{\beta}\|_{\ell_2}^2 = \|(y - X\beta) - Xh\|_{\ell_2}^2 = \|Xh\|_{\ell_2}^2 + \|y - X\beta\|_{\ell_2}^2 - 2\langle Xh, y - X\beta\rangle.$$

Plugging this identity with $z = y - X\beta$ into the above inequality and rearranging the terms gives

$$\frac{1}{2}\|Xh\|_{\ell_2}^2 \leq \langle Xh, z\rangle + 2\lambda_p\left(\|\beta\|_{\ell_1} - \|\hat{\beta}\|_{\ell_1}\right). \tag{3.3.9}$$

Next, break h up into h_T and h_{I^c} (observe that $\hat{\beta}_{I^c} = h_{I^c}$) and rewrite (3.3.9) as

$$\frac{1}{2}\|Xh\|_{\ell_2}^2 \leq \langle h, X^*z\rangle + 2\lambda_p\left(\|\beta_T\|_{\ell_1} - \|\beta_T + h_T\|_{\ell_1} - \|h_{I^c}\|_{\ell_1}\right).$$

For each $i \in I$, we have

$$|\hat{\beta}_i| = |\beta_i + h_i| \geq |\beta_i| + \operatorname{sgn}(\beta_i)\, h_i$$

and thus, $\|\beta_T + h_T\|_{\ell_1} \geq \|\beta\|_{\ell_1} + \langle h_T, \operatorname{sgn}(\beta_T)\rangle$. Inserting this inequality above yields

$$\frac{1}{2}\|Xh\|_{\ell_2}^2 \leq \langle h, X^*z\rangle - 2\lambda_p(\langle h_T, \operatorname{sgn}(\beta_T)\rangle + \|h_{I^c}\|_{\ell_1}). \tag{3.3.10}$$

Observe now that $\langle h, X^*z\rangle = \langle h_T, X_T^*z\rangle + \langle h_{I^c}, X_{I^c}^*z\rangle$ and that the orthogonality condition implies

$$\langle h_{I^c}, X_{I^c}^*z\rangle \leq \|h_{I^c}\|_{\ell_1}\|X_{I^c}^*z\|_{\ell_\infty} \leq \sqrt{2}\lambda_p\|h_{I^c}\|_{\ell_1}.$$

The conclusion is the following useful estimate

$$\frac{1}{2}\|Xh\|_{\ell_2}^2 \leq \langle h_T, v\rangle - (2 - \sqrt{2})\lambda_p\|h_{I^c}\|_{\ell_1}, \tag{3.3.11}$$

where $v \equiv X_T^*z - 2\lambda_p \operatorname{sgn}(\beta_T)$.

We complete the argument by bounding $\langle h_T, v\rangle$. The key here is to use the fact that $\|X^*Xh\|_{\ell_\infty}$ is known to be small as pointed out by Terence Tao. We have

$$\begin{aligned}\langle h_T, v\rangle &= \langle (X_T^*X_T)^{-1} X_T^*X_T h_T, v\rangle \\ &= \langle X_T^*X_T h_T, (X_T^*X_T)^{-1} v\rangle \\ &= \langle X_T^*Xh, (X_T^*X_T)^{-1} v\rangle - \langle X_T^*X_{I^c} h_{I^c}, (X_T^*X_T)^{-1} v\rangle \equiv A_1 - A_2. \end{aligned} \tag{3.3.12}$$

We address each of the two terms individually. First,

$$A_1 \le \|X_T^*Xh\|_{\ell_\infty} \cdot \|(X_T^*X_T)^{-1}v\|_{\ell_1}$$

and

$$\begin{aligned}\|(X_T^*X_T)^{-1}v\|_{\ell_1} &\le \sqrt{s}\cdot \|(X_T^*X_T)^{-1}v\|_{\ell_2} \\ &\le \sqrt{s}\cdot \|(X_T^*X_T)^{-1}\|\,\|v\|_{\ell_2} \\ &\le s\cdot \|(X_T^*X_T)^{-1}\|\,\|v\|_{\ell_\infty}.\end{aligned}$$

Because 1) $\|X_T^*Xh\|_{\ell_\infty} \le (2+\sqrt{2})\,\lambda_p$ by Lemma 3.3.1 together with the orthogonality condition (see (3.3.6)) and 2) $\|(X_T^*X_T)^{-1}\|_{\ell_2} \le 2$ by the invertibility condition, we have

$$A_1 \le 2(2+\sqrt{2})\,\lambda_p\, s\|v\|_{\ell_\infty}.$$

However,

$$\|v\|_{\ell_\infty} \le \|X_T^*z\|_{\ell_\infty} + 2\lambda_p \le (2+\sqrt{2})\lambda_p.$$

so that

$$A_1 \le 2\,(2+\sqrt{2})^2\,\lambda_p^2 \cdot s. \tag{3.3.13}$$

Second, we simply bound the other term $A_2 = \langle h_{I^c}, X_{I^c}^*X_T(X_T^*X_T)^{-1}v\rangle$ by

$$|A_2| \le \|h_{I^c}\|_{\ell_1}\,\|X_{I^c}^*X_T(X_T^*X_T)^{-1}v\|_{\ell_\infty}$$

with $v = X_T^*z - 2\lambda_p\,\mathrm{sgn}(\beta_T)$. Since

$$\begin{aligned}\|X_{I^c}^*X_T(X_T^*X_T)^{-1}v\|_{\ell_\infty} &\le \|X_{I^c}^*X_T(X_T^*X_T)^{-1}X_T^*z\|_{\ell_\infty} + 2\lambda_p\|X_{I^c}^*X_T(X_T^*X_T)^{-1}\mathrm{sgn}(\beta_T)\|_{\ell_\infty} \\ &\le (2-\sqrt{2})\,\lambda_p\end{aligned}$$

because of the complementary size condition, we have

$$|A_2| \le (2-\sqrt{2})\lambda_p \|h_{I^c}\|_{\ell_1}.$$

To summarize,

$$|\langle h_T, v\rangle| \le 2\,(2+\sqrt{2})^2\,\lambda_p^2 \cdot s + (2-\sqrt{2})\lambda_p \|h_{I^c}\|_{\ell_1}. \tag{3.3.14}$$

We conclude by inserting (3.3.14) into (3.3.11) which gives

$$\frac{1}{2}\|X(\hat{\beta}-\beta)\|_{\ell_2}^2 \le 2\,(2+\sqrt{2})^2\,\lambda_p^2 \cdot s,$$

which is what we needed to prove.

3.3.3 Norms of random submatrices

In this section we establish that the invertibility and the complementary size conditions hold with large probability. These essentially rely on a recent result of Joel Tropp, which we state first.

Theorem 3.3.2 *[152] Suppose that a set T of predictors is sampled using a Bernoulli model by first creating a sequence $(\delta_j)_{1\le j\le p}$ of i.i.d. random variables with $\delta_j = 1$ w.p. s/p and $\delta_j = 0$ w.p. $1 - s/p$, and then setting $I \equiv \{j : \delta_j = 1\}$ so that $\mathbb{E}\,|T| = s$. Then for $q = 2\log p$,*

$$(\mathbb{E}\,\|X_T^* X_T - \mathrm{Id}\|^q)^{1/q} \le 30\mu(X)\,\log p + 13\sqrt{\frac{2s\,\|X\|^2\,\log p}{p}} \tag{3.3.15}$$

provided that $s\|X\|^2/p \le 1/4$. In addition, for the same value of q

$$(\mathbb{E}\max_{i\in I^c}\|X_T^* X_i\|_{\ell^2}^q)^{1/q} \le 4\mu(X)\,\sqrt{\log p} + \sqrt{s\,\|X\|^2/p}. \tag{3.3.16}$$

The first inequality (3.3.15) can be derived from the last equation in Section 4 of [152]. To be sure, using the notations of that chapter and letting $H \equiv X^*X - \mathrm{Id}$, Tropp shows that

$$\mathbb{E}_q\,\|RHR\| \le 15\bar{q}\,\mathbb{E}_q\,\|RHR'\|_{\max} + 12\sqrt{\delta\bar{q}}\|HR\|_{1\to 2} + 2\delta\|H\|, \quad \delta = s/p,$$

where $\bar{q} = \max\{q, 2\log p\}$. Now consider the following three facts: 1) $\|RHR'\|_{\max} \le \mu(X)$; 2) $\|HR\|_{1\to 2} \le \|X\|$; and 3) $\|H\| \le \|X\|^2$. The first assertion is immediate. The second is justified in [152]. For the third, observe that $\|X^*X - \mathrm{Id}\| \le \max\{\|X\|^2 - 1, 1\}$ (this is an equality when $p > n$) and the claim follows from $\|X\| \ge 1$, which holds since X has unit-normed columns. With $q = 2\log p$, this gives

$$\mathbb{E}_q\,\|RHR\| \le 30\mu(X)\,\log p + 12\sqrt{\frac{2s\log p\,\|X\|^2}{p}} + \frac{2s\|X\|^2}{p}.$$

Suppose that $s\|X\|^2/p \le 1/4$, then we can simplify the above inequality and obtain

$$\mathbb{E}_q \|RHR\| \le 30\mu(X)\log p + (12\sqrt{2\log p}+1)\sqrt{s\|X\|^2/p},$$

which implies (3.3.15). The second inequality (3.3.16) is exactly Corollary 5.1 in [152].

The inequalities (3.3.15) and (3.3.16) also hold for our slightly different model in which $I \subset \{1,\dots,p\}$ is a random subset of predictors with s elements provided that the right-hand side of both inequalities be multiplied by $2^{1/q}$. This follows from a simple Poissonization argument, which is similar to that posed in the proof of Lemma 3.3.6.

It is now time to investigate how these results imply our conditions, and we first examine how (3.3.15) implies the invertibility condition. Let T be a random set and put $Z = \|X_T^* X_T - \mathrm{Id}\|$. Clearly, if $Z \le 1/2$, then all the eigenvalues of $X_T^* X_T$ are in the interval $[1/2, 3/2]$ and $\|(X_T^* X_T)^{-1}\| \le 2$. Suppose that $\mu(X)$ and s are sufficiently small so that the right-hand side of (3.3.15) is less than $1/4$, say. This happens when the coherence $\mu(X)$ and s obey the hypotheses of the theorem. Then by Markov's inequality, we have that for $q = 2\log p$,

$$\mathbb{P}(Z > 1/2) \le 2^q\ \mathbb{E}\, Z^q \le (1/2)^q.$$

In other words the invertibility condition holds with probability exceeding $1 - p^{-2\log 2}$.

Recalling that the signs of the nonzero entries of β are i.i.d. symmetric variables, we now examine the complementary size condition and begin with a simple lemma.

Lemma 3.3.3 *Let $(W_j)_{j\in J}$ be a fixed collection of vectors in $\ell_2(I)$ and consider the random variable Z_0 defined by $Z_0 = \max_{j\in J} |\langle W_j, sgn(\beta_T)\rangle|$. Then*

$$\mathbb{P}(Z_0 \ge t) \le 2|J| \cdot e^{-t^2/2\kappa^2}, \tag{3.3.17}$$

for any κ obeying $\kappa \ge \max_{j\in J} \|W_j\|_{\ell_2}$. Similarly, letting $(W_j')_{j\in J}$ be a fixed collection of vectors in $\mathbb{R}^n$ and setting $Z_1 = \max_{j\in J} |\langle W_j', z\rangle|$, we have

$$\mathbb{P}(Z_1 \ge t) \le 2|J| \cdot e^{-t^2/2\kappa^2}, \tag{3.3.18}$$

for any κ obeying $\kappa \ge \max_{j\in J} \|W_j'\|_{\ell_2}$.[8]

[8]Note that this lemma also holds if the collection of vectors $(W_j)_{j\in J}$ is random, as long as it is independent of $\mathrm{sgn}(\beta_T)$ and z.

Proof The first inequality is an application of Hoeffding's inequality. Indeed, letting $Z_{0,j} = \langle W_j, \mathrm{sgn}(\beta_T)\rangle$, Hoeffding's inequality gives

$$\mathbb{P}(|Z_{0,j}| > t) \le 2e^{-t^2/2\|W_j\|_{\ell_2}^2} \le 2e^{-t^2/2\max_j \|W_j\|_{\ell_2}^2}. \tag{3.3.19}$$

Inequality (3.3.17) then follows from the union bound. The second part is even easier since $Z_{1,j} = \langle W_j', z\rangle \sim \mathcal{N}(0, \|W_j'\|_{\ell_2}^2)$ and thus

$$\mathbb{P}(|Z_{1,j}| > t) \le 2e^{-t^2/2\|W_j'\|_{\ell_2}^2} \le 2e^{-t^2/2\max_j \|W_j'\|_{\ell_2}^2}. \tag{3.3.20}$$

Again, the union bound gives (3.3.18). ∎

For each $i \in I^c$, define $Z_{0,i}$ and $Z_{1,i}$ as

$$Z_{0,i} = X_i^* X_T (X_T^* X_T)^{-1} \mathrm{sgn}(\beta_T) \quad \text{and} \quad Z_{1,i} = X_i^* X_T (X_T^* X_T)^{-1} X_T^* z.$$

With these notations, in order to prove the complementary size condition, it is sufficient to show that with large probability,

$$2\lambda_p Z_0 + Z_1 \le (2 - \sqrt{2})\lambda_p,$$

where $Z_0 = \max_{i\in I^c} |Z_{0,i}|$ and likewise for Z_1. Therefore, it is sufficient to prove that with large probability

$$Z_0 \le 1/4 \quad \text{and} \quad Z_1 \le (3/2 - \sqrt{2})\lambda_p.$$

The idea is of course to apply Lemma 5.2.5 together with Theorem 3.3.2. We have

$$Z_{0,i} = \langle W_i, \mathrm{sgn}(\beta_T)\rangle \quad \text{and} \quad Z_{1,i} = \langle W_i', z\rangle,$$

where

$$W_i = (X_T^* X_T)^{-1} X_T^* X_i \quad \text{and} \quad W_i' = X_T (X_T^* X_T)^{-1} X_T^* X_i.$$

Recall the definition of Z above and consider the event $E = \{Z \le 1/2\} \cap \{\max_{i\in I^c} \|X_T^* X_i\| \le \gamma\}$ for some positive γ. On this event, all the singular values of X_T are between $1/\sqrt{2}$ and $\sqrt{3/2}$, and thus $\|(X_T^* X_T)^{-1}\| \le 2$ and $\|X_T (X_T^* X_T)^{-1}\| \le \sqrt{2}$, which gives

$$\|W_i\| \le 2\gamma, \quad \text{and} \quad \|W_i'\| \le \sqrt{2}\gamma.$$

Applying (3.3.17) and (3.3.18) gives

$$\begin{aligned}\mathbb{P}(\{Z_0 \geq t\} \cup \{Z_1 \geq u\}) &\leq \mathbb{P}(\{Z_0 \geq t\} \cup \{Z_1 \geq u\} \mid E) + \mathbb{P}(E^c) \\ &\leq \mathbb{P}(Z_0 \geq t \mid E) + \mathbb{P}(Z_1 \geq u \mid E) + \mathbb{P}(E^c) \\ &\leq 2p\, e^{-t^2/8\gamma^2} + 2p\, e^{-u^2/4\gamma^2} + \mathbb{P}(Z > 1/2) + \mathbb{P}(\max_{i \in I^c} \|X_T^* X_i\| > \gamma).\end{aligned}$$

We already know that the second-to-last term of the right-hand side is polynomially small in p provided that $\mu(X)$ and s obey the conditions of the theorem. For the other three terms let γ_0 be the right-hand side of (3.3.16). For $t = 1/4$, one can find a constant c_0 such that if $\gamma < c_0/\sqrt{\log p}$, then $2pe^{-t^2/8\gamma^2} \leq 2p^{-2\log 2}$, say. Likewise, for $u = (3/2 - \sqrt{2})\lambda_p$, we may have $2pe^{-u^2/4\gamma^2} \leq 2p^{-2\log 2}$. The last term is treated by Markov's inequality since for $q = 2\log p$, (3.3.16) gives

$$\mathbb{P}(\max_{i \in I^c} \|X_T^* X_i\| > \gamma) \leq \gamma^{-q} \cdot \mathbb{E}(\max_{i \in I^c} \|X_T^* X_i\|^q) \leq (\gamma_0/\gamma)^q.$$

Therefore, if $\gamma_0 \leq \gamma/2 = c_0/2\sqrt{\log p}$, we have that this last term does not exceed $1 - p^{-2\log 2}$. For $\mu(X)$ and s obeying the hypotheses of Theorem 3.1.4, it is indeed the case that $\gamma_0 \leq c_0/2\sqrt{\log p}$. In conclusion, we have shown that all three conditions hold under our hypotheses with probability at least $1 - 6p^{-2\log 2} - p^{-1}(2\pi \log p)^{-1/2}$.

In passing, we would like to remark that proving that $Z_0 \leq 1/4$ establishes that the strong irrepresentable condition from [169] holds (with high probability). This condition states if T is the support of β

$$\|X_{I^c}^* X_T (X_T^* X_T)^{-1} \mathrm{sgn}(\beta_T)\|_{\ell_\infty} \leq 1 - \nu$$

where ν is any (small) constant greater than zero (this condition is used to show the asymptotic recovery of the support of β).

3.3.4 Proof of Theorem 3.1.6

The proof of Theorem 3.1.6 parallels that of Theorem 3.1.4 and we only sketch it although we carefully detail the main differences. Let T_0 be the support of β_0. Just as before, all three conditions of Section 3.3.2 with T_0 in place of T and β_0 in place of β hold with overwhelming probability. From now on, we just assume that they are all true.

Since $\hat{\beta}$ minimizes $K(y, b)$, we have $K(y, \hat{\beta}) \leq K(y, \beta_0)$ or equivalently

$$\frac{1}{2}\|y - X\hat{\beta}\|_{\ell_2}^2 + 2\lambda_p \|\hat{\beta}\|_{\ell_1} \leq \frac{1}{2}\|y - X\beta_0\|_{\ell_2}^2 + 2\lambda_p \|\beta_0\|_{\ell_1}. \tag{3.3.21}$$

Expand $\|y - X\hat{\beta}\|_{\ell_2}^2$ as

$$\|y - X\hat{\beta}\|_{\ell_2}^2 = \|z - (X\hat{\beta} - X\beta)\|_{\ell_2}^2 = \|z\|_{\ell_2}^2 - 2\langle z, X\hat{\beta} - X\beta\rangle + \|X\hat{\beta} - X\beta\|_{\ell_2}^2$$

and $\|y - X\beta_0\|_{\ell_2}^2$ in the same way. Then plug these identities in (3.3.21) to obtain

$$\frac{1}{2}\|X\hat{\beta} - X\beta\|_{\ell_2}^2 \le \frac{1}{2}\|X\beta_0 - X\beta\|_{\ell_2}^2 + \langle z, X\hat{\beta} - X\beta_0\rangle + 2\lambda_p\left(\|\beta_0\|_{\ell_1} - \|\hat{\beta}\|_{\ell_1}\right). \tag{3.3.22}$$

Put $h = \hat{\beta} - \beta_0$. We follow the same steps as in Section 3.3.2 to arrive at

$$\frac{1}{2}\|X\hat{\beta} - X\beta\|_{\ell_2}^2 \le \frac{1}{2}\|X\beta_0 - X\beta\|_{\ell_2}^2 + \langle h_{T_0}, v\rangle - (2 - \sqrt{2})\lambda_p \|h_{T_0^c}\|_{\ell_1},$$

where $v = X_{T_0}^* z - 2\lambda_p \mathrm{sgn}(\beta_{T_0})$. Just as before,

$$\langle h_{T_0}, v\rangle = \langle X_{T_0}^* X h, \left(X_{T_0}^* X_{T_0}\right)^{-1} v\rangle - \langle h_{T_0^c}, X_{T_0}^* X_{T_0^c}\left(X_{T_0}^* X_{T_0}\right)^{-1} v\rangle \equiv A_1 - A_2.$$

By assumption $|A_2| \le (2 - \sqrt{2})\lambda_p \cdot \|h_{T_0^c}\|_{\ell_1}$. The difference is now in A_1 since we can no longer claim that $\|X^* X h\|_{\ell_\infty} \le (2 + \sqrt{2})\lambda_p$. Decompose A_1 as

$$A_1 = \langle X_{T_0}^* X(\hat{\beta} - \beta), \left(X_{T_0}^* X_{T_0}\right)^{-1} v\rangle + \langle X_{T_0}^* X(\beta - \beta_0), \left(X_{T_0}^* X_{T_0}\right)^{-1} v\rangle \equiv A_1^0 + A_1^1.$$

Because $\|X^* X(\hat{\beta} - \beta)\|_{\ell_\infty} \le (2 + \sqrt{2})\lambda_p$, one can use the same argument as before to obtain

$$A_1^0 \le 2(2 + \sqrt{2})^2 \lambda_p^2 s.$$

We now look at the other term. Since $\|X_{T_0}\left(X_{T_0}^* X_{T_0}\right)^{-1}\| \le \sqrt{2}$ by assumption, we have

$$\begin{aligned} |A_1^1| &= \langle X(\beta - \beta_0), X_{T_0}\left(X_{T_0}^* X_{T_0}\right)^{-1} v\rangle \\ &\le \|X(\beta - \beta_0)\|_{\ell_2} \|X_{T_0}\left(X_{T_0}^* X_{T_0}\right)^{-1} v\|_{\ell_2} \\ &\le \sqrt{2}\|X(\beta - \beta_0)\|_{\ell_2} \|v\|_{\ell_2}. \end{aligned}$$

Using $ab \le (a^2 + b^2)/2$ and $\|v\|_{\ell_2}^2 \le (2 + \sqrt{2})^2 \lambda_p^2 s$ gives

$$|A_1^1| \le \frac{\sqrt{2}}{2}\|X(\beta - \beta_0)\|_{\ell_2}^2 + \frac{\sqrt{2}}{2}(2 + \sqrt{2})^2 \lambda_p^2 s.$$

To summarize

$$\langle h_{T_0}, v\rangle \le \frac{\sqrt{2}}{2}\|X(\beta - \beta_0)\|_{\ell_2}^2 + \left(2 + \frac{\sqrt{2}}{2}\right)(2 + \sqrt{2})^2 \lambda_p^2 s + (2 - \sqrt{2})\lambda_p \cdot \|h_{T_0^c}\|_{\ell_1}.$$

It follows that

$$\frac{1}{2}\|X\hat{\beta} - X\beta\|_{\ell_2}^2 \leq \frac{1+\sqrt{2}}{2}\|X\beta_0 - X\beta\|_{\ell_2}^2 + (4+\sqrt{2})(1+\sqrt{2})^2\lambda_p^2 s.$$

This concludes the proof.

We close this section by arguing about (3.1.20) and (3.1.21). First, it follows from our proof that (3.1.20) holds. And second, our analysis also shows that the set $\mathcal{A}_{0,s}$ is very large and obeys (3.1.21).

3.3.5 Proof of Theorem 3.1.5

Just as with our other claims, we begin by stating a few assumptions which hold with very large probability, and then show that under these conditions, the conclusions of the theorem hold. These assumptions are stated below.

(i) The matrix $X_T^* X_T$ is invertible and obeys $\|(X_T^* X_T)^{-1}\| \leq 2$.

(ii) $\|X_{I^c}^* X_T \left(X_T^* X_T\right)^{-1} \mathrm{sgn}(\beta_T)\|_{\ell_\infty} < \frac{1}{4}$.

(iii) $\| \left(X_T^* X_T\right)^{-1} X_T^* z\|_{\ell_\infty} \leq 2\lambda_p$.

(iv) $\|X_{I^c}^*(I - P[T])z\|_{\ell_\infty} \leq \sqrt{2}\lambda_p$.

(v) The matrix-vector product $\left(X_T^* X_T\right)^{-1} \mathrm{sgn}(\beta_T)$ obeys

$$\| \left(X_T^* X_T\right)^{-1} \mathrm{sgn}(\beta_T)\|_{\ell_\infty} \leq 3. \tag{3.3.23}$$

We already know that conditions (i) and (ii) hold with large probability, see Section 3.3.3 (the change from 1/2 to 1/4 in (ii) is unessential). As before, we let E be the event $\{\|X_T^* X_T - \mathrm{Id}\| \leq 1/2\}$. For (iii), the idea is the same and we express $\| \left(X_T^* X_T\right)^{-1} X_T^* z\|_{\ell_\infty}$ as $\max_{i\in I} |\langle W_i, z\rangle|$, where W_i is now the ith row of $\left(X_T^* X_T\right)^{-1} X_T^*$. On E, $\max_i \|W_i\| \leq \| \left(X_T^* X_T\right)^{-1} X_T^*\| \leq \sqrt{2}$ and the claim now follows from (3.3.5). Indeed, one can check that conditional on E

$$\mathbb{P}(\| \left(X_T^* X_T\right)^{-1} X_T^* z\|_{\ell_\infty} > 2\,\lambda_p) \leq |T| \cdot p^{-2} \cdot (2\pi \log p)^{-1/2}.$$

For (iv), we write $\|X_{I^c}^*(I - P[T])z\|_{\ell_\infty}$ as $\max_{i\in I^c} |\langle W_i, z\rangle|$ where $W_i = (I - P[T])X_i$. We have $\|W_i\| \leq \|X_i\| = 1$ and conditional on E, it follows from (3.3.5)

$$\mathbb{P}(\|X_{I^c}^*(I - P[T])z\|_{\ell_\infty} > \sqrt{2}\lambda_p) \leq |I^c| \cdot p^{-2} \cdot (2\pi \log p)^{-1/2}.$$

The subtle estimate is (v) and is proven in the next section. There, we show that (3.3.23) holds with probability at least $1 - 2p^{-2\log 2} - 2|T|\,p^{-2}$. Hence, under the assumptions of Theorem 3.1.5, (i)–(v) hold with probability at least $1 - 2p^{-1}\left((2\pi\log p)^{-1/2} + |T|/p\right) - O(p^{-2\log 2})$.

Lemma 3.3.4 *Suppose that the assumptions (i)–(v) hold and assume that* $\min_{i\in I}|\beta_i|$ *obeys the condition of Theorem 3.1.5. Then the LASSO solution is given by* $\hat\beta \equiv \beta + h$ *with*

$$\begin{aligned} h_T &= (X_T^* X_T)^{-1}\left[X_T^* z - 2\lambda_p sgn(\beta_T)\right], \\ h_{I^c} &= 0. \end{aligned} \tag{3.3.24}$$

Proof The point $\hat\beta$ is the unique solution to the LASSO functional if

$$\begin{aligned} X_i^*(y - X\hat\beta) &= 2\lambda_p \operatorname{sgn}(\hat\beta_i), && \hat\beta_i \neq 0, \\ |X_i^*(y - X\hat\beta)| &< 2\lambda_p, && \hat\beta_i = 0, \end{aligned} \tag{3.3.25}$$

and the columns of X_T are linearly independent where T is the support of $\hat\beta$. Consider then h as in (3.3.24) and observe that

$$\|h_T\|_{\ell_\infty} \le \|(X_T^* X_T)^{-1} X_T^* z\|_{\ell_\infty} + 2\lambda_p \|(X_T^* X_T)^{-1}\operatorname{sgn}(\beta_T)\|_{\ell_\infty} \le 2\lambda_p + 6\lambda_p.$$

It follows that $\|h_T\|_{\ell_\infty} < \min_{i\in I}|\beta_i|$ and, therefore, $\hat\beta = \beta + h$ obeys

$$\begin{aligned} \operatorname{supp}(\hat\beta) &= \operatorname{supp}(\beta), \\ \operatorname{sgn}(\hat\beta_T) &= \operatorname{sgn}(\beta_T). \end{aligned}$$

We now check that $\hat\beta = \beta + h$ obeys (3.3.25). By definition, we have

$$y - X\hat\beta = z - Xh = z - X_T(X_T^* X_T)^{-1}\left[X_T^* z - 2\lambda_p \operatorname{sgn}(\hat\beta_T)\right]$$

since β and $\hat\beta$ share the same support and the same signs. Clearly,

$$X_T^*(y - X\hat\beta) = 2\lambda_p \operatorname{sgn}(\hat\beta_T),$$

which is the first half of (3.3.25). For the second half, let $P[T] = X_T(X_T^*X_T)^{-1}X_T^*$ be the orthonormal projection onto the span of X_T. Then

$$\begin{aligned}\|X_{I^c}^*(y - X\hat{\beta})\|_{\ell_\infty} &= \|X_{I^c}^*(I - P[T])z + 2\lambda_p X_{I^c}^* X_T (X_T^*X_T)^{-1}\mathrm{sgn}(\beta_T)\|_{\ell_\infty} \\ &\le \|X_{I^c}^*(I - P[T])z\|_{\ell_\infty} + 2\lambda_p \|X_{I^c}^* X_T (X_T^*X_T)^{-1}\mathrm{sgn}(\beta_T)\|_{\ell_\infty} \\ &< \sqrt{2}\lambda_p + \frac{1}{2}\lambda_p \\ &< 2\,\lambda_p.\end{aligned}$$

Finally, note that $X_T^*X_T$ is indeed invertible since $T = I$; this is just our invertibility condition. This concludes the proof. ∎

Lemma 3.3.4 proves that $\hat{\beta}$ has the same support as β and the same signs as β, which is of course the content of Theorem 3.1.5.

