Structured Estimation in High-Dimensions

Sahand N Negahban

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-110
http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-110.html

May 11, 2012

Structured Estimation In High-Dimensions

by

Sahand N. Negahban

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

in

Engineering-Electrical Engineering & Computer Sciences

and the Designated Emphasis

in

Communication, Computation, and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martin J. Wainwright, Chair
Professor Peter Bickel
Professor Bin Yu

Spring 2012

Structured Estimation In High-Dimensions

Abstract

Structured Estimation In High-Dimensions

by

Sahand N. Negahban

Doctor of Philosophy in Engineering-Electrical Engineering & Computer Sciences

and the Designated Emphasis

in

Communication, Computation, and Statistics

University of California, Berkeley

Professor Martin J. Wainwright, Chair

High-dimensional statistical inference deals with models in which the number of parameters p is comparable to or larger than the sample size n. Since it is usually impossible to obtain consistent procedures unless $p/n \to 0$, a line of recent work has studied models with various types of low-dimensional structure, including sparse vectors, sparse and structured matrices, low-rank matrices, and combinations thereof. Such structure arises in problems found in compressed sensing, sparse graphical model estimation, and matrix completion. In such settings, a general approach to estimation is to solve a regularized optimization problem, which combines a loss function measuring how well the model fits the data with some regularization function that encourages the assumed structure. We will present a unified framework for establishing consistency and convergence rates for such regularized M-estimators under high-dimensional scaling. We will then show how this framework can be utilized to re-derive a few existing results and also to obtain a number of new results on consistency and convergence rates, in both ℓ_2-error and related norms.

An equally important consideration is the computational efficiency in performing inference in the high-dimensional setting. This high-dimensional structure precludes the usual global assumptions—namely, strong convexity and smoothness conditions—that underlie much of classical optimization analysis. We will discuss ties between the statistical inference problem itself and efficient computational methods for performing the estimation. In particular, we will show that the same underlying statistical structure can be exploited to prove global geometric convergence of the gradient descent procedure up to *statistical accuracy*. This analysis reveals interesting connections between statistical precision and computational efficiency in high-dimensional estimation.

Contents

List of Figures

Acknowledgments

At this time I am pleased to have the opportunity to acknowledge the vast community that has provided me support and encouragement throughout my early academic career and graduate school.

My adviser Martin J. Wainwright has been a tremendous force and influential figure for me. Working with Martin on high-dimensional statistics was awesome. From the beginnings of working on the perils of certain convex-relaxations to general methods for statistical recovery I have gained an incredible amount of experience and knowledge working with Martin. His energy carried over from the white board to the tables outside of the Northside food court where we would often discuss problems over fried rice, soup, and tea. Through my work with Martin I have also had the opportunity to collaborate with Bin Yu. I am very grateful and thankful for the opportunities I have had working with Bin. Our collaborations led to the work presented in Chapter 3 and the ideas appear throughout the subsequent chapters as well. The many discussions that we had regarding ways to improve a paper or her general ideas on statistics, optimization, and machine learning have all helped build upon my appreciation and understanding of statistics. I am also thankful to Peter Bickel, who served as an extremely helpful member of my qualifying exam committee and my dissertation committee. He always took the time to meet with me to discuss my dissertation and my qualifying examination. His insights into statistical inference and recovery has left an impact on my understanding of the field. Finally, I would like to thank Laurent El Ghaoui. He served on my qualifying examination and has provided a number of helpful insights regarding convex optimization for machine learning throughout my time as a graduate student. I also thank Michael Gastpar, David Tse, Anant Sahai, and Shankar Sastry for providing me with exceptional advice and guidance throughout my time as an undergraduate at Berkeley.

My years as a graduate student would not have been complete without my friends, colleagues, and collaborators. I must thank Alekh Agarwal for being an extremely sharp and enthusiastic coauthor and friend. Working with Alekh has taught me a tremendous amount about convex optimization algorithms. The last portion of this dissertation is based on joint work with Alekh and builds on the relationships between convex optimization and statistical inference. I also thank Pradeep Ravikumar for all of the discussions involved in developing much of the theory of regularized M-estimators, which are discussed throughout this dissertation. I thank John Duchi; our long discussions regarding optimization methods and our bike rides through the Berkeley Hills have all been tremendously helpful for me. My friend Dapo Omidiran deserves thanks for numerous astute and enlightening comments regarding compressed sensing, politics, and basketball. My discussions with Garvesh Raskutti helped me gain a deeper understanding of non-parametric regression as well as minimax optimality and I thank him for those chats. Additionally, I thank Galen Reeves for a number of helpful conversations providing me with a different perspective on compressed sensing. My other group members, Po-Ling Loh, Miles Lopes, and Nima Noorshams have also provided perceptive feedback regarding my talks and presentations and I thank them for their advice and

interest. I also thank Hao Zhang and Jiening Zhan for being great course project partners and friends. My work with Hao during a class project piqued my interest in problems regarding matrix completion and collaborative filtering. Pursuing a theoretical understanding of these ideas constitutes a significant portion of this dissertation and I credit him for a number of helpful conversations. I am grateful to Arash Ali Amini for helpful discussions regarding random matrix theory and for assistance with the formatting of this dissertation. I also thank Gireeja Ranade, Kristen (Kris) Woyach, Venkatesan (Venky) Ekambaram, and Pulkit Grover for the support they provided to Wireless Foundations (wifo). I thank the Wireless Foundations community: Guy Bresler, Naveen Goela, Nebojsa Milosavljevic, Bobak Nazer, Alex Dimakis, Hari Palaiyanur, Sae-Young Park, Changho Suh, I-Hsiang Wang, and also all of the other members for making wifo a pleasant community to be involved in and to work at. Furthermore, I would like to thank Amy Ng and Kim Kail for being so wonderful and helpful. I also owe much gratitude to the EECS staff for all for their support. I would like to thank those past and current members of the EECS Graduate division, Ruth Gjerde, Mary Byrnes, Dana Jantz, and Shirley Salanio, for their assistance throughout the years. I am extremely appreciative of all of my friends that have provided me with support, cupcakes, pizza, snowboarding trips, and inspiration throughout these years. I have been at Berkeley for almost a decade and before that spent my time in Irvine. I thank the vast number of students, teachers, faculty, and advisers that have provided me with advice, kindness, and support throughout the years.

I am thankful to my brother Makan and the wonderful Ms. Jenny Conway for their constant advice and encouragement during my last few years of graduate school. Finally, I thank my parents for their constant love and support; and for instilling in me a sense of passion and desire for pursuing my academic goals throughout my life.

Chapter 1

Introduction

"The goals of science and society, which statisticians share, are to draw useful information from data using everything that we know." [18] In this dissertation we will aim to understand how we may exploit the "low-dimensional" underlying structure of a high-dimensional estimation problem in order to obtain similar statistical and computational performance as the low-dimensional version. We will begin by first reviewing some ideas in statistical inference followed by a more concrete discussion of high-dimensional statistical estimation. After this discussion we will introduce some of the computational complexities involved in performing high-dimensional inference and consider how we might overcome these computational challenges.

1.1 Statistical Inference

One of the fundamental problems in statistics is that of *statistical estimation*, i.e. we wish to recover or extract information given a set of unorganized observations as *efficiently* as possible. In this thesis we will be addressing efficiency with respect to the number of observations required as well as the computational costs. Our goal will be to develop computationally tractable methods to efficiently exploit the structure underlying our data in order to extract the information. Statistical inference problems arise in a wide variety of settings including: astronomy [128], econometrics, epidemiology (for example John Snow's famous work), statistical signal processing [41, 45], medical image processing [85, 144], gene expression arrays [123], hand-written digits (Post-office data), social-network analysis (disease spread), and neuroscience [37]. For example, a common problem in statistical signal processing is for a receiver to estimate a transmitted signal corrupted by a noisy wireless channel. In the trivial case the signal is known by both the transmitter and receiver. In a more interesting setting the transmitter might send a signal from a set of possible signals. The receiver must then decode the transmission based on the noisy observations. Alternatively, in movie recommendations, we might be interested in learning the average sentiment for a movie by

analyzing the average rating. This piece of data is not immediately accessible without asking every person to rate the movie–a costly and time consuming task. Instead, we simply consider the sample of the population that has watched and rated the movie already in order to obtain an accurate estimate of the average score. The average movie rating can be useful to recommend a movie, however the single score does not account for variations in movie goer preferences. Hence, making accurate predictions for movie preferences can become more challenging. Instead, authors have shown that making use of all movies that a user has rated can greatly increase the recommendation accuracy [126]. That is, we wish to recover estimates for all possible pairs of movie and user ratings based on only a small fraction of rated films. More abstractly, we can consider the setting that we have n observations and we wish to estimate d parameters. The above stated example elicits a setting where we simply wish to understand a given population, and hence, we wish to extract pieces of information that describe the group. In general, such pieces of information can then be utilized for policy making, health-care decisions, or advertising.

A common challenge in the aforementioned instances of estimation is that the data can be non-uniform and exhibit randomness. That is, there is a level of noise or uncertainty in our observations. For example our estimate of the average rating can vary based on the sample of the population that we select; the same hand-written digits can appear differently; or signal measurements can be corrupted by thermal noise in the silicon sensors. Other challenges also exist when our desired parameters are not directly observable, for example learning if two individuals are friends based on their interactions on a social network, inferring an efficient representation of an image, or understanding what economic and political indicators are related to changes in stock prices.

In order to better assess and analyze these problems, we frequently rely on mathematical models that help us understand the objects in question. These models allow us to isolate the crucial components of a problem so that we can formulate a better understanding of the challenges involved. We refer the reader to the existing literature on methods of modeling in various statistical and engineering contexts [18, 136].

We may now present a more precise mathematical formulation for discussing the statistical estimation problem. In general we will suppose that we observe n observations $Z_1^n = \{Z_i\}_{i=1}^n \in \mathcal{Z}^n$. When we make statements regarding the probabilistic behavior of our methods, we will assume that the samples are drawn from a distribution $\mathbb{P}$ over the data space $\mathcal{Z}^n$. Furthermore, we assume that $\mathbb{P}$ can be well approximated by another distribution that lies in a family of probability measures $\mathcal{P}$ denoted as the *model* [18]. Each distribution $\mathbb{P} \in \mathcal{P}$ will be indexed by a set of parameters $\theta \in \Omega$, that is $\mathbb{P} \in \mathcal{P} := \{\mathbb{P}_\theta \mid \theta \in \Omega\}$. In order to help distinguish parameters, we will equip the space Ω with a metric $e : \Omega \times \Omega \to \mathbb{R}$ to compare two parameters and define the *loss function* $\mathcal{L} : \Omega \times \mathcal{Z}^n \to \mathbb{R}$ that will measure how well our choice of parameter fits the data. We will take $\theta^* \in \operatorname{argmin}_{\theta \in \Omega} \bar{\mathcal{L}}(\theta)$ to be any minimizer of the population risk $\bar{\mathcal{L}}(\theta) := \mathbb{E}_{Z_1^n}[\mathcal{L}(\theta; Z_1^n)]$. We wish to *infer* the parameter θ^* from the observations Z_1^n. Our estimate of θ^* will be denoted $\widehat{\theta} = \widehat{\theta}(Z_1^n)$ and is a function of

Z_1^n.

We will frequently wish to understand how the random object $e(\widehat{\theta}, \theta^*)$ behaves. For example, in classical asymptotic statistics one problem setting of interest is the case that Ω is a fixed set of d parameters, that is $\Omega \subset \mathbb{R}^d$, and we want to understand if $e(\widehat{\theta}, \theta^*)$ converges to zero in probability, expectation, or almost surely as the number of observations n goes to ∞. A more delicate analysis can further tell us the rate at which the above quantities converge to zero. In fact, under suitable regularity conditions, we can frequently show that the above quantities will converge with the rate $1/\sqrt{n}$ when the data are independently and identically distributed according to some product distribution $\mathbb{P}_{\theta^*}$. We will refer to the rate of convergence of our estimate to the true set of parameters as the *statistical performance* of our method. The book by van der Vaart [141] presents a more thorough discussion of classical asymptotic statistics and its vast applications to analyzing the performance of statistical estimators.

One question that we may wish to ask is: how useful is a theory that treats d as fixed whereas the number of observations n is taken to ∞? Put another way, will the asymptotic behavior be applicable to a specific problem with a fixed d and n when $d > n$? In the classical setting, this theory can be quite elucidating as we frequently have more observations than parameters, for example estimating the average movie rating of a population. However, in the modern statistical setting, we are pressed to carefully assess the validity of the above questions. With the advancements of modern data acquisition techniques we are faced with a plethora of data sources (the Internet), advanced scientific equipment that can produce massive amounts of data daily (LHC), and a requirement to analyze larger more complex problems with far fewer example as in gene expression array studies [81]. One of the most poignant instances of these challenges surfaces in the problem of collaborative filtering or matrix completion [124].

Consider the collaborative filtering problem of estimating how d_1 users will rate d_2 movies. Clearly, we will have $d = d_1 \times d_2$ parameters to estimate and can take $\Omega = \mathbb{R}^{d_1 \times d_2}$ to be the set of $d_1 \times d_2$ matrices. The classical statistical setting would require that we observe $n \gg d$ observations in order to make an accurate prediction of the ratings. In essence, we would require every user to watch and record a rating for every possible movie, which would be very impractical. In general, there is no way to overcome this. Intuitively, if we only observe a subset of the entries, then there are an infinite number of matrices that can fit the same data observations.

This hurdle is not unique to just the collaborative filtering problem. It also arises in linear regression, graphical model selection, and system identification. The underlying difficulty is owing to the fact that without sufficient observations, our loss function cannot generally distinguish between the various parameters, and is hence "flat" in some areas. A theme that will appear throughout this thesis is overcoming the "flatness" problem. However, unless we make further structural assumptions this problem is hopeless. This observation leads us to structured high-dimensional statistical estimation.

1.2 Structured high-dimensional statistical estimation

We will refer to high-dimensional estimation as instances of statistical estimation in which the ambient dimension of the data is comparable to (or possibly larger than) the sample size. From another perspective, we may also consider non-parametric inference problems as high-dimensional in nature. Such problems also suffer from the "curse of dimensionality" since even estimating a univariate function can be challenging without additional assumptions on the function class. Hence, a line of research has focused on exploiting certain sparsity and smoothness assumptions on the functions in order to efficiently estimate them [110]. We refer the interested reader to the work by Yu [151], which discusses some of the complexities and challenges with non-parametric regression. The primary focus of this paper will be on parametric models, and hence the complexity of our model classes will be related to the ambient dimensionality of the parameter space as well as the lower-dimensional underlying structure. This focus is not necessarily a weaker assumption: we may consider approximating a univariate function as the sum of polynomials up to an arbitrary order. Doing so allows us to approximate the non-parametric problem as a parametric one. The parameters in this case will be the coefficients in the polynomial expansion.

Problems with a high-dimensional character arise in a variety of applications in science and engineering, including analysis of gene array data, medical imaging, remote sensing, and astronomical data analysis. In settings where the number of parameters may be large relative to the sample size, the utility of classical (fixed dimension) results is questionable, and accordingly, a line of on-going statistical research seeks to obtain results that hold under high-dimensional scaling, meaning that both the problem size and sample size (as well as other problem parameters) may tend to infinity simultaneously. It is usually impossible to obtain consistent procedures in such settings without imposing some sort of additional constraints. Accordingly, there are now various lines of work on high-dimensional inference based on imposing different types of structural constraints. A substantial body of past work has focused on models with sparsity constraints, including the problem of sparse linear regression [131, 39, 44, 94, 20], banded or sparse covariance matrices [19, 16, 69], sparse inverse covariance matrices [153, 53, 120, 111], sparse eigenstructure [67, 4, 106], and sparse regression matrices [104, 83, 152, 63]. A theme common to much of this work is the use of ℓ_1-penalty as a surrogate function to enforce the sparsity constraint.

High-dimensional statistics is concerned with models in which the ambient dimension of the problem d may be of the same order as—or substantially larger than—the sample size n. Hence, we must consider new methods to model such problem instances. Classical work considers asymptotic analysis of such problems as $n \to \infty$ and $d \to \infty$. Our focus in this thesis will be on establishing error bounds with respect to some error metric that will hold with high probability for a finite number of observations n and demonstrate the dependency on d as well as other structural parameters.

As alluded to above, the roots of high-dimensional statistics are quite old, dating back to work on random matrix theory and high-dimensional testing problems (e.g, [55, 91, 105, 147]).

However, the past decade has witnessed a tremendous surge of research activity as the classical asymptotic assumptions prove no longer valid. Rapid development of data collection technology is a major driving force: it allows for more observations to be collected (larger n), and also for more variables to be measured (larger d).

In the regime $d \gg n$, it is well known that consistent estimators cannot be obtained unless additional constraints are imposed on the model. Accordingly, there are now several lines of work within high-dimensional statistics, all of which are based on imposing some type of low-dimensional constraint on the model space, and then studying the behavior of different estimators. Examples include linear regression with sparsity constraints, estimation of structured covariance or inverse covariance matrices, graphical model selection, sparse principal component analysis, low-rank matrix estimation, matrix decomposition problems, and estimation of sparse additive non-parametric models. The classical technique of regularization has proven fruitful in all of these contexts. Many well-known estimators are based on solving a convex optimization problem formed by the sum of a loss function (c.f. Section 1.1) with a weighted *regularizer*; we refer to any such method as a *regularized M-estimator*. The purpose of the regularizer is to encourage are estimates to satisfy our structural assumptions by penalizing deviations away from our assumptions. For instance, in application to linear models, the Lasso or basis pursuit approach [131, 39] is based on a combination of the least-squares loss with ℓ_1-regularization, and so involves solving a quadratic program. Similar approaches have been applied to generalized linear models, resulting in more general (non-quadratic) convex programs with ℓ_1-constraints. Several types of regularization have been used for estimating matrices, including standard ℓ_1-regularization, a wide range of sparse group-structured regularizers, as well as regularization based on the nuclear norm (sum of singular values).

Returning to the collaborative filtering example we see that any number of matrices could potentially fit our observations. The problem with such a model is that it necessarily treats all movies and users as completely unique. However, such an assumption is overly complex and does not model reality: various users (and movies) share similar characteristics, hence potentially reducing the number of effective parameters. Taking this point to the extreme, suppose that all movies and users are identical. In this case, rather than taking Ω to be the set of all matrices, we take it to be the set of all matrices M such that $M_{i,j} = c$ for some real number $c \in \mathbb{R}$. Hence, we have effectively reduced the number of parameters from $d_1\, d_2$ to one.

We have just presented an example where we have imposed an implicit structural constraint on our parameter set, thus reducing the effective size of the parameter space. In the chapters to follow we will observe various structural constraints that still admit effective modeling of our data. As will shall see in the sequel, enforcing such constraints can come at a cost; if the constraints do not model our data well, then we will be forced to pay a penalty in the statistical performance of our estimator.

Returning to the general model at hand, there are a large number of theoretical results in place for various types of regularized M-estimators that apply to various structural

constraints. Sparse linear regression has perhaps been the most active area, and multiple bodies of work can be differentiated by the error metric under consideration. They include work on exact recovery for noiseless observations (e.g., [46, 44, 31]), prediction error consistency (e.g., [56, 25, 138, 139, 155]), consistency of the parameter estimates in ℓ_2 or some other norm (e.g., [25, 26, 139, 155, 95, 20, 32]), as well as variable selection consistency (e.g., [94, 145, 156]). The information-theoretic limits of sparse linear regression are also well-understood, and ℓ_1-based methods are known to be optimal for ℓ_q-ball sparsity [109], and near-optimal for model selection [146]. For generalized linear models (GLMs), estimators based on ℓ_1-regularized maximum likelihood have also been studied, including results on risk consistency [140], consistency in ℓ_2 or ℓ_1-norm [10, 68, 92], and model selection consistency [113, 24]. Sparsity has also proven useful in application to different types of matrix estimation problems, among them banded and sparse covariance matrices (e.g., [19, 28, 69]). Another line of work has studied the problem of estimating Gaussian Markov random fields, or equivalently inverse covariance matrices with sparsity constraints. Here there are a range of results, including convergence rates in Frobenius, operator and other matrix norms [120, 114, 75, 158], as well as results on model selection consistency [114, 75, 94]. Motivated by applications in which sparsity arises in a structured manner, other researchers have proposed different types of block-structured regularizers (e.g., [135, 149, 137, 157, 152, 8, 11, 64]), among them the group Lasso based on ℓ_1/ℓ_2 regularization. High-dimensional consistency results have been obtained for exact recovery based on noiseless observations [129, 11], convergence rates in ℓ_2-norm (e.g., [96, 63, 83, 11]) as well as model selection consistency (e.g., [104, 97, 96]). Problems of low-rank matrix estimation also arise in numerous applications. Techniques based on nuclear norm regularization have been studied for different statistical models, including compressed sensing [117, 79], matrix completion [33, 71, 115, 99], multitask regression [154, 98, 119, 27, 9], and system identification [51, 98, 82]. Finally, although we primarily focus on high-dimensional parametric models in this thesis, regularization methods have also proven effective for a class of high-dimensional non-parametric models that have a sparse additive decomposition (e.g., [112, 93, 72, 73]), and shown to achieve minimax-optimal rates [110]. With this, the primary focus of the first half of the thesis will be in exploring the structural assumptions that arise in statistics and understanding how we may exploit them in order to obtain error-bounds that are applicable in the high-dimensional setting. However, it is not immediately clear that such high-dimensional problems lend themselves to computationally efficient solutions. In the second half of the thesis we will focus on methods for performing *computationally efficient* inference and explore the difficulties and solutions that arise.

1.3 Computational Considerations

Understanding the statistical behavior of a method is important in understanding its applicability to a problem domain. However, without acknowledging the computational aspects

of a problem, our solutions may prove unusable in practice. The regularized M-estimation techniques noted above are based on optimizing convex objectives. In principle, solutions can be found up to ϵ-accuracy in polynomial time by using interior point methods and other standard semi-definite program solvers [21, 15]. However, with the advent of high-dimensional problems, it has become increasingly clear that standard off-the-shelf or Newton-based approaches to convex optimization can be prohibitively expensive for the very large-scale problems that arise from high-dimensional data sets. Consequently, there has been a resurgence in research activity aimed at developing efficient first-order optimization based methods for large scale statistics and machine learning applications, e.g. projected gradient descent and mirror descent.

Several authors (e.g., [14, 66, 13]) have used variants of Nesterov's accelerated gradient method [102] to obtain algorithms for high-dimensional statistical problems with a sublinear rate of convergence. Note that an optimization algorithm, generating a sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$, is said to exhibit *sublinear convergence* to an optimum $\widehat{\theta}$ if the optimization error $\|\theta^t - \widehat{\theta}\|$ decays at the rate $1/t^{\kappa}$, for some exponent $\kappa > 0$ and norm $\|\cdot\|$. Although this type of convergence is quite slow, it is the best possible with gradient descent-type methods for convex programs under only Lipschitz conditions [101].

It is known that much faster global rates—in particular, a linear or geometric rate—can be achieved if global regularity conditions like strong convexity and smoothness are imposed [101]. An optimization algorithm is said to exhibit *linear or geometric* convergence if the optimization error $\|\theta^t - \widehat{\theta}\|$ decays at a rate κ^t, for some contraction coefficient $\kappa \in (0, 1)$. Note that such convergence is exponentially faster than sub-linear convergence. For certain classes of problems involving polyhedral constraints and global smoothness, Tseng and Luo [84] have established geometric convergence. However, a challenging aspect of statistical estimation in high dimensions is that the underlying optimization problems can never be strongly convex in a global sense when $d > n$ (since the $d \times d$ Hessian matrix is rank-deficient), and global smoothness conditions cannot hold when $d/n \to +\infty$. Some more recent work has exploited structure specific to the optimization problems that arise in statistical settings. For the special case of sparse linear regression with random isotropic designs (also referred to as compressed sensing), some authors have established fast convergence rates in a local sense, meaning guarantees that apply once the iterates are close enough to the optimum [22, 58]. The intuition underlying these results is that once an algorithm identifies the support set of the optimal solution, the problem is then effectively reduced to a lower-dimensional subspace, and thus fast convergence can be guaranteed in a local sense. Also in the setting of compressed sensing, Tropp and Gilbert [134] studied finite convergence of greedy algorithms based on thresholding techniques, and showed linear convergence up to a certain tolerance. For the same class of problems, Garg and Khandekar [54] showed that a thresholded gradient algorithm converges rapidly up to some tolerance. In both of these results, the convergence tolerance is of the order of the noise variance, and hence substantially larger than the true statistical precision of the problem.

In the second half of the thesis, we will focus on the convergence rate of two simple gradient-based algorithms for solving the convex programs that arise when using regularized M-estimators. Our goal will be to exploit the statistical structure that we used to obtain good statistical behavior in order to obtain good computational behavior. For a constrained problem with a differentiable objective function, the projected gradient method generates a sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$ by taking a step in the negative gradient direction, and then projecting the result onto the constraint set. The composite gradient method of Nesterov [102] is well-suited to solving regularized problems formed by the sum of a differentiable and (potentially) non-differentiable component. The main contribution of this paper is to establish a form of global geometric convergence for these algorithms that holds for a broad class of high-dimensional statistical problems. In order to provide intuition for this guarantee, Figure 1.1 shows the performance of projected gradient descent for a Lasso problem (ℓ_1-constrained least-squares). In panel (a), we have plotted the logarithm of the optimization error, measured in terms of the Euclidean norm $\|\theta^t - \widehat{\theta}\|$ between the current iterate θ^t and an optimal solution $\widehat{\theta}$, versus the iteration number t. The plot includes three different curves, corresponding to sparse regression problems in dimension $d \in \{5000, 10000, 20000\}$, and a fixed sample size $n = 2500$. Note that all curves are linear (on this logarithmic scale), revealing the geometric convergence predicted by our theory. Such convergence is not predicted by classical optimization theory, since the objective function cannot be strongly convex whenever $n < d$. Moreover, the convergence is geometric even at early iterations, and takes place to a precision far less than the noise level ($\nu^2 = 0.25$ in this example). We also note that the design matrix does not satisfy the restricted isometry property, as assumed in some past work.

The results in panel (a) exhibit an interesting property: the convergence rate is *dimension-dependent*, meaning that for a fixed sample size, projected gradient descent converges more slowly for a large problem than a smaller problem—compare the squares for $d = 20000$ to the diamonds for $d = 5000$. This phenomenon reflects the natural intuition that larger problems are, in some sense, "harder" than smaller problems. A notable aspect of our theory is that in addition to guaranteeing geometric convergence, it makes a quantitative prediction regarding the extent to which a larger problem is harder than a smaller one. In particular, our convergence rates suggest that if the sample size n is re-scaled in a certain way according to the dimension d and also other model parameters such as sparsity, then convergence rates should be roughly similar. Panel (b) provides a confirmation of this prediction: when the sample size is rescaled according to our theory (in particular, see Corollary 6.2 in Section 6.3.2), then all three curves lie essentially on top of another.

Although high-dimensional optimization problems are typically neither strongly convex nor smooth, this paper shows that it is fruitful to consider suitably restricted notions of strong convexity and smoothness. The notion of restricted strong convexity (RSC) is is related to but slightly different than that introduced in Chapter 3 for establishing statistical consistency. As we discuss in the sequel, bounding the optimization error introduces new

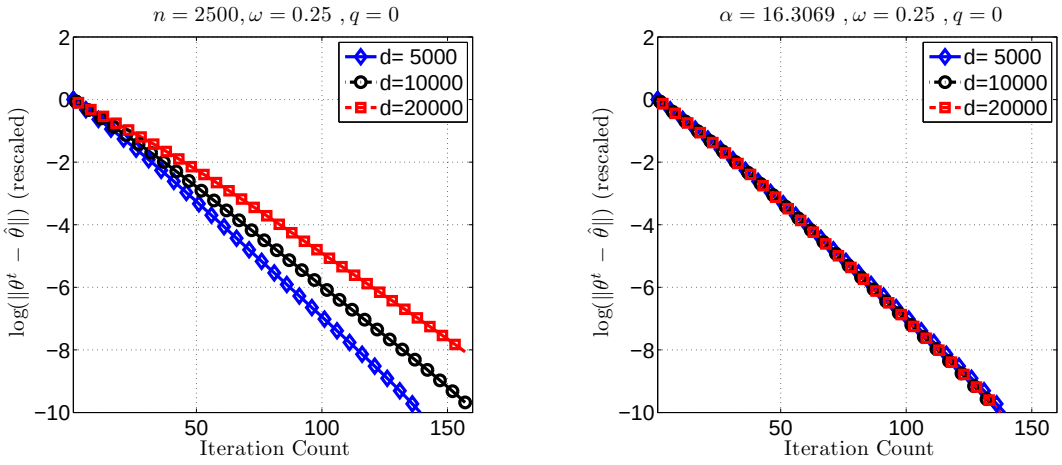


Figure 1.1. Convergence rates of projected gradient descent in application to Lasso programs (ℓ_1-constrained least-squares). Each panel shows the log optimization error $\log \|\theta^t - \widehat{\theta}\|$ versus the iteration number t. Panel (a) shows three curves, corresponding to dimensions $d \in \{5000, 10000, 20000\}$, sparsity $k = \lceil \sqrt{d} \rceil$, and all with the same sample size $n = 2500$. All cases show geometric convergence, but the rate for larger problems becomes progressively slower. (b) For an appropriately rescaled sample size ($\alpha = \frac{n}{k \log d}$), all three convergence rates should be roughly the same, as predicted by the theory.

challenges not present when analyzing the statistical error. We also introduce a related notion of restricted smoothness (RSM), not needed for proving statistical rates but essential in the setting of optimization. Our analysis consists of two parts. We first show that for optimization problems underlying many regularized M-estimators, appropriately modified notions of restricted strong convexity (RSC) and smoothness (RSM) are sufficient to guarantee global linear convergence of projected gradient descent. Our second contribution is to prove that for the iterates generated by our first-order method, these RSC/RSM assumptions do indeed hold with high probability for a broad class of statistical models, among them sparse linear models, models with group sparsity constraints, and various classes of matrix estimation problems, including matrix completion and matrix decomposition.

An interesting aspect of our results is that the global geometric convergence is not guaranteed to an arbitrary numerical precision, but only to an accuracy related to *statistical precision* of the problem. For a given error norm $\|\cdot\|$, given by the Euclidean or Frobenius norm for most examples in this paper, the statistical precision is given by the mean-squared error $\mathbb{E}[\|\widehat{\theta} - \theta^*\|^2]$ between the true parameter θ^* and the estimate $\widehat{\theta}$ obtained by solving the optimization problem, where the expectation is taken over randomness in the statistical model. Note that this is very natural from the statistical perspective, since it is the true parameter θ^* itself (as opposed to the solution $\widehat{\theta}$ of the M-estimator) that is of primary interest, and our analysis allows us to approach it as close as is statistically possible. Our analysis shows that

we can geometrically converge to a parameter θ such that $\|\theta - \theta^*\| = \|\widehat{\theta} - \theta^*\| + o(\|\widehat{\theta} - \theta^*\|)$, which is the best we can hope for statistically, ignoring lower order terms. Overall, our results reveal an interesting connection between the statistical and computational properties of M-estimators—that is, the properties of the underlying statistical model that make it favorable for estimation also render it more amenable to optimization procedures.

1.3.1 Organization of the thesis

The remainder of this dissertation is organized as follows. We will begin with an overview of the general background and present the notation used throughout this thesis in Chapter 2. We will then go into a discussion of general regularized M-estimators in Chapter 3 and demonstrate how we may exploit the regularizer to enforce our desired structural assumptions. We will then establish a few key properties that allow us to evaluate the statistical performance of a given estimator. Next, in Chapters 4 and 5 we will use the ideas from Chapter 3 to show how we may perform efficient low-rank matrix estimation for a variety of observation models: including those arising in system identification, matrix completion, and multi-task learning. Finally, in Chapter 6 we will discuss first-order gradient methods for solving the convex problems presented in Chapter 3.

Chapter 2

Background

This chapter highlights the mathematical concepts that we will employ throughout the text. These will include concepts in matrix analysis, probability theory, empirical process theory, and convex analysis. A number of theorems throughout this thesis make crucial use of the properties of convex functions and sets. Hence, the discussion presented in Section 2.2 is paramount to the development of the ideas in this thesis. Empirical process theory plays a second fundamental role in allowing us to make concrete probabilistic statements regarding the performance of our methods, and these statements will be presented in Section 2.3. Finally, underlying much of our presentation will be basic concepts in matrix analysis. We will begin by establishing a few theorems and setting down our notation for the remainder of our development.

2.1 Matrix analysis

This section presents some standard and fundamental notation from which we may build the rest of the thesis. We will denote a real d-dimensional vector as $\beta = (\beta_1, \beta_2, \ldots, \beta_d) \in \mathbb{R}^d$ where β_i is the i^{th} component of the vector. With a slight abuse of notation, we will take $e_i \in \mathbb{R}^d$ to be the standard basis vector where $(e_i)_i = 1$ and $(e_i)_j = 0$ for all $i \neq j$. Given two vectors $\beta, \zeta \in \mathbb{R}^d$ we define the inner product between the two vectors as

$$\langle \beta, \zeta \rangle := \sum_{i=1}^{d} \beta_i \zeta_i. \tag{2.1}$$

Furthermore, for any $q \in [1, \infty]$ we define the ℓ_q norm $\|\cdot\|_q$ as

$$\|\beta\|_q := \begin{cases} \left(\sum_{i=1}^{d} |\beta_i|^q \right)^{\frac{1}{q}} & q < \infty \\ \max_{1 \leq i \leq n} (|\beta_i|) & q = \infty. \end{cases} \tag{2.2}$$

We note that in the special case of $q = 2$ we have that the $\|\beta\|_2^2 = \langle \beta, \beta \rangle$. For a parameter $q \in [0, 1]$ and a radius $R_q > 0$, we may also define the ℓ_q "ball"

$$\mathbb{B}_q(R_q) := \left\{ \beta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\beta_j|^q \leq R_q \right\}.$$

We note that when $q \in [0, 1)$ the above sets are not convex in contrast to the ℓ_q balls defined as $\{\beta \mid \|\beta\|_q \leq R_q\}$ when $q \in [1, \infty]$. For $q = 0$ the ℓ_0-norm counts the total number of non-zero entries. Hence, any vector $\beta \in \mathbb{B}_0(R_0)$ is supported on a set of cardinality at most R_0 and we will denote the support of a vector β as $\text{supp}(\beta)$. We will denote a vector as "sparse" when $\|\beta\|_0 \ll d$. For $q \in (0, 1]$, membership in the set $\mathbb{B}_q(R_q)$ enforces a decay rate on the ordered coefficients, thereby modelling approximate sparsity. The parameter q will play a crucial role in our statistical error rates as the difficulty of the inference problem increases as the set of parameters become "less sparse."

Given the basic structure of vectors, we may now introduce matrices, which will serve as another useful parametric representation for a number of our models introduced in Chapters 4 and 5. We denote the set of real-valued $d_1 \times d_2$-dimensional matrices as $\mathbb{R}^{d_1 \times d_2}$. Given a matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$ we will denote the j^{th} column vector as Θ_j. Furthermore, we let the entry in the i^{th} row and j^{th} column be denoted $\Theta_{i,j}$. Such a parameter space can be thought of as a $\mathbb{R}^{d_1 d_2}$-dimensional vector space that is equipped with additional structure. In order to make that analogy concrete, we let the vectorized version of the matrix Θ to be $\text{vec}(\Theta) \in \mathbb{R}^{d_1 d_2}$. More precisely, we take

$$(\text{vec}(\Theta))_i := \Theta_{a(i),b(i)},$$

where $b(i) = \lfloor (i-1)/d_1 \rfloor$[1] $+ 1$ and $a(i) = i - (b(i) - 1)d_1$. With this basic structure in hand, we may now discuss particular operations that we can take on matrices.

The transpose of a matrix is the $d_2 \times d_1$-dimensional matrix Θ^T such that $(\Theta^T)_{i,j} = \Theta_{j,i}$. We will call a matrix symmetric when $\Theta^T = \Theta$. A matrix U will be referred to as orthogonal or orthonormal when $UU^T = I$, where I is the identity matrix. Given a matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$, we recall that it admits its singular value decomposition [60] as

$$\Theta = USV^T, \tag{2.3}$$

where $U \in \mathbb{R}^{d_1 \times d_1}$ and $V \in \mathbb{R}^{d_2 \times d_2}$ are both orthogonal and $S \in \mathbb{R}^{d_1 \times d_2}$ is diagonal. The entries $S_{i,i}$ are called the *singular values* of the matrix Θ and by definition are positive. We will frequently denote them as $\sigma_i(\Theta) = S_{i,i}$. The column vectors U_i and V_i are the respective left and right singular vectors corresponding to the i^{th} singular value. If we let $m = \min(d_1, d_2)$ and $r = \text{rank}(\Theta) \leq m$ then the matrix Θ has at most r non-zero singular values while the remaining $m - r$ singular values are zero. We call a matrix low-rank

[1] The operator $\lfloor x \rfloor$ is the largest integer less than or equal to x.

when $r \ll \min(d_1, d_2)$. Furthermore, we will assume that the singular values are sorted in decreasing order so that

$$\sigma_1(\Theta) \geq \sigma_2(\Theta) \geq \cdots \geq \sigma_r(\Theta) \geq 0 = \sigma_{r+1}(\Theta) = \cdots = \sigma_m(\Theta).$$

Given the singular value decomposition we may define the nuclear or trace norm as

$$|||\Theta|||_{\text{nuc}} := \sum_{i=1}^{r} \sigma_i(\Theta). \tag{2.4}$$

The nuclear norm will appear in Chapters 4 and 5 as the primary regularizer used to encourage our parameter estimates to be low-rank. We also introduce the operator norm of a matrix as

$$|||\Theta|||_2 := \max_{1 \leq i \leq r}(\sigma_i(\Theta)) \tag{2.5}$$

and the Frobenius norm as

$$|||\Theta|||_F := (\sum_{i=1}^{r} \sigma_i^2(\Theta))^{1/2}. \tag{2.6}$$

These norms have a natural analog with the ℓ_p-norms introduced in equation (2.2)–they can be viewed as ℓ_p norms taken on the vector of singular values. Such norms are referred to as Schatten-p norms and the Frobenius norm is the analog of the ℓ_2 norm while the operator norm is the analog of the ℓ_∞ norm and the analog of the nuclear norm is the ℓ_1 norm. More generally, we may define the Schatten-p norm as

$$|||\Theta|||_p := (\sum_{i=1}^{r} \sigma_i^q(\Theta))^{1/p}.$$

Analogous to the ℓ_q "balls" that we defined for $q \in [0, 1]$, we may also define

$$\mathbb{M}_q(R_q) := \left\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \sum_{j=1}^{r} \sigma_i(\Theta)^q \leq R_q\right\}. \tag{2.7}$$

As with the ball $\mathbb{B}_0(R_q)$ defined for vectors, $\mathbb{M}_0(R_q)$ denotes the set of all matrices with rank at most R_q. In general, the above quantity will allow us to discuss approximate low-rankedness as it will impose a decay rate on the singular values of the matrix $\Theta \in \mathbb{M}_q$.

Now, consider two matrices $X, Y \in \mathbb{R}^{d_1 \times d_2}$. Recalling that the trace of a square matrix $M \in \mathbb{R}^{d_1 \times d_1}$ is $\text{trace}(M) = \sum_{i=1}^{d_1} M_{i,i}$ we define the inner product between the matrices X and Y as

$$\langle\langle X, Y \rangle\rangle := \text{trace}(XY^T). \tag{2.8}$$

The above inner product is known as the trace inner product and forms the natural analog to the inner product defined above for vectors. A simple exercise shows that $\text{trace}(XY^T) = \langle \text{vec}(X), \text{vec}(Y) \rangle$ and the Euclidean norm induced by the trace inner product is equal to the Frobenius norm. This fact can be easily verified by noting that the singular value decomposition of the matrix $X = USV^T$ so that

$$\begin{aligned} \text{trace}(XX^T) &= \text{trace}(USV^TVS^TU^T) \\ &= \text{trace}(USS^TU^T), \end{aligned}$$

where the second inequality follows from the fact that $V^TV = I$. Finally, an elementary inequality yields that

$$\begin{aligned} \text{trace}(USS^TU^T) &= \text{trace}(SS^TU^TU) \\ &= \text{trace}(SS^T) \\ &= \sum_{i=1}^{r} \sigma_i^2(X). \end{aligned}$$

Therefore, $\langle\langle X,\ X \rangle\rangle = |||X|||_F^2$, which establishes our desired result.

Given a norm $\|\cdot\|$, we may define the dual norm $\|\cdot\|_*$ as

$$\|v\|_* := \sup_{\|u\| \leq 1} \langle u,\, v \rangle. \tag{2.9}$$

We assume that for matrices, the above inner product is taken as the trace inner product defined in equation (2.8). For any fixed $p \in [1, \infty]$, the dual to the ℓ_p norm is the $\ell_{p'}$ norm, where p' is the *Hölder conjugate* to p and satisfies $\frac{1}{p} + \frac{1}{p'} = 1$. For example the dual norm to the ℓ_2 norm is again the ℓ_2 norm. The dual norm to the ℓ_1 norm is the ℓ_∞ norm. Additionally, the dual norm to the Schatten-p norm is the Schatten-p' norm. For example, the dual norm to the Frobenius norm is again the Frobenius norm, while the dual norm to the nuclear norm is the operator norm. Finally, by equation (2.9) we have that

$$\langle u,\, v \rangle \leq \|u\|\, \|v\|_*. \tag{2.10}$$

As a specific instance the above statement we have Hölder's inequality, which states that

$$\langle u,\, v \rangle \leq \|u\|_p\, \|v\|_{p'}$$

for $p \in [1, \infty]$ and its Hölder conjugate p'. When $p = 2$ and $p' = 2$, the above inequality is known as the *Cauchy-Schwarz inequality*.

Finally, for $v \in \mathbb{R}^d$ and $1 \leq p \leq q \leq \infty$ we have the following chain of inequalities,

$$\frac{1}{d^{1/p - 1/q}} \|v\|_p \ \leq\ \|v\|_q \ \leq\ \|v\|_p. \tag{2.11}$$

The above inequalities are tight. The inequality on the left achieves equality if we take v to be the all ones vector and the inequality on the right achieves equality when we take $v = e_i$. However, the above inequalities are not tight when we consider restricted examples of vectors. For example, if $v \in \mathbb{R}^d$ is k-sparse, i.e. v has k non-zero entries, then

$$\|v\|_1 \leq \sqrt{k}\, \|v\|_2. \tag{2.12}$$

Indeed, for k-sparse vectors we may replace d in the above inequalities with k. We also note that analogous bounds hold for the Schatten-p norms. We will make use of such inequalities in establishing our error bounds in order to compare two norms.

The first inequality in equation (2.11) follows by an immediate application of Jensen's inequality [48]. In order to establish the second inequality we first assume that without loss of generality that $\|v\|_q = 1$ and $v_i \geq 0$. Therefore,

$$\|v\|_q \;=\; \sum_{i=1}^{d} v_i^q \;=\; 1$$

so that

$$(\sum_{i=1}^{d} v_i^q)^{1/p} = 1.$$

Now, by the assumption that $\|v\|_q = 1$ and $v_i \geq 0$, we immediately have that $0 \leq v_i \leq 1$. Thus, $v_i^q \leq v_i^p$ since $q \geq p$ and $v_i^x \leq v_i$ whenever $0 \leq v_i \leq 1$ and $x \geq 1$. Hence,

$$\sum_{i=1}^{d} v_i^q \leq \sum_{i=1}^{d} v_i^p.$$

Now the function $x \mapsto x^{1/p}$ is monotonic so that

$$(\sum_{i=1}^{d} v_i^q)^{1/p} \leq (\sum_{i=1}^{d} v_i^p)^{1/p},$$

thus establishing that $\|v\|_p \geq 1$.

Given these fundamental ideas we recall that our inference techniques will be based on solving regularized M-estimators. We will be restricting our attention to the setting that such estimators are convex. Hence, in the next section we will present some of the necessary background in convex analysis that will allow us to analyze the statistical and computational properties of our M-estimators.

2.2 Convex Analysis

Throughout our later developments we will take our estimators to be convex. That is, we will obtain our estimate of θ^* by solving

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta} \mathcal{L}(\theta),$$

for some loss function $\mathcal{L}$. Convex functions have been used throughout the statistics and machine learning [12]. Convex functions have a number of favorable properties. For instance, a convex function has a single global optima and there has been a vast amount of literature committed to developing efficient algorithms for solving convex optimization problems [101, 15, 21]. Additionally, in Chapter 6 we discuss favorable computationally properties of large-scale convex optimization procedures used for solving M-estimators. To that end, this section will focus on the background in convex analysis that will be applied in later developments.

2.2.1 Convex Regularized M-estimators

The inference algorithms that we discuss throughout this thesis are based on optimizing over a loss function $\mathcal{L}(\theta)$ plus a regularizer $\mathcal{R}(\theta)$ in order to obtain an estimate $\widehat{\theta}$ of θ^*. That is, we will focus on methods such that

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta} \mathcal{L}(\theta) + \mathcal{R}(\theta)$$

and that the loss function and regularizers are both *convex*. A function $f : \Omega \mapsto \mathbb{R}$ is convex if for any $v, w \in \Omega$ and for any $\alpha \in [0, 1]$

$$f(\alpha v + (1 - \alpha)w) \leq \alpha f(v) + (1 - \alpha) f(w).$$

For example, given $X \in \mathbb{R}^{n \times d}$ with rows $x_i \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$, the function

$$\begin{aligned} \mathcal{L}(\beta) &= \sum_{i=1}^{n} (\langle x_i, \beta \rangle - y_i)^2 \\ &= \|X\beta - y\|_2^2 \end{aligned}$$

is a convex function and the function $\mathcal{L}(\beta)$ is our first example of a loss function. Namely, given observations of the form $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, we wish to find a $\beta \in \mathbb{R}^d$ such that $\langle x_i, \beta \rangle$ is a good approximate of y_i. Hence, for each observation (x_i, y_i) we penalize the choice of β by squaring the error. Thus, our M-estimator in this setting will be based on minimizing the squared error over $\beta \in \Omega$. Our goal will be to understand the statistical properties of an optimal solution[2] $\widehat{\theta}$ as well as the computational complexity in obtaining such an estimate.

[2]We say *an* optimal solution rather than *the* optimal solution as we do not assume there is a unique solution. Instead, our theory will establish desirable statistical properties for all possible solutions to the inference problem.

In subsequent developments we will analyze regularized M-estimators $\mathcal{L}(\beta) + \mathcal{R}(\beta)$ for a suitably chosen regularizer $\mathcal{R}$ that will encourage the estimate to satisfy specified structural assumptions, such as sparsity. We will continue to see more examples of convex M-estimators Chapter 3. In the later chapters, we will make use of an equivalent definition of convexity in the setting that f is differentiable at a point w so that the gradient $\nabla f(w)$ exists, then a function f is convex if and only if

$$f(v) - f(w) - \langle \nabla f(w), v - w \rangle \geq 0,$$

for all $v, w \in \Omega$ [118, 21]. As a consequence of the above inequality we see that if $\nabla f(w) = 0$, then $f(w) \leq f(v)$ for all v, thus establishing that w is a global optimum of f. In general, if f is convex but not differentiable at a point w, we may still find a set of vectors g such that

$$f(v) \geq f(w) + \langle g, v - w \rangle.$$

Such vectors g are called the *subradients* of f, and we denote the set of all subgradients at a point w as $\partial f(w)$. If f is differentiable at the point w, then the set $\partial f(w)$ contains the single point $\nabla f(w)$ [118].

Proposition 2.1 (Theorem 3.1.15 [101])**.** *Suppose that we solve the convex program* $\min_{v \in \mathbb{R}^d} f(v)$. *Then a point* $w^* \in \mathbb{R}^d$ *is an optimal solution to the convex program if and only if* $0 \in \partial f(w^*)$.

We now introduce the definition of *strong-convexity*, which will be a crucial idea throughout much of the chapters.

Definition 2.1 (**Strong convexity**)**.** A function f is strongly convex with parameter $\mu \geq 0$ over the set Ω if

$$f(v) - f(w) - \langle \nabla f(w), v - w \rangle \geq \frac{\mu}{2} \|v - w\|_2^2 \qquad \text{for all } v, w \in \Omega. \tag{2.13}$$

Strong-convexity immediately implies standard convexity and imposed a lower-curvature condition on the function f. Furthermore, we may use strong-convexity in order to establish error bounds on our parameters. Suppose that we have a loss-function $\mathcal{L}(\beta)$ that we wish to use in order to estimate some parameter θ^*. If we let $\widehat{\theta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \mathcal{L}(\beta)$ (where we assume that $\mathcal{R}$, then by strong convexity we know that

$$\mathcal{L}(\widehat{\theta}) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), \widehat{\theta} - \theta^* \rangle \geq \frac{\mu}{2} \|\widehat{\theta} - \theta^*\|_2^2.$$

Furthermore, since $\widehat{\theta}$ is an optimal solution of the optimization we know that $\mathcal{L}(\widehat{\theta}) - \mathcal{L}(\theta^*) \leq 0$, so that

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{2}{\mu} \langle \nabla \mathcal{L}(\theta^*), \widehat{\theta} - \theta^* \rangle. \tag{2.14}$$

Furthermore, by the Hölder's inequality,

$$-\langle \nabla \mathcal{L}(\theta^*), \widehat{\theta} - \theta^* \rangle \leq \|\nabla \mathcal{L}(\theta^*)\|_{p'} \|\widehat{\theta} - \theta^*\|_p$$

where $p \in [1, \infty]$ and p' is its H—older conjugate. Now, for $p \in [1, 2]$ we have by equation (2.11)

$$\|\widehat{\theta} - \theta^*\|_p \leq d^{1/p - 1/2} \|\widehat{\theta} - \theta^*\|_2.$$

Therefore, combining the last two equations with equation (2.14)

$$\|\widehat{\theta} - \theta^*\| \leq \frac{2}{\mu} d^{1/p - 1/2} \|\nabla \mathcal{L}(\theta^*)\|_{p'}.$$

Hence, we may establish error bounds by analyzing the behavior of $\|\nabla \mathcal{L}(\theta^*)\|_{p'}$. This form of analysis will be generalized in order to help establish error bounds in the high-dimensional setting since, as alluded to, strong-convexity does not hold for a number of statistical problem in the high-dimensional setting. Returning to the example presented with $\mathcal{L}(\beta) = \|X\beta - y\|_2^2$ then strong convexity requires that

$$\begin{aligned} \|X\beta - y\|_2^2 - \|X\beta' - y\|_2^2 - \langle X^T(X\beta' - y), \beta - \beta' \rangle = \|X(\beta - \beta')\|_2^2 \\ \geq \frac{\mu}{2} \|\beta - \beta'\|. \end{aligned}$$

Hence, if $n \leq d$ the matrix X will have a non-trivial nullspace, so there exists a v such that $Xv = 0$. Therefore, we may find a β and β' such that $X(\beta - \beta') = 0$, which implies that strong convexity cannot hold. In the sequel we will see that we may employ regularized M-estimators in order to guarantee that the error $\widehat{\theta} - \theta^*$ will not be in the kernel of X.

Next, we define *smoothness*, which serves to provide upper-curvature control on a function.

Definition 2.2 (Smoothness). A function f is smooth with parameter $\gamma \geq 0$ over the set Ω if

$$f(v) - f(w) - \langle \nabla f(w), v - w \rangle \leq \frac{\gamma}{2} \|v - w\|_2^2 \qquad \text{for all } v, w \in \Omega. \tag{2.15}$$

Smoothness will come to play an important role in Chapter 6. However, we will see in the sequel that an altered version of smoothness must be applied in the high-dimensional setting.

The loss functions and regularized M-estimators that we introduce throughout will be examples of convex functions. We have introduced a few of the topics that will prove useful throughout our later discussions and our overview of convex analysis has been necessarily brief. We refer the reader to the existing literature [118] for a more thorough discussion. Another key idea that we must discuss is that our available data is necessarily random. Hence, we only have empirical quantities available to us, such as the empirical loss, while in the ideal setting we would like access to the *population* versions of these empirical quantities. Concentration inequalities provide us with a way to concretely discuss the relationships between empirical quantities and their population counterparts.

2.3 Concentration Inequalities

Suppose that we are given some function f of a random variable X; we wish to understand the deviations of $f(X)$ around its expected value $\mathbb{E}f(X)$. More concretely, we aim to categorize the behavior of

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t)$$

for all $t \geq 0$. For example, if $f(x) = x$ and X is a normal random variable with mean μ and variance σ^2 so that $X \sim N(\mu, \sigma^2)$, then

$$\mathbb{P}(|X - \mu| \geq t) \ \leq 2\exp(-\frac{t^2}{2\sigma^2}). \tag{2.16}$$

We may establish the above bound by recalling Chernoff's inequality

$$\mathbb{P}(X - \mu \geq t) \ \leq \inf_{\lambda > 0} \mathbb{E}\exp(\lambda(X - \mu))\exp(-\lambda t).$$

Furthermore, a simple calculation yields that $\mathbb{E}\exp(\lambda(X - \mu)) = \exp((\sigma^2\lambda^2)/2)$, so that

$$\mathbb{P}(X - \mu \geq t) \ \leq \inf_{\lambda > 0} \exp(\frac{\sigma^2\lambda^2}{2})\exp(-\lambda t).$$

Setting $\lambda = \frac{t}{\sigma^2}$ minimizes the above bound yielding

$$\mathbb{P}(X - \mu \geq t) \ \leq \exp(-\frac{t^2}{2\sigma^2}). \tag{2.17}$$

A similar calculation shows that

$$\mathbb{P}(X - \mu \leq -t) \ \leq \exp(-\frac{t^2}{2\sigma^2}),$$

which then implies the inequality (2.16) by an application of the union bound[3].

The above derivation simply used the fact that $\mathbb{E}\exp(\lambda(X-\mu)) \leq \exp((\sigma^2\lambda^2)/2)$. Hence, we may generalize the above derivation to any random variable that satisfies the latter inequality. Thus, we have the following definition

Definition 2.3. A random variable X with mean μ is *sub-Gaussian* with parameter σ^2 if

$$\mathbb{E}\exp(\lambda(X - \mu)) \leq \exp((\sigma^2\lambda^2)/2)$$

for all $\lambda \in \mathbb{R}$.

[3]Recall that the union bound states that for a finite set of events $\{A_i\}_{i=1}^n$ we have that $\mathbb{P}(\bigcup_{i=1}^n A_i) \leq n \max_i \mathbb{P}(A_i)$.

Therefore, any sub-Gaussian random variable with parameter σ^2 satisfies inequality (2.16). Given the above definition we may now present the next proposition which states that the sum of independent sub-Gaussian random variables is still sub-Gaussian.

Proposition 2.2. *Consider a collection of n independent sub-Gaussian random-variables $\{X_i\}_{i=1}^n$ each with sub-Gaussian parameter σ_i^2 and mean μ_i. Then for all $t \geq 0$*

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mu_i) \geq t\right] \leq \exp(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}).$$

The proof of the above result follows from the fact that

$$\begin{aligned}\mathbb{E}[\exp(\lambda \sum_{i=1}^n (X_i - \mu_i))] &= \prod_{i=1}^n \mathbb{E}[\exp(\lambda(X_i - \mu_i))] \\ &\leq \prod_{i=1}^n \exp(\lambda^2 \sigma_i^2/2) \\ &= \exp(\lambda^2 \sum_{i=1}^n \sigma_i^2/2).\end{aligned}$$

The first inequality follows by independence and the second inequality follows from the definition of sub-Gaussianity. Thus, the sum $\sum_{i=1}^n (X_i - \mu_i)$ is mean zero with sub-Gaussian parameter $\sum_{i=1}^n \sigma_i^2$. Therefore establishing our desired result after appealing to equation (2.17).

The next proposition is classical [78, 87] and yields sharp concentration of a Lipschitz function of Gaussian random variables around its mean.

Proposition 2.3. *Let $X \in \mathbb{R}^n$ have i.i.d. $N(0,1)$ entries, and let and $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz with constant L (i.e., $|f(x) - f(y)| \leq L\|x - y\|_2 \ \forall x, y \in \mathbb{R}^n$). Then for all $t > 0$, we have*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| > t) \leq 2\exp\big(-\frac{t^2}{2L^2}\big). \tag{2.18}$$

We now present another classical result [77] on the concentration of functions with bounded differences. This proposition will allow us to establish that a function of a set of random variables $f(X_1, X_2, \ldots, X_n)$ concentrates around its mean when

$$|f(X_1, X_2, \ldots, X_i, \ldots, X_n) - f(X_1, X_2, \ldots, Y_i, \ldots, X_n)| \leq c_i. \tag{2.19}$$

That is, the function can change by at most c_i when we vary the i^{th} coordinate in the function. The above condition is called the bounded differences property.

Proposition 2.4. *Suppose the collection of random variables $\{X_i\}_{i=1}^n$ are independent and that the function f satisfies equation (2.19) with parameters $(c_1, \ldots, c_n)$. Then*

$$\mathbb{P}(|f(X_1, X_2, \ldots, X_n) - \mathbb{E}f(X_1, X_2, \ldots, X_n)| \geq t) \leq 2\exp\Big(-\frac{t^2}{2\sum_{i=1}^n c_i^2}\Big).$$

Chapter 3

Regularized M-estimators

3.1 Introduction

There has been a tremendous amount of work in analyzing various types of regularized M-estimators, with the choice of loss function, regularizer and statistical assumptions changing according to the model. This methodological similarity suggests an intriguing possibility: is there a *common set of theoretical principles* that underlies analysis of all these estimators? If so, it could be possible to gain a unified understanding of a large collection of techniques for high-dimensional estimation, and afford some insight into the literature.

The main contribution of this chapter is to provide an affirmative answer to this question. In particular, we isolate and highlight two key properties of a regularized M-estimator—namely, a *decomposability property* for the regularizer, and a notion of *restricted strong convexity* that depends on the interaction between the regularizer and the loss function. For loss functions and regularizers satisfying these two conditions, we prove a general result (Theorem 3.1) about consistency and convergence rates for the associated estimators. This result provides a family of bounds indexed by subspaces, and each bound consists of the sum of approximation error and estimation error. This general result, when specialized to different statistical models, yields in a direct manner a large number of corollaries, some of them known and others novel. This framework can be applied to prove several results on low-rank matrix estimation using the nuclear norm, that we discuss in more detail in Chapter 4, as well as minimax-optimal rates for noisy matrix completion, discussed in Chapter 5, and noisy matrix decomposition [2]. Finally, en route to establishing these corollaries, we also prove some new technical results that are of independent interest, including guarantees of restricted strong convexity for group-structured regularization (Proposition 3.1). These ideas will then be later exploited in Chapter 6 in order to obtain computational gains.

The remainder of this chapter is organized as follows. We begin in Section 3.2 by formulating the class of regularized M-estimators that we consider, and then defining the notions of decomposability and restricted strong convexity. Section 3.3 is devoted to the statement

of our main result (Theorem 3.1), and discussion of its consequences. Subsequent sections are devoted to corollaries of this main result for different statistical models, including sparse linear regression (Section 3.4) and estimators based on group-structured regularizers (Section 3.5).

3.2 Problem formulation and some key properties

In this section, we begin with a precise formulation of the problem, and then develop some key properties of the regularizer and loss function.

3.2.1 A family of M-estimators

Let $Z_1^n := \{Z_1, \ldots, Z_n\}$ denote n observations with marginal distribution $\mathbb{P}$. Recall that we are interested in estimating some parameter θ of the distribution $\mathbb{P}$. Let $\mathcal{L} : \Omega \times \mathcal{Z}^n \to \mathbb{R}$ be a convex and differentiable loss function that, for a given set of observations Z_1^n, assigns a cost $\mathcal{L}(\theta; Z_1^n)$ to any parameter $\theta \in \mathbb{R}^d$. Take $\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \bar{\mathcal{L}}(\theta)$ be any minimizer of the population risk $\bar{\mathcal{L}}(\theta) := \mathbb{E}_{Z_1^n}[\mathcal{L}(\theta; Z_1^n)]$. In order to estimate this quantity based on the data Z_1^n, we solve the convex optimization problem

$$\widehat{\theta}_{\lambda_n} \quad \in \quad \arg\min_{\theta \in \mathbb{R}^d} \big\{ \mathcal{L}(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta) \big\}, \tag{3.1}$$

where $\lambda_n > 0$ is a user-defined regularization penalty, and $\mathcal{R} : \Omega \to \mathbb{R}_+$ is a norm. Note that this set-up allows for the possibility of mis-specified models as well.

Our goal is to provide general techniques for deriving bounds on the difference between any solution $\widehat{\theta}_{\lambda_n}$ to the convex program (3.1) and the unknown vector θ^*. In this chapter, we derive bounds on the quantity $\|\widehat{\theta}_{\lambda_n} - \theta^*\|$, where the error norm $\|\cdot\|$ is induced by some inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^d$. Most often, this error norm will either be the Euclidean ℓ_2-norm on vectors, or the analogous Frobenius norm for matrices, but our theory also applies to certain types of weighted norms. In addition, we provide bounds on the quantity $\mathcal{R}(\widehat{\theta}_{\lambda_n} - \theta^*)$, which measures the error in the regularizer norm. In the classical setting, the ambient dimension d stays fixed while the number of observations n tends to infinity. Under these conditions, there are standard techniques for proving consistency and asymptotic normality for the error $\widehat{\theta}_{\lambda_n} - \theta^*$. In contrast, the analysis presented throughout this thesis is all within a high-dimensional framework, in which the tuple (n, d), as well as other problem parameters, such as vector sparsity or matrix rank etc., are all allowed to tend to infinity. In contrast to asymptotic statements, our goal is to obtain explicit finite sample error bounds that hold with high probability.

3.2.2 Decomposability of $\mathcal{R}$

The first ingredient in our analysis is a property of the regularizer known as decomposability, defined in terms of a pair of subspaces $\mathcal{M} \subseteq \overline{\mathcal{M}}$ of $\mathbb{R}^d$. The role of the *model subspace* $\mathcal{M}$ is to capture the constraints specified by the model; for instance, it might be the subspace of vectors with a particular support (see Example 3.1), or a subspace of low-rank matrices (see Example 3.3). The orthogonal complement of the space $\overline{\mathcal{M}}$, namely the set

$$\overline{\mathcal{M}}^{\perp} := \left\{ v \in \mathbb{R}^d \mid \langle u, v \rangle = 0 \quad \text{for all } u \in \overline{\mathcal{M}} \right\} \tag{3.2}$$

is referred to as the *perturbation subspace*, representing deviations away from the model subspace $\mathcal{M}$. In the ideal case, we have $\overline{\mathcal{M}}^{\perp} = \mathcal{M}^{\perp}$, but our definition allows for the possibility that $\overline{\mathcal{M}}$ is strictly larger than $\mathcal{M}$, so that $\overline{\mathcal{M}}^{\perp}$ is strictly smaller than $\mathcal{M}^{\perp}$. This generality is needed for treating the case of low-rank matrices and nuclear norm, as discussed in Example 3.3 to follow.

Definition 3.1. Given a pair of subspaces $\mathcal{M} \subseteq \overline{\mathcal{M}}$, a norm-based regularizer $\mathcal{R}$ is ***decomposable*** with respect to $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ if

$$\mathcal{R}(\theta + \gamma) = \mathcal{R}(\theta) + \mathcal{R}(\gamma) \quad \text{for all } \theta \in \mathcal{M} \text{ and } \gamma \in \overline{\mathcal{M}}^{\perp}. \tag{3.3}$$

In order to build some intuition, let us consider the ideal case $\mathcal{M} = \overline{\mathcal{M}}$ for the time being, so that the decomposition (3.3) holds for all pairs $(\theta, \gamma) \in \mathcal{M} \times \mathcal{M}^{\perp}$. For any given pair (θ, γ) of this form, the vector $\theta + \gamma$ can be interpreted as perturbation of the model vector θ away from the subspace $\mathcal{M}$, and it is desirable that the regularizer penalize such deviations as much as possible. By the triangle inequality for a norm, we always have $\mathcal{R}(\theta + \gamma) \leq \mathcal{R}(\theta) + \mathcal{R}(\gamma)$, so that the decomposability condition (3.3) holds if and only if the triangle inequality is tight for all pairs $(\theta, \gamma) \in (\mathcal{M}, \overline{\mathcal{M}}^{\perp})$. It is exactly in this setting that the regularizer penalizes deviations away from the model subspace $\mathcal{M}$ as much as possible.

In general, it is not difficult to find subspace pairs that satisfy the decomposability property. As a trivial example, any regularizer is decomposable with respect to $\mathcal{M} = \mathbb{R}^d$ and its orthogonal complement $\mathcal{M}^{\perp} = \{0\}$. As will be clear in our main theorem, it is of more interest to find subspace pairs in which the model subspace $\mathcal{M}$ is "small", so that the orthogonal complement $\mathcal{M}^{\perp}$ is "large". To formalize this intuition, let us define the projection operator

$$\Pi_{\mathcal{M}}(u) := \arg\min_{v \in \mathcal{M}} \|u - v\|, \tag{3.4}$$

with the projection $\Pi_{\mathcal{M}^{\perp}}$ defined in an analogous manner. To simplify notation, we frequently use the shorthand $u_{\mathcal{M}} = \Pi_{\mathcal{M}}(u)$ and $u_{\mathcal{M}^{\perp}} = \Pi_{\mathcal{M}^{\perp}}(u)$.

Of interest to us are the action of these projection operators on the unknown parameter $\theta^* \in \mathbb{R}^d$. In the most desirable setting, the model subspace $\mathcal{M}$ can be chosen such that

$\theta^*_{\mathcal{M}} \approx \theta^*$, or equivalently, such that $\theta^*_{\mathcal{M}^\perp} \approx 0$. If this can be achieved with the model subspace $\mathcal{M}$ remaining relatively small, then our main theorem guarantees that it is possible to estimate θ^* at a relatively fast rate. The following examples illustrate suitable choices of the spaces $\mathcal{M}$ and $\overline{\mathcal{M}}$ in three concrete settings, beginning with the case of sparse vectors.

Example 3.1. *Sparse vectors and ℓ_1-norm regularization.* Suppose the error norm $\|\cdot\|$ is the usual ℓ_2-norm, and that the model class of interest is the set of k-sparse vectors in d dimensions. For any particular subset $S \subseteq \{1, 2, \ldots, d\}$ with cardinality k, we define the model subspace

$$\mathcal{M}(S) := \big\{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \notin S\big\}. \tag{3.5}$$

Here our notation reflects the fact that $\mathcal{M}$ depends explicitly on the chosen subset S. By construction, we have $\Pi_{\mathcal{M}(S)}(\theta^*) = \theta^*$ for any vector θ^* that is supported on S.

In this case, we may define $\overline{\mathcal{M}}(S) = \mathcal{M}(S)$, and note that the orthogonal complement with respect to the Euclidean inner product is given by

$$\overline{\mathcal{M}}^\perp(S) = \mathcal{M}^\perp(S) = \big\{\gamma \in \mathbb{R}^d \mid \gamma_j = 0 \quad \text{for all } j \in S\big\}. \tag{3.6}$$

This set corresponds to the perturbation subspace, capturing deviations away from the set of vectors with support S. We claim that for any subset S, the ℓ_1-norm $\mathcal{R}(\theta) = \|\theta\|_1$ is decomposable with respect to the pair $(\mathcal{M}(S), \mathcal{M}^\perp(S))$. Indeed, by construction of the subspaces, any $\theta \in \mathcal{M}(S)$ can be written in the partitioned form $\theta = (\theta_S, 0_{S^c})$, where $\theta_S \in \mathbb{R}^k$ and $0_{S^c} \in \mathbb{R}^{d-k}$ is a vector of zeros. Similarly, any vector $\gamma \in \mathcal{M}^\perp(S)$ has the partitioned representation $(0_S, \gamma_{S^c})$. Putting together the pieces, we obtain

$$\|\theta + \gamma\|_1 = \|(\theta_S, 0) + (0, \gamma_{S^c})\|_1 \;=\; \|\theta\|_1 + \|\gamma\|_1,$$

showing that the ℓ_1-norm is decomposable as claimed. ♢

As a follow-up to the previous example, it is also worth noting that the same argument shows that for a strictly positive weight vector ω, the *weighted ℓ_1-norm* $\|\theta\|_\omega := \sum_{j=1}^d \omega_j |\theta_j|$ is also decomposable with respect to the pair $(\mathcal{M}(S), \overline{\mathcal{M}}(S))$. For another natural extension, we now turn to the case of sparsity models with more structure.

Example 3.2. *Group-structured norms.* In many applications, sparsity arises in a more structured fashion, with groups of coefficients likely to be zero (or non-zero) simultaneously. In order to model this behavior, suppose that the index set $\{1, 2, \ldots, d\}$ can be partitioned into a set of $N_{\mathcal{G}}$ disjoint groups, say $\mathcal{G} = \{G_1, G_2, \ldots, G_{N_{\mathcal{G}}}\}$. With this set-up, for a given vector $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{N_{\mathcal{G}}}) \in [1, \infty]^{N_{\mathcal{G}}}$, the associated $(1, \vec{\alpha})$-*group norm* takes the form

$$\|\theta\|_{\mathcal{G},\vec{\alpha}} \;:=\; \sum_{t=1}^{N_{\mathcal{G}}} \|\theta_{G_t}\|_{\alpha_t}. \tag{3.7}$$

For instance, with the choice $\vec{\alpha} = (2, 2, \ldots, 2)$, we obtain the group ℓ_1/ℓ_2-norm, corresponding to the regularizer that underlies the group Lasso [152]. On the other hand, the choice $\vec{\alpha} = (\infty, \ldots, \infty)$, corresponding to a form of block ℓ_1/ℓ_∞ regularization, has also been studied in past work [137, 97, 157]. Note that for $\vec{\alpha} = (1, 1, \ldots, 1)$, we obtain the standard ℓ_1 penalty. Interestingly, our analysis shows that setting $\vec{\alpha} \in [2, \infty]^{N_\mathcal{G}}$ can often lead to superior statistical performance.

We now show that the norm $\|\cdot\|_{\mathcal{G},\vec{\alpha}}$ is again decomposable with respect to appropriately defined subspaces. Indeed, given any subset $S_\mathcal{G} \subseteq \{1, \ldots, N_\mathcal{G}\}$ of group indices, say with cardinality $k_\mathcal{G} = |S_\mathcal{G}|$, we can define the subspace

$$\mathcal{M}(S_\mathcal{G}) := \big\{\theta \in \mathbb{R}^d \mid \theta_{G_t} = 0 \quad \text{for all } t \notin S_\mathcal{G}\big\}, \tag{3.8}$$

as well as its orthogonal complement with respect to the usual Euclidean inner product

$$\mathcal{M}^\perp(S_\mathcal{G}) = \overline{\mathcal{M}}^\perp(S_\mathcal{G}) := \big\{\theta \in \mathbb{R}^d \mid \theta_{G_t} = 0 \quad \text{for all } t \in S_\mathcal{G}\big\}. \tag{3.9}$$

With these definitions, for any pair of vectors $\theta \in \mathcal{M}(S_\mathcal{G})$ and $\gamma \in \overline{\mathcal{M}}^\perp(S_\mathcal{G})$, we have

$$\|\theta + \gamma\|_{\mathcal{G},\vec{\alpha}} = \sum_{t \in S_\mathcal{G}} \|\theta_{G_t} + 0_{G_t}\|_{\alpha_t} + \sum_{t \notin S_\mathcal{G}} \|0_{G_t} + \gamma_{G_t}\|_{\alpha_t} = \|\theta\|_{\mathcal{G},\vec{\alpha}} + \|\gamma\|_{\mathcal{G},\vec{\alpha}}, \tag{3.10}$$

thus verifying the decomposability condition. ♢

In the preceding example, we exploited the fact that the groups were non-overlapping in order to establish the decomposability property. Therefore, some modifications would be required in order to choose the subspaces appropriately for overlapping group regularizers proposed in past work [64, 65].

Example 3.3. *Low-rank matrices and nuclear norm.* Now suppose that each parameter $\Theta \in \mathbb{R}^{d_1 \times d_2}$ is a matrix; this corresponds to an instance of our general set-up with $d = d_1 d_2$, as long as we identify the space $\mathbb{R}^{d_1 \times d_2}$ with $\mathbb{R}^{d_1 d_2}$ in the usual way. We equip this space with the inner product $\langle\langle \Theta, \Gamma \rangle\rangle := \text{trace}(\Theta \Gamma^T)$, a choice which yields (as the induced norm) the *Frobenius norm*

$$|||\Theta|||_F := \sqrt{\langle\langle \Theta, \Theta \rangle\rangle} = \sqrt{\sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \Theta_{jk}^2}. \tag{3.11}$$

In many settings, it is natural to consider estimating matrices that are low-rank; examples include principal component analysis, spectral clustering, collaborative filtering, and matrix completion. With certain exceptions, it is computationally expensive to enforce a rank-constraint in a direct manner, so that a variety of researchers have studied the *nuclear*

norm, also known as the trace norm, as a surrogate for a rank constraint. More precisely, the nuclear norm is given by

$$|||\Theta|||_{\text{nuc}} := \sum_{j=1}^{\min\{d_1,d_2\}} \sigma_j(\Theta), \tag{3.12}$$

where $\{\sigma_j(\Theta)\}$ are the singular values of the matrix Θ.

The nuclear norm is decomposable with respect to appropriately chosen subspaces. Let us consider the class of matrices $\Theta \in \mathbb{R}^{d_1 \times d_2}$ that have rank $r \leq \min\{d_1, d_2\}$. For any given matrix Θ, we let $\text{row}(\Theta) \subseteq \mathbb{R}^{d_2}$ and $\text{col}(\Theta) \subseteq \mathbb{R}^{d_1}$ denote its row space and column space respectively. Let U and V be a given pair of r-dimensional subspaces $U \subseteq \mathbb{R}^{d_1}$ and $V \subseteq \mathbb{R}^{d_2}$; these subspaces will represent left and right singular vectors of the target matrix Θ^* to be estimated. For a given pair (U, V), we can define the subspaces $\mathcal{M}(U, V)$ and $\overline{\mathcal{M}}^{\perp}(U, V)$ of $\mathbb{R}^{d_1 \times d_2}$ given by

$$\mathcal{M}(U, V) := \big\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \text{row}(\Theta) \subseteq V, \ \text{col}(\Theta) \subseteq U\big\}, \quad \text{and} \tag{3.13a}$$

$$\overline{\mathcal{M}}^{\perp}(U, V) := \big\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \text{row}(\Theta) \subseteq V^{\perp}, \ \text{col}(\Theta) \subseteq U^{\perp}\big\}. \tag{3.13b}$$

So as to simplify notation, we omit the indices (U, V) when they are clear from context. Unlike the preceding examples, in this case the set $\mathcal{M}$ is not[1] equal to $\overline{\mathcal{M}}$.

Finally, we claim that the nuclear norm is decomposable with respect to the pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$. By construction, any pair of matrices $\Theta \in \mathcal{M}$ and $\Gamma \in \overline{\mathcal{M}}^{\perp}$ have orthogonal row and column spaces, which implies the required decomposability condition—namely $|||\Theta + \Gamma|||_{\text{nuc}} = |||\Theta|||_{\text{nuc}} + |||\Gamma|||_{\text{nuc}}$. Please see Appendix B.1 for a more detailed discussion of the decomposability of the nuclear norm. ♢

A line of recent work (e.g., [38, 148, 35, 2, 62, 90]) has studied matrix problems involving the sum of a low-rank matrix with a sparse matrix, along with the regularizer formed by a weighted sum of the nuclear norm and the elementwise ℓ_1-norm. By a combination of Examples 3.1 and Example 3.3, this regularizer also satisfies the decomposability property with respect to appropriately defined subspaces.

3.2.3 A key consequence of decomposability

Thus far, we have specified a class (3.1) of M-estimators based on regularization, defined the notion of decomposability for the regularizer and worked through several illustrative examples. We now turn to the statistical consequences of decomposability—more specifically,

[1] However, as is required by our theory, we do have the inclusion $\mathcal{M} \subseteq \overline{\mathcal{M}}$. Indeed, given any $\Theta \in \mathcal{M}$ and $\Gamma \in \overline{\mathcal{M}}^{\perp}$, we have $\Theta^T\Gamma = 0$ by definition, which implies that $\langle\langle \Theta, \ \Gamma \rangle\rangle = \text{trace}(\Theta^T\Gamma) = 0$. Since $\Gamma \in \overline{\mathcal{M}}^{\perp}$ was arbitrary, we have shown that Θ is orthogonal to the space $\overline{\mathcal{M}}^{\perp}$, meaning that it must belong to $\overline{\mathcal{M}}$.

its implications for the error vector $\widehat{\Delta}_{\lambda_n} = \widehat{\theta}_{\lambda_n} - \theta^*$, where $\widehat{\theta} \in \mathbb{R}^d$ is any solution of the regularized M-estimation procedure (3.1). For a given inner product $\langle \cdot, \cdot \rangle$, the dual norm of $\mathcal{R}$ is given by

$$\mathcal{R}^*(v) := \sup_{u \in \mathbb{R}^d \setminus \{0\}} \frac{\langle u, v \rangle}{\mathcal{R}(u)} = \sup_{\mathcal{R}(u) \leq 1} \langle u, v \rangle. \tag{3.14}$$

This notion is best understood by working through some examples.

Dual of ℓ_1-norm: For the ℓ_1-norm $\mathcal{R}(u) = \|u\|_1$ previously discussed in Example 3.1, let us compute its dual norm with respect to the Euclidean inner product on $\mathbb{R}^d$. For any vector $v \in \mathbb{R}^d$, we have

$$\sup_{\|u\|_1 \leq 1} \langle u, v \rangle \leq \sup_{\|u\|_1 \leq 1} \sum_{k=1}^{d} |u_k||v_k| \leq \sup_{\|u\|_1 \leq 1} \Big(\sum_{k=1}^{d} |u_k| \Big) \max_{k=1,\ldots,d} |v_k| = \|v\|_\infty.$$

We claim that this upper bound actually holds with equality. In particular, letting j be any index for which $|v_j|$ achieves the maximum $\|v\|_\infty = \max_{k=1,\ldots,d} |v_k|$, suppose that we form a vector $\bar{u} \in \mathbb{R}^d$ with $\bar{u}_j = \operatorname{sign}(v_j)$, and $\bar{u}_k = 0$ for all $k \neq j$. With this choice, we have $\|\bar{u}\|_1 \leq 1$, and hence $\sup_{\|u\|_1 \leq 1} \langle u, v \rangle \geq \sum_{k=1}^{d} \bar{u}_k v_k = \|v\|_\infty$, showing that the dual of the ℓ_1-norm is the ℓ_∞-norm.

Dual of group norm: Now recall the group norm from Example 3.2, specified in terms of a vector $\vec{\alpha} \in [2, \infty]^{N_\mathcal{G}}$. A similar calculation shows that its dual norm, again with respect to the Euclidean norm on $\mathbb{R}^d$, is given by

$$\|v\|_{\mathcal{G}, \vec{\alpha}^*} = \max_{t=1,\ldots,N_\mathcal{G}} \|v\|_{\alpha_t^*} \qquad \text{where } \tfrac{1}{\alpha_t} + \tfrac{1}{\alpha_t^*} = 1 \text{ are dual exponents.} \tag{3.15}$$

As special cases of this general duality relation, the block $(1, 2)$ norm that underlies the usual group Lasso leads to a block $(\infty, 2)$ norm as the dual, whereas the the block $(1, \infty)$ norm leads to a block $(\infty, 1)$ norm as the dual.

Dual of nuclear norm: For the nuclear norm, the dual is defined with respect to the trace inner product on the space of matrices. For any matrix $N \in \mathbb{R}^{d_1 \times d_2}$, it can be shown that

$$\mathcal{R}^*(N) = \sup_{|||M|||_{\text{nuc}} \leq 1} \langle\langle M, N \rangle\rangle = |||N|||_2 = \max_{j=1,\ldots,\min\{d_1, d_2\}} \sigma_j(N),$$

corresponding to the ℓ_∞-norm applied to the vector $\sigma(N)$ of singular values. In the special case of diagonal matrices, this fact reduces to the dual relationship between the vector ℓ_1

and ℓ_∞ norms.

The dual norm plays a key role in our general theory, in particular by specifying a suitable choice of the regularization weight λ_n. We summarize in the following:

Lemma 3.1. *Suppose that $\mathcal{L}$ is a convex and differentiable function, and consider any optimal solution $\widehat{\theta}$ to the optimization problem (3.1) with a strictly positive regularization parameter satisfying*

$$\lambda_n \geq 2\,\mathcal{R}^*(\nabla \mathcal{L}(\theta^*; Z_1^n)). \tag{3.16}$$

Then for any pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ over which $\mathcal{R}$ is decomposable, the error $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$ belongs to the set

$$\mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^\perp; \theta^*) := \big\{\Delta \in \mathbb{R}^d \mid \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \leq 3\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\big\}. \tag{3.17}$$

We prove this result in Appendix A.1.1. It has the following important consequence: for any decomposable regularizer and an appropriate choice (3.16) of regularization parameter, we are guaranteed that the error vector $\widehat{\Delta}$ belongs to a very specific set, depending on the unknown vector θ^*. As illustrated in Figure 3.1, the geometry of the set $\mathbb{C}$ depends on the relation between θ^* and the model subspace $\mathcal{M}$. When $\theta^* \in \mathcal{M}$, then we are guaranteed that $\mathcal{R}(\theta^*_{\mathcal{M}^\perp}) = 0$. In this case, the constraint (3.17) reduces to $\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \leq 3\mathcal{R}(\Delta_{\bar{\mathcal{M}}})$, so that $\mathbb{C}$ is a cone, as illustrated in panel (a). In the more general case when $\theta^* \notin \mathcal{M}$ so that $\mathcal{R}(\theta^*_{\mathcal{M}^\perp}) \neq 0$, the set $\mathbb{C}$ is *not* a cone, but rather a star-shaped set (panel (b)). As will be clarified in the sequel, the case $\theta^* \notin \mathcal{M}$ requires a more delicate treatment.

3.2.4 Restricted strong convexity

We now turn to an important requirement of the loss function, and its interaction with the statistical model. Recall that $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$ is the difference between an optimal solution $\widehat{\theta}_{\lambda_n}$ and the true parameter, and consider the loss difference[2] $\mathcal{L}(\widehat{\theta}_{\lambda_n}) - \mathcal{L}(\theta^*)$. In the classical setting, under fairly mild conditions, one expects that that the loss difference should converge to zero as the sample size n increases. It is important to note, however, that such convergence on its own is *not sufficient* to guarantee that $\widehat{\theta}_{\lambda_n}$ and θ^* are close, or equivalently that $\widehat{\Delta}$ is small. Rather, the closeness depends on the curvature of the loss function, as illustrated in Figure 3.2. In a desirable setting (panel (a)), the loss function is sharply curved around its optimum $\widehat{\theta}_{\lambda_n}$, so that having a small loss difference $|\mathcal{L}(\theta^*) - \mathcal{L}(\widehat{\theta}_{\lambda_n})|$ translates to a small error $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$. Panel (b) illustrates a less desirable setting, in which the loss function is relatively flat, so that the loss difference can be small while the error $\widehat{\Delta}$ is relatively large.

[2]To simplify notation, we frequently write $\mathcal{L}(\theta)$ as shorthand for $\mathcal{L}(\theta; Z_1^n)$ when the underlying data Z_1^n is clear from context.

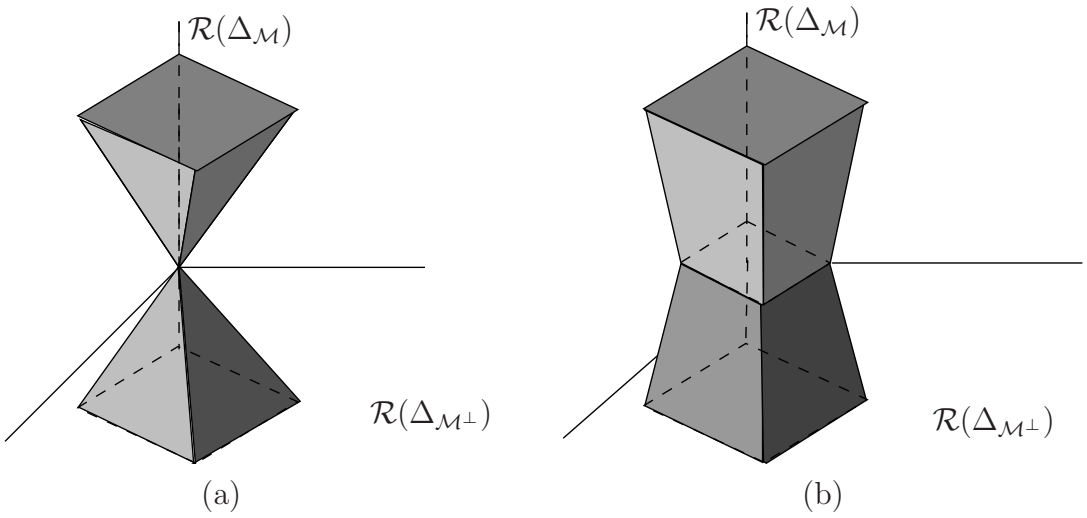


Figure 3.1. Illustration of the set $\mathbb{C}(\mathcal{M}, \mathcal{M}^\perp; \theta^*)$ in the special case $\Delta = (\Delta_1, \Delta_2, \Delta_3) \in \mathbb{R}^3$ and regularizer $\mathcal{R}(\Delta) = \|\Delta\|_1$, relevant for sparse vectors (Example 3.1). This picture shows the case $S = \{3\}$, so that the model subspace is $\mathcal{M}(S) = \{\Delta \in \mathbb{R}^3 \mid \Delta_1 = \Delta_2 = 0\}$, and its orthogonal complement is given by $\mathcal{M}^\perp(S) = \{\Delta \in \mathbb{R}^3 \mid \Delta_3 = 0\}$. (a) In the special case when $\theta^*_1 = \theta^*_2 = 0$, so that $\theta^* \in \mathcal{M}$, the set $\mathbb{C}(\mathcal{M}, \mathcal{M}^\perp; \theta^*)$ is a cone. (b) When θ^* does not belong to $\mathcal{M}$, the set $\mathbb{C}(\mathcal{M}, \mathcal{M}^\perp; \theta^*)$ is enlarged in the co-ordinates (Δ_1, Δ_2) that span $\mathcal{M}^\perp$. It is no longer a cone, but is still a star-shaped set.

The standard way to ensure that a function is "not too flat" is via the notion of strong convexity. Since $\mathcal{L}$ is differentiable by assumption, we may perform a first-order Taylor series expansion at θ^*, and in some direction Δ; the error in this Taylor series is given by

$$\delta\mathcal{L}(\Delta, \theta^*) \; := \; \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \langle \nabla\mathcal{L}(\theta^*), \, \Delta \rangle. \tag{3.18}$$

One way in which to enforce that $\mathcal{L}$ is strongly convex is to require the existence of some positive constant $\kappa > 0$ such that $\delta\mathcal{L}(\Delta, \theta^*) \geq \kappa\|\Delta\|^2$ for all $\Delta \in \mathbb{R}^d$ in a neighborhood of θ^*. When the loss function is twice differentiable, strong convexity amounts to lower bound on the eigenvalues of the Hessian $\nabla^2\mathcal{L}(\theta)$, holding uniformly for all θ in a neighborhood of θ^*.

Under classical "fixed d, large n" scaling, the loss function will be strongly convex under mild conditions. For instance, suppose that population risk $\bar{\mathcal{L}}$ is strongly convex, or equivalently, that the Hessian $\nabla^2\bar{\mathcal{L}}(\theta)$ is strictly positive definite in a neighborhood of θ^*. As a concrete example, when the loss function $\mathcal{L}$ is defined based on negative log likelihood of a statistical model, then the Hessian $\nabla^2\bar{\mathcal{L}}(\theta)$ corresponds to the Fisher information matrix, a quantity which arises naturally in asymptotic statistics. If the dimension d is fixed while the sample size n goes to infinity, standard arguments can be used to show that (under mild regularity conditions) the random Hessian $\nabla^2\mathcal{L}(\theta)$ converges to $\nabla^2\bar{\mathcal{L}}(\theta)$ uniformly for all θ in

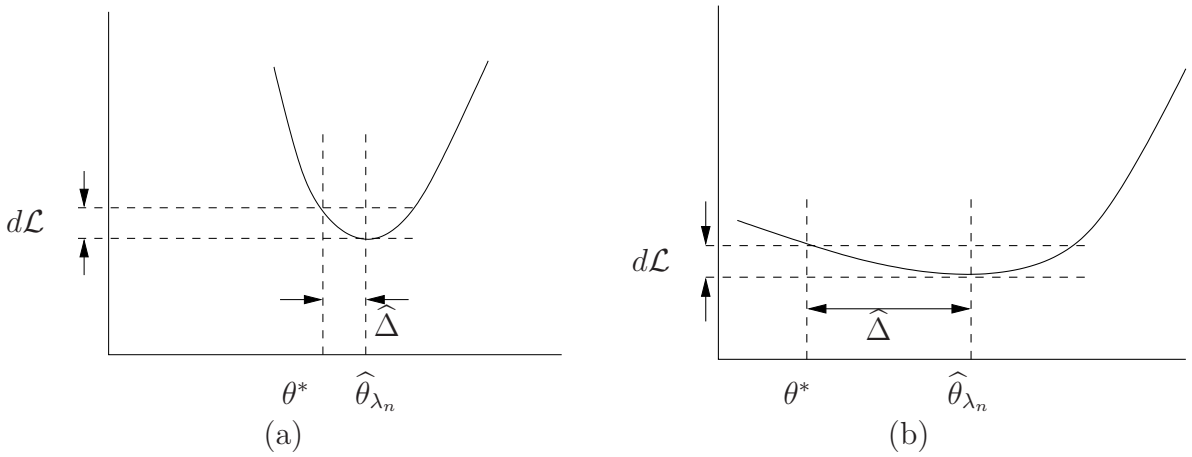


Figure 3.2. Role of curvature in distinguishing parameters. (a) Loss function has high curvature around $\widehat{\Delta}$. A small excess loss $d\mathcal{L} = |\mathcal{L}(\widehat{\theta}_{\lambda_n}) - \mathcal{L}(\theta^*)|$ guarantees that the parameter error $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$ is also small. (b) A less desirable setting, in which the loss function has relatively low curvature around the optimum.

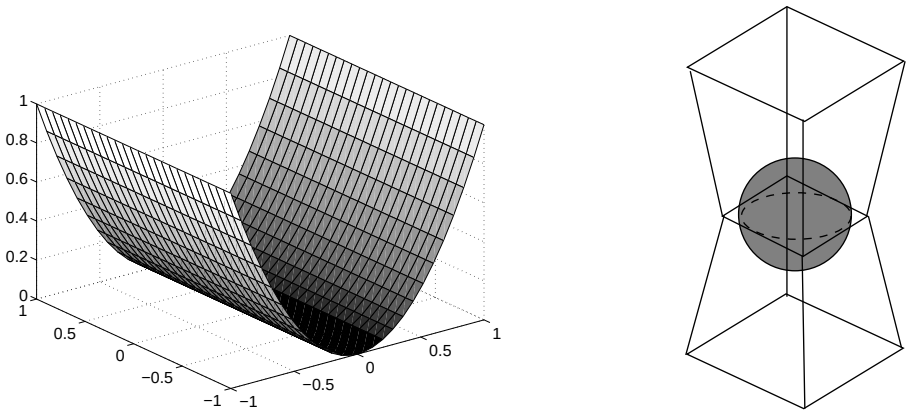
an open neighborhood of θ^*. In contrast, whenever the pair (n, d) both increase in such a way that $d > n$, the situation is drastically different: the Hessian matrix $\nabla^2\mathcal{L}(\theta)$ is often singular. As a concrete example, consider linear regression based on samples $Z_i = (y_i, x_i) \in \mathbb{R} \times \mathbb{R}^d$, for $i = 1, 2, \ldots, n$. Using the least-squares loss $\mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$, the $d \times d$ Hessian matrix $\nabla^2\mathcal{L}(\theta) = \frac{1}{n}X^TX$ has rank at most n, meaning that the loss cannot be strongly convex when $d > n$. Consequently, it impossible to guarantee global strong convexity, so that we need to restrict the set of directions Δ in which we require a curvature condition.

Ultimately, the only direction of interest is given by the error vector $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$. Recall that Lemma 3.1 guarantees that, for suitable choices of the regularization parameter λ_n, this error vector must belong to the set $\mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \theta^*)$, as previously defined (3.17). Consequently, it suffices to ensure that the function is strongly convex over this set, as formalized in the following:

Definition 3.2. The loss function satisfies a ***restricted strong convexity*** (RSC) condition with *curvature* $\kappa_{\mathcal{L}} > 0$ and *tolerance function* $\tau_{\mathcal{L}}$ if

$$\delta\mathcal{L}(\Delta, \theta^*) \;\geq\; \kappa_{\mathcal{L}}\,\|\Delta\|^2 \;-\; \tau_{\mathcal{L}}^2(\theta^*) \qquad \text{for all } \Delta \in \mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \theta^*). \tag{3.19}$$

In the simplest of cases—in particular, when $\theta^* \in \mathcal{M}$—there are many statistical models for which this RSC condition holds with tolerance $\tau_{\mathcal{L}}(\theta^*) = 0$. In the more general setting, it can hold only with a non-zero tolerance term, as illustrated in Figure 3.3(b). As our proofs will clarify, we in fact require only the lower bound (3.19) to hold for the intersection of $\mathbb{C}$ with a local ball $\{\|\Delta\| \leq R\}$ of some radius centered at zero. As will be clarified later, this



(a) (b)

Figure 3.3. (a) Illustration of a generic loss function in the high-dimensional $d > n$ setting: it is curved in certain directions, but completely flat in others. (b) When $\theta^* \notin \mathcal{M}$, the set $\mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \theta^*)$ contains a ball centered at the origin, which necessitates a tolerance term $\tau_{\mathcal{L}}(\theta^*) > 0$ in the definition of restricted strong convexity.

restriction is not necessary for the least-squares loss, but is essential for more general loss functions, such as those that arise in generalized linear models.

We will see in the sequel that for many loss functions, it is possible to prove that with high probability the first-order Taylor series error satisfies a lower bound of the form

$$\delta\mathcal{L}(\Delta, \theta^*) \;\geq\; \kappa_1 \|\Delta\|^2 \;-\; \kappa_2 \, g(n,d) \mathcal{R}^2(\Delta) \qquad \text{for all } \|\Delta\| \leq 1, \tag{3.20}$$

where κ_1, κ_2 are positive constants, and $g(n,d)$ is a function of the sample size n and ambient dimension d, decreasing in the sample size. For instance, in the case of ℓ_1-regularization, for covariates with suitably controlled tails, this type of bound holds for the least squares loss with the function $g(n,d) = \frac{\log d}{n}$; see equation (3.31) to follow. For generalized linear models and the ℓ_1-norm, a similar type of bound is given in equation (3.43). We also provide a bound of this form for the least-squares loss group-structured norms in equation (3.46), with a different choice of the function g depending on the group structure.

A bound of the form (3.20) implies a form of restricted strong convexity as long as $\mathcal{R}(\Delta)$ is not "too large" relative to $\|\Delta\|$. In order to formalize this notion, we define a quantity that relates the error norm and the regularizer:

Definition 3.3 (Subspace compatibility constant)**.** For any subspace $\mathcal{M}$ of $\mathbb{R}^d$, the ***subspace compatibility constant*** with respect to the pair $(\mathcal{R}, \|\cdot\|)$ is given by

$$\Psi(\mathcal{M}) := \sup_{u \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{R}(u)}{\|u\|}. \tag{3.21}$$

This quantity reflects the degree of compatibility between the regularizer and the error norm over the subspace $\mathcal{M}$. In alternative terms, it is the Lipschitz constant of the regularizer with respect to the error norm, restricted to the subspace $\mathcal{M}$. As a simple example, if $\mathcal{M}$ is a k-dimensional co-ordinate subspace, with regularizer $\mathcal{R}(u) = \|u\|_1$ and error norm $\|u\| = \|u\|_2$, then we have $\Psi(\mathcal{M}) = \sqrt{k}$.

This compatibility constant appears explicitly in the bounds of our main theorem, and also arises in establishing restricted strong convexity. Let us now illustrate how it can be used to show that the condition (3.20) implies a form of restricted strong convexity. To be concrete, let us suppose that θ^* belongs to a subspace $\mathcal{M}$; in this case, membership of Δ in the set $\mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \theta^*)$ implies that $\mathcal{R}(\Delta_{\bar{\mathcal{M}}^{\perp}}) \leq 3\mathcal{R}(\Delta_{\bar{\mathcal{M}}})$. Consequently, by triangle inequality and the definition (3.21), we have

$$\mathcal{R}(\Delta) \leq \mathcal{R}(\Delta_{\bar{\mathcal{M}}^{\perp}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) \;\leq\; 4\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) \leq 4\Psi(\overline{\mathcal{M}})\|\Delta\|.$$

Therefore, whenever a bound of the form (3.20) holds and $\theta^* \in \mathcal{M}$, we are guaranteed that

$$\delta\mathcal{L}(\Delta, \theta^*) \;\geq\; \big\{\kappa_1 - 16\,\kappa_2\,\Psi^2(\overline{\mathcal{M}})g(n,d)\big\}\|\Delta\|^2 \qquad \text{for all } \|\Delta\| \leq 1.$$

Consequently, as long as the sample size is large enough that $16\kappa_2\,\Psi^2(\overline{\mathcal{M}})g(n,d) < \frac{\kappa_1}{2}$, the restricted strong convexity condition will hold with $\kappa_{\mathcal{L}} = \frac{\kappa_1}{2}$ and $\tau_{\mathcal{L}}(\theta^*) = 0$. We make use of arguments of this flavor throughout this thesis.

3.3 Bounds for general M-estimators

We are now ready to state a general result that provides bounds and hence convergence rates for the error $\|\widehat{\theta}_{\lambda_n} - \theta^*\|$, where $\widehat{\theta}_{\lambda_n}$ is any optimal solution of the convex program (3.1). Although it may appear somewhat abstract at first sight, this result has a number of concrete and useful consequences for specific models. In particular, we recover as an immediate corollary the best known results about estimation in sparse linear models with general designs [20, 95], as well as a number of new results, including minimax-optimal rates for estimation under ℓ_q-sparsity constraints and estimation of block-structured sparse matrices. We also apply these theorems to establishing results for sparse generalized linear models [100], matrix decomposition problems [2], and sparse non-parametric regression models [110]. Results for the estimation of low-rank matrices are presented in more detail in Chapters 4 and 5.

Let us recall our running assumptions on the structure of the convex program (3.1).

(G1) The regularizer $\mathcal{R}$ is a norm, and is decomposable with respect to the subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, where $\mathcal{M} \subseteq \overline{\mathcal{M}}$.

(G2) The loss function $\mathcal{L}$ is convex and differentiable, and satisfies restricted strong convexity with curvature $\kappa_{\mathcal{L}}$ and tolerance $\tau_{\mathcal{L}}$.

The reader should also recall the definition (3.21) of the subspace compatibility constant. With this notation, we can now state the main result of this chapter:

Theorem 3.1 (Bounds for general models)**.** *Under conditions (G1) and (G2), consider the problem* (3.1) *based on a strictly positive regularization constant* $\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\theta^*))$. *Then any optimal solution* $\widehat{\theta}_{\lambda_n}$ *to the convex program* (3.1) *satisfies the bound*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|^2 \leq 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2}\,\Psi^2(\overline{\mathcal{M}}) + \frac{\lambda_n}{\kappa_{\mathcal{L}}}\big\{2\tau_{\mathcal{L}}^2(\theta^*) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\big\}, \tag{3.22}$$

Remarks: Let us consider in more detail some different features of this result.

(a) It should be noted that Theorem 3.1 is actually a *deterministic* statement about the set of optimizers of the convex program (3.1) for a fixed choice of λ_n. Although the program is convex, it need not be strictly convex, so that the global optimum might be attained at more than one point $\widehat{\theta}_{\lambda_n}$. The stated bound holds for any of these optima. Probabilistic analysis is required when Theorem 3.1 is applied to particular statistical models, and we need to verify that the regularizer satisfies the condition

$$\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\theta^*)), \tag{3.23}$$

and that the loss satisfies the RSC condition. A challenge here is that since θ^* is unknown, it is usually impossible to compute the right-hand side of the condition (3.23). Instead, when we derive consequences of Theorem 3.1 for different statistical models, we use concentration inequalities in order to provide bounds that hold with high probability over the data.

(b) Second, note that Theorem 3.1 actually provides a *family of bounds,* one for each pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ of subspaces for which the regularizer is decomposable. Ignoring the term involving $\tau_{\mathcal{L}}$ for the moment, for any given pair, the error bound is the sum of two terms, corresponding to estimation error $\mathcal{E}_{\text{err}}$ and approximation error $\mathcal{E}_{\text{app}}$, given by (respectively)

$$\mathcal{E}_{\text{err}} := 9\,\frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2}\Psi^2(\overline{\mathcal{M}}), \quad \text{and} \quad \mathcal{E}_{\text{app}} := 4\,\frac{\lambda_n}{\kappa_{\mathcal{L}}}\mathcal{R}(\theta^*_{\mathcal{M}^\perp}). \tag{3.24}$$

As the dimension of the subspace $\mathcal{M}$ increases (so that the dimension of $\mathcal{M}^\perp$ decreases), the approximation error tends to zero. But since $\mathcal{M} \subseteq \overline{\mathcal{M}}$, the estimation error is increasing at the same time. Thus, in the usual way, optimal rates are obtained by choosing $\mathcal{M}$ and $\overline{\mathcal{M}}$ so as to balance these two contributions to the error. We illustrate such choices for various specific models to follow.

(c) As will be clarified in the sequel, many high-dimensional statistical models have an unidentifiable component, and the tolerance term $\tau_{\mathcal{L}}$ reflects the degree of this non-identifiability.

A large body of past work on sparse linear regression has focused on the case of exactly sparse regression models for which the unknown regression vector θ^* is k-sparse. For this special case, recall from Example 3.1 in Section 3.2.2 that we can define an k-dimensional subspace $\mathcal{M}$ that contains θ^*. Consequently, the associated set $\mathbb{C}(\mathcal{M}, \mathcal{M}^{\perp}; \theta^*)$ is a cone (see Figure 3.1(a)), and it is thus possible to establish that restricted strong convexity (RSC) holds with tolerance parameter $\tau_{\mathcal{L}}(\theta^*) = 0$. This same reasoning applies to other statistical models, among them group-sparse regression, in which a small subset of groups are active, as well as low-rank matrix estimation. The following corollary provides a simply stated bound that covers all of these models:

Corollary 3.1. *Suppose that, in addition to the conditions of Theorem 3.1, the unknown θ^* belongs to $\mathcal{M}$ and the RSC condition holds over $\mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}, \theta^*)$ with $\tau_{\mathcal{L}}(\theta^*) = 0$. Then any optimal solution $\widehat{\theta}_{\lambda_n}$ to the convex program (3.1) satisfies the bounds*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\| \leq 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}}\Psi^2(\overline{\mathcal{M}}), \qquad \text{and} \tag{3.25a}$$

$$\mathcal{R}(\widehat{\theta}_{\lambda_n} - \theta^*) \leq 12\,\frac{\lambda_n}{\kappa_{\mathcal{L}}}\Psi^2(\overline{\mathcal{M}}). \tag{3.25b}$$

Focusing first on the bound (3.25a), it consists of three terms, each of which has a natural interpretation. First, it is inversely proportional to the RSC constant $\kappa_{\mathcal{L}}$, so that higher curvature guarantees lower error, as is to be expected. The error bound grows proportionally with the subspace compatibility constant $\Psi(\overline{\mathcal{M}})$, which measures the compatibility between the regularizer $\mathcal{R}$ and error norm $\|\cdot\|$ over the subspace $\overline{\mathcal{M}}$ (see Definition 3.3). This term increases with the size of subspace $\overline{\mathcal{M}}$, which contains the model subspace $\mathcal{M}$. Third, the bound also scales linearly with the regularization parameter λ_n, which must be strictly positive and satisfy the lower bound (3.23). The bound (3.25b) on the error measured in the regularizer norm is similar, except that it scales quadratically with the subspace compatibility constant. As the proof clarifies, this additional dependence arises since the regularizer over the subspace $\overline{\mathcal{M}}$ is larger than the norm $\|\cdot\|$ by a factor of at most $\Psi(\overline{\mathcal{M}})$ (see Definition 3.3).