3.3.6 Proof of (3.3.23)

We need to show that $\|\,(X_T^*X_T)^{-1}\,\mathrm{sgn}(\beta_T)\|_{\ell_\infty}$ is small with high probability and write

$$\begin{aligned}\|\,(X_T^*X_T)^{-1}\,\mathrm{sgn}(\beta_T)\|_{\ell_\infty} &\le \|\mathrm{sgn}(\beta_T)\|_{\ell_\infty} + \|((X_T^*X_T)^{-1} - \mathrm{Id})\mathrm{sgn}(\beta_T)\|_{\ell_\infty} \\ &\le 1 + \max_{i\in I} |\langle W_i, \mathrm{sgn}(\beta_T)\rangle|,\end{aligned}$$

where W_i is the ith row of $(X_T^*X_T)^{-1} - \mathrm{Id}$ (or column since this is a symmetric matrix).

Lemma 3.3.5 *Let W_i be the ith row of $(X_T^*X_T)^{-1} - \mathrm{Id}$. Under the hypotheses of Theorem 3.1.5, we have*

$$\mathbb{P}(\max_{i\in I} \|W_i\| \ge (\log p)^{-1/2}) \le 2p^{-2\log 2}.$$

Proof Set $A \equiv \mathrm{Id} - X_T^*X_T$. On the event $E \equiv \{\|\mathrm{Id} - X_T^*X_T\| \le 1/2\}$ (which holds w. p. at least $1 - p^{-2\log 2}$), we have

$$(X_T^*X_T)^{-1} = I + A + A^2 + \dots.$$

Therefore, since $W_i = ((X_T^*X_T)^{-1} - \mathrm{Id})e_i$ where e_i is the vector whose ith component is 1 and the others 0, $W_i = Ae_i + A^2e_i + \dots$ and

$$\begin{aligned}\|W_i\| &\le \|Ae_i\| + \|A\|\|Ae_i\| + \|A^2\|\|Ae_i\| + \dots \\ &\le \|Ae_i\| \sum_{k=0}^{\infty} \|A\|^k \\ &\le \|Ae_i\|/(1 - \|A\|).\end{aligned}$$

Hence on E, $\|W_i\| \le 2\|Ae_i\|$.

For each $i \in I$, Ae_i is the ith row or column of $\mathrm{Id} - X_T^* X_T$ and for each $j \in I$, its jth component is equal to $-\langle X_i, X_j\rangle$ if $j \ne i$, and 0 for $j = i$ since $\|X_i\| = 1$. Thus,

$$\|W_i\|^2 \le 4 \sum_{j\in I: j\ne i} |\langle X_i, X_j\rangle|^2.$$

Now it follows from Lemma 3.3.6 that

$$\sum_{j\in I: j\ne i} |\langle X_i, X_j\rangle|^2 \le s\|X\|^2/p + t$$

with probability at least $1 - 2e^{-t^2/[2\mu^2(X)(s\|X\|^2/p+t/3)]}$. Under the assumptions of Theorem 3.1.5, we have $s\|X\|^2/p \le c_0(\log p)^{-1} \le (8\log p)^{-1}$ provided that $c_0 \le 1/8$. With $t = (8\log p)^{-1}$, this gives

$$\sum_{j\in I: j\ne i} |\langle X_i, X_j\rangle|^2 \le 1/(4\log p) \tag{3.3.26}$$

with probability at least $1 - 2e^{-3/[64\mu^2(X)\log p]}$. Now the assumption about the coherence guarantees that $\mu(X) \le A_0/\log p$ so that (3.3.26) holds with probability at least $1 - 2e^{-3\log p/[64A_0^2]}$. Hence, by choosing A_0 sufficiently small, the lemma follows from the union bound. ■

Lemma 3.3.6 *Suppose that $I \subset \{1,\dots,p\}$ is a random subset of predictors with at most s elements. For each i, $1 \le i \le p$, we have*

$$\mathbb{P}\left(\sum_{j\in I: j\ne i} |\langle X_i, X_j\rangle|^2 > \frac{s}{p}\|X\|^2 + t\right) \le 2\exp\left(-\frac{t^2}{2\mu^2(X)(s\|X\|^2/p + t/3)}\right). \tag{3.3.27}$$

Proof The inequality (3.3.27) is essentially an application of Bernstein's inequality, which states that for a sum of uniformly bounded independent random variables with $|Y_k - \mathbb{E}\,Y_k| < c$,

$$\mathbb{P}\left(\sum_{k=1}^n (Y_k - \mathbb{E}\,Y_k) > t\right) \le e^{-t^2/(2\sigma^2 + 2ct/3)}, \tag{3.3.28}$$

where σ^2 is the sum of the variances, $\sigma^2 \equiv \sum_{k=1}^n \mathrm{Var}(Y_k)$. The issue here is that $\sum_{j\in I: j\ne i} |\langle X_i, X_j\rangle|^2$ is not a sum of independent variables and we need to use a kind of Poissonization argument to reduce this to a sum of independent terms.

A set I' of predictors is sampled using a Bernoulli model by first creating the sequence

$$\delta_j = \begin{cases} 1 & \text{w. p. } s/p, \\ 0 & \text{w. p. } 1 - s/p \end{cases}$$

and then setting $I' \equiv \{j \in \{1,\dots,p\} : \delta_j = 1\}$. The size of the set I' follows a binomial distribution, and $\mathbb{E}\,|I'| = s$. We make two claims: first, for each $t > 0$, we have

$$\mathbb{P}\Big(\sum_{j\in I: j\neq i} |\langle X_i, X_j\rangle|^2 > t\Big) \le 2\,\mathbb{P}\Big(\sum_{j\in I': j\neq i} |\langle X_i, X_j\rangle|^2 > t\Big); \tag{3.3.29}$$

second, for each $t > 0$

$$\mathbb{P}\Big(\sum_{j\in I': j\neq i} |\langle X_i, X_j\rangle|^2 > \frac{s}{p}\|X\|^2 + t\Big) \le \exp\left(-\frac{t^2}{2\mu^2(X)(s\|X\|^2/p + t/3)}\right). \tag{3.3.30}$$

Clearly, (3.3.29) and (3.3.30) give (3.3.27).

To justify the first claim, observe that

$$\begin{aligned}
\mathbb{P}\Big(\sum_{j\in I': j\neq i} |\langle X_i, X_j\rangle|^2 > t\Big) &= \sum_{k=0}^{p} \mathbb{P}\Big(\sum_{j\in I': j\neq i} |\langle X_i, X_j\rangle|^2 > t \mid |I'| = k\Big)\, P(|I'| = k) \\
&\ge \sum_{k=s}^{p} \mathbb{P}\Big(\sum_{j\in I': j\neq i} |\langle X_i, X_j\rangle|^2 > t \mid |I'| = k\Big)\, P(|I'| = k) \\
&= \sum_{k=s}^{p} \mathbb{P}\Big(\sum_{j\in I_k: j\neq i} |\langle X_i, X_j\rangle|^2 > t\Big)\, P(|I'| = k),
\end{aligned}$$

where I_k is selected uniformly at random with $|I_k| = k$. We make two observations: 1) since s is an integer, it is the median of $|I'|$ and $P(|I'| \ge s) \ge 1/2$; and 2) $\mathbb{P}(\sum_{j\in I_k: j\neq i} |\langle X_i, X_j\rangle|^2 > t)$ is a nondecreasing function of k. To see why this is true, consider that a subset I_{k+1} of size $k+1$ can be sampled by first choosing a subset I_k of size k uniformly, and then choosing the remaining entry uniformly at random from the complement of I_k. It follows that with $Z_k = \sum_{j\in I_k} |\langle X_i, X_j\rangle|^2\, 1_{\{i\neq j\}}$, we have that Z_{k+1} and $Z_k + Y_k$, where Y_k is a nonnegative random variable, have the same distribution. Hence $\mathbb{P}(Z_{k+1} \ge t) \ge \mathbb{P}(Z_k \ge t)$. With these two observations in mind, we continue

$$\begin{aligned}
\mathbb{P}\Big(\sum_{j\in I': j\neq i} |\langle X_i, X_j\rangle|^2 > t\Big) &\ge \mathbb{P}\Big(\sum_{j\in I: j\neq i} |\langle X_i, X_j\rangle|^2 > t\Big) \sum_{k=s}^{p} P(|I'| = k) \\
&\ge \frac{1}{2}\,\mathbb{P}\Big(\sum_{j\in I: j\neq i} |\langle X_i, X_j\rangle|^2 > t\Big),
\end{aligned}$$

which is the first claim (3.3.29).

For the second claim (3.3.30), observe that

$$\sum_{j\in I': j\neq i} |\langle X_i, X_j\rangle|^2 = \sum_{1\le j\le p: j\neq i} \delta_j\, |\langle X_i, X_j\rangle|^2 \equiv \sum_{1\le j\le p: j\neq i} Y_j.$$

The Y_j are independent and obey:

1. $|Y_j - \mathbb{E}\,Y_j| \le \sup_{j\neq i} |\langle X_i, X_j\rangle|^2 \le \mu^2(X)$.

2. The sum of means is bounded by

$$\sum_{1\le j\le p:j\neq i} \mathbb{E}\, Y_j = \frac{s}{p} \sum_{1\le j\le p:j\neq i} |\langle X_i, X_j\rangle|^2 \le \frac{s\|X\|^2}{p}.$$

   The last inequality follows from $\sum_{1\le j\le p:j\neq i} |\langle X_i, X_j\rangle|^2 \le \sum_{1\le j\le p} |\langle X_i, X_j\rangle|^2$ where the right-hand side is equal to $\|X^* X_i\|^2 \le \|X^*\|^2 \|X_i\|^2 = \|X\|^2$ since the columns are unit-normed.

3. The sum of variances is bounded by

$$\sum_{1\le j\le p:j\neq i} \mathrm{Var}(Y_j) = \frac{s}{p}\left(1-\frac{s}{p}\right) \sum_{1\le j\le p:j\neq i} |\langle X_i, X_j\rangle|^4 \le \frac{s\mu^2(X)\|X\|^2}{p}.$$

   The last inequality follows from $\sum_{1\le j\le p:j\neq i} |\langle X_i, X_j\rangle|^4 \le \mu^2(X) \sum_{1\le j\le p} |\langle X_i, X_j\rangle|^2$, which is less or equal to $\mu^2(X)\,\|X\|^2$ as before.

The claim (3.3.30) is now a simple application of Bernstein's inequality (3.3.27). ■

Lemma 3.3.5 establishes that (3.3.23) holds with probability at least $1-2p^{-2\log 2}-2|T|\,p^{-2}$. Indeed, on the event $\max_i \|W_i\| \le (\log p)^{-1/2}$, it follows from Lemma 5.2.5 that

$$\mathbb{P}(\max_{i\in I} |\langle W_i, \mathrm{sgn}(\beta_T)\rangle| \ge 2) \le 2|T|\, e^{-2\log p} \le 2|T|\, p^{-2}.$$

3.4 Discussion

3.4.1 Comparison to related theoretical results

As described in Section 3.1.2, a common assumption in the literature is that the sparsity is smaller than the inverse of the coherence, $1/\mu$. This assumption essentially requires $s \lesssim \sqrt{n}$. However, the LASSO is known to work very well empirically when the sparsity far exceeds this threshold [59]. Thus there is a disconnect between what experience shows and the requirements in the majority of the literature.

In the noiseless setting, this gap was bridged in Tropp's paper [153]. Our work bridges this gap in the noisy setting. We do so in the same way as Tropp: by considering the performance of the LASSO one expects in almost all cases but not all. By considering statistical ensembles much as in [37,153], one shows that in the above examples, the LASSO works provided that the sparsity level is bounded by about $n/\log p$; that is, for generic signals, the sparsity can grow almost linearly with the sample size. We also prove that under these conditions, the irrepresentable condition holds with high probability and we show that, as long as the entries of β are not too small, one can recover the exact support of β with high probability.

Finally, there does not seem much room for improvement as all of our conditions appear necessary as well. In Section 3.2, we have proposed special examples in which the LASSO performs poorly. On the one hand, these examples show that even with highly incoherent matrices, one cannot expect good performance in all cases unless the sparsity level is very small. And on the other hand, one cannot really eliminate our assumption about the coherence since we have shown that with coherent matrices, the LASSO would fail to work well on generically sparse objects.

One could of course consider other statistical descriptions of sparse β's and/or ideal models, and leave this issue open for further research. Further, one could consider the error in estimating β under weak coherence conditions; this appears to be a difficult theoretical problem.

Chapter 4

Low-rank matrix estimation with the RIP

4.1 Introduction

Low-rank matrix recovery is a burgeoning topic drawing the attention of many researchers in the closely related field of sparse approximation and compressive sensing. As noted in Chapter 1, in the matrix recovery problem, the signal to be recovered is a low-rank matrix $M \in \mathbb{R}^{n_1 \times n_2}$, about which we have information supplied by means of a linear operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ (typically, m is far less than $n_1 n_2$),

$$y = \mathcal{A}(M) + z, \quad y \in \mathbb{R}^m.$$

Here z is a stochastic noise term. Signal recovery appears to be an ill-posed problem because there are many more unknowns than equations. However, analogously to the results demonstrated for sparse approximation, taking into account the parsimony of the model may cause recovery to become feasible.

In this chapter, we derive similar results for matrix recovery to those available for compressed sensing and sparse approximation in general. However, in contrast to results available in the literature on compressive sensing or sparse regression, we show that the error bound is within a constant factor (rather than a log factor) of an idealized oracle error bound achieved by projecting the data onto a smaller subspace given by the oracle (and also within a constant of the minimax error bound). This error bound also applies to full-rank matrices (which are well-approximated by low-rank matrices), and there appears to be no analogue of this in the compressive sensing world.

Another contribution is to lower the number of measurements to stably recover a matrix of rank r by convex programming. It is not hard to see that we need at least $m \geq (n_1 + n_2 - r)r$ measurements to recover matrices of rank r, by any method whatsoever. To be sure, if $m < (n_1 + n_2 - r)r$, we will always have two distinct matrices M and M' or rank at most r with the property $\mathcal{A}(M) = \mathcal{A}(M')$

no matter what $\mathcal{A}$ is. To see this, fix two matrices $U \in \mathbb{R}^{n_1 \times r}$, $V \in \mathbb{R}^{n_2 \times r}$ with orthonormal columns, and consider the linear space of matrices of the form

$$T = \{UX^* - YV^* : X \in \mathbb{R}^{n_2 \times r}, Y \in \mathbb{R}^{n_1 \times r}\}.$$

The dimension of T is $r(n_1 + n_2 - r)$. Thus, if $m < (n_1 + n_2 - r)r$, there exists $M = UX^* - YV^* \neq 0$ in T such that $\mathcal{A}(M) = 0$. This proves the claim since $\mathcal{A}(UX^*) = \mathcal{A}(YV^*)$ for two distinct matrices of rank at most r. Now a novel result discussed in this chapter is that, even without knowing that $M \in T$, one can stably recover M from a constant times $(n_1 + n_2)r$ measurements via nuclear-norm minimization. Once again, in contrast to similar results in compressive sensing, the number of measurements required is within a constant of the theoretical lower limit—there is no extra log factor.

4.1.1 A few applications

Following a series of advances in the theory of low-rank matrix recovery from undersampled linear measurements [32,34,36,44,96,97,101,110,129], a number of new applications have sprung up to join ranks with the already established ones. A quick survey shows that low-rank modeling is getting very popular in science and engineering, and we present a few eclectic examples to illustrate this point.

- **Quantum state tomography [88].** In quantum state tomography, a mixed quantum state is represented as a square positive semidefinite matrix, M (with trace 1). If M is actually a pure state, then it has rank 1, and more generally, if it is approximately pure then it will be well approximated by a low-rank matrix [88].

- **Face recognition [13,32].** Let $\{y_i\}_{i=1}^n$ be a sequence of images (in vector form) of the same face under varying illumination. In theory and under idealized circumstances (the images are assumed to be convex, Lambertian objects), these faces all reside near the same nine-dimensional linear subspace [13]. Thus, the matrix created by stacking together the images is well approximated by a rank-9 matrix. In practice, face-recognition techniques based on the assumption that these images reside in a low-dimensional subspace are highly successful [13,32].

Quantum state tomography lends itself perfectly to the compressive sensing framework. On an abstract level, one sees measurements consisting of linear combinations of the unknown quantum state M—inner products with certain observables which can be chosen with some flexibility by the physicist—and the goal is to recover a good approximation of M. The size of M grows exponentially with the number of particles in the system, so one would like to use the structure of M to reduce the number of measurements required, thus necessitating compressive sensing (see [88] for a more

in depth discussion and a specific analysis of this problem). Depending upon the measurements used, the RIP may or may not be applicable. For the specific case of random Pauli measurements suggested in [88] it is presently unknown whether the RIP holds (with high probability). [1]

In an important application using the face recognition model, one sees the entirety of every face, except that a small subset of pixels may have very large errors (this may be caused by shadows or occlusions). In this case, the sampling operator is the identity. While our results can apply to this problem, the algorithms discussed in this chapter are intended for bounded, dense errors. (We include face recognition in our discussion to illustrate the different uses of the low-rank matrix model.) However, there is another nuclear-norm based minimization approach, specialized to deal with sparse errors, with strong guarantees for this model [32].

4.1.2 Related literature

There has recently been an explosion of literature regarding low-rank matrix recovery, with special attention given to the matrix completion subproblem (as made famous by the million dollar Netflix Prize). Several different algorithms have been proposed, with many drawing their roots from standard compressive sensing techniques [25, 34, 36, 52, 97, 101, 104, 110, 129]. For example, nuclear-norm minimization is highly analogous to ℓ_1 minimization (as a convex relaxation to an intractable problem), and the algorithms analyzed in this chapter are analogous to the Dantzig selector and the LASSO.

The efficacy of nuclear-norm minimization is a focus of the thesis of Fazel [72] who outlined several of its various applications from control theory to covariance estimation. Aside from an emphasis on applications, Fazel et al. [72, 74] demonstrated through numerical simulations that nuclear-norm minimization, and also a somewhat similar log-det heuristic, are quite effective in recovering low-rank matrices.

The theory regarding the power of nuclear-norm minimization to recover low-rank matrices from subsampled measurements began with two papers by Srebro et al. [142,143] in the matrix completion setup. General (noiseless) linear measurements were then considered by Recht et al. [129], a paper which created a bridge between compressive-sensing and low-rank matrix recovery via the RIP (to be defined in Section 4.2.1). Bach [8] then considered the asymptotics, and demonstrated consistency of the matrix LASSO estimate in the noisy case, as the number of measurements approached infinity (the matrix LASSO estimate is defined in Section 4.1.4). Subsequently, several papers revisited the matrix competion setup which turns out to be RIP-less and further developed the theory of nuclear-norm minimization [32, 34, 36, 44, 89]; this literature is motivated by very clear applications such as

[1] An interesting point about quantum state tomography is that if one enforces the constraints that M is positive semidefinite and $\mathrm{trace}(M) = 1$ then this ensures that $\|M\|_* = 1$, and the scientist is left with a feasibility problem. In [88] the authors suggest to solve this feasibility problem by removing a constraint and then performing nuclear-norm minimization and they show that under certain conditions this is sufficient for exact recovery (and thus of course the solution obeys the unenforced constraint).

recommender systems and network localizations, and has required very sophisticated mathematical techniques.

Outside of matrix completion, another route used to prove the effectiveness of nuclear-norm minimization is the consideration of null space conditions. Recht et al. [130] introduced a necessary and sufficient condition for the noiseless recovery of all low-rank matrices by nuclear-norm minimization; this condition is the clear analog of the null space condition for sparse recovery (see Section 2.2.3). Recht et al. [130] considered Gaussian measurements, and proved that the null space condition is satisfied with high probability above a certain threshold of measurements. He also augmented these results with an analysis of the weak threshold, i.e., the number of measurements necessary to recover a fixed low-rank matrix, rather than what is necessary for universal recovery. Further, he showed through numerical simulations that for matrices with relatively large rank his threshold appears to be fairly tight. Oymak et al. [121] improved these thresholds particularly in the case when the rank is an arbitrarily small proportion of the matrix length or width.

With the recent increase in attention given to the low-rank matrix model, which we surmise is due to the spring of new theory, new applications are being quickly discovered that deviate from the matrix completion setup (such as quantum state tomography [88]), and could benefit from a different analysis. In this chapter, we once again consider measurement ensembles obeying the RIP as in [129], which are of a different nature than those involved in matrix completion. As in compressive sensing, the only known measurement ensembles which provably satisfy the RIP at a nearly minimal sampling rate are random (such as the Gaussian measurement ensemble in Section 4.2.1). Having said this, two comments are in order. First, our results provide an absolute benchmark of what is achievable, thus allowing direct comparisons with other methods and other sampling operators $\mathcal{A}$. For instance, one can quantify how far the error bounds for the RIP-less matrix completion are from what is then known to be essentially unimprovable. Second, since our results imply that the restricted isometry property *alone* guarantees a near-optimal accuracy, we hope that this will encourage more applications with random ensembles, and also encourage researchers to establish whether or not their measurements obey this desirable property. Finally, we hope that our analysis offers insights for applications with nonrandom measurement ensembles.

While the results discussed in this chapter are novel, recent similar and complementary results are available in the literature [116–118, 131]. Negahban et al. [117] and [118] proved strong results under the *restricted strong convexity* condition, which is different than the RIP, but similar in nature. In fact, in [116], Negahban et al. extend this same analysis to the CS problem. In [117] the authors showed that this condition holds with high probability for the Gaussian measurement ensemble (see Section 4.2.2) as long as there is a constant number of measurements per degree of freedom; thus, this mirrors one of our results. Tsybakov et al. [131] assumed the RIP plus another hypothesis, and established noisy error bounds. When the unknown matrix is exactly—not approximately—

low rank and when its nonzero singular values stand far above the noise level,[2] the error bounds in [117, 118, 131], holding under the respective conditions of those papers, are the same as those presented in Theorem 4.2.4 in Section 4.2.2 below (to within a constant factor). However, when the unknown matrix is well approximated by a matrix of much lower rank, [118, 131] present error bounds of a very different nature than those discussed in this thesis; we invite readers to explore these complementary results. Last, a unique feature of the research discussed in this chapter, is that we give lower bounds matching our intuitive error bounds (Theorems 4.2.4 and 4.2.7).

4.1.3 Problem setup

As mentioned above, we observe data y from the model

$$y = \mathcal{A}(M) + z, \tag{4.1.1}$$

where M is an unknown $n_1 \times n_2$ matrix, $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a linear mapping, and z is an m-dimensional noise term. The synthesized versions of our error bounds assume that z is a Gaussian vector with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, written as $z \sim \mathcal{N}(0, \sigma^2 I)$. The goal is to recover a good approximation of M while requiring as few measurements as possible.

We pause to demonstrate the form of $\mathcal{A}(X)$ explicitly: the ith entry of $\mathcal{A}(X)$ is $[\mathcal{A}(X)]_i = \langle A_i, X \rangle$ for some sequence of matrices $\{A_i\}$ and with the standard inner product $\langle A, X \rangle = \text{trace}(A^* X)$. ($A^*$ is the adjoint of A.) Each A_i can be likened to a row of a compressive sensing matrix, and in fact it can aid the intuition to think of $\mathcal{A}$ as a large matrix, i.e., one could write $\mathcal{A}(X)$ as

$$\mathcal{A}(X) = \begin{bmatrix} \text{vec}(A_1) \\ \text{vec}(A_2) \\ \vdots \\ \text{vec}(A_m) \end{bmatrix} \text{vec}(X), \tag{4.1.2}$$

where $\text{vec}(X)$ is a long vector obtained by stacking the columns of X. In the common matrix completion problem, each A_i is of the form $e_k e_j^*$ so that the ith component of $\mathcal{A}(X)$ is of the form $\langle e_k e_j^*, M \rangle = e_k^* M e_j = M_{kj}$ for some (j, k).

[2] This can be interpreted as having a signal well separated in amplitude from the noise floor.

4.1.4 Algorithms

We analyze the theoretical properties of two different nuclear-norm-minimization-based programs. The first is an analogue to the Dantzig selector from compressive sensing [43], defined as follows:

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & \|\mathcal{A}^*(q)\| \le \lambda \\ & q = y - \mathcal{A}(X), \end{array} \tag{4.1.3}$$

where the optimal solution is our estimate $\hat{M}$, $\|\cdot\|$ is the operator norm and $\|\cdot\|_*$ is its dual, i.e., the nuclear norm. (The nuclear norm of a matrix, X, is the sum of the singular values of X and the operator norm is its largest singular value.) $\mathcal{A}^*$ is the adjoint of $\mathcal{A}$. We call this convex program the *matrix Dantzig selector*.

To pick a useful value for the parameter λ in (4.1.3), we stipulate that the 'true' matrix M should be feasible (this is a necessary condition for our proofs). In other words, one should have $\|\mathcal{A}^*(z)\| \le \lambda$; Section 4.2.2 provides further intuition about this requirement. In the case of Gaussian noise, this corresponds to $\lambda = C\sqrt{n}\sigma$ for some numerical constant C as in the following lemma.

Lemma 4.1.1 *Suppose z is a Gaussian vector with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries and let $n = \max(n_1, n_2)$. Then if $C_0 > 4\sqrt{(1+\delta_1)\log 12}$*

$$\|\mathcal{A}^*(z)\| \le C_0\sqrt{n}\sigma, \tag{4.1.4}$$

with probability at least $1 - 2e^{-cn}$ for a fixed numerical constant $c > 0$.

This lemma is proved in Section 4.3 using a standard covering argument. The scalar δ_1 is the isometry constant at rank 1, as defined in Section 4.2.1, but suffice for now that it is a very small constant bounded by $\sqrt{2} - 1$ (with high probability) under the assumptions of all of our theorems.

The optimization program (4.1.3) may be formulated as a semidefinite program (SDP) and can thus be solved by any of the standard SDP solvers. To see this, we first recall that the nuclear norm admits an SDP characterization since $\|X\|_*$ is the optimal value of the SDP

$$\begin{array}{ll} \text{minimize} & \big(\text{trace}(W_1) + \text{trace}(W_2)\big)/2 \\ \text{subject to} & \begin{bmatrix} W_1 & X \\ X^* & W_2 \end{bmatrix} \succeq 0 \end{array}$$

with optimization variables $X, W_1, W_2 \in \mathbb{R}^{n\times n}$. (We say that a matrix $Q \succeq 0$ if Q is positive semidefinite.) Second, the constraint $\|\mathcal{A}^*(q)\| \le \lambda$ is an SDP constraint since it can be expressed as the

linear matrix inequality (LMI)

$$\begin{bmatrix} \lambda I_n & \mathcal{A}^*(q) \\ [\mathcal{A}^*(q)]^* & \lambda I_n \end{bmatrix} \succeq 0.$$

This shows that (4.1.3) can be formulated as the SDP

$$\begin{aligned} \text{minimize} \quad & \big(\operatorname{trace}(W_1) + \operatorname{trace}(W_2)\big)/2 \\ \text{subject to} \quad & \begin{bmatrix} W_1 & X & 0 & 0 \\ X^* & W_2 & 0 & 0 \\ 0 & 0 & \lambda I_n & \mathcal{A}^*(q) \\ 0 & 0 & [\mathcal{A}^*(q)]^* & \lambda I_n \end{bmatrix} \succeq 0 \\ & q = y - \mathcal{A}(X), \end{aligned}$$

with optimization variables $X, W_1, W_2 \in \mathbb{R}^{n \times n}$.