Obtaining concrete rates using Corollary 3.1 requires some work in order to verify the conditions of Theorem 3.1, and to provide control on the three quantities in the bounds (3.25a) and (3.25b), as illustrated in the examples to follow.

3.4 Convergence rates for sparse regression

As an illustration, we begin with one of the simplest statistical models, namely the standard linear model. It is based on n observations $Z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ of covariate-response pairs.

Let $y \in \mathbb{R}^n$ denote a vector of the responses, and let $X \in \mathbb{R}^{n \times d}$ be the design matrix, where $x_i \in \mathbb{R}^d$ is the i^{th} row. This pair is linked via the linear model

$$y \;=\; X\theta^* + w, \tag{3.26}$$

where $\theta^* \in \mathbb{R}^d$ is the unknown regression vector, and $w \in \mathbb{R}^n$ is a noise vector. To begin, we focus on this simple linear set-up, and describe extensions to generalized models in Section 3.4.4.

Given the data set $Z_1^n = (y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times d}$, our goal is to obtain a "good" estimate $\widehat{\theta}$ of the regression vector θ^*, assessed either in terms of its ℓ_2-error $\|\widehat{\theta} - \theta^*\|_2$ or its ℓ_1-error $\|\widehat{\theta} - \theta^*\|_1$. It is worth noting that whenever $d > n$, the standard linear model (3.26) is unidentifiable in a certain sense, since the rectangular matrix $X \in \mathbb{R}^{n \times d}$ has a nullspace of dimension at least $d - n$. Consequently, in order to obtain an identifiable model—or at the very least, to bound the degree of non-identifiability—it is essential to impose additional constraints on the regression vector θ^*. One natural constraint is some type of sparsity in the regression vector; for instance, one might assume that θ^* has at most k non-zero coefficients, as discussed at more length in Section 3.4.2. More generally, one might assume that although θ^* is not exactly sparse, it can be well-approximated by a sparse vector, in which case one might say that θ^* is "weakly sparse", "sparsifiable" or "compressible". Section 3.4.3 is devoted to a more detailed discussion of this weakly sparse case.

A natural M-estimator for this problem is the Lasso [39, 131], obtained by solving the ℓ_1-penalized quadratic program

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \Big\}, \tag{3.27}$$

for some choice $\lambda_n > 0$ of regularization parameter. Note that this Lasso estimator is a particular case of the general M-estimator (3.1), based on the loss function and regularization pair $\mathcal{L}(\theta; Z_1^n) = \frac{1}{2n}\|y - X\theta\|_2^2$ and $\mathcal{R}(\theta) = \sum_{j=1}^d |\theta_j| \;=\; \|\theta\|_1$. We now show how Theorem 3.1 can be specialized to obtain bounds on the error $\widehat{\theta}_{\lambda_n} - \theta^*$ for the Lasso estimate.

3.4.1 Restricted eigenvalues for sparse linear regression

For the least-squares loss function that underlies the Lasso, the first-order Taylor series expansion from Definition 3.2 is exact, so that

$$\delta\mathcal{L}(\Delta, \theta^*) \;=\; \langle \Delta, \frac{1}{n} X^T X \Delta \rangle \;=\; \frac{1}{n} \|X\Delta\|_2^2.$$

Thus, in this special case, the Taylor series error is independent of θ^*, a fact which allows for substantial theoretical simplification. More precisely, in order to establish restricted strong convexity, it suffices to establish a lower bound on $\|X\Delta\|_2^2/n$ that holds uniformly for an appropriately restricted subset of d-dimensional vectors Δ.

As previously discussed in Example 3.1, for any subset $S \subseteq \{1, 2, \ldots, d\}$, the ℓ_1-norm is decomposable with respect to the subspace $\mathcal{M}(S) = \{\theta \in \mathbb{R}^d \mid \theta_{S^c} = 0\}$ and its orthogonal complement. When the unknown regression vector $\theta^* \in \mathbb{R}^d$ is exactly sparse, it is natural to choose S equal to the support set of θ^*. By appropriately specializing the definition (3.17) of $\mathbb{C}$, we are led to consider the cone

$$\mathbb{C}(S) := \big\{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\big\}. \tag{3.28}$$

See Figure 3.1(a) for an illustration of this set in three dimensions. With this choice, restricted strong convexity with respect to the ℓ_2-norm is equivalent to requiring that the design matrix X satisfy the condition

$$\frac{\|X\theta\|_2^2}{n} \geq \kappa_{\mathcal{L}} \|\theta\|_2^2 \qquad \text{for all } \theta \in \mathbb{C}(S). \tag{3.29}$$

This lower bound is a type of *restricted eigenvalue* (RE) condition, and has been studied in past work on basis pursuit and the Lasso (e.g., [20, 95, 109, 139]). One could also require that a related condition hold with respect to the ℓ_1-norm—viz.

$$\frac{\|X\theta\|_2^2}{n} \geq \kappa'_{\mathcal{L}} \frac{\|\theta\|_1^2}{|S|} \qquad \text{for all } \theta \in \mathbb{C}(S). \tag{3.30}$$

This type of ℓ_1-based RE condition is less restrictive than the corresponding ℓ_2-version (3.29). We refer the reader to the paper by van de Geer and Bühlmann [139] for an extensive discussion of different types of restricted eigenvalue or compatibility conditions.

It is natural to ask whether there are many matrices that satisfy these types of RE conditions. If X has i.i.d. entries following a sub-Gaussian distribution (including Gaussian and Bernoulli variables as special cases), then known results in random matrix theory imply that the restricted isometry property [32] holds with high probability, which in turn implies that the RE condition holds [20, 139]. Since statistical applications involve design matrices with substantial dependency, it is natural to ask whether an RE condition also holds for more general random designs. This question was addressed by Raskutti et al. [109, 108], who showed that if the design matrix $X \in \mathbb{R}^{n \times d}$ is formed by independently sampling each row $X_i \sim N(0, \Sigma)$, referred to as the Σ-*Gaussian ensemble*, then there are strictly positive constants (κ_1, κ_2), depending only on the positive definite matrix Σ, such that

$$\frac{\|X\theta\|_2^2}{n} \geq \kappa_1 \|\theta\|_2^2 - \kappa_2 \frac{\log d}{n} \|\theta\|_1^2 \qquad \text{for all } \theta \in \mathbb{R}^d \tag{3.31}$$

with probability greater than $1 - c_1 \exp(-c_2 n)$. The bound (3.31) has an important consequence: it guarantees that the RE property (3.29) holds[3] with $\kappa_{\mathcal{L}} = \frac{\kappa_1}{2} > 0$ as long as

[3]To see this fact, note that for any $\theta \in \mathbb{C}(S)$, we have $\|\theta\|_1 \leq 4\|\theta_S\|_1 \leq 4\sqrt{k}\|\theta_S\|_2$. Given the lower bound (3.31), for any $\theta \in \mathbb{C}(S)$, we have the lower bound $\frac{\|X\theta\|_2}{\sqrt{n}} \geq \big\{\kappa_1 - 4\kappa_2\sqrt{\frac{k \log d}{n}}\big\} \|\theta\|_2 \geq \frac{\kappa_1}{2}\|\theta\|_2$, where final inequality follows as long as $n > 64(\kappa_2/\kappa_1)^2\, k \log d$.

$n > 64(\kappa_2/\kappa_1)\, k \log d$. Therefore, not only do there exist matrices satisfying the RE property (3.29), but any matrix sampled from a Σ-Gaussian ensemble will satisfy it with high probability. Related analysis by Rudelson and Zhou [121] extends these types of guarantees to the case of sub-Gaussian designs, also allowing for substantial dependencies among the covariates. We refer the reader to a more detailed discussion in Appendix D.4, which presents a discussion for general Gaussian observation operators with an arbitrary regularizer $\mathcal{R}$.

3.4.2 Lasso estimates with exact sparsity

We now show how Corollary 3.1 can be used to derive convergence rates for the error of the Lasso estimate when the unknown regression vector θ^* is k-sparse. In order to state these results, we require some additional notation. Using $X_j \in \mathbb{R}^n$ to denote the j^{th} column of X, we say that X is *column-normalized* if

$$\frac{\|X_j\|_2}{\sqrt{n}} \leq 1 \qquad \text{for all } j = 1, 2, \ldots, d. \tag{3.32}$$

Here we have set the upper bound to one in order to simplify notation. This particular choice entails no loss of generality, since we can always rescale the linear model appropriately (including the observation noise variance) so that it holds.

In addition, we assume that the noise vector $w \in \mathbb{R}^n$ is zero-mean and has *sub-Gaussian tails*, meaning that there is a constant $\sigma > 0$ such that for any fixed $\|v\|_2 = 1$,

$$\mathbb{P}\big[|\langle v, w\rangle| \geq t\big] \leq 2 \exp\big(-\frac{\delta^2}{2\sigma^2}\big) \qquad \text{for all } \delta > 0. \tag{3.33}$$

For instance, this condition holds when the noise vector w has i.i.d. $N(0,1)$ entries, or consists of independent bounded random variables. Under these conditions, we recover as a corollary of Theorem 3.1 the following result:

Corollary 3.2. *Consider an k-sparse instance of the linear regression model (3.26) such that X satisfies the RE condition (3.29), and the column normalization condition (3.32). Given the Lasso program (3.27) with regularization parameter $\lambda_n = 4\,\sigma\sqrt{\frac{\log d}{n}}$, then with probability at least $1 - c_1 \exp(-c_2 n \lambda_n^2)$, any optimal solution $\widehat{\theta}_{\lambda_n}$ satisfies the bounds*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq \frac{64\,\sigma^2}{\kappa_{\mathcal{L}}^2}\,\frac{k \log d}{n}, \quad \text{and} \quad \|\widehat{\theta}_{\lambda_n} - \theta^*\|_1 \leq \frac{24\,\sigma}{\kappa_{\mathcal{L}}}\, k \sqrt{\frac{\log d}{n}}. \tag{3.34}$$

Although error bounds of this form are known from past work (e.g., [20, 32, 95]), our proof illuminates the underlying structure that leads to the different terms in the bound—in particular, see equations (3.25a) and (3.25b) in the statement of Corollary 3.1.

Proof. We first note that the RE condition (3.30) implies that RSC holds with respect to the subspace $\mathcal{M}(S)$. As discussed in Example 3.1, the ℓ_1-norm is decomposable with respect to $\mathcal{M}(S)$ and its orthogonal complement, so that we may set $\overline{\mathcal{M}}(S) = \mathcal{M}(S)$. Since any vector $\theta \in \mathcal{M}(S)$ has at most k non-zero entries, the subspace compatibility constant is given by $\Psi(\mathcal{M}(S)) = \sup_{\theta \in \mathcal{M}(S)\setminus\{0\}} \frac{\|\theta\|_1}{\|\theta\|_2} = \sqrt{k}$.

The final step is to compute an appropriate choice of the regularization parameter. The gradient of the quadratic loss evaluated at θ^* is given by $\nabla\mathcal{L}(\theta; (y, X)) = \frac{1}{n} X^T w$, whereas the dual norm of the ℓ_1-norm is the ℓ_∞-norm. Consequently, we need to specify a choice of $\lambda_n > 0$ such that

$$\lambda_n \geq 2\,\mathcal{R}^*(\nabla\mathcal{L}(\theta^*)) = 2\big\|\frac{1}{n} X^T w\big\|_\infty$$

with high probability. Using the column normalization (3.32) and sub-Gaussian (3.33) conditions, for each $j = 1, \ldots, d$, we have the tail bound $\mathbb{P}\big[|\langle X_j,\, w\rangle/n| \geq t\big] \leq 2\exp\big(-\frac{nt^2}{2\sigma^2}\big)$. Consequently, by union bound, we conclude that $\mathbb{P}\big[\|X^T w/n\|_\infty \geq t\big] \leq 2\exp\big(-\frac{nt^2}{2\sigma^2} + \log d\big)$. Setting $t^2 = \frac{4\sigma^2 \log d}{n}$, we see that the choice of λ_n given in the statement is valid with probability at least $1 - c_1\exp(-c_2 n\lambda_n^2)$. Consequently, the claims (3.34) follow from the bounds (3.25a) and (3.25b) in Corollary 3.1. □

3.4.3 Lasso estimates with weakly sparse models

We now consider regression models for which θ^* is not exactly sparse, but rather can be approximated well by a sparse vector. One way in which to formalize this notion is by considering the ℓ_q "ball" of radius R_q, given by

$$\mathbb{B}_q(R_q) := \{\theta \in \mathbb{R}^d \mid \sum_{i=1}^d |\theta_i|^q \leq R_q\}, \qquad \text{where } q \in [0, 1] \text{ is fixed.}$$

In the special case $q = 0$, this set corresponds to an exact sparsity constraint—that is, $\theta^* \in \mathbb{B}_0(R_0)$ if and only if θ^* has at most R_0 non-zero entries. More generally, for $q \in (0, 1]$, the set $\mathbb{B}_q(R_q)$ enforces a certain decay rate on the ordered absolute values of θ^*.

In the case of weakly sparse vectors, the constraint set $\mathbb{C}$ takes the form

$$\mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}; \theta^*) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1 + 4\|\theta^*_{S^c}\|_1\}. \tag{3.35}$$

In contrast to the case of exact sparsity, the set $\mathbb{C}$ is no longer a cone, but rather contains a ball centered at the origin— compare panels (a) and (b) of Figure 3.1. As a consequence, it is *never* possible to ensure that $\|X\theta\|_2/\sqrt{n}$ is uniformly bounded from below for all vectors θ in the set (3.35), and so a strictly positive tolerance term $\tau_{\mathcal{L}}(\theta^*) > 0$ is required. The random matrix result (3.31), stated in the previous section, allows us to establish a form of RSC that is appropriate for the setting of ℓ_q-ball sparsity. We summarize our conclusions in the following:

Corollary 3.3. *Suppose that X satisfies the RE condition (3.31) as well as the column normalization condition (3.32), the noise w is sub-Gaussian (3.33), and θ^* belongs to $\mathbb{B}_q(R_q)$ for a radius R_q such that $\sqrt{R_q}\left(\frac{\log d}{n}\right)^{\frac{1}{2}-\frac{q}{4}} \leq 1$. Then if we solve the Lasso with regularization parameter $\lambda_n = 4\sigma\sqrt{\frac{\log d}{n}}$, there are universal positive constants (c_0, c_1, c_2) such that any optimal solution $\widehat{\theta}_{\lambda_n}$ satisfies*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq c_0\, R_q \left(\frac{\sigma^2}{\kappa_1^2}\frac{\log d}{n}\right)^{1-\frac{q}{2}} \tag{3.36}$$

with probability at least $1 - c_1\exp(-c_2 n\lambda_n^2)$.

Remarks: Note that this corollary is a strict generalization of Corollary 3.2, to which it reduces when $q = 0$. More generally, the parameter $q \in [0, 1]$ controls the relative "sparsifiability" of θ^*, with larger values corresponding to lesser sparsity. Naturally then, the rate slows down as q increases from 0 towards 1. In fact, Raskutti et al. [109] show that the rates (3.36) are minimax-optimal over the ℓ_q-balls—implying that not only are the consequences of Theorem 3.1 sharp for the Lasso, but more generally, no algorithm can achieve faster rates.

Proof. Since the loss function $\mathcal{L}$ is quadratic, the proof of Corollary 3.2 shows that the stated choice $\lambda_n = 4\sqrt{\frac{\sigma^2\log d}{n}}$ is valid with probability at least $1 - c\exp(-c' n\lambda_n^2)$. Let us now show that the RSC condition holds. We do so via condition (3.31) applied to equation (3.35). For a threshold $\mu > 0$ to be chosen, define the thresholded subset

$$S_\mu := \big\{j \in \{1, 2, \ldots, d\} \;\mid\; |\theta_j^*| > \mu\big\}. \tag{3.37}$$

Now recall the subspaces $\mathcal{M}(S_\mu)$ and $\mathcal{M}^\perp(S_\mu)$ previously defined in equations (3.5) and (3.6) of Example 3.1, where we set $S = S_\mu$. The following lemma, proved in Appendix A.2, provides sufficient conditions for restricted strong convexity with respect to these subspace pairs:

Lemma 3.2. *Suppose that the conditions of Corollary 3.3 hold, and $n > 9\kappa_2|S_\mu|\log d$. Then with the choice $\mu = \frac{\lambda_n}{\kappa_1}$, the RSC condition holds over $\mathbb{C}(\mathcal{M}(S_\mu), \mathcal{M}^\perp(S_\mu), \theta^*)$ with $\kappa_{\mathcal{L}} = \kappa_1/4$ and $\tau_{\mathcal{L}}^2 = 8\kappa_2\frac{\log d}{n}\|\theta^*_{S_\mu^c}\|_1^2$.*

Consequently, we may apply Theorem 3.1 with $\kappa_{\mathcal{L}} = \kappa_1/4$ and $\tau_{\mathcal{L}}^2(\theta^*) = 8\kappa_2\frac{\log d}{n}\|\theta^*_{S_\mu^c}\|_1^2$ to conclude that

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq 144\frac{\lambda_n^2}{\kappa_1^2}\,|S_\mu| + \frac{4\,\lambda_n}{\kappa_1}\left\{16\kappa_2\frac{\log d}{n}\|\theta^*_{S_\mu^c}\|_1^2 + 4\|\theta^*_{S_\mu^c}\|_1\right\}, \tag{3.38}$$

where we have used the fact that $\Psi^2(S_\mu) = |S_\mu|$, as noted in the proof of Corollary 3.2.

It remains to upper bound the cardinality of S_μ in terms of the threshold μ and ℓ_q-ball radius R_q. Note that we have

$$R_q \geq \sum_{j=1}^{d} |\theta_j^*|^q \;\geq\; \sum_{j \in S_\mu} |\theta_i^*|^q \;\geq\; \mu^q |S_\mu|, \tag{3.39}$$

whence $|S_\mu| \leq \mu^{-q} R_q$ for any $\mu > 0$. Next we upper bound the approximation error $\|\theta^*_{S_\mu^c}\|_1$, using the fact that $\theta^* \in \mathbb{B}_q(R_q)$. Letting S_μ^c denote the complementary set $S_\mu \backslash \{1, 2, \ldots, d\}$, we have

$$\|\theta^*_{S_\mu^c}\|_1 = \sum_{j \in S_\mu^c} |\theta_j^*| \;=\; \sum_{j \in S_\mu^c} |\theta_j^*|^q |\theta_j^*|^{1-q} \leq R_q \, \mu^{1-q}. \tag{3.40}$$

Setting $\mu = \lambda_n / \kappa_1$ and then substituting the bounds (3.39) and (3.40) into the bound (3.38) yields

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq 160 \big(\frac{\lambda_n^2}{\kappa_1^2}\big)^{1-\frac{q}{2}} \, R_q + 64 \kappa_2 \big\{ \big(\frac{\lambda_n^2}{\kappa_1^2}\big)^{1-\frac{q}{2}} R_q \big\}^2 \, \frac{(\log d)/n}{\lambda_n / \kappa_1}.$$

For any fixed noise variance, our choice of regularization parameter ensures that the ratio $\frac{(\log d)/n}{\lambda_n/\kappa_1}$ is of order one, so that the claim follows. □

3.4.4 Extensions to generalized linear models

In this section, we briefly outline extensions of the preceding results to the family of generalized linear models (GLM). Suppose that conditioned on a vector $x \in \mathbb{R}^d$ of covariates, a response variable $y \in \mathcal{Y}$ has the distribution

$$\mathbb{P}_{\theta^*}(y \mid x) \propto \exp\Big\{ \frac{y \, \langle \theta^*, x \rangle - \Phi(\langle \theta^*, x \rangle)}{c(\sigma)} \Big\}. \tag{3.41}$$

Here the quantity $c(\sigma)$ is a fixed and known scale parameter, and the function $\Phi : \mathbb{R} \to \mathbb{R}$ is the link function, also known. The family (3.41) includes many well-known classes of regression models as special cases, including ordinary linear regression (obtained with $\mathcal{Y} = \mathbb{R}$, $\Phi(t) = t^2/2$ and $c(\sigma) = \sigma^2$), and logistic regression (obtained with $\mathcal{Y} = \{0, 1\}$, $c(\sigma) = 1$ and $\Phi(t) = \log(1 + \exp(t))$).

Given samples $Z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$, the goal is to estimate the unknown vector $\theta^* \in \mathbb{R}^d$. Under a sparsity assumption on θ^*, a natural estimator is based on minimizing the

(negative) log likelihood, combined with an ℓ_1-regularization term. This combination leads to the convex program

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \underbrace{\frac{1}{n} \sum_{i=1}^{n} \big\{ -y_i \langle \theta, x_i \rangle + \Phi(\langle \theta, x_i \rangle) \big\}}_{\mathcal{L}(\theta; Z_1^n)} + \lambda_n \|\theta\|_1 \Big\}. \tag{3.42}$$

In order to extend the error bounds from the previous section, a key ingredient is to establish that this GLM-based loss function satisfies a form of restricted strong convexity. Along these lines, Negahban et al. [100] proved the following result: suppose that the covariate vectors x_i are zero-mean with covariance matrix $\Sigma \succ 0$, and are drawn i.i.d. from a distribution with sub-Gaussian tails (see equation (3.33)). Then there are constants κ_1, κ_2 such that the first-order Taylor series error for the GLM-based loss (3.42) satisfies the lower bound

$$\delta\mathcal{L}(\Delta, \theta^*) \geq \kappa_1 \|\Delta\|_2^2 - \kappa_2 \frac{\log d}{n} \|\Delta\|_1^2 \qquad \text{for all } \|\Delta\|_2 \leq 1. \tag{3.43}$$

As discussed following Definition 3.2, this type of lower bound implies that $\mathcal{L}$ satisfies a form of RSC, as long as the sample size scales as $n = \Omega(k \log d)$, where k is the target sparsity. Consequently, this lower bound (3.43) allows us to recover analogous bounds on the error $\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2$ of the GLM-based estimator (3.42).

3.5 Convergence rates for group-structured norms

The preceding two sections addressed M-estimators based on ℓ_1-regularization, the simplest type of decomposable regularizer. We now turn to some extensions of our results to more complex regularizers that are also decomposable. Various researchers have proposed extensions of the Lasso based on regularizers that have more structure than the ℓ_1 norm (e.g., [137, 152, 157, 92, 11]). Such regularizers allow one to impose different types of block-sparsity constraints, in which groups of parameters are assumed to be active (or inactive) simultaneously. These norms arise in the context of multivariate regression, where the goal is to predict a multivariate output in $\mathbb{R}^m$ on the basis of a set of d covariates. Here it is appropriate to assume that groups of covariates are useful for predicting the different elements of the m-dimensional output vector. We refer the reader to the papers [137, 152, 157, 92, 11] for further discussion of and motivation for the use of block-structured norms.

Given a collection $\mathcal{G} = \{G_1, \ldots, G_{N_\mathcal{G}}\}$ of groups, recall from Example 3.2 in Section 3.2.2 the definition of the group norm $\|\cdot\|_{\mathcal{G},\vec{\alpha}}$. In full generality, this group norm is based on a weight vector $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{N_\mathcal{G}}) \in [2, \infty]^{N_\mathcal{G}}$, one for each group. For simplicity, here we consider the case when $\alpha_t = \alpha$ for all $t = 1, 2, \ldots, N_\mathcal{G}$, and we use $\|\cdot\|_{\mathcal{G},\alpha}$ to denote the associated group norm. As a natural extension of the Lasso, we consider the *block Lasso*

estimator

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \frac{1}{n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_{\mathcal{G},\alpha} \Big\}, \tag{3.44}$$

where $\lambda_n > 0$ is a user-defined regularization parameter. Different choices of the parameter α yield different estimators, and in this section, we consider the range $\alpha \in [2, \infty]$. This range covers the two most commonly applied choices, $\alpha = 2$, often referred to as the group Lasso, as well as the choice $\alpha = +\infty$.

3.5.1 Restricted strong convexity for group sparsity

As a parallel to our analysis of ordinary sparse regression, our first step is to provide a condition sufficient to guarantee restricted strong convexity for the group-sparse setting. More specifically, we state the natural extension of condition (3.31) to the block-sparse setting, and prove that it holds with high probability for the class of Σ-Gaussian random designs. Recall from Theorem 3.1 that the dual norm of the regularizer plays a central role. As discussed previously, for the block-$(1,\alpha)$-regularizer, the associated dual norm is a block-(∞, α^*) norm, where (α, α^*) are conjugate exponents satisfying $\frac{1}{\alpha} + \frac{1}{\alpha^*} = 1$.

Letting $\varepsilon \sim N(0, I_{d \times d})$ be a standard normal vector, we consider the following condition. Suppose that there are strictly positive constants (κ_1, κ_2) such that, for all $\Delta \in \mathbb{R}^d$, we have

$$\frac{\|X\Delta\|_2^2}{n} \geq \kappa_1 \|\Delta\|_2^2 - \kappa_2 \, \rho_{\mathcal{G}}^2(\alpha^*) \, \|\Delta\|_{1,\alpha}^2 \qquad \text{where } \rho_{\mathcal{G}}(\alpha^*) := \mathbb{E}\Big[\max_{t=1,2,\ldots,N_{\mathcal{G}}} \frac{\|\varepsilon_{G_t}\|_{\alpha^*}}{\sqrt{n}} \Big]. \tag{3.45}$$

To understand this condition, first consider the special case of $N_{\mathcal{G}} = d$ groups, each of size one, so that the group-sparse norm reduces to the ordinary ℓ_1-norm, and its dual is the ℓ_∞-norm. Using $\alpha = 2$ for concreteness, we have $\rho_{\mathcal{G}}(2) = \mathbb{E}[\|\varepsilon\|_\infty]/\sqrt{n} \leq \sqrt{\frac{3 \log d}{n}}$, using standard bounds on Gaussian maxima. Therefore, condition (3.45) reduces to the earlier condition (3.31) in this special case.

Let us consider a more general setting, say with $\alpha = 2$ and $N_{\mathcal{G}}$ groups each of size m, so that $d = N_{\mathcal{G}} m$. For this choice of groups and norm, we have $\rho_{\mathcal{G}}(2) = \mathbb{E}\big[\max_{t=1,\ldots,N_{\mathcal{G}}} \frac{\|\varepsilon_{G_t}\|_2}{\sqrt{n}} \big]$ where each sub-vector w_{G_t} is a standard Gaussian vector with m elements. Since $\mathbb{E}[\|\varepsilon_{G_t}\|_2] \leq \sqrt{m}$, tail bounds for χ^2-variates yield $\rho_{\mathcal{G}}(2) \leq \sqrt{\frac{m}{n}} + \sqrt{\frac{3 \log N_{\mathcal{G}}}{n}}$, so that the condition (3.45) is equivalent to

$$\frac{\|X\Delta\|_2^2}{n} \geq \kappa_1 \|\Delta\|_2^2 - \kappa_2 \left[\sqrt{\frac{m}{n}} + \sqrt{\frac{3 \log N_{\mathcal{G}}}{n}} \right]^2 \|\Delta\|_{\mathcal{G},2}^2 \qquad \text{for all } \Delta \in \mathbb{R}^d. \tag{3.46}$$

Thus far, we have seen the form that condition (3.45) takes for different choices of the groups and parameter α. It is natural to ask whether there are any matrices that satisfy the

condition (3.45). As shown in the following result, the answer is affirmative—more strongly, almost every matrix satisfied from the Σ-Gaussian ensemble will satisfy this condition with high probability. (Here we recall that for a non-degenerate covariance matrix, a random design matrix $X \in \mathbb{R}^{n \times d}$ is drawn from the Σ-Gaussian ensemble if each row $x_i \sim N(0, \Sigma)$, i.i.d. for $i = 1, 2, \ldots, n$.)

Proposition 3.1. *For a design matrix $X \in \mathbb{R}^{n \times d}$ from the Σ-ensemble, there are constants (κ_1, κ_2) depending only Σ such that condition (3.45) holds with probability greater than $1 - c_1 \exp(-c_2 n)$.*

We provide the proof of this result in Appendix A.3.1. This condition can be used to show that appropriate forms of RSC hold, for both the cases of exactly group-sparse and weakly sparse vectors. As with ℓ_1-regularization, these RSC conditions are milder than analogous group-based RIP conditions (e.g., [63, 129, 11]), which require that all sub-matrices up to a certain size are close to isometries.

3.5.2 Convergence rates

Apart from RSC, we impose one additional condition on the design matrix. For a given group G of size m, let us view the matrix $X_G \in \mathbb{R}^{n \times m}$ as an operator from $\ell_\alpha^m \to \ell_2^n$, and define the associated operator norm $|||X_G|||_{\alpha \to 2} := \max_{\|\theta\|_\alpha = 1} \|X_G\, \theta\|_2$. We then require that

$$\frac{|||X_{G_t}|||_{\alpha \to 2}}{\sqrt{n}} \leq 1 \qquad \text{for all } t = 1, 2, \ldots, N_{\mathcal{G}}. \tag{3.47}$$

Note that this is a natural generalization of the column normalization condition (3.32), to which it reduces when we have $N_{\mathcal{G}} = d$ groups, each of size one. As before, we may assume without loss of generality, rescaling X and the noise as necessary, that condition (3.47) holds with constant one. Finally, we define the maximum group size $m = \max_{t=1,\ldots,N_{\mathcal{G}}} |G_t|$. With this notation, we have the following novel result:

Corollary 3.4. *Suppose that the noise w is sub-Gaussian (3.33), and the design matrix X satisfies condition (3.45) and the block normalization condition (3.47). If we solve the group Lasso with*

$$\lambda_n \geq 2\sigma \left\{ \frac{m^{1-1/\alpha}}{\sqrt{n}} + \sqrt{\frac{\log N_{\mathcal{G}}}{n}} \right\}, \tag{3.48}$$

then with probability at least $1 - 2/{N_{\mathcal{G}}}^2$, for any group subset $S_{\mathcal{G}} \subseteq \{1, 2, \ldots, N_{\mathcal{G}}\}$ with cardinality $|S_{\mathcal{G}}| = k_{\mathcal{G}}$, any optimal solution $\widehat{\theta}_{\lambda_n}$ satisfies

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq \frac{4\lambda_n^2}{\kappa_{\mathcal{L}}^2}\, k_{\mathcal{G}} + \frac{4\lambda_n}{\kappa_{\mathcal{L}}} \sum_{t \notin S_{\mathcal{G}}} \|\theta^*_{G_t}\|_\alpha. \tag{3.49}$$

Remarks: Since the result applies to any $\alpha \in [2, \infty]$, we can observe how the choices of different group-sparse norms affect the convergence rates. So as to simplify this discussion, let us assume that the groups are all of equal size m, so that $d = mN_{\mathcal{G}}$ is the ambient dimension of the problem.

Case $\alpha = 2$: The case $\alpha = 2$ corresponds to the block $(1, 2)$ norm, and the resulting estimator is frequently referred to as the group Lasso. For this case, we can set the regularization parameter as $\lambda_n = 2\sigma\{\sqrt{\frac{m}{n}} + \sqrt{\frac{\log N_{\mathcal{G}}}{n}}\}$. If we assume moreover that θ^* is exactly group-sparse, say supported on a group subset $S_{\mathcal{G}} \subseteq \{1, 2, \ldots, N_{\mathcal{G}}\}$ of cardinality $k_{\mathcal{G}}$, then the bound (3.49) takes the form

$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim \frac{k_{\mathcal{G}}\, m}{n} + \frac{k_{\mathcal{G}}\, \log N_{\mathcal{G}}}{n}. \tag{3.50}$$

Similar bounds were derived in independent work by Lounici et al. [83] and Huang and Zhang [63] for this special case of exact block sparsity. The analysis here shows how the different terms arise, in particular via the noise magnitude measured in the dual norm of the block regularizer.

In the more general setting of weak block sparsity, Corollary 3.4 yields a number of novel results. For instance, for a given set of groups $\mathcal{G}$, we can consider the block sparse analog of the ℓ_q-"ball"—namely the set

$$\mathbb{B}_q(R_q; \mathcal{G}, 2) := \Big\{ \theta \in \mathbb{R}^d \mid \sum_{t=1}^{N_{\mathcal{G}}} \|\theta_{G_t}\|_2^q \leq R_q \Big\}.$$

In this case, if we optimize the choice of S in the bound (3.49) so as to trade off the estimation and approximation errors, then we obtain

$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim R_q \left(\frac{m}{n} + \frac{\log N_{\mathcal{G}}}{n} \right)^{1 - \frac{q}{2}},$$

which is a novel result. This result is a generalization of our earlier Corollary 3.3, to which it reduces when we have $N_{\mathcal{G}} = d$ groups each of size $m = 1$.

Case $\alpha = +\infty$: Now consider the case of ℓ_1/ℓ_∞ regularization, as suggested in past work [137]. In this case, Corollary 3.4 implies that $\|\widehat{\theta} - \theta^*\|_2^2 \precsim \frac{k\, m^2}{n} + \frac{k \log N_{\mathcal{G}}}{n}$. Similar to the case $\alpha = 2$, this bound consists of an estimation term, and a search term. The estimation term $\frac{km^2}{n}$ is larger by a factor of m, which corresponds to amount by which an ℓ_∞-ball in m dimensions is larger than the corresponding ℓ_2-ball.

We provide the proof of Corollary 3.4 in Appendix A.3.2. It is based on verifying the conditions of Theorem 3.1: more precisely, we use Proposition 3.1 in order to establish RSC, and we provide a lemma that shows that the regularization choice (3.48) is valid in the context of Theorem 3.1.

3.6 Discussion

In this chapter, we have presented a unified framework for deriving error bounds and convergence rates for a class of regularized M-estimators. The theory is high-dimensional and non-asymptotic in nature, meaning that it yields explicit bounds that hold with high probability for finite sample sizes, and reveals the dependence on dimension and other structural parameters of the model. Two properties of the M-estimator play a central role in our framework. We isolated the notion of a regularizer being *decomposable* with respect to a pair of subspaces, and showed how it constrains the error vector—meaning the difference between any solution and the nominal parameter—to lie within a very specific set. This fact is significant, because it allows for a fruitful notion of *restricted strong convexity* to be developed for the loss function. Since the usual form of strong convexity cannot hold under high-dimensional scaling, this interaction between the decomposable regularizer and the loss function is essential.

Our main result (Theorem 3.1) provides a deterministic bound on the error for a broad class of regularized M-estimators. By specializing this result to different statistical models, we derived various explicit convergence rates for different estimators, including some known results and a range of novel results. We derived convergence rates for sparse linear models, both under exact and approximate sparsity assumptions, and these results have been shown to be minimax optimal [109]. In the case of sparse group regularization, we established a novel upper bound of the oracle type, with a separation between the approximation and estimation error terms. This framework has also been applied to obtain minimax-optimal rates for noisy matrix decomposition, which involves using a combination of the nuclear norm and elementwise ℓ_1-norm. Finally, in a result that we report elsewhere, we have also applied these results to deriving convergence rates on generalized linear models. Doing so requires leveraging that restricted strong convexity can also be shown to hold for these models, as stated in the bound (3.43).

There are a variety of interesting open questions associated with our work. In this chapter, for simplicity of exposition, we have specified the regularization parameter in terms of the dual norm $\mathcal{R}^*$ of the regularizer. In many cases, this choice leads to optimal convergence rates, including linear regression over ℓ_q-balls (Corollary 3.3) for sufficiently small radii, and various instances of low-rank matrix regression. In other cases, some refinements of our convergence rates are possible; for instance, in the special case of linear sparsity regression (i.e., an exactly sparse vector, with a constant fraction of non-zero elements), our rates can be sharpened by a more careful analysis of the noise term, which allows for a slightly smaller choice of the regularization parameter. Similarly, there are other non-parametric settings in which a more delicate choice of the regularization parameter is required [73, 110]. Last, we suspect that there are many other statistical models, not discussed in this chapter, for which this framework can yield useful results. Some examples include different types of hierarchical regularizers and/or overlapping group regularizers [64, 65], as well as methods using combinations of decomposable regularizers, such as the fused Lasso [132].

Chapter 4

Low-rank matrix estimation

4.1 Introduction

In this chapter, we focus on the problem of high-dimensional inference in the setting of matrix estimation. As mentioned in the previous chapter, there is already a substantial body of work on the problem of sparse matrix recovery. In contrast, our interest in this chapter is the problem of estimating a matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ that is either *exactly low rank*, meaning that it has at most $r \ll \min\{d_1, d_2\}$ non-zero singular values, or more generally is *near low-rank*, meaning that it can be well-approximated by a matrix of low rank. As we discuss at more length in the sequel, such exact or approximate low-rank conditions are appropriate for many applications, including multivariate or multi-task forms of regression, system identification for autoregressive processes, collaborative filtering, and matrix recovery from random projections. Analogous to the use of an ℓ_1-regularizer for enforcing sparsity, we consider the use of the nuclear norm (also known as the trace norm) for enforcing a rank constraint in the matrix setting. By definition, the nuclear norm is the sum of the singular values of a matrix, and so encourages sparsity in the vector of singular values, or equivalently for the matrix to be low-rank. The problem of low-rank matrix approximation and the use of nuclear norm regularization have been studied by various researchers. In her Ph.D. thesis, Fazel [51] discusses the use of nuclear norm as a heuristic for restricting the rank of a matrix, showing that in practice it is often able to yield low-rank solutions. Other researchers have provided theoretical guarantees on the performance of nuclear norm and related methods for low-rank matrix approximation. Srebro et al. [127] proposed nuclear norm regularization for the collaborative filtering problem, and established risk consistency under certain settings. Recht et al. [117] provided sufficient conditions for exact recovery using the nuclear norm heuristic when observing random projections of a low-rank matrix, a set-up analogous to the compressed sensing model in sparse linear regression [44, 31]. Other researchers have studied a version of matrix completion in which a subset of entries are revealed, and the goal is to obtain perfect reconstruction either via the nuclear norm heuristic [33] or by other

SVD-based methods [71]. We will elaborate on these added complexities in Chapter 5. For general observation models, Bach [9] has provided results on the consistency of nuclear norm minimization in noisy settings, but applicable to the classical "fixed p" setting. In addition, Yuan et al. [154] provide non-asymptotic bounds on the operator norm error of the estimate in the multi-task setting, provided that the design matrices are orthogonal. Under the assumption of RIP, Lee and Bresler [79] prove stability properties of least-squares under nuclear norm constraint when a form of restricted isometry property is imposed on the sampling operator. Liu and Vandenberghe [82] develop an efficient interior-point method for solving nuclear-norm constrained problems, and illustrate its usefulness for problems of system identification, an application also considered in this chapter. Finally, in related work, Rohde and Tsybakov [119] and Candes and Plan [30] have studied certain aspects of nuclear norm minimization under high-dimensional scaling. We discuss connections to this work at more length in Section 4.3.2 following the statement of our main results.

The goal of this chapter is to analyze the nuclear norm relaxation for a general class of noisy observation models, and obtain non-asymptotic error bounds on the Frobenius norm that hold under high-dimensional scaling, and are applicable to both exactly and approximately low-rank matrices. We begin by presenting a generic observation model, and illustrating how it can be specialized to the several cases of interest, including low-rank multivariate regression, estimation of autoregressive processes, and random projection (compressed sensing) observations. In particular, this model is specified in terms of an operator $\mathfrak{X}$, which may be deterministic or random depending on the setting, that maps any matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ to a vector of N noisy observations. We then present a single main theorem (Theorem 4.1) followed by two corollaries that cover the cases of exact low-rank constraints (Corollary 4.1) and near low-rank constraints (Corollary 4.2) respectively. These results demonstrate that high-dimensional error rates are controlled by two key quantities. First, the (random) observation operator $\mathfrak{X}$ is required to satisfy *restricted strong convexity* (RSC), introduced in Chapter 3, which ensures that the loss function has sufficient curvature to guarantee consistent recovery of the unknown matrix Θ^*. As we show via various examples, this RSC condition is weaker than the RIP property, which requires that the sampling operator behave very much like an isometry on low-rank matrices. Second, our theory provides insight into the *choice of regularization parameter* that weights the nuclear norm, showing that an appropriate choice is to set it proportional to the spectral norm of a random matrix defined by the adjoint of the observation operator $\mathfrak{X}$, and the observation noise in the problem.

This initial set of results, though appealing in terms of their simple statements and generality, are somewhat abstractly formulated. Our next contribution is to show that by specializing our main result (Theorem 4.1) to three classes of models, we can obtain some concrete results based on readily interpretable conditions. In particular, Corollary 4.3 deals with the case of low-rank multivariate regression, relevant for applications in multitask learning. We show that the random operator $\mathfrak{X}$ satisfies the RSC property for a broad class of observation models, and we use random matrix theory to provide an appropriate choice of

the regularization parameter. Our next result, Corollary 4.4, deals with the case of estimating the matrix of parameters specifying a vector autoregressive (VAR) process [6, 86]. The usefulness of the nuclear norm in this context has been demonstrated by Liu and Vandenberghe [82]. Here we also establish that a suitable RSC property holds with high probability for the random operator $\mathfrak{X}$, and also specify a suitable choice of the regularization parameter. We note that the technical details here are considerably more subtle than the case of low-rank multivariate regression, due to dependencies introduced by the autoregressive sampling scheme. Accordingly, in addition to terms that involve the size, the matrix dimensions and rank, our bounds also depend on the mixing rate of the VAR process. Finally, we turn to the compressed sensing observation model for low-rank matrix recovery, as introduced by Recht and colleagues [117, 116]. In this setting, we again establish that the RSC property holds with high probability, specify a suitable choice of the regularization parameter, and thereby obtain a Frobenius error bound for noisy observations (Corollary 4.5). A technical result that we prove en route—namely, Proposition 4.1—is of possible independent interest, since it provides a bound on the constrained norm of a random Gaussian operator. In particular, this proposition allows us to obtain a sharp result (Corollary 4.6) for the problem of recovering a low-rank matrix from perfectly observed random Gaussian projections with a general dependency structure.

The remainder of this chapter is organized as follows. Section 4.2 is devoted to background material, and the set-up of the problem. We present a generic observation model for low-rank matrices, and then illustrate how it captures various cases of interest. We then define the convex program based on nuclear norm regularization that we analyze in this chapter. In Section 4.3, we state our main theoretical results and discuss their consequences for different model classes. Section 4.4 is devoted to the proofs of our results; in each case, we break down the key steps in a series of lemmas, with more technical details deferred to the appendices. In Section 4.5, we present the results of various simulations that illustrate excellent agreement between the theoretical bounds and empirical behavior.

Notation: For the convenience of the reader, we collect standard pieces of notation here. For a pair of matrices Θ and Γ with commensurate dimensions, we let $\langle\langle \Theta, \Gamma \rangle\rangle = \text{trace}(\Theta^T \Gamma)$ denote the trace inner product on matrix space. For a matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$, we define $m = \min\{d_1, d_2\}$, and denote its (ordered) singular values by $\sigma_1(\Theta) \geq \sigma_2(\Theta) \geq \ldots \geq \sigma_m(\Theta) \geq 0$. We also use the notation $\sigma_{\max}(\Theta) = \sigma_1(\Theta)$ and $\sigma_{\min}(\Theta) = \sigma_m(\Theta)$ to refer to the maximal and minimal singular values respectively. We use the notation $|||\cdot|||$ for various types of matrix norms based on these singular values, including the *nuclear norm* $|||\Theta|||_{\text{nuc}} = \sum_{j=1}^m \sigma_j(\Theta)$, the *spectral or operator norm* $|||\Theta|||_2 = \sigma_1(\Theta)$, and the *Frobenius norm* $|||\Theta|||_F = \sqrt{\text{trace}(\Theta^T\Theta)} = \sqrt{\sum_{j=1}^m \sigma_j^2(\Theta)}$. We refer the reader to Horn and Johnson [60, 61] for more background on these matrix norms and their properties.

4.2 Background and problem set-up

We begin with some background on problems and applications in which rank constraints arise, before describing a generic observation model. We then introduce the concrete convex program based on nuclear norm regularization that we study in this chapter.

4.2.1 Models with rank constraints

Imposing a rank r constraint on a matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ is equivalent to requiring the rows (or columns) of Θ^* lie in some r-dimensional subspace of $\mathbb{R}^{d_2}$ (or $\mathbb{R}^{d_1}$ respectively). Such types of rank constraints (or approximate forms thereof) arise in a variety of applications, as we discuss here. In some sense, rank constraints are a generalization of sparsity constraints; rather than assuming that the data is sparse in a known basis, a rank constraint implicitly imposes sparsity but without assuming the basis.

We first consider the problem of multivariate regression, also referred to as multi-task learning in statistical machine learning. The goal of *multivariate regression* is to estimate a prediction function that maps covariates $Z_j \in \mathbb{R}^m$ to multi-dimensional output vectors $Y_j \in \mathbb{R}^{d_1}$. More specifically, let us consider the linear model, specified by a matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, of the form

$$Y_a = \Theta^* Z_a + W_a, \qquad \text{for } a = 1, \ldots, n, \tag{4.1}$$

where $\{W_a\}_{a=1}^n$ is an i.i.d. sequence of d_1-dimensional zero-mean noise vectors. Given a collection of observations $\{Z_a, Y_a\}_{a=1}^n$ of covariate-output pairs, our goal is to estimate the unknown matrix Θ^*. This type of model has been used in many applications, including analysis of fMRI image data [59], analysis of EEG data decoding [5], neural response modeling [23] and analysis of financial data. This model and closely related ones also arise in the problem of collaborative filtering [127], in which the goal is to predict users' preferences for items (such as movies or music) based on their and other users' ratings of related items. The papers [1, 7] discuss additional instances of low-rank decompositions. In all of these settings, the low-rank condition translates into the existence of a smaller set of "features" that are actually controlling the prediction.

As a second (not unrelated) example, we now consider the problem of system identification in vector autoregressive processes (see the book [86] for detailed background). A *vector autoregressive* (VAR) process in m-dimensions is a stochastic process $\{Z_t\}_{t=1}^\infty$ specified by an initialization $Z_1 \in \mathbb{R}^m$, followed by the recursion

$$Z_{t+1} = \Theta^* Z_t + W_t, \qquad \text{for } t = 1, 2, 3, \ldots. \tag{4.2}$$

In this recursion, the sequence $\{W_t\}_{t=1}^\infty$ consists of i.i.d. samples of innovations noise. We assume that each vector $W_t \in \mathbb{R}^m$ is zero-mean with covariance matrix $C \succ 0$, so that the

process $\{Z_t\}_{t=1}^{\infty}$ is zero-mean, and has a covariance matrix Σ given by the solution of the discrete-time Ricatti equation

$$\Sigma = \Theta^* \Sigma (\Theta^*)^T + C. \tag{4.3}$$

The goal of system identification in a VAR process is to estimate the unknown matrix $\Theta^* \in \mathbb{R}^{m \times m}$ on the basis of a sequence of samples $\{Z_t\}_{t=1}^{n}$. In many application domains, it is natural to expect that the system is controlled primarily by a low-dimensional subset of variables. For instance, models of financial data might have an ambient dimension m of thousands (including stocks, bonds, and other financial instruments), but the behavior of the market might be governed by a much smaller set of macro-variables (combinations of these financial instruments). Similar statements apply to other types of time series data, including neural data [23, 52], subspace tracking models in signal processing, and motion models models in computer vision. While the form of system identification formulated here assumes direct observation of the state variables $\{Z_t\}_{t=1}^{n}$, it is also possible to tackle the more general problem when only noisy versions are observed (e.g., see Liu and Vandenberghe [82]). An interesting feature of the system identification problem is that the matrix Θ^*, in addition to having low rank, might also be required to satisfy some type of structural constraint (e.g., having a Hankel-type structure), and the estimator that we consider here allows for this possibility.

A third example that we consider in this chapter is a *compressed sensing* observation model, in which one observes random projections of the unknown matrix Θ^*. This observation model has been studied extensively in the context of estimating sparse vectors [44, 31], and Recht and colleagues [117] suggested and studied its extension to low-rank matrices. In their set-up, one observes trace inner products of the form $\langle\langle X_i, \Theta^* \rangle\rangle = \text{trace}(X_i^T \Theta^*)$, where $X_i \in \mathbb{R}^{d_1 \times d_2}$ is a random matrix (for instance, filled with standard normal $N(0,1)$ entries), so that $\langle\langle X_i, \Theta^* \rangle\rangle$ is a standard random projection. In the sequel, we consider this model with a more general family of random projections involving matrices with dependent entries. Like compressed sensing for sparse vectors, applications of this model include computationally efficient updating in large databases (where the matrix Θ^* measures the difference between the data base at two different time instants), and matrix denoising.

4.2.2 A generic observation model

We now introduce a generic observation model that will allow us to deal with these different observation models in an unified manner. For pairs of matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$, recall the Frobenius or trace inner product $\langle\langle A, B \rangle\rangle := \text{trace}(BA^T)$. We then consider a linear observation model of the form

$$y_i = \langle\langle X_i, \Theta^* \rangle\rangle + \varepsilon_i, \qquad \text{for } i = 1, 2, \ldots, N, \tag{4.4}$$

which is specified by the sequence of observation matrices $\{X_i\}_{i=1}^N$ and observation noise $\{\varepsilon_i\}_{i=1}^N$. This observation model can be written in a more compact manner using operator-theoretic notation. In particular, let us define the observation vector

$$\vec{y} = \begin{bmatrix} y_1 & \ldots & y_N \end{bmatrix}^T \in \mathbb{R}^N,$$

with a similar definition for $\vec{\varepsilon} \in \mathbb{R}^N$ in terms of $\{\varepsilon_i\}_{i=1}^N$. We then use the observation matrices $\{X_i\}_{i=1}^N$ to define an operator $\mathfrak{X} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^N$ via $\big[\mathfrak{X}(\Theta)\big]_i = \langle\langle X_i, \Theta \rangle\rangle$. With this notation, the observation model (4.4) can be re-written as

$$\vec{y} = \mathfrak{X}(\Theta^*) + \vec{\varepsilon}. \tag{4.5}$$

Let us illustrate the form of the observation model (4.5) for some of the applications that we considered earlier.

Example 4.1 (Multivariate regression)**.** Recall the observation model (4.1) for multivariate regression. In this case, we make n observations of vector pairs $(Y_a, Z_a) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. Accounting for the d_1-dimensional nature of the output, after the model is scalarized, we receive a total of $N = d_1 n$ observations. Let us introduce the quantity $b = 1, \ldots, d_1$ to index the different elements of the output, so that we can write

$$Y_{ab} = \langle\langle e_b Z_a^T, \Theta^* \rangle\rangle + W_{ab}, \qquad \text{for } b = 1, 2, \ldots, d_1. \tag{4.6}$$

By re-indexing this collection of $N = nd_1$ observations via the mapping

$$(a, b) \mapsto i = a + (b-1)\, d_1,$$

we recognize multivariate regression as an instance of the observation model (4.4) with observation matrix $X_i = e_b Z_a^T$ and scalar observation $y_i = Y_{ab}$.

Example 4.2 (Vector autoregressive processes)**.** Recall that a vector autoregressive (VAR) process is defined by the recursion (4.2), and suppose that we observe an n-sequence $\{Z_t\}_{t=1}^n$ produced by this recursion. Since each $Z_t = \begin{bmatrix} Z_{t1} & \ldots & Z_{tm} \end{bmatrix}^T$ is m-variate, the scalarized sample size is $N = nm$. Letting $b = 1, 2, \ldots, m$ index the dimension, we have

$$Z_{(t+1)\,b} = \langle\langle e_b Z_t^T, \Theta^* \rangle\rangle + W_{tb}. \tag{4.7}$$

In this case, we re-index the collection of $N = nm$ observations via the mapping

$$(t, b) \mapsto i = t + (b-1)\, m.$$

After doing so, we see that the autoregressive problem can be written in the form (4.4) with $y_i = Z_{(t+1)\,b}$ and observation matrix $X_i = e_b Z_t^T$.

Example 4.3 (Compressed sensing)**.** As mentioned earlier, this is a natural extension of the compressed sensing observation model for sparse vectors to the case of low-rank matrices [117, 116]. In a typical form of compressed sensing, the observation matrix $X_i \in \mathbb{R}^{d_1 \times d_2}$ has i.i.d. standard normal $N(0,1)$ entries, so that one makes observations of the form

$$y_i = \langle\langle X_i, \Theta^* \rangle\rangle + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots, N. \tag{4.8}$$

By construction, these observations are an instance of the model (4.4). In the sequel, we study a more general observation model, in which the entries of X_i are allowed to have general Gaussian dependencies. For this problem, the more compact form (4.5) involves a random Gaussian operator mapping $\mathbb{R}^{d_1 \times d_2}$ to $\mathbb{R}^N$, and we study some of its properties in the sequel.

4.2.3 Regression with nuclear norm regularization

We now consider an estimator that is naturally suited to the problems described in the previous section. Recall that the *nuclear or trace norm* of a matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$ is given by $|||\Theta|||_{\text{nuc}} = \sum_{j=1}^{m} \sigma_j(\Theta)$, corresponding to the sum of its singular values. Given a collection of observations $(y_i, X_i) \in \mathbb{R} \times \mathbb{R}^{d_1 \times d_2}$, for $i = 1, \ldots, N$ from the observation model (4.4), we consider estimating the unknown $\Theta^* \in \Omega$ by solving the following optimization problem

$$\widehat{\Theta} \in \arg\min_{\Theta \in \Omega} \left\{ \frac{1}{2N} \|\vec{y} - \mathfrak{X}(\Theta)\|_2^2 + \lambda_N |||\Theta|||_{\text{nuc}} \right\}, \tag{4.9}$$

where Ω is a convex subset of $\mathbb{R}^{d_1 \times d_2}$, and $\lambda_N > 0$ is a regularization parameter. When $\Omega = \mathbb{R}^{d_1 \times d_2}$, the optimization problem (4.9) can be viewed as the analog of the Lasso estimator [131], tailored to low-rank matrices as opposed to sparse vectors. We include the possibility of a more general convex set Ω since they arise naturally in certain applications (e.g., Hankel-type constraints in system identification [82]). When Ω is a polytope (with $\Omega = \mathbb{R}^{d_1 \times d_2}$ as a special case), then the optimization problem (4.9) can be solved in time polynomial in the sample size N and the matrix dimensions d_1 and d_2. Indeed, the optimization problem (4.9) is an instance of a *semidefinite program* [142], a class of convex optimization problems that can be solved efficiently by various polynomial-time algorithms [21]. For instance, Liu and Vandenberghe [82] develop an efficient interior point method for solving constrained versions of nuclear norm programs. Moreover, as we discuss in Section 4.5, there are a variety of first-order methods for solving the semidefinite program (SDP) defining our M-estimator [102, 66]. These first-order methods are well-suited to the high-dimensional problems arising in statistical settings, and we make use of one in performing our simulations.

Like in any typical M-estimator for statistical inference, the regularization parameter λ_N is specified by the statistician. As part of the theoretical results in the next section, we provide suitable choices of this parameter so that the estimate $\widehat{\Theta}$ is close in Frobenius norm

to the unknown matrix Θ^*. The setting of the regularizer depends on the knowledge of the noise variance. While in general one might need to estimate this parameter through cross validation [50, 17], we assume knowledge of the noise variance in order to most succinctly demonstrate the empirical behavior of our results through the experiments.

4.3 Main results and some consequences

In this section, we state our main results and discuss some of their consequences. Section 4.3.1 is devoted to results that apply to generic instances of low-rank problems, whereas Section 4.3.3 is devoted to the consequences of these results for more specific problem classes, including low-rank multivariate regression, estimation of vector autoregressive processes, and recovery of low-rank matrices from random projections.

4.3.1 Results for general model classes

We begin recalling the definition of restricted strong convexity (RSC) presented in equation (3.19). Recall that RSC is the key technical condition that allows us to control the error $\widehat{\Theta} - \Theta^*$ between an optimal solution $\widehat{\Theta}$ and the unknown matrix Θ^*. Restricted strong convexity amounts to guaranteeing that the quadratic loss function in the convex program (4.9) is strictly convex over a restricted set of directions. Letting $\mathbb{C} \subseteq \mathbb{R}^{d_1 \times d_2}$ denote the restricted set of directions, we say that the operator $\mathfrak{X}$ satisfies RSC over the set $\mathbb{C}$ if there exists some $\kappa(\mathfrak{X}) > 0$ such that

$$\frac{1}{2N}\|\mathfrak{X}(\Delta)\|_2^2 \geq \kappa(\mathfrak{X}) \, |||\Delta|||_F^2 \qquad \text{for all } \Delta \in \mathbb{C}. \tag{4.10}$$

We note that analogous conditions have been used to establish error bounds in the context of sparse linear regression [20, 40], in which case the set $\mathbb{C}$ corresponded to certain subsets of sparse vectors. These types of conditions are weaker than restricted isometry properties, since they involve only lower bounds on the operator $\mathfrak{X}$, and the constant $\kappa(\mathfrak{X})$ can be arbitrarily small.

Of course, the definition (4.10) hinges on the choice of the restricted set $\mathbb{C}$. In order to specify some appropriate sets for the case of (near) low-rank matrices, we require some additional notation. Any matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ has a singular value decomposition of the form $\Theta^* = UDV^T$, where $U \in \mathbb{R}^{d_1 \times d_1}$ and $V \in \mathbb{R}^{d_2 \times d_2}$ are orthonormal matrices. For each integer $r \in \{1, 2, \ldots, m\}$, we let $U^r \in \mathbb{R}^{d_1 \times r}$ and $V^r \in \mathbb{R}^{d_2 \times r}$ be the sub-matrices of singular vectors associated with the top r singular values of Θ^*. We recall the following two subspaces of $\mathbb{R}^{d_1 \times d_2}$ from equation (3.13)

$$\mathcal{M}(U^r, V^r) := \big\{\Delta \in \mathbb{R}^{d_1 \times d_2} \mid \text{row}(\Delta) \subseteq V^r \text{ and } \text{col}(\Delta) \subseteq U^r\big\}, \qquad \text{and} \tag{4.11a}$$

$$\overline{\mathcal{M}}^{\perp}(U^r, V^r) := \big\{\Delta \in \mathbb{R}^{d_1 \times d_2} \mid \text{row}(\Delta) \perp V^r \text{ and } \text{col}(\Delta) \perp U^r\big\}, \tag{4.11b}$$

where $\mathrm{row}(\Delta) \subseteq \mathbb{R}^{d_2}$ and $\mathrm{col}(\Delta) \subseteq \mathbb{R}^{d_1}$ denote the row space and column space, respectively, of the matrix Δ. When (U^r, V^r) are clear from the context, we adopt the shorthand notation $\mathcal{M}$ and $\overline{\mathcal{M}}^{\perp}$.

We can now define the subsets of interest. Let $\Pi_{\overline{\mathcal{M}}^{\perp}}$ denote the projection operator onto the subspace $\overline{\mathcal{M}}^{\perp}$, and define $\Delta'' = \Pi_{\overline{\mathcal{M}}^{\perp}}(\Delta)$ and $\Delta' = \Delta - \Delta''$. For a positive integer $r \leq m = \min\{d_1, d_2\}$ and a tolerance parameter $\delta \geq 0$, consider the following subset of matrices

$$\mathbb{C}(r; \delta) := \left\{ \Delta \in \mathbb{R}^{d_1 \times d_2} \mid |||\Delta|||_F \geq \delta, \; |||\Delta''|||_{\mathrm{nuc}} \leq 3|||\Delta'|||_{\mathrm{nuc}} + 4 \sum_{j=r+1}^{m} \sigma_j(\Theta^*) \right\}. \tag{4.12}$$

Note that this set corresponds to matrices Δ for which the quantity $|||\Delta''|||_{\mathrm{nuc}}$ is relatively small compared to $\Delta - \Delta''$ and the remaining $m - r$ singular values of Θ^*.

The next ingredient is the choice of the regularization parameter λ_N used in solving the SDP (4.9). Our theory specifies a choice for this quantity in terms of the adjoint of the operator $\mathfrak{X}$—namely, the operator $\mathfrak{X}^* : \mathbb{R}^N \to \mathbb{R}^{d_1 \times d_2}$ defined by

$$\mathfrak{X}^*(\vec{\varepsilon}) := \sum_{i=1}^{N} \varepsilon_i X_i. \tag{4.13}$$

With this notation, we come to the first result of the chapter. The statement is a specialization of Theorem 3.1 to the low-rank matrix inference setting. The result is deterministic, which specifies two conditions—namely, an RSC condition and a choice of the regularizer—that suffice to guarantee that any solution of the convex program (4.9) falls within a certain radius.

Theorem 4.1. *Suppose $\Theta^* \in \Omega$ and that the operator $\mathfrak{X}$ satisfies restricted strong convexity with parameter $\kappa(\mathfrak{X}) > 0$ over the set $\mathbb{C}(r; \delta)$, and that the regularization parameter λ_N is chosen such that $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_2/N$. Then any solution $\widehat{\Theta}$ to the semidefinite program (4.9) satisfies*

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \max\left\{ \delta, \; \frac{32\lambda_N \sqrt{r}}{\kappa(\mathfrak{X})}, \; \left[\frac{16 \, \lambda_N \, \sum_{j=r+1}^{m} \sigma_j(\Theta^*)}{\kappa(\mathfrak{X})} \right]^{1/2} \right\}. \tag{4.14}$$

Apart from the tolerance parameter δ, the two main terms in the bound (4.14) have a natural interpretation. The first term (involving $\sqrt{r}$) corresponds to *estimation error*, capturing the difficulty of estimating a rank r matrix. The second is an *approximation error* that describes the gap between the true matrix Θ^* and the best rank r approximation. Understanding the magnitude of the tolerance parameter δ is a bit more subtle, and it depends on the geometry of the set $\mathbb{C}(r; \delta)$, and more specifically, the inequality

$$|||\Delta''|||_{\mathrm{nuc}} \leq 3|||\Delta'|||_{\mathrm{nuc}} + 4 \sum_{j=r+1}^{m} \sigma_j(\Theta^*). \tag{4.15}$$

In the simplest case, when Θ^* is at most rank r, then we have $\sum_{j=r+1}^{m} \sigma_j(\Theta^*) = 0$, so the constraint (4.15) defines a cone. This cone completely excludes certain directions, and thus it is possible that the operator $\mathfrak{X}$, while failing RSC in a global sense, can satisfy it over the cone. Therefore, there is no need for a non-zero tolerance parameter δ in the exact low-rank case. In contrast, when Θ^* is only approximately low-rank, then the constraint (4.15) no longer defines a cone; rather, it includes an open ball around the origin. Thus, if $\mathfrak{X}$ fails RSC in a global sense, then it will also fail it under the constraint (4.15). The purpose of the additional constraint $|||\Delta|||_F \geq \delta$ is to eliminate the open ball centered at the origin, so that it is possible that $\mathfrak{X}$ satisfies RSC over $\mathbb{C}(r, \delta)$.

Let us now illustrate the consequences of Theorem 4.1 when the true matrix Θ^* has exactly rank r, in which case the approximation error term is zero. For the technical reasons mentioned above, it suffices to set $\delta = 0$ in the case of exact rank constraints, and we thus obtain the following result:

Corollary 4.1 (Exact low-rank recovery). *Suppose that $\Theta^* \in \Omega$ has rank r, and $\mathfrak{X}$ satisfies RSC with respect to $\mathbb{C}(r; 0)$. Then as long as $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_2/N$, any optimal solution $\widehat{\Theta}$ to the SDP (4.9) satisfies the bound*

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \frac{32\sqrt{r}\,\lambda_N}{\kappa(\mathfrak{X})}. \tag{4.16}$$

Like Theorem 4.1, Corollary 4.1 is a deterministic statement on the SDP error. It takes a much simpler form since when Θ^* is exactly low rank, then neither tolerance parameter δ nor the approximation term are required.

As a more delicate example, suppose instead that Θ^* is *nearly low-rank*, an assumption that we can formalize by requiring that its singular value sequence $\{\sigma_i(\Theta^*)\}_{i=1}^m$ decays quickly enough. In particular, for a parameter $q \in [0, 1]$ and a positive radius R_q, we define the set

$$\mathbb{B}_q(R_q) := \Big\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \sum_{i=1}^{m} |\sigma_i(\Theta)|^q \leq R_q\Big\}, \tag{4.17}$$

where $m = \min\{d_1, d_2\}$. Note that when $q = 0$, the set $\mathbb{B}_0(R_0)$ corresponds to the set of matrices with rank at most R_0.

Corollary 4.2 (Near low-rank recovery). *Suppose that $\Theta^* \in \mathbb{B}_q(R_q) \cap \Omega$, the regularization parameter is lower bounded as $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_2/N$, and the operator $\mathfrak{X}$ satisfies RSC with parameter $\kappa(\mathfrak{X}) \in (0, 1]$ over the set $\mathbb{C}(R_q\lambda_N^{-q}; \delta)$. Then any solution $\widehat{\Theta}$ to the SDP (4.9) satisfies*

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \max\Big\{\delta,\ 32\,\sqrt{R_q}\Big(\frac{\lambda_N}{\kappa(\mathfrak{X})}\Big)^{1-q/2}\Big\}. \tag{4.18}$$

Note that the error bound (4.18) reduces to the exact low rank case (4.16) when $q = 0$, and $\delta = 0$. The quantity $\lambda_N^{-q} R_q$ acts as the "effective rank" in this setting; as clarified by our proof in Section 4.4.2. This particular choice is designed to provide an optimal trade-off between the approximation and estimation error terms in Theorem 4.1. Since λ_N is chosen to decay to zero as the sample size N increases, this effective rank will increase, reflecting the fact that as we obtain more samples, we can afford to estimate more of the smaller singular values of the matrix Θ^*.

4.3.2 Comparison to related work

Past work by Lee and Bresler [79] provides stability results on minimizing the nuclear norm with a quadratic constraint, or equivalently, performing least-squares with nuclear norm constraints. Their results are based on the restricted isometry property (RIP), which is more restrictive than than the RSC condition given here; see Example 4.4 and Example 4.5 for concrete examples of operators $\mathfrak{X}$ that satisfy RSC but fail RIP. In our notation, their stability results guarantee that the error $|||\widehat{\Theta} - \Theta^*|||_F$ is bounded by a quantity proportional $t := \|y - \mathfrak{X}(\Theta^*)\|_2/\sqrt{N}$. Given the observation model (4.5) with a noise vector $\vec{\varepsilon}$ in which each entry is i.i.d., zero mean with variance ν^2, note that we have $t \approx \nu$ with high probability. Thus, although such a result guarantees stability, it does not guarantee consistency, since for any fixed noise variance $\nu^2 > 0$, the error bound does not tend to zero as the sample size N increases. In contrast, our bounds all depend on the noise and sample size via the regularization parameter, whose optimal choice is $\lambda_N^* = 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_2/N$. As will be clarified in Corollaries 4.3 through 4.5 to follow, for noise $\vec{\varepsilon}$ with variance ν and various choices of $\mathfrak{X}$, this regularization parameter satisfies the scaling $\lambda_N^* \asymp \nu\sqrt{\frac{d_1+d_2}{N}}$. Thus, our results guarantee consistency of the estimator, meaning that the error tends to zero as the sample size increases.

As previously noted, some concurrent work [30, 119] has also provided results on estimation of high-dimensional matrices in the noisy and statistical setting. Rohde and Tsybakov [119] derive results for estimating low-rank matrices based on a quadratic loss term regularized by the Schatten-q norm for $0 < q \leq 1$. Note that the the nuclear norm ($q = 1$) is a convex program, whereas the values $q \in (0, 1)$ provide analogs on concave regularized least squares [50] in the linear regression setting. They provide results on both multivariate regression and matrix completion; most closely related to our work are the results on multivariate regression, which we discuss at more length following Corollary 4.3 below. On the other hand, Candes and Plan [30] present error rates in the Frobenius norm for estimating approximately low-rank matrices under the compressed sensing model, and we discuss below the connection to our Corollary 4.5 for this particular observation model. A major difference between our work and this body of work lies in the assumptions imposed on the observation operator $\mathfrak{X}$. All of the papers [79, 30, 119] impose the restricted isometry property (RIP), which requires that all restricted singular values of $\mathfrak{X}$ very close to 1 (so that it is a near-

isometry). In contrast, we require only the restricted strong convexity (RSC) condition, which imposes only an arbitrarily small but positive lower bound on the operator. It is straightforward to construct operators $\mathfrak{X}$ that satisfy RSC while failing RIP, as we discuss in Examples 4.4 and Example 4.5 to follow.

4.3.3 Results for specific model classes

As stated, Corollaries 4.1 and 4.2 are fairly abstract in nature. More importantly, it is not immediately clear how the key underlying assumption—namely, the RSC condition—can be verified, since it is specified via subspaces that depend on Θ^*, which is itself the unknown quantity that we are trying to estimate. Nonetheless, we now show how, when specialized to more concrete models, these results yield concrete and readily interpretable results. As will be clear in the proofs of these results, each corollary requires overcoming two main technical obstacles: establishing that the appropriate form of the RSC property holds in a uniform sense (so that a priori knowledge of Θ^* is not required), and specifying an appropriate choice of the regularization parameter λ_N. Each of these two steps is non-trivial, requiring some random matrix theory, but the end results are simply stated upper bounds that hold with high probability.

We begin with the case of rank-constrained multivariate regression. As discussed earlier in Example 4.1, recall that we observe pairs $(Y_i, Z_i) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ linked by the linear model $Y_i = \Theta^* Z_i + W_i$, where $W_i \sim N(0, \nu^2 I_{d_1 \times d_1})$ is observation noise. Here we treat the case of *random design regression*, meaning that the covariates Z_i are modeled as random. In particular, in the following result, we assume that $Z_i \sim N(0, \Sigma)$, i.i.d. for some d_2-dimensional covariance matrix $\Sigma \succ 0$. Recalling that $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ denote the maximum and minimum eigenvalues respectively, we have:

Corollary 4.3 (Low-rank multivariate regression)**.** *Consider the random design multivariate regression model where $\Theta^* \in \mathbb{B}_q(R_q) \cap \Omega$. There are universal constants $\{c_i, i = 1, 2, 3\}$ such that if we solve the SDP (4.9) with regularization parameter $\lambda_N = 10 \frac{\nu}{d_1} \sqrt{\sigma_{\max}(\Sigma)} \sqrt{\frac{(d_1+d_2)}{n}}$, we have*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_1 \left(\frac{\nu^2 \sigma_{\max}(\Sigma)}{\sigma_{\min}^2(\Sigma)} \right)^{1-q/2} R_q \left(\frac{d_1 + d_2}{n} \right)^{1-q/2} \tag{4.19}$$

with probability greater than $1 - c_2 \exp(-c_3(d_1 + d_2))$.

Remarks: Corollary 4.3 takes a particularly simple form when $\Sigma = I_{d_2 \times d_2}$: then there exists a constant c_1' such that $|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_1' \nu^{2-q} R_q \left(\frac{d_1+d_2}{n}\right)^{1-q/2}$. When Θ^* is exactly low rank—that is, $q = 0$, and Θ^* has rank $r = R_0$—this simplifies even further to

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_1' \frac{\nu^2 \, r \, (d_1 + d_2)}{n}.$$

The scaling in this error bound is easily interpretable: naturally, the squared error is proportional to the noise variance ν^2, and the quantity $r(d_1 + d_2)$ counts the number of degrees of freedom of a $d_1 \times d_2$ matrix with rank r. Note that if we did not impose any constraints on Θ^*, then since a $d_1 \times d_2$ matrix has a total of $d_1 d_2$ free parameters, we would expect at best[1] to obtain rates of the order $|||\widehat{\Theta} - \Theta^*|||_F^2 = \Omega(\frac{\nu^2 \, d_1 \, d_2}{n})$. Note that when Θ^* is low rank—in particular, when $r \ll \min\{d_1, d_2\}$—then the nuclear norm estimator achieves substantially faster rates.[2]

It is worth comparing this corollary to a result on multivariate regression due to Rohde and Tsybakov [119]. Their result applies to exactly low-rank matrices (say with rank r), but provides bounds on general Schatten norms (including the Frobenius norm). In this case, it provides a comparable rate when we make the setting $q = 0$ and $R_0 = r$ in the bound (4.19), namely showing that we require roughly $n \approx r\,(d_1 + d_2)$ samples, corresponding to the number of degrees of freedom. A significant difference lies in the conditions imposed on the design matrices: whereas their result is derived under RIP conditions on the design matrices, we require only the milder RSC condition. The following example illustrates the distinction for this model.

Example 4.4 (Failure of RIP for multivariate regression)**.** Under the random design model for multivariate regression, we have

$$F(\Theta) := \frac{\mathbb{E}[\|\mathfrak{X}(\Theta)\|_2^2]}{n|||\Theta|||_F^2} \;=\; \frac{\sum_{j=1}^{d_2} \|\sqrt{\Sigma}\Theta_j\|_2^2}{|||\Theta|||_F^2}, \tag{4.20}$$

where Θ_j is the j^{th} row of Θ. In order for RIP to hold, it is necessary that quantity $F(\Theta)$ is extremely close to 1—certainly less than two—for all low-rank matrices. We now show that this cannot hold unless Σ has a small condition number. Let $v \in \mathbb{R}^{d_2}$ and $v' \in \mathbb{R}^{d_2}$ denote the minimum and maximum eigenvectors of Σ. By setting $\Theta = e_1 v^T$, we obtain a rank one matrix for which $F(\Theta) = \sigma_{\min}(\Sigma)$, and similarly, setting $\Theta' = e_1 (v')^T$ yields another rank one matrix for which $F(\Theta') = \sigma_{\max}(\Sigma)$. The preceding discussion applies to the average $\mathbb{E}[\|\mathfrak{X}(\Theta)\|_2^2]/n$, but since the individual matrices matrices X_i are i.i.d. and Gaussian, we have

$$\frac{\|\mathfrak{X}(\Theta)\|_2^2}{n} = \frac{1}{n}\sum_{i=1}^{n} \langle\!\langle X_i, \, \Theta \rangle\!\rangle^2 \;\leq 2F(\Theta) \;=\; 2\,\sigma_{\min}(\Sigma)$$

[1] To clarify our use of sample size, we can either view the multivariate regression model as consisting of n samples with a constant SNR, or as N samples with SNR of order $1/d_1$. We adopt the former interpretation here.

[2] We also note that as stated, the result requires that $(d_1 + d_2)$ tend to infinity in order for the claim to hold with high probability. Although such high-dimensional scaling is the primary focus of this chapter, we note that for application to the classical setting of fixed (d_1, d_2), the same statement (with different constants) holds with $d_1 + d_2$ replaced by $\log n$.

with high probability, using χ^2-tail bounds. Similarly, $\|\mathfrak{X}(\Theta')\|_2^2/n \geq (1/2)\sigma_{\max}(\Sigma)$ with high probability. Thus, we have exhibited a pair of rank one matrices with $|||\Theta|||_F = |||\Theta'|||_F = 1$ for which

$$\frac{\|\mathfrak{X}(\Theta')\|_2^2}{\|\mathfrak{X}(\Theta)\|_2^2} \geq \frac{1}{4}\frac{\sigma_{\max}(\Sigma)}{\sigma_{\min}(\Sigma)}.$$

Consequently, unless $\sigma_{\max}(\Sigma)/\sigma_{\min}(\Sigma) \leq 64$, it is not possible for RIP to hold with constant $\delta \leq 1/2$. In contrast, as our results show, the RSC will hold w.h.p. whenever $\sigma_{\min}(\Sigma) > 0$, and the error is allowed to scale with the ratio $\sigma_{\max}(\Sigma)/\sigma_{\min}(\Sigma)$.

Next we turn to the case of estimating the system matrix Θ^* of an autoregressive (AR) model, as discussed in Example 4.2.

Corollary 4.4 (Autoregressive models)**.** *Suppose that we are given n samples $\{Z_t\}_{t=1}^n$ from a m-dimensional autoregressive process (4.2) that is stationary, based on a system matrix that is stable ($|||\Theta^*|||_2 \leq \gamma < 1$), and approximately low-rank ($\Theta^* \in \mathbb{B}_q(R_q) \cap \Omega$). Then there are universal constants $\{c_i, i = 1, 2, 3\}$ such that if we solve the SDP (4.9) with regularization parameter $\lambda_N = \frac{2c_0 |||\Sigma|||_2}{m(1-\gamma)}\sqrt{\frac{m}{n}}$, then any solution $\widehat{\Theta}$ satisfies*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_1 \left[\frac{\sigma_{\max}^2(\Sigma)}{\sigma_{\min}^2(\Sigma)(1-\gamma)^2}\right]^{1-q/2} R_q \left(\frac{m}{n}\right)^{1-q/2} \tag{4.21}$$

with probability greater than $1 - c_2\exp(-c_3 m)$.

Remarks: Like Corollary 4.3, the result as stated requires that the matrix dimension m tends to infinity, but the same bounds hold with m replaced by $\log n$, yielding results suitable for classical (fixed dimension) scaling. Second, the factor $(m/n)^{1-q/2}$, like the analogous term[3] in Corollary 4.3, shows that faster rates are obtained if Θ^* can be well-approximated by a low rank matrix, namely for choices of the parameter $q \in [0, 1]$ that are closer to zero. Indeed, in the limit $q = 0$, we again reduce to the case of an exact rank constraint $r = R_0$, and the corresponding squared error scales as rm/n. In contrast to the case of multivariate regression, the error bound (4.21) also depends on the upper bound $|||\Theta^*|||_2 = \gamma < 1$ on the operator norm of the system matrix Θ^*. Such dependence is to be expected since the quantity γ controls the (in)stability and mixing rate of the autoregressive process. As clarified in the proof, the dependence of the sampling in the AR model also presents some technical challenges not present in the setting of multivariate regression.

Finally, we turn to the analysis of the compressed sensing model for matrix recovery, as initially described in Example 4.3. Although standard compressed sensing is based on

[3]The term in Corollary 4.3 has a factor $d_1 + d_2$, since the matrix in that case could be non-square in general.

observation matrices X_i with i.i.d. elements, here we consider a more general model that allows for dependence between the entries of X_i. First defining the quantity $M = d_1\, d_2$, we use $\mathrm{vec}(X_i) \in \mathbb{R}^M$ to denote the vectorized form of the $d_1 \times d_2$ matrix X_i. Given a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{M \times M}$, we say that the observation matrix X_i is sampled from the Σ-ensemble if $\mathrm{vec}(X_i) \sim N(0, \Sigma)$. Finally, we define the quantity

$$\zeta_{\mathrm{mat}}(\Sigma) := \sup_{\|u\|_2=1, \|v\|_2=1} \mathrm{var}(u^T X v), \tag{4.22}$$

where the random matrix $X \in \mathbb{R}^{d_1 \times d_2}$ is sampled from the Σ-ensemble. In the special case $\Sigma = I$, corresponding to the usual compressed sensing model, we have $\zeta_{\mathrm{mat}}(I) = 1$.

The following result applies to any observation model in which the noise vector $\vec{\varepsilon} \in \mathbb{R}^N$ satisfies the bound $\|\vec{\varepsilon}\|_2 \leq 2\nu\sqrt{N}$ for some constant ν. This assumption that holds for any bounded noise, and also holds with high probability for any random noise vector with sub-Gaussian entries with parameter ν. (The simplest example is that of Gaussian noise $N(0, \nu^2)$.)

Corollary 4.5 (Compressed sensing with dependent sampling)**.** *Suppose that the matrices $\{X_i\}_{i=1}^N$ are drawn i.i.d. from the Σ-Gaussian ensemble, and that the unknown matrix $\Theta^* \in \mathbb{B}_q(R_q) \cap \Omega$ for some $q \in (0, 1]$. Then there are universal constants c_i such that for a sample size $N > c_1 \zeta_{mat}(\Sigma)\; R_q^{1-q/2}\; (d_1 + d_2)$, any solution $\widehat{\Theta}$ to the SDP (4.9) with regularization parameter $\lambda_N = c_0 \sqrt{\zeta_{mat}(\Sigma)}\nu \sqrt{\frac{d_1+d_2}{N}}$ satisfies the bound*

$$\|\widehat{\Theta} - \Theta^*\|_F^2 \leq c_2 R_q \left(\frac{(\nu^2 \vee 1) \frac{\zeta_{mat}(\Sigma)}{\sigma^2_{\min}(\Sigma)}\; (d_1 + d_2)}{N} \right)^{1-\frac{q}{2}} \tag{4.23}$$

with probability greater than $1 - c_3 \exp(-c_4(d_1 + d_2))$. In the special case $q = 0$ and Θ^* of rank r, we have

$$\|\widehat{\Theta} - \Theta^*\|_F^2 \leq c_2 \frac{\zeta_{mat}(\Sigma)\nu^2}{\sigma^2_{\min}(\Sigma)}\; \frac{r\,(d_1 + d_2)}{N}. \tag{4.24}$$

The central challenge in proving this result is in proving an appropriate form of the RSC property. The following result on the random operator $\mathfrak{X}$ may be of independent interest here:

Proposition 4.1. *Consider the random operator $\mathfrak{X} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^N$ formed by sampling from the Σ-ensemble. Then it satisfies*

$$\frac{\|\mathfrak{X}(\Theta)\|_2}{\sqrt{N}} \geq c_1 \|\sqrt{\Sigma}\,\mathrm{vec}(\Theta)\|_2 - c_2 \sqrt{\zeta_{mat}(\Sigma)}\; \left(\sqrt{\frac{d_1}{N}} + \sqrt{\frac{d_2}{N}} \right) |||\Theta|||_{\mathrm{nuc}} \qquad \text{for all } \Theta \in \mathbb{R}^{d_1 \times d_2} \tag{4.25}$$

with probability at least $1 - 2\exp(-N/32)$.

The proof of this result, provided in Appendix B.5, makes use of the Gordon-Slepian inequalities for Gaussian processes, and concentration of measure. As we show in Section 4.4.5, it implies the form of the RSC property needed to establish Corollary 4.5.

In concurrent work, Candes and Plan [30] derived a result similar to Corollary 4.5 for the compressed sensing observation model. Their result applies to matrices with i.i.d. elements with sub-Gaussian tail behavior. While the analysis given here is specific to Gaussian random matrices, it allows for general dependence among the entries. Their result applies only under certain restrictions on the sample size relative to matrix dimension and rank, whereas our result holds more generally without these extra conditions. Moreover, their proof relies on an application of RIP, which is in general more restrictive than the RSC condition used in our analysis. The following example provides a concrete illustration of a matrix family where the restricted isometry constants are unbounded as the rank r grows, but RSC still holds.

Example 4.5 (RSC holds when RIP violated)**.** Here we consider a family of random operators $\mathfrak{X}$ for which RSC holds with high probability, while RIP fails. Consider generating an i.i.d. collection of design matrices $X_i \in \mathbb{R}^{m \times m}$, each of the form

$$X_i = z_i I_{m \times m} + G_i, \qquad \text{for } i = 1, 2, \ldots, N, \tag{4.26}$$

where $z_i \sim N(0,1)$ and $G_i \in \mathbb{R}^{m \times m}$ is a standard Gaussian random matrix, independent of z_i. Note that we have $\operatorname{vec}(X_i) \sim N(0, \Sigma)$, where the $m^2 \times m^2$ covariance matrix has the form

$$\Sigma = \operatorname{vec}(I_{m \times m}) \operatorname{vec}(I_{m \times m})^T + I_{m^2 \times m^2}. \tag{4.27}$$

Let us compute the quantity $\zeta_{\text{mat}}(\Sigma) = \sup_{\substack{\|u\|_2 = 1 \\ \|v\|_2 = 1}} \operatorname{var}(u^T X v)$. By the independence of z and G in the model (4.26), we have

$$\zeta_{\text{mat}}(\Sigma) \leq \operatorname{var}(z) \sup_{u \in S^{d_1 - 1}, v \in S^{d_2 - 1}} u^T v + \sup_{u \in S^{d_1 - 1}, v \in S^{d_2 - 1}} \operatorname{var}(u^T G v) \leq 2.$$

Letting $\mathfrak{X}$ be the associated random operator, we observe that for any $\Theta \in \mathbb{R}^{m \times m}$, the independence of z_i and G_i implies that

$$\mathbb{E}\left[\frac{\|\mathfrak{X}(\Theta)\|_2^2}{N}\right] = \|\sqrt{\Sigma} \operatorname{vec}(\Theta)\|_2^2 = \operatorname{trace}(\Theta)^2 + |\!|\!|\Theta|\!|\!|_F^2 \geq |\!|\!|\Theta|\!|\!|_F^2.$$

Consequently, Proposition 4.1 implies that

$$\frac{\|\mathfrak{X}(\Theta)\|_2}{\sqrt{N}} \geq \frac{1}{4} |\!|\!|\Theta|\!|\!|_F - 48 \sqrt{\frac{m}{N}} |\!|\!|\Theta|\!|\!|_{\text{nuc}} \qquad \text{for all } \Theta \in \mathbb{R}^{m \times m}, \tag{4.28}$$

with high probability. As mentioned previously, we show in Section 4.4.5 how this type of lower bound implies the RSC condition needed for our results.

On the other hand, the random operator can never satisfy RIP (with the rank r increasing), as the following calculation shows. In this context, RIP requires that bounds of the form

$$\frac{\|\mathfrak{X}(\Theta)\|_2^2}{N\,|||\Theta|||_F^2} \in [1-\delta, 1+\delta] \qquad \text{for all } \Theta \text{ with rank at most } r,$$

where $\delta \in (0,1)$ is a constant independent of r. Note that the bound (4.28) implies that a *lower bound* of this form holds as long as $N = \Omega(rm)$. Moreover, this lower bound cannot be substantially sharpened, since the trace term plays no role for matrices with zero diagonals.

We now show that no such upper bound can ever hold. For a rank $1 \leq r < m$, consider the $m \times m$ matrix of the form

$$\Gamma := \begin{bmatrix} I_{r\times r}/\sqrt{r} & 0_{r\times(m-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(m-r)} \end{bmatrix}.$$

By construction, we have $|||\Gamma|||_F = 1$ and $\operatorname{trace}(\Gamma) = \sqrt{r}$. Consequently, we have

$$\mathbb{E}\big[\frac{\|\mathfrak{X}(\Gamma)\|_2^2}{N}\big] = \operatorname{trace}(\Gamma)^2 + |||\Gamma|||_F^2 \;=\; r+1.$$

The independence of the matrices X_i implies that $\frac{\|\mathfrak{X}(\Gamma)\|_2^2}{N}$ is sharply concentrated around this expected value, so that we conclude that

$$\frac{\|\mathfrak{X}(\Gamma)\|_2^2}{N\,|||\Gamma|||_F^2} \geq \frac{1}{2}\big[1+r\big],$$

with high probability, showing that RIP cannot hold with upper and lower bounds of the same order.

Finally, we note that Proposition 4.1 also implies an interesting property of the null space of the operator $\mathfrak{X}$, one which can be used to establish a corollary about recovery of low-rank matrices when the observations are noiseless. In particular, suppose that we are given the noiseless observations $y_i = \langle\!\langle X_i,\, \Theta^* \rangle\!\rangle$ for $i = 1, \ldots, N$, and that we try to recover the unknown matrix Θ^* by solving the SDP

$$\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} |||\Theta|||_{\text{nuc}} \quad \text{such that } \langle\!\langle X_i,\, \Theta \rangle\!\rangle = y_i \text{ for all } i = 1, \ldots, N. \tag{4.29}$$

This recovery procedure was studied by Recht and colleagues [117, 116] in the special case that X_i is formed of i.i.d. $N(0,1)$ entries. Proposition 4.1 allows us to obtain a sharp result on recovery using this method for Gaussian matrices with general dependencies.

Corollary 4.6 (Exact recovery with dependent sampling). *Suppose that the matrices* $\{X_i\}_{i=1}^N$ *are drawn i.i.d. from the* Σ*-Gaussian ensemble, and that* $\Theta^* \in \Omega$ *has rank* r*. Given* $N > c_0 \zeta_{mat}(\Sigma) r(d_1 + d_2)$ *noiseless samples, then with probability at least* $1 - 2\exp(-N/32)$*, the SDP* (4.29) *recovers the matrix* Θ^* *exactly.*

This result removes some extra logarithmic factors that were included in initial work [117] and provides the appropriate analog to compressed sensing results for sparse vectors [44, 31]. Note that (like in most of our results) we have made little effort to obtain good constants in this result: the important property is that the sample size N scales linearly in both r and $d_1 + d_2$. We refer the reader to Recht et al. [116], who study the standard Gaussian model under the scaling $r = \Theta(m)$ and obtain sharp results on the constants.

4.4 Proofs

We now turn to the proofs of Theorem 4.1, and Corollaries 4.1 through 4.6. In each case, we provide the primary steps in the main text, with more technical details stated as lemmas and proved in the Appendix.

4.4.1 Proof of Theorem 4.1

By the optimality of $\widehat{\Theta}$ and feasibility of Θ^* for the SDP (4.9), we have

$$\frac{1}{2N}\|\vec{y} - \mathfrak{X}(\widehat{\Theta})\|_2^2 + \lambda_N |||\widehat{\Theta}|||_{\text{nuc}} \leq \frac{1}{2N}\|\vec{y} - \mathfrak{X}(\Theta^*)\|_2^2 + \lambda_N |||\Theta^*|||_{\text{nuc}}.$$

Defining the error matrix $\Delta = \Theta^* - \widehat{\Theta}$ and performing some algebra yields the inequality

$$\frac{1}{2N}\|\mathfrak{X}(\Delta)\|_2^2 \leq \frac{1}{N}\langle \vec{\varepsilon}, \mathfrak{X}(\Delta)\rangle + \lambda_N \big\{ |||\widehat{\Theta} + \Delta|||_{\text{nuc}} - |||\widehat{\Theta}|||_{\text{nuc}} \big\}. \tag{4.30}$$

By definition of the adjoint and Hölder's inequality, we have

$$\frac{1}{N}\big|\langle \vec{\varepsilon}, \mathfrak{X}(\Delta)\rangle\big| = \frac{1}{N}\big|\langle \mathfrak{X}^*(\vec{\varepsilon}), \Delta\rangle\big| \leq \frac{1}{N} |||\mathfrak{X}^*(\vec{\varepsilon})|||_2 \, |||\Delta|||_{\text{nuc}}. \tag{4.31}$$

By the triangle inequality, we have $|||\widehat{\Theta} + \Delta|||_{\text{nuc}} - |||\widehat{\Theta}|||_{\text{nuc}} \leq |||\Delta|||_{\text{nuc}}$. Substituting this inequality and the bound (4.31) into the inequality (4.30) yields

$$\frac{1}{2N}\|\mathfrak{X}(\Delta)\|_2^2 \leq \frac{1}{N} |||\mathfrak{X}^*(\vec{\varepsilon})|||_2 |||\Delta|||_{\text{nuc}} + \lambda_N |||\Delta|||_{\text{nuc}} \leq 2\lambda_N |||\Delta|||_{\text{nuc}},$$

where the second inequality makes use of our choice $\lambda_N \geq \frac{2}{N} |||\mathfrak{X}^*(\vec{\varepsilon})|||_2$.

It remains to lower bound the term on the left-hand side, while upper bounding the quantity $|||\Delta|||_{\text{nuc}}$ on the right-hand side. The following technical result allows us to do so. Recall our earlier definition (4.11) of the sets $\mathcal{M}$ and $\overline{\mathcal{M}}^{\perp}$ associated with a given subspace pair.

Lemma 4.1. *Let $U^r \in \mathbb{R}^{d_1 \times r}$ and $V^r \in \mathbb{R}^{d_2 \times r}$ be matrices consisting of the top r left and right (respectively) singular vectors of Θ^*. Then there exists a matrix decomposition $\Delta = \Delta' + \Delta''$ of the error Δ such that*

(a) The matrix Δ' satisfies the constraint $\operatorname{rank}(\Delta') \leq 2r$, and

(b) If $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_2/N$, then the nuclear norm of Δ'' is bounded as

$$|||\Delta''|||_{\text{nuc}} \;\leq\; 3|||\Delta'|||_{\text{nuc}} + 4\sum_{j=r+1}^{m} \sigma_j(\Theta^*) \tag{4.32}$$

Compare the above result to Lemma 3.1. In order to establish the above lemma we exploit the decomposability of the nuclear norm (see Example 3.3) and apply the results from Lemma 3.1. See Appendix B.1 for the detailed proof of the above claim. Using Lemma 4.1, we can complete the proof of the theorem. In particular, from the bound (4.32) and the RSC assumption, we find that for $|||\Delta|||_F \geq \delta$, we have

$$\frac{1}{2N}\|\mathfrak{X}(\Delta)\|_2^2 \geq \kappa(\mathfrak{X})\,|||\Delta|||_F^2.$$

Using the triangle inequality together with inequality (4.32), we obtain

$$|||\Delta|||_{\text{nuc}} \leq |||\Delta'|||_{\text{nuc}} + |||\Delta''|||_{\text{nuc}} \;\leq\; 4|||\Delta'|||_{\text{nuc}} + 4\sum_{j=r+1}^{m} \sigma_j(\Theta^*).$$

From the rank constraint in Lemma 4.1(a), we have $|||\Delta'|||_{\text{nuc}} \leq \sqrt{2r}|||\Delta'|||_F$. Putting together the pieces, we find that either $|||\Delta|||_F \leq \delta$, or

$$\kappa(\mathfrak{X})\,|||\Delta|||_F^2 \leq \max\big\{32\lambda_N\sqrt{r}\,|||\Delta|||_F,\; 16\,\lambda_N \sum_{j=r+1}^{m} \sigma_j(\Theta^*)\big\},$$

which implies that

$$|||\Delta|||_F \leq \max\left\{\delta,\; \frac{32\lambda_N\sqrt{r}}{\kappa(\mathfrak{X})},\; \left(\frac{16\,\lambda_N\,\sum_{j=r+1}^{m}\sigma_j(\Theta^*)}{\kappa(\mathfrak{X})}\right)^{1/2}\right\},$$

as claimed.

4.4.2 Proof of Corollary 4.2

Let $d = \min\{d_1, d_2\}$. In this case, we consider the singular value decomposition $\Theta^* = UDV^T$, where $U \in \mathbb{R}^{d_1 \times d}$ and $V \in \mathbb{R}^{d_2 \times d}$ are orthogonal, and we assume that D is diagonal with the

singular values in non-increasing order $\sigma_1(\Theta^*) \geq \sigma_2(\Theta^*) \geq \ldots \sigma_d(\Theta^*) \geq 0$. For a parameter $\tau > 0$ to be chosen, we define

$$K := \big\{ i \in \{1, 2, \ldots, d\} \mid \sigma_i(\Theta^*) > \tau \big\},$$

and we let U^K (respectively V^K) denote the $d_1 \times |K|$ (respectively the $d_2 \times |K|$) orthogonal matrix consisting of the first $|K|$ columns of U (respectively V). With this choice, the matrix $\Theta^*_{K^c} := \Pi_{\underset{\mathcal{M}}{\perp}}(\Theta^*)$ has rank at most $d - |K|$, with singular values $\{\sigma_i(\Theta^*), i \in K^c\}$. Moreover, since $\sigma_i(\Theta^*) \leq \tau$ for all $i \in K^c$, we have

$$|||\Theta^*_{K^c}|||_{\text{nuc}} = \tau \sum_{i=|K|+1}^{d} \frac{\sigma_i(\Theta^*)}{\tau} \leq \tau \sum_{i=|K|+1}^{d} \Big(\frac{\sigma_i(\Theta^*)}{\tau}\Big)^q \leq \tau^{1-q} R_q.$$

On the other hand, we also have $R_q \geq \sum_{i=1}^{d} |\sigma_i(\Theta^*)|^q \geq |K|\, \tau^q$, which implies that $|K| \leq \tau^{-q} R_q$. From the general error bound with $r = |K|$, we obtain

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \max\Big\{\delta, \ \frac{32\lambda_N \sqrt{R_q}\, \tau^{-q/2}}{\kappa(\mathfrak{X})}, \ \Big[\frac{16\, \lambda_N\, \tau^{1-q} R_q}{\kappa(\mathfrak{X})}\Big]^{1/2}\Big\},$$

Setting $\tau = \lambda_N/\kappa$ yields that

$$\begin{aligned} |||\widehat{\Theta} - \Theta^*|||_F &\leq \max\Big\{\delta, \ \frac{32\lambda_N^{1-q/2} \sqrt{R_q}}{\kappa^{1-q/2}}, \ \Big[\frac{16\, \lambda_N^{2-q} R_q}{\kappa^{2-q}}\Big]^{1/2}\Big\} \\ &= \max\Big\{\delta, \ 32\, \sqrt{R_q}\Big(\frac{\lambda_N}{\kappa(\mathfrak{X})}\Big)^{1-q/2}\Big\}, \end{aligned}$$

as claimed.

4.4.3 Proof of Corollary 4.3

For the proof of this corollary, we adopt the following notation. We first define the three matrices

$$X = \begin{bmatrix} Z_1^T \\ Z_2^T \\ \cdots \\ Z_n^T \end{bmatrix} \in \mathbb{R}^{n \times d_2}, \quad Y = \begin{bmatrix} Y_1^T \\ Y_2^T \\ \cdots \\ Y_n^T \end{bmatrix} \in \mathbb{R}^{n \times d_1}, \quad \text{and} \quad W = \begin{bmatrix} W_1^T \\ W_2^T \\ \cdots \\ W_n^T \end{bmatrix} \in \mathbb{R}^{n \times d_1}. \tag{4.33}$$

With this notation and using the relation $N = nd_1$, the SDP objective function (4.9) can be written as $\frac{1}{d_1}\big\{\frac{1}{2n}|||Y - X\Theta^T|||_F^2 + \lambda_n |||\Theta|||_{\text{nuc}}\big\}$, where we have defined $\lambda_n = \lambda_N\, d_1$.

In order to establish the RSC property for this model, some algebra shows that we need to establish a lower bound on the quantity

$$\frac{1}{2n}\|X\Delta\|_F^2 = \frac{1}{2n}\sum_{j=1}^{m}\|(X\Delta)_j\|_2^2 \;\geq\; \frac{\sigma_{\min}(X^TX)}{2n}\,\|\Delta\|_F^2,$$

where $\sigma_{\min}$ denotes the minimum eigenvalue. The following lemma follows by adapting known concentration results for random matrices (see the paper [145] for details):

Lemma 4.2. *Let $X \in \mathbb{R}^{n\times m}$ be a random matrix with i.i.d. rows sampled from a m-variate $N(0,\Sigma)$ distribution. Then for $n \geq 2m$, we have*

$$\mathbb{P}\left[\sigma_{\min}\Big(\frac{1}{n}X^TX\Big) \geq \frac{\sigma_{\min}(\Sigma)}{9},\; \sigma_{\max}\Big(\frac{1}{n}X^TX\Big) \leq 9\sigma_{\max}(\Sigma)\right] \;\geq\; 1 - 4\exp(-n/2).$$

As a consequence, we have $\frac{\sigma_{\min}(X^TX)}{2n} \geq \frac{\sigma_{\min}(\Sigma)}{18}$ with probability at least $1-4\exp(-n)$ for all $n \geq 2m$, which establishes that the RSC property holds with $\kappa(\mathfrak{X}) = \frac{1}{20d_1}\sigma_{\min}(\Sigma)$.

Next we need to upper bound the quantity $\|\mathfrak{X}^*(\vec{\varepsilon})\|_2$ for this model, so as to verify that the stated choice for λ_N is valid. Following some algebra, we find that

$$\frac{1}{n}\,\|\mathfrak{X}^*(\vec{\varepsilon})\|_2 = \frac{1}{n}\|X^TW\|_2.$$

The following lemma is proved in Appendix B.3:

Lemma 4.3. *There are constants $c_i > 0$ such that*

$$\mathbb{P}\left[\Big|\frac{1}{n}\|X^TW\|_2\Big| \geq 5\nu\sqrt{\sigma_{\max}(\Sigma)}\,\sqrt{\frac{d_1+d_2}{n}}\right] \leq c_1\exp(-c_2(d_1+d_2)). \tag{4.34}$$

Using these two lemmas, we can complete the proof of Corollary 4.3. First, recalling the scaling $N = d_1 n$, we see that Lemma 4.3 implies that the choice $\lambda_n = 10\nu\sqrt{\sigma_{\max}(\Sigma)}\sqrt{\frac{d_1+d_2}{n}}$ satisfies the conditions of Corollary 4.2 with high probability. Lemma 4.2 shows that the RSC property holds with $\kappa(\mathfrak{X}) = \sigma_{\min}(\Sigma)/(20d_1)$, again with high probability. Consequently, Corollary 4.2 implies that

$$\begin{aligned}\|\widehat{\Theta} - \Theta^*\|_F^2 &\leq 32^2 R_q\left(10\nu\sqrt{\sigma_{\max}(\Sigma)}\sqrt{\frac{d_1+d_2}{n}}\;\;\frac{20}{\sigma_{\min}(\Sigma)}\right)^{2-q}\\ &= c_1\left(\frac{\nu^2\sigma_{\max}(\Sigma)}{\sigma^2_{\min}(\Sigma)}\right)^{1-q/2} R_q\left(\frac{d_1+d_2}{n}\right)^{1-q/2}\end{aligned}$$

with probability greater than $1 - c_2\exp(-c_3(d_1+d_2))$, as claimed.

4.4.4 Proof of Corollary 4.4

For the proof of this corollary, we adopt the notation

$$X = \begin{bmatrix} Z_1^T \\ Z_2^T \\ \cdots \\ Z_n^T \end{bmatrix} \in \mathbb{R}^{n \times m}, \quad \text{and} \quad Y = \begin{bmatrix} Z_2^T \\ Z_2^T \\ \cdots \\ Z_{n+1}^T \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

Finally, we let $W \in \mathbb{R}^{n \times m}$ be a matrix where each row is sampled i.i.d. from the $N(0, C)$ distribution corresponding to the innovations noise driving the VAR process. With this notation and using the relation $N = nm$, the SDP objective function (4.9) can be written as $\frac{1}{m}\{\frac{1}{2n}|||Y - X\Theta^T|||_F^2 + \lambda_n |||\Theta|||_{\text{nuc}}\}$, where we have defined $\lambda_n = \lambda_N m$. At a high level, the proof of this corollary is similar to that of Corollary 4.3, in that we use random matrix theory to establish the required RSC property, and to justify the choice of λ_n, or equivalently λ_N. However, it is considerably more challenging, due to the dependence in the rows of the random matrices, and the cross-dependence between the two matrices X and W (which were independent in the setting of multivariate regression).