While this SDP formulation implies that the program can be solved in polynomial time, a few algorithms have recently been developed to solve similar nuclear-norm minimization problems without using interior-point methods which work especially efficiently in practice [25,104]. The nuclear-norm minimization problem solved using fixed-point continuation in [104] is an analogue to the LASSO, and is defined as follows:

$$\underset{X}{\text{minimize}} \quad \frac{1}{2}\|\mathcal{A}(X) - y\|_{\ell_2}^2 + \mu\|X\|_*. \tag{4.1.5}$$

We call this convex program the *matrix LASSO* and it is the second convex program whose theoretical properties are analyzed in this chapter. To our knowledge, this program was first proposed in [143] (as a specific case of a more general program).

4.1.5 Organization

The results in this chapter mostly concern random measurements and random noise and so they hold with high probability. In Section 4.2.1, we show that certain classes of random measurements satisfy the RIP when only sampling a constant number of measurements per degree of freedom. In Section 4.2.2 we present the simplest of our error bounds, demonstrating that when the RIP holds, the solution to (4.1.3) is within a constant of the minimax risk. This error bound is refined in Section 5.3.2 to provide a more adaptive error that holds improvements when the singular values of M decay below the noise level. It is shown that this error bound is within a constant of the expected value of a certain oracle error. In Section 4.2.4, we present an error bound handling the case when M has full rank but is well approximated by a low-rank matrix. Section 4.3 contains the proofs and we finish with some concluding remarks in Section 4.4.

4.1.6 Notation

We review all notation used in this chapter in order to ease readability. We assume $M \in \mathbb{R}^{n_1 \times n_2}$ and let $n = \max(n_1, n_2)$. A variety of norms are used throughout: $\|X\|_*$ is the nuclear norm (the sum of the singular values); $\|X\|$ is the operator norm of X (the top singular value); $\|X\|_F$ is the Frobenius norm (the ℓ_2-norm of the vector of singular values). The matrix X^* is the adjoint of X, and for the linear operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^{n_1 \times n_2}$ is the adjoint operator. Specifically, if $[\mathcal{A}(X)]_i = \langle A_i, X \rangle$ for all matrices $X \in \mathbb{R}^{n_1 \times n_2}$, then

$$\mathcal{A}^*(q) = \sum_{i=1}^{m} q_i A_i$$

for all vectors $q \in \mathbb{R}^m$.

4.2 Main Results

4.2.1 Matrix RIP

The matrix version of the RIP is an integral tool in proving our theoretical results and we begin by defining the RIP in this setting and describing measurement ensembles that satisfy it. As discussed in Section 4.1.2, the RIP was first considered in low-rank matrix recovery by Recht et al. in [129]. To characterize the RIP, we introduce the isometry constants of a linear map $\mathcal{A}$.

Definition 4.2.1 *For each integer $r = 1, 2, \ldots, n$, the isometry constant δ_r of $\mathcal{A}$ is the smallest quantity such that*

$$(1 - \delta_r)\|X\|_F^2 \leq \|\mathcal{A}(X)\|_{\ell_2}^2 \leq (1 + \delta_r)\|X\|_F^2 \tag{4.2.1}$$

holds for all matrices X of rank at most r.

We say that $\mathcal{A}$ satisfies the RIP at rank r if δ_r (or δ_{4r}) is bounded by a sufficiently small constant between 0 and 1, the value of which will become apparent in further sections (see, e.g., Theorem 4.2.4).

Which linear maps $\mathcal{A}$ satisfy the RIP? As a quintessential example, we introduce the Gaussian measurement ensemble.

Definition 4.2.2 *$\mathcal{A}$ is a Gaussian measurement ensemble if each 'row' A_i, $1 \leq i \leq m$, contains i.i.d. $\mathcal{N}(0, 1/m)$ entries (and the A_i's are independent from each other).*

This is of course highly analogous to the Gaussian random matrices in compressive sensing. Having said this, a comment about the normalization is in order. We have selected the variance of the entries to be $1/m$ so that for a fixed matrix X, $\mathbb{E}\,\|\mathcal{A}(X)\|_{\ell_2}^2 = \|X\|_F^2$. However, we could instead require that the entries be standard normal. Suppose one observes $y = \mathcal{A}(M) + z$, where $\mathcal{A}$ has standard normal

entries and $z \sim \mathcal{N}(0, \sigma^2 I)$. A simple rescaling gives $m^{-1/2} y = m^{-1/2}\mathcal{A}(M) + m^{-1/2} z$, and the entries of $m^{-1/2}\mathcal{A}$ have variance $1/m$. Hence, one would just need to replace σ^2 appearing in our bounds by σ^2/m.

Our first result refines a result by Recht et al. [129] (see below for a comparison). It demonstrates that Gaussian measurement ensembles, along with many other random measurement ensembles, satisfy the RIP when $m \geq C\, nr$ (with high probability) for some constant $C > 0$.

Theorem 4.2.3 *Fix $0 \leq \delta < 1$ and let $\mathcal{A}$ be a random measurement ensemble obeying the following condition: for any given $X \in \mathbb{R}^{n_1 \times n_2}$ and any fixed $0 < t < 1$,*

$$P(|\|\mathcal{A}(X)\|_{\ell_2}^2 - \|X\|_F^2| > t\|X\|_F^2) \leq C \exp(-cm) \tag{4.2.2}$$

for fixed constants $C, c > 0$ (which may depend on t). Then if $m \geq Dnr$, $\mathcal{A}$ satisfies the RIP with isometry constant $\delta_r \leq \delta$ with probability exceeding $1 - Ce^{-dm}$ for fixed constants $D, d > 0$.

The many unspecified constants involved in the presentation of Theorem 4.2.3 are meant to allow for general use with many random measurement ensembles. However, to make the presentation more concrete we describe the constants involved in the concentration bound (4.2.2) for a few special random measurement ensembles. If $\mathcal{A}$ is a Gaussian random measurement ensemble, $\|\mathcal{A}(X)\|_{\ell_2}^2$ is distributed as $m^{-1}\|X\|_F^2$ times a chi-squared random variable with m degrees of freedom and (4.2.2) follows from standard concentration inequalities [99, 129, 164]. Specifically, we have

$$P(|\|\mathcal{A}(X)\|_{\ell_2}^2 - \|X\|_F^2| > t\|X\|_F^2) \leq 2\exp\left(-\frac{m}{2}(t^2/2 - t^3/3)\right). \tag{4.2.3}$$

Similarly, $\mathcal{A}$ satisfies equation (4.2.3) in the case when each entry of each 'row' A_i has i.i.d. entries that are equally likely to take the value $1/\sqrt{m}$ or $-1/\sqrt{m}$, or if $\mathcal{A}$ is a random projection [2, 129]. Further, $\mathcal{A}$ satisfies (4.2.2) if the 'rows' A_i contain sub-Gaussian entries (properly normalized) [160], although in this case the constants involved depend on the parameters of the sub-Gaussian entries.

In order to ascertain the strength of Theorem 4.2.3, note that the number of degrees of freedom of an $n_1 \times n_2$ matrix of rank r is equal to $r(n_1 + n_2 - r)$. (This can be seen by counting the number of equations and unknowns in the singular value decomposition.) Thus, one may expect that if $m < r(n_1 + n_2 - r)$, there should be a rank-r matrix in the null space of $\mathcal{A}$ leading to a failure to achieve the lower bound in (5.3.4). In order to make this intuition rigorous (to within a constant) assume without loss of generality that $n_2 \geq n_1$, and observe that the set of rank-r matrices contains all those matrices restricted to have nonzero entries only in the first r rows. This is an $n \times r$ dimensional vector space and thus we must have $m \geq nr$ or otherwise there will be a rank-r matrix in the null space of $\mathcal{A}$ regardless of what measurements are used. (This is a similar alternative to the null-space argument posed in the introduction.)

Theorem 4.2.3 is inspired by a similar theorem in [129][Theorem 4.2] and refines this result in two ways. First, it shows that one only needs a constant number of measurements per degree of freedom of the underlying rank-r matrix in order to obtain the RIP at rank r (which improves on the result in [129] by a factor of $\log n$ and also achieves the theoretical lower bound to within a constant). Second, it shows that one must only require a single concentration bound on $\mathcal{A}$, removing another assumption required in [129]. A possible third benefit is that the proof follows simply and quickly from a specialized covering argument. The novelty is in the method used to cover low-rank matrices.

4.2.2 The matrix Dantzig selector and the matrix LASSO are nearly minimax

In this section, we present our first and simplest error bound, which only requires that $\mathcal{A}$ satisfies the RIP.

Theorem 4.2.4 *Assume the measurement operator $\mathcal{A}$ is fixed and satisfies the RIP, and that* $\operatorname{rank}(M) \leq r$. *Let $\hat{M}_{DS}$ be the solution to the matrix Dantzig selector* (4.1.3) *and $\hat{M}_L$ be the solution to the matrix LASSO* (4.1.5). *If $\delta_{4r} < \sqrt{2} - 1$ and $\|\mathcal{A}^*(z)\| \leq \lambda$ then*

$$\|\hat{M}_{DS} - M\|_F^2 \leq C_0\, r\lambda^2, \tag{4.2.4}$$

and if $\delta_{4r} < (3\sqrt{2} - 1)/17$ and $\|\mathcal{A}^*(z)\| \leq \mu/2$, then

$$\|\hat{M}_L - M\|_F^2 \leq C_1\, r\mu^2; \tag{4.2.5}$$

above, C_0 and C_1 are small constants depending only on the isometry constant δ_{4r}. In particular, if $z \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$ and $\hat{M}$ is either $\hat{M}_{DS}$ with $\lambda = 8\sqrt{n}\sigma$, or $\hat{M}_L$ with $\mu = 16\sqrt{n}\sigma$, we have

$$\|\hat{M} - M\|_F^2 \leq C_0'\, nr\sigma^2 \tag{4.2.6}$$

with probability at least $1 - 2e^{-cn}$ for a constant C_0' (depending only on δ_{4r}).

Note that (4.2.6) follows from (4.2.4) and (4.2.5) simply by plugging $\lambda, \mu/2 = 8\sqrt{n}\sigma$ into Lemma 4.1.1. In a nutshell, the error is proportional to the number of degrees of freedom times the noise level.

An important point is that one may expect the error to be reduced when further measurements are taken, i.e., one may expect the error to be inversely proportional to m. In fact, this is the case for the Gaussian measurement ensemble, but this extra factor is absorbed into the definition in order to normalize the measurements so that they satisfy the RIP (see Section 4.2.1). If instead, each row 'A_i' in the Gaussian measurement ensemble is defined to have i.i.d. standard normal entries, then it

follows from the discussion immediately after Definition 4.2.2 that the error bound reads

$$\|\hat{M} - M\|_F^2 \leq C_0' \, nr\sigma^2/m. \tag{4.2.7}$$

Of course, this rescaling argument applies in general to non-Gaussian measurements.

A second remark is that exploiting the low-rank structure helps to denoise. For example, if we measured every entry of M (a measurement ensemble with isometry constant $\delta_r = 0$), but with each measurement corrupted by a $\mathcal{N}(0, \sigma^2)$ noise term, then taking the measurements as they are as the estimate of M would lead to an expected error equal to

$$\mathbb{E}\,\|\hat{M} - M\|_F^2 = n^2\sigma^2.$$

Nuclear-norm minimization[3] reduces this error by a factor of about n/r.

The strength of Theorem 4.2.4 is that the error bound (4.2.6) is nearly optimal in the sense that no estimator can do essentially better without further assumptions, as seen by lower-bounding the expected minimax error.

Theorem 4.2.5 *Suppose that the measurement operator $\mathcal{A}$ is fixed and satisfies the RIP, and that $z \sim \mathcal{N}(0, \sigma^2\mathrm{I})$. Then any estimator $\hat{M}(y)$ obeys*

$$\sup_{M:\mathrm{rank}(M)\leq r} \mathbb{E}\,\|\hat{M}(y) - M\|_F^2 \geq \frac{1}{1+\delta_r}\, nr\sigma^2. \tag{4.2.8}$$

In other words, the minimax error over the class of matrices of rank at most r is lower bounded by about $nr\sigma^2$.

The exacting reading may argue that while this lower bound is in expectation, the upper bound holds with high probability. To address this, we also prove the following complementary theorem.

Theorem 4.2.6 *Under the assumptions of Theorem 4.2.5, any estimator $\hat{M}(y)$ obeys*

$$\sup_{M:\mathrm{rank}(M)\leq r} \mathbb{P}\left(\|\hat{M}(y) - M\|_F^2 \geq \frac{1}{2(1+\delta_r)} nr\sigma^2\right) \geq 1 - e^{-nr/16}. \tag{4.2.9}$$

Before continuing, it may be helpful to analyze the solutions to the matrix Dantzig selector and the matrix LASSO in a simple case in order to understand the error bounds in Theorem 4.2.4 intuitively, and also to understand our choice of λ and μ. Suppose $\mathcal{A}$ is the identity so that changing the notation a bit, the model is $Y = M + Z$, where Z is an $n \times n$ matrix with i.i.d. Gaussian entries. We would like the unknown matrix M to be a feasible point, which requires that $\|Z\| \leq \lambda$ (for example,

[3] Of course if one sees all of the entries of the matrix plus noise, nuclear-norm minimization is unnecessary, and one can achieve minimax error bounds by truncating the singular values.

if $\|Z\| > \lambda$, we already have problems when $M = 0$). It is well known that the top singular value of a square $n \times n$ Gaussian matrix, with per-entry variance σ^2, is concentrated around $\sqrt{2n}\sigma$, and thus we require $\lambda \geq \sqrt{2n}\sigma$ (this provides a slightly sharper bound than Lemma 4.1.1). Let $T_\lambda(X)$ denote the singular value thresholding operator given by

$$T_\lambda(X) = \sum_i \max(\sigma_i(X) - \lambda, 0) u_i v_i^*,$$

where $X = \sum_i \sigma_i(X) u_i v_i^*$ is any singular value decomposition. In this simple setting, the solution to (4.1.3) and (4.1.5) can be explicitly calculated, and for $\lambda = \mu$ they are both equal to $T_\lambda(M + Z)$. If λ is too large, then $T_\lambda(M + Z)$ becomes strongly biased towards zero, and thus (loosely) λ should be as small as possible while still allowing M to be feasible for the matrix Dantzig selector (4.1.3), leading to the choice $\lambda \approx \sqrt{2n}\sigma$.

Further, in this simple case we can calculate the error bound in a few lines. We have

$$\begin{aligned} \left\|\hat{M} - M\right\| &= \|T_\lambda(Y) - Y + Z\| \\ &\leq \|T_\lambda(Y) - Y\| + \|Z\| \\ &\leq 2\lambda \end{aligned}$$

assuming that $\lambda \geq \|Z\|$. Then

$$\begin{aligned} \|\hat{M} - M\|_F^2 &\leq \left\|\hat{M} - M\right\|^2 \operatorname{rank}(\hat{M} - M) \\ &\leq 4\lambda^2 \operatorname{rank}(\hat{M} - M). \end{aligned} \tag{4.2.10}$$

Once again, assuming that $\lambda \geq \|Z\|$, we have $\operatorname{rank}(\hat{M} - M) \leq \operatorname{rank}(\hat{M}) + \operatorname{rank}(M) \leq 2r$. Plugging this in with $\lambda = C\sqrt{n}\sigma$ gives the error bound (4.2.6).

4.2.3 Oracle inequalities

Showing that an estimator achieves the minimax risk is reassuring but is sometimes not considered completely satisfactory. As is frequently discussed in the literature, the minimax approach focuses on the worst-case performance and it is quite reasonable to expect that for matrices of general interest, better performances are possible. In fact, a recent trend in statistical estimation is to compare the performance of an estimator with what is achievable with the help of an oracle that reveals extra information about the problem. A good match indicates an overall excellent performance.

What information should the oracle reveal in this problem? A first thought is that the oracle could reveal the rank of M. However, this information could be quite suboptimal as shown by the following toy example. Imagine that M has 2 large singular values, far above the noise level, and 20

other small singular values, far below the noise level. The rank of M is 22, but under the presence of noise it ostensibly has rank 2. Thus, ostensibly it would be best to attempt to approximate it with a rank−2 matrix, and one would hope to set $r = 2$ in the error bound from Theorem 4.2.4. Thus, perhaps the oracle should give the 'ostensible' rank of M. To make this rigorous, we examine the following family of least squares estimators.[4]

$$\hat{M}[r] = \arg\min\{\|y - \mathcal{A}(\hat{M})\|_{\ell_2} : \operatorname{rank}\hat{M} = r\}. \tag{4.2.11}$$

Knowing the true matrix M, an oracle or a genie would then select the best rank to use as to minimize the mean-squared error (MSE)

$$\mathbb{E}\,\|\hat{M}[r] - M\|_F^2. \tag{4.2.12}$$

In fact, this expected error is difficult to analyze due to the nonlinear structure of the manifold of matrices at a fixed rank, r. Instead, we introduce an oracle which gives more information, and is easier to analyze. Instead of just the 'best' rank, this oracle gives the 'best' column space as well.

To develop this oracle bound, assume w.l.o.g. that $n_2 \geq n_1$ so that $n = n_2$ and let $r_M = \operatorname{rank}(M)$. Suppose $\mathcal{A}$ satisfies the RIP at rank r, and consider the family of estimators defined as follows: for each $n_1 \times r$ orthogonal matrix U with $r \leq r_M$, define

$$\hat{M}[U] = \arg\min_R\{\|y - \mathcal{A}(\hat{M})\|_{\ell_2} : \hat{M} = UR \text{ for some } R\}. \tag{4.2.13}$$

In other words, we fix the column space (the linear space spanned by the columns of the matrix U), and then find the matrix with that column space which best fits the data. The oracle then supplies the column space that minimizes the MSE.

$$\inf_{U,r} \mathbb{E}\,\|M - \hat{M}[U]\|_F^2. \tag{4.2.14}$$

The question is whether it is possible to mimic the performance of the oracle and achieve a MSE close to (4.2.14) with a real estimator.

Before giving a precise answer to this question, it is useful to determine how large the oracle risk is. To this end, consider a fixed orthogonal matrix U, and write the least-squares estimate (4.2.13) as

$$\hat{M}[U] := U\mathcal{H}_U(y), \quad \mathcal{H}_U = (\mathcal{A}_U^*\mathcal{A}_U)^{-1}\mathcal{A}_U^*,$$

[4]One must make a choice of estimators when examining oracle errors, and we use the standard choice as in [43].

where $\mathcal{A}_U$ is the linear map

$$\begin{array}{rccc} \mathcal{A}_U & : & \mathbb{R}^{r\times n} & \to & \mathbb{R}^m \\ & & R & \mapsto & \mathcal{A}(UR), \end{array} \tag{4.2.15}$$

and

$$\begin{array}{rccc} \mathcal{A}_U^* & : & \mathbb{R}^{m} & \to & \mathbb{R}^{r\times n} \\ & & y & \mapsto & U^*\mathcal{A}^*(y). \end{array}$$

Then decompose the MSE as the sum of the squared bias and variance

$$\begin{aligned} \mathbb{E}\,\|M-\hat{M}[U]\|_F^2 &= \|\text{bias}\|_F^2 + \text{variance} \\ &= \|\,\mathbb{E}\,\hat{M}[U]-M\|_F^2 + \mathbb{E}\,\|U\mathcal{H}_U(z)\|_F^2. \end{aligned}$$

The variance term is classically equal to

$$\mathbb{E}\,\|U\mathcal{H}_U(z)\|_F^2 = \mathbb{E}\,\|\mathcal{H}_U(z)\|_F^2 = \sigma^2\,\text{trace}(\mathcal{H}_U^*\mathcal{H}_U) = \sigma^2\,\text{trace}((\mathcal{A}_U^*\mathcal{A}_U)^{-1}).$$

Due to the restricted isometry property, all the eigenvalues of the linear operator $\mathcal{A}_U^*\mathcal{A}_U$ belong to the interval $[1-\delta_{r_M}, 1+\delta_{r_M}]$, see Lemma 4.3.13. Therefore, the variance term obeys

$$\sigma^2\,\text{trace}((\mathcal{A}_U^*\mathcal{A}_U)^{-1}) \geq \frac{1}{1+\delta_{r_M}}\,nr\sigma^2.$$

For the bias term, we have

$$\mathbb{E}\,\hat{M}[U]-M = U(\mathcal{A}_U^*\mathcal{A}_U)^{-1}\mathcal{A}_U^*\mathcal{A}(M)-M,$$

which we rewrite as

$$\begin{aligned} \mathbb{E}\,\hat{M}[U]-M &= U(\mathcal{A}_U^*\mathcal{A}_U)^{-1}\mathcal{A}_U^*\mathcal{A}((I-UU^*+UU^*)M)-M \\ &= U(\mathcal{A}_U^*\mathcal{A}_U)^{-1}\mathcal{A}_U^*\mathcal{A}((I-UU^*)M)+U(\mathcal{A}_U^*\mathcal{A}_U)^{-1}\mathcal{A}_U^*\mathcal{A}_U(U^*M)-M \\ &= U(\mathcal{A}_U^*\mathcal{A}_U)^{-1}\mathcal{A}_U^*\mathcal{A}((I-UU^*)M)-(I-UU^*)M. \end{aligned}$$

Hence, the bias is the sum of two matrices: the first has a column space included in the span of the columns of U while the column space of the other is orthogonal to this span. Put $P_{U^\perp}(M) = (I-UU^*)M$; that is, $P_{U^\perp}(M)$ is the (left) multiplication with the orthogonal projection matrix $(I-UU^*)$. We have

$$\begin{aligned} \|\,\mathbb{E}\,\hat{M}[U]-M\|_F^2 &= \|U(\mathcal{A}_U^*\mathcal{A}_U)^{-1}\mathcal{A}_U^*\mathcal{A}(P_{U^\perp}(M))\|_F^2 + \|P_{U^\perp}(M)\|_F^2 \\ &\geq \|P_{U^\perp}(M)\|_F^2. \end{aligned}$$

To summarize, the oracle bound obeys

$$\inf_{U,r} \mathbb{E}\,\|M - \hat{M}[U]\|_F^2 \geq \inf_U \left[\|P_{U^\perp}(M)\|_F^2 + \frac{nr\sigma^2}{1+\delta_{r_M}} \right].$$

Now for a given dimension r, the best U—that minimizing the squared bias term or its proxy $\|P_{U^\perp}(M)\|_F^2$—spans the top r singular vectors of the matrix M. Denoting the singular values of M by $\sigma_i(M)$, we obtain

$$\inf_{U,r} \mathbb{E}\,\|M - \hat{M}[U]\|_F^2 \geq \inf_r \left[\sum_{i>r} \sigma_i^2(M) \; + \frac{1}{2} nr\sigma^2 \right],$$

which for convenience we simplify to

$$\inf_{U,r} \mathbb{E}\,\|M - \hat{M}[U]\|_F^2 \geq \frac{1}{2} \sum_i \min(\sigma_i^2(M), n\sigma^2). \tag{4.2.16}$$

The right-hand side has a nice interpretation. Write the SVD of M as $M = \sum_{i=1}^r \sigma_i(M) u_i v_i^*$. Then if $\sigma_i^2(M) > n\sigma^2$, one should try to estimate the rank-1 contribution $\sigma_i(M)\, u_i v_i^*$ and pay the variance term (which is about $n\sigma^2$) whereas if $\sigma_i^2(M) \leq n\sigma^2$, we should not try to estimate this component, and pay a squared bias term equal to $\sigma_i^2(M)$. In other words, the right-hand side may be interpreted as an *ideal* bias-variance trade-off. Note that when all of the singular values are far above the noise level, the oracle (nearly) matches the minimax bound of Theorem 4.2.5.)

The main result of this section is that the matrix Dantzig selector and matrix LASSO achieve this same ideal bias-variance trade-off to within a constant.

Theorem 4.2.7 *Assume the measurement operator $\mathcal{A}$ is fixed and satisfies the RIP, and that* $\operatorname{rank}(M) \leq r$. *Let $\hat{M}_{DS}$ be the solution to the matrix Dantzig selector* (4.1.3) *and $\hat{M}_L$ be the solution to the matrix LASSO* (4.1.5). *Suppose $z \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$ and let $\lambda = 16\sqrt{n}\sigma$ and $\mu = 32\sqrt{n}\sigma^2$. If $\delta_{4r} < \sqrt{2} - 1$, then*

$$\|\hat{M}_{DS} - M\|_F^2 \leq C_0 \sum_i \min(\sigma_i^2(M), n\sigma^2), \tag{4.2.17}$$

and if $\delta_{4r} < (3\sqrt{2} - 1)/17$, then

$$\|\hat{M}_L - M\|_F^2 \leq C_1 \sum_i \min(\sigma_i^2(M), n\sigma^2) \tag{4.2.18}$$

with probability at least $1 - 2e^{-cn}$ for constants C_0 and C_1 (depending only on δ_{4r}).

In other words, not only does nuclear-norm minimization mimic the performance that one would achieve with an oracle that gives the exact column space of M (as in Theorem 4.2.5), but in fact the error bound is within a constant of what one would achieve by projecting onto the optimal column space corresponding only to the significant singular values.

While a similar result holds in the compressive sensing literature [43], we derive the result here

using a novel technique. We use a middle estimate $\bar{M}$ which is the optimal solution to a certain rank-minimization problem (see Section 4.3) and is provably near $\hat{M}$ and M. With this technique, the proof is a fairly simple extension of Theorem 4.2.4.

4.2.4 Extension to full-rank matrices

In some applications, such as sensor localization, M has exactly low rank, i.e., only the top few of its singular values are nonzero. However, in many applications, such as quantum state tomography, M has full rank, but is well approximated by a low-rank matrix. In this section, we demonstrate an extension of the preceding error bound when M has full rank.

First, suppose $n_1 \leq n_2$ and note that a result of the form

$$\|\hat{M} - M\|_F^2 \leq C \sum_{i=1}^{n_1} \min(\sigma_i^2(M), n\sigma^2) \tag{4.2.19}$$

would be impossible when undersampling M because it would imply that as the noise level σ approaches zero, an arbitrary full-rank $n \times n$ matrix could be exactly reconstructed from fewer than n^2 linear measurements. Instead, our result essentially splits M into two parts,

$$M = \sum_{i=1}^{\bar{r}} \sigma_i(M) u_i v_i^* + \sum_{i=\bar{r}+1}^{n_1} \sigma_i(M) u_i v_i^* = M_{\bar{r}} + M_c$$

where $\bar{r} \approx m/n$, and $M_{\bar{r}}$ is the best rank-$\bar{r}$ approximation to M. The error bound in the theorem reflects a near-optimal bias-variance trade-off in recovering $M_{\bar{r}}$, but an inability to recover M_c (and indeed the proof essentially considers M_c as non-Gaussian noise). Note that $\bar{r}(n_1 + n_2 - \bar{r})$ is of the same order as m so that the part of the matrix which is well recovered has about as many degrees of freedom as the number of measurements. In other words, even in the noiseless case this theorem demonstrates instance optimality, i.e., the error bound is proportional to the norm of the part of M that is irrecoverable given the number of measurements (see [166] for an analogous result in compressive sensing). In the noisy case there does not seem to be any current analogue to this error bound in compressive sensing.

Theorem 4.2.8 *Fix M. Suppose that $\mathcal{A}$ is sampled from the Gaussian measurement ensemble with $m \leq c_0 n^2 / \log(m/n)$ and let $\bar{r} \leq c_1 m/n$ for some fixed numerical constants c_0 and c_1. Let $\hat{M}$ be the solution to the matrix Dantzig selector* (4.1.3) *with $\lambda = 16\sqrt{n}\sigma$ or the solution to the matrix LASSO* (4.1.5) *with $\mu = 32\sqrt{n}\sigma$. Suppose that $z \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$. Then*

$$\|\hat{M} - M\|_F^2 \leq C \left(\sum_{i=1}^{\bar{r}} \min(\sigma_i^2(M), n\sigma^2) + \sum_{i=\bar{r}+1}^{n} \sigma_i^2(M) \right) \tag{4.2.20}$$

with probability greater than $1 - De^{-dn}$ for fixed numerical constants $C, D, d > 0$. Roughly, the same

conclusion extends to operators obeying the NNQ condition, see below.

First note that $\bar{r}$ is small enough so that the RIP holds with high probability (see Lemma 4.2.3). However, the theorem requires more than just the RIP. The other main requirement is a certain NNQ condition, which holds for Gaussian measurement ensembles and is introduced in Section 4.3. It is an analogous requirement to the LQ condition introduced by Wojtaszczyk [166] in compressive sensing. To keep the presentation of the theorem simple, we defer the explanation of the NNQ condition to the proofs section and simply state the theorem for the Gaussian measurement ensemble. However, the proof is not sensitive to the use of this ensemble (for example sub-Gaussian measurements yield the same result). Many generalizations of this theorem are available and the lemmas necessary to make such generalizations are spelled out in Section 4.3.

The assumption that $m \leq cn^2/\log(m/n)$ seems to be an artifact of the proof technique. Indeed, one would not expect further measurements to negatively impact performance. In fact, when $m \geq c'n^2$ for a fixed constant c', one can use Lemma 4.3.2 from Section 4.3 to derive the error bound (4.2.20) (with high probability), leaving the necessity for a small 'patch' in the theory when $cn^2/\log(m/n) \leq n \leq c'n^2$. However, our results intend to address the situation in which M is significantly undersampled, i.e., $m \ll n^2$, so the requirement that $m \leq cn^2/\log(m/n)$ should be intrinsic to the problem setup.

4.3 Proofs

The proofs of several of the theorems use ϵ-nets. For a set S, an ϵ-net S_ϵ with respect to a norm $\|\cdot\|$ satisfies the following property: for any $v \in S$, there exists $v_0 \in S_\epsilon$ with $\|v_0 - v\| \leq \epsilon$. (We abuse notation in this paragraph and let $\|\cdot\|$ denote any norm, rather than just the operator norm.) In other words, S_ϵ approximates S to within distance ϵ with respect to the norm $\|\cdot\|$. As shown in [161][Lecture 6], there always exists an ϵ-net S_ϵ satisfying $S_\epsilon \subset S$ and

$$|S_\epsilon| \leq \frac{\mathrm{Vol}\left(S + \frac{1}{2}D\right)}{\mathrm{Vol}\left(\frac{1}{2}D\right)}$$

where $\frac{1}{2}D$ is an $\epsilon/2$ ball (with respect to the norm $\|\cdot\|$) and $S + \frac{1}{2}D = \{x + y : x \in S, y \in \frac{1}{2}D\}$. [5] In particular, if S is a unit ball in n dimensions (with respect to the norm $\|\cdot\|$) or if it is the surface of the unit ball or any other subset of the unit ball, then $S + \frac{1}{2}D$ is contained in the $1 + \epsilon/2$ ball, and thus

$$|S_\epsilon| \leq \frac{(1 + \epsilon/2)^n}{(\epsilon/2)^n} = \left(\frac{2 + \epsilon}{\epsilon}\right)^n \leq (3/\epsilon)^n \tag{4.3.1}$$

[5] The proposition in [161] does not state that one can take $S_\epsilon \subset S$, but the proof remains identical when enforcing this constraint.

where the last inequality follows because we always take $\epsilon \leq 1$. See [161] for a more detailed argument. We will require in all of our proofs that $S_\epsilon \subset S$.

4.3.1 Proof of Lemma 4.1.1

We assume that $\sigma = 1$ without loss of generality. Put $Z = \mathcal{A}^*(z)$. The norm of Z is given by

$$\|Z\| = \sup \langle w, Zv \rangle,$$

where the supremum is taken over all pairs of vectors on the unit sphere S^{n-1}. Consider a 1/4-net $\mathcal{N}_{1/4}$ of S^{n-1} (under the ℓ_2 norm) with $|\mathcal{N}_{1/4}| \leq 12^n$. For each $v, w \in S^{n-1}$,

$$\begin{aligned}\langle w, Zv \rangle &= \langle w - w_0, Zv \rangle + \langle w_0, Z(v - v_0) \rangle + \langle w_0, Zv_0 \rangle \\ &\leq \|Z\| \|w - w_0\|_{\ell_2} + \|Z\| \|v - v_0\|_{\ell_2} + \langle w_0, Zv_0 \rangle\end{aligned}$$

for some $v_0, w_0 \in \mathcal{N}_{1/4}$ obeying $\|v - v_0\|_{\ell_2} \leq 1/4$, $\|w - w_0\|_{\ell_2} \leq 1/4$. Hence,

$$\|Z\| \leq 2 \sup_{v_0, w_0 \in \mathcal{N}_{1/4}} \langle w_0, Zv_0 \rangle.$$

Now for a fixed pair (v_0, w_0),

$$\langle w_0, Zv_0 \rangle = \operatorname{trace}(w_0^* \mathcal{A}^*(z) v_0) = \operatorname{trace}(v_0 w_0^* \mathcal{A}^*(z)) = \langle w_0 v_0^*, \mathcal{A}^*(z) \rangle = \langle \mathcal{A}(w_0 v_0^*), z \rangle.$$

We deduce from this that $\langle w_0, Zv_0 \rangle \sim \mathcal{N}(0, \|\mathcal{A}(w_0 v_0^*)\|_{\ell_2}^2)$. Now

$$\|\mathcal{A}(w_0 v_0^*)\|_{\ell_2}^2 \leq (1 + \delta_1) \|w_0 v_0^*\|_F^2 = (1 + \delta_1)$$

so that by a standard tail bound for Gaussian random variables

$$\mathbb{P}(|\langle w_0, Zv_0 \rangle| \geq \lambda) \leq 2e^{-\frac{1}{2} \frac{\lambda^2}{1 + \delta_1}}.$$

Therefore,

$$\mathbb{P}(\max |\langle w_0, Zv_0 \rangle| \geq \gamma \sqrt{(1 + \delta_1) n}) \leq 2|\mathcal{N}_{1/4}|^2 e^{-\frac{1}{2}\gamma^2 n} \leq 2e^{2n \log 12 - \frac{1}{2}\gamma^2 n},$$

which is bounded by $2e^{-cn}$ with $c = \gamma^2/2 - 2 \log 12$ (we require $\gamma > 2\sqrt{\log 12}$ so that $c > 0$).

4.3.2 Proof of Theorem 4.2.3

The proof uses a covering argument, starting with the following lemma. Throughout the proof, we make use the of the covering number bound (4.3.1).

Lemma 4.3.1 (Covering number for low-rank matrices) *Let* $S_r = \{X \in \mathbb{R}^{n_1 \times n_2} : \mathrm{rank}(X) \leq r, \|X\|_F = 1\}$. *Then there exists an* ϵ*-net* $\bar{S}_r \subset S_r$ *with respect to the Frobenius norm obeying*

$$|\bar{S}_r| \leq (9/\epsilon)^{(n_1+n_2+1)r}.$$

Proof Recall the SVD $X = U\Sigma V^*$ of any $X \in S_r$ obeying $\|\Sigma\|_F = 1$ (Σ is a diagonal matrix with singular values on the diagonal, and U and V are rectangular matrices of left- and right-singular vectors). Our argument constructs an ϵ-net for S_r by covering the set of permissible U, V, and Σ. We work in the simpler case where $n_1 = n_2 = n$ since the general case is a straightforward modification.

Let D be the set of diagonal matrices with nonnegative diagonal entries and Frobenius norm equal to one. We take $\bar{D}$ to be an $\epsilon/3$-net for D with $|\bar{D}| \leq (9/\epsilon)^r$. Next, let $O_{n,r} = \{U \in \mathbb{R}^{n \times r} : U^*U = I\}$. To cover $O_{n,r}$, it is beneficial to use the $\|\cdot\|_{1,2}$ norm defined as

$$\|X\|_{1,2} = \max_i \|X_i\|_{\ell_2},$$

where X_i denotes the ith column of X. Clearly, $O_{n,r}$ is a subset of the unit ball under the norm $\|\cdot\|_{1,2}$ since the columns of an orthogonal matrix are all unit normed. Hence, by (4.3.1), there is an $\epsilon/3$-net $\bar{O}_{n,r}$ for $O_{n,r}$ obeying $|\bar{O}_{n,r}| \leq (9/\epsilon)^{nr}$. We now let $\bar{S}_r = \{\bar{U}\bar{\Sigma}\bar{V}^* : \bar{U}, \bar{V} \in \bar{O}_{n,r}, \bar{\Sigma} \in \bar{D}\}$, and remark that $|\bar{S}_r| \leq |\bar{O}_{n,r}|^2 |\bar{D}| \leq (9/\epsilon)^{(2n+1)r}$. It remains to show that for all $X \in S_r$ there exists $\bar{X} \in \bar{S}_r$ with $\|X - \bar{X}\|_F \leq \epsilon$.

Fix $X \in S_r$ and decompose X as $X = U\Sigma V^*$ as above. Then there exist $\bar{X} = \bar{U}\bar{\Sigma}\bar{V}^* \in \bar{S}_r$ with $\bar{U}, \bar{V} \in \bar{O}_{n,r}$, $\bar{\Sigma} \in \bar{D}$ obeying $\|U - \bar{U}\|_{1,2} \leq \epsilon/3, \|V - \bar{V}\|_{1,2} \leq \epsilon/3$, and $\|\Sigma - \bar{\Sigma}\|_F \leq \epsilon/3$. This gives

$$\begin{aligned}
\|X - \bar{X}\|_F &= \|U\Sigma V^* - \bar{U}\bar{\Sigma}\bar{V}^*\|_F \\
&= \|U\Sigma V^* - \bar{U}\Sigma V^* + \bar{U}\Sigma V^* - \bar{U}\bar{\Sigma}V^* + \bar{U}\bar{\Sigma}V^* - \bar{U}\bar{\Sigma}\bar{V}^*\|_F \\
&\leq \|(U - \bar{U})\Sigma V^*\|_F + \|\bar{U}(\Sigma - \bar{\Sigma})V^*\|_F + \|\bar{U}\bar{\Sigma}(V - \bar{V})^*\|_F. \qquad (4.3.2)
\end{aligned}$$

For the first term, note that since V is an orthogonal matrix, $\|(U - \bar{U})\Sigma V^*\|_F = \|(U - \bar{U})\Sigma\|_F$, and

$$\begin{aligned}
\|(U - \bar{U})\Sigma\|_F^2 &= \sum_{1 \leq i \leq r} \Sigma_{i,i}^2 \|\bar{U}_i - U_i\|_{\ell_2}^2 \\
&\leq \|\Sigma\|_F^2 \|U - \bar{U}\|_{1,2}^2 \\
&\leq (\epsilon/3)^2.
\end{aligned}$$

Hence, $\|(U - \bar{U})\Sigma V^*\|_F \leq \epsilon/3$. The same argument gives $\|\bar{U}\bar{\Sigma}(V - \bar{V})^*\|_F \leq \epsilon/3$. To bound the middle term, observe that $\|\bar{U}(\Sigma - \bar{\Sigma})V^*\|_F = \|\Sigma - \bar{\Sigma}\|_F \leq \epsilon/3$. This completes the proof. ∎

We now prove Theorem 4.2.3. It is a standard argument from this point, and is essentially the same as the proof of Lemma 4.3 in [129], but we repeat it here to keep the results self-contained. We begin by showing that $\mathcal{A}$ is an approximate isometry on the covering set $\bar{S}_r$. Lemma 4.3.1 with $\epsilon = \delta/(4\sqrt{2})$ gives

$$|\bar{S}_r| \leq (36\sqrt{2}/\delta)^{(n_1+n_2+1)r}. \tag{4.3.3}$$

Then it follows from (4.2.2) together with the union bound that

$$\begin{aligned}
\mathbb{P}\left(\max_{\bar{X}\in\bar{S}_r} |\|\mathcal{A}(\bar{X})\|_{\ell_2}^2 - \|\bar{X}\|_F^2| > \delta/2\right) &\leq |\bar{S}_r| C e^{-cm} \\
&\leq 2(36\sqrt{2}/\delta)^{(n_1+n_2+1)r} C e^{-cm} \\
&= C\exp\Big((n_1+n_2+1)r\log(36\sqrt{2}/\delta) - cm\Big) \\
&\leq 2\exp(-dm)
\end{aligned}$$

where $d = c - \frac{\log(36\sqrt{2}/\delta)}{C}$ and we plugged in both requirements $m \geq C(n_1+n_2+1)r$ and $C > \log(36\sqrt{2}/\delta)/c$.

Now suppose that

$$\max_{\bar{X}\in\bar{S}_r} |\,\|\mathcal{A}(\bar{X})\|_{\ell_2}^2 - \|\bar{X}\|_F^2| \leq \delta/2$$

(which occurs with probability at least $1 - C\exp(-dm)$). We begin by showing that the upper bound in the RIP condition holds. Set

$$\kappa_r = \sup_{X\in S_r} \|\mathcal{A}(X)\|_{\ell_2}.$$

For any $X \in S_r$, there exists $\bar{X} \in \bar{S}_r$ with $\|X - \bar{X}\|_F \leq \delta/(4\sqrt{2})$ and, therefore,

$$\|\mathcal{A}(X)\|_{\ell_2} \leq \|\mathcal{A}(X-\bar{X})\|_{\ell_2} + \|\mathcal{A}(\bar{X})\|_{\ell_2} \leq \|\mathcal{A}(X-\bar{X})\|_{\ell_2} + 1 + \delta/2. \tag{4.3.4}$$

Put $\Delta X = X - \bar{X}$ and note that $\operatorname{rank}(\Delta X) \leq 2r$. Write $\Delta X = \Delta X_1 + \Delta X_2$, where $\langle \Delta X_1, \Delta X_2\rangle = 0$, and $\operatorname{rank}(\Delta X_i) \leq r$, $i = 1, 2$ (for example by splitting the SVD). Note that $\Delta X_1/\|\Delta X_1\|_F$, $\Delta X_2/\|\Delta X_2\|_F \in S_r$ and, thus,

$$\|\mathcal{A}(\Delta X)\|_{\ell_2} \leq \|\mathcal{A}(\Delta X_1)\|_{\ell_2} + \|\mathcal{A}(\Delta X_2)\|_{\ell_2} \leq \kappa_r(\|\Delta X_1\|_F + \|\Delta X_2\|_F). \tag{4.3.5}$$

Now $\|\Delta X_1\|_F + \|\Delta X_2\|_F \leq \sqrt{2}\|\Delta X\|_F$ which follows from $\|\Delta X_1\|_F^2 + \|\Delta X_2\|_F^2 = \|\Delta X\|_F^2$. Also, $\|\Delta X\|_F \leq \delta/(4\sqrt{2})$ leading to $\|\mathcal{A}(\Delta X)\|_{\ell_2} \leq \delta/4$. Plugging this into (4.3.4) gives

$$\|\mathcal{A}(X)\|_{\ell_2} \leq \kappa_r\delta/4 + 1 + \delta/2.$$

Since this holds for all $X \in S_r$, we have $\kappa_r \le \kappa_r \delta/4 + 1 + \delta/2$ and thus $\kappa_r \le (1+\delta/2)/(1-\delta/4) \le 1+\delta$ which essentially completes the upper bound. Now that this is established, the lower bound now follows from

$$\|\mathcal{A}(X)\|_{\ell_2} \ge \|\mathcal{A}(\bar{X})\|_{\ell_2} - \|\mathcal{A}(\Delta X)\|_{\ell_2} \ge 1 - \delta/2 - (1+\delta)\sqrt{2}\delta/(4\sqrt{2}) \ge 1 - \delta.$$

Note that we have shown

$$(1-\delta)\|X\|_F \le \|\mathcal{A}(X)\|_{\ell_2} \le (1+\delta)\|X\|_F,$$

which can then be easily translated into the desired version of the RIP bound.

4.3.3 Proof of Theorem 4.2.4

We prove Theorems 4.2.4, 4.2.7, and 4.2.8 for the matrix Dantzig selector (4.1.3) and describe in Section 4.3.7 how to extend these proofs to the matrix LASSO. We also assume that we are dealing with square matrices from this point forward ($n = n_1 = n_2$) for notational simplicity; the generalizations of the proofs to rectangular matrices are straightforward.

We begin by a lemma, which applies to full-rank matrices, and contains Theorem 4.2.4 as a special case.[6]

Lemma 4.3.2 *Suppose $\delta_{4r} < \sqrt{2} - 1$ and let M_r be any rank-r matrix. Let $M_c = M - M_r$. Suppose λ obeys $\|\mathcal{A}^*(z)\| \le \lambda$. Then the solution $\hat{M}$ to (4.1.3) obeys*

$$\|\hat{M} - M\|_F \le C_0 \sqrt{r}\lambda + C_1 \|M_c\|_*/r, \tag{4.3.6}$$

where C_0 and C_1 are small constants depending only on the isometry constant δ_{4r}.

We shall use the fact that $\mathcal{A}$ maps low-rank orthogonal matrices to approximately orthogonal vectors.

Lemma 4.3.3 *[41] For all X, X' obeying $\langle X, X' \rangle = 0$, and $\mathrm{rank}(X) \le r$, $\mathrm{rank}(X') \le r'$,*

$$|\langle \mathcal{A}(X), \mathcal{A}(X') \rangle| \le \delta_{r+r'} \|X\|_F \|X'\|_F.$$

Proof This is a simple application of the parallelogram identity. Suppose without loss of generality that X and X' have unit Frobenius norms. Then

$$(1-\delta_{r+r'})\|X \pm X'\|_F^2 \le \|\mathcal{A}(X \pm X')\|_{\ell_2}^2 \le (1+\delta_{r+r'})\|X \pm X'\|_F^2,$$

[6] We did not present this lemma in the main portion of the chapter because it does not seem to have an intuitive interpretation.

since $\operatorname{rank}(X \pm X') \leq r + r'$. We have $\|X \pm X'\|_F^2 = \|X\|_F^2 + \|X'\|_F^2 = 2$ and the parallelogram identity asserts that

$$|\langle \mathcal{A}(X), \mathcal{A}(X') \rangle| = \frac{1}{4} \left| \|\mathcal{A}(X + X')\|_{\ell_2}^2 - \|\mathcal{A}(X - X')\|_{\ell_2}^2 \right| \leq \delta_{r+r'},$$

which concludes the proof. ∎

The proof of Lemma 4.3.2 parallels that of Candès and Tao about the recovery of nearly sparse vectors from a limited number of measurements [43]. It is also inspired by the work of Fazel, Recht, Candès, and Parrilo [73, 129]. Set $H = \hat{M} - M$ and observe that by the triangle inequality,

$$\|\mathcal{A}^* \mathcal{A}(H)\| \leq \|\mathcal{A}^*(\mathcal{A}(\hat{M}) - y)\| + \|\mathcal{A}^*(y - \mathcal{A}(M))\| \leq 2\lambda, \tag{4.3.7}$$

since M is feasible for the problem (4.1.3). Decompose H as

$$H = H_0 + H_c,$$

where $\operatorname{rank}(H_0) \leq 2r$, $M_r H_c^* = 0$ and $M_r^* H_c = 0$ (see [129]). We have

$$\begin{aligned} \|M + H\|_* &\geq \|M_r + H_c\|_* - \|M_c\|_* - \|H_0\|_* \\ &= \|M_r\|_* + \|H_c\|_* - \|M_c\|_* - \|H_0\|_*. \end{aligned}$$

Since by definition, $\|M + H\|_* \leq \|M\|_* \leq \|M_r\|_* + \|M_c\|_*$, this gives

$$\|H_c\|_* \leq \|H_0\|_* + 2\|M_c\|_*. \tag{4.3.8}$$

Next, we use an estimate developed in [40] (see also [129]). This also seems to have deeper roots in approximation theory. Let $H_c = U \operatorname{diag}(\boldsymbol{\sigma}) V^*$ be the SVD of H_c, where $\boldsymbol{\sigma}$ is the list of ordered singular values (not to be confused with the noise standard deviation). Decompose H_c into a sum of matrices $H_1, H_2, \ldots$, each of rank at most $2r$ as follows. For each i define the index set $I_i = \{2r(i-1)+1, ..., 2ri\}$, and let $H_i := U_{I_i} \operatorname{diag}(\boldsymbol{\sigma}_{I_i}) V_{I_i}^*$; that is, H_1 is the part of H_c corresponding to the $2r$ largest singular values, H_2 is the part corresponding to the next $2r$ largest and so on. A now standard computation shows that

$$\sum_{j \geq 2} \|H_j\|_F \leq \frac{1}{\sqrt{2r}} \|H_c\|_*, \tag{4.3.9}$$

and thus

$$\sum_{j \geq 2} \|H_j\|_F \leq \|H_0\|_F + \sqrt{\frac{2}{r}} \|M_c\|_*$$

since $\|H_0\|_* \leq \sqrt{2r}\, \|H_0\|_F$ by Cauchy-Schwarz.

Now the restricted isometry property gives

$$(1-\delta_{4r})\|H_0+H_1\|_F^2 \le \|\mathcal{A}(H_0+H_1)\|_{\ell_2}^2, \tag{4.3.10}$$

and observe that

$$\|\mathcal{A}(H_0+H_1)\|_{\ell_2}^2 = \langle \mathcal{A}(H_0+H_1), \mathcal{A}(H-\sum_{j\ge 2} H_j)\rangle.$$

We first argue that

$$\langle \mathcal{A}(H_0+H_1), \mathcal{A}(H)\rangle \le \|H_0+H_1\|_F \sqrt{4r}\|\mathcal{A}^*\mathcal{A}(H)\|. \tag{4.3.11}$$

To see why this is true, let $U\Sigma V^*$ be the reduced SVD of H_0+H_1 in which U and V are $n\times r'$, and Σ is $r'\times r'$ with $r' = \operatorname{rank}(H_0+H_1) \le 4r$. We have

$$\begin{aligned}
\langle \mathcal{A}(H_0+H_1), \mathcal{A}(H)\rangle &= \langle H_0+H_1, \mathcal{A}^*\mathcal{A}(H)\rangle \\
&= \langle \Sigma, U^*[\mathcal{A}^*\mathcal{A}(H)]V\rangle \\
&\le \|\Sigma\|_F \|U^*[\mathcal{A}^*\mathcal{A}(H)]V\|_F \\
&= \|H_0+H_1\|_F \|U^*[\mathcal{A}^*\mathcal{A}(H)]V\|_F.
\end{aligned}$$

The claim follows from $\|U^*[\mathcal{A}^*\mathcal{A}(H)]V\|_F \le \sqrt{r'}\|\mathcal{A}^*\mathcal{A}(H)\|$, which holds since $U^*[\mathcal{A}^*\mathcal{A}(H)]V$ is an $r'\times r'$ matrix with spectral norm bounded by $\|\mathcal{A}^*\mathcal{A}(H)\|$. Second, Lemma 4.3.3 implies that for $j \ge 2$

$$\langle \mathcal{A}(H_0), \mathcal{A}(H_j)\rangle \le \delta_{4r}\|H_0\|_F\,\|H_j\|_F, \tag{4.3.12}$$

and similarly with H_1 in place of H_0. Note that because H_0 is orthogonal to H_1, we have that $\|H_0+H_1\|_F^2 = \|H_0\|_F^2 + \|H_1\|_F^2$ and thus $\|H_0\|_F + \|H_1\|_F \le \sqrt{2}\|H_0+H_1\|_F$. This gives

$$\langle \mathcal{A}(H_0+H_1), \mathcal{A}(H_j)\rangle \le \sqrt{2}\delta_{4r}\|H_0+H_1\|_F\,\|H_j\|_F. \tag{4.3.13}$$

Taken together, (4.3.10), (4.3.11), and (4.3.13) yield

$$\begin{aligned}
(1-\delta_{4r})\|H_0+H_1\|_F &\le \sqrt{4r}\|\mathcal{A}^*\mathcal{A}(H)\| + \sqrt{2}\delta_{4r}\sum_{j\ge 2}\|H_j\|_F \\
&\le \sqrt{4r}\|\mathcal{A}^*\mathcal{A}(H)\| + \sqrt{2}\delta_{4r}\|H_0\|_F + \frac{2\delta_{4r}}{\sqrt{r}}\|M_c\|_*.
\end{aligned}$$

To conclude, we have that

$$\|H_0+H_1\|_F \le C_1\sqrt{4r}\|\mathcal{A}^*\mathcal{A}(H)\| + C_1\frac{2\delta_{4r}}{\sqrt{r}}\|M_c\|_*, \quad C_1 = 1/[1-(\sqrt{2}+1)\delta_{4r}],$$

provided that $C_1 > 0$. Our claim (4.2.4) then follows from (5.3.6) together with

$$\|H\|_F \le \|H_0 + H_1\|_F + \sum_{j\ge 2} \|H_j\|_F \le 2\|H_0 + H_1\|_F + \sqrt{\frac{2}{r}}\|M_c\|_*.$$

4.3.4 Proof of Theorem 4.2.4

Theorem 4.2.4 follows by simply plugging $M_r = M$ into Theorem 4.3.2. To generalize the results, note that there are only two requirements on $M, \mathcal{A}$, and y used in the proof.

- $\|\mathcal{A}^*(\mathcal{A}(M) - y)\| \le \lambda$.
- $\operatorname{rank}(M) = r$ and $\delta_{4r} < \sqrt{2} - 1$.

Thus, the steps above also prove the following lemma which is useful in proving Theorem 4.2.7.

Lemma 4.3.4 *Assume that X is of rank at most r and that $\delta_{4r} < \sqrt{2} - 1$. Suppose λ obeys $\|\mathcal{A}^*(y - \mathcal{A}(X))\| \le \lambda$. Then the solution $\hat{M}$ to (4.1.3) obeys*

$$\|\hat{M} - X\|_F^2 \le C_0\, r\lambda^2, \tag{4.3.14}$$

where C_0 is a small constant depending only on the isometry constant δ_{4r}.

4.3.5 Proof of Theorem 4.2.7

In this section, $\lambda = 16\sqrt{n}\sigma$ and we take as given that $\|\mathcal{A}^*(z)\| \le \lambda/2$ (and thus, by Lemma 4.1.1, the end result holds with probability at least $1 - 2e^{-cn}$). The novelty in this proof—the way it differs from analogous proofs in compressive sensing—is in the use of a middle estimate $\bar{M}$. Define K as

$$K(X; M) \equiv \gamma \operatorname{rank}(X) + \|\mathcal{A}(X) - \mathcal{A}(M)\|_{\ell_2}^2, \qquad \gamma = \frac{\lambda^2}{4(1 + \delta_1)} \tag{4.3.15}$$

and let $\bar{M} = \operatorname{argmin}_X K(X, M)$. In words, $\bar{M}$ achieves a compromise between goodness of fit and parsimony in the model with noiseless data. The factor γ could be replaced by λ^2, but the derivations are cleanest in the present form. We note that similar minimizations have been considered in sparse approximation and compressed sensing, see [151, Appendix V] and the works discussed in the introduction of Chapter 3. In fact, an analogous result to Lemma 4.3.5 below is established in [151, Appendix V].

We begin by bounding the distance between M and $\bar{M}$ using the RIP, and obtain

$$\|\bar{M} - M\|_F^2 \le \frac{1}{1 - \delta_{2r}} \|\mathcal{A}(\bar{M}) - \mathcal{A}(M)\|_{\ell_2}^2 \tag{4.3.16}$$

where the use of the isometry constant δ_{2r} follows from the fact that $\operatorname{rank}(\bar{M}) \le \operatorname{rank}(M)$.

We now state a simple lemma which will be useful in our derivations. We defer the proof until later in the section.