The following lemma provides the lower bound needed to establish RSC for the autoregressive model:

Lemma 4.4. *The eigenspectrum of the matrix $X^T X/n$ is well-controlled in terms of the stationary covariance matrix: in particular, as long as $n > c_3 m$, we have*

$$\sigma_{\max}\Big(\Big(\frac{1}{n}X^T X\Big)\Big) \overset{(a)}{\leq} \frac{24\,\sigma_{\max}(\Sigma)}{1-\gamma}, \quad \text{and} \quad \sigma_{\min}\Big(\Big(\frac{1}{n}X^T X\Big)\Big) \overset{(b)}{\geq} \frac{\sigma_{\min}(\Sigma)}{4}, \tag{4.35}$$

both with probability greater than $1 - 2c_1 \exp(-c_2\, m)$.

Thus, from the bound (4.35)(b), we see with the high probability, the RSC property holds with $\kappa(\mathfrak{X}) = \sigma_{\min}(\Sigma)/(4d_2)$ as long as $n > c_3 m$.

As before, in order to verify the choice of λ_n, we need to control the quantity $\frac{1}{n}|||X^T W|||_2$. The following inequality, proved in Appendix B.4.2, yields a suitable upper bound:

Lemma 4.5. *There exist constants $c_i > 0$, independent of n, m, Σ etc. such that*

$$\mathbb{P}\big[\frac{1}{n}|||X^T W|||_2 \geq \frac{c_0\,|||\Sigma|||_2}{1-\gamma}\sqrt{\frac{m}{n}}\big] \leq c_2 \exp(-c_3 m). \tag{4.36}$$

From Lemma 4.5, we see that it suffices to choose $\lambda_n = \frac{2c_0\,|||\Sigma|||_2}{1-\gamma}\sqrt{\frac{m}{n}}$. With this choice, Corollary 4.2 of Theorem 4.1 yields that

$$|||\Theta - \Theta^*|||_F^2 \leq c_1\, R_q \left[\frac{\sigma_{\max}(\Sigma)}{\sigma_{\min}(\Sigma)\,(1-\gamma)}\right]^{2-q} \left(\frac{m}{n}\right)^{1-q/2}$$

with probability greater than $1 - c_2 \exp(-c_3 m)$, as claimed.

4.4.5 Proof of Corollary 4.5

Recall that for this model, the observations are of the form $y_i = \langle\langle X_i, \Theta^* \rangle\rangle + \varepsilon_i$, where $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ is the unknown matrix, and $\{\varepsilon_i\}_{i=1}^N$ is an associated noise sequence.

We now show how Proposition 4.1 implies the RSC property with an appropriate tolerance parameter $\delta > 0$ to be defined. Observe that the bound (4.25) implies that for any $\Delta \in \mathbb{C}$, we have

$$\begin{aligned} \frac{\|\mathfrak{X}(\Delta)\|_2}{\sqrt{N}} &\geq \frac{\sqrt{\sigma_{\min}(\Sigma)}}{4} |||\Delta|||_F - c_2 \sqrt{\zeta_{\text{mat}}(\Sigma)} \left(\sqrt{\frac{d_1}{N}} + \sqrt{\frac{d_2}{N}} \right) |||\Delta|||_{\text{nuc}} \\ &= \frac{\sqrt{\sigma_{\min}(\Sigma)}}{4} \left\{ |||\Delta|||_F - \underbrace{\frac{48\sqrt{\zeta_{\text{mat}}(\Sigma)}}{\sqrt{\sigma_{\min}(\Sigma)}} \left(\sqrt{\frac{d_1}{N}} + \sqrt{\frac{d_2}{N}} \right)}_{\tau} |||\Delta|||_{\text{nuc}} \right\}, \end{aligned} \tag{4.37}$$

where we have defined the quantity $\tau > 0$. Following the arguments used in the proofs of Theorem 4.1 and Corollary 4.2, we find that

$$|||\Delta|||_{\text{nuc}} \leq 4|||\Delta'|||_{\text{nuc}} + 4 \sum_{j=r+1}^{m} \sigma_j(\Theta^*) \leq 4\sqrt{2R_q \tau^{-q}} |||\Delta'|||_F + 4 R_q \tau^{1-q}. \tag{4.38}$$

Note that this corresponds to truncating the matrices at effective rank $r = 2R_q\tau^{-q}$. Combining this bound with the definition of τ, we obtain

$$\tau |||\Delta|||_{\text{nuc}} \leq 4\sqrt{2R_q}\tau^{1-q/2} |||\Delta'|||_F + 4 R_q \tau^{2-q} \leq 4\sqrt{2R_q}\tau^{1-q/2} |||\Delta|||_F + 4 R_q \tau^{2-q}.$$

Substituting this bound into equation (4.37) yields

$$\frac{\|\mathfrak{X}(\Delta)\|_2}{\sqrt{N}} \geq \frac{\sqrt{\sigma_{\min}(\Sigma)}}{4} \left\{ |||\Delta|||_F - 4\sqrt{2R_q}\tau^{1-q/2} |||\Delta'|||_F - 4 R_q \tau^{2-q} \right\}.$$

As long $N > c_0 R_q^{2/(2-q)} \frac{\zeta_{\text{mat}}(\Sigma)}{\sigma_{\min}(\Sigma)} (d_1 + d_2)$ for a sufficiently large constant c_0, we can ensure that $4\sqrt{2R_q}\tau^{1-q/2} < 1/2$, and hence that

$$\frac{\|\mathfrak{X}(\Delta)\|_2}{\sqrt{N}} \geq \frac{\sqrt{\sigma_{\min}(\Sigma)}}{4} \left\{ \frac{1}{2} |||\Delta|||_F - 4 R_q \tau^{2-q} \right\}.$$

Consequently, if we define $\delta := 16 R_q \tau^{2-q}$, then we are guaranteed that for all $|||\Delta|||_F \geq \delta$, we have $4 R_q \tau^{2-q} \leq |||\Delta|||_F/4$, and hence

$$\frac{\|\mathfrak{X}(\Delta)\|_2}{\sqrt{N}} \geq \frac{\sqrt{\sigma_{\min}(\Sigma)}}{16} |||\Delta|||_F$$

for all $|||\Delta|||_F \geq \delta$. We have thus shown that $\mathbb{C}(2 R_q \tau^{-q}; \delta)$ with parameter $\kappa(\mathfrak{X}) = \frac{\sigma_{\min}(\Sigma)}{256}$.

The next step is to control the quantity $|||\mathfrak{X}^*(\vec{\varepsilon})|||_2/N$, required for specifying a suitable choice of λ_N.

Lemma 4.6. *If $\|\vec{\varepsilon}\|_2 \leq 2\nu\sqrt{N}$, then there are universal constants c_i such that*

$$\mathbb{P}\Big[\frac{|||\mathfrak{X}^*(\vec{\varepsilon})|||_2}{N} \geq c_0\sqrt{\zeta_{mat}(\Sigma)}\nu\Big(\sqrt{\frac{d_1}{N}} + \sqrt{\frac{d_2}{N}}\Big)\Big] \leq c_1\exp(-c_2(d_1+d_2)). \tag{4.39}$$

Proof. By the definition of the adjoint operator, we have $Z = \frac{1}{N}\mathfrak{X}^*(\vec{\varepsilon}) = \frac{1}{N}\sum_{i=1}^N \varepsilon_i X_i$. Since the observation matrices $\{X_i\}_{i=1}^N$ are i.i.d. Gaussian, if the sequence $\{\varepsilon_i\}_{i=1}^N$ is viewed as fixed (by conditioning as needed), then the random matrix Z is a sample from the Γ-ensemble with covariance matrix $\Gamma = \frac{\|\vec{\varepsilon}\|^2}{N^2}\Sigma \preceq \frac{2\nu^2}{N}\Sigma$. Therefore, letting $\widetilde{Z} \in \mathbb{R}^{d_1\times d_2}$ be a random matrix drawn from the $2\nu^2\Sigma/N$-ensemble, we have

$$\mathbb{P}\big[|||Z|||_2 \geq t\big] \leq \mathbb{P}\big[|||\widetilde{Z}|||_2 \geq t\big].$$

Using Lemma B.1 from Appendix B.5, we have

$$\mathbb{E}[|||\widetilde{Z}|||_2] \leq \frac{12\sqrt{2}\nu\sqrt{\zeta_{\mathrm{mat}}(\Sigma)}}{\sqrt{N}}\big(\sqrt{d_1}+\sqrt{d_2}\big)$$

and

$$\mathbb{P}\big[|||\widetilde{Z}|||_2 \geq \mathbb{E}[|||\widetilde{Z}|||_2] + t\big] \leq \exp\big(-c_1\frac{Nt^2}{\nu^2\zeta_{\mathrm{mat}}(\Sigma)}\big)$$

for a universal constant c_1. Setting $t^2 = \Omega(\frac{\nu^2\zeta_{\mathrm{mat}}(\Sigma)(\sqrt{d_1}+\sqrt{d_2})^2}{N}$ yields the claim. □

4.4.6 Proof of Corollary 4.6

This corollary follows from a combination of Proposition 4.1 and Lemma 4.1. Let $\widehat{\Theta}$ be an optimal solution to the SDP (4.29), and let $\Delta = \widehat{\Theta} - \Theta^*$ be the error. Since $\widehat{\Theta}$ is optimal and Θ^* is feasible for the SDP, we have $|||\widehat{\Theta}|||_{\mathrm{nuc}} = |||\Theta^* + \Delta|||_{\mathrm{nuc}} \leq |||\Theta^*|||_{\mathrm{nuc}}$. Using the decomposition $\Delta = \Delta' + \Delta''$ from Lemma 4.1 and applying triangle inequality, we have

$$|||\Theta^* + \Delta' + \Delta''|||_{\mathrm{nuc}} \geq |||\Theta^* + \Delta''|||_{\mathrm{nuc}} - |||\Delta'|||_{\mathrm{nuc}}.$$

From the properties of the decomposition in Lemma 4.1 (see Appendix B.1), we find that

$$|||\widehat{\Theta}|||_{\mathrm{nuc}} = |||\Theta^* + \Delta' + \Delta''|||_{\mathrm{nuc}} \geq |||\Theta^*|||_{\mathrm{nuc}} + |||\Delta''|||_{\mathrm{nuc}} - |||\Delta'|||_{\mathrm{nuc}}.$$

Combining the pieces yields that $|||\Delta''|||_{\mathrm{nuc}} \leq |||\Delta'|||_{\mathrm{nuc}}$, and hence $|||\Delta|||_{\mathrm{nuc}} \leq 2|||\Delta'|||_{\mathrm{nuc}}$. By Lemma 4.1(a), the rank of Δ' is at most $2r$, so that we obtain $|||\Delta|||_{\mathrm{nuc}} \leq 2\sqrt{2r}|||\Delta|||_F \leq 4\sqrt{r}|||\Delta|||_F$.

Note that $\mathfrak{X}(\Delta) = 0$, since both $\widehat{\Theta}$ and Θ^* agree with the observations. Consequently, from Proposition 4.1, we have that

$$\begin{aligned} 0 \;=\; \frac{\|\mathfrak{X}(\Delta)\|_2}{\sqrt{N}} &\geq c_1 |\!|\!| \Delta |\!|\!|_F - c_2 \sqrt{\zeta_{\text{mat}}(\Sigma)} \left(\sqrt{\frac{d_1}{N}} + \sqrt{\frac{d_2}{N}} \right) |\!|\!| \Delta |\!|\!|_{\text{nuc}} \\ &\geq |\!|\!| \Delta |\!|\!|_F \left(c_1 - 12 \sqrt{\zeta_{\text{mat}}(\Sigma)} \sqrt{\frac{r d_1}{N}} + 12 \sqrt{\zeta_{\text{mat}}(\Sigma)} \sqrt{\frac{r d_2}{N}} \right) \\ &\geq \frac{1}{20} |\!|\!| \Delta |\!|\!|_F \end{aligned}$$

where the final inequality as long as $N > c_0 \zeta_{\text{mat}}(\Sigma) r(d_1 + d_2)$ for a sufficiently large constant c_0. We have thus shown that $\Delta = 0$, which implies that $\widehat{\Theta} = \Theta^*$ as claimed.

4.5 Experimental results

In this section, we report the results of various simulations that demonstrate the close agreement between the scaling predicted by our theory, and the actual behavior of the SDP-based M-estimator (4.9) in practice. In all cases, we solved the convex program (4.9) by using our own implementation in MATLAB of an accelerated gradient descent method which adapts a non-smooth convex optimization procedure [102] to the nuclear-norm [66]. We chose the regularization parameter λ_N in the manner suggested by our theoretical results; in doing so, we assumed knowledge of the noise variance ν^2. In practice, one would have to estimate such quantities from the data using methods such as cross-validation, as has been studied in the context of the Lasso, and we leave this as an interesting direction for future research.

We report simulation results for three of the running examples discussed in this chapter: low-rank multivariate regression, estimation in vector autoregressive processes, and matrix recovery from random projections (compressed sensing). In each case, we solved instances of the SDP for a square matrix $\Theta^* \in \mathbb{R}^{m \times m}$, where $m \in \{40, 80, 160\}$ for the first two examples, and $m \in \{20, 40, 80\}$ for the compressed sensing example. In all cases, we considered the case of exact low rank constraints, with $\text{rank}(\Theta^*) = r = 10$, and we generated Θ^* by choosing the subspaces of its left and right singular vectors uniformly at random from the Grassman manifold.[4] The observation noise had variance $\nu^2 = 1$, and we chose $C = \nu^2 I$ for the VAR process. The VAR process was generated by first solving for the covariance matrix Σ using the MATLAB function dylap and then generating a sample path. For each setting of (r, m), we solved the SDP for a range of sample sizes N.

Figure 4.1 shows results for a multivariate regression model with the covariates chosen randomly from a $N(0, I)$ distribution. Panel (a) plots the Frobenius error $|\!|\!| \widehat{\Theta} - \Theta^* |\!|\!|_F$ on a logarithmic scale versus the sample size N for three different matrix sizes, $m \in \{40, 80, 160\}$.

[4] More specifically, we let $\Theta^* = XY^T$, where $X, Y \in \mathbb{R}^{m \times r}$ have i.i.d. $N(0, 1)$ elements.

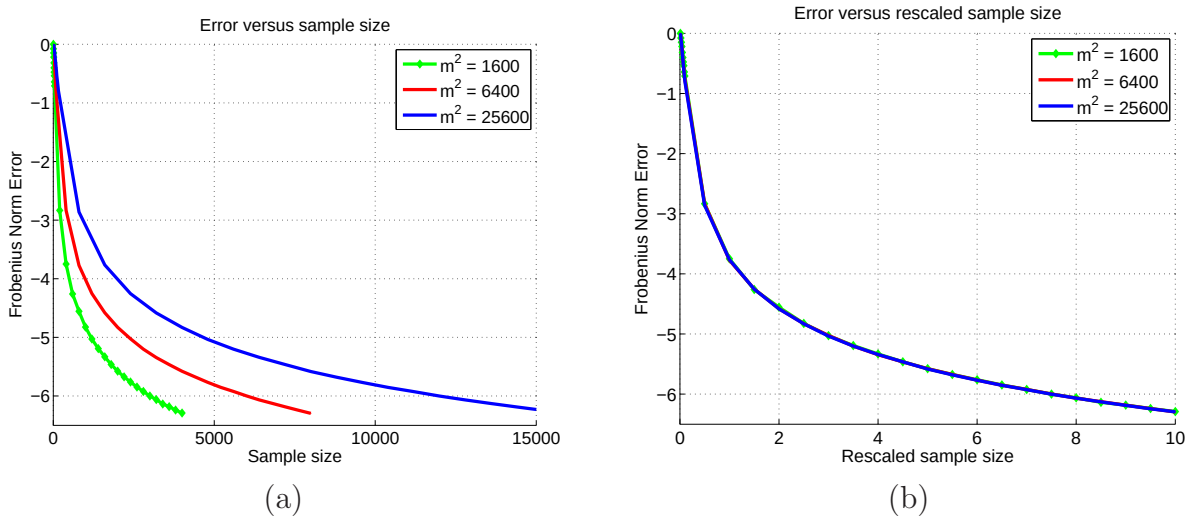
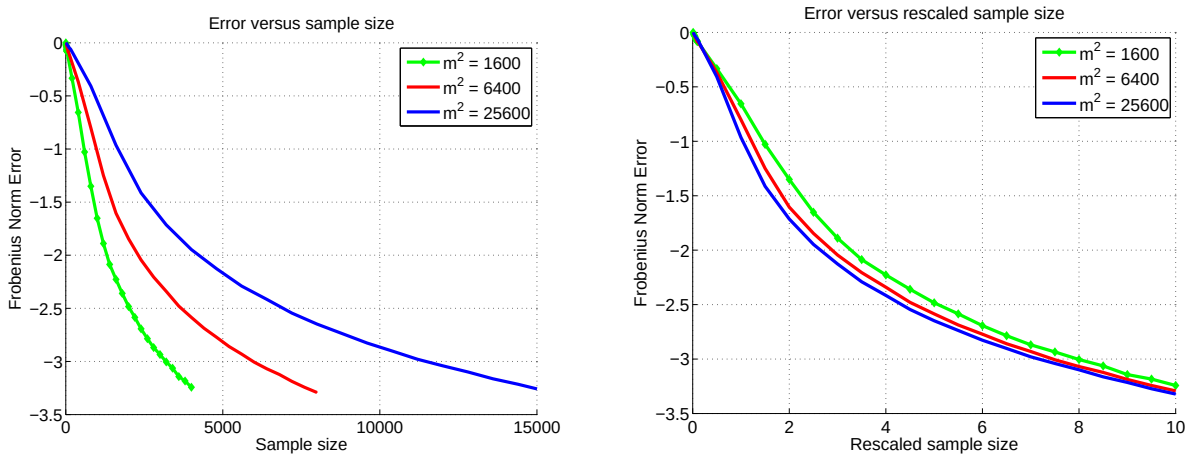


Figure 4.1. Results of applying the SDP (4.9) with nuclear norm regularization to the problem of low-rank multivariate regression. (a) Plots of the Frobenius error $|||\widehat{\Theta} - \Theta^*|||_F$ on a logarithmic scale versus the sample size N for three different matrix sizes $m^2 \in \{1600, 6400, 25600\}$, all with rank $r = 10$. (b) Plots of the same Frobenius error versus the rescaled sample size $N/(rm)$. Consistent with theory, all three plots are now extremely well-aligned.

Naturally, in each case, the error decays to zero as N increases, but larger matrices require larger sample sizes, as reflected by the rightward shift of the curves as m is increased. Panel (b) of Figure 4.1 shows the exact same set of simulation results, but now with the Frobenius error plotted versus the rescaled sample size $\widetilde{N} := N/(rm)$. As predicted by Corollary 4.3, the error plots now are all aligned with one another; the degree of alignment in this particular case is so close that the three plots are now indistinguishable. (The blue curve is the only one visible since it was plotted last by our routine.) Consequently, Figure 4.1 shows that $N/(rm)$ acts as the effective sample size in this high-dimensional setting.

Figure 4.2 shows similar results for the autoregressive model discussed in Example 4.2. As shown in panel (a), the Frobenius error again decays as the sample size is increased, although problems involving larger matrices are shifted to the right. Panel (b) shows the same Frobenius error plotted versus the rescaled sample size $N/(rm)$; as predicted by Corollary 4.4, the errors for different matrix sizes m are again quite well-aligned. In this case, we find (both in our theoretical analysis and experimental results) that the dependence in the autoregressive process slows down the rate at which the concentration occurs, so that the results are not as crisp as the low-rank multivariate setting in Figure 4.1.

Finally, Figure 4.3 presents the same set of results for the compressed sensing observation model discussed in Example 4.3. Even though the observation matrices X_i here are qualitatively different (in comparison to the multivariate regression and autoregressive examples),



(a) (b)

Figure 4.2. Results of applying the SDP (4.9) with nuclear norm regularization to estimating the system matrix of a vector autoregressive process. (a) Plots of the Frobenius error $|||\widehat{\Theta} - \Theta^*|||_F$ on a logarithmic scale versus the sample size N for three different matrix sizes $m^2 \in \{1600, 6400, 25600\}$, all with rank $r = 10$. (b) Plots of the same Frobenius error versus the rescaled sample size $N/(rm)$. Consistent with theory, all three plots are now reasonably well-aligned.

we again see the "stacking" phenomenon of the curves when plotted versus the rescaled sample size $N/(rm)$, as predicted by Corollary 4.5.

4.6 Discussion

In this chapter, we have analyzed the nuclear norm relaxation for a general class of noisy observation models, and obtained non-asymptotic error bounds on the Frobenius norm that hold under high-dimensional scaling. In contrast to most past work, our results are applicable to both exactly and approximately low-rank matrices. We stated a main theorem that provides high-dimensional rates in a fairly general setting, and then showed how by specializing this result to some specific model classes—namely, low-rank multivariate regression, estimation of autoregressive processes, and matrix recovery from random projections—it yields concrete and readily interpretable rates. Lastly, we provided some simulation results that showed excellent agreement with the predictions from our theory.

This chapter has focused on achievable results for low-rank matrix estimation using a particular polynomial-time method. It would be interesting to establish matching lower bounds, showing that the rates obtained by this estimator are minimax-optimal. We suspect that this should be possible, for instance by using the techniques exploited in Raskutti et al. [109] in analyzing minimax rates for regression over ℓ_q-balls.

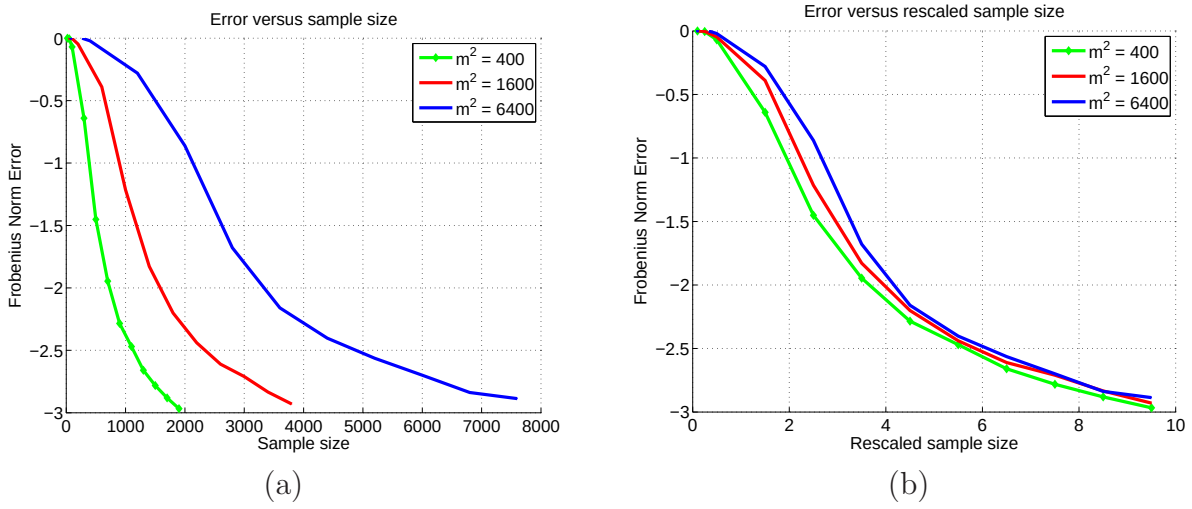


Figure 4.3. Results of applying the SDP (4.9) with nuclear norm regularization to recovering a low-rank matrix on the basis of random projections (compressed sensing model) (a) Plots of the Frobenius error $|||\widehat{\Theta} - \Theta^*|||_F$ on a logarithmic scale versus the sample size N for three different matrix sizes $m^2 \in \{400, 1600, 6400\}$, all with rank $r = 10$. (b) Plots of the same Frobenius error versus the rescaled sample size $N/(rm)$. Consistent with theory, all three plots are now reasonably well-aligned.

Chapter 5

Matrix Completion

5.1 Introduction

Matrix completion problems correspond to reconstructing matrices, either exactly or approximately, based on observing a subset of their entries [76, 43]. In the simplest formulation of matrix completion, the observations are assumed to be uncorrupted, whereas a more general formulation (as considered in this chapter) allows for noisiness in these observations. As noted in Chapter 1, matrix recovery based on only partial information is an ill-posed problem, and accurate estimates are possible only if the matrix satisfies additional structural constraints, with examples including bandedness, positive semidefiniteness, Euclidean distance measurements, Toeplitz, and low-rank structure (see the survey paper [76] and references therein for more background).

The focus of this chapter is low-rank matrix completion based on noisy observations. This problem is motivated by a variety of applications where an underlying matrix is likely to have low-rank, or near low-rank structure. The archetypal example is the Netflix challenge, a version of the collaborative filtering problem, in which the unknown matrix is indexed by individuals and movies, and each observed entry of the matrix corresponds to the rating assigned to the associated movie by the given individual. Since the typical person only watches a tiny number of movies (compared to the total Netflix database), it is only a sparse subset of matrix entries that are observed. In this context, one goal of collaborative filtering is to use the observed entries to make recommendations to a person regarding movies that they have *not* yet seen. We refer the reader to Srebro's thesis [124] (and references therein) for further discussion and motivation for collaborative filtering and related problems.

In this chapter, we analyze a method for approximate low-rank matrix recovery using an M-estimator that is a combination of a data term, and a weighted nuclear norm as a regularizer, as discussed in the previous chapter. Recall that the nuclear norm is the sum of the singular values of a matrix [60], and has been studied in a body of past work, both on matrix completion and more general problems of low-rank matrix estimation (e.g., Chap-

ter 4 and [51, 124, 126, 125, 117, 9, 34, 115, 70, 71, 119]). A parallel line of work has studied computationally efficient algorithms for solving problems with nuclear norm constraints (e.g, [89, 102, 80]). Here we limit our detailed discussion to those papers that study various aspects of the matrix completion problem. Motivated by various problems in collaborative filtering, Srebro and colleagues [124, 126, 125] studied various aspects nuclear norm regularization; among various other contributions, Srebro et al. [126] established generalization error bounds under certain conditions. Candes and Recht [33] studied the exact reconstruction of a low-rank matrix given perfect (noiseless) observations of a subset of entries, and provided sufficient conditions for exact recovery via nuclear norm relaxation, with later refinements [34, 115]. Gross [57] recognized the utility of the Ahlswede-Winter matrix concentration bounds, and the simplest argument to date is provided by Recht [115]. In a parallel line of work, Keshavan et al. [70, 71] have studied a method based on thresholding and singular value decomposition, and established various results on its behavior, both for noiseless and noisy matrix completion. Among other results, Rohde and Tsybakov [119] establish prediction error bounds for matrix completion, a different metric than the matrix recovery problem of interest here. In recent work, Salakhutdinov and Srebro [122] provided various motivations for the use of weighted nuclear norms, in particular showing that the standard nuclear norm relaxation can behave very poorly when the sampling is non-uniform. The analysis presented in this chapter applies to both uniform and non-uniform sampling, as well as a form of reweighted nuclear norm as suggested by these authors, one which includes the ordinary nuclear norm as a special case. We provide a more detailed comparison between our results and some aspects of past work in Section 5.3.4.

As has been noted before [29], a significant theoretical challenge is that conditions that have proven very useful for sparse linear regression—among them the restricted isometry property—are *not* satisfied for the matrix completion problem. For this reason, it is natural to seek an alternative and less restrictive property that might be satisfied in the matrix completion setting. In Chapter 3 we isolate a weaker condition known as *restricted strong convexity* (RSC). In both Chapters 3 and 4 we prove that certain statistical models satisfy RSC with high probability when the associated regularizer satisfies the *decomposability* condition 3.1. When an M-estimator satisfies the RSC condition, it is relatively straightforward to derive non-asymptotic error bounds on parameter estimates as shown in Chapter 3. The class of decomposable regularizers includes the nuclear norm as particular case, and the RSC/decomposability approach has been exploited in Chapter 4 to derive bounds for various matrix estimation problems, among them multi-task learning, autoregressive system identification, and compressed sensing.

To date, however, an open question is whether or not an appropriate form of RSC holds for the matrix completion problem. If it did hold, then it would be possible to derive non-asymptotic error bounds (in Frobenius norm) for matrix completion based on noisy observations. Within this context, the main contribution of presented in this chapter is to prove that with high probability, a form of the RSC condition holds for the matrix completion problem, in particular over an interesting set of matrices $\mathcal{D}$, as defined in equation (5.8) to

follow, that have both low nuclear/Frobenius norm ratio and low "spikiness". Exploiting this RSC condition then allows us to derive non-asymptotic error bounds on matrix recovery in weighted Frobenius norms, both for exactly and approximately low-rank matrices. The theoretical core of this chapter consists of three main results. Our first result (Theorem 5.1) proves that the matrix completion loss function satisfies restricted strong convexity with high probability over the set $\mathcal{D}$. Our second result (Theorem 5.2) exploits this fact to derive a non-asymptotic error bound for matrix recovery in the weighted Frobenius norm, one applicable to general matrices. We then specialize this result to the problem of estimating exactly low-rank matrices (with a small number of non-zero singular values), as well as near low-rank matrices characterized by relatively swift decay of their singular values. To the best of our knowledge, our results on near low-rank matrices are the first for approximate matrix recovery in the noisy setting, and as we discuss at more length in Section 5.3.4, our results on the exactly low-rank case are sharper than past work on the problem. Indeed, our final result (Theorem 5.3) uses information-theoretic techniques to establish that up to logarithmic factors, no algorithm can obtain faster rates than our method over the ℓ_q-balls of matrices with bounded spikiness treated in this chapter.

The remainder of this chapter is organized as follows. We begin in Section 5.2 with background and a precise formulation of the problem. Section 5.3 is devoted to a statement of our main results, and discussion of some of their consequences. In Sections 5.4 and Section 5.5, we prove our main results, with more technical aspects of the arguments deferred to appendices. We conclude with a discussion in Section 5.6.

5.2 Background and problem formulation

In this section, we introduce background on low-rank matrix completion problem, and also provide a precise statement of the problem studied in this chapter.

5.2.1 Uniform and weighted sampling models

Let $\Theta^* \in \mathbb{R}^{m_r \times m_c}$ be an unknown matrix, and consider an observation model in which we make n i.i.d. observations of the form

$$\widetilde{y}_i = \Theta^*_{j(i)k(i)} + \frac{\nu}{\sqrt{m_r m_c}} \widetilde{\xi}_i, \tag{5.1}$$

Here the quantities $\frac{\nu}{\sqrt{m_r m_c}} \widetilde{\xi}_i$ correspond to additive observation noises with variance appropriately scaled according to the matrix dimensions. In defining the observation model, one can either allow the Frobenius norm of Θ^* to grow with the dimension, as in done in other work [29, 71], or rescale the noise as we have done here. This choice is consistent with our assumption that Θ^* has constant Frobenius norm regardless of its rank or dimensions.

With this scaling, each observation in the model (5.1) has a constant signal-to-noise ratio regardless of matrix dimensions.

In the simplest model, the row $j(i)$ and column $k(i)$ indices are chosen uniformly at random from the sets $\{1, 2, \ldots, m_r\}$ and $\{1, 2, \ldots, m_c\}$ respectively. In this chapter, we consider a somewhat more general weighted sampling model. In particular, let $R \in \mathbb{R}^{m_r \times m_r}$ and $C \in \mathbb{R}^{m_c \times m_c}$ be diagonal matrices, with rescaled diagonals $\{R_j/m_r, j = 1, 2, \ldots, m_r\}$ and $\{C_k/m_c, k = 1, 2, \ldots, m_c\}$ representing probability distributions over the rows and columns of an $m_r \times m_c$ matrix. We consider the weighted sampling model in which we make a noisy observation of entry (j, k) with probability $R_j C_k/(m_r m_c)$, meaning that the row index $j(i)$ (respectively column index $k(i)$) is chosen according to the probability distribution R/m_r (respectively C/m_c). Note that in the special case that $R = \mathbf{1}_{m_r}$ and $C = \mathbf{1}_{m_c}$, the observation model (5.1) reduces to the usual model of uniform sampling.

We assume that each row and column is sampled with positive probability, in particular that there is some constant $1 \leq L < \infty$ such that $R_a \geq 1/L$ and $C_b \geq 1/L$ for all rows and columns. However, apart from the constraints $\sum_{a=1}^{m_r} R_{aa} = m_r$ and $\sum_{b=1}^{m_c} C_{bb} = m_c$, we do not require that the row and column weights remain bounded as m_r and m_c tend to infinity.

5.2.2 The observation operator and restricted strong convexity

We now describe an alternative formulation of the observation model (5.1) that, while statistically equivalent to the original, turns out to be more natural for analysis. For each $i = 1, 2, \ldots, n$, define the matrix

$$X_i = \sqrt{m_r m_c}\, \varepsilon_i \, e_{a(i)} e_{b(i)}^T, \tag{5.2}$$

where $\varepsilon_i \in \{-1, +1\}$ is a random sign, and consider the observation model

$$y_i = \langle\langle X_i, \, \Theta^* \rangle\rangle + \nu \, \xi_i, \qquad \text{for } i = 1, \ldots, n, \tag{5.3}$$

where $\langle\langle A, \, B \rangle\rangle := \sum_{j,k} A_{jk} B_{jk}$ is the trace inner product, and ξ_i is an additive noise from the same distribution as the original model. The model (5.3) is can be obtained from the original model (5.1) by rescaling all terms by the factor $\sqrt{m_r m_c}$, and introducing the random signs ε_i. The rescaling has no statistical effect, and nor do the random signs, since the noise is symmetric (so that $\xi_i = \varepsilon_i \widetilde{\xi}_i$ has the same distribution as $\widetilde{\xi}_i$). Thus, the observation model (5.3) is statistically equivalent to the original one (5.1).

In order to specify a vector form of the observation model, let us define an operator $\mathfrak{X} : \mathbb{R}^{m_r \times m_c} \to \mathbb{R}^n$ via

$$[\mathfrak{X}(\Theta)]_i := \langle\langle X_i, \, \Theta \rangle\rangle, \quad \text{for } i = 1, 2, \ldots n.$$

We refer to $\mathfrak{X}$ as the *observation operator*, since it maps any matrix $\Theta \in \mathbb{R}^{m_r \times m_c}$ to an n-vector of samples. With this notation, we can write the observations (5.3) in a vectorized form as $y = \mathfrak{X}(\Theta^*) + \nu \xi$.

The reformulation (5.3) is convenient for various reasons. For any matrix $\Theta \in \mathbb{R}^{m_r \times m_c}$, we have $\mathbb{E}[\langle\langle X_i, \Theta \rangle\rangle] = 0$ and

$$\mathbb{E}\big[\langle\langle X_i, \Theta \rangle\rangle^2\big] = \sum_{j=1}^{m_r} \sum_{k=1}^{m_c} R_j \Theta_{jk}^2 C_k = \underbrace{|||\sqrt{R}\Theta\sqrt{C}|||_F^2}_{|||\Theta|||_{\omega(F)}^2}, \tag{5.4}$$

where we have defined the *weighted Frobenius norm* $|||\cdot|||_{\omega(F)}$ in terms of the row R and column C weights. As a consequence, the signal-to-noise ratio in the observation model (5.3) is given by the ratio $\text{SNR} = \frac{|||\Theta^*|||_{\omega(F)}^2}{\nu^2}$.

As shown in Chapter 3 a key ingredient in establishing error bounds for the observation model (5.3) is obtaining lower bounds on the restricted curvature of the sampling operator—in particular, to establish the existence of a constant $c > 0$, which may be arbitrarily small as long as it is positive, such that

$$\frac{\|\mathfrak{X}(\Theta)\|_2}{\sqrt{n}} \geq c\, |||\Theta|||_{\omega(F)}. \tag{5.5}$$

For sample sizes of interest for matrix completion ($n \ll m_r m_c$) , one cannot expect such a bound to hold uniformly over all matrices $\Theta \in \mathbb{R}^{m_r \times m_c}$, even when rank constraints are imposed. Indeed, as noted by Candes and Plan [29], the condition (5.5) is violated with high probability by the rank one matrix Θ^* such that $\Theta^*_{11} = 1$ with all other entries zero. Indeed, for a sample size $n \ll m_r m_c$, we have a vanishing probability of observing the entry Θ^*_{11}, so that $\mathfrak{X}(\Theta^*) = 0$ with high probability.

5.2.3 Controlling the spikiness and rank

Intuitively, one must exclude matrices that are overly "spiky" in order to avoid the phenomenon just described. Past work has relied on fairly restrictive matrix incoherence conditions (see Section 5.3.4 for more discussion), based on specific conditions on singular vectors of the unknown matrix Θ^*. In this chapter, we formalize the notion of "spikiness" in a natural and less restrictive way—namely by comparing a weighted form of ℓ_∞-norm to the weighted Frobenius norm. In particular, for any non-zero matrix Θ, let us define (for any non-zero matrix) the *weighted spikiness ratio*

$$\alpha_{\text{sp}}(\Theta) := \sqrt{m_r m_c}\, \frac{|||\Theta|||_{\omega(\infty)}}{|||\Theta|||_{\omega(F)}}, \tag{5.6}$$

where $|||\Theta|||_{\omega(\infty)} := \|\sqrt{R}\Theta\sqrt{C}\|_\infty$ is the weighted elementwise ℓ_∞-norm. Note that this ratio is invariant to the scaling of Θ, and satisfies the inequalities $1 \leq \alpha_{\text{sp}}(\Theta) \leq \sqrt{m_r m_c}$. We have $\alpha_{\text{sp}}(\Theta) = 1$ for any non-zero matrix whose entries are all equal, whereas the opposite extreme

$\alpha_{\mathrm{sp}}(\Theta) = \sqrt{m_r m_c}$ is achieved by the "maximally spiky" matrix that is zero everywhere except for a single position.

In order to provide a tractable measure of how close Θ is to a low-rank matrix, we define (for any non-zero matrix) the ratio

$$\beta_{\mathrm{ra}}(\Theta) := \frac{|||\Theta|||_{\omega(1)}}{|||\Theta|||_{\omega(F)}} \tag{5.7}$$

which satisfies the inequalities $1 \leq \beta_{\mathrm{ra}}(\Theta) \leq \sqrt{\min\{m_r, m_c\}}$. By definition of the (weighted) nuclear and Frobenius norms, note that $\beta_{\mathrm{ra}}(\Theta)$ is simply the ratio of the ℓ_1 to ℓ_2 norms of the singular values of the weighted matrix $\sqrt{R}\Theta\sqrt{C}$. This measure can also be upper bounded by the rank of Θ: indeed, since R and C are full-rank, we always have

$$\beta^2_{\mathrm{ra}}(\Theta) \leq \mathrm{rank}(\sqrt{R}\Theta\sqrt{C}) \;=\; \mathrm{rank}(\Theta),$$

with equality holding if all the non-zero singular values of $\sqrt{R}\Theta\sqrt{C}$ are identical.

5.3 Main results and their consequences

We now turn to the statement of our main results, and discussion of their consequences. Section 5.3.1 is devoted to a result showing that a suitable form of restricted strong convexity holds for the random sampling operator $\mathfrak{X}$, as long as we restrict it to matrices Δ for which $\beta_{\mathrm{ra}}(\Delta)$ and $\alpha_{\mathrm{sp}}(\Delta)$ are not "overly large". In Section 5.3.2, we develop the consequences of the RSC condition for noisy matrix completion, and in Section 5.3.3, we prove that our error bounds are minimax-optimal up to logarithmic factors. In Section 5.3.4, we provide a detailed comparison of our results with past work.

5.3.1 Restricted strong convexity for matrix sampling

Introducing the convenient shorthand $m = \frac{1}{2}(m_r + m_c)$, let us define the constraint set

$$\mathcal{D}(n; c_0) := \left\{ \Delta \in \mathbb{R}^{m_r \times m_c},\, \Delta \neq 0 \;\middle|\; \alpha_{\mathrm{sp}}(\Delta)\, \beta_{\mathrm{ra}}(\Delta) \leq \frac{1}{c_0\, L} \sqrt{\frac{n}{m \log m}} \right\}, \tag{5.8}$$

where c_0 is a universal constant. Note that as the sample size n increases, this set allows for matrices with larger values of the spikiness and/or rank measures, $\alpha_{\mathrm{sp}}(\Delta)$ and $\beta_{\mathrm{ra}}(\Delta)$ respectively.

Theorem 5.1. *There are universal constants (c_0, c_1, c_2, c_3) such that as long as $n > c_3\, m \log m$, we have*

$$\frac{\|\mathfrak{X}(\Delta)\|_2}{\sqrt{n}} \geq \frac{1}{8} |||\Delta|||_{\omega(F)} \left\{ 1 - \frac{128\, \alpha_{\mathrm{sp}}(\Delta) L}{\sqrt{n}} \right\} \qquad \textit{for all } \Delta \in \mathcal{D}(n; c_0) \tag{5.9}$$

with probability greater than $1 - c_1 \exp(-c_2 m \log m)$.

Roughly speaking, this bound guarantees that the observation operator captures a substantial component of any matrix $\Delta \in \mathcal{D}(n; c_0)$ that is not overly spiky. More precisely, as long as $\frac{128 L \alpha_{\mathrm{sp}}(\Delta)}{\sqrt{n}} \leq \frac{1}{2}$, the bound (5.9) implies that

$$\frac{\|\mathfrak{X}(\Delta)\|_2^2}{n} \geq \frac{1}{256} |||\Delta|||_{\omega(F)}^2 \qquad \text{for any } \Delta \in \mathcal{D}(n; c_0). \tag{5.10}$$

This bound can be interpreted in terms of *restricted strong convexity* 3.2.4. In particular, given a vector $y \in \mathbb{R}^n$ of noisy observations, consider the quadratic loss function

$$\mathcal{L}(\Theta; y) = \frac{1}{2n} \|y - \mathfrak{X}(\Theta)\|_2^2.$$

Since the Hessian matrix of this function is given by $\mathfrak{X}^* \mathfrak{X}/n$, the bound (5.10) implies that the quadratic loss is strongly convex in a restricted set of directions Δ.

As discussed previously, the worst-case value of the "spikiness" measure is $\alpha_{\mathrm{sp}}(\Delta) = \sqrt{m_r m_c}$, achieved for a matrix that is zero everywhere except a single position. In this most degenerate of cases, the combination of the constraints $\frac{\alpha_{\mathrm{sp}}(\Delta)}{\sqrt{n}} < 1$ and the membership condition $\Delta \in \mathcal{D}(n; c_0)$ imply that even for a rank one matrix (so that $\beta_{\mathrm{ra}}(\Delta) = 1$), we need sample size $n \gg m^2$ for Theorem 5.1 to provide a non-trivial result, as is to be expected.

5.3.2 Consequences for noisy matrix completion

We now turn to some consequences of Theorem 5.1 for matrix completion in the noisy setting. In particular, assume that we are given n i.i.d. samples from the model (5.3), and let $\widehat{\Theta}$ be some estimate of the unknown matrix Θ^*. Our strategy is to exploit the lower bound (5.9) in application to the error matrix $\widehat{\Theta} - \Theta^*$, and accordingly, we need to ensure that it has relatively low-rank and spikiness. Based on this intuition, it is natural to consider the estimator

$$\widehat{\Theta} \in \arg \min_{|||\Theta|||_{\omega(\infty)} \leq \frac{\alpha^*}{\sqrt{m_r m_c}}} \Big\{ \frac{1}{2n} \|y - \mathfrak{X}(\Theta)\|_2^2 + \lambda_n |||\Theta|||_{\omega(1)} \Big\}, \tag{5.11}$$

where $\alpha^* \geq 1$ is a measure of spikiness, and the regularization parameter $\lambda_n > 0$ serves to control the nuclear norm of the solution. In the special case when both R and C are identity matrices (of the appropriate dimensions), this estimator is closely related to the standard one considered in past work on the problem, with the only difference between the additional ℓ_∞-norm constraint. In the more general weighted case, an M-estimator of the form (5.11) using the weighted nuclear norm (but without the elementwise constraint) was recently suggested by Salakhutdinov and Srebro [122], who provided empirical results to show superiority of the weighted nuclear norm over the standard choice for the Netflix problem.

Past work on matrix completion has focused on the case of exactly low-rank matrices. Here we consider the more general setting of approximately low-rank matrices, including the exact setting as a particular case. We begin by stating a general upper bound that applies to any matrix Θ^*, and involves a natural decomposition into estimation and approximation error terms. The only relevant quantity is the signal-to-noise ratio, as measured by the ratio of the Frobenius norm of Θ^* to the noise variance, so that we allow the noise variance to be free, while assuming that $|||\widetilde{\Delta}|||_{\omega(F)}$ remains bounded.

Theorem 5.2. *Suppose that $n \geq Lm \log m$, and consider any solution $\widehat{\Theta}$ to the weighted SDP (5.11) using regularization parameter*

$$\lambda_n \geq 2\nu \, |||\frac{1}{n}\sum_{i=1}^{n} \xi_i R^{-\frac{1}{2}} X_i C^{-\frac{1}{2}}|||_2, \tag{5.12}$$

and define $\lambda_n^* = \max\{\lambda_n, L\sqrt{\frac{m \log m}{n}}\}$. Then with probability greater than $1 - c_2 \exp(-c_2 \log m)$, for each $r = 1, \ldots, m_r$, the error $\widetilde{\Delta} = \widehat{\Theta} - \Theta^*$ satisfies

$$|||\widetilde{\Delta}|||_{\omega(F)}^2 \leq c_1 \, \alpha^* \, \lambda_n^* \left[\sqrt{r}|||\widetilde{\Delta}|||_{\omega(F)} + \sum_{j=r+1}^{m_r} \sigma_j(\sqrt{R}\Theta^*\sqrt{C})\right] + \frac{c_1(\alpha^* L)^2}{n}. \tag{5.13}$$

Apart from the trailing $\mathcal{O}(n^{-1})$ the term, the bound (5.13) shows a natural splitting into two terms. The first can be interpreted as the *estimation error* associated with a rank r matrix, whereas the second term corresponds to *approximation error*, measuring how far $\sqrt{R}\Theta^*\sqrt{C}$ is from a rank r matrix. Of course, the bound holds for any choice of r, and in the corollaries to follow, we choose r optimally so as to balance the estimation and approximation error terms.

In order to provide concrete rates using Theorem 5.2, it remains to address two issues. First, we need to specify an explicit choice of λ_n by bounding the operator norm of the matrix $\frac{1}{n}\sum_{i=1}^{n} \xi_i \sqrt{R} X_i \sqrt{C}$, and secondly, we need to understand how to choose the parameter r so as to achieve the tightest possible bound. When Θ^* is exactly low-rank, then it is obvious that we should choose $r = \text{rank}(\Theta^*)$, so that the approximation error vanishes—viz. $\sum_{j=r+1}^{m_r} \sigma_j(\sqrt{R}\Theta^*\sqrt{C})_j = 0$. Doing so yields the following result:

Corollary 5.1 (Exactly low-rank matrices)**.** *Suppose that the noise sequence $\{\xi_i\}$ is i.i.d., zero-mean and sub-exponential, and Θ^* has rank at most r, Frobenius norm at most 1, and spikiness at most $\alpha_{\text{sp}}(\Theta^*) \leq \alpha^*$. If we solve the SDP (5.11) with $\lambda_n = 4\nu\sqrt{\frac{m \log m}{n}}$ then there is a numerical constant c_1' such that*

$$|||\widehat{\Theta} - \Theta^*|||_{\omega(F)}^2 \leq c_1' \, (\nu^2 \vee L^2) \, (\alpha^*)^2 \, \frac{rm \log m}{n} \tag{5.14}$$

with probability greater than $1 - c_2 \exp(-c_3 \log m)$.

Note that this rate has a natural interpretation: since a rank r matrix of dimension $m_r \times m_c$ has roughly $r(m_r + m_c)$ free parameters, we require a sample size of this order (up to logarithmic factors) so as to obtain a controlled error bound. An interesting feature of the bound (5.14) is the term $\nu^2 \vee 1 = \max\{\nu^2, 1\}$, which implies that we do not obtain exact recovery as $\nu \to 0$. As we discuss at more length in Section 5.3.4, under the mild spikiness condition that we have imposed, this behavior is unavoidable due to lack of identifiability within a certain radius, as specified in the set $\mathcal{D}$. For instance, consider the matrix Θ^* and the perturbed version $\widetilde{\Theta} = \Theta^* + \frac{1}{\sqrt{m_r m_c}} e_1 e_1^T$. With high probability, we have $\mathfrak{X}(\Theta^*) = \mathfrak{X}(\widetilde{\Theta})$, so that the observations—even if they were noiseless—fail to distinguish between these two models. These types of examples, leading to non-identifiability, cannot be overcome without imposing fairly restrictive matrix incoherence conditions, as we discuss at more length in Section 5.3.4.

As with past work [29, 71], Corollary 5.1 applies to the case of matrices that have exactly rank r. In practical settings, it is more realistic to assume that the unknown matrix is not exactly low-rank, but rather can be well approximated by a matrix with low rank. One way in which to formalize this notion is via the ℓ_q-"ball" of matrices

$$\mathbb{B}_q(R_q) := \left\{ \Theta \in \mathbb{R}^{m_r \times m_c} \mid \sum_{j=1}^{\min\{m_r, m_c\}} |\sigma_j(\sqrt{R}\Theta\sqrt{C})|^q \leq R_q \right\}. \tag{5.15}$$

For $q = 0$, this set corresponds to the set of matrices with rank at most $r = \rho_0$, whereas for values $q \in (0, 1]$, it consists of matrices whose (weighted) singular values decay at a relatively fast rate. By applying Theorem 5.2 to this matrix family, we obtain the following corollary:

Corollary 5.2 (Estimation of near low-rank matrices)**.** *Suppose that the noise $\{\xi_i\}$ is zero-mean and sub-exponential, Consider a matrix $\Theta^* \in \mathbb{B}_q(R_q)$ with spikiness at most $\alpha_{\mathrm{sp}}(\Theta^*) \leq \alpha^*$, and Frobenius norm at most one. With the same choice of λ_n as Corollary 5.1, there is a universal constant c_1' such that*

$$\|\widehat{\Theta} - \Theta^*\|^2_{\omega(F)} \leq c_1 R_q \left((\nu^2 \vee L^2)(\alpha^*)^2 \frac{m \log m}{n} \right)^{1 - \frac{q}{2}} + \frac{c_1 (\alpha^* L)^2}{n} \tag{5.16}$$

with probability greater than $1 - c_2 \exp(-c_3 \log m)$.

Note that this result is a strict generalization of Corollary 5.1, to which it reduces in the case $q = 0$. (When $q = 0$, we have $\rho_0 = r$ so that the bound has the same form.) Note that the price that we pay for approximately low rank is a smaller exponent—namely, $1 - q/2$ as opposed to 1 in the case $q = 0$. The proof of Corollary 5.2 is based on a more subtle application of Theorem 5.2, one which chooses the effective rank r in the bound (5.13) so as to trade off between the estimation and approximation errors. In particular, the choice $r \asymp R_q \, (\frac{n}{m \log m})^{q/2}$ turns out to yield the optimal trade-off, and hence the given error

bound (5.16).