Lemma 4.3.5 *The minimizer $\bar{M}$ obeys*

$$\left\|\mathcal{A}^*\mathcal{A}(\bar{M}-M)\right\| \le \lambda/2.$$

We use this lemma to develop a bound about $\|\hat{M}-\bar{M}\|_F^2$. Lemma 4.3.5 gives

$$\left\|\mathcal{A}^*(y-\mathcal{A}(\bar{M}))\right\| \le \|\mathcal{A}^*(z)\| + \left\|\mathcal{A}^*\mathcal{A}(M-\bar{M})\right\| \le \lambda,$$

i.e., $\bar{M}$ is feasible for (4.1.3). Also, $\operatorname{rank}(\bar{M}) \le \operatorname{rank}(M)$ and, thus, plugging $\bar{M}$ into Lemma 4.3.4 gives

$$\|\hat{M}-\bar{M}\|_F^2 \le C\lambda^2 \operatorname{rank}(\bar{M}).$$

Combining this with (4.3.16) gives

$$\begin{aligned}\|\hat{M}-M\|_F^2 &\le 2\|\hat{M}-\bar{M}\|_F^2 + 2\|\bar{M}-M\|_F^2 \\ &\le 2C\lambda^2 \operatorname{rank}(\bar{M}) + \frac{2}{1-\delta_{2r}}\|\mathcal{A}(\bar{M})-\mathcal{A}(M)\|_{\ell_2}^2 \\ &\le C'K(\bar{M};M) \qquad (4.3.17)\end{aligned}$$

where $C' = \max(8C(1+\delta_1), 2/(1-\delta_{2r}))$.

Now $\bar{M}$ is the minimizer of $K(\cdot;M)$, and so $K(\bar{M};M) \le K(M_0;M)$, where

$$M_0 = \sum_i \sigma_i(M) 1_{\{\sigma_i(M)>\lambda\}} u_i v_i^*. \qquad (4.3.18)$$

We have

$$\begin{aligned}K(M_0;M) &\le \gamma \sum_{i=1}^r 1_{\{\sigma_i(M)>\lambda\}} + \|\mathcal{A}(M-M_0)\|_{\ell_2}^2 \\ &\le \gamma \sum_{i=1}^r 1_{\{\sigma_i(M)>\lambda\}} + (1+\delta_r)\|M-M_0\|_F^2 \\ &\le (1+\delta_r)\sum_{i=1}^r \min(\lambda^2, \sigma_i^2(M)).\end{aligned}$$

In conclusion, the proof follows from $\lambda = 16\sqrt{n}\sigma$ since

$$\|\hat{M}-M\|_F^2 \le C' \sum_{i=1}^r \min(\lambda^2, \sigma_i^2(M)).$$

We now prove Lemma 4.3.5.

Proof [Lemma 4.3.5] Suppose not. Then there are unit-normed vectors $u, v \in \mathbb{R}^n$ obeying

$$\langle uv^*, \mathcal{A}^*\mathcal{A}(\bar{M} - M)\rangle > \lambda/2.$$

We construct the rank-1 perturbation $M' = \bar{M} - \alpha uv^*$, $\alpha = \langle uv^*, \mathcal{A}^*\mathcal{A}(\bar{M} - M)\rangle / \|\mathcal{A}(uv^*)\|_{\ell_2}^2$, and claim that $K(M' : M) < K(\bar{M}; M)$ thus providing the contradiction. We have

$$\begin{aligned}\|\mathcal{A}(M' - M)\|_{\ell_2}^2 &= \|\mathcal{A}(\bar{M} - M)\|_{\ell_2}^2 - 2\alpha\langle \mathcal{A}(uv^*), \mathcal{A}(\bar{M} - M)\rangle + \alpha^2\|\mathcal{A}(uv^*)\|_{\ell_2}^2 \\ &= \|\mathcal{A}(\bar{M} - M)\|_{\ell_2}^2 - \alpha^2\|\mathcal{A}(uv^*)\|_{\ell_2}^2.\end{aligned}$$

It then follows that

$$\begin{aligned}K(M'; M) &\le \gamma(\mathrm{rank}(M) + 1) + \|\mathcal{A}(\bar{M} - M)\|_{\ell_2}^2 - \alpha^2\|\mathcal{A}(uv^*)\|_{\ell_2}^2 \\ &= K(\bar{M}; M) + \gamma - \alpha^2\|\mathcal{A}(uv^*)\|_{\ell_2}^2.\end{aligned}$$

However, $\|\mathcal{A}(uv^*)\|_{\ell_2}^2 \le (1+\delta_1)\|uv^*\|_F^2 = 1+\delta_1$ and, therefore, $\alpha^2\|\mathcal{A}(uv^*)\|_{\ell_2}^2 > \gamma$ since $\langle uv^*, \mathcal{A}^*\mathcal{A}(\bar{M} - M)\rangle > \lambda/2$. ∎

4.3.6 Proof of Theorem 4.2.8

Three useful lemmas are established in the course of the proof of this more involved result, and we would like to point out that these can be used as powerful error bounds themselves. Throughout the proof, C is a constant that may depend on δ_{4r} only, and whose value may change from line to line. An important fact to keep in mind is that under the assumptions of the theorem, $\delta_{4\bar{r}}$ can be bounded, with high probability, by an arbitrarily small constant depending on the size of the scalar c_1 appearing in the condition $\bar{r} \le c_1 m/n$. This is a consequence of Theorem 4.2.3. In particular, $\delta_{4\bar{r}} \le (\sqrt{2} - 1)/2$ with probability at least $1 - De^{-dm}$.

Lemma 4.3.6 *Let $\bar{M}$ and M_0 be defined via* (4.3.15) *and* (4.3.18)*, and set*

$$r = \max(\mathrm{rank}(\bar{M}), \mathrm{rank}(M_0)).$$

Suppose that $\delta_{4r} < \sqrt{2} - 1$ and that λ obeys $\|\mathcal{A}^*(z)\| \le \lambda/2$. Then the solution $\hat{M}$ to (4.1.3) *obeys*

$$\|\hat{M} - M\|_F^2 \le C_0\left(\sum_{i=1}^{n} \min(\lambda^2, \sigma_i^2(M)) + \|\mathcal{A}(M - M_0)\|_{\ell_2}^2\right), \tag{4.3.19}$$

where C_0 is a small constant depending only on the isometry constant δ_{4r}.

Before we begin the proof, set $M_c = M - M_0$ so that M_c only contains the singular values below the noise level.

Proof The proof is essentially the same as that of Theorem 4.2.7, and so we quickly go through the main steps. First,

$$\begin{aligned}\|\bar{M} - M\|_F^2 &\leq 2\|\bar{M} - M_0\|_F^2 + 2\|M_c\|_F^2 \\ &\leq \frac{2}{1-\delta_{2r}}\|\mathcal{A}(\bar{M} - M_0)\|_{\ell_2}^2 + 2\|M_c\|_F^2 \\ &\leq \frac{4}{1-\delta_{2r}}\|\mathcal{A}(\bar{M} - M)\|_{\ell_2}^2 + \frac{4}{1-\delta_{2r}}\|\mathcal{A}(M_c)\|_{\ell_2}^2 + 2\|M_c\|_F^2.\end{aligned}$$

Second, we bound $\|\hat{M} - \bar{M}\|_F$ using the exact same steps as in the proof of Theorem 4.2.7, and obtain

$$\|\hat{M} - \bar{M}\|_F^2 \leq Cr\lambda^2.$$

Hence,

$$\|\hat{M} - M\|_F^2 \leq C(K(\bar{M}; M) + \|\mathcal{A}(M_c)\|_{\ell_2}^2 + \|M_c\|_F^2).$$

Finally, use $K(\bar{M}; M) \leq K(M_0; M)$ as before, and simplify to attain (4.3.19). ∎

The factor $\|\mathcal{A}(M_c)\|_F^2$ in (4.3.19) prevents us from stating the bound as the near-ideal bias-variance-trade-off (4.2.19). However, many random measurement ensembles obeying the RIP are also unlikely to drastically change the norm of *any* fixed matrix (see (4.2.2)). Thus, we expect that $\|\mathcal{A}(M_c)\|_{\ell_2} \approx \|M_c\|_{\ell_2}$ with high probability. This idea is not novel and has been developed in [115] and [166]. Specifically, if $\mathcal{A}$ obeys (4.2.2), then

$$\|\mathcal{A}(M_c)\|_{\ell_2}^2 \leq 1.5\|M_c\|_F^2 \tag{4.3.20}$$

with probability at least $1 - De^{-cm}$ for fixed constants D, c. An important point here is that this inequality only holds (with high probability) when M_c is fixed, and $\mathcal{A}$ is chosen randomly (independently). In the worst-case-scenario, one could have

$$\|\mathcal{A}(M_c)\|_{\ell_2} = \|\mathcal{A}\| \cdot \|M_c\|_F$$

where $\|\mathcal{A}\|$ is the operator norm of $\mathcal{A}$. Thus we emphasize that the bound holds with high probability for a given M verifying our conditions, but may not hold uniformly over all such M's.

Returning to the proof, (4.3.25) together with

$$\|M_c\|_F^2 = \sum_{i=1}^{n} \sigma_i^2(M) 1_{\{\sigma_i(M) < \lambda\}}$$

give the following lemma:

Lemma 4.3.7 *Fix M and suppose $\mathcal{A}$ obeys* (4.2.2)*. Then under the assumptions of Lemma 4.3.6, the solution $\hat{M}$ to* (4.1.3) *obeys*

$$\|\hat{M} - M\|_F^2 \leq C_0 \sum_{i=1}^{n} \min(\lambda^2, \sigma_i^2(M)) \tag{4.3.21}$$

with probability at least $1 - De^{-cn}$ where C_0 is a small constant depending only on the isometry constant δ_{4r}, and c, D are fixed constants.

The above two lemmas require a bound on the rank of M_0. However, as the noise level approaches zero, the rank of M_0 approaches the rank of M, which can be as large as the dimension. This requires further analysis, and in order to provide theoretical error bounds when the noise level is low (and M has full rank, say), a certain property of many measurement operators is useful. We call it the NNQ property, and is inspired by a similar property from compressive sensing, see [166].

Definition 4.3.8 (NNQ) *Let $B_*^{n\times n}$ be the set of $n \times n$ matrices with nuclear norm bounded 1. Let $B_{\ell_2}^m$ be the standard ℓ_2 unit ball for vectors in $\mathbb{R}^m$. We say that $\mathcal{A}$ satisfies NNQ(α) if*

$$\mathcal{A}(B_*^{n\times n}) \supseteq \alpha B_{\ell_2}^m. \tag{4.3.22}$$

This condition may appear cryptic at the moment. To give a taste of why it may be useful, note that Lemma 4.3.2 includes $\|M - M_r\|_*$ as part of the error bound. The point is that using the NNQ condition, we can find a proxy for $M - M_r$, which we call $\tilde{M}$, satisfying $\mathcal{A}(\tilde{M}) = \mathcal{A}(M - M_r)$, but also $\|\tilde{M}\|_* \leq \|\mathcal{A}(M - M_r)\|_{\ell_2}/\alpha$. Before continuing this line of thought, we prove that Gaussian measurement ensembles satisfy NNQ($\mu\sqrt{n/m}$) with high probability for some fixed constant $\mu > 0$.

Theorem 4.3.9 (NNQ for Gaussian measurements) *Suppose $\mathcal{A}$ is a Gaussian measurement ensemble and $m \leq Cn^2/\log(m/n)$ for some fixed constant $C > 0$. Then $\mathcal{A}$ satisfies NNQ($\mu\sqrt{n/m}$) with probability at least $1 - 3e^{-cn}$ for fixed constants c and μ.*

Proof Put $\alpha = \mu\sqrt{n/m}$ and suppose $\mathcal{A}$ does not satisfy NNQ(α). Then there exists a vector $x \in \mathbb{R}^m$ with $\|x\|_{\ell_2} = 1$ such that

$$\langle \mathcal{A}(M), x\rangle \leq \alpha \quad \text{for all } M \in B_*^{n\times n}.$$

In particular,

$$\|\mathcal{A}^*(x)\| \leq \alpha.$$

Let $\bar{B}_{\ell_2}^m \subset B_{\ell_2}^m$ be an α-net for $B_{\ell_2}^m$ with $|\bar{B}_{\ell_2}^m| \leq (3/\alpha)^m$. Then there exists $\bar{x} \in \bar{B}_{\ell_2}^m$ with $\|\bar{x} - x\|_{\ell_2} \leq \alpha$ satisfying

$$\|\mathcal{A}^*(\bar{x})\| \leq \|\mathcal{A}^*(\bar{x} - x)\| + \|\mathcal{A}(x)\| \leq \langle uv^*, \mathcal{A}^*(\bar{x} - x)\rangle + \alpha,$$

where u, v are the left and right singular vectors of $\mathcal{A}^*(\bar{x} - x)$ corresponding to the top singular value. Then

$$\langle uv^*, \mathcal{A}^*(\bar{x} - x)\rangle = \langle \mathcal{A}(uv^*), \bar{x} - x\rangle \le \|\mathcal{A}(uv^*)\|_{\ell_2} \|\bar{x} - x\|_{\ell_2} \le \sqrt{1 + \delta_1}\,\alpha$$

and, therefore,

$$\|\mathcal{A}^*(\bar{x})\| \le 3\alpha$$

assuming $\delta_1 \le 1$ (this occurs with probability at least $1 - 2e^{-cn}$ when $m \ge Cn$ for fixed constants c, C).

We will provide the contradiction by showing that with high probability, $\|\mathcal{A}^*(\bar{x})\| > 3\alpha$, for all $\bar{x} \in \bar{B}_*^{n\times n}$. For each $\bar{x}$, $\mathcal{A}^*(\bar{x})$ is equal in distribution to $\frac{1}{\sqrt{m}} Z$, where Z is a matrix with i.i.d. standard normal entries. Let Z_i be the ith column of Z. Then

$$\begin{aligned}\mathbb{P}(\|\mathcal{A}^*(\bar{x})\| \le 3\alpha) &\le \mathbb{P}(\|Z\| \le 3\sqrt{m}\alpha) \\ &\le \mathbb{P}(\max_{i=1,\dots,n} \|Z_i\|_{\ell_2} \le 3\sqrt{m}\alpha);\end{aligned}$$

the second step uses the fact that the operator norm of Z is always larger or equal to the ℓ_2 norm of any column. With $\alpha = \mu\sqrt{n/m}$ and using the fact that the columns are independent, this yields

$$\mathbb{P}(\|\mathcal{A}^*(\bar{x})\| \le 3\alpha) \le \mathbb{P}(\|Z_1\|_{\ell_2}^2 \le 9\mu^2 n)^n.$$

However, $\|Z_1\|_{\ell_2}^2$ is a chi-squared random variable with n degrees of freedom, and can be bounded using a standard concentration of measure result [99, Lemma 1]:

$$\mathbb{P}(\|Z_1\|_{\ell_2}^2 - n \le -t\sqrt{2n}) \le e^{-t^2/2}.$$

Hence,

$$\mathbb{P}(\|\mathcal{A}^*(\bar{x})\| \le 3\alpha) \le e^{-cn^2},$$

where $c = (1 - 9\mu^2)^2/4$ (we require $\mu < 1/3$ here). Thus, by the union bound,

$$\mathbb{P}\left(\min_{\bar{x}\in B_{\ell_2}^m} \|\mathcal{A}^*(\bar{x})\| \le 3\alpha\right) \le (3/\alpha)^m e^{-cn^2} = \exp\left(m \log\left(\frac{3\sqrt{m}}{\mu\sqrt{n}}\right) - cn^2\right) \le e^{-c'n^2}$$

provided that $m \le Cn^2/\log(m/n)$ for fixed constants, C, c'. The theorem is established. ∎

Note that the preceding proof can be repeated when $\mathcal{A}$ is a sub-Gaussian measurement ensemble; the only difference is that Z above will contain sub-Gaussian entries, rather than Gaussian entries.

Using the NNQ property, we can now bound the error when the noise level is low; this does not involve any condition on the rank of M_0, and does not involve a term in the bound depending on $\|M - M_0\|_*$.

Lemma 4.3.10 *Suppose that $\mathcal{A}$ satisfies NNQ($\mu\sqrt{n/m}$) for a fixed constant μ and that $\|\mathcal{A}^*(z)\| \le \lambda$. Let $\bar{r} \ge cm/n$ for some fixed numerical constant c, and suppose that that $\delta_{4\bar{r}} \le \frac{1}{2}(\sqrt{2}-1)$. Let*

$$M_{\bar{r}} = \sum_{i=1}^{\bar{r}} \sigma_i(M) u_i v_i^*.$$

Let $\hat{M}$ be the solution to (4.1.3)*. Then*

$$\|\hat{M} - M\|_F \le C(\lambda\sqrt{\bar{r}} + \|\mathcal{A}(M - M_{\bar{r}})\|_{\ell_2}) + \|M - M_{\bar{r}}\|_F. \tag{4.3.23}$$

Proof Set $M_c = M - M_{\bar{r}} = \sum_{i=\bar{r}+1}^{n} \sigma_i(M) u_i v_i^*$. The NNQ($\alpha$) property with $\alpha = \mu\sqrt{n/m}$ gives

$$\mathcal{A}(M_c) = \mathcal{A}(\tilde{M})$$

for some $\tilde{M}$ obeying $\|\tilde{M}\|_* \le \|\mathcal{A}(M_c)\|_{\ell_2}/\alpha$. We also take note of the identity $\mathcal{A}(M_{\bar{r}} + \tilde{M}) = \mathcal{A}(M)$. It follows from Lemma 4.3.2 that

$$\|\hat{M} - (M_{\bar{r}} + \tilde{M})\|_F \le C(\lambda\sqrt{\bar{r}} + \|\tilde{M}\|_*/\sqrt{\bar{r}}).$$

Plugging in $\|\tilde{M}\|_* \le \|\mathcal{A}(M_c)\|_{\ell_2}/\alpha$, along with $\bar{r} \ge cm/n$, we obtain

$$\|\hat{M} - (M_{\bar{r}} + \tilde{M})\|_F \le C(\lambda\sqrt{\bar{r}} + \|\mathcal{A}(M_c)\|_{\ell_2}).$$

Therefore,

$$\|\hat{M} - M\|_F \le C(\lambda\sqrt{\bar{r}} + \|\mathcal{A}(M_c)\|_{\ell_2}) + \|\tilde{M}\|_F + \|M_c\|_F. \tag{4.3.24}$$

It remains to bound $\|\tilde{M}\|_F$. As in the proof of Lemma 4.3.2, decompose $\tilde{M}$ as $\tilde{M} = \tilde{M}_1 + \tilde{M}_2 + \dots$ so that $\tilde{M}_1$ corresponds to the largest $\bar{r}$ singular values of $\tilde{M}$, $\tilde{M}_2$ corresponds with the next $\bar{r}$ largest, and so on. Just as before,

$$\|\tilde{M}\|_F \le \sum_i \|\tilde{M}_i\|_F \le \|\tilde{M}_1\|_F + \|\tilde{M}\|_*/\sqrt{\bar{r}}.$$

We now bound $\|\tilde{M}_1\|_F$. By the RIP,

$$\begin{aligned}\|\tilde{M}_1\|_F &\le \frac{1}{\sqrt{1-\delta_{\bar{r}}}}\|\mathcal{A}(\tilde{M}_1)\|_{\ell_2} \\ &= \frac{1}{\sqrt{1-\delta_{\bar{r}}}}\Big(\|\mathcal{A}(\tilde{M}) - \sum_{i\ge 2}\mathcal{A}(\tilde{M}_i)\|_{\ell_2}\Big) \\ &\le \frac{1}{\sqrt{1-\delta_{\bar{r}}}}\Big(\|\mathcal{A}(\tilde{M})\|_{\ell_2} + \sum_{i\ge 2}\|\mathcal{A}(\tilde{M}_i)\|_{\ell_2}\Big).\end{aligned}$$

By the RIP again, $\|\mathcal{A}(\tilde{M}_i)\|_{\ell_2} \le \sqrt{1+\delta_{\bar{r}}}\|\tilde{M}_i\|_F$, and so

$$\sum_{i\ge 2}\|\mathcal{A}(\tilde{M}_i)\|_{\ell_2} \le \sqrt{1+\delta_{\bar{r}}}\sum_{i\ge 2}\|\tilde{M}_i\|_F \le \sqrt{1+\delta_{\bar{r}}}\,\frac{\|\tilde{M}\|_*}{\sqrt{\bar{r}}}.$$

This together with $\mathcal{A}(\tilde{M}) = \mathcal{A}(M_c)$ give

$$\|\tilde{M}\|_F \le \sqrt{\frac{1+\delta_r}{1-\delta_r}}\Big(\|\mathcal{A}(M_c)\|_{\ell_2} + \frac{\|\tilde{M}\|_*}{\sqrt{\bar{r}}}\Big).$$

However, $\|\tilde{M}\|_* \le \|\mathcal{A}(M)\|_{\ell_2}/\alpha \le \sqrt{\bar{r}}\|\mathcal{A}(M)\|_{\ell_2}/(\mu\sqrt{c})$ and, therefore,

$$\|\tilde{M}\|_F \le C\|\mathcal{A}(M_c)\|_{\ell_2}.$$

Inserting this into (4.3.24) completes the proof of the lemma. ∎

We are now in position to prove our main theorem concerning the recovery of matrices with decaying singular values (Theorem 4.2.8). There are three cases to consider depending on the number of singular values of M standing above the noise level. In each case, we need the inequality

$$\|\mathcal{A}(M_c)\|_F^2 \le 1.5\|M_c\|_F^2 \tag{4.3.25}$$

which holds with probability at least $1 - De^{-cn}$ for any measurement ensemble satisfying (4.2.2) (including the Gaussian measurement ensemble). Put $\lambda = 16\sqrt{n}\sigma$ and recall the definition of M_0:

$$M_0 = \sum_{i=1}^{n}\sigma_i(M)1_{\{\sigma_i(M)\ge\lambda\}}\,u_i v_i^*$$

whose rank is exactly the number of singular values of M above the noise level. There are three cases to consider depending mostly on the interplay between the singular values of M and the noise level.

Case 1: high noise level

Suppose $K(M_0; M) \le \frac{\lambda^2}{4(1+\delta_1)}\bar{r}$. Then $\operatorname{rank}(M_0) \le \bar{r}$ and $\operatorname{rank}(\bar{M}) \le \bar{r}$ by definition of $\bar{M}$. Hence, Lemma 4.3.7 gives

$$\|\hat{M} - M\|_F^2 \le C \sum_{i=1}^{n} \min(n\sigma^2, \sigma_i^2(M))$$

with probability at least $1 - 2e^{-cn}$.

Case 2: low noise level

Suppose $K(M_0; M) > \frac{\lambda^2}{4(1+\delta_1)}\bar{r}$ and $\operatorname{rank}(M_0) \ge \bar{r}$. It follows from (4.2.2) that

$$\|\mathcal{A}(M - M_{\bar{r}})\|_{\ell_2}^2 \le \sqrt{1.5}\|M - M_{\bar{r}}\|_F^2 \tag{4.3.26}$$

with probability at least $1 - De^{-cn}$. Now, for the Gaussian measurement ensemble, the requirements of Lemma 4.3.10 are met with probability at least $1 - Ce^{-cn}$. Combining (4.3.26) with Lemma 4.3.10 yields

$$\|\hat{M} - M\|_F \le C(\lambda\sqrt{\bar{r}} + \|M - M_{\bar{r}}\|_F)$$

and thus

$$\|\hat{M} - M\|_F^2 \le 2C^2(\lambda^2\bar{r} + \|M - M_{\bar{r}}\|_F^2) = 2C^2\left(\sum_{i=1}^{\bar{r}} \min(\lambda^2, \sigma_i^2(M)) + \sum_{i=\bar{r}+1}^{n} \sigma_i^2(M)\right).$$

Since $\lambda = 16\sqrt{n}\sigma$, this is (4.2.20).

Case 3: medium noise level

Suppose $K(M_0; M) > \frac{\lambda^2}{4(1+\delta_1)}\bar{r}$ and $\operatorname{rank}(M_0) < \bar{r}$. As in Case 2, we have

$$\|\hat{M} - M\|_F^2 \le 2C^2(\lambda^2\bar{r} + \|M - M_{\bar{r}}\|_F^2).$$

From $\lambda^2\bar{r} < 4(1 + \delta_1)K(M_0; M)$, it follows that

$$\|\hat{M} - M\|_F^2 \le 2C^2(\lambda^2 \operatorname{rank}(M_0) + 4(1 + \delta_1)\|\mathcal{A}(M - M_0)\|_{\ell_2}^2 + \|M - M_{\bar{r}}\|_F^2).$$

We also have $\|\mathcal{A}(M - M_0)\|_{\ell_2}^2 \le 1.5\|M - M_0\|_F^2$ with probability at least $1 - De^{-cn}$. Inserting this bound into the previous equation, along with $\|M - M_{\bar{r}}\|_F \le \|M - M_0\|_F$, gives the desired conclusion.

These three cases comprise all possibilities. In short, the proof of Theorem 4.2.8 is complete.

4.3.7 Extension of proofs to the solution to the LASSO (4.1.5)

In the sparse regression setup, Bickel et al. [17] showed that the Dantzig selector and the LASSO have analogous properties, leading to analogous error bounds. The analogies still hold in the low-rank matrix recovery problem (for similar reasons). In fact, all of the theorems above also hold for the solution to (4.1.5) aside from a shift in those constants appearing in the assumptions, and those appearing in the error bounds. To see this, note that our proofs only used two crucial properties about $\hat{M}$:

1. $\|\hat{M}\|_* \le \|M\|_*$,
2. $\left\|\mathcal{A}^*(\mathcal{A}(\hat{M}) - y)\right\| \le \lambda$.

The second property automatically holds for the solution to (4.1.5) (but with λ replaced by μ). This follows from the optimality conditions which state that $\mathcal{A}^*(y - \mathcal{A}(\hat{M})) \in \mu\partial\|\hat{M}\|_*$ where $\|\hat{M}\|_*$ is the family of subgradients to the nuclear norm at the minimizer. Formally, let $U\Sigma V^*$ be the SVD of $\hat{M}$, then

$$\mathcal{A}^*(y - \mathcal{A}(\hat{M})) = \mu(UV^* + W)$$

for some W obeying $\|W\| \le 1$ and $U^*W = 0, WV = 0$ (see, e.g. [36]). Hence, the second property follows from $\|UV^* + W\| = \max(\|UV^*\|, \|W\|) \le 1$ (we use $U^*W, WV = 0$ to obtain the equality in the last equation).

The first property does not necessarily hold for the matrix LASSO, but a close enough approximation is verified (this is analogous to an argument made in [17]). Suppose that $\|\mathcal{A}^*(z)\| \le c_0\mu$ for a small constant c_0 (which, by Lemma 4.1.1, holds with high probability for Gaussian noise if $\mu^2 = Cn\sigma^2$). Then since $\hat{M}$ minimizes (4.1.5), we have

$$\frac{1}{2}\|\mathcal{A}(\hat{M}) - y\|_{\ell_2}^2 + \mu\|\hat{M}\|_* \le \frac{1}{2}\|\mathcal{A}(M) - y\|_{\ell_2}^2 + \mu\|M\|_*.$$

Plug in $y = \mathcal{A}(M) + z$ and rearrange terms to give

$$\mu\|\hat{M}\|_* \le \frac{1}{2}\|\mathcal{A}(\hat{M} - M)\|_{\ell_2}^2 + \mu\|\hat{M}\|_* \le \langle \hat{M} - M, \mathcal{A}^*(z)\rangle + \mu\|M\|_*.$$

Since the nuclear norm and the operator norm are dual to each other, we have $\langle \hat{M} - M, \mathcal{A}^*(z)\rangle \le \|\hat{M} - M\|_* \cdot \|\mathcal{A}^*(z)\| \le c_0\mu\|H\|_*$, where we use the notation $H = \hat{M} - M$ as in the proof of Lemma 4.3.2. This gives

$$\|\hat{M}\|_* \le c_0\|H\|_* + \|M\|_*,$$

which nearly is the first property. When c_0 is chosen to be a small constant, this factor has no essential detrimental effects on the proof. In particular, (5.3.5) in the proof of Lemma 4.3.2 is

replaced by

$$(1-c_0)\|H_c\|_* \le (1+c_0)\|H_0\|_* + 2\|M_c\|_*.$$

In particular, for $c_0 = 1/2$,

$$\|H_c\|_* \le 3\|H_0\|_* + 4\|M_c\|_*.$$

The rest of the proofs follow.

4.3.8 Proof of Theorem 4.2.5

We begin with a well-known lemma which gives the minimax risk for estimating the vector $x \in \mathbb{R}^n$ from the data $y \in \mathbb{R}^m$ and the linear model

$$y = Ax + z, \tag{4.3.27}$$

where $A \in \mathbb{R}^{m\times n}$ and the z_i's are i.i.d. $\mathcal{N}(0,\sigma^2)$.

Lemma 4.3.11 *Let $\lambda_i(A^*A)$ be the eigenvalues of the matrix A^*A. Then*[7]

$$\inf_{\hat{x}} \sup_{x\in\mathbb{R}^n} \mathbb{E}\,\|\hat{x}-x\|_{\ell_2}^2 = \sigma^2 \operatorname{trace}((A^*A)^{-1}) = \sum_i \frac{\sigma^2}{\lambda_i(A^*A)}. \tag{4.3.28}$$

In particular, if one of the eigenvalues vanishes (as in the case in which $m < n$), then the minimax risk is unbounded.

This result can be found as Exercise 5.8, p. 403, in the textbook [102]. We sketch the important steps. First, the minimax risk is lower bounded by the Bayes risk. Consider then the prior which assumes that all the components of x are i.i.d. $\mathcal{N}(0,\tau^2)$. At this point, to simplify the derivation it is convenient to diagonalize the problem—this is achievable since Gaussian vectors are invariant in distribution with respect to multiplication by orthogonal matrices. Now, the estimator which minimizes the Bayes risk can be explicitly calculated as the conditional expectation of x given y, and thus the Bayes risk itself can be explicitly calculated. Last, take the limit as $\tau \to \infty$; the estimator converges to the least-squares estimator, and the risk converges to the right-hand side of (4.3.28), thus completing the proof. In fact, these exact steps give the following more general lemma, which states that the least squares estimate, $\hat{x}_{\mathrm{LS}} := (A^*A)^{-1}A^*y$, is often minimax. (This lemma will be useful in the proof of Theorem 4.2.6.)

Lemma 4.3.12 *Let f be a monotonically increasing function. Then*

$$\inf_{\hat{x}} \sup_{x\in\mathbb{R}^n} \mathbb{E}\, f(\|\hat{x}-x\|_{\ell_2}) = \mathbb{E}\, f(\|\hat{x}_{LS}-x\|_{\ell_2}) = \mathbb{E}\, f\left(\|(A^*A)^{-1}A^*z\|_{\ell_2}\right). \tag{4.3.29}$$

[7] The infimum is over all measurable functions $\hat{x}(y)$ of y.

We are now in position to prove Theorem 4.2.5. The set of rank-r matrices is (much) larger than the set of matrices of the form

$$M = UR,$$

where U is a *fixed* orthogonal $n \times r$ matrix with orthonormal columns (note that the matrices of this form have a fixed r-dimensional column space). Thus,

$$\inf_{\hat{M}} \sup_{M:\mathrm{rank}(M)=r} \mathbb{E}\,\|\hat{M} - M\|_F^2 \geq \inf_{\hat{M}} \sup_{M:M=UR} \mathbb{E}\,\|\hat{M} - M\|_F^2.$$

Knowing that $M = UR$ for some unknown $r \times n$ matrix R, one can of course limit ourselves to estimators of the form $\hat{M} = U\hat{R}$, and since

$$\mathbb{E}\,\|\hat{M} - M\|_F^2 = \mathbb{E}\,\|U\hat{R} - UR\|_F^2 = \mathbb{E}\,\|\hat{R} - R\|_F^2,$$

the minimax risk is lower bounded by that of estimating R from the data

$$y = \mathcal{A}_U(R) + z,$$

where $\mathcal{A}_U$ is the linear map (4.2.15). We then apply Lemma 4.3.11 to conclude that the minimax rate is lower bounded by

$$\sum_i \frac{\sigma^2}{\lambda_i(\mathcal{A}_U^* \mathcal{A}_U)}.$$

The claim follows from the simple lemma below.

Lemma 4.3.13 *Let U be an $n \times r$ matrix with orthonormal columns. Then all the eigenvalues of $\mathcal{A}_U^* \mathcal{A}_U$ belong to the interval $[1 - \delta_r, 1 + \delta_r]$.*

Proof By definition,

$$\lambda_{\min}(\mathcal{A}_U^* \mathcal{A}_U) = \inf_{\|R\|_F \leq 1} \langle R, \mathcal{A}_U^* \mathcal{A}_U(R) \rangle$$

and similarly for $\lambda_{\max}(\mathcal{A}_U^* \mathcal{A}_U)$ with a sup in place of inf. Since

$$\langle R, \mathcal{A}_U^* \mathcal{A}_U(R) \rangle = \|\mathcal{A}_U(R)\|_{\ell_2}^2 = \|\mathcal{A}(UR)\|_{\ell_2}^2,$$

the claim follows from

$$(1 - \delta_r)\|UR\|_F^2 \leq \|\mathcal{A}(UR)\|_{\ell_2}^2 \leq (1 + \delta_r)\|UR\|_F^2,$$

which is valid since $\mathrm{rank}(UR) \leq r$ together with $\|UR\|_F^2 = \|R\|_F^2$. ∎

4.3.9 Proof of Theorem 4.2.6

The proof is similar to that of Theorem 4.2.5 and we begin with a lemma.

Lemma 4.3.14 *Suppose that x, y, A, z follow the linear model (4.3.27), with $z \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$. Then*

$$\inf_{\hat{x}} \sup_{x \in \mathbb{R}^n} \mathbb{P}\left(\|\hat{x} - x\|_{\ell_2}^2 \geq \frac{1}{2\|A\|^2} n\sigma^2 \right) \geq 1 - e^{-n/16}. \tag{4.3.30}$$

The Lemma is a straightforward application of Lemma 4.3.12, as follows:

Proof We set f to be the indicator function

$$f(t) := \mathbb{1}_{[t^2 \geq t_0^2]}, \qquad t_0^2 := \frac{1}{2\|A\|^2} n\sigma^2,$$

and we use Lemma 4.3.12 to give

$$\inf_{\hat{x}} \sup_{x \in \mathbb{R}^n} \mathbb{P}\left(\|\hat{x} - x\|_{\ell_2}^2 \geq t_0^2 \right) = \mathbb{P}\left(\|(A^*A)^{-1} A^* z\|_{\ell_2}^2 \geq t_0^2 \right) = \mathbb{P}\left(\|\mathcal{N}(0, (A^*A)^{-1})\|_{\ell_2}^2 \cdot \sigma^2 \geq t_0^2 \right). \tag{4.3.31}$$

However, $\|\mathcal{N}(0, (A^*A)^{-1})\|_{\ell_2}^2$ is equal in distribution to $\|\mathcal{N}(0, \Sigma^{-1})\|_{\ell_2}^2$ where Σ is the diagonal matrix of eigenvalues $\{\lambda_i\}$ of A^*A. Now let $z_1, \cdots, z_n$ be a sequence of iid standard normal random variables, so that we may write

$$\|\mathcal{N}(0, \Sigma^{-1})\|_{\ell_2}^2 = \sum_{i=1}^n \frac{z_i^2}{\lambda_i} \geq \frac{1}{\max_i \lambda_i} \sum_{i=1}^n z_i^2 = \frac{1}{\|A\|^2} \chi_n^2$$

where χ_n^2 is a chi-squared random variable with n degrees of freedom. Plug this into the right-hand side of (4.3.31) to give

$$\inf_{\hat{x}} \sup_{x \in \mathbb{R}^n} \mathbb{P}\left(\|\hat{x} - x\|_{\ell_2}^2 \geq t_0^2 \right) \geq \mathbb{P}\left(\chi_n^2 \cdot \sigma^2 \geq t_0^2 \|A\|^2 \right) = \mathbb{P}\left(\chi_n^2 \geq n/2 \right).$$

The result now follows from a standard concentration bound for χ^2 variables as in [99] which states that

$$\mathbb{P}(\chi_n^2 - n \leq -c\sqrt{2n}) \leq e^{-c^2/2}.$$

Taking $c = \sqrt{n}/(2\sqrt{2})$ finishes the proof. ∎

The proof of Theorem 4.2.6 now follows from the same steps as in the proof of Theorem 4.2.5. Once again, note that the worst-case probability (the LHS of equation (4.2.9)) can only decrease when we limit the space that M can dwell in. As before, constrain M to be of the form $M = UR$, so that

$$y = \mathcal{A}_U(R) + z.$$

The proof now follows from Lemma 4.3.13 which shows that $\|A_U\| \leq \sqrt{1+\delta_r}$.

4.4 Discussion

Using RIP-based analysis, we have shown that low-rank matrices can be stably recovered via nuclear-norm minimization from nearly the minimal possible number of linear samples. Further, the error bound is within a constant of the expected minimax error, and of an expected oracle error, and extends to the case when M has full rank. We pause to note that similar ideas with strong error bounds have been developed in randomized linear algebra (in this case the measurements consist of multiplications of M with random vectors, and are seen one vector at a time). See [92].

This work differs from the main thrust of the recent literature on low-rank matrix recovery, which has concentrated on the RIP-less matrix completion problem. An interesting observation regarding matrix completion is that when the measurements are randomly chosen entries of M, one requires at least about $nr \log n$ measurements to recover M *by any method* when $\operatorname{rank}(M) = O(1)$ [36,44]. In contrast, the results in this chapter show that on the order of nr measurements are enough provided these are sufficiently random.

The popularity of the matrix completion model stems from the fact that this setup currently dominates the applications of low-rank matrix recovery. There are far fewer applications in which the measurements are random linear combinations of many entries of M (quantum-state tomography is a notable application though). As a great deal of attention is given to low-rank matrix modeling these days, with new applications being discovered all the time, this may change rapidly. We hope that our theory encourages further applications and research in this direction.

Chapter 5

Matrix completion with noise

5.1 Introduction

Imagine that we only observe a few samples of a signal. Is it possible to reconstruct this signal exactly or at least accurately? For example, suppose we observe a few entries of a vector $x \in \mathbb{R}^n$, which we can think of as a digital signal or image. Can we recover the large fraction of entries—of pixels if you will—that we have not seen? In general, everybody would agree that this is impossible. However, if the signal is known to be sparse in the Fourier domain and, by extension, in an incoherent domain, then accurate—and even exact—recovery is possible by ℓ_1 minimization as shown in [39] and also discussed in Chapter 2.

Imagine now that we only observe a few entries of a data matrix. Then is it possible to accurately—or even exactly—guess the entries that we have not seen? For example, suppose we observe a few movie ratings from a large data matrix in which rows are users and columns are movies (we can only observe a few ratings because each user is typically rating a few movies as opposed to the tens of thousands of movies which are available). Can we predict the rating a user would hypothetically assign to a movie he/she has not seen? In general, everybody would agree that recovering a data matrix from a subset of its entries is impossible. However, if the unknown matrix is known to have low rank or approximately low rank, then accurate and even exact recovery is possible by nuclear-norm minimization [36, 44]. This revelation, which to some extent is inspired by the great body of work in compressed sensing, is the subject of this chapter.

From now on, we will refer to the problem of inferring the many missing entries as the *matrix completion* problem. Now just as sparse signal recovery is arguably of paramount importance these days, we do believe that matrix completion will become increasingly studied in years to come. For now, we give a few examples of applications in which these problems do come up.

- *Collaborative filtering.* In a few words, collaborative filtering is the task of making automatic predictions about the interests of a user by collecting taste information from many users [84].

Perhaps the most well-known implementation of collaborating filtering is the Netflix recommendation system alluded to earlier, which seeks to make rating predictions about unseen movies. This is a matrix completion problem in which the unknown full matrix has approximately low rank because only a few factors typically contribute to an individual's tastes or preferences. In the new economy, companies are interested predicting musical preferences (Apple, Inc.), literary preferences (Amazon, Barnes and Noble), and many other such things.

- *Global positioning.* Finding the global positioning of points in Euclidean space from a local or partial set of pairwise distances is a problem in geometry that emerges naturally in sensor networks [19,140,141]. For example, because of power constraints, sensors may only be able to construct reliable distance estimates from their immediate neighbors. From these estimates, we can form a partially observed distance matrix, and the problem is to infer all the pairwise distances from just a few observed ones so that locations of the sensors can be reliably estimated. This reduces to a matrix completion problem where the unknown matrix is of rank two if the sensors are located in the plane, and three if they are located in space.

- *Remote sensing.* The MUSIC algorithm [138] is frequently used to determine the direction of arrival of incident signals in a coherent radio-frequency environment. In a typical application, incoming signals are being recorded at various sensor locations, and this algorithm operates by extracting the directions of wave arrivals from the covariance matrix obtained by computing the correlations of the signals received at all sensor pairs. In remote sensing applications, one may not be able to estimate or transmit all correlations because of power constraints [165]. In this case, we would like to infer a full covariance matrix from just a few observed partial correlations. This is a matrix completion problem in which the unknown signal covariance matrix has low rank since it is equal to the number of incident waves, which is usually much smaller than the number of sensors.

There are of course many other examples of matrix completion and low-rank matrix recovery in general, including the structure-from-motion problem [46, 148] in computer vision, multi-class learning in data analysis [5, 6], and so on.

This chapter investigates whether or not one can recover a low-rank matrix from a small subsets of its entries, and if so, how and how well. In Section 5.2, we will study the noiseless problem in which the observed entries are precisely those of the unknown matrix; this section is a review of known results. Section 5.3 examines the more common situation in which the few available entries are corrupted with noise, and in this chapter offers novel results. We complement our study with a few numerical experiments demonstrating the empirical performance of our methods in Section 5.4 and conclude with a discussion of open problems and the most recent related literature, which became available after the material in this chapter was published (Section 5.5).

Before we begin, it is best to provide a brief summary of the notations used throughout the chapter. We shall use three norms of a matrix $X \in \mathbb{R}^{n_1 \times n_2}$ with singular values $\{\sigma_k\}$. The *spectral norm* is denoted by $\|X\|$ and is the largest singular value. The Euclidean inner product between two matrices is defined by the formula $\langle X, Y \rangle := \text{trace}(X^* Y)$, and the corresponding Euclidean norm is called the *Frobenius norm* and denoted by $\|X\|_F$ (note that this is the ℓ_2 norm of the vector of singular values). The *nuclear norm* is denoted by $\|X\|_* := \sum_k \sigma_k$ and is the sum of singular values (the ℓ_1 norm of the vector $\{\sigma_k\}$). As is standard, $X \succeq Y$ means that $X - Y$ is positive semidefinite.

Further, we will also manipulate linear transformations which act on the space $\mathbb{R}^{n_1 \times n_2}$, and we will use calligraphic letters for these operators as in $\mathcal{A}(X)$. In particular, the identity operator on this space will be denoted by $\mathcal{I} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$. We use the same convention as above, and $\mathcal{A} \succeq \mathcal{I}$ means that $\mathcal{A} - \mathcal{I}$ (seen as a big matrix) is positive semidefinite.

We use the usual asymptotic notation, for instance writing $O(M)$ to denote a quantity bounded in magnitude by CM for some absolute constant $C > 0$.

5.2 Exact Matrix Completion

From now on, $M \in \mathbb{R}^{n_1 \times n_2}$ is a matrix we would like to know as precisely as possible. However, the only information available about M is a sampled set of entries M_{ij}, $(i, j) \in \Omega$, where Ω is a subset of the complete set of entries $[n_1] \times [n_2]$. (Here and in the sequel, $[n]$ denotes the list $\{1, \ldots, n\}$.) It will be convenient to summarize the information available via $\mathcal{P}_\Omega(M)$, where the sampling operator $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ is defined by

$$[\mathcal{P}_\Omega(X)]_{ij} = \begin{cases} X_{ij}, & (i, j) \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the question is whether it is possible to recover our matrix only from the information $\mathcal{P}_\Omega(M)$. We will assume that the entries are selected at random without replacement as to avoid trivial situations in which a row or a column is unsampled, since matrix completion is clearly impossible in such cases. (If we have no data about a specific user, how can we guess his/her preferences? If we have no distance estimates about a specific sensor, how can we guess its distances to all the sensors?)

Even with the information that the unknown matrix M has low rank, this problem may be severely ill posed. Here is an example that shows why: let x be a vector in $\mathbb{R}^n$ and consider the $n \times n$

rank-1 matrix

$$M = e_1 x^* = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_{n-1} & x_n \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where e_1 is the first vector in the canonical basis of $\mathbb{R}^n$. Clearly, this matrix cannot be recovered from a subset of its entries. Even if one sees 95% of the entries sampled at random, then we will miss elements in the first row with very high probability, which makes the recovery of the vector x, and by extension of M, impossible. The analogy in compressed sensing is that one obviously cannot recover a signal assumed to be sparse in the time domain, by subsampling in the time domain!

This example shows that one cannot hope to complete the matrix if some of the singular vectors of the matrix are extremely sparse—above, one cannot recover M without sampling all the entries in the first row, see [36] for other related pathological examples. More generally, if a row (or column) has no relationship to the other rows (or columns) in the sense that it is approximately orthogonal, then one would basically need to see all the entries in that row to recover the matrix M. Such informal considerations led the authors of [36] to introduce a geometric incoherence assumption, but for the moment, we will discuss an even simpler notion which forces the singular vectors of M to be spread across all coordinates. To express this condition, recall the singular value decomposition (SVD) of a matrix of rank r,

$$M = \sum_{k \in [r]} \sigma_k u_k v_k^*, \tag{5.2.1}$$

in which $\sigma_1, \ldots, \sigma_r \geq 0$ are the singular values, and $u_1, \ldots, u_r \in \mathbb{R}^{n_1}$, $v_1, \ldots, v_r \in \mathbb{R}^{n_2}$ are the singular vectors. Our assumption—which is also considered in [36] and following works—is as follows:

$$\|u_k\|_{\ell_\infty} \leq \sqrt{\mu_B / n_1}, \quad \|v_k\|_{\ell_\infty} \leq \sqrt{\mu_B / n_2}, \tag{5.2.2}$$

for some $\mu_B \geq 1$, where the ℓ_∞ norm is of course defined by $\|x\|_{\ell_\infty} = \max_i |x_i|$. We think of μ_B as being small, e.g., $O(1)$, so that the singular vectors are not too spiky as explained above.

If the singular vectors of M are sufficiently spread, the hope is that there is a unique low-rank matrix which is consistent with the observed entries. If this is the case, one could, in principle, recover the unknown matrix by solving

$$\begin{aligned} &\text{minimize} && \operatorname{rank}(X) \\ &\text{subject to} && \mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(M), \end{aligned} \tag{5.2.3}$$

where $X \in \mathbb{R}^{n_1 \times n_2}$ is the decision variable. Unfortunately, not only is this problem NP-hard, but all

known algorithms for exactly solving it are doubly exponential in theory and in practice [50]. This is analogous to the intractability of ℓ_0-minimization in sparse signal recovery.

A popular alternative is the convex relaxation [36, 44, 72, 74, 129]

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & \mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(M), \end{array} \tag{5.2.4}$$

(see [15, 111] for the earlier related trace heuristic). Just as ℓ_1-minimization is the tightest convex relaxation of the combinatorial ℓ_0-minimization problem in the sense that the ℓ_1 ball of $\mathbb{R}^n$ is the convex hull of unit-normed 1-sparse vectors (i.e., vectors with at most one nonzero entry), nuclear-norm minimization is the tightest convex relaxation of the NP-hard rank minimization problem. To be sure, the nuclear ball $\{X \in \mathbb{R}^{n_1 \times n_2} : \|X\|_* \leq 1\}$ is the convex hull of the set of rank-one matrices with spectral norm bounded by one. Moreover, in compressed sensing, ℓ_1 minimization subject to linear equality constraints can be cast as a linear program (LP) for the ℓ_1 norm has an LP characterization: indeed for each $x \in \mathbb{R}^n$, $\|x\|_{\ell_1}$ is the optimal value of

$$\begin{array}{ll} \text{maximize} & \langle u, x \rangle \\ \text{subject to} & \|u\|_{\ell_\infty} \leq 1, \end{array}$$

with decision variable $u \in \mathbb{R}^n$. In the same vein, the nuclear norm of $X \in \mathbb{R}^{n_1 \times n_2}$ has the SDP characterization

$$\begin{array}{ll} \text{maximize} & \langle W, X \rangle \\ \text{subject to} & \|W\| \leq 1, \end{array} \tag{5.2.5}$$

with decision variable $W \in \mathbb{R}^{n_1 \times n_2}$. This expresses the fact that the spectral norm is dual to the nuclear norm. The constraint on the spectral norm of W is an SDP constraint since it is equivalent to

$$\begin{bmatrix} I_{n_1} & W \\ W^* & I_{n_2} \end{bmatrix} \succeq 0,$$

where I_n is the $n \times n$ identity matrix. Hence, (5.2.4) is an SDP, which one can express by writing $\|X\|_*$ as the optimal value of the SDP dual to (5.2.5). Moreover, specialized algorithms that take advantage of the structure of the problem have been shown to outperform interior-point methods by several orders of magnitude (see [25, 104]).

In [44], it is proven that nuclear-norm minimization succeeds nearly as soon as recovery is possible by any method whatsoever.

Theorem 5.2.1 *[44] Let $M \in \mathbb{R}^{n_1 \times n_2}$ be a fixed matrix of rank $r = O(1)$ obeying (5.2.2) and set $n := \max(n_1, n_2)$. Suppose we observe m entries of M with locations sampled uniformly at random.*

Then there is a positive numerical constant C such that if

$$m \geq C\,\mu_B^4\, n \log^2 n, \tag{5.2.6}$$

then M is the unique solution to (5.2.4) *with probability at least $1 - n^{-3}$. In other words: with high probability, nuclear-norm minimization recovers all the entries of M with no error.*

As a side remark, one can obtain a probability of success at least $1 - n^{-\beta}$ for a given β by taking C in (5.2.6) of the form $C'\beta$ for some universal constant C'. The probabilistic nature of this result stems from the assumption that the revealed entries of M are sampled from the uniform distribution. Another interpretation is that matrix completion is exact for 'most' sampling sets obeying (5.2.6).

An $n_1 \times n_2$ matrix of rank r depends upon $r(n_1 + n_2 - r)$ degrees of freedom[1]. When r is small, the number of degrees of freedom is much less than $n_1 n_2$ and this is the reason why subsampling is possible. (In compressed sensing, the number of degrees of freedom corresponds to the sparsity of the signal; i.e., the number of nonzero entries.) What is remarkable here, is that exact recovery by nuclear-norm minimization occurs as soon as the sample size exceeds the number of degrees of freedom by a couple of logarithmic factors. Further, observe that if Ω completely misses one of the rows (e.g., one has no rating about one user) or one of the columns (e.g., one has no rating about one movie), then one cannot hope to recover even a matrix of rank 1 of the form $M = xy^*$. Thus one needs to sample every row (and also every column) of the matrix. When Ω is sampled at random, it is well established that one needs at least on the order $O(n \log n)$ for this to happen as this is the famous coupon collector's problem. Hence, (5.2.6) misses the information theoretic limit by at most a logarithmic factor.

To obtain similar results for all values of the rank, [44] introduces the *strong incoherence property* with parameter μ stated below.

A1 Let P_U (resp. P_V) be the orthogonal projection onto the singular vectors $u_1, \ldots, u_r$ (resp. $v_1, \ldots, v_r$). For all pairs $(a, a') \in [n_1] \times [n_1]$ and $(b, b') \in [n_2] \times [n_2]$,

$$\left| \langle e_a, P_U e_{a'} \rangle - \frac{r}{n_1} 1_{a=a'} \right| \leq \mu \frac{\sqrt{r}}{n_1},$$
$$\left| \langle e_b, P_V e_{b'} \rangle - \frac{r}{n_2} 1_{b=b'} \right| \leq \mu \frac{\sqrt{r}}{n_2}.$$

A2 Let E be the 'sign matrix' defined by

$$E = \sum_{k \in [r]} u_k v_k^*. \tag{5.2.7}$$

[1] This can be seen by counting the number of parameters in the singular value decomposition.

For all $(a, b) \in [n_1] \times [n_2]$,

$$|E_{ab}| \leq \mu \frac{\sqrt{r}}{\sqrt{n_1 n_2}}.$$

These conditions do not assume anything about the singular values. As we will see, incoherent matrices with a small value of the strong incoherence parameter μ can be recovered from a minimal set of entries. Before we state this result, it is important to note that many model matrices obey the strong incoherence property with a small value of μ.

- Suppose the singular vectors obey (5.2.2) with $\mu_B = O(1)$ (which informally says that the singular vectors are not spiky), then with the exception of a very few peculiar matrices, M obeys the strong incoherence property with $\mu = O(\sqrt{\log n})$. Specifically, there is a generic random model under which $\mu = O(\sqrt{\log n})$ with very high probability, see [36].

- Assume that the column matrices $[u_1, \ldots, u_r]$ and $[v_1, \ldots, v_r]$ are independent random orthogonal matrices, then with high probability, M obeys the strong incoherence property with $\mu = O(\sqrt{\log n})$, at least when $r \geq \log n$ as to avoid small samples effects.

The sampling result below is general, nonasymptotic and optimal up to a few logarithmic factors.

Theorem 5.2.2 *[44] Let $M \in \mathbb{R}^{n_1 \times n_2}$ be a fixed rank-r matrix with strong incoherence parameter μ, and set $n := \max(n_1, n_2)$. Suppose we observe m entries of M with locations sampled uniformly at random. Then there is a numerical constant C such that if*

$$m \geq C \mu^2 n r \log^6 n, \tag{5.2.8}$$

M is the unique solution to (5.2.4) *with probability at least $1 - n^{-3}$.*

In other words, if a matrix is strongly incoherent and the cardinality of the sampled set is about the number of degrees of freedom times a few logarithmic factors, then nuclear-norm minimization is exact. This improves on an earlier result of Candès and Recht [36] who proved—under slightly different assumptions—that on the order of $n^{6/5} r \log n$ samples were sufficient, at least for values of the rank obeying $r \leq n^{1/5}$.

More recently, a new matrix-completion result has further reduced the number of measurements required. This is encapsulated in the following theorem, which is due to the theoretical developments of Gross [88], and is also derived by mainly the same techniques in Recht's paper [128]. In fact, [88] allows for much more general measurement bases—for example, it handles the Pauli measurements discussed in Chapter 4—while Recht's paper specializes these results to matrix completion. We present the version in [128]; the parameter μ_1 appearing in this theorem is quite similar to the strong coherence parameter μ, and will be defined just below.

Theorem 5.2.3 *Let $M \in \mathbb{R}^{n_1 \times n_2}$ be a fixed rank-r matrix with incoherence parameter μ_1, and set $n := \max(n_1, n_2)$. Suppose we observe m entries of M with locations sampled uniformly at random. Then if*

$$m \geq 32\mu_1 r(n_1 + n_2)\beta log^2(2n) \tag{5.2.9}$$

M is the unique solution to (5.2.4) with probability at least $1 - 6\log(n_2)(n_1 + n_2)^{2-2\beta} - n^{2-2\beta^{1/2}}$.

Above, μ_1 is the smallest scalar value such that

A3

$$\sqrt{\frac{n_1}{r}} \max_{1 \leq a \leq n_1} \|P_U e_a\|_{\ell_2} \leq \mu_1$$
$$\sqrt{\frac{n_2}{r}} \max_{1 \leq a \leq n_2} \|P_V e_a\|_{\ell_2} \leq \mu_1$$

A4 (This is the same as A2.)Let E be the sign matrix defined in A2. For all $(a, b) \in [n_1] \times [n_2]$,

$$|E_{ab}| \leq \mu_1 \frac{\sqrt{r}}{\sqrt{n_1 n_2}}.$$

We would now like to point out a result of a broadly similar nature, but with a completely different recovery algorithm and with a somewhat different range of applicability, which was recently established by Keshavan, Oh, and Montanari [96]. Their conditions are related to the incoherence property introduced in [36], and are also satisfied by a number of reasonable random matrix models. There is, however, another condition which states that the singular values of the unknown matrix cannot be too large or too small (the ratio between the top and lowest value must be bounded). This algorithm 1) trims each row and column with too many entries; i.e., replaces the entries in those rows and columns by zero and 2) computes the SVD of the trimmed matrix, truncates it as to only keep the top r singular values (note that the value of r is needed here), and rescales. The result is that under some suitable conditions discussed above, this recovers a good approximation to the matrix M provided that the number of samples be on the order of nr. The recovery at this point is not exact, but one can now perform local minimization to achieve exact recovery provided that one has more samples (on the order of $nr \max(\log n, r)$), and in fact the recovery is stable provided that the noise level is small [97]. This builds upon an earlier spectral technique suggested by the computer science community [7], which also proves stability results, but under stronger conditions.