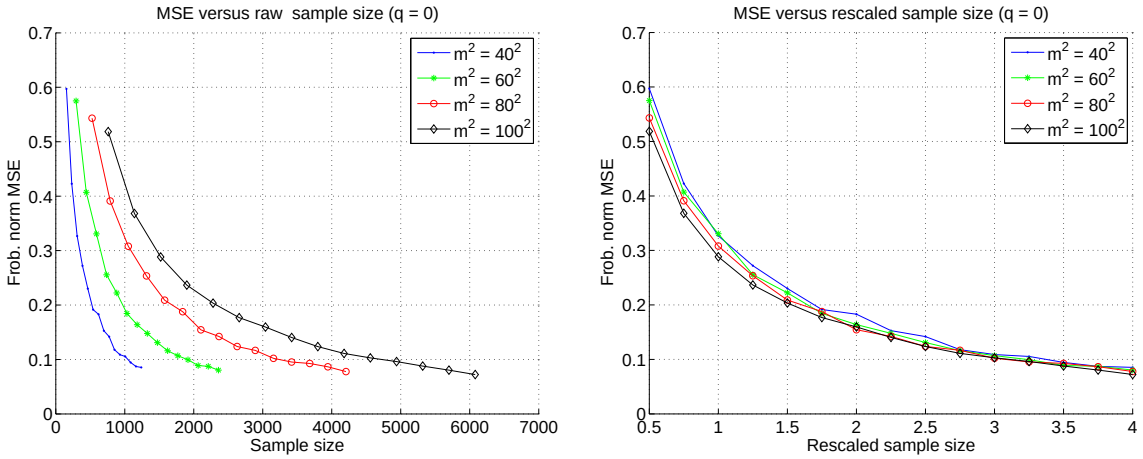
Although we have stated our results in terms of bounds on the weighted squared Frobenius norm $|||\Theta|||^2_{\omega(F)} = |||\sqrt{R}\Theta\sqrt{C}|||^2_F$, our assumed lower bound on the entries R and C implies that $|||\Theta|||^2_{\omega(F)} \geq \frac{|||\Theta|||^2_F}{L^2}$. Consequently, as long as each row and column is sampled a constant fraction of the time, our results also yield bounds on the Frobenius norm. In some applications, certain rows and columns might be heavily sampled, meaning that some entries of R and/or C could be relatively large. Since we require *only a lower bound* on the row/column sampling frequencies, our Frobenius norm bounds would not degrade if some rows and/or columns were heavily sampled. In contrast, a RIP-type analysis would not be valid in this setting, since heavy sampling means that the Frobenius norm could not be uniformly bounded from above.

In order to illustrate the sharpness of our theory, let us compare the predictions of our two corollaries to the empirical behavior of the M-estimator. In particular, we applied the nuclear norm SDP to simulated data, using Gaussian observation noise with variance $\nu^2 = 0.25$ and the uniform sampling model. In all cases, we solved the nuclear norm SDP using a non-smooth optimization procedure due to Nesterov [102], via our own implementation in MATLAB. For a given problem size m, we ran $T = 25$ trials and computed the squared Frobenius norm error $|||\widehat{\Theta} - \Theta^*|||^2_F$ averaged over the trials.

Figure 5.1 shows the results in the case of exactly low-rank matrices ($q = 0$), with the matrix rank given by $r = \lceil \log^2(m) \rceil$. Panel (a) shows plots of the mean-squared Frobenius error versus the raw sample size, for three different problem sizes with the number of matrix elements sizes $m^2 \in \{40^2, 60^2, 80^2, 100^2\}$. These plots show that the M-estimator is consistent, since each of the curves decreases to zero as the sample size n increases. Note that the curves shift to the right as the matrix dimension m increases, reflecting the natural intuition that larger matrices require more samples. Based on the scaling predicted by Corollary 5.1, we expect that the mean-squared Frobenius error should exhibit the scaling $|||\widehat{\Theta} - \Theta^*|||^2_F \asymp \frac{rm \log m}{n}$. Equivalently, if we plot the MSE versus the *rescaled sample size* $N := \frac{n}{rm \log m}$, then all the curves should be relatively well aligned, and decay at the rate $1/N$. Panel (b) of Figure 5.1 shows the same simulation results re-plotted versus this rescaled sample size. Consistent with the prediction of Corollary 5.1, all four plots are now relatively well-aligned. Figure 5.2 shows the same plots for the case of approximately low-rank matrices ($q = 0.5$). Again, consistent with the prediction of Corollary 5.2, we see qualitatively similar behavior in the plots of the MSE versus sample size (panel (a)), and the rescaled sample size (panel (b)).

5.3.3 Information-theoretic lower bounds

The results of the previous section are achievable results, based on a particular polynomial-time estimator. It is natural to ask how these bounds compare to the fundamental limits of the problem, meaning the best performance achievable by any algorithm. As various authors



(a) (b)

Figure 5.1. Plots of the mean-squared error in Frobenius norm for $q = 0$. Each curve corresponds to a different problem size $m^2 \in \{40^2, 60^2, 80^2, 100^2\}$. (a) MSE versus the raw sample size n. As expected, the curves shift to the right as m increases, since more samples should be required to achieve a given MSE for larger problems. (b) The same MSE plotted versus the rescaled sample size $n/(rm \log m)$. Consistent with Corollary 5.1, all the plots are now fairly well-aligned.

have noted [29, 71], a parameter counting argument indicates that roughly $n \approx r\,(m_r + m_c)$ samples are required to estimate an $m_r \times m_c$ matrix with rank r. This calculation can be made more formal by metric entropy calculations for the Grassman manifold (e.g., [130]); see also Rohde and Tsybakov [119] for results on approximation numbers for the more general ℓ_q-balls of matrices. Such calculations, while accounting for the low-rank conditions, do *not* address the additional "spikiness" constraints that are essential to the setting of matrix completion. It is conceivable that these additional constraints could lead to a substantial volume reduction in the allowable class of matrices, so that the scalings suggested by parameter counting or metric entropy calculation for Grassman manifolds would be overly conservative.

Accordingly, in this section, we provide a direct and constructive argument to lower bound the minimax rates of Frobenius norm over classes of matrices that are near low-rank and not overly spiky. This argument establishes that the bounds established in Corollaries 5.1 and 5.2 are sharp up to logarithmic factors, meaning that no estimator performs substantially better than the one considered here. More precisely, consider the matrix classes

$$\widetilde{\mathbb{B}}(R_q) = \left\{ \Theta \in \mathbb{R}^{m \times m} \;\mid\; \sum_{j=1}^{m} \sigma_j(\Theta)^q \leq R_q,\; \alpha_{\mathrm{sp}}(\Theta) \leq \sqrt{32 \log m} \right\}, \tag{5.17}$$

corresponding to square $m \times m$ matrices that are near low-rank (belonging to the ℓ_q-balls previously defined (5.15)), and have a logarithmic spikiness ratio. The following result applies

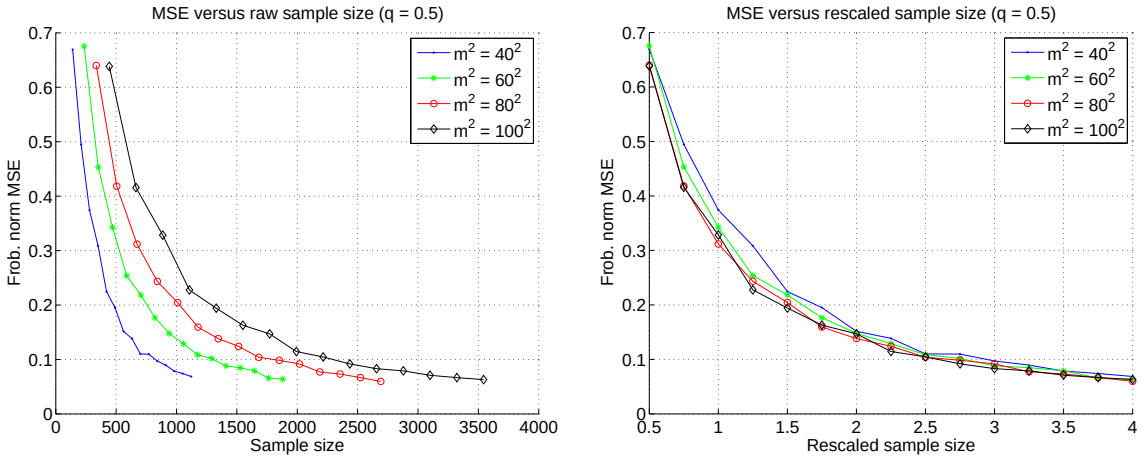


Figure 5.2. Plots of the mean-squared error in Frobenius norm for $q = 0.5$. Each curve corresponds to a different problem size $m^2 \in \{40^2, 60^2, 80^2, 100^2\}$. (a) MSE versus the raw sample size n. As expected, the curves shift to the right as m increases, since more samples should be required to achieve a given MSE for larger problems. (b) The same MSE plotted versus the rescaled sample size $n/(R_q^{\frac{1}{1-q/2}}\, m \log m)$. Consistent with Corollary 5.2, all the plots are now fairly well-aligned.

to the *minimax risk* in Frobenius norm, namely the quantity

$$\mathfrak{M}_n(\widetilde{\mathbb{B}}(R_q)) := \inf_{\widetilde{\Theta}} \sup_{\Theta^* \in \widetilde{\mathbb{B}}(R_q)} \mathbb{E}\big[\|\widetilde{\Theta} - \Theta^*\|_F^2\big], \tag{5.18}$$

where the infimum is taken over all estimators $\widetilde{\Theta}$ that are measurable functions of n samples.

Theorem 5.3. *There is a universal numerical constant $c_5 > 0$ such that*

$$\mathfrak{M}_n(\widetilde{\mathbb{B}}(R_q)) \geq c_5 \ \min\left\{ R_q \left(\frac{\nu^2 m}{n} \right)^{1-\frac{q}{2}}, \ \frac{\nu^2 m^2}{n} \right\}. \tag{5.19}$$

The term of primary interest in this bound is the first one—namely, $R_q \left(\frac{\nu^2 m}{n}\right)^{1-\frac{q}{2}}$. It is the dominant term in the bound whenever the ℓ_q-radius satisfies the bound

$$R_q \leq \left(\frac{\nu^2 m}{n} \right)^{\frac{q}{2}} m. \tag{5.20}$$

In the special case $q = 0$, corresponding the exactly low-rank case, the bound (5.20) always holds, since it reduces to requiring that the rank $r = \rho_0$ is less than or equal to m. In these regimes, Theorem 5.3 establishes that the upper bounds obtained in Corollaries 5.1 and 5.2 are minimax-optimal up to factors logarithmic in matrix dimension m.

5.3.4 Comparison to other work

We now turn to a detailed comparison of our bounds to those obtained in past work on noisy matrix completion, in particular the papers by Candes and Plan [29] (hereafter CP) and Keshavan et al. [71] (hereafter KMO). Both papers considered only the case of exactly low-rank matrices, corresponding to the special case of $q = 0$ in our notation. Since neither paper provided results for the general case of near-low rank matrices, nor the general result (with estimation and approximation errors) stated in Theorem 5.2, our discussion is mainly limited to comparing Corollary 5.1 to their results. So as to simplify discussion, we restate all results under the scalings used in this chapter[1] (i.e., with $|||\Theta^*|||_F = 1$).

Comparison of rates

Under the strong incoherence conditions required for exact matrix recovery (see below for discussion), Theorem 7 in CP give an bound on $|||\widehat{\Theta} - \Theta^*|||_F$ that depends on the Frobenius norm of the potentially adversarial error matrix $\Xi \in \mathbb{R}^{d_1 \times d_2}$, as defined by the noise variables $[\Xi]_{j(i)\,k(i)} = \widetilde{\xi}_i$ in our case. In the special case of stochastic noise, under the observation model (5.1) and the scalings of our chapter, as long as $n > m$, where $m = d_1 + d_2$—a condition certainly required for Frobenius norm consistency—we have $|||\Xi|||_F = \Theta(\nu\sqrt{n}/m)$ with high probability. Given this scaling, the CP upper bound takes the form

$$|||\widehat{\Theta} - \Theta^*|||_F \lesssim \nu \left\{ \sqrt{m} + \frac{\sqrt{n}}{m} \right\}. \tag{5.21}$$

Note that if the noise standard deviation ν tends to zero while the sample size n, matrix size d and rank r all remain fixed, then this bound guarantees that the Frobenius error tends to zero. This behavior as $\nu \to 0$ is intuitively reasonable, given that their proof technique is an extrapolation from the case of exact recovery for noiseless observations ($\nu = 0$). However, note that for any fixed noise deviation $\nu > 0$, the first term increases to infinity as the matrix dimension m increases, whereas the second term actually grows as the sample size n increases. Consequently, the CP results do not guarantee statistical consistency, unlike the bounds proved here.

Turning to a setting with adversarial noise, suppose that the error vector has Frobenius norm at most δ. A modification of our analysis yields error bounds of the form $|||\widehat{\Theta} - \Theta^*|||_F \lesssim \left\{ \frac{m^2}{\sqrt{n}}\delta + \sqrt{\frac{rm \log m}{n}} \right\}$. In the setting of square matrices with $\delta \geq \sqrt{\frac{r \log m}{m}}$, our result yields an upper bound tighter by a factor of order $\sqrt{m}$ better than those presented in CP. Last, as pointed out by a reviewer, the CP analysis does yield bounds for approximately low-rank matrices, in particular by writing $\Theta^* = \Pi_r(\Theta^*) + \Delta$, where Π_r is the Frobenius norm projection onto the space of rank r matrices, and $\Delta = \Theta^* - \Pi_r(\Theta^*)$ is the approximation

[1] The paper CP and KMO use two different sets of scaling, one with $|||\Theta^*|||_F = \Theta(m)$ and the other with $|||\Theta^*|||_F = \sqrt{r}$, so that some care is required in converting between results.

error. With this notation, their analysis guarantees error bounds of the form $\sqrt{m}|||\Delta|||_F$ with high probability, which is a weaker guarantee than our bound whenever $|||\Delta|||_F \geq c\sqrt{\frac{r \log m}{n}}$ and $n = \Omega(m \log m)$.

Keshavan et al. [71] analyzed alternative methods based on trimming and applying the SVD. For Gaussian noise, their methods guarantee bounds (with high probability) of the form

$$|||\widehat{\Theta} - \Theta^*|||_F \lesssim \nu \min\left\{\alpha\sqrt{\frac{d_2}{d_1}}, \kappa^2(\Theta^*)\right\}\sqrt{\frac{rd_2}{n}}, \tag{5.22}$$

where d_2/d_1 is the aspect ratio of Θ^*, and $\kappa(\Theta^*) = \frac{\sigma_{\max}(\Theta^*)}{\sigma_{\min}(\Theta^*)}$ is the condition number of Θ^*. This result is more directly comparable to our Corollary 5.1; apart from the additional factor involving either the aspect ratio or the condition number, it is sharper since it does not involve the factor $\log m$ present in our bound. For a fixed noise standard deviation ν, the bound (5.22) guarantees statistical consistency as long as $\frac{rd_2}{n}$ tends to zero. The most significant differences are the presence of the aspect ratio d_2/d_1 or the condition number $\kappa(\Theta^*)$ in the upper bound (5.22). The aspect ratio is a quantity that can be as small as one, or as large as d_2, so that the pre-factor in the bound (5.22) can scale in a dimension-dependent way. Similarly, for any matrix with rank larger than one, the condition number can be made arbitrarily large. For instance, in the rank two case, define a matrix with $\sigma_{\max}(\Theta^*) = \sqrt{1-\delta^2}$ and $\sigma_{\min}(\Theta^*) = \delta$, and consider the behavior as $\delta \to 0$. In contrast, our bounds are invariant to both the aspect ratio and the condition number of Θ^*.

Comparison of matrix conditions

We now turn to a comparison of the various *matrix incoherence assumptions* invoked in the analysis of CP and KMO, and comparison to our spikiness condition. As before, for clarity, we specialize our discussion to the square case ($m_r = m_c = m$), since the rectangular case is not essentially different. The matrix incoherence conditions are stated in terms of the singular value decomposition $\Theta^* = U\Sigma V^T$ of the target matrix. Here $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{m\times r}$ are matrices of the left and right singular vectors respectively, satisfying $U^TU = V^TV = I_{r\times r}$, whereas $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix of the singular values. The purpose of matrix incoherence is to enforce that the left and right singular vectors should not be aligned with the standard basis. Among other assumptions, the CP analysis imposes the incoherence conditions

$$\|UU^T - \frac{r}{m}I_{m\times m}\|_\infty \leq \mu\frac{\sqrt{r}}{m}, \qquad \|VV^T - \frac{r}{m}I_{m\times m}\|_\infty \leq \mu\frac{\sqrt{r}}{m}, \quad \text{and} \quad \|UV^T\|_\infty \leq \mu\frac{\sqrt{r}}{m}, \tag{5.23}$$

for some constant $\mu > 0$. Parts of the KMO analysis impose the related incoherence condition

$$\max_{j=1,\ldots,m}|UU^T|_{jj} \leq \mu_0\frac{r}{m}, \quad \text{and} \max_{j=1,\ldots,m}|VV^T|_{jj} \leq \mu_0\frac{r}{m}. \tag{5.24}$$

Both of these conditions ensure that the singular vectors are sufficiently "spread-out", so as not to be aligned with the standard basis.

A remarkable property of conditions (5.23) and (5.24) is that they exhibit *no dependence* on the singular values of Θ^*. If one is interested only in exact recovery in the noiseless setting, then this lack of dependence is reasonable. However, if approximate recovery is the goal—as is necessarily the case in the more realistic setting of noisy observations—then it is clear that a minimal set of sufficient conditions should also involve the singular values, as is the case for our spikiness measure $\alpha_{\mathrm{sp}}(\Theta^*)$. The following example gives a concrete demonstration of an instance where our conditions are satisfied, so that approximate recovery is possible, whereas the incoherence conditions are violated.

Example. Let $\Gamma \in \mathbb{R}^{m \times m}$ be a positive semidefinite symmetric matrix with rank $r-1$, Frobenius norm $|||\Gamma|||_F = 1$ and $\|\Gamma\|_\infty \leq c_0/m$. For a scalar parameter $t > 0$, consider the matrix

$$\Theta^* := \Gamma + te_1 e_1^T \tag{5.25}$$

where $e_1 \in \mathbb{R}^m$ is the canonical basis vector with one in its first entry, and zero elsewhere. By construction, the matrix Θ^* has rank at most r. Moreover, as long as $t = \mathcal{O}(1/m)$, we are guaranteed that our spikiness measure satisfies the bound $\alpha_{\mathrm{sp}}(\Theta^*) = \mathcal{O}(1)$. Indeed, we have $|||\Theta^*|||_F \geq |||\Gamma|||_F - t \ = 1 - t$, and hence

$$\alpha_{\mathrm{sp}}(\Theta^*) = \frac{m\|\Theta^*\|_\infty}{|||\Theta^*|||_F} \ \leq \ \frac{m\big(\|\Gamma\|_\infty + t\big)}{1-t} \ \leq \ \frac{c_0 + mt}{1-t} \ = \ \mathcal{O}(1).$$

Consequently, for any choice of Γ as specified above, Corollary 5.1 implies that the SDP will recover the matrix Θ^* up to a tolerance $\mathcal{O}(\sqrt{\frac{rm\log m}{n}})$. This captures the natural intuition that "poisoning" the matrix Γ with the term $te_1^T e_1$ should have essentially no effect, as long as t is not too large.

On the other hand, suppose that we choose the matrix Γ such that its $r-1$ eigenvectors are orthogonal to e_1. In this case, we have $\Theta^* e_1 = te_1$, so that e_1 is also an eigenvector of Θ^*. Letting $U \in \mathbb{R}^{m \times r}$ be the matrix of eigenvectors, we have $e_1^T UU^T e_1 = 1$. Consequently, for any fixed μ (or μ_0) and rank $r \ll m$, conditions (5.23) and (5.24) are violated.

$\diamondsuit$

5.4 Proofs for noisy matrix completion

We now turn to the proofs of our results. This section is devoted to the results that apply directly to noisy matrix completion, in particular the achievable result given in Theorem 5.2, its associated Corollaries 5.1 and 5.2, and the information-theoretic lower bound given in Theorem 5.3. The proof of Theorem 5.1 is provided in Section 5.5 to follow.

5.4.1 A useful transformation

We begin by describing a transformation that is useful both in these proofs, and the later proof of Theorem 5.1. In particular, we consider the mapping $\Theta \mapsto \Gamma := \sqrt{R}\Theta\sqrt{C}$, as well as the modified observation operator $\mathfrak{X}' : \mathbb{R}^{m\times m} \to \mathbb{R}^n$ with elements

$$[\mathfrak{X}'(\Gamma)]_i = \langle\langle \widetilde{X}^{(i)}, \Gamma \rangle\rangle, \quad \text{for } i = 1, 2, \ldots, n,$$

where $\widetilde{X}^{(i)} := R^{-1/2} X_i C^{-1/2}$. Note that $\mathfrak{X}'(\Gamma) = \mathfrak{X}(\Theta)$ by construction, and moreover

$$|||\Gamma|||_F = |||\Theta|||_{\omega(F)}, \quad |||\Gamma|||_{\text{nuc}} = |||\Theta|||_{\omega(1)}, \quad \text{and} \quad |||\Gamma|||_\infty = |||\Theta|||_{\omega(\infty)},$$

which implies that

$$\beta_{\text{ra}}(\Theta) = \underbrace{\frac{|||\Gamma|||_{\text{nuc}}}{|||\Gamma|||_F}}_{\beta'_{\text{ra}}(\Gamma)}, \quad \text{and} \quad \alpha_{\text{sp}}(\Theta) = \underbrace{\frac{m\,\|\Gamma\|_\infty}{|||\Gamma|||_F}}_{\alpha'_{\text{sp}}(\Gamma)}. \tag{5.26}$$

Based on this change of variables, let us define a modified version of the constraint set (5.8) as follows

$$\mathfrak{C}'(n; c_0) = \Big\{ 0 \neq \Gamma \in \mathbb{R}^{m\times m} \;\mid\; \alpha'_{\text{sp}}(\Gamma)\,\beta'_{\text{ra}}(\Gamma) \leq \frac{1}{c_0\, L}\sqrt{\frac{n}{m \log m}} \Big\}. \tag{5.27}$$

In this new notation, the lower bound (5.9) from Theorem 5.1 can be re-stated as

$$\frac{\|\mathfrak{X}'(\Gamma)\|_2}{\sqrt{n}} \geq \frac{1}{8}|||\Gamma|||_F \Big\{1 - \frac{128 L \alpha'_{\text{sp}}(\Gamma)}{\sqrt{n}}\Big\} \quad \text{for all } \Gamma \in \mathfrak{C}'(n; c_0). \tag{5.28}$$

5.4.2 Proof of Theorem 5.2

We now turn to the proof of Theorem 5.2. Defining the estimate $\widehat{\Gamma} := \sqrt{R}\widehat{\Theta}\sqrt{C}$, we have

$$\widehat{\Gamma} \in \arg \min_{\|\Gamma\|_\infty \leq \frac{\alpha^*}{\sqrt{m_r m_c}}} \Big\{ \frac{1}{2n}\|y - \mathfrak{X}'(\Gamma)\|_2^2 + \lambda_n |||\Gamma|||_{\text{nuc}} \Big\}, \tag{5.29}$$

and our goal is to upper bound the ordinary Frobenius norm $|||\widehat{\Gamma} - \Gamma^*|||_F$.

We now recall Lemma 4.1 from Chapter 4 and note that we adopt the shorthand $\Delta = \widehat{\Gamma} - \Gamma^*$ throughout the analysis. Lemma 4.1 establishes that there exists a matrix decomposition $\Delta = \Delta' + \Delta''$ of the error Δ such that

(a) The matrix Δ' satisfies the constraint $\text{rank}(\Delta') \leq 2r$, and

(b) Given the choice (5.12), the nuclear norm of Δ'' is bounded as

$$|||\Delta''|||_{\text{nuc}} \leq 3|||\Delta'|||_{\text{nuc}} + 4\sum_{j=r+1}^{m_r}\sigma_j(\Gamma^*).$$

The above bound combined with triangle inequality, implies that

$$\begin{aligned}|||\widehat{\Delta}|||_{\text{nuc}} \leq |||\Delta'|||_{\text{nuc}} + |||\Delta''|||_{\text{nuc}} &\leq 4|||\Delta'|||_{\text{nuc}} + 4\sum_{j=r+1}^{m_r}\sigma_j(\Gamma^*) \\ &\leq 8\sqrt{r}|||\widehat{\Delta}|||_F + 4\sum_{j=r+1}^{m_r}\sigma_j(\Gamma^*) \qquad (5.30)\end{aligned}$$

where the second inequality uses the fact that $\text{rank}(\Delta') \leq 2r$.

We now split into two cases, depending on whether or not the error $\widehat{\Delta}$ belongs to the set $\mathfrak{C}'(n; c_0)$.

Case 1: First suppose that $\widehat{\Delta} \notin \mathfrak{C}'(n; c_0)$. In this case, by the definition (5.27), we have

$$\begin{aligned}|||\widehat{\Delta}|||_F^2 &\leq c_0 L\left(\sqrt{m_r m_c}\|\widehat{\Delta}\|_\infty\right)|||\widehat{\Delta}|||_{\text{nuc}}\sqrt{\frac{m\log m}{n}} \\ &\leq 2c_0 L\alpha^*|||\widehat{\Delta}|||_{\text{nuc}}\sqrt{\frac{m\log m}{n}},\end{aligned}$$

since $\|\widehat{\Delta}\|_\infty \leq \|\Gamma^*\|_\infty + \|\widehat{\Gamma}\|_\infty \leq \frac{2\alpha^*}{\sqrt{m_r m_c}}$. Now applying the bound (5.30), we obtain

$$|||\widehat{\Delta}|||_F^2 \leq 2c_0\, L\,\alpha^*\sqrt{\frac{m\log m}{n}}\left\{8\sqrt{r}|||\widehat{\Delta}|||_F + 4\sum_{j=r+1}^{m_r}\sigma_j(\Gamma^*)\right\}. \qquad (5.31)$$

Case 2: Otherwise, we must have $\widehat{\Delta} \in \mathfrak{C}'(n; c_0)$. Recall the reformulated lower bound (5.28). On one hand, if $\frac{128L\alpha'_{\text{sp}}(\widehat{\Delta})}{\sqrt{n}} > 1/2$, then we have

$$|||\widehat{\Delta}|||_F \leq \frac{256L\sqrt{m_r m_c}\|\widehat{\Delta}\|_\infty}{\sqrt{n}} \leq \frac{512\, L\,\alpha^*}{\sqrt{n}}. \qquad (5.32)$$

On the other hand, if $\frac{128L\alpha'_{\text{sp}}(\widehat{\Delta})}{\sqrt{n}} \leq 1/2$, then from the bound (5.28), we have

$$\frac{\|\mathfrak{X}'(\widehat{\Delta})\|_2}{\sqrt{n}} \geq \frac{|||\widehat{\Delta}|||_F}{16} \qquad (5.33)$$

with high probability. Note that $\widehat{\Gamma}$ is optimal and Γ^* is feasible for the convex program (5.29), so that we have the basic inequality

$$\frac{1}{2n}\|y - \mathfrak{X}'(\widehat{\Gamma})\|_2^2 + \lambda_n |\!|\!|\widehat{\Gamma}|\!|\!|_{\text{nuc}} \leq \frac{1}{2n}\|y - \mathfrak{X}'(\Gamma^*)\|_2^2 + \lambda_n |\!|\!|\Gamma^*|\!|\!|_{\text{nuc}}.$$

Some algebra then yields the inequality

$$\frac{1}{2n}\|\mathfrak{X}'(\widehat{\Delta})\|_2^2 \leq \nu \, \langle\!\langle \widehat{\Delta}, \, \frac{1}{n}\sum_{i=1}^n \xi_i \widetilde{X}^{(i)} \rangle\!\rangle + \lambda_n |\!|\!|\Gamma^*|\!|\!|_{\text{nuc}} - \lambda_n |\!|\!|\Gamma^* + \widehat{\Delta}|\!|\!|_{\text{nuc}},$$

Substituting the lower bound (5.33) into this inequality yields

$$\frac{\|\widehat{\Delta}\|_F^2}{512} \leq \nu \, \langle\!\langle \widehat{\Delta}, \, \frac{1}{n}\sum_{i=1}^n \xi_i \widetilde{X}^{(i)} \rangle\!\rangle + \lambda_n |\!|\!|\Gamma^*|\!|\!|_{\text{nuc}} - \lambda_n |\!|\!|\Gamma^* + \widehat{\Delta}|\!|\!|_{\text{nuc}}.$$

From this point onwards, the proof is identical (apart from constants) to Theorem 4.1, and we obtain that there is a numerical constant c_1 such that

$$|\!|\!|\Delta|\!|\!|_F^2 \leq c_1 \, \alpha^* \, \lambda_n \Big\{ \sqrt{r} |\!|\!|\Delta|\!|\!|_F + \sum_{j=r+1}^{m_r} \sigma_j(\Gamma^*) \Big\}. \tag{5.34}$$

Putting together the pieces: Summarizing our results, we have shown that with high probability, one of the three bounds (5.31), (5.32) or (5.34) must hold. These claims can be summarized in the form

$$|\!|\!|\Delta|\!|\!|_F^2 \leq c_1 \, \alpha^* \, \max\Big\{ \lambda_n, \sqrt{\frac{m \log m}{n}} \Big\} \, \big[\sqrt{r} |\!|\!|\Delta|\!|\!|_F + \sum_{j=r+1}^{m_r} \sigma_j(\Gamma^*) \big].$$

for a universal positive constant c_1. Translating this result back to the original co-ordinate system ($\Gamma^* = \sqrt{R}\Theta^*\sqrt{C}$) yields the claim (5.13).

5.4.3 Proof of Corollary 5.1

When Θ^* (and hence $\sqrt{R}\Theta^*\sqrt{C}$) has rank $r < m_r$, then we have $\sum_{j=r+1}^{m_r} \sigma_j(\sqrt{R}\Theta^*\sqrt{C}) = 0$. Consequently, the bound (5.13) reduces to $|\!|\!|\widetilde{\Delta}|\!|\!|_{\omega(F)} \leq c_1 \, \alpha^* \, \lambda_n^* \sqrt{r}$. To complete the proof, it suffices to show that

$$\mathbb{P}\big[|\!|\!| \frac{1}{n}\sum_{i=1}^n \xi_i R^{-1/2} X_i C^{-1/2} |\!|\!|_2 \geq c_1 \, \nu \sqrt{\frac{m \log m}{n}} \big] \leq c_2 \exp(-c_2 m \log m).$$

We do so via the Ahlswede-Winter matrix bound, as stated in Appendix C.6. Defining the random matrix $Y^{(i)} := \xi_i R^{-1/2} \, X_i \, C^{-1/2}$, we first note that ξ_i is sub-exponential with

parameter 1, and $|R^{-1/2} X_i C^{-1/2}|$ has a single entry with magnitude at most $L\sqrt{m_r m_c}$, which implies that

$$\|Y^{(i)}\|_{\psi_1} \leq L\,\nu\,\sqrt{m_r m_c} \leq 2\nu\, Lm$$

(Here $\|\cdot\|_{\psi_1}$ denotes the Orlicz norm [78] of a random variable, as defined by the function $\psi_1(x) = \exp(x) - 1$; see Appendix C.6). Moreover, we have

$$\begin{aligned}
\mathbb{E}[(Y^{(i)})^T Y^{(i)}] &= \nu^2\,\mathbb{E}\Big[\frac{m_r m_c}{R_{j(i)}\,C_{k(i)}} e_{k(i)} e_{j(i)}^T e_{j(i)} e_{k(i)}^T\Big] \\
&= \nu^2\,\mathbb{E}\Big[\frac{m_r m_c}{R_{j(i)}\,C_{k(i)}} e_{k(i)} e_{k(i)}^T\Big] \\
&= \nu^2\, m_r I_{m_c \times m_c}.
\end{aligned}$$

so that $\|\mathbb{E}[(Y^{(i)})^T Y^{(i)}]\|_2 \leq 2\nu^2 m$, recalling that $2m = m_r + m_c \geq m_r$. The same bound applies to $\|\mathbb{E}[Y^{(i)}(Y^{(i)})^T]\|_2$, so that applying Lemma C.2 with $t = n\delta$, we conclude that

$$\mathbb{P}\big[\|\frac{1}{n}\sum_{i=1}^n \xi_i R^{-1/2} X_i C^{-1/2}\|_2 \geq \delta\big] \leq (m_r \times m_c)\,\max\big\{\exp(-n\delta^2/(16\nu^2 m)),\ \exp(-\frac{n\delta}{4\nu\, Lm})\big\}$$

Since $\sqrt{m_r m_c} \leq m_r + m_c = 2m$, if we set $\delta^2 = c_1^2 \nu^2 \frac{m\log m}{n}$ for a sufficiently large constant c_1, the result follows. (Here we also use the assumption that $n = \Omega(Lm\log m)$, so that the term $\sqrt{\frac{m\log m}{n}}$ is dominant.)

5.4.4 Proof of Corollary 5.2

For this corollary, we need to determine an appropriate choice of r so as to optimize the bound (5.13). To ease notation, let us make use of the shorthand notation $\Gamma^* = \sqrt{R}\Theta^*\sqrt{C}$. With the singular values of Γ^* ordered in non-increasing order, fix some threshold $\tau > 0$ to be determined, and set $r = \max\{j \mid \sigma_j(\Gamma^*) > \tau\}$. This choice ensures that

$$\sum_{j=r+1}^{m_r} \sigma_j(\Gamma^*) = \tau \sum_{j=r+1}^{m_r} \frac{\sigma_j(\Gamma^*)}{\tau} \leq \tau \sum_{j=r+1}^{m_r} \big(\frac{\sigma_j(\Gamma^*)}{\tau}\big)^q \leq \tau^{1-q} R_q.$$

Moreover, we have $r\,\tau^q \leq \sum_{j=1}^r \big\{\sigma_j(\Gamma^*)\big\}^q \leq R_q$, which implies that $\sqrt{r} \leq \sqrt{R_q}\tau^{-q/2}$. Substituting these relations into the upper bound (5.13) leads to

$$\|\widetilde{\Delta}\|_{\omega(F)}^2 \leq c_1\ \alpha^*\ \lambda_n^*\ \big[\sqrt{R_q}\tau^{-q/2}\|\widetilde{\Delta}\|_{\omega(F)} + \tau^{1-q} R_q\big\}$$

In order to obtain the sharpest possible upper bound, we set $\tau = \alpha^*\lambda_n^*$. Following some algebra, we find that there is a universal constant c_1 such that

$$\|\widetilde{\Delta}\|_{\omega(F)}^2 \leq c_1 R_q\,\big((\alpha^*)^2(\lambda_n^*)^2\big)^{1-\frac{q}{2}}.$$

As in the proof of Corollary 5.1, it suffices to choose $\lambda_n = \Omega(\nu\sqrt{\frac{m\log m}{n}})$, so that $\lambda_n^* = \mathcal{O}\sqrt{(\nu^2 + L)\frac{m\log m}{n}})$, from which the claim follows.

5.4.5 Proof of Theorem 5.3

Our proof of this lower bound based on a combination of information-theoretic methods [151, 150], which allow us to reduce to a multiway hypothesis test, and an application of the probabilistic method so as to construct a suitably large packing set. By Markov's inequality, it suffices to prove that

$$\sup_{\Theta^* \in \widetilde{\mathbb{B}}(R_q)} \mathbb{P}\left[|||\widehat{\Theta} - \Theta^*|||_F^2 \geq \frac{\delta^2}{4} \right] \geq \frac{1}{2}.$$

In order to do so, we proceed in a standard way—namely, by reducing the estimation problem to a testing problem over a suitably constructed packing set contained within $\widetilde{\mathbb{B}}(R_q)$. In particular, consider a set $\{\Theta^1, \ldots, \Theta^{M(\delta)}\}$ of matrices, contained within $\widetilde{\mathbb{B}}(R_q)$, such that $|||\Theta^k - \Theta^\ell|||_F \geq \delta$ for all $\ell \neq k$. To ease notation, we use M as shorthand for $M(\delta)$ through much of the argument. Suppose that we choose an index $V \in \{1, 2, \ldots, M\}$ uniformly at random (u.a.r.), and we are given observations $y \in \mathbb{R}^n$ from the observation model (5.3) with $\Theta^* = \Theta^V$. Then triangle inequality yields the lower bound

$$\sup_{\Theta^* \in \widetilde{\mathbb{B}}(R_q)} \mathbb{P}\left[|||\widehat{\Theta} - \Theta^*|||_F \geq \frac{\delta}{2} \right] \geq \mathbb{P}[\widehat{V} \neq V].$$

If we condition on $\mathfrak{X}$, a variant of Fano's inequality yields

$$\mathbb{P}[\widehat{V} \neq V \mid \mathfrak{X}] \geq 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell \neq k} D(\Theta^k \,\|\, \Theta^\ell) + \log 2}{\log M}, \tag{5.35}$$

where $D(\Theta^k \,\|\, \Theta^\ell)$ denotes the Kullback-Leibler divergence between the distributions of $(y|\mathfrak{X}, \Theta^k)$ and $(y|\mathfrak{X}, \Theta^\ell)$. In particular, for additive Gaussian noise with variance ν^2, we have

$$D(\Theta^k \,\|\, \Theta^\ell) = \frac{1}{2\nu^2} \|\mathfrak{X}(\Theta^k) - \mathfrak{X}(\Theta^\ell)\|_2^2,$$

and moreover,

$$\mathbb{E}_{\mathfrak{X}}\big[D(\Theta^k \,\|\, \Theta^\ell)\big] = \frac{1}{2\nu^2} |||\Theta^k - \Theta^\ell|||_F^2.$$

Combined with the bound (5.35), we obtain the bound

$$\begin{aligned} \mathbb{P}[\widehat{V} \neq V] &= \mathbb{E}_{\mathfrak{X}}\big\{\mathbb{P}[\widehat{V} \neq V \mid \mathfrak{X}]\big\} \\ &\geq 1 - \frac{\binom{M}{2}^{-1} \sum_{\ell \neq k} \frac{n}{2\nu^2} |||\Theta^k - \Theta^\ell|||_F^2 + \log 2}{\log M}, \end{aligned} \tag{5.36}$$

The remainder of the proof hinges on the following technical lemma, which we prove in Appendix C.1.

Lemma 5.1. *Let $m \geq 10$ be a positive integer, and let $\delta > 0$. Then for each $r = 1, 2, \ldots, m$, there exists a set of m-dimensional matrices $\{\Theta^1, \ldots, \Theta^M\}$ with cardinality $M = \lfloor \frac{1}{4} \exp\left(\frac{rm}{128}\right) \rfloor$ such that each matrix has rank r, and moreover*

$$|||\Theta^\ell|||_F = \delta \qquad \text{for all } \ell = 1, 2, \ldots, M, \tag{5.37a}$$

$$|||\Theta^\ell - \Theta^k|||_F \geq \delta \qquad \text{for all } \ell \neq k, \tag{5.37b}$$

$$\alpha_{\mathrm{sp}}(\Theta^\ell) \leq \sqrt{32 \log m} \quad \text{for all } \ell = 1, 2, \ldots, M, \text{ and} \tag{5.37c}$$

$$|||\Theta^\ell|||_2 \leq \frac{4\delta}{\sqrt{r}} \quad \text{for all } \ell = 1, 2, \ldots, M. \tag{5.37d}$$

We now show how to use this packing set in our Fano bound. To avoid technical complications, we assume throughout that $rm > 1024 \log 2$. Note that packing set from Lemma 5.1 satisfies $|||\Theta^k - \Theta^\ell|||_F \leq 2\delta$ for all $k \neq \ell$, and hence from Fano bound (5.36), we obtain

$$\begin{aligned}
\mathbb{P}[\widehat{V} \neq V] &\geq 1 - \frac{2\frac{n\delta^2}{\nu^2} + \log 2}{\frac{rm}{128} - \log 4} \\
&\geq 1 - \frac{2\frac{n\delta^2}{\nu^2} + \log 2}{\frac{rm}{256}} \\
&\geq 1 - \frac{512\frac{n\delta^2}{\nu^2} + 256 \log 2}{rm}.
\end{aligned}$$

If we now choose $\delta^2 = \frac{\nu^2}{2048}\frac{rm}{n}$, then

$$\mathbb{P}[\widehat{V} \neq V] \geq 1 - \frac{\frac{rm}{4} + 256 \log 2}{rm} \geq \frac{1}{2},$$

where the final inequality again uses the bound $rm \geq 1024 \log 2$.

In the special case $q = 0$, the proof is complete, since the elements Θ^ℓ all have rank $r = R_0$, and satisfy the bound $\alpha_{\mathrm{sp}}(\Theta^\ell) \leq \sqrt{32 \log m}$. For $q \in (0, 1]$, consider the matrix class $\widetilde{\mathbb{B}}(R_q)$, and let us set $r = \min\{m, \lceil R_q \left(\frac{m}{n}\right)^{-\frac{q}{2}} \rceil\}$ in Lemma 5.1. With this choice, since each matrix Θ^ℓ has rank r, we have

$$\sum_{j=1}^{d} \sigma_i(\Theta^\ell)^q \leq r \left(\frac{\delta}{\sqrt{r}}\right)^q = r\left(\frac{1}{2048}\sqrt{\frac{m}{n}}\right)^q \leq R_q,$$

so that we are guaranteed that $\Theta^\ell \in \widetilde{\mathbb{B}}(R_q)$. Finally, we note that

$$\frac{rm}{n} \geq \min\left\{R_q \left(\frac{m}{n}\right)^{1-\frac{q}{2}}, \frac{m^2}{n}\right\},$$

so that we conclude that the minimax error is lower bounded by

$$\frac{1}{4096}\min\left\{R_q\left(\frac{\nu^2 m}{n}\right)^{1-\frac{q}{2}}, \frac{\nu^2 m^2}{n}\right\}$$

for mr sufficiently large. (At the expense of a worse pre-factor, the same bound holds for all $m \geq 10$.)

5.5 Proof of Theorem 5.1

We now turn to the proof that the sampling operator in weighted matrix completion satisfies restricted strong convexity over the set $\mathcal{D}$, as stated in Theorem 5.1. In order to lighten notation, we prove the theorem in the case $m_r = m_c$. In terms of rates, this is a worst-case assumption, effectively amounting to replacing both m_r and m_c by the worst-case $\max\{m_r, m_c\}$. However, since our rates are driven by $m = \frac{1}{2}(m_r + m_c)$ and we have the inequalities

$$\frac{1}{2}\max\{m_r, m_c\} \leq \frac{1}{2}(m_r + m_c) \leq \max\{m_r, m_c\},$$

this change has only an effect on the constant factors. The proof can be extended to the general setting $m_r \neq m_c$ by appropriate modifications if these constant factors are of interest.

5.5.1 Reduction to simpler events

In order to prove Theorem 5.1, it is equivalent to show that, with high probability, we have

$$\frac{\|\mathfrak{X}'(\Gamma)\|_2}{\sqrt{n}} \geq \frac{1}{8}|||\Gamma|||_F - \frac{48L\, m\, \|\Gamma\|_\infty}{\sqrt{n}} \qquad \text{for all } \Gamma \in \mathfrak{C}'(n; c_0). \tag{5.38}$$

The remainder of the proof is devoted to studying the "bad" event

$$\mathcal{E}(\mathfrak{X}') := \left\{\exists \quad \Gamma \in \mathfrak{C}'(n; c_0) \;\mid\; \left|\frac{\|\mathfrak{X}'(\Gamma)\|_2}{\sqrt{n}} - |||\Gamma|||_F\right| > \frac{7}{8}|||\Gamma|||_F + \frac{48L\, m\, \|\Gamma\|_\infty}{\sqrt{n}}\right\}. \tag{5.39}$$

Suppose that $\mathcal{E}(\mathfrak{X}')$ does *not* hold: then we have

$$\left|\frac{\|\mathfrak{X}'(\Gamma)\|_2}{\sqrt{n}} - |||\Gamma|||_F\right| \leq \frac{7}{8}|||\Gamma|||_F + \frac{48L\, m\, \|\Gamma\|_\infty}{\sqrt{n}} \qquad \text{for all } \Gamma \in \mathfrak{C}'(n; c_0),$$

which implies that the bound (5.38) holds. Consequently, in terms of the "bad" event, the claim of Theorem 5.1 is implied by the tail bound $\mathbb{P}[\mathcal{E}(\mathfrak{X}')] \leq 16\exp(-c' m \log m)$.

We now show that in order to establish a tail bound on $\mathcal{E}(\mathfrak{X}')$, it suffices to bound the probability of some simpler events $\mathcal{E}(\mathfrak{X}'; R_{\mathrm{P}})$, defined below. Since the definition of the

set $\mathfrak{C}'(n; c_0)$ and event $\mathcal{E}(\mathfrak{X}')$ is invariant to rescaling of Γ, we may assume without loss of generality that $\|\Gamma\|_\infty = \frac{1}{m}$. The remaining degrees of freedom in the set $\mathfrak{C}'(n; c_0)$ can be parameterized in terms of the quantities $R_{\mathrm{P}} = |\!|\!|\Gamma|\!|\!|_F$ and $\Upsilon = |\!|\!|\Gamma|\!|\!|_{\mathrm{nuc}}$. For any $\Gamma \in \mathfrak{C}'(n; c_0)$ with $\|\Gamma\|_\infty = \frac{1}{m}$ and $|\!|\!|\Gamma|\!|\!|_F \leq R_{\mathrm{P}}$, we have $|\!|\!|\Gamma|\!|\!|_{\mathrm{nuc}} \leq \Upsilon(R_{\mathrm{P}})$, where

$$\Upsilon(R_{\mathrm{P}}) := \frac{R_{\mathrm{P}}^2}{c_0 \, L \sqrt{\frac{m \log m}{n}}}.$$

For each radius $R_{\mathrm{P}} > 0$, consider the set

$$\mathcal{R}(R_{\mathrm{P}}) := \big\{\Gamma \in \mathfrak{C}'(n; c_0) \;\mid\; \|\Gamma\|_\infty = \frac{1}{m}, \; |\!|\!|\Gamma|\!|\!|_F \leq R_{\mathrm{P}}, \; |\!|\!|\Gamma|\!|\!|_{\mathrm{nuc}} \leq \Upsilon(R_{\mathrm{P}})\big\}, \tag{5.40}$$

and the associated event

$$\mathcal{E}(\mathfrak{X}'; R_{\mathrm{P}}) := \Big\{\exists \quad \Gamma \in \mathcal{R}(R_{\mathrm{P}}) \;\mid\; \Big|\frac{\|\mathfrak{X}'(\Gamma)\|_2}{\sqrt{n}} - |\!|\!|\Gamma|\!|\!|_F\Big| \geq \frac{3}{4} R_{\mathrm{P}} + \frac{48 L}{\sqrt{n}}\Big\}. \tag{5.41}$$

The following lemma shows that it suffices to upper bound the probability of the event $\mathcal{E}(\mathfrak{X}'; R_{\mathrm{P}})$ for each fixed $R_{\mathrm{P}} > 0$.

Lemma 5.2. *Suppose that are universal constants (c_1, c_2) such that*

$$\mathbb{P}[\mathcal{E}(\mathfrak{X}'; R_{\mathrm{P}})] \leq c_1 \exp(-c_2 n R_{\mathrm{P}}^2) \tag{5.42}$$

for each fixed $R_{\mathrm{P}} > 0$. Then there is a universal constant c_2' such that

$$\mathbb{P}[\mathcal{E}(\mathfrak{X}')] \leq c_1 \frac{\exp(-c_2' m \log m)}{1 - \exp(-c_2' m \log m)}. \tag{5.43}$$

The proof of this claim, provided in Appendix C.2, follows by a peeling argument.

5.5.2 Bounding the probability of $\mathcal{E}(\mathfrak{X}'; R_{\mathrm{P}})$

Based on Lemma 5.2, it suffices to prove the tail bound (5.42) on the event $\mathcal{E}(\mathfrak{X}'; R_{\mathrm{P}})$ for each fixed $R_{\mathrm{P}} > 0$. Let us define

$$Z_n(R_{\mathrm{P}}) := \sup_{\Gamma \in \overline{\mathcal{R}}(R_{\mathrm{P}})} \left| \frac{\|\mathfrak{X}'(\Gamma)\|_2}{\sqrt{n}} - |\!|\!|\Gamma|\!|\!|_F \right|, \tag{5.44}$$

where

$$\overline{\mathcal{R}}(R_{\mathrm{P}}) := \big\{\Gamma \in \mathfrak{C}'(n; c_0) \;\mid\; \|\Gamma\|_\infty \leq \frac{1}{m}, \; |\!|\!|\Gamma|\!|\!|_F \leq R_{\mathrm{P}}, \; |\!|\!|\Gamma|\!|\!|_{\mathrm{nuc}} \leq \Upsilon(R_{\mathrm{P}})\big\}. \tag{5.45}$$

(The only difference from $\mathcal{R}(R_{\mathrm{P}})$ is that we have relaxed to the inequality $\|\Gamma\|_\infty \leq \frac{1}{m}$.) In the remainder of this section, we prove that there are universal constants (c_1, c_2) such that

$$\mathbb{P}\big[Z_n(R_{\mathrm{P}}) \geq \frac{3}{4}R_{\mathrm{P}} + \frac{48L}{\sqrt{n}}\big] \leq c_1 \exp(-c_2 \frac{nR_{\mathrm{P}}^2}{L^2}) \quad \text{for each fixed } R_{\mathrm{P}} > 0. \tag{5.46}$$

This tail bound means that the condition of Lemma 5.2 is satisfied, and so completes the proof of Theorem 5.1.

In order to prove (5.46), we begin with a discretization argument. Let $\Gamma^1, \ldots, \Gamma^{N(\delta)}$ be a δ-covering of $\overline{\mathcal{R}}(R_{\mathrm{P}})$ in the Frobenius norm. By definition, given an arbitrary $\Gamma \in \overline{\mathcal{R}}(R_{\mathrm{P}})$, there exists some index $k \in \{1, \ldots, N(\delta)\}$ and a matrix $\Delta \in \mathbb{R}^{m \times m}$ with $|||\Delta|||_F \leq \delta$ such that $\Gamma = \Gamma^k + \Delta$. Therefore, we have

$$\begin{aligned}
\frac{\|\mathfrak{X}'(\Gamma)\|_2}{\sqrt{n}} - |||\Gamma|||_F &= \frac{\|\mathfrak{X}'(\Gamma^k + \Delta)\|_2}{\sqrt{n}} - |||\Gamma^k + \Delta|||_F \\
&\leq \frac{\|\mathfrak{X}'(\Gamma^k)\|_2}{\sqrt{n}} + \frac{\|\mathfrak{X}'(\Delta)\|_2}{\sqrt{n}} - |||\Gamma^k|||_F + |||\Delta|||_F \\
&\leq \Big|\frac{\|\mathfrak{X}'(\Gamma^k)\|_2}{\sqrt{n}} - |||\Gamma^k|||_F\Big| + \frac{\|\mathfrak{X}'(\Delta)\|_2}{\sqrt{n}} + \delta,
\end{aligned}$$

where we have used the triangle inequality. Following the same steps establishes that this inequality holds for the absolute value of the difference.

Moreover, since $\Delta = \Gamma^k - \Gamma$ with both Γ^k and Γ belonging to $\overline{\mathcal{R}}(R_{\mathrm{P}})$, we have $|||\Delta|||_{\mathrm{nuc}} \leq 2\Upsilon(R_{\mathrm{P}})$ and $\|\Delta\|_\infty \leq \frac{2}{m}$, where we have used the definition (5.40). Putting together the pieces, we conclude that

$$Z_n(R_{\mathrm{P}}) \quad \leq \quad \delta \quad + \quad \max_{k=1,\ldots,N(\delta)} \Big|\frac{\|\mathfrak{X}'(\Gamma^k)\|_2}{\sqrt{n}} \quad - \quad |||\Gamma^k|||_F\Big| \quad + \quad \sup_{\Delta \in \mathfrak{D}(\delta, R)} \Big|\frac{\|\mathfrak{X}'(\Delta)\|_2}{\sqrt{n}}\Big|, \tag{5.47}$$

where

$$\mathfrak{D}(\delta, R) := \big\{\Delta \in \mathbb{R}^{m \times m} \mid |||\Delta|||_F \leq \delta,\ |||\Delta|||_{\mathrm{nuc}} \leq 2\Upsilon(R_{\mathrm{P}}),\ \|\Delta\|_\infty \leq \frac{2}{m}\big\}. \tag{5.48}$$

Note that the bound (5.47) holds for any choice of $\delta > 0$. We establish the tail bound (5.46) with the choice $\delta = R_{\mathrm{P}}/8$, and using the following two lemmas. The first lemma provides control of the maximum over the covering set:

Lemma 5.3. *As long $m \geq 10$, we have*

$$\max_{k=1,\ldots,N(R_{\mathrm{P}}/8)} \Big|\frac{\|\mathfrak{X}'(\Gamma^k)\|_2}{\sqrt{n}} - |||\Gamma^k|||_F\Big| \leq \frac{R_{\mathrm{P}}}{8} + \frac{48L}{\sqrt{n}} \tag{5.49}$$

with probability greater than $1 - c\exp\big(-\frac{nR_{\mathrm{P}}^2}{2048\,L^2}\big)$.