5.2.1 Geometry and dual certificates

We cannot rehash the proof of Theorem 5.2.2 from [44] in this chapter, or even explain the main technical steps, because of space limitations. We will, however, detail sufficient and almost necessary

conditions for the low-rank matrix M to be the unique solution to the SDP (5.2.4). This will be useful to establish stability results.

The recovery is exact if the feasible set is tangent to the nuclear ball at the point M. To express this mathematically[2], standard duality theory asserts that M is a solution to (5.2.4) if and only if there exists a dual matrix Λ such that $\mathcal{P}_\Omega(\Lambda)$ is a subgradient of the nuclear norm at M, written as

$$\mathcal{P}_\Omega(\Lambda) \in \partial\|M\|_*. \tag{5.2.10}$$

Recall the SVD (5.2.1) of M and the 'sign matrix' E (5.2.7). It is well-known that $Z \in \partial\|M\|_*$ if and only if Z is of the form,

$$Z = E + W, \tag{5.2.11}$$

where

$$P_U W = 0, \quad W P_V = 0, \quad \|W\| \le 1. \tag{5.2.12}$$

In English, Z is a subgradient if it can be decomposed as the sign matrix plus another matrix with spectral norm bounded by one, whose column (resp. row) space is orthogonal to the span of $u_1, \ldots, u_r$, (resp. of $v_1, \ldots, v_r$). Another way to put this is by using notations introduced in [36]. Let T be the linear space spanned by elements of the form $u_k x^*$ and $y v_k^*$, $k \in [r]$, and let $T^\perp$ be the orthogonal complement to T. Note that $T^\perp$ is the set of matrices obeying $P_U W = 0$ and $W P_V = 0$. Then, $Z \in \partial\|M\|_*$ if and only if

$$Z = E + \mathcal{P}_{T^\perp}(Z), \quad \|\mathcal{P}_{T^\perp}(Z)\| \le 1.$$

This motivates the following definition.

Definition 5.2.4 (Dual certificate) *We say that Λ is a dual certificate if Λ is supported on Ω ($\Lambda = \mathcal{P}_\Omega(\Lambda)$), $\mathcal{P}_T(\Lambda) = E$ and $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1$.*

Before continuing, we would like to pause to observe the relationship with ℓ_1 minimization. The point $x^\star \in \mathbb{R}^n$ is solution to

$$\begin{aligned} &\text{minimize} && \|x\|_{\ell_1} \\ &\text{subject to} && Ax = b, \end{aligned} \tag{5.2.13}$$

with $A \in \mathbb{R}^{m\times n}$ if and only if there exists $\lambda \in \mathbb{R}^m$ such that $A^*\lambda \in \partial\|x^\star\|_{\ell_1}$. Note that if $S^\star$ is the

[2]In general, M minimizes the nuclear norm subject to the linear constraints $\mathcal{A}(X) = b$, $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$, if and only if there is $\lambda \in \mathbb{R}^m$ such that $\mathcal{A}^*(\lambda) \in \partial\|M\|_*$.

support of $x^\star$, $z \in \partial\|x^\star\|_{\ell_1}$ is equivalent to

$$z = e + w, \quad e = \begin{cases} \operatorname{sgn}(x_i^\star), & i \in S^\star, \\ 0, & i \notin S^*, \end{cases}$$

and

$$w_i = 0 \text{ for all } i \in S, \quad \|w\|_{\ell_\infty} \le 1.$$

Hence, there is a clear analogy and one can think of T defined above as playing the role of the support set in the sparse recovery problem.

With this in place, we shall make use of the following lemma from [36]:

Lemma 5.2.5 *[36] Suppose there exists a dual certificate Λ and consider any H obeying $\mathcal{P}_\Omega(H) = 0$. Then*

$$\|M + H\|_* \ge \|M\|_* + (1 - \|\mathcal{P}_{T^\perp}(\Lambda)\|)\|\mathcal{P}_{T^\perp}(H)\|_*.$$

Proof For any $Z \in \partial\|M\|_*$, we have

$$\|M + H\|_* \ge \|M\|_* + \langle Z, H\rangle.$$

With $\Lambda = E + \mathcal{P}_{T^\perp}(\Lambda)$ and $Z = E + \mathcal{P}_{T^\perp}(Z)$, we have

$$\begin{aligned} \|M + H\|_* &\ge \|M\|_* + \langle \Lambda, H\rangle + \langle \mathcal{P}_{T^\perp}(Z - \Lambda), H\rangle \\ &= \|M\|_* + \langle Z - \Lambda, \mathcal{P}_{T^\perp}(H)\rangle \end{aligned}$$

since $\mathcal{P}_\Omega(H) = 0$. Now we use the fact that the nuclear and spectral norms are dual to one another. In particular, there exists $\|\bar{Z}\| \le 1$ such that $\langle \bar{Z}, \mathcal{P}_{T^\perp}(H)\rangle = \|\mathcal{P}_{T^\perp}(H)\|_*$. Now pick Z such that $\mathcal{P}_{T^\perp}(Z) = \mathcal{P}_{T^\perp}(\bar{Z})$ so that $\langle Z, \mathcal{P}_{T^\perp}(H)\rangle = \|\mathcal{P}_{T^\perp}(H)\|_*$. Second, note that $|\langle \Lambda, \mathcal{P}_{T^\perp}(H)\rangle| = |\langle \mathcal{P}_{T^\perp}(\Lambda), \mathcal{P}_{T^\perp}(H)\rangle| \le \|\mathcal{P}_{T^\perp}(\Lambda)\|\|\mathcal{P}_{T^\perp}(H)\|_*$. Therefore,

$$\|M + H\|_* \ge \|M\|_* + (1 - \|\mathcal{P}_{T^\perp}(\Lambda)\|)\|\mathcal{P}_{T^\perp}(H)\|_*,$$

which concludes the proof. ∎

A consequence of this lemma are the sufficient conditions below.

Lemma 5.2.6 *[36] Suppose there exists a dual certificate obeying $\|\mathcal{P}_{T^\perp}(\Lambda)\| < 1$ and that the restriction $\mathcal{P}_\Omega \restriction_T : T \to \mathcal{P}_\Omega(\mathbb{R}^{n\times n})$ of the (sampling) operator $\mathcal{P}_\Omega$ restricted to T is injective. Then M is the unique solution to the convex program* (5.2.4).

Proof Consider any feasible perturbation $M + H$ obeying $\mathcal{P}_\Omega(H) = 0$. Then by assumption, Lemma

5.2.5 gives

$$\|M + H\|_* > \|M\|_*$$

unless $\mathcal{P}_{T^\perp}(H) = 0$. Assume then that $\mathcal{P}_{T^\perp}(H) = 0$; that is to say, $H \in T$. Then $\mathcal{P}_\Omega(H) = 0$ implies that $H = 0$ by the injectivity assumption. The conclusion is that M is the unique minimizer since any nontrivial perturbation increases the nuclear norm. ■

The methods for proving that matrix completion by nuclear minimization is exact, consist in constructing a dual certificate.

Theorem 5.2.7 *[44] Under the assumptions of either Theorem 5.2.1 or Theorem 5.2.2, there exists a dual certificate obeying $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1/2$. In addition, if $p = m/(n_1 n_2)$ is the fraction of observed entries, the operator $\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T : T \to T$ is one-to-one and obeys*

$$\frac{p}{2}\mathcal{I} \preceq \mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T \preceq \frac{3p}{2}\mathcal{I}, \tag{5.2.14}$$

where $\mathcal{I} : T \to T$ is the identity operator.

The second part, namely, (5.2.14) shows that the mapping $\mathcal{P}_\Omega : T \to \mathbb{R}^{n_1 \times n_2}$ is injective. Hence, the sufficient conditions of Lemma 5.2.6 are verified, and the recovery is exact. What is interesting, is that the existence of a dual certificate together with the near-isometry (5.2.14)—in fact, the lower bound—are sufficient to establish the robustness of matrix completion vis a vis noise.

5.3 Stable Matrix Completion

In any real world application, one will only observe a few entries corrupted at least by a small amount of noise. In the Netflix problem, users' ratings are uncertain. In the system identification problem, one cannot determine the locations $y(t)$ with infinite precision. In the global positioning problem, local distances are imperfect. And finally, in the remote sensing problem, the signal covariance matrix is always modeled as being corrupted by the covariance of noise signals. Hence, to be broadly applicable, we need to develop results which guarantee that reasonably accurate matrix completion is possible from noisy sampled entries. This section develops novel results showing that this is, indeed, the case.

Our noisy model assumes that we observe

$$Y_{ij} = M_{ij} + Z_{ij}, \quad (i,j) \in \Omega, \tag{5.3.1}$$

where $\{Z_{ij} : (i,j) \in \Omega\}$ is a noise term which may be stochastic or deterministic (adversarial).

Another way to express this model is as

$$\mathcal{P}_\Omega(Y) = \mathcal{P}_\Omega(M) + \mathcal{P}_\Omega(Z),$$

where Z is an $n \times n$ matrix with entries Z_{ij} for $(i,j) \in \Omega$ (note that the values of Z outside of Ω are irrelevant). All we assume is that $\|\mathcal{P}_\Omega(Z)\|_F \le \delta$ for some $\delta > 0$. For example, if $\{Z_{ij}\}$ is a white noise sequence with standard deviation σ, then $\delta^2 \le (m + \sqrt{8m})\sigma^2$ with high probability, say. To recover the unknown matrix, we propose solving the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & \|\mathcal{P}_\Omega(X - Y)\|_F \le \delta. \end{array} \tag{5.3.2}$$

Among all matrices consistent with the data, find the one with minimum nuclear norm. This is also an SDP, and let $\hat{M}$ be the solution to this problem.

Our main result is that this reconstruction is accurate.

Theorem 5.3.1 *With the notations of Theorem 5.2.7, suppose there exists a dual certificate obeying* $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1/2$ *and that* $\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T \succeq \frac{p}{2}\mathcal{I}$ *(both these conditions are true with very large probability under the assumptions of the noiseless recovery Theorems 5.2.1 and 5.2.2). Then* $\hat{M}$ *obeys*

$$\|M - \hat{M}\|_F \le 4\sqrt{\frac{C_p \min(n_1, n_2)}{p}}\,\delta + 2\delta, \tag{5.3.3}$$

with $C_p = 2 + p$.

The same error bound also holds when there exists an 'inexact dual certificate' (see [88]); this was proved by an adaptation of our techniques in [90]. It is significant because the tightest matrix completion result (Theorem 5.2.3) is proved by constructing an inexact dual certificate.

For small values of p (recall this is the fraction of observed entries), the error is of course at most just about $4\sqrt{\frac{2\min(n_1,n_2)}{p}}\,\delta$. As we will see from the proof, there is nothing special about $1/2$ in the condition $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le 1/2$. All we need is that there is a dual certificate obeying $\|\mathcal{P}_{T^\perp}(\Lambda)\| \le a$ for some $a < 1$ (the value of a only influences the numerical constant in (5.3.3)). Further, when Z is random, (5.3.3) holds on the event $\|\mathcal{P}_\Omega(Z)\|_F \le \delta$.

Roughly speaking, our theorem states the following: *when perfect noiseless recovery occurs, then matrix completion is stable vis a vis perturbations.* To be sure, the error is proportional to the noise level δ; when the noise level is small, the error is small. Moreover, improving conditions under which noiseless recovery occurs, has automatic consequences for the more realistic recovery from noisy samples.

A significant novelty here is that there is no equivalent of this result in the compressed sensing

or statistical literature; in particular our matrix completion problem since it does not obey the *restricted isometry property* (RIP) [41]. For matrices, the RIP would assume that the sampling operator obeys

$$(1-\delta)\|X\|_F^2 \leq \frac{1}{p}\|\mathcal{P}_\Omega(X)\|_F^2 \leq (1+\delta)\|X\|_F^2 \tag{5.3.4}$$

for all matrices X with sufficiently small rank and $\delta < 1$ sufficiently small [129]. However, the RIP does not hold here. To see why, let the sampled set Ω be arbitrarily chosen and fix $(i,j) \notin \Omega$. Then the rank-1 matrix $e_i e_j^*$ whose (i,j)th entry is 1, and vanishes everywhere else, obeys $\mathcal{P}_\Omega(e_i e_j^*) = 0$. Clearly, this violates (5.3.4).

It is nevertheless instructive to compare (5.3.3) with the bound one would achieve if the RIP (5.3.4) were true. In this case, [73] would give

$$\|\hat{M} - M\|_F \leq C_0 p^{-1/2}\,\delta$$

for some numerical constant C_0. That is, an estimate which would be better by a factor proportional to $1/\sqrt{\min(n_1,n_2)}$.

We close this section by emphasizing that our methods are also applicable to sparse signal recovery problems in which the RIP does not hold.

5.3.1 Proof of Theorem 5.3.1

We use the notation of the previous section, and begin the proof by observing two elementary properties. The first is that since M is feasible for (5.3.2), we have the *cone constraint*

$$\|\hat{M}\|_* \leq \|M\|_*. \tag{5.3.5}$$

The second is that the triangle inequality implies the *tube constraint*

$$\begin{aligned}\|\mathcal{P}_\Omega(\hat{M} - M)\|_F &\leq \|\mathcal{P}_\Omega(\hat{M} - Y)\|_F + \|\mathcal{P}_\Omega(Y - M)\|_F \\ &\leq 2\delta,\end{aligned} \tag{5.3.6}$$

since M is feasible. We will see that under our hypotheses, (5.3.5) and (5.3.6) imply that $\hat{M}$ is close to M. Set $\hat{M} = M + H$ and put $H_\Omega := \mathcal{P}_\Omega(H)$, $H_{\Omega^c} := \mathcal{P}_{\Omega^c}(H)$ for short. We need to bound $\|H\|_F^2 = \|H_\Omega\|_F^2 + \|H_{\Omega^c}\|_F^2$, and since (5.3.6) gives $\|H_\Omega\|_F \leq 2\delta$, it suffices to bound $\|H_{\Omega^c}\|_F$. Note that by the Pythagorean identity, we have

$$\|H_{\Omega^c}\|_F^2 = \|\mathcal{P}_T(H_{\Omega^c})\|_F^2 + \|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F^2, \tag{5.3.7}$$

and it is thus sufficient to bound each term in the right-hand side.

We start with the second term. Let Λ be a dual certificate obeying $\|\mathcal{P}_{T^\perp}(\Lambda)\| \leq 1/2$, we have

$$\|M + H\|_* \geq \|M + H_{\Omega^c}\|_* - \|H_\Omega\|_*$$

and

$$\|M + H_{\Omega^c}\|_* \geq \|M\|_* + [1 - \|\mathcal{P}_{T^\perp}(\Lambda)\|]\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_*.$$

The second inequality follows from Lemma 5.2.5. Therefore, with $\|\mathcal{P}_{T^\perp}(\Lambda)\| \leq 1/2$, the cone constraint gives

$$\|M\|_* \geq \|M\|_* + \frac{1}{2}\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_* - \|H_\Omega\|_*,$$

or, equivalently,

$$\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_* \leq 2\|H_\Omega\|_*.$$

Since the nuclear norm dominates the Frobenius norm, $\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F \leq \|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_*$, we have

$$\begin{aligned}\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F &\leq 2\|H_\Omega\|_* \\ &\leq 2\sqrt{n}\|H_\Omega\|_F \leq 4\sqrt{n}\delta,\end{aligned} \tag{5.3.8}$$

where the second inequality follows from the Cauchy-Schwarz inequality, and the last from (5.3.6).

To develop a bound on $\|\mathcal{P}_T(H_{\Omega^c})\|_F$, observe that the assumption $\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T \geq \frac{p}{2}\mathcal{I}$ together with $\mathcal{P}_T^2 = \mathcal{P}_T$, $\mathcal{P}_\Omega^2 = \mathcal{P}_\Omega$ give

$$\begin{aligned}\|\mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c})\|_F^2 &= \langle \mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c}), \mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c})\rangle \\ &= \langle \mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c}), \mathcal{P}_T(H_{\Omega^c})\rangle \\ &\geq \frac{p}{2}\|\mathcal{P}_T(H_{\Omega^c})\|_F^2.\end{aligned}$$

But since $\mathcal{P}_\Omega(H_{\Omega^c}) = 0 = \mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c}) + \mathcal{P}_\Omega\mathcal{P}_{T^\perp}(H_{\Omega^c})$, we have

$$\begin{aligned}\|\mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c})\|_F &= \|\mathcal{P}_\Omega\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F \\ &\leq \|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F.\end{aligned}$$

Hence, the last two inequalities give

$$\|\mathcal{P}_T(H_{\Omega^c})\|_F^2 \leq \frac{2}{p}\|\mathcal{P}_\Omega\mathcal{P}_T(H_{\Omega^c})\|_F^2 \leq \frac{2}{p}\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F^2. \tag{5.3.9}$$

As a consequence of this and (5.3.7), we have

$$\|H_{\Omega^c}\|_F^2 \le \Big(\frac{2}{p}+1\Big)\|\mathcal{P}_{T^\perp}(H_{\Omega^c})\|_F^2.$$

The theorem then follows from this inequality together with (5.3.8).

5.3.2 Comparison with an oracle

We would like to return to discussing the best possible accuracy one could ever hope for. For simplicity, assume that $n_1 = n_2 = n$, and suppose that we have an oracle informing us about T. In many ways, going back to the discussion from Section 5.2.1, this is analogous to giving away the support of the signal in compressed sensing [43]. With this precious information, we would know that M lives in a linear space of dimension $2nr - r^2$ and would probably solve the problem by the method of least squares:

$$\begin{aligned} &\text{minimize} && \|\mathcal{P}_\Omega(X) - \mathcal{P}_\Omega(Y)\|_F \\ &\text{subject to} && X \in T. \end{aligned} \tag{5.3.10}$$

That is, we would find the matrix in T, which best fits the data in a least-squares sense. Let $\mathcal{A} : T \to \Omega$ (we abuse notations and let Ω be the range of $\mathcal{P}_\Omega$) be defined by $\mathcal{A} := \mathcal{P}_\Omega\mathcal{P}_T$. Then assuming that the operator $\mathcal{A}^*\mathcal{A} = \mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T$ mapping T onto T is invertible (which is the case under the hypotheses of Theorem 5.3.1), the least-squares solution is given by

$$\begin{aligned} M^{\text{Oracle}} &:= (\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Y) \\ &= M + (\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Z). \end{aligned} \tag{5.3.11}$$

Hence,

$$\|M^{\text{Oracle}} - M\|_F = \|(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Z)\|_F.$$

Let Z' be the minimal (normalized) eigenvector of $\mathcal{A}^*\mathcal{A}$ with minimum eigenvalue $\lambda_{\min}$, and set $Z = \delta\lambda_{\min}^{-1/2}\mathcal{A}(Z')$ (note that by definition $\mathcal{P}_\Omega(Z) = Z$ since Z is in the range of $\mathcal{A}$).[3] By construction, $\|Z\|_F = \delta$, and

$$\|(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Z)\|_F = \lambda_{\min}^{-1/2}\delta \gtrsim p^{-1/2}\,\delta$$

since by assumption, all the eigenvalues of $\mathcal{A}^*\mathcal{A} = \mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T$ lie in the interval $[p/2, 3p/2]$. The matrix Z defined above also maximizes $\|(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Z)\|_F$ among all matrices bounded by δ and so the oracle achieves

$$\|M^{\text{Oracle}} - M\|_F \approx p^{-1/2}\delta \tag{5.3.12}$$

[3] To clarify, Z' is itself a matrix but it may be useful to picture it as a vector with $n_1 \cdot n_2$ entries.

with adversarial noise. Consequently, our analysis loses a $\sqrt{n}$ factor vis a vis an optimal bound that is achievable via the help of an oracle.

The diligent reader may argue that the least-squares solution above may not be of rank r (it is at most of rank $2r$) and may thus argue that this is not the strongest possible oracle. However, as explained below, if the oracle gave T and r, then the best fit in T of rank r would not do much better than (5.3.12). In fact, there is an elegant way to understand the significance of this oracle which we now present. Consider a stronger oracle which reveals the row space of the unknown matrix M (and thus the rank of the matrix). Then we would know that the unknown matrix is of the form

$$M = M_C R^*,$$

where M_C is an $n \times r$ matrix, and R is an $n \times r$ matrix whose columns form an orthobasis for the row space (which we can build since the oracle gave us perfect information). We would then fit the nr unknown entries by the method of least squares and find $X \in \mathbb{R}^{n\times r}$ minimizing

$$\|\mathcal{P}_\Omega(XR^*) - \mathcal{P}_\Omega(Y)\|_F.$$

Using our previous notations, the oracle gives away $T_0 \subset T$ where T_0 is the span of elements of the form yv_k^*, $k \in [r]$, and is more precise. If $\mathcal{A}_0 : T_0 \to \Omega$ is defined by $\mathcal{A}_0 := \mathcal{P}_\Omega \mathcal{P}_{T_0}$, then the least-squares solution is now

$$(\mathcal{A}_0^*\mathcal{A}_0)^{-1}\mathcal{A}_0^*(Y).$$

Because all the eigenvalues of $\mathcal{A}_0^*\mathcal{A}_0$ belong to $[\lambda_{\min}(\mathcal{A}^*\mathcal{A}), \lambda_{\max}(\mathcal{A}^*\mathcal{A})]$, the previous analysis applies and this stronger oracle would also achieve an error of size about $p^{-1/2}\delta$. In conclusion, when all we know is $\|\mathcal{P}_\Omega(Z)\|_F \le \delta$, one cannot hope for a root-mean squared error better than $p^{-1/2}\delta$.

Note that when the noise is stochastic, e.g., when Z_{ij} is white noise with standard deviation σ, the oracle gives an error bound which is adaptive, and is smaller as the rank gets smaller. Indeed, $\mathbb{E}\,\|(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*(Z)\|_F^2$ is equal to

$$\sigma^2 \operatorname{trace}((\mathcal{A}^*\mathcal{A})^{-1}) \approx \frac{2nr - r^2}{p}\sigma^2 \approx \frac{2nr}{p}\sigma^2, \tag{5.3.13}$$

since all the $2nr - r^2$ eigenvalues of $(\mathcal{A}^*\mathcal{A})^{-1}$ are just about equal to p^{-1}. When $nr \ll m$, this is better than (5.3.12).

5.4 Numerical Experiments

We have seen that matrix completion is stable amid noise. To emphasize the practical nature of this result, a series of numerical matrix completion experiments were run with noisy data. To be

n	100	200	500	1000
RMS error	.99	.61	.34	.24

Table 5.1: RMS error ($\|\hat{M} - M\|_F/n$) as a function of n when subsampling 20% of an $n \times n$ matrix of rank two. Each RMS error is averaged over 20 experiments.

precise, for several values of the dimension n (our first experiments concern $n \times n$ matrices), the rank r, and the fraction of observed entries $p = m/n^2$, the following numerical simulations were repeated 20 times, and the errors averaged. A rank-r matrix M is created as the product of two rectangular matrices, $M = M_L M_R^*$, where the entries of $M_L, M_R \in \mathbb{R}^{n\times r}$ are iid $N(0, \sigma_n^2 := 20/\sqrt{n})$[4]. The sampled set Ω is picked uniformly at random among all sets with m entries. The observations $\mathcal{P}_\Omega(Y)$ are corrupted by noise as in (5.3.1), where $\{Z_{ij}\}$ is iid $N(0, \sigma^2)$; here, we take $\sigma = 1$. Last, $\hat{M}$ is recovered as the solution to (5.4.1) below.

For a peek at the results, consider Table 5.1. The RMS error defined as $\|\hat{M} - M\|_F/n$, measures the root-mean squared error per entry. From the table, one can see that even though each entry is corrupted by noise with variance 1, when M is a 1000 by 1000 matrix, the RMS error per entry is .24. To see the significance of this, suppose one had the chance to see *all* the entries of the noisy matrix $Y = M + Z$. Naively accepting Y as an estimate of M would lead to an expected MS error of $\mathbb{E}\,\|Y - M\|_F^2/n^2 = \mathbb{E}\,\|Z\|_F^2/n^2 = 1$, whereas the MS error achieved from only viewing 20% of the entries is $\|\hat{M} - M\|_F^2/n^2 = .24^2 = .0576$ when solving the SDP (5.4.1)! Not only are we guessing accurately the entries we have not seen, but we also 'denoise' those we have seen.

In order to stably recover M from a fraction of noisy entries, the following regularized nuclear-norm minimization problem was solved using the FPC algorithm from [104],

$$\text{minimize} \quad \frac{1}{2}\|\mathcal{P}_\Omega(X - Y)\|_F^2 + \mu\|X\|_*. \tag{5.4.1}$$

It is a standard duality result that (5.4.1) is equivalent to (5.3.2), for some value of μ, and thus one could use (5.4.1) to solve (5.3.2) by searching for the value of $\mu(\delta)$ giving $\|\mathcal{P}_\Omega(\hat{M} - Y)\|_F = \delta$ (assuming $\|\mathcal{P}_\Omega(Y)\|_F > \delta$). We use (5.4.1) because it works well in practice, and because the FPC algorithm solves (5.4.1) nicely and accurately. We also remark that a variation on our stability proof could also give a stable error bound when using the SDP (5.4.1).

It is vital to choose a suitable value of μ, which we do with the following heuristic argument: first, simplifying to the case when Ω is the set of all elements of the matrix, note that the solution of (5.4.1) is equal to Y but with singular values shifted towards zero by μ (soft-thresholding), as can be seen from the optimality conditions of Section 5.2 by means of subgradients, or see [25].

[4]The value of σ_n is rather arbitrary. Here, it is set so that the singular values of M are quite larger than the singular values of $\mathcal{P}_\Omega(Z)$ so that M can be distinguished from the null matrix. Having said that, note that for large n and small r, the entries of M are much smaller than those of the noise, and thus the signal appears to be completely buried in noise.

When Ω is not the entire set, the solution is no longer exactly a soft-thresholded version of Y, but experimentally, it is generally close. Thus, we want to pick μ large enough to threshold away the noise (keep the variance low), and small enough not to overshrink the original matrix (keep the bias low). To this end, μ is set to be the smallest possible value such that if $M = 0$ and $Y = Z$, then it is likely that the minimizer of (5.4.1) satisfies $\hat{M} = 0$. It can be seen that the solution to (5.4.1) is $\hat{M} = 0$ if $\|\mathcal{P}_\Omega(Y)\| \leq \mu$ (once again, check the subgradient or [25]). Then the question is: what is $\|\mathcal{P}_\Omega(Z)\|$? If we make a nonessential change in the way Ω is sampled, then the answer follows from random matrix theory. Rather than picking Ω uniformly at random, choose Ω by selecting each entry with probability p, independently of the others. With this modification, each entry of $\mathcal{P}_\Omega(Z)$ is iid with variance $p\sigma^2$. Then if $Z \in \mathbb{R}^{n\times n}$, it is known that $n^{-1/2}\|\mathcal{P}_\Omega(Z)\| \to \sqrt{2p}\sigma$, almost surely as $n \to \infty$. Thus we pick $\mu = \sqrt{2np}\sigma$, where $p = m/n^2$. In practice, this value of μ seems to work very well for square matrices. For $n_1 \times n_2$ matrices, based on the same considerations, the proposal is $\mu = (\sqrt{n_1} + \sqrt{n_2})\sqrt{p}\sigma$ with $p = m/(n_1 n_2)$.

In order to interpret our numerical results, they are compared to those achieved by the oracle, see Section 5.3.2. To this end, Figure 5.1 plots three curves for varying values of n, p, and r: 1) the RMS error introduced above, 2) the RMS error achievable when the oracle reveals T, and the problem is solved using least squares, 3) the estimated oracle root expected MS error derived in Section 5.3.2, i.e., $\sqrt{\mathrm{df}/[n^2 p]} = \sqrt{\mathrm{df}/m}$, where $\mathrm{df} = r(2n - r)$. In our experiments, as n and m/df increased, with $r = 2$, the RMS error of the nuclear norm problem appeared to be fit very well by $1.68\sqrt{\mathrm{df}/m}$. Thus, to compare the oracle error to the actual recovered error, we plotted the oracle errors times 1.68. We also note that in our experiments, the RMS error was never greater than $2.25\sqrt{\mathrm{df}/m}$.