See Appendix C.3 for the proof of this claim.

Our second lemma, proved in Appendix C.4, provides control over the final term in the upper bound (5.47).

Lemma 5.4.

$$\sup_{\Delta \in \mathfrak{D}(\frac{R_{\mathrm{P}}}{8}, R)} \Big| \frac{\|\mathfrak{X}'(\Delta)\|_2}{\sqrt{n}} \Big| \leq \frac{R_{\mathrm{P}}}{2}$$

with probability at least $1 - 2\exp\big(-\frac{nR_{\mathrm{P}}^2}{8192L^2}\big)$.

Combining these two lemmas with the upper bound (5.47) with $\delta = R_{\mathrm{P}}/8$, we obtain

$$\begin{aligned} Z_n(R_{\mathrm{P}}) &\leq \frac{R_{\mathrm{P}}}{8} + \frac{R_{\mathrm{P}}}{8} + \frac{48L}{\sqrt{n}} + \frac{R_{\mathrm{P}}}{2} \\ &\leq \frac{3R_{\mathrm{P}}}{4} + \frac{48L}{\sqrt{n}} \end{aligned}$$

with probability at least $1 - 4\exp\big(-\frac{nR_{\mathrm{P}}^2}{8192}\big)$, thereby establishing the tail bound (5.46) and completing the proof of Theorem 5.1.

5.6 Discussion

In this chapter, we have established error bounds for the problem of weighted matrix completion based on partial and noisy observations. We proved both a general result, one which applies to any matrix, and showed how it yields corollaries for both the cases of exactly low-rank and approximately low-rank matrices. A key technical result is establishing that the matrix sampling operator satisfies a suitable form of restricted strong convexity 3.2.4 over a set of matrices with controlled rank and spikiness. Since more restrictive properties such as RIP do not hold for matrix completion, this RSC ingredient is essential to our analysis. Our proof of the RSC condition relied on a number of techniques from empirical process and random matrix theory, including concentration of measure, contraction inequalities and the Ahlswede-Winter bound. Using information-theoretic methods, we also proved that up to logarithmic factors, our error bounds cannot be improved upon by any algorithm, showing that our method is essentially minimax-optimal.

There are various open questions that remain to be studied. Although our analysis applies to both uniform and non-uniform sampling models, it is limited to the case where each row (or column) is sampled with a certain probability. It would be interesting to consider extensions to settings in which the sampling probability differed from entry to entry, as investigated empirically by Salakhutdinov and Srebro [122]. Although we have focused on least-squares

losses in this chapter, the notion of restricted strong convexity applies to more general loss functions. Indeed, it should be possible to combine the results of this paper with Proposition 2 in Negahban et al. [100] so as to obtain bounds for matrix completion with general losses.

Chapter 6

Structured Optimization

6.1 Introduction

We now present the results that allow us to exploit the same statistical structures utilized above for computational efficiency rather than statistical improvements. The crux of our argument is leveraging the statistical structure to allow us to relate ill-behaved empirical quantities with their well-conditioned population counterparts. In the remainder of this chapter we begin in Section 6.2 with a precise formulation of the class of convex programs analyzed. We will then recall certain desired properties of the loss function. Section 6.3 is devoted to the statement of our main convergence result, as well as to the development and discussion of its various corollaries for specific statistical models. In Section 6.4, we provide a number of empirical results that confirm the sharpness of our theoretical predictions. Finally, Section 6.5 contains the proofs, with more technical aspects of the arguments deferred to the Appendix D.

6.2 Background and problem formulation

In this section, we begin by describing the class of regularized M-estimators to which our analysis applies, as well as the optimization algorithms that we analyze. Finally, we introduce some important notions that underlie our analysis, including the notions of a decomposable regularization, and the properties of restricted strong convexity and smoothness.

6.2.1 Loss functions, regularization and gradient-based methods

We recall that given a random variable $Z \sim \mathbb{P}$ taking values in some set $\mathcal{Z}$, let $Z_1^n = \{Z_1, \ldots, Z_n\}$ be a collection of n observations. Here the integer n is the *sample size* of the problem. Assuming that $\mathbb{P}$ lies within some indexed family $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$, the goal is to recover an estimate of the unknown true parameter $\theta^* \in \Omega$ generating the data. Here Ω is some subset

of $\mathbb{R}^d$, and the integer d is known as the *ambient dimension* of the problem. In order to measure the "fit" of any given parameter $\theta \in \Omega$ to a given data set Z_1^n, we introduce a loss function $\mathcal{L}_n : \Omega \times \mathcal{Z}^n \to \mathbb{R}_+$. By construction, for any given n-sample data set $Z_1^n \in \mathcal{Z}^n$, the loss function assigns a cost $\mathcal{L}_n(\theta; Z_1^n) \geq 0$ to the parameter $\theta \in \Omega$. In many (but not all) applications, the loss function has a separable structure across the data set, meaning that $\mathcal{L}_n(\theta; Z_1^n) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i)$ where $\ell : \Omega \times \mathcal{Z} :\to \mathbb{R}_+$ is the loss function associated with a single data point.

Of primary interest in this chapter are estimation problems that are under-determined, meaning that the number of observations n is smaller than the ambient dimension d. In such settings, without further restrictions on the parameter space Ω, there are various impossibility theorems, asserting that consistent estimates of the unknown parameter θ^* cannot be obtained. For this reason, it is necessary to assume that the unknown parameter θ^* either lies within a smaller subset of Ω, or is well-approximated by some member of such a subset. In order to incorporate these types of structural constraints, we introduce a *regularizer* $\mathcal{R} : \Omega \to \mathbb{R}_+$ over the parameter space. With these ingredients, the analysis of this chapter applies to the *constrained M-estimator*

$$\widehat{\theta}_\rho \in \arg\min_{\mathcal{R}(\theta) \leq \rho} \{\mathcal{L}_n(\theta; Z_1^n)\}, \tag{6.1}$$

where $\rho > 0$ is a user-defined radius, as well as to the *regularized M-estimator*

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\mathcal{R}(\theta) \leq \bar{\rho}} \big\{ \underbrace{\mathcal{L}_n(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta)}_{\phi_n(\theta)} \big\} \tag{6.2}$$

where the regularization weight $\lambda_n > 0$ is user-defined. Note that the radii ρ and $\bar{\rho}$ may be different in general. Throughout this chapter, we impose the following two conditions:

(a) for any data set Z_1^n, the function $\mathcal{L}_n(\cdot; Z_1^n)$ is convex and differentiable over Ω, and

(b) the regularizer $\mathcal{R}$ is a norm.

These conditions ensure that the overall problem is convex, so that by Lagrangian duality, the optimization problems (6.1) and (6.2) are equivalent. However, as our analysis will show, solving one or the other can be computationally more preferable depending upon the assumptions made. Some remarks on notation: when the radius ρ or the regularization parameter λ_n is clear from the context, we will drop the subscript on $\widehat{\theta}$ to ease the notation. Similarly, we frequently adopt the shorthand $\mathcal{L}_n(\theta)$, with the dependence of the loss function on the data being implicitly understood. Procedures based on optimization problems of either form are known as M-estimators in the statistics literature.

The focus of this chapter is on two simple algorithms for solving the above optimization problems. The method of *projected gradient descent* applies naturally to the constrained

problem (6.1), whereas the *composite gradient descent* method due to Nesterov [102] is suitable for solving the regularized problem (6.2). Each routine generates a sequence $\{\theta^t\}_{t=0}^{\infty}$ of iterates by first initializing to some parameter $\theta^0 \in \Omega$, and then applying the recursive update

$$\theta^{t+1} = \arg\min_{\theta \in \mathbb{B}_{\mathcal{R}}(\rho)} \left\{ \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t), \theta - \theta^t \rangle + \frac{\gamma_u}{2} \|\theta - \theta^t\|^2 \right\}, \qquad \text{for } t = 0, 1, 2, \ldots, \quad (6.3)$$

in the case of projected gradient descent, or the update

$$\theta^{t+1} = \arg\min_{\theta \in \mathbb{B}_{\mathcal{R}}(\bar{\rho})} \left\{ \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t), \theta - \theta^t \rangle + \frac{\gamma_u}{2} \|\theta - \theta^t\|^2 + \lambda_n \mathcal{R}(\theta) \right\}, \qquad \text{for } t = 0, 1, 2, \ldots, \quad (6.4)$$

for the composite gradient method. Note that the only difference between the two updates is the addition of the regularization term in the objective. These updates have a natural intuition: the next iterate θ^{t+1} is obtained by constrained minimization of a first-order approximation to the loss function, combined with a smoothing term that controls how far one moves from the current iterate in terms of Euclidean norm. Moreover, it is easily seen that the update (6.3) is equivalent to

$$\theta^{t+1} = \Pi\left(\theta^t - \frac{1}{\gamma_u} \nabla \mathcal{L}_n(\theta^t) \right), \quad (6.5)$$

where $\Pi \equiv \Pi_{\mathbb{B}_{\mathcal{R}}(\rho)}$ denotes Euclidean projection onto the ball $\mathbb{B}_{\mathcal{R}}(\rho) = \{\theta \in \Omega \mid \mathcal{R}(\theta) \leq \rho\}$ of radius ρ. In this formulation, we see that the algorithm takes a step in the gradient direction, using the quantity $1/\gamma_u$ as stepsize parameter, and then projects the resulting vector onto the constraint set. The update (6.4) takes an analogous form, however, the projection will depend on both λ_n and γ_u. As will be illustrated in the examples to follow, for many problems, the updates (6.3) and (6.4), or equivalently (6.5), have a very simple solution. For instance, in the case of ℓ_1-regularization, it can be obtained by an appropriate form of the soft-thresholding operator.

6.2.2 Restricted strong convexity and smoothness

In this section, we define the conditions on the loss function and regularizer that underlie our analysis. Global smoothness and strong convexity assumptions play an important role in the classical analysis of optimization algorithms [15, 21, 101]. In application to a differentiable loss function $\mathcal{L}_n$, both of these properties are defined in terms of a first-order Taylor series expansion around a vector θ' in the direction of θ—namely, the quantity

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') := \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta') - \langle \nabla \mathcal{L}_n(\theta'), \theta - \theta' \rangle. \quad (6.6)$$

By the assumed convexity of $\mathcal{L}_n$, this error is always non-negative, and global strong convexity is equivalent to imposing a stronger condition, namely that for some parameter $\gamma_\ell > 0$, the first-order Taylor error $\mathcal{T}_{\mathcal{L}}(\theta; \theta')$ is lower bounded by a quadratic term $\frac{\gamma_\ell}{2} \|\theta - \theta'\|^2$ for all $\theta, \theta' \in \Omega$. Global smoothness is defined in a similar way, by imposing a quadratic upper bound on the Taylor error. It is known that under global smoothness and strong convexity assumptions, the method of projected gradient descent (6.3) enjoys a *globally geometric convergence rate*, meaning that there is some $\kappa \in (0, 1)$ such that[1]

$$\|\theta^t - \widehat{\theta}\|^2 \lesssim \kappa^t \, \|\theta^0 - \widehat{\theta}\|^2 \qquad \text{for all iterations } t = 0, 1, 2, \ldots. \tag{6.7}$$

We refer the reader to Bertsekas [15, Prop. 1.2.3, p. 145], or Nesterov [101, Thm. 2.2.8, p. 88] for such results on projected gradient descent, and to Nesterov [102] for composite gradient descent.

Unfortunately, in the high-dimensional setting ($d > n$), it is usually impossible to guarantee strong convexity of the problem (6.1) in a global sense. For instance, when the data is drawn i.i.d., the loss function consists of a sum of n terms. If the loss is twice differentiable, the resulting $d \times d$ Hessian matrix $\nabla^2 \mathcal{L}(\theta; Z_1^n)$ is often a sum of n matrices each with rank one, so that the Hessian is rank-degenerate when $n < d$. However, as we show in this chapter, in order to obtain fast convergence rates for the optimization method (6.3), it is sufficient that (a) the objective is strongly convex and smooth in a restricted set of directions, and (b) the algorithm approaches the optimum $\widehat{\theta}$ only along these directions. Let us now formalize these ideas.

Definition 6.1 (Restricted strong convexity (RSC)). The loss function $\mathcal{L}_n$ satisfies restricted strong convexity with respect to $\mathcal{R}$ and with parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ over the set Ω' if

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') \geq \frac{\gamma_\ell}{2} \|\theta - \theta'\|^2 - \tau_\ell(\mathcal{L}_n) \, \mathcal{R}^2(\theta - \theta') \qquad \text{for all } \theta, \theta' \in \Omega'. \tag{6.8}$$

We refer to the quantity γ_ℓ as the *(lower) curvature parameter*, and to the quantity τ_ℓ as the *tolerance parameter*. The set Ω' corresponds to a suitably chosen subset of the space Ω of all possible parameters.

In order to gain intuition for this definition, first suppose that the condition (6.8) holds with tolerance parameter $\tau_\ell = 0$. In this case, the regularizer plays no role in the definition, and condition (6.8) is equivalent to the usual definition of strong convexity on the optimization set Ω. As discussed previously, this type of global strong convexity typically *fails* to hold for high-dimensional inference problems. In contrast, when tolerance parameter τ_ℓ is

[1] In this statement (and throughout the chapter), we use $\lesssim$ to mean an inequality that holds with some universal constant c, independent of the problem parameters.

strictly positive, the condition (6.8) is much milder, in that it only applies to a *limited set* of vectors. For a given pair $\theta \neq \theta'$, consider the inequality

$$\frac{\mathcal{R}^2(\theta - \theta')}{\|\theta - \theta'\|^2} < \frac{\gamma_\ell}{2\,\tau_\ell(\mathcal{L}_n)}. \tag{6.9}$$

If this inequality is violated, then the right-hand side of the bound (6.8) is non-positive, in which case the RSC constraint (6.8) is vacuous. Thus, restricted strong convexity imposes a non-trivial constraint only on pairs $\theta \neq \theta'$ for which the inequality (6.8) holds, and a central part of our analysis will be to prove that, for the sequence of iterates generated by projected gradient descent, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ satisfies a constraint of the form (6.9). We note that since the regularizer $\mathcal{R}$ is convex, strong convexity of the loss function $\mathcal{L}_n$ also implies the strong convexity of the regularized loss ϕ_n as well.

For the least-squares loss, the RSC definition depends purely on the direction (and not the magnitude) of the difference vector $\theta - \theta'$. For other types of loss functions—such as those arising in generalized linear models—it is essential to localize the RSC definition, requiring that it holds only for pairs for which the norm $\|\theta - \theta'\|_2$ is not too large. We refer the reader to Section 6.2.4 for further discussion of this issue.

Finally, as pointed out by a reviewer, our restricted version of strong convexity can be seen as an instance of the general theory of paraconvexity (e.g., [103]); however, we are not aware of convergence rates for minimizing general paraconvex functions.

We also specify an analogous notion of restricted smoothness:

Definition 6.2 (Restricted smoothness (RSM)). We say the loss function $\mathcal{L}_n$ satisfies restricted smoothness with respect to $\mathcal{R}$ and with parameters $(\gamma_u, \tau_u(\mathcal{L}_n))$ over the set Ω' if

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') \leq \frac{\gamma_u}{2} \|\theta - \theta'\|^2 + \tau_u(\mathcal{L}_n)\, \mathcal{R}^2(\theta - \theta') \qquad \text{for all } \theta, \theta' \in \Omega'. \tag{6.10}$$

As with our definition of restricted strong convexity, the additional tolerance $\tau_u(\mathcal{L}_n)$ is not present in analogous smoothness conditions in the optimization literature, but it is essential in our set-up.

6.2.3 Decomposable regularizers

We saw in Chapter 3, that the notion of a decomposable regularizer can be quite useful. Although the focus of this chapter is a rather different set of questions—namely, optimization as opposed to statistics—decomposability also plays an important role here. In particular, for any M-estimator involving a decomposable regularizer, the *optimization error* $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ belongs to exactly the limited set of directions for which the RSC and RSM conditions apply. (For a precise statement of this connection, see Lemma 6.1 in Section 6.5).

Recall that decomposability is defined given a pair of subspaces defined with respect to the parameter space $\Omega \subseteq \mathbb{R}^d$: the model subspace $\mathcal{M}$ and the perturbation subspace $\overline{\mathcal{M}}^{\perp}$. Furthermore, for any vector $\alpha \in \mathcal{M}$ and $\beta \in \overline{\mathcal{M}}^{\perp}$ the regularizer $\mathcal{R}$ satisfies $\mathcal{R}(\alpha + \beta) = \mathcal{R}(\alpha) + \mathcal{R}(\beta)$. Additionally, recall that for a given error norm $\|\cdot\|$, the subspace compatibility constant $\Psi(\overline{\mathcal{M}})$ defined in equation (3.21) acts as the Lipschitz constant of the regularizer against the norm $\|\cdot\|$. Namely that for any vector $\alpha \in \overline{\mathcal{M}}$, $\mathcal{R}(\alpha) \leq \Psi(\overline{\mathcal{M}})\|\alpha\|$. For a more detailed discussion, see Definition 3.3.

6.2.4 Some illustrative examples

We now describe some particular examples of M-estimators with decomposable regularizers, and discuss the form of the projected gradient updates as well as RSC/RSM conditions. We cover two main families of examples: log-linear models with sparsity constraints and ℓ_1-regularization (Section 6.2.4), and matrix regression problems with nuclear norm regularization (Section 6.2.4).

Sparse log-linear models and ℓ_1-regularization

Suppose that each sample Z_i consists of a scalar-vector pair $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^d$, corresponding to the scalar response $y_i \in \mathcal{Y}$ associated with a vector of predictors $x_i \in \mathbb{R}^d$. A log-linear model with canonical link function assumes that the response y_i is linked to the covariate vector x_i via a conditional distribution of the form $\mathbb{P}(y_i \mid x_i; \theta^*, \sigma) \propto \exp\left\{\frac{y_i \langle \theta^*, x_i \rangle - \Phi(\langle \theta^*, x_i \rangle)}{c(\sigma)}\right\}$, where $c(\sigma)$ is a known quantity, $\Phi(\cdot)$ is the log-partition function to normalize the density, and $\theta^* \in \mathbb{R}^d$ is an unknown regression vector. In many applications, the regression vector θ^* is relatively sparse, so that it is natural to impose an ℓ_1-constraint. Computing the maximum likelihood estimate subject to such a constraint involves solving the convex program[2]

$$\widehat{\theta} \in \arg\min_{\theta \in \Omega} \underbrace{\left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i \langle \theta, x_i \rangle - \Phi(\langle \theta, x_i \rangle) \right\} \right\}}_{\mathcal{L}_n(\theta; Z_1^n)} \quad \text{such that } \|\theta\|_1 \leq \rho, \tag{6.11}$$

with $x_i \in \mathbb{R}^d$ as its i^{th} row. We refer to this estimator as the log-linear Lasso; it is a special case of the M-estimator (6.1), with the loss function $\mathcal{L}_n(\theta; Z_1^n) = \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i \langle \theta, x_i \rangle - \Phi(\langle \theta, x_i \rangle) \right\}$ and the regularizer $\mathcal{R}(\theta) = \|\theta\|_1 = \sum_{j=1}^{d} |\theta_j|$.

Ordinary linear regression is the special case of the log-linear setting with $\Phi(t) = t^2/2$ and $\Omega = \mathbb{R}^d$, and in this case, the estimator (6.11) corresponds to ordinary least-squares version of Lasso [39, 131]. Other forms of log-linear Lasso that are of interest include logistic regression, Poisson regression, and multinomial regression.

[2]The link function Φ is convex since it is the log-partition function of a canonical exponential family.

Projected gradient updates: Computing the gradient of the log-linear loss from equation (6.11) is straightforward: we have $\nabla \mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n x_i\{y_i - \Phi'(\langle \theta, x_i\rangle)\}$, and the update (6.5) corresponds to the Euclidean projection of the vector $\theta^t - \frac{1}{\gamma_u}\nabla \mathcal{L}_n(\theta^t)$ onto the ℓ_1-ball of radius ρ. It is well-known that this projection can be characterized in terms of soft-thresholding, and that the projected update (6.5) can be computed easily. We refer the reader to Duchi et al. [47] for an efficient implementation requiring $\mathcal{O}(d)$ operations.

Composite gradient updates: The composite gradient update for this problem amounts to solving

$$\theta^{t+1} = \arg\min_{\|\theta\|_1 \leq \bar{\rho}} \left\{ \langle \theta, \nabla \mathcal{L}_n(\theta)\rangle + \frac{\gamma_u}{2}\|\theta - \theta^t\|_2^2 + \lambda_n \|\theta\|_1 \right\}.$$

The update can be computed by two soft-thresholding operations. The first step is soft thresolding the vector $\theta^t - \frac{1}{\gamma_u}\nabla \mathcal{L}_n(\theta^t)$ at a level λ_n/γ_u. If the resulting vector has ℓ_1-norm greater than $\bar{\rho}$, then we project on to the ℓ_1-ball just like before. Overall, the complexity of the update is still $\mathcal{O}(d)$ as before.

RSC/RSM conditions: A calculation using the mean-value theorem shows that for the loss function (6.11), the error in the first-order Taylor series, as previously defined in equation (6.6), can be written as

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') = \frac{1}{n}\sum_{i=1}^n \Phi''\big(\langle \theta_t, x_i\rangle\big)\,\big(\langle x_i, \theta - \theta'\rangle\big)^2$$

where $\theta_t = t\theta + (1-t)\theta'$ for some $t \in [0,1]$. When $n < d$, then we can always find pairs $\theta \neq \theta'$ such that $\langle x_i, \theta - \theta'\rangle = 0$ for all $i = 1, 2, \ldots, n$, showing that the objective function can never be strongly convex. On the other hand, restricted strong convexity for log-linear models requires only that there exist positive numbers $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ such that

$$\frac{1}{n}\sum_{i=1}^n \Phi''\big(\langle \theta_t, x_i\rangle\big)\,\big(\langle x_i, \theta - \theta'\rangle\big)^2 \geq \frac{\gamma_\ell}{2}\,\|\theta - \theta'\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\theta - \theta') \qquad \text{for all } \theta, \theta' \in \Omega', \tag{6.12}$$

where $\Omega' := \Omega \cap \mathbb{B}_2(R)$ is the intersection of the parameter space Ω with a Euclidean ball of some fixed radius R around zero. This restriction is essential because for many generalized linear models, the Hessian function Φ'' approaches zero as its argument diverges. For instance, for the logistic function $\Phi(t) = \log(1 + \exp(t))$, we have $\Phi''(t) = \exp(t)/[1 + \exp(t)]^2$, which tends to zero as $t \to +\infty$. Restricted smoothness imposes an analogous upper bound on the Taylor error. For a broad class of log-linear models, such bounds hold with with tolerance $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ of the order $\sqrt{\frac{\log d}{n}}$. Further details on such results

are provided in the corollaries to follow our main theorem. A brief discussion of RSC for exponential families in statistical problems can be found in Chapter 3.

In order to ensure RSC/RSM conditions on the iterates θ^t of the updates (6.3) or (6.4), we also need to ensure that $\theta^t \in \Omega'$. This can be done by defining $\mathcal{L}_n^{'} = \mathcal{L}_n + \mathbb{I}_{\Omega'}(\theta)$, where $\mathbb{I}_{\Omega'}(\theta)$ is zero when $\theta \in \Omega'$ and ∞ otherwise. This is equivalent to projection on the intersection of ℓ_1-ball with Ω' in the updates (6.3) and (6.4) and can be done efficiently with Dykstra's algorithm [49], for instance, as long as the individual projections are efficient.

In the special case of linear regression, we have $\Phi''(t) = 1$ for all $t \in \mathbb{R}$, so that the lower bound (6.12) involves only the Gram matrix $X^T X/n$. (Here $X \in \mathbb{R}^{n \times d}$ is the usual design matrix, with $x_i \in \mathbb{R}^d$ as its i^{th} row.) For linear regression and ℓ_1-regularization, the RSC condition is equivalent to the lower bound

$$\frac{\|X(\theta - \theta')\|_2^2}{n} \geq \frac{\gamma_\ell}{2}\|\theta - \theta'\|_2^2 - \tau_\ell(\mathcal{L}_n)\, \|\theta - \theta'\|_1^2 \qquad \text{for all } \theta, \theta' \in \Omega. \tag{6.13}$$

Such a condition corresponds to a variant of the restricted eigenvalue (RE) conditions that have been studied in the literature [20, 139]. Such RE conditions are significantly milder than the restricted isometry property; we refer the reader to van de Geer and Buhlmann [139] for an in-depth comparison of different RE conditions. From past work, the condition (6.13) is satisfied with high probability for a broad classes of anisotropic random design matrices [108, 121], and parts of our analysis make use of this fact.

Matrices and nuclear norm regularization

We now discuss a general class of matrix regression problems that falls within our framework. Consider the space of $d_1 \times d_2$ matrices endowed with the trace inner product $\langle\langle A,\ B \rangle\rangle := \text{trace}(A^T B)$. In order to ease notation, we define $d := \min\{d_1, d_2\}$. Let $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ be an unknown matrix and suppose that for $i = 1, 2, \ldots, n$, we observe a scalar-matrix pair $Z_i = (y_i, X_i) \in \mathbb{R} \times \mathbb{R}^{d_1 \times d_2}$ linked to Θ^* via the linear model

$$y_i = \langle\langle X_i,\ \Theta^* \rangle\rangle + w_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{6.14}$$

where w_i is an additive observation noise. In many contexts, it is natural to assume that Θ^* is exactly low-rank, or approximately so, meaning that it is well-approximated by a matrix of low rank. In such settings, a number of authors (e.g., [51, 119]) have studied the M-estimator

$$\widehat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \langle\langle X_i,\ \Theta \rangle\rangle \right)^2 \right\} \quad \text{such that } |||\Theta|||_{\text{nuc}} \leq \rho, \tag{6.15}$$

or the corresponding regularized version (see Chapter 4). Here the *nuclear or trace norm* is given by $|||\Theta|||_{\text{nuc}} := \sum_{j=1}^{d} \sigma_j(\Theta)$, corresponding to the sum of the singular values. This optimization problem is an instance of a semidefinite program. As discussed in more detail in Chapter 4, there are various applications in which this estimator and variants thereof have proven useful.

Form of projected gradient descent: For the M-estimator (6.15), the projected gradient updates take a very simple form—namely

$$\Theta^{t+1} = \Pi\Big(\Theta^t - \frac{1}{\gamma_u}\frac{\sum_{i=1}^n \big(y_i - \langle\!\langle X_i,\ \Theta^t \rangle\!\rangle\big)\, X_i}{n}\Big), \tag{6.16}$$

where Π denotes Euclidean projection onto the nuclear norm ball

$$\mathbb{B}_1(\rho) = \{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid |\!|\!|\Theta|\!|\!|_{\text{nuc}} \leq \rho\}.$$

This nuclear norm projection can be obtained by first computing the singular value decomposition (SVD), and then projecting the vector of singular values onto the ℓ_1-ball. The latter step can be achieved by the fast projection algorithms discussed earlier, and there are various methods for fast computation of SVDs. The composite gradient update also has a simple form, requiring at most two singular value thresholding operations as was the case for linear regression.

In some special cases such as matrix completion or matrix decomposition that we describe in the sequel, Ω' will involve an additional bound on the entries of Θ^* as well as the iterates Θ^t to establish RSC/RSM conditions. This can be done by augmenting the loss with an indicator of the constraint and using cyclic projections for computing the updates as mentioned earlier in Example 6.2.4.

6.3 Main results and some consequences

We are now equipped to state the two main results of this chapter, and discuss some of their consequences. We illustrate its application to several statistical models, including sparse regression (Section 6.3.2), matrix estimation with rank constraints (Section 6.3.3), and matrix decomposition problems (Section 6.3.4).

6.3.1 Geometric convergence

Recall that the projected gradient algorithm (6.3) is well-suited to solving an M-estimation problem in its constrained form, whereas the composite gradient algorithm (6.4) is appropriate for a regularized problem. Accordingly, let $\widehat{\theta}$ be any optimal solution to the constrained problem (6.1), or the regularized problem (6.2), and let $\{\theta^t\}_{t=0}^{\infty}$ be a sequence of iterates generated by generated by the projected gradient updates (6.3), or the the composite gradient updates (6.4), respectively. Of primary interest to us in this chapter are bounds on the *optimization error*, which can be measured either in terms of the error vector $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$, or the difference between the cost of θ^t and the optimal cost defined by $\widehat{\theta}$. In this section, we state two main results —-Theorems 6.1 and 6.2—corresponding to the constrained and regularized cases respectively. In addition to the optimization error previously discussed,

both of these results involve the *statistical error* $\Delta^* := \widehat{\theta} - \theta^*$ between the optimum $\widehat{\theta}$ and the nominal parameter θ^*. At a high level, these results guarantee that under the RSC/RSM conditions, the optimization error shrinks geometrically, with a contraction coefficient that depends on the the loss function $\mathcal{L}_n$ via the parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ and $(\gamma_u, \tau_u(\mathcal{L}_n))$. An interesting feature is that the contraction occurs only up to a certain tolerance parameter ϵ^2 depending on these same parameters, and the statistical error. However, as we discuss, for many statistical problems of interest, we can show that this tolerance parameter is of lower order than the intrinsic statistical error, and hence can be neglected from the statistical point of view. Consequently, our theory gives an explicit upper bound on the number of iterations required to solve an M-estimation problem up to statistical precision.

Convergence rates for projected gradient: We now provide the notation necessary for a precise statement of this claim. Our main result actually involves a family of upper bounds on the optimization error, one for each pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ of $\mathcal{R}$-decomposable subspaces (see Definition 3.1). As will be clarified in the sequel, this subspace choice can be optimized for different models so as to obtain the tightest possible bounds. For a given pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ such that $16\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n) < \gamma_u$, let us define the *contraction coefficient*

$$\kappa(\mathcal{L}_n; \overline{\mathcal{M}}) := \Big\{1 - \frac{\gamma_\ell}{\gamma_u} + \frac{16\Psi^2(\overline{\mathcal{M}})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)}{\gamma_u}\Big\}\Big\{1 - \frac{16\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n)}{\gamma_u}\Big\}^{-1}. \quad (6.17)$$

In addition, we define the *tolerance parameter*

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) := \frac{32\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)\,\big(2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \Psi(\overline{\mathcal{M}})\|\Delta^*\| + 2\mathcal{R}(\Delta^*)\big)^2}{\gamma_u}, \quad (6.18)$$

where $\Delta^* = \widehat{\theta} - \theta^*$ is the statistical error, and $\Pi_{\mathcal{M}^\perp}(\theta^*)$ denotes the Euclidean projection of θ^* onto the subspace $\mathcal{M}^\perp$.

In terms of these two ingredients, we now state our first main result:

Theorem 6.1. *Suppose that the loss function $\mathcal{L}_n$ satisfies the RSC/RSM condition with parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ and $(\gamma_u, \tau_u(\mathcal{L}_n))$ respectively. Let $(\mathcal{M}, \overline{\mathcal{M}})$ be any $\mathcal{R}$-decomposable pair of subspaces such that $\mathcal{M} \subseteq \overline{\mathcal{M}}$ and $0 < \kappa \equiv \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) < 1$. Then for any optimum $\widehat{\theta}$ of the problem (6.1) for which the constraint is active, we have*

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \|\theta^0 - \widehat{\theta}\|^2 + \frac{\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})}{1 - \kappa} \qquad \text{for all iterations } t = 0, 1, 2, \ldots. \quad (6.19)$$

Remarks: Theorem 6.1 actually provides a family of upper bounds, one for each $\mathcal{R}$-decomposable pair $(\mathcal{M}, \overline{\mathcal{M}})$ such that $0 < \kappa \equiv \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) < 1$. This condition is always satisfied by setting $\overline{\mathcal{M}}$ equal to the trivial subspace $\{0\}$: indeed, by definition (3.21) of the

subspace compatibility, we have $\Psi(\overline{\mathcal{M}}) = 0$, and hence $\kappa(\mathcal{L}_n; \{0\}) = \left(1 - \frac{\gamma_\ell}{\gamma_u}\right) < 1$. Although this choice of $\overline{\mathcal{M}}$ minimizes the contraction coefficient, it will lead[3] to a very large tolerance parameter $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$. A more typical application of Theorem 6.1 involves non-trivial choices of the subspace $\overline{\mathcal{M}}$.

The bound (6.19) guarantees that the optimization error decreases geometrically, with contraction factor $\kappa \in (0, 1)$, up to a certain tolerance proportional to $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, as illustrated in Figure 6.1(a). The contraction factor κ approaches the $1 - \gamma_\ell/\gamma_u$ as the number of samples grows. The appearance of the ratio γ_ℓ/γ_u is natural since it measures the conditioning of the objective function; more specifically, it is essentially a restricted condition number of the Hessian matrix. On the other hand, the tolerance parameter ϵ depends on the choice of decomposable subspaces, the parameters of the RSC/RSM conditions, and the statistical error $\Delta^* = \widehat{\theta} - \theta^*$ (see equation (6.18)). In the corollaries of Theorem 6.1 to follow, we show that the subspaces can often be chosen such that $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) = o(\|\widehat{\theta} - \theta^*\|^2)$. Consequently, the bound (6.19) guarantees geometric convergence up to a tolerance *smaller than statistical precision*, as illustrated in Figure 6.1(b). This is sensible, since in statistical settings, there is no point to optimizing beyond the statistical precision.

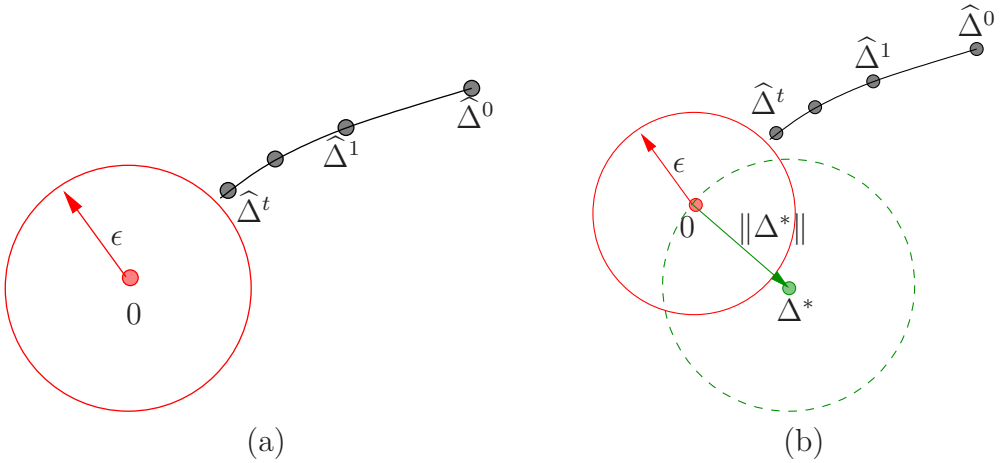


Figure 6.1. (a) Generic illustration of Theorem 6.1. The optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ is guaranteed to decrease geometrically with coefficient $\kappa \in (0, 1)$, up to the tolerance $\epsilon^2 = \epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, represented by the circle. (b) Relation between the optimization tolerance $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ (solid circle) and the statistical precision $\|\Delta^*\| = \|\theta^* - \widehat{\theta}\|$ (dotted circle). In many settings, we have $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \ll \|\Delta^*\|^2$, so that convergence is guaranteed up to a tolerance lower than statistical precision.

The result of Theorem 6.1 takes a simpler form when there is a subspace $\mathcal{M}$ that includes

[3]Indeed, the setting $\mathcal{M}^\perp = \mathbb{R}^d$ means that the term $\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) = \mathcal{R}(\theta^*)$ appears in the tolerance; this quantity is far larger than statistical precision.

θ^*, and the $\mathcal{R}$-ball radius is chosen such that $\rho \leq \mathcal{R}(\theta^*)$. In this case, by appropriately controlling the error term, we can establish that it is of lower order than the statistical precision —namely, the squared difference $\|\widehat{\theta} - \theta^*\|^2$ between an optimal solution $\widehat{\theta}$ to the convex program (6.1), and the unknown parameter θ^*.

Corollary 6.1. *In addition to the conditions of Theorem 6.1, suppose that $\theta^* \in \mathcal{M}$ and $\rho \leq \mathcal{R}(\theta^*)$. Then as long as $\Psi^2(\overline{\mathcal{M}})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big) = o(1)$, we have*

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \, \|\theta^0 - \widehat{\theta}\|^2 + o\big(\|\widehat{\theta} - \theta^*\|^2\big) \qquad \text{for all iterations } t = 0, 1, 2, \ldots. \tag{6.20}$$

Thus, Corollary 6.1 guarantees that the optimization error decreases geometrically, with contraction factor κ, up to a tolerance that is of strictly lower order than the statistical precision $\|\widehat{\theta} - \theta^*\|^2$. As will be clarified in several examples to follow, the condition $\Psi^2(\overline{\mathcal{M}})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big) = o(1)$ is satisfied for many statistical models, including sparse linear regression and low-rank matrix regression. This result is illustrated in Figure 6.1(b), where the solid circle represents the optimization tolerance, and the dotted circle represents the statistical precision. In the results to follow, we will quantify the term $o\big(\|\widehat{\theta} - \theta^*\|^2\big)$ in a more precise manner for different statistical models.

Convergence rates for composite gradient: We now present our main result for the composite gradient iterates (6.4) that are suitable for the Lagrangian-based estimator (6.2). As before, our analysis yields a range of bounds indexed by subspace pairs $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ that are $\mathcal{R}$-decomposable. For any subspace $\overline{\mathcal{M}}$ such that $64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}}) < \gamma_\ell$, we define *effective RSC coefficient* as

$$\overline{\gamma_\ell} := \gamma_\ell - 64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}}). \tag{6.21}$$

This coefficient accounts for the residual amount of strong convexity after accounting for the lower tolerance terms. In addition, we define the *compound contraction coefficient* as

$$\kappa(\mathcal{L}_n; \overline{\mathcal{M}}) := \left\{1 - \frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{64\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n)}{\overline{\gamma_\ell}}\right\} \xi(\overline{\mathcal{M}}) \tag{6.22}$$

where $\xi(\overline{\mathcal{M}}) := \big(1 - \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\bar{\mathcal{M}})}{\overline{\gamma_\ell}}\big)^{-1}$, and $\Delta^* = \widehat{\theta}_{\lambda_n} - \theta^*$ is the statistical error vector[4] for a specific choice of $\bar{\rho}$ and λ_n. As before, the coefficient κ measures the geometric rate of convergence for the algorithm. Finally, we define the *compound tolerance parameter*

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) := 8\,\xi(\overline{\mathcal{M}})\,\beta(\overline{\mathcal{M}})\,\big(6\Psi(\overline{\mathcal{M}})\|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))\big)^2, \tag{6.23}$$

[4]When the context is clear, we remind the reader that we drop the subscript λ_n on the parameter $\widehat{\theta}$.

where $\beta(\overline{\mathcal{M}}) := 2\left(\frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{128\tau_u(\mathcal{L}_n)\Psi^2(\bar{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\tau_\ell(\mathcal{L}_n) + 8\tau_u(\mathcal{L}_n) + 2\tau_\ell(\mathcal{L}_n)$. As with our previous result, the tolerance parameter determines the radius up to which geometric convergence can be attained.

Recall that the regularized problem (6.2) involves both a regularization weight λ_n, and a constraint radius $\bar{\rho}$. Our theory requires that the constraint radius is chosen such that $\bar{\rho} \geq \mathcal{R}(\theta^*)$, which ensures that θ^* is feasible. In addition, the regularization parameter should be chosen to satisfy the constraint

$$\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}_n(\theta^*)), \tag{6.24}$$

where $\mathcal{R}^*$ is the dual norm of the regularizer. This constraint is known to play an important role in proving bounds on the statistical error of regularized M-estimators (see Chapter 3 and references therein for further details). Recalling the definition (6.2) of the overall objective function $\phi_n(\theta)$, the following result provides bounds on the *excess loss* $\phi_n(\theta^t) - \phi_n(\widehat{\theta}_{\lambda_n})$.

Theorem 6.2. *Consider the optimization problem (6.2) for a radius $\bar{\rho}$ such that θ^* is feasible, and a regularization parameter λ_n satisfying the bound (6.24), and suppose that the loss function $\mathcal{L}_n$ satisfies the RSC/RSM condition with parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ and $(\gamma_u, \tau_u(\mathcal{L}_n))$ respectively. Let $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ be any $\mathcal{R}$-decomposable pair such that*

$$\kappa \equiv \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) \in [0,1), \quad \text{and} \quad \frac{32\,\bar{\rho}}{1-\kappa(\mathcal{L}_n;\overline{\mathcal{M}})}\xi(\overline{\mathcal{M}})\beta(\overline{\mathcal{M}}) \leq \lambda_n. \tag{6.25}$$

Then for any tolerance parameter $\delta^2 \geq \frac{\epsilon^2(\Delta^*;\mathcal{M},\overline{\mathcal{M}})}{(1-\kappa)}$, we have

$$\phi_n(\theta^t) - \phi_n(\widehat{\theta}_{\lambda_n}) \leq \delta^2 \quad \text{for all} \quad t \geq \frac{2\log\frac{\phi_n(\theta^0)-\phi_n(\widehat{\theta}_{\lambda_n})}{\delta^2}}{\log(1/\kappa)} + \log_2\log_2\left(\frac{\bar{\rho}\lambda_n}{\delta^2}\right)\left(1 + \frac{\log 2}{\log(1/\kappa)}\right). \tag{6.26}$$

Remarks: Note that the bound (6.26) guarantees the excess loss $\phi_n(\theta^t) - \phi_n(\widehat{\theta})$ decays geometrically up to any squared error δ^2 larger than the compound tolerance (6.23). Moreover, the RSC condition also allows us to translate this bound on objective values to a bound on the optimization error $\theta^t - \widehat{\theta}$. In particular, for any iterate θ^t such that $\phi_n(\theta^t) - \phi_n(\widehat{\theta}) \leq \delta^2$, we are guaranteed that

$$\|\theta^t - \widehat{\theta}_{\lambda_n}\|^2 \leq \frac{2\delta^2}{\overline{\gamma_\ell}} + \frac{16\delta^2\tau_\ell(\mathcal{L}_n)}{\overline{\gamma_\ell}\lambda_n^2} + \frac{4\tau_\ell(\mathcal{L}_n)(6\Psi(\overline{\mathcal{M}}) + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)))^2}{\overline{\gamma_\ell}}. \tag{6.27}$$

In conjunction with Theorem 6.2, we see that it suffices to take a number of steps that is logarithmic in the inverse tolerance $(1/\delta)$, again showing a geometric rate of convergence.

Whereas Theorem 6.1 requires setting the radius so that the constraint is active, Theorem 6.2 has only a very mild constraint on the radius $\bar{\rho}$, namely that it be large enough such

that $\bar{\rho} \geq \mathcal{R}(\theta^*)$. The reason for this much milder requirement is that the additive regularization with weight λ_n suffices to constrain the solution, whereas the extra side constraint is only needed to ensure good behavior of the optimization algorithm in the first few iterations. The regularization parameter λ_n must satisfy the so-called dual norm condition (6.24), which was required in establishing the statistical bounds presented throughout this thesis.

Step-size setting: It seems that the updates (6.3) and (6.4) need to know the smoothness bound γ_u in order to set the step-size for gradient updates. However, we can use the same doubling trick as described in Algorithm (3.1) of Nesterov [102]. At each step, we check if the smoothness upper bound holds at the current iterate relative to the previous one. If the condition does not hold, we double our estimate of γ_u and resume. This guarantees a geometric convergence with a contraction factor worse at most by a factor of 2, compared to the knowledge of γ_u. We refer the reader to Nesterov [102] for details.

The following subsections are devoted to the development of some consequences of Theorems 6.1 and 6.2 and Corollary 6.1 for some specific statistical models, among them sparse linear regression with ℓ_1-regularization, and matrix regression with nuclear norm regularization. In contrast to the entirely deterministic arguments that underlie the Theorems 6.1 and 6.2, these corollaries involve probabilistic arguments, more specifically in order to establish that the RSC and RSM properties hold with high probability.

6.3.2 Sparse vector regression

Recall from Section 6.2.4 the observation model for sparse linear regression. In a variety of applications, it is natural to assume that θ^* is sparse. For a parameter $q \in [0, 1]$ and radius $R_q > 0$, let us define the ℓ_q "ball"

$$\mathbb{B}_q(R_q) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\beta_j|^q \leq R_q\big\}. \tag{6.28}$$

Note that $q = 0$ corresponds to the case of "hard sparsity", for which any vector $\beta \in \mathbb{B}_0(R_0)$ is supported on a set of cardinality at most R_0. For $q \in (0, 1]$, membership in the set $\mathbb{B}_q(R_q)$ enforces a decay rate on the ordered coefficients, thereby modelling approximate sparsity. In order to estimate the unknown regression vector $\theta^* \in \mathbb{B}_q(R_q)$, we consider the least-squares Lasso estimator from Section 6.2.4, based on the quadratic loss function $\mathcal{L}(\theta; Z_1^n) := \frac{1}{2n}\|y - X\theta\|_2^2$, where $X \in \mathbb{R}^{n \times d}$ is the design matrix. In order to state a concrete result, we consider a random design matrix X, in which each row $x_i \in \mathbb{R}^d$ is drawn i.i.d. from a $N(0, \Sigma)$ distribution, where Σ is a positive definite covariance matrix. We refer to this as the Σ-*ensemble of random design matrices*, and use $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ to refer the maximum and minimum eigenvalues of Σ respectively, and $\zeta(\Sigma) := \max_{j=1,2,\ldots,d} \Sigma_{jj}$

for the maximum variance. We also assume that the observation noise is zero-mean and sub-Gaussian with parameter ν^2.

Guarantees for constrained Lasso: Our convergence rate on the optimization error $\theta^t - \widehat{\theta}$ is stated in terms of the contraction coefficient

$$\kappa := \left\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\right\}\left\{1 - \chi_n(\Sigma)\right\}^{-1}, \tag{6.29}$$

where we have adopted the shorthand

$$\chi_n(\Sigma) := \begin{cases} \frac{c_0\zeta(\Sigma)}{\sigma_{\max}(\Sigma)}\, R_q \left(\frac{\log d}{n}\right)^{1-q/2} & \text{for } q > 0 \\ \frac{c_0\zeta(\Sigma)}{\sigma_{\max}(\Sigma)}\, k \left(\frac{\log d}{n}\right) & \text{for } q = 0 \end{cases}, \qquad \text{for a numerical constant } c_0, \tag{6.30}$$

We assume that $\chi_n(\Sigma)$ is small enough to ensure that $\kappa \in (0,1)$; in terms of the sample size, this amounts to a condition of the form $n = \Omega(R_q^{1/(1-q/2)} \log d)$. Such a scaling is sensible, since it is known from minimax theory on sparse linear regression [109] to be necessary for any method to be statistically consistent over the ℓ_q-ball.

With this set-up, we have the following consequence of Theorem 6.1:

Corollary 6.2 (Sparse vector recovery)**.** *Under conditions of Theorem 6.1, suppose that we solve the constrained Lasso with $\rho \leq \|\theta^*\|_1$.*

(a) Exact sparsity: *If θ^* is supported on a subset of cardinality k, then with probability at least $1 - \exp(-c_1 \log d)$, the iterates (6.3) with $\gamma_u = 2\sigma_{\max}(\Sigma)$ satisfy*

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq \kappa^t \|\theta^0 - \widehat{\theta}\|_2^2 + c_2\, \chi_n(\Sigma)\, \|\widehat{\theta} - \theta^*\|_2^2 \qquad \text{for all } t = 0, 1, 2, \ldots. \tag{6.31}$$

(b) Weak sparsity: *Suppose that $\theta^* \in \mathbb{B}_q(R_q)$ for some $q \in (0,1]$. Then with probability at least $1 - \exp(-c_1 \log d)$, the iterates (6.3) with $\gamma_u = 2\sigma_{\max}(\Sigma)$ satisfy*

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq \kappa^t\, \|\theta^0 - \widehat{\theta}\|_2^2 + c_2\, \chi_n(\Sigma) \left\{R_q\Big(\frac{\log d}{n}\Big)^{1-q/2} + \|\widehat{\theta} - \theta^*\|_2^2\right\}. \tag{6.32}$$

We provide the proof of Corollary 6.2 in Section 6.5.4. Here we compare part (a), which deals with the special case of exactly sparse vectors, to some past work that has established convergence guarantees for optimization algorithms for sparse linear regression. Certain methods are known to converge at sublinear rates (e.g., [14]), more specifically at the rate $\mathcal{O}(1/t^2)$. The geometric rate of convergence guaranteed by Corollary 6.2 is exponentially faster. Other work on sparse regression has provided geometric rates of convergence that hold once the iterates are close to the optimum [22, 58], or geometric convergence up to

the noise level ν^2 using various methods, including greedy methods [134] and thresholded gradient methods [54]. In contrast, Corollary 6.2 guarantees geometric convergence for all iterates up to a precision below that of statistical error. For these problems, the statistical error $\frac{\nu^2 k \log d}{n}$ is typically much smaller than the noise variance ν^2, and decreases as the sample size is increased.

In addition, Corollary 6.2 also applies to the case of approximately sparse vectors, lying within the set $\mathbb{B}_q(R_q)$ for $q \in (0, 1]$. There are some important differences between the case of exact sparsity (Corollary 6.2(a)) and that of approximate sparsity (Corollary 6.2(b)). Part (a) guarantees geometric convergence to a tolerance depending only on the statistical error $\|\widehat{\theta} - \theta^*\|_2$. In contrast, the second result also has the additional term $R_q\big(\frac{\log d}{n}\big)^{1-q/2}$. This second term arises due to the statistical non-identifiability of linear regression over the ℓ_q-ball, and it is no larger than $\|\widehat{\theta} - \theta^*\|_2^2$ with high probability. This assertion follows from known results [109] about minimax rates for linear regression over ℓ_q-balls; these unimprovable rates include a term of this order.

Guarantees for regularized Lasso: Using similar methods, we can also use Theorem 6.2 to obtain an analogous guarantee for the regularized Lasso estimator. Here focus only on the case of exact sparsity, although the result extends to approximate sparsity in a similar fashion. Letting $c_i, i = 0, 1, 2, 3, 4$ be universal positive constants, we define the modified curvature constant $\overline{\gamma_\ell} := \gamma_\ell - c_0 \frac{k \log d}{n} \zeta(\Sigma)$. Our results assume that $n = \Omega(k \log d)$, a condition known to be necessary for statistical consistency, so that $\overline{\gamma_\ell} > 0$. The contraction factor then takes the form

$$\kappa := \big\{1 - \frac{\sigma_{\min}(\Sigma)}{16\sigma_{\max}(\Sigma)} + c_1\chi_n(\Sigma)\big\}\,\big\{1 - c_2\chi_n(\Sigma)\big\}^{-1}, \quad \text{where} \quad \chi_n(\Sigma) = \frac{\zeta(\Sigma)}{\overline{\gamma_\ell}}\,\frac{k\,\log d}{n}.$$

The tolerance factor in the optimization is given by

$$\epsilon_{\text{tol}}^2 := \frac{5 + c_2\chi_n(\Sigma)}{1 - c_3\chi_n(\Sigma)}\,\frac{\zeta(\Sigma)\,k \log d}{n}\|\theta^* - \widehat{\theta}\|_2^2, \tag{6.33}$$

where $\theta^* \in \mathbb{R}^d$ is the unknown regression vector, and $\widehat{\theta}$ is any optimal solution. With this notation, we have the following corollary.