No one can predict the weather. We conclude the numerical section with a real world example. We retrieved from the website [1] a 366×1472 matrix whose entries are daily average temperatures at 1472 different weather stations throughout the world in 2008. Checking its SVD reveals that this is an approximately low rank matrix as expected. In fact, letting M be the temperature matrix, and calling M_2 the matrix created by truncating the SVD after the top two singular values gives $\|M_2\|_F/\|M\|_F = .9927$.

We first tested whether the incoherence assumptions described above were satisfied. Since M_2 contained almost all of the energy in M, we measured μ_B in terms of the singular vectors of M_2 and found $\mu_B = 3.83$. We considered this to be small because μ_B is bounded as $1 \leq \mu_B \leq 1472{\cdot}5 \approx 38.4$.

To test the performance of our matrix completion algorithm, we subsampled 30% of M and then recovered an estimate, $\hat{M}$, using (5.4.1). Note that this is a much different problem than those proposed earlier in this section. Here, we attempt to recover a matrix that is not exactly low rank, but only approximately. The solution gives a relative error of $\|\hat{M} - M\|_F/\|M\|_F = .166$. For

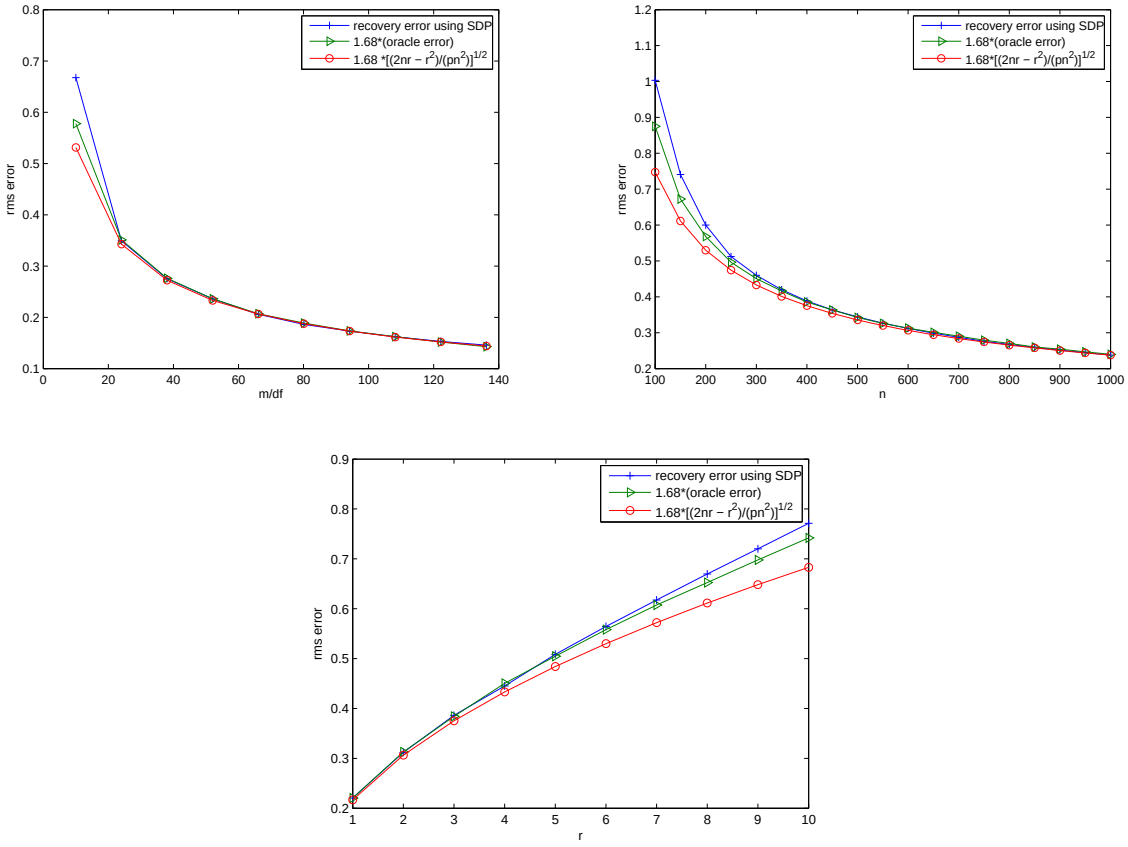


Figure 5.1: Comparison between the recovery error, the oracle error times 1.68, and the estimated oracle error times 1.68. Each point on the plot corresponds to an average over 20 trials. Top left: in this experiment $n = 600, r = 2$, and p varies. The x-axis is the number of measurements per degree of freedom (df). Top right: n varies, whereas $r = 2$, $p = .2$. Bottom: $n = 600$, r varies, and $p = .2$.

comparison[5], exact knowledge of the best rank-2 approximation achieves $\|M_2 - M\|_F / \|M\|_F = .121$. Here μ has been selected to give a good cross-validated error and is about 535.

5.5 Discussion

This chapter reviewed and developed some new results about matrix completion. By and large, matrix completion is a field in complete infancy abounding with interesting and open questions, and if the recent avalanche of results in compressed sensing is any indication, it is likely that this field will experience tremendous growth in the next few years.

Noisy matrix completion appears to be a difficult problem, in particular because it is RIP-less. Thus it is difficult to achieve near-optimal error bounds under general assumptions, and indeed the error bounds presented in this chapter were clearly non-optimal. Nevertheless, a few very recent papers solving somewhat different tractable problems [97,98,119] have demonstrated near-ideal error bounds for matrix completion, but under restrictive conditions. In particular, Montanari et al. [97] demonstrate near-ideal error bounds but under a requirement that the condition number of M be bounded by a constant, and certain coherence-type requirements hold as well. Wainwright et al. [119] and Koltchinskii et al. [98] also demonstrate near-ideal error bounds under coherence-type requirements when the noise level is high. However, it remains an open problem to weaken these assumptions, and in particular to give near-ideal error bounds that apply when the noise level is low and without bounding the condition number.

[5]The number 2 is somewhat arbitrary here, although we picked it because there is a large drop-off in the size of the singular values after the second. If, for example, M_{10} is the best rank-10 approximation, then $\|M_{10} - M\|_F / \|M\|_F = .081$.

Chapter 6

Conclusion

In Chapter 2 we demonstrated the efficacy of ℓ_1-minimization-based programs, namely the LASSO and Dantzig selector, in recovering sparse signals from the linear model $y = Ax + \sigma z$. We showed that in many cases the number of measurements necessary (i.e., the length of y) for stable signal recovery is no more than about $s \log n$. Moreover, the error achieved is within a polylogarithmic factor of the oracle error achieved by regressing y onto A_T (where T contains the support of x). These results were also generalized to the the case when x is approximately sparse; none of these results required the restricted isometry property.

One key motivation for this work in CS is the application to MRI. However, besides for angiography, many important MRI applications are outside of the scope of the current CS theory. In particular, the current theory requires a notion of incoherence, as described in Chapter 2, but the wavelet bases which are commonly used in MRI are in fact quite coherent with Fourier measurements. Further, it is of interest to also consider the expansion of MRI images in overcomplete dictionaries (e.g., curvelets). While there are some recent results for CS with overcomplete dictionaries [31], these are RIP-based in a sense, and do not directly apply to the MRI setup.

In Chapter 3, we studied the ability of the LASSO to recover $X\beta$ and the support of β from the linear model $y = X\beta + z$ in a RIP-less setting. Under mild assumptions on the collinearity between columns of X, we once again demonstrated near-ideal error bounds, this time regarding the error in estimating the mean vector, $X\beta$. Further we demonstrated the perfect recovery of the support of β as long as all of its entries stood above the noise. As a consequence of this second result, one may achieve accurate recovery of β by regressing y onto its support. However, in the case when the entries of β do not all stand above the noise, our results do not apply to the recovery β. Demonstrating the accurate recovery of β—with oracle error bounds—under weaker conditions is still an important problem with many applications.

In Chapter 4, we turned to a RIP-based analysis of low-rank matrix recovery. We showed that the matrix version of the RIP holds at rank r with high probability for certain random measurement ensembles as long as the measurement operator provides at least $O(nr)$ measurements. This is

optimal up to a constant because the set of $n \times n$ matrices with rank r, has many linear subsets of dimension nr. Further, we used the RIP to give oracle error bounds for the matrix LASSO and the matrix Dantzig selector. In contrast to CS and SA results, the error bounds and number of measurements needed are within a constant factor of the oracle error, rather than a logarithmic factor.

While the results in this chapter were theoretically optimal up to constants, they are less applicable than the results in the other chapters due to the strong requirements on the measurement ensembles—the main ones considered were Gaussian or subgaussian. An important open problem is to weaken requirements on the random measurement ensembles. In particular, in the quantum state tomography problem discussed in this chapter, it is unknown whether the RIP holds.

In Chapter 5, we turned to the RIP-less matrix completion subproblem. We demonstrated stability of the matrix LASSO from about $nr \log^2 n$ measurements, but this time without near-optimal error bounds. While there is a growing literature about matrix completion [97,98,119] with exciting new results, it has not yet been proven that convex optimization provides near-ideal error bounds without restrictive assumptions. Such a result would be of great practical relevance due to the many applications of the matrix completion problem, and also of great theoretical interest due to its apparent difficulty.

Bibliography

[1] National climatic data center. http://www.ncdc.noaa.gov/oa/ncdc.html.

[2] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comp. System Sci.*, 66(4):671–687, Jun. 2003.

[3] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inf. Theory*, 48(3):569–579, 2002.

[4] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974. System identification and time-series analysis.

[5] Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. *Proc. Twenty-fourth Int. Conf. Mach. Learn.*, 2007.

[6] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *Neural Inf. Proc. Systems*, 2007.

[7] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *Proc. Thirty-Third Annual ACM Symp. on Theory of Computing*, pages 619–626. ACM New York, NY, USA, 2001.

[8] F. R. Bach. Consistency of trace norm minimization. *J. Mach. Learn. Research*, 9:1019–1048, 2008.

[9] B. Bah and J. Tanner. Improved bounds on restricted isometry constants for Gaussian matrices. 2010. Available at `http://arxiv.org/abs/1003.3299`.

[10] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Construct. Approx.*, 28(3):253–263, 2008.

[11] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theory*, 56(4):1982–2001, 2010.

[12] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113:301–413, 1999.

[13] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Analysis Mach. Intelligence*, 25(2):218–233, Feb. 2003.

[14] M. Bayati and A. Montanari. The LASSO risk for Gaussian matrices. 2010. Available at `http://arxiv.org/abs/1008.2581`.

[15] C. Beck and R. D'Andrea. Computational study and comparisons of LFT reducibility methods. In *Proc. American Control Conf.*, 1998.

[16] A. Beurling. Sur les intégrales de fourier absolument convergentes et leur application à une transformation functionelle. *Proc. Scandinavian Math. Congress*, 1938.

[17] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[18] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[19] P. Biswas, T-C. Lian, T-C. Wang, and Y. Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Trans. Sen. Netw.*, 2(2):188–220, 2006.

[20] J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Math.*, 162(1):73–141, 1989.

[21] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[22] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation and Sparsity Via ℓ_1 Penalized Least Squares. *Learning theory*, pages 379–391, 2006.

[23] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[24] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007.

[25] J. F. Cai, E. J. Candès, and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM J. Opt.*, 20:1956, 2010.

[26] T. T. Cai, L. Wang, and G. Xu. New bounds for restricted isometry constants. *IEEE Trans. Inf. Theory*, 56(9):4388–4394, 2010.

[27] E. Candès and J. Romberg. Practical signal recovery from random projections. *IEEE Trans. Sign. Proc.*, 2005.

[28] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris, Serie I*, 346:589–92, 2008.

[29] E. J. Candès and D. L. Donoho. Curvelets—a surprisingly effective nonadaptive representation for objects with edges. In C. Rabut A. Cohen and L. L. Schumaker, editors, *Curves and Surfaces*, pages 105–120, Vanderbilt University Press, 2000. Nashville, TN.

[30] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities. *Comm. Pure and Appl. Math.*, 57(2):219–266, 2004.

[31] E. J. Candès, Y. C. Eldar, and D. Needell. Compressed sensing with coherent and redundant dictionaries. 2010. Available at `http://arxiv.org/PS_cache/arxiv/pdf/1005/1005.2613v3.pdf`.

[32] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? 2009. Available at `http://statistics.stanford.edu/~ckirby/techreports/GEN/2009/2009-13.pdf`.

[33] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.*, 37:2145–2177, 2009.

[34] E. J. Candès and Y. Plan. Matrix completion with noise. *Proc. IEEE*, 98(6):925–936, 2010.

[35] E. J. Candès, Y. Plan, and J. A. Tropp, 2010. Working draft.

[36] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comp. Math*, 9(6):717–772, 2009.

[37] E. J. Candès and J. Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Comput. Math.*, 6(2):227–254, 2006.

[38] E. J. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Probl.*, 23(3):969–985, 2007.

[39] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[40] E. J. Candès, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math*, 59(8):1207, 2006.

[41] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.

[42] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[43] E. J. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[44] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory*, 56(5):2053–2080, 2010.

[45] B. Carl. Inequalities of Bernstein–Jackson type and the degree of compactness of operators in Banach spaces. *Ann. Inst. Fourier*, 35(3):79–118, 1985.

[46] P. Chen and D. Suter. Recovering the missing components in a large noisy low-rank matrix: application to SFM source. *IEEE Trans. Pattern Analysis Mach. Intelligence*, 26(8):1051–1063, 2004.

[47] S. S. Chen. *Basis pursuit*. PhD thesis, Stanford University, 1995.

[48] S. S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. on Sci. Comp.*, 20(1):33–61, 1998.

[49] B. Cheng and D. M. Titterington. Neural networks: a review from a statistical perspective. With comments and a rejoinder by the authors. *Stat. Sci.*, 9:2–54, 1994.

[50] A. L. Chistov and D. Yu. Grigoriev. Complexity of quantifier elimination in the theory of algebraically closed fields. In *Proceedings of the 11th Symposium on Mathematical Foundations of Computer Science*, volume 176 of *Lecture Notes in Computer Science*, pages 17–31. Springer Verlag, 1984.

[51] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22(1):211–231, 2009.

[52] W. Dai and O. Milenkovic. SET: an algorithm for consistent matrix completion. 2009. Available at `http://arxiv.org/abs/0909.2705`.

[53] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inf. Theory*, 55(5):2230–2249, 2009.

[54] G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. *J. Constr. Approx*, 13:57–98, 1997.

[55] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf.*, 52(4):1289–1306, 2006.

[56] D. L. Donoho. For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.*, 59(7):907–934, 2006.

[57] D. L. Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Disc. Comp. Geom.*, 35(4):617–652, 2006.

[58] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

[59] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

[60] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on*, 47(7):2845–2862, 2002.

[61] D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA*, 106(45):18914, 2009.

[62] D. L. Donoho, A. Maleki, and A. Montanari. The noise sensitivity phase transition in compressed sensing. 2010. Available at `http://arxiv.org/abs/1004.1218`.

[63] D. L. Donoho and J. L. Starck. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. 2007. Available at `http://www-stat.stanford.edu/~donoho/Reports/2006/StOMP-20060403.pdf`.

[64] D. L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA*, 102(27):9452, 2005.

[65] D. L. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *J. Amer. Math. Soc.*, 2006. To appear.

[66] D. L. Donoho and J. Tanner. Counting the faces of randomly-projected hypercubes and orthants, with applications. *Disc. Comp. Geom.*, 43(3):522–541, 2010.

[67] D.L. Donoho and M. Elad. Maximal sparsity representation via ℓ_1 minimization. *Proc. National Acad. Sci.*, 100(5):2197, 2003.

[68] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Magn.*, 25(2):83–91, 2008.

[69] M. Elad and A. M. Bruckstein. On sparse representations. In *International Conf. Image. Proc.*, 2001.

[70] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inf. Theory*, 48(9):2558–2567, 2002.

[71] J. Ellenberg. Fill in the blanks: Using math to turn lo-res datasets into hi-res samples, March 2010. Available at `http://www.wired.com/magazine/2010/02/ff_algorithm/all/1`.

[72] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[73] M. Fazel, E. J. Candès, B. Recht, and P. Parrilo. Compressed sensing and robust recovery of low rank matrices. In *42nd Asilomar Conference on Signals, Systems, and Computers*, pages 1043–1047. IEEE, 2009.

[74] M. Fazel, H. Hindi, and S. Boyd. Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. *Proc. Am. Control Conf*, June 2003.

[75] A. Feuer and A. Nemirovski. On sparse representation in pairs of bases. *IEEE Trans. Inf. Theoryn*, 49(6):1579–1581, 2003.

[76] A. K. Fletcher, S. Rangan, and V. K. Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Trans. Inf. Theory*, 55(12):5758–5772, 2009.

[77] D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *Ann. Statist.*, 22(4):1947–1975, 1994.

[78] S. Foucart and M. J. Lai. Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q \leq 1$. *Appl. Comput. Harmon. Anal.*, 26(3):395–407, 2009.

[79] J. J. Fuchs. More on sparse representations in arbitrary bases. *IEEE Trans. Inf. Theory*, 50(6):1341–1344, 2004.

[80] J. J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inf. Theory*, 50(6):1341–1344, 2004.

[81] A. Gilbert and P. Indyk. Sparse recovery using sparse matrices. *Proc. IEEE*, 98(6):937–947, 2010.

[82] ED Gluskin. On some finite-dimensional problems in the theory of widths. *Vestnik Leningrad Univ. Math*, 14:163–170, 1982.

[83] ED Gluskin. Norms of random matrices and widths of finite-dimensional sets. *Math. of the USSR-Sbornik*, 48:173, 1984.

[84] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Comm. ACM*, 35:61–70, 1992.

[85] E. Greenshtein. Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.*, 34(5):2367–2386, 2006.

[86] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

[87] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *Information Theory, IEEE Transactions on*, 49(12):3320–3325, 2004.

[88] D. Gross. Recovering low-rank matrices from few coefficients in any basis. 2009. Available at `http://arxiv.org/abs/0910.1879`.

[89] D. Gross. Recovering low-rank matrices from few coefficients in any basis, 2009. Available at `http://arxiv.org/PS_cache/arxiv/pdf/0910/0910.1879v4.pdf`.

[90] D. Gross, Y.K. Liu, S.T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Physical Review Letters*, 105(15):150401, 2010.

[91] O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Majorizing measures and proportional subsets of bounded orthonormal systems. 2008. Available at `http://arxiv.org/PS_cache/arxiv/pdf/0801/0801.3556v1.pdf`.

[92] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. Available at `http://arxiv.org/PS_cache/arxiv/pdf/0909/0909.4061v2.pdf`.

[93] J. Huang, S. Ma, and C.-H. Zhang. Adaptive lasso for sparse high-dimensional regression models. Technical report, University of Iowa, 2006.

[94] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank. Efficient and robust compressed sensing using optimized expander graphs. *IEEE Trans. Inf. Theory*, 55(9):4299–4308, 2009.

[95] B. S. Kashin. Diameters of certain finite-dimensional sets in classes of smooth functions. *Izv. Akad. Nauk SSSR, Ser. Mat*, 41(2):334–351, 1977.

[96] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inf. Theory*, 56(6):2980–2998, 2010.

[97] R. H. Keshavan, A. Montanari, and S. Oh. Matrix Completion from Noisy Entries. *J. Mach. Learn. Research*, 11:2057–2078, 2010.

[98] V. Koltchinskii, A. B. Tsybakov, and K. Lounici. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. 2010. Available at `http://arxiv.org/PS_cache/arxiv/pdf/1011/1011.6256v2.pdf`.

[99] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Stat.*, 28(5):1302–1338, 2000.

[100] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes.* Springer, 1991.

[101] K. Lee, Y. Bresler, H. Munthe-Kaas, A. Lundervold, S. Gaubert, M. Sharify, C. J. Cotter, M. Colombeau, M. Hutzenthaler, A. Jentzen, et al. Admira: Atomic decomposition for minimum rank approximation. 2009. Available at `http://arxiv.org/PS_cache/arxiv/pdf/0905/0905.0044v2.pdf`.

[102] E. L. Lehmann and G. Casella. *Theory of point estimation.* Springer Verlag, 1998.

[103] S. Levy and P. K. Fullagar. Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution. *Geophysics*, 46:1235, 1981.

[104] S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. Technical report, 2008.

[105] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, San Diego, Calif., 2nd edition, 1999.

[106] C. L. Mallows. Some comments on c_p. *Technometrics*, 15:661–676, 1973.

[107] C. McDiarmid. *Concentration. Probabilistic methods for algorithmic discrete mathematics.* Springer, 1998.

[108] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[109] N. Meinshausen and B. Yu. Lasso type recovery of sparse representations for high dimensional data. Technical report, University of California, 2006. Revised, August 2007.

[110] R. Meka, P. Jain, and I.S. Dhillon. Guaranteed rank minimization via singular value projection. 2009. Available at `http://arxiv.org/PS_cache/arxiv/pdf/0909/0909.5457v3.pdf`.

[111] M. Mesbahi and G. P. Papavassilopoulos. On the rank minimization problem over a positive semidefinite linear matrix inequality. *IEEE Trans. Automatic Control*, 42(2):239–243, 1997.

[112] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

[113] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comp. Harm. Anal.*, 26(3):301–321, 2009.

[114] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Comp. Math.*, 9(3):317–334, 2009.

[115] D. Needell and R. Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE J. Sel. Topics in Sign. Proc.*, 4(2):310–316, 2010.

[116] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. 2009. Available at `http://www.cs.utexas.edu/~pradeepr/paperz/highdimrates.pdf`.

[117] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Adv. in Neural Inf. Proc. Systems*, 2009.

[118] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. 2009. Submitted for publication and preprint available at `http://arxiv.org/PS_cache/arxiv/pdf/0912/0912.5100v1.pdf`.

[119] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Available at `http://arxiv.org/PS_cache/arxiv/pdf/1009/1009.2118v1.pdf`*, 2010.

[120] R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electronic J. Probab.*, 15:203–212, 2010.

[121] S. Oymak and B. Hassibi. New Null Space Results and Recovery Thresholds for Matrix Rank Minimization. *Available at `http://arxiv.org/abs/arXiv:1011.6326`*, 2010.

[122] A. Pajor and N. Tomczak-Jaegermann. Subspaces of small codimension of finite-dimensional Banach spaces. *Proc. Amer. Math. Soc.*, 97(4):637–642, 1986.

[123] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proc. 27th Annu. Asilomar Conf. Signals, Systems, and Computers*, volume 1, pages 40–44, Pacific Grove, CA, Nov. 1993.

[124] M. Raginsky, S. Jafarpour, Z. Harmany, R. Marcia, R. Willett, and R. Calderbank. Performance bounds for expander-based compressed sensing in Poisson noise. Available at `http://arxiv.org/abs/1007.2377`.

[125] H. Rauhut. Random sampling of sparse trigonometric polynomials. *Appl. Comput. Harmon. Anal.*, 22(1):16–42, 2007.

[126] H. Rauhut. Compressive sensing and structured random matrices. In *Theoretical Found. and Numer. Methods for Sparse Recovery*, 2010.

[127] H. Rauhut, J. Romberg, and J. A. Tropp. Restricted isometries for partial random circulant matrices. 2010. Available at `http://arxiv.org/abs/arXiv:1010.1847`.

[128] B. Recht. A simpler approach to matrix completion. 2009. Available at `http://arxiv.org/PS_cache/arxiv/pdf/0910/0910.0651v2.pdf`.

[129] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(471), 2010.

[130] B. Recht, W. Xu, and B. Hassibi. Null space conditions and thresholds for rank minimization. *Math. Programming*, 2010.

[131] A. Rohde and A. B. Tsybakov. Estimation of High-Dimensional Low-Rank Matrices. 2009. Available at `http://arxiv.org/PS_cache/arxiv/pdf/0912/0912.5338v2.pdf`.

[132] M. Rosenfeld. In praise of the Gram matrix. *Alg. and Combinat.*, 14:318–323, 1997.

[133] M. Rudelson. Almost orthogonal submatrices of an orthogonal matrix. *Israel J. Math.*, 111(1):143–155, 1999.

[134] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999.

[135] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Commun. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[136] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.*, 7(4):1307–1330, 1986.

[137] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comp.*, 7:1307, 1986.

[138] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Ant. and Prop.*, 34(3):276–280, 1986.

[139] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[140] A. Singer. A remark on global positioning from local distances. *Proc. Natl. Acad. Sci. USA*, 105(28):9507–9511, 2008.

[141] A. Singer and M. Cucuringu. Uniqueness of low-rank matrix completion by rigidity theory. Submitted for publication, 2009.

[142] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17:1329–1336, 2005.

[143] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. *Learning Theory*, pages 545–560, 2005.

[144] T. Strohmer and R. W. Heath. Grassmannian frames with applications to coding and communication. *Appl. Comp. Harm. Anal.*, 14(3):257–275, 2003.

[145] M. Talagrand. Majorizing measures: the generic chaining. *Ann. Prob.*, 24(3):1049–1103, 1996.

[146] H. L. Taylor, S. C. Banks, and J. F. McCoy. Deconvolution with the ℓ_1 norm. *Geophysics*, 44(1):39, 1979.

[147] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[148] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[149] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inf. Theory*, 50(10):2231–2242, 2004.

[150] J. A. Tropp. Recovery of short, complex linear combinations via ℓ_1 minimization. *IEEE Trans. Inf. Theory*, 51(4):1568–1570, 2005.

[151] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inf. Theory*, 52(3):1030–1051, 2006.

[152] J. A. Tropp. Norms of random submatrices and sparse approximation. *C. R. Acad. Sci. Paris*, 346(23-24):1271–1274, 2008.

[153] J. A. Tropp. On the conditioning of random subdictionaries. *Appl. Comp. Harm. Anal.*, 25(1):1–24, 2008.

[154] J. A. Tropp. On the linear independence of spikes and sines. *J. Fourier Anal. Appl.*, 14(5):838–858, 2008.

[155] J. A. Tropp. User-friendly tail bounds for matrix martingales. Available at `http://arxiv.org/abs/1004.4389`, 2010.

[156] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory*, 53(12):4655–4666, 2007.

[157] J. A. Tropp, J.N. Laska, M.F. Duarte, J.K. Romberg, and R.G. Baraniuk. Beyond Nyquist: Efficient sampling of sparse bandlimited signals. *IEEE Trans. Inf. Theory*, 56(1):520–544, 2009.

[158] S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.

[159] S. S. Vasanawala, M. T. Alley, B. A. Hargreaves, R. A. Barth, J. M. Pauly, and M. Lustig. Improved Pediatric MR Imaging with Compressed Sensing. *Radiology*, 256(2):607, 2010.

[160] R. Vershynin. On large random almost Euclidean bases. *Acta Mathematica Universitatis Comenianae*, 69(2):137–144, 2000.

[161] R. Vershynin. Math 280 lecture notes, 2007. Available at `http://www-personal.umich.edu/~romanv/teaching/2006-07/280/lec6.pdf`.

[162] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity, 2006. Available at `http://www.citebase.org/abstract?id=oai:arXiv.org:math/0605740`.

[163] M. J. Wainwright. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using. *IEEE Trans. Inf. Theory*, 55(5):2183, 2009.

[164] D. L. Wallace. Bounds on normal approximations to Student's and the chi-square distributions. *The Annals of Mathematical Statistics*, 30(4):1121–1130, 1959.

[165] C. C. Weng and P. P. Vaidyanathan. Matrix completion for DOA estimation, 2009. In preparation.

[166] P. Wojtaszczyk. Stability and instance optimality for Gaussian measurements in compressed sensing. *Found. Comput. Math.*, 10(1):1–13, 2009.

[167] T. Zhang. Some sharp performance bounds for least squares regression with ℓ_1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009.

[168] Y. Zhang. Theory of compressive sensing via ℓ_1 minimization: a non-rip analysis and extensions. Technical report, Rice University, 2008. Available at `http://www.caam.rice.edu/~zhang/reports/tr0811_revised.pdf`.

[169] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.