Corollary 6.3 (Regularized Lasso)**.** *Under conditions of Theorem 6.2, suppose that we solve the regularized Lasso with $\lambda_n = 6\sqrt{\frac{\nu \log d}{n}}$, and that θ^* is supported on a subset of cardinality at most k. Then with probability at least $1 - \exp(-c_4 \log d)$, for any $\delta^2 \geq \epsilon_{tol}^2$, for any optimum $\widehat{\theta}_{\lambda_n}$, we have*

$$\|\theta^t - \widehat{\theta}_{\lambda_n}\|_2^2 \leq \delta^2 \qquad \textit{for all iterations } t \;\geq\; \big(\log \tfrac{\phi_n(\theta^0) - \phi_n(\widehat{\theta}_{\lambda_n})}{\delta^2}\big)/\big(\log \tfrac{1}{\kappa}\big).$$

As with Corollary 6.2(a), this result guarantees that $\mathcal{O}(\log(1/\epsilon_{\text{tol}}^2))$ iterations are sufficient to obtain an iterate θ^t that is within squared error $\mathcal{O}(\epsilon_{\text{tol}}^2)$ of any optimum $\widehat{\theta}_{\lambda_n}$. Moreover, whenever $\frac{k \log d}{n} = o(1)$—a condition that is required for statistical consistency of *any method*—the optimization tolerance ϵ_{tol}^2 is of lower order than the statistical error $\|\theta^* - \theta\|_2^2$.

6.3.3 Matrix regression with rank constraints

We now turn estimation of matrices under various types of "soft" rank constraints. Recall the model of matrix regression from Section 6.2.4, and the M-estimator based on least-squares regularized with the nuclear norm (6.15). So as to reduce notational overhead, here we specialize to square matrices $\Theta^* \in \mathbb{R}^{m \times m}$, so that our observations are of the form

$$y_i = \langle\langle X_i, \Theta^* \rangle\rangle + w_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{6.34}$$

where $X_i \in \mathbb{R}^{m \times m}$ is a matrix of covariates, and $w_i \sim N(0, \nu^2)$ is Gaussian noise. As discussed in Section 6.2.4, the nuclear norm $\mathcal{R}(\Theta) = |||\Theta|||_{\text{nuc}} = \sum_{j=1}^m \sigma_j(\Theta)$ is decomposable with respect to appropriately chosen matrix subspaces, and we exploit this fact heavily in our analysis.

We model the behavior of both exactly and approximately low-rank matrices by enforcing a sparsity condition on the vector $\sigma(\Theta) = \begin{bmatrix} \sigma_1(\Theta) & \sigma_2(\Theta) & \cdots & \sigma_d(\Theta) \end{bmatrix}$ of singular values. In particular, for a parameter $q \in [0, 1]$, we define the ℓ_q-"ball" of matrices

$$\mathbb{B}_q(R_q) := \big\{ \Theta \in \mathbb{R}^{m \times m} \mid \sum_{j=1}^m |\sigma_j(\Theta)|^q \leq R_q \big\}. \tag{6.35}$$

Note that if $q = 0$, then $\mathbb{B}_0(R_0)$ consists of the set of all matrices with rank at most $r = R_0$. On the other hand, for $q \in (0, 1]$, the set $\mathbb{B}_q(R_q)$ contains matrices of all ranks, but enforces a relatively fast rate of decay on the singular values.

Bounds for matrix compressed sensing

We begin by considering the compressed sensing version of matrix regression discussed in Chapter 4, a model first introduced by Recht et al. [117], and later studied by other authors (e.g., [79]). In this model, the observation matrices $X_i \in \mathbb{R}^{m \times m}$ are dense and drawn from some random ensemble. The simplest example is the standard Gaussian ensemble, in which each entry of X_i is drawn i.i.d. as standard normal $N(0, 1)$. Note that X_i is a dense matrix in general; this in an important contrast with the matrix completion setting to follow shortly.

Here we consider a more general ensemble of random matrices X_i, in which each matrix $X_i \in \mathbb{R}^{m \times m}$ is drawn i.i.d. from a zero-mean normal distribution in $\mathbb{R}^{m^2}$ with covariance matrix $\Sigma \in \mathbb{R}^{m^2 \times m^2}$. The setting $\Sigma = I_{m^2 \times m^2}$ recovers the standard Gaussian ensemble studied in past work. As usual, we let $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ define the maximum and minimum

eigenvalues of Σ, and we define $\zeta_{\text{mat}}(\Sigma) = \sup_{\|u\|_2=1} \sup_{\|v\|_2=1} \text{var}\big(\langle\langle X,\ uv^T \rangle\rangle\big)$, corresponding to the maximal variance of X when projected onto rank one matrices. For the identity ensemble, we have $\zeta_{\text{mat}}(I) = 1$.

We now state a result on the convergence of the updates (6.16) when applied to a statistical problem involving a matrix $\Theta^* \in \mathbb{B}_q(R_q)$. The convergence rate depends on the contraction coefficient

$$\kappa := \left\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\right\}\left\{1 - \chi_n(\Sigma)\right\}^{-1},$$

where $\chi_n(\Sigma) := \frac{c_1 \zeta_{\text{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} R_q\big(\frac{m}{n}\big)^{1-q/2}$ for some universal constant c_1. In the case $q = 0$, corresponding to matrices with rank at most r, note that we have $R_0 = r$. With this notation, we have the following convergence guarantee:

Corollary 6.4 (Low-rank matrix recovery)**.** *Under conditions of Theorem 6.1, consider the semidefinite program (6.15) with $\rho \leq |||\Theta^*|||_{\text{nuc}}$, and suppose that we apply the projected gradient updates (6.16) with $\gamma_u = 2\sigma_{\max}(\Sigma)$.*

(a) Exactly low-rank: *In the case $q = 0$, if Θ^* has rank $r < d$, then with probability at least $1 - \exp(-c_0 m)$, the iterates (6.16) satisfy the bound*

$$|||\Theta^t - \widehat{\Theta}|||_F^2 \leq \kappa^t |||\Theta^0 - \widehat{\Theta}|||_F^2 + c_2\ \chi_n(\Sigma)\ |||\widehat{\Theta} - \Theta^*|||_F^2 \qquad \text{for all } t = 0, 1, 2, \ldots. \quad (6.36)$$

(b) Approximately low-rank: *If $\Theta^* \in \mathbb{B}_q(R_q)$ for some $q \in (0, 1]$, then with probability at least $1 - \exp(-c_0 m)$, the iterates (6.16) satisfy*

$$|||\Theta^t - \widehat{\Theta}|||_F^2 \leq \kappa^t\, |||\Theta^0 - \widehat{\Theta}|||_F^2 + c_2 \chi_n(\Sigma) \left\{R_q\left(\frac{m}{n}\right)^{1-q/2} + |||\widehat{\Theta} - \Theta^*|||_F^2\right\}, \quad (6.37)$$

Although quantitative aspects of the rates are different, Corollary 6.4 is analogous to Corollary 6.2. For the case of exactly low rank matrices (part (a)), geometric convergence is guaranteed up to a tolerance involving the statistical error $|||\widehat{\Theta} - \Theta^*|||_F^2$. For the case of approximately low rank matrices (part (b)), the tolerance term involves an additional factor of $R_q\big(\frac{m}{n}\big)^{1-q/2}$. Again, from known results on minimax rates for matrix estimation [119], this term is known to be of comparable or lower order than the quantity $|||\widehat{\Theta} - \Theta^*|||_F^2$. As before, it is also possible to derive an analogous corollary of Theorem 6.2 for estimating low-rank matrices; in the interests of space, we leave such a development to the reader.

Bounds for matrix completion

In this model, observation y_i is a noisy version of a randomly selected entry $\Theta^*_{a(i),b(i)}$ of the unknown matrix Θ^*. Applications of this matrix completion problem include collaborative filtering [126], where the rows of the matrix Θ^* correspond to users, and the columns correspond to items (e.g., movies in the Netflix database), and the entry Θ^*_{ab} corresponds to user's a rating of item b. Given observations of only a subset of the entries of Θ^*, the goal is to fill in, or complete the matrix, thereby making recommendations of movies that a given user has not yet seen.

Matrix completion can be viewed as a particular case of the matrix regression model (6.14), in particular by setting $X_i = E_{a(i)b(i)}$, corresponding to the matrix with a single one in position $(a(i), b(i))$, and zeroes in all other positions. Note that these observation matrices are extremely sparse, in contrast to the compressed sensing model. As shown in Chapter 5, nuclear-norm based estimators for matrix completion are known to have good statistical properties (see also the papers [33, 115, 126]). Here we consider the M-estimator

$$\widehat{\Theta} \in \arg\min_{\Theta \in \Omega} \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - \Theta_{a(i)b(i)}\big)^2 \quad \text{such that } |||\Theta|||_{\text{nuc}} \leq \rho, \tag{6.38}$$

where $\Omega = \{\Theta \in \mathbb{R}^{m \times m} \mid \|\Theta\|_\infty \leq \frac{\alpha}{m}\}$ is the set of matrices with bounded elementwise ℓ_∞ norm. This constraint eliminates matrices that are overly "spiky" (i.e., concentrate too much of their mass in a single position); as discussed in Chapter 5, such spikiness control is necessary in order to bound the non-identifiable component of the matrix completion model.

Corollary 6.5 (Matrix completion)**.** *Under the conditions of Theorem 6.1, suppose that $\Theta^* \in \mathbb{B}_q(R_q)$, and that we solve the program (6.38) with $\rho \leq |||\Theta^*|||_{\text{nuc}}$. As long as $n > c_0 R_q^{1/(1-q/2)}\, m \log m$ for a sufficiently large constant c_0, then with probability at least $1 - \exp(-c_1 m \log m)$, there is a contraction coefficient $\kappa_t \in (0,1)$ that decreases with t such that for all iterations $t = 0, 1, 2, \ldots$,*

$$|||\Theta^{t+1} - \widehat{\Theta}|||_F^2 \leq \kappa_t^t \, |||\Theta^0 - \widehat{\Theta}|||_F^2 + c_2 \Big\{ R_q \big(\frac{\alpha^2 m \log m}{n}\big)^{1-q/2} + |||\widehat{\Theta} - \Theta^*|||_F^2 \Big\}. \tag{6.39}$$

In some cases, the bound on $\Theta\|_\infty$ in the algorithm (6.38) might be unknown, or undesirable. While this constraint is necessary in general 5.3.2, it can be avoided if more information such as the sampling distribution (that is, the distribution of X_i) is known and used to construct the estimator. In this case, Koltchinskii et al. [74] show error bounds on a nuclear norm penalized estimator without requiring ℓ_∞ bound on $\widehat{\Theta}$.

Again a similar corollary of Theorem 6.2 can be derived by combining the proof of Corollary 6.5 with that of Theorem 6.2. An interesting aspect of this problem is that the condition 6.24(b) takes the form $\lambda_n > \frac{c\alpha\sqrt{m \log m/n}}{1-\kappa}$, where α is a bound on $\|\Theta\|_\infty$. This condition is independent of $\bar{\rho}$, and hence, given a sample size as stated in the corollary, the algorithm always converges geometrically for any radius $\bar{\rho} \geq |||\Theta^*|||_{\text{nuc}}$.

6.3.4 Matrix decomposition problems

In recent years, various researchers have studied methods for solving the problem of matrix decomposition (e.g., [38, 36, 148, 2, 62]). The basic problem has the following form: given a pair of unknown matrices Θ^* and Γ^*, both lying in $\mathbb{R}^{d_1 \times d_2}$, suppose that we observe a third matrix specified by the model $Y = \Theta^* + \Gamma^* + W$, where $W \in \mathbb{R}^{d_1 \times d_2}$ represents observation noise. Typically the matrix Θ^* is assumed to be low-rank, and some low-dimensional structural constraint is assumed on the matrix Γ^*. For example, the papers [38, 36, 62] consider the setting in which Γ^* is sparse, while Xu et al. [148] consider a column-sparse model, in which only a few of the columns of Γ^* have non-zero entries. In order to illustrate the application of our general result to this setting, here we consider the low-rank plus column-sparse framework [148]. (We note that since the ℓ_1-norm is decomposable, similar results can easily be derived for the low-rank plus entrywise-sparse setting as well.)

Since Θ^* is assumed to be low-rank, as before we use the nuclear norm $|||\Theta|||_{\text{nuc}}$ as a regularizer (see Section 6.2.4). We assume that the unknown matrix $\Gamma^* \in \mathbb{R}^{d_1 \times d_2}$ is column-sparse, say with at most $k < d_2$ non-zero columns. A suitable convex regularizer for this matrix structure is based on the *columnwise* $(1,2)$*-norm*, given by

$$\|\Gamma\|_{1,2} := \sum_{j=1}^{d_2} \|\Gamma_j\|_2, \tag{6.40}$$

where $\Gamma_j \in \mathbb{R}^{d_1}$ denotes the j^{th} column of Γ. Note also that the dual norm is given by the *elementwise* $(\infty, 2)$*-norm* $\|\Gamma\|_{\infty,2} = \max_{j=1,\ldots,d_2} \|\Gamma_j\|_2$, corresponding to the maximum ℓ_2-norm over columns.

In order to estimate the unknown pair (Θ^*, Γ^*), we consider the M-estimator

$$(\widehat{\Theta}, \widehat{\Gamma}) := \arg\min_{\Theta,\Gamma} |||Y - \Theta - \Gamma|||_F^2 \quad \text{such that} \quad |||\Theta|||_{\text{nuc}} \leq \rho_\Theta, \quad \|\Gamma\|_{1,2} \leq \rho_\Gamma \text{ and } \|\Theta\|_{\infty,2} \leq \frac{\alpha}{\sqrt{d_2}} \tag{6.41}$$

The first two constraints restrict Θ and Γ to a nuclear norm ball of radius ρ_Θ and a $(1,2)$-norm ball of radius ρ_Γ, respectively. The final constraint controls the "spikiness" of the low-rank component Θ, as measured in the $(\infty, 2)$-norm, corresponding to the maximum ℓ_2-norm over the columns. As with the elementwise ℓ_∞-bound for matrix completion, this additional constraint is required in order to limit the non-identifiability in matrix decomposition. (See the paper [2] for more discussion of non-identifiability issues in matrix decomposition.)

With this set-up, consider the projected gradient algorithm when applied to the matrix decomposition problem: it generates a sequence of matrix pairs (Θ^t, Γ^t) for $t = 0, 1, 2, \ldots$, and the optimization error is characterized in terms of the matrices $\widehat{\Delta}^t_\Theta := \Theta^t - \widehat{\Theta}$ and $\widehat{\Delta}^t_\Gamma := \Gamma^t - \widehat{\Gamma}$. Finally, we measure the optimization error at time t in terms of the squared

Frobenius error $e^2(\widehat{\Delta}^t_\Theta, \widehat{\Delta}^t_\Gamma) := \|\widehat{\Delta}^t_\Theta\|_F^2 + \|\widehat{\Delta}^t_\Gamma\|_F^2$, summed across both the low-rank and column-sparse components.

Corollary 6.6 (Matrix decomposition)**.** *Under the conditions of Theorem 6.1, suppose that* $\|\Theta^*\|_{\infty,2} \leq \frac{\alpha}{\sqrt{d_2}}$ *and* Γ^* *has at most* k *non-zero columns. If we solve the convex program* (6.41) *with* $\rho_\Theta \leq |||\Theta^*|||_{\text{nuc}}$ *and* $\rho_\Gamma \leq \|\Gamma^*\|_{1,2}$*, then for all iterations* $t = 0, 1, 2, \ldots,$

$$e^2(\widehat{\Delta}^t_\Theta, \widehat{\Delta}^t_\Gamma) \leq \left(\frac{3}{4}\right)^t e^2(\widehat{\Delta}^0_\Theta, \widehat{\Delta}^0_\Gamma) + c\left(|||\widehat{\Gamma} - \Gamma^*|||_F^2 + \alpha^2 \frac{k}{d_2}\right).$$

This corollary has some unusual aspects, relative to the previous corollaries. First of all, in contrast to the previous results, the guarantee is a deterministic one (as opposed to holding with high probability). More specifically, the RSC/RSM conditions hold deterministic sense, which should be contrasted with the high probability statements given in Corollaries 6.2-6.5. Consequently, the effective conditioning of the problem does not depend on sample size and we are guaranteed geometric convergence at a fixed rate, independent of sample size. The additional tolerance term is completely independent of the rank of Θ^* and only depends on the column-sparsity of Γ^*.

6.4 Simulation results

In this section, we provide some experimental results that confirm the accuracy of our theoretical results, in particular showing excellent agreement with the linear rates predicted by our theory. In addition, the rates of convergence slow down for smaller sample sizes, which lead to problems with relatively poor conditioning. In all the simulations reported below, we plot the log error $\|\theta^t - \widehat{\theta}\|$ between the iterate θ^t at time t versus the final solution $\widehat{\theta}$. Each curve provides the results averaged over five random trials, according to the ensembles which we now describe.

6.4.1 Sparse regression

We begin by considering the linear regression model $y = X\theta^* + w$ where θ^* is the unknown regression vector belonging to the set $\mathbb{B}_q(R_q)$, and i.i.d. observation noise $w_i \sim N(0, 0.25)$. We consider a family of ensembles for the random design matrix $X \in \mathbb{R}^{n \times d}$. In particular, we construct X by generating each row $x_i \in \mathbb{R}^d$ independently according to following procedure. Let $z_1, \ldots, z_n$ be an i.i.d. sequence of $N(0, 1)$ variables, and fix some correlation parameter $\omega \in [0, 1)$. We first initialize by setting $x_{i,1} = z_1/\sqrt{1 - \omega^2}$, and then generate the remaining entries by applying the recursive update $x_{i,t+1} = \omega x_{i,t} + z_t$ for $t = 1, 2, \ldots, d - 1$, so that $x_i \in \mathbb{R}^d$ is a zero-mean Gaussian random vector. It can be verified that all the eigenvalues of $\Sigma = \text{cov}(x_i)$ lie within the interval $[\frac{1}{(1+\omega)^2}, \frac{2}{(1-\omega)^2(1+\omega)}]$, so that Σ has a a finite condition number for all $\omega \in [0, 1)$. At one extreme, for $\omega = 0$, the matrix Σ is the identity, and so

has condition number equal to 1. As $\omega \to 1$, the matrix Σ becomes progressively more ill-conditioned, with a condition number that is very large for ω close to one. As a consequence, although incoherence conditions like the restricted isometry property can be satisfied when $\omega = 0$, they will fail to be satisfied (w.h.p.) once ω is large enough.

For this random ensemble of problems, we have investigated convergence rates for a wide range of dimensions d and radii R_q. Since the results are relatively uniform across the choice of these parameters, here we report results for dimension $d = 20,000$, and radius $R_q = \lceil (\log d)^2 \rceil$. In the case $q = 0$, the radius $R_0 = k$ corresponds to the sparsity level. The per iteration cost in this case is $\mathcal{O}(nd)$. In order to reveal dependence of convergence rates on sample size, we study a range of the form $n = \lceil \alpha \; k \log d \rceil$, where the *order parameter* $\alpha > 0$ is varied.

Our first experiment is based on taking the correlation parameter $\omega = 0$, and the ℓ_q-ball parameter $q = 0$, corresponding to exact sparsity. We then measure convergence rates for sample sizes specified by $\alpha \in \{1, 1.25, 5, 25\}$. As shown by the results plotted in panel (a) of Figure 6.2, projected gradient descent fails to converge for $\alpha = 1$ or $\alpha = 1.25$; in both these cases, the sample size n is too small for the RSC and RSM conditions to hold, so that a constant step size leads to oscillatory behavior in the algorithm. In contrast, once the order parameter α becomes large enough to ensure that the RSC/RSM conditions hold (w.h.p.), we observe a geometric convergence of the error $\|\theta^t - \widehat{\theta}\|_2$. Moreover the convergence rate is faster for $\alpha = 25$ compared to $\alpha = 5$, since the RSC/RSM constants are better with larger sample size. Such behavior is in agreement with the conclusions of Corollary 6.2, which predicts that the the convergence rate should improve as the number of samples n is increased.

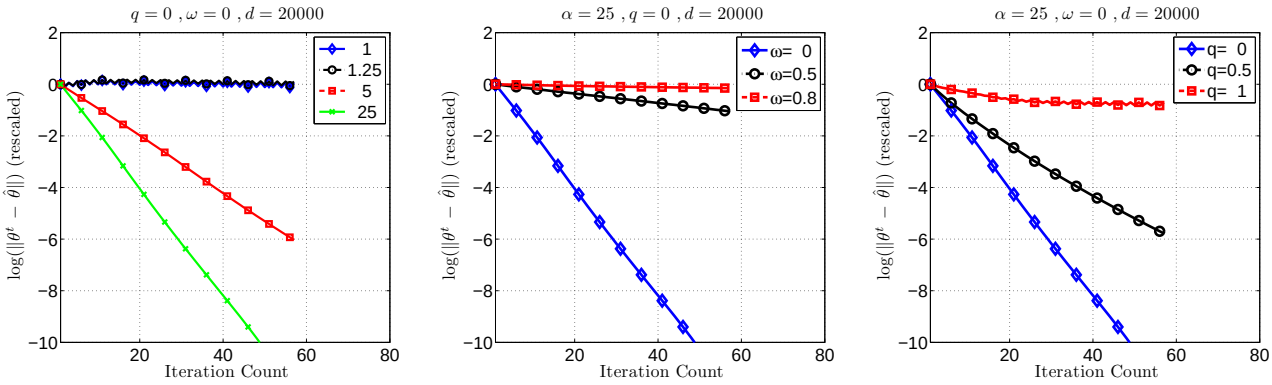


Figure 6.2. Plot of the log of the optimization error $\log(\|\theta^t - \widehat{\theta}\|_2)$ in the sparse linear regression problem, rescaled so the plots start at 0. In this problem, $d = 20000$, $k = \lceil \log d \rceil$, $n = \alpha k \log d$. Plot (a) shows convergence for the exact sparse case with $q = 0$ and $\Sigma = I$ (i.e. $\omega = 0$). In panel (b), we observe how convergence rates change as the correlation parameter ω is varied for $q = 0$ and $\alpha = 25$. Plot (c) shows the convergence rates when $\omega = 0$, $\alpha = 25$ and q is varied.

On the other hand, Corollary 6.2 also predicts that convergence rates should be slower when the condition number of Σ is worse. In order to test this prediction, we again studied an exactly sparse problem ($q = 0$), this time with the fixed sample size $n = \lceil 25k \log d \rceil$, and we varied the correlation parameter $\omega \in \{0, 0.5, 0.8\}$. As shown in panel (b) of Figure 6.2, the convergence rates slow down as the correlation parameter is increased and for the case of extremely high correlation of $\omega = 0.8$, the optimization error curve is almost flat—the method makes very slow progress in this case.

A third prediction of Corollary 6.2 is that the convergence of projected gradient descent should become slower as the sparsity parameter q is varied between exact sparsity ($q = 0$), and the least sparse case ($q = 1$). (In particular, note for $n > \log d$, the quantity χ_n from equation (6.30) is monotonically increasing with q.) Panel (c) of Figure 6.2 shows convergence rates for the fixed sample size $n = 25k \log d$ and correlation parameter $\omega = 0$, and with the sparsity parameter $q \in \{0, 0.5, 1.0\}$. As expected, the convergence rate slows down as q increases from 0 to 1. Corollary 6.2 further captures how the contraction factor changes as the problem parameters (k, d, n) are varied. In particular, it predicts that as we change the triplet simultaneously, while holding the ratio $\alpha = k \log d/n$ constant, the convergence rate should stay the same. We recall that this phenomenon was indeed demonstrated in Figure 1.1 in Section 1.3.

6.4.2 Low-rank matrix estimation

We also performed experiments with two different versions of low-rank matrix regression. Our simulations applied to instances of the observation model $y_i = \langle\langle X_i, \Theta^* \rangle\rangle + w_i$, for $i = 1, 2, \ldots, n$, where $\Theta^* \in \mathbb{R}^{200 \times 200}$ is a fixed unknown matrix, $X_i \in \mathbb{R}^{200 \times 200}$ is a matrix of covariates, and $w_i \sim N(0, 0.25)$ is observation noise. In analogy to the sparse vector problem, we performed simulations with the matrix Θ^* belonging to the set $\mathbb{B}_q(R_q)$ of approximately low-rank matrices, as previously defined in equation (6.35) for $q \in [0, 1]$. The case $q = 0$ corresponds to the set of matrices with rank at most $r = R_0$, whereas the case $q = 1$ corresponds to the ball of matrices with nuclear norm at most R_1.

In our first set of matrix experiments, we considered the matrix version of compressed sensing [116], in which each matrix $X_i \in \mathbb{R}^{200 \times 200}$ is randomly formed with i.i.d. $N(0, 1)$ entries, as described in Section 6.3.3. In the case $q = 0$, we formed a matrix $\Theta^* \in \mathbb{R}^{200 \times 200}$ with rank $R_0 = 5$, and performed simulations over the sample sizes $n = \alpha R_0 m$, with the parameter $\alpha \in \{1, 1.25, 5, 25\}$. The per iteration cost in this case is $\mathcal{O}(nm^2)$. As seen in panel (a) of Figure 6.3, the projected gradient descent method exhibits behavior that is qualitatively similar to that for the sparse linear regression problem. More specifically, it fails to converge when the sample size (as reflected by the order parameter α) is too small, and converges geometrically with a progressively faster rate as α is increased. We have also observed similar types of scaling as the matrix sparsity parameter is increased from $q = 0$ to $q = 1$.

In our second set of matrix experiments, we studied the behavior of projected gradient

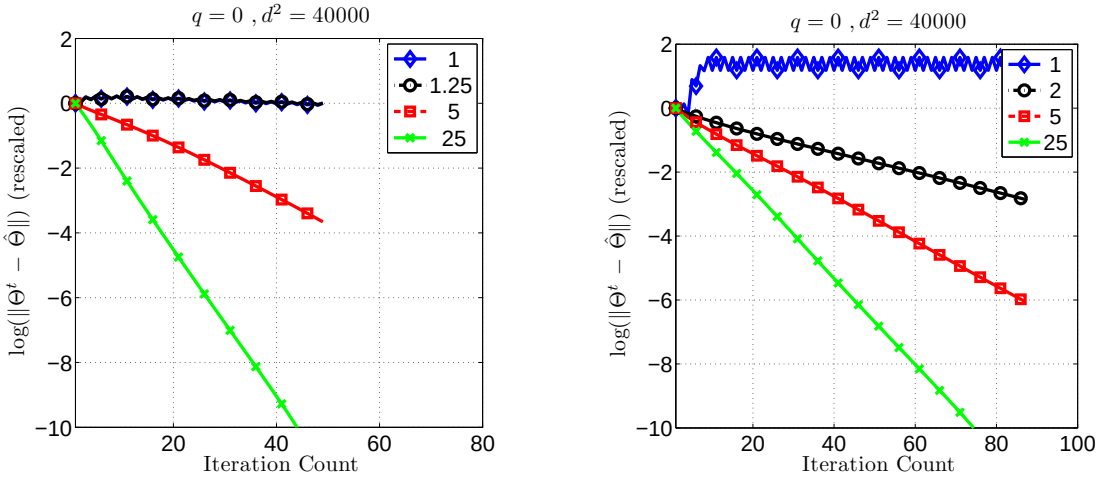


Figure 6.3. (a) Plot of log Frobenius error $\log(\|\Theta^t - \widehat{\Theta}\|_F)$ versus number of iterations in matrix compressed sensing for a matrix size $m = 200$ with rank $R_0 = 5$, and sample sizes $n = \alpha R_0 m$. For $\alpha \in \{1, 1.25\}$, the algorithm oscillates, whereas geometric convergence is obtained for $\alpha \in \{5, 25\}$, consistent with the theoretical prediction. (b) Plot of log Frobenius error $\log(\|\Theta^t - \widehat{\Theta}\|_F)$ versus number of iterations in matrix completion with $m = 200$, $R_0 = 5$, and $n = \alpha R_o m \log(m)$ with $\alpha \in \{1, 2, 5, 25\}$. For $\alpha \in \{2, 5, 25\}$ the algorithm enjoys geometric convergence.

descent for the problem of matrix completion, as described in Section 6.3.3. For this problem, we again studied matrices of dimension $m = 200$ and rank $R_0 = 5$, and we varied the sample size as $n = \alpha\, R_0\, m \log m$ for $\alpha \in \{1, 2, 5, 25\}$. As shown in panel (b) of Figure 6.3, projected gradient descent for matrix completion also enjoys geometric convergence for α large enough.

6.5 Proofs

In this section, we provide the proofs of our results. Recall that we use $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ to denote the optimization error, and $\Delta^* = \widehat{\theta} - \theta^*$ to denote the statistical error. For future reference, we point out a slight weakening of restricted strong convexity (RSC), useful for obtaining parts of our results. As the to follow reveals, it is only necessary to enforce an RSC condition of the form

$$\mathcal{T}_{\mathcal{L}}(\theta^t; \widehat{\theta}) \geq \frac{\gamma_\ell}{2} \|\theta^t - \widehat{\theta}\|^2 - \tau_\ell(\mathcal{L}_n)\, \mathcal{R}^2(\theta^t - \widehat{\theta}) - \delta^2, \tag{6.42}$$

which is milder than the original RSC condition (6.8), in that it applies only to differences of the form $\theta^t - \widehat{\theta}$, and allows for additional slack δ. We make use of this refined notion in the proofs of various results to follow.

With this relaxed RSC condition and the same RSM condition as before, our proof shows that

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \|\theta^0 - \widehat{\theta}\|^2 + \frac{\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) + 2\delta^2/\gamma_u}{1-\kappa} \qquad \text{for all iterations } t = 0, 1, 2, \ldots. \tag{6.43}$$

Note that this result reduces to the previous statement when $\delta = 0$. This extension of Theorem 6.1 is used in the proofs of Corollaries 6.5 and 6.6.

We will assume without loss of generality that all the iterates lie in the subset Ω' of Ω. This can be ensured by augmenting the loss with the indicator of Ω' or equivalently performing projections on the set $\Omega' \cap \mathbb{B}_{\mathcal{R}}(\rho)$ as mentioned earlier.

6.5.1 Proof of Theorem 6.1

Recall that Theorem 6.1 concerns the constrained problem (6.1). The proof is based on two technical lemmas. The first lemma guarantees that at each iteration $t = 0, 1, 2, \ldots$, the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ belongs to an interesting constraint set defined by the regularizer.

Lemma 6.1. *Let $\widehat{\theta}$ be any optimum of the constrained problem (6.1) for which $\mathcal{R}(\widehat{\theta}) = \rho$. Then for any iteration $t = 1, 2, \ldots$ and for any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ belongs to the set*

$$\mathbb{S}(\mathcal{M}; \overline{\mathcal{M}}; \theta^*) := \left\{ \Delta \in \Omega \mid \mathcal{R}(\Delta) \leq 2\,\Psi(\overline{\mathcal{M}})\,\|\Delta\| + 2\mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)) + 2\mathcal{R}(\Delta^*) + \Psi(\overline{\mathcal{M}})\|\Delta^*\| \right\}. \tag{6.44}$$

The proof of this lemma, provided in Appendix D.1.1, exploits the decomposability of the regularizer in an essential way.

The structure of the set (6.44) takes a simpler form in the special case when $\mathcal{M}$ is chosen to contain θ^* and $\overline{\mathcal{M}} = \mathcal{M}$. In this case, we have $\mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)) = 0$, and hence the optimization error $\widehat{\Delta}^t$ satisfies the inequality

$$\mathcal{R}(\widehat{\Delta}^t) \leq 2\,\Psi(\mathcal{M})\left\{\|\widehat{\Delta}^t\| + \|\Delta^*\|\right\} + 2\mathcal{R}(\Delta^*). \tag{6.45}$$

An inequality of this type, when combined with the definitions of RSC/RSM, allows us to establish the curvature conditions required to prove globally geometric rates of convergence.

We now state a second lemma under the more general RSC condition (6.42):

Lemma 6.2. *Under the RSC condition (6.42) and RSM condition (6.10), for all $t = 0, 1, 2, \ldots$, we have*

$$\gamma_u \langle \theta^t - \theta^{t+1}, \theta^t - \widehat{\theta} \rangle \geq \Big\{ \frac{\gamma_u}{2} \|\theta^t - \theta^{t+1}\|^2 - \tau_u(\mathcal{L}_n) \mathcal{R}^2(\theta^{t+1} - \theta^t) \Big\} + \Big\{ \frac{\gamma_\ell}{2} \|\theta^t - \widehat{\theta}\|^2 - \tau_\ell(\mathcal{L}_n)\, \mathcal{R}^2(\theta^t - \widehat{\theta}) - \delta^2 \Big\}. \tag{6.46}$$

The proof of this lemma, provided in Appendix D.1.2, follows along the lines of the intermediate result within Theorem 2.2.8 of Nesterov [101], but with some care required to handle the additional terms that arise in our weakened forms of strong convexity and smoothness.

Using these auxiliary results, let us now complete the the proof of Theorem 6.1. We first note the elementary relation

$$\|\theta^{t+1} - \widehat{\theta}\|^2 = \|\theta^t - \widehat{\theta} - \theta^t + \theta^{t+1}\|^2 = \|\theta^t - \widehat{\theta}\|^2 + \|\theta^t - \theta^{t+1}\|^2 - 2\langle \theta^t - \widehat{\theta}, \theta^t - \theta^{t+1} \rangle. \tag{6.47}$$

We now use Lemma 6.2 and the more general form of RSC (6.42) to control the cross-term, thereby obtaining the upper bound

$$\begin{aligned} \|\theta^{t+1} - \widehat{\theta}\|^2 &\leq \|\theta^t - \widehat{\theta}\|^2 - \frac{\gamma_\ell}{\gamma_u} \|\theta^t - \widehat{\theta}\|^2 + \frac{2\tau_u(\mathcal{L}_n)}{\gamma_u} \mathcal{R}^2(\theta^{t+1} - \theta^t) + \frac{2\tau_\ell(\mathcal{L}_n)}{\gamma_u} \mathcal{R}^2(\theta^t - \widehat{\theta}) + \frac{2\delta^2}{\gamma_u} \\ &= \big(1 - \frac{\gamma_\ell}{\gamma_u}\big) \|\theta^t - \widehat{\theta}\|^2 + \frac{2\tau_u(\mathcal{L}_n)}{\gamma_u} \mathcal{R}^2(\theta^{t+1} - \theta^t) + \frac{2\tau_\ell(\mathcal{L}_n)}{\gamma_u} \mathcal{R}^2(\theta^t - \widehat{\theta}) + \frac{2\delta^2}{\gamma_u}. \end{aligned}$$

We now observe that by triangle inequality and the Cauchy-Schwarz inequality,

$$\mathcal{R}^2(\theta^{t+1} - \theta^t) \leq \big(\mathcal{R}(\theta^{t+1} - \widehat{\theta}) + \mathcal{R}(\widehat{\theta} - \theta^t)\big)^2 \leq 2\mathcal{R}^2(\theta^{t+1} - \widehat{\theta}) + 2\mathcal{R}^2(\theta^t - \widehat{\theta}).$$

Recall the definition of the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$, we have the upper bound

$$\|\widehat{\Delta}^{t+1}\|^2 \leq \big(1 - \frac{\gamma_\ell}{\gamma_u}\big) \|\widehat{\Delta}^t\|^2 + \frac{4\tau_u(\mathcal{L}_n)}{\gamma_u} \mathcal{R}^2(\widehat{\Delta}^{t+1}) + \frac{4\tau_u(\mathcal{L}_n) + 2\tau_\ell(\mathcal{L}_n)}{\gamma_u} \mathcal{R}^2(\widehat{\Delta}^t) + \frac{2\delta^2}{\gamma_u}. \tag{6.48}$$

We now apply Lemma 6.1 to control the terms involving $\mathcal{R}^2$. In terms of squared quantities, the inequality (6.44) implies that

$$\mathcal{R}^2(\widehat{\Delta}^t) \leq 4\, \Psi^2(\overline{\mathcal{M}}^\perp)\, \|\widehat{\Delta}^t\|^2 + 2\nu^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \qquad \text{for all } t = 0, 1, 2, \ldots,$$

where we recall that $\Psi^2(\overline{\mathcal{M}}^{\perp})$ is the subspace compatibility (3.21) and $\nu^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ accumulates all the residual terms. Applying this bound twice—once for t and once for $t+1$—and substituting into equation (6.48) yields that $\{1 - \frac{16\Psi^2(\overline{\mathcal{M}}^{\perp})\tau_u(\mathcal{L}_n)}{\gamma_u}\}\|\Delta^{t+1}\|^2$ is upper bounded by

$$\Big\{1 - \frac{\gamma_\ell}{\gamma_u} + \frac{16\Psi^2(\overline{\mathcal{M}}^{\perp})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)}{\gamma_u}\Big\}\|\Delta^t\|^2 + \frac{16\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)\nu^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})}{\gamma_u} + \frac{2\delta^2}{\gamma_u}.$$

Under the assumptions of Theorem 6.1, we are guaranteed that $\frac{16\Psi^2(\overline{\mathcal{M}}^{\perp})\tau_u(\mathcal{L}_n)}{\gamma_u} < 1/2$, and so we can re-arrange this inequality into the form

$$\|\Delta^{t+1}\|^2 \leq \kappa\,\|\Delta^t\|^2 + \epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) + \frac{2\delta^2}{\gamma_u} \tag{6.49}$$

where κ and $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ were previously defined in equations (6.17) and (6.18) respectively. Iterating this recursion yields

$$\|\Delta^{t+1}\|^2 \leq \kappa^t\,\|\Delta^0\|^2 + \left(\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) + \frac{2\delta^2}{\gamma_u}\right)\Big(\sum_{j=0}^{t}\kappa^j\Big).$$

The assumptions of Theorem 6.1 guarantee that $\kappa \in (0, 1)$, so that summing the geometric series yields the claim (6.19).

6.5.2 Proof of Theorem 6.2

The Lagrangian version of the optimization program is based on solving the convex program (6.2), with the objective function $\phi(\theta) = \mathcal{L}_n(\theta) + \lambda_n\mathcal{R}(\theta)$. Our proof is based on analyzing the error $\phi(\theta^t) - \phi(\widehat{\theta})$ as measured in terms of this objective function. It requires two technical lemmas, both of which are stated in terms of a given tolerance $\bar{\eta} > 0$, and an integer $T > 0$ such that

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \bar{\eta} \qquad \text{for all } t \geq T. \tag{6.50}$$

Our first technical lemma is analogous to Lemma 6.1, and restricts the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ to a cone-like set.

Lemma 6.3 (Iterated Cone Bound (ICB)). *Let $\widehat{\theta}$ be any optimum of the regularized M-estimator (6.2). Under condition (6.50) with parameters $(T, \bar{\eta})$, for any iteration $t \geq T$ and for any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ satisfies*

$$\mathcal{R}(\widehat{\Delta}^t) \leq 4\Psi(\overline{\mathcal{M}})\|\widehat{\Delta}^t\| + 8\Psi(\overline{\mathcal{M}})\|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)) + 2\min\left(\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\right) \tag{6.51}$$

Our next lemma guarantees sufficient decrease of the objective value difference $\phi(\theta^t) - \phi(\widehat{\theta})$. Lemma 6.3 plays a crucial role in its proof. Recall the definition (6.22) of the compound contraction coefficient $\kappa(\mathcal{L}_n; \overline{\mathcal{M}})$, defined in terms of the related quantities $\xi(\overline{\mathcal{M}})$ and $\beta(\overline{\mathcal{M}})$. Throughout the proof, we drop the arguments of κ, ξ and β so as to ease notation.

Lemma 6.4. *Under the RSC (6.42) and RSM conditions (6.10), as well as assumption (6.50) with parameters $(\bar{\eta}, T)$, for all $t \geq T$, we have*

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^{t-T}(\phi(\theta^T) - \phi(\widehat{\theta})) + \frac{2}{1-\kappa}\xi(\mathcal{M})\,\beta(\mathcal{M})(\varepsilon^2 + \bar{\epsilon}^2_{stat}),$$

where $\varepsilon := 2\min(\bar{\eta}/\lambda_n, \bar{\rho})$ and $\bar{\epsilon}_{stat} := 8\Psi(\overline{\mathcal{M}})\|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))$.

We are now in a position to prove our main theorem, in particular via a recursive application of Lemma 6.4. At a high level, we divide the iterations $t = 0, 1, 2, \ldots$ into a series of disjoint epochs $[T_k, T_{k+1})$ with $0 = T_0 \leq T_1 \leq T_2 \leq \cdots$. Moreover, we define an associated sequence of tolerances $\bar{\eta}_0 > \bar{\eta}_1 > \cdots$ such that at the end of epoch $[T_{k-1}, T_k)$, the optimization error has been reduced to $\bar{\eta}_k$. Our analysis guarantees that $\phi(\theta^t) - \phi(\widehat{\theta}) \leq \bar{\eta}_k$ for all $t \geq T_k$, allowing us to apply Lemma 6.4 with smaller and smaller values of $\bar{\eta}$ until it reduces to the statistical error $\bar{\epsilon}_{\text{stat}}$.

At the first iteration, we have no a priori bound on the error $\bar{\eta}_0 = \phi(\theta^0) - \phi(\widehat{\theta})$. However, since Lemma 6.4 involves the quantity $\varepsilon = \min(\bar{\eta}/\lambda_n, \bar{\rho})$, we may still apply it[5] at the first epoch with $\varepsilon_0 = \bar{\rho}$ and $T_0 = 0$. In this way, we conclude that for all $t \geq 0$,

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^t(\phi(\theta^0) - \phi(\widehat{\theta})) + \frac{2}{1-\kappa}\xi\beta(\bar{\rho}^2 + \bar{\epsilon}^2_{\text{stat}}).$$

Now since the contraction coefficient $\kappa \in (0,1)$, for all iterations $t \geq T_1 := (\lceil \log(2\,\bar{\eta}_0/\bar{\eta}_1)/\log(1/\kappa)\rceil)_+$, we are guaranteed that

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \underbrace{\frac{4\,\xi\beta}{1-\kappa}(\bar{\rho}^2 + \bar{\epsilon}^2_{\text{stat}})}_{\bar{\eta}_1} \leq \frac{8\xi\beta}{1-\kappa}\max(\bar{\rho}^2, \bar{\epsilon}^2_{\text{stat}}).$$

This same argument can now be applied in a recursive manner. Suppose that for some $k \geq 1$, we are given a pair $(\bar{\eta}_k, T_k)$ such that condition (6.50) holds. An application of Lemma 6.4 yields the bound

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^{t-T_k}(\phi(\theta^{T_k}) - \phi(\widehat{\theta})) + \frac{2\,\xi\beta}{1-\kappa}(\varepsilon_k^2 + \bar{\epsilon}^2_{\text{stat}}) \qquad \text{for all } t \geq T_k.$$

[5]It is for precisely this reason that our regularized M-estimator includes the additional side-constraint defined in terms of $\bar{\rho}$.

We now define $\bar{\eta}_{k+1} := \frac{4\xi\beta}{1-\kappa}(\varepsilon_k^2 + \bar{\epsilon}_{\text{stat}}^2)$. Once again, since $\kappa < 1$ by assumption, we can choose $T_{k+1} := \lceil \log(2\bar{\eta}_k/\bar{\eta}_{k+1})/\log(1/\kappa) \rceil + T_k$, thereby ensuring that for all $t \geq T_{k+1}$, we have

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \frac{8\xi\beta}{1-\kappa} \max(\varepsilon_k^2, \bar{\epsilon}_{\text{stat}}^2).$$

In this way, we arrive at recursive inequalities involving the tolerances $\{\bar{\eta}_k\}_{k=0}^{\infty}$ and time steps $\{T_k\}_{k=0}^{\infty}$—namely

$$\bar{\eta}_{k+1} \leq \frac{8\,\xi\beta}{1-\kappa} \max(\varepsilon_k^2, \bar{\epsilon}_{\text{stat}}^2), \qquad \text{where } \varepsilon_k = 2\,\min\{\bar{\eta}_k/\lambda_n, \bar{\rho}\}, \text{ and} \tag{6.52a}$$

$$T_k \leq k + \frac{\log(2^k\bar{\eta}_0/\bar{\eta}_k)}{\log(1/\kappa)}. \tag{6.52b}$$

Now we claim that the recursion (6.52a) can be unwrapped so as to show that

$$\bar{\eta}_{k+1} \leq \frac{\bar{\eta}_k}{4^{2^{k-1}}} \quad \text{and} \quad \frac{\bar{\eta}_{k+1}}{\lambda_n} \leq \frac{\bar{\rho}}{4^{2^k}} \qquad \text{for all } k = 1, 2, \ldots. \tag{6.53}$$

Taking these statements as given for the moment, let us now show how they can be used to upper bound the smallest k such that $\bar{\eta}_k \leq \delta^2$. If we are in the first epoch, the claim of the theorem is straightforward from equation (6.52a). If not, we first use the recursion (6.53) to upper bound the number of epochs needed and then use the inequality (6.52b) to obtain the stated result on the total number of iterations needed. Using the second inequality in the recursion (6.53), we see that it is sufficient to ensure that $\frac{\bar{\rho}\lambda_n}{4^{2^{k-1}}} \leq \delta^2$. Rearranging this inequality, we find that the error drops below δ^2 after at most

$$k_\delta \;\geq\; \log\left(\log\left(\frac{\bar{\rho}\lambda_n}{\delta^2}\right)/\log(4)\right)/\log(2) + 1 = \log_2\log_2\left(\frac{\bar{\rho}\lambda_n}{\delta^2}\right)$$

epochs. Combining the above bound on k_δ with the recursion 6.52b, we conclude that the inequality $\phi(\theta^t) - \phi(\widehat{\theta}) \leq \delta^2$ is guaranteed to hold for all iterations

$$t \geq k_\delta\left(1 + \frac{\log 2}{\log(1/\kappa)}\right) + \frac{\log\frac{\bar{\eta}_0}{\delta^2}}{\log(1/\kappa)},$$

which is the desired result.

It remains to prove the recursion (6.53), which we do via induction on the index k. We begin with base case $k = 1$. Recalling the setting of $\bar{\eta}_1$ and our assumption on λ_n in the theorem statement (6.25), we are guaranteed that $\bar{\eta}_1/\lambda_n \leq \bar{\rho}/4$, so that $\varepsilon_1 \leq \varepsilon_0 = \bar{\rho}$. By applying equation (6.52a) with $\varepsilon_1 = 2\bar{\eta}_1/\lambda_n$ and assuming $\varepsilon_1 \geq \bar{\epsilon}_{\text{stat}}$, we obtain

$$\bar{\eta}_2 \;\leq\; \frac{32\xi\beta\bar{\eta}_1^2}{(1-\kappa)\lambda_n^2} \overset{(i)}{\leq} \frac{32\xi\beta\bar{\rho}\bar{\eta}_1}{(1-\kappa)4\lambda_n} \overset{(ii)}{\leq} \frac{\bar{\eta}_1}{4}, \tag{6.54}$$

where step (i) uses the fact that $\frac{\bar{\eta}_1}{\lambda_n} \leq \frac{\bar{\rho}}{4}$, and step (ii) uses the condition (6.25) on λ_n. We have thus verified the first inequality (6.53) for $k = 1$. Turning to the second inequality in the statement (6.53), using equation 6.54, we have

$$\frac{\bar{\eta}_2}{\lambda_n} \leq \frac{\bar{\eta}_1}{4\lambda_n} \overset{(iii)}{\leq} \frac{\bar{\rho}}{16},$$

where step (iii) follows from the assumption (6.25) on λ_n. Turning to the inductive step, we again assume that $2\bar{\eta}_k/\lambda_n \geq \bar{\epsilon}_{\text{stat}}$ and obtain from inequality (6.52a)

$$\bar{\eta}_{k+1} \leq \frac{32\xi\beta\bar{\eta}_k^2}{(1-\kappa)\lambda_n^2} \overset{(iv)}{\leq} \frac{32\xi\beta\bar{\eta}_k\bar{\rho}}{(1-\kappa)\lambda_n 4^{2^{k-1}}} \overset{(v)}{\leq} \frac{\bar{\eta}_k}{4^{2^{k-1}}}.$$

Here step (iv) uses the second inequality of the inductive hypothesis (6.53) and step (v) is a consequence of the condition on λ_n as before. The second part of the induction is similarly established, completing the proof.

6.5.3 Proof of Corollary 6.1

In order to prove this claim, we must show that $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, as defined in equation (6.18), is of order lower than $\mathbb{E}[\|\widehat{\theta} - \theta^*\|^2] = \mathbb{E}[\|\Delta^*\|^2]$. We make use of the following lemma, proved in Appendix D.3:

Lemma 6.5. *If $\rho \leq \mathcal{R}(\theta^*)$, then for any solution $\widehat{\theta}$ of the constrained problem (6.1) and any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, the statistical error $\Delta^* = \widehat{\theta} - \theta^*$ satisfies the inequality*

$$\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\| + \mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)). \tag{6.55}$$

Using this lemma, we can complete the proof of Corollary 6.1. Recalling the form (6.18), under the condition $\theta^* \in \mathcal{M}$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) := \frac{32\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)\,\big(2\mathcal{R}(\Delta^*) + \Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|\big)^2}{\gamma_u}.$$

Using the assumption $\frac{(\tau_u(\mathcal{L}_n)+\tau_\ell(\mathcal{L}_n))\Psi^2(\overline{\mathcal{M}}^{\perp})}{\gamma_u} = o(1)$, it suffices to show that $\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|$. Since Corollary 6.1 assumes that $\theta^* \in \mathcal{M}$ and hence that $\Pi_{\mathcal{M}^{\perp}}(\theta^*) = 0$, Lemma 6.5 implies that $\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|$, as required.

6.5.4 Proofs of Corollaries 6.2 and 6.3

The central challenge in proving this result is verifying that suitable forms of the RSC and RSM conditions hold with sufficiently small parameters $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$.

Lemma 6.6. *Define the maximum variance* $\zeta(\Sigma) := \max_{j=1,2,\ldots,d} \Sigma_{jj}$. *Under the conditions of Corollary 6.2, there are universal positive constants* (c_0, c_1) *such that for all* $\Delta \in \mathbb{R}^d$, *we have*

$$\frac{\|X\Delta\|_2^2}{n} \geq \frac{1}{2}\|\Sigma^{1/2}\Delta\|_2^2 - c_1\zeta(\Sigma)\frac{\log d}{n}\|\Delta\|_1^2, \qquad \text{and} \tag{6.56a}$$

$$\frac{\|X\Delta\|_2^2}{n} \leq 2\|\Sigma^{1/2}\Delta\|_2^2 + c_1\zeta(\Sigma)\frac{\log d}{n}\|\Delta\|_1^2, \tag{6.56b}$$

with probability at least $1 - \exp(-c_0 n)$.

Note that this lemma implies that the RSC and RSM conditions both hold with high probability, in particular with parameters

$$\gamma_\ell = \frac{1}{2}\sigma_{\min}(\Sigma), \text{ and } \quad \tau_\ell(\mathcal{L}_n) = c_1\zeta(\Sigma)\frac{\log d}{n}, \qquad \text{for RSC, and}$$

$$\gamma_u = 2\sigma_{\max}(\Sigma) \text{ and } \quad \tau_u(\mathcal{L}_n) = c_1\zeta(\Sigma)\frac{\log d}{n} \qquad \text{for RSM.}$$

This lemma has been proved by Raskutti et al. [108] for obtaining minimax rates in sparse linear regression.

Let us first prove Corollary 6.2 in the special case of hard sparsity ($q = 0$), in which θ^* is supported on a subset S of cardinality k. Let us define the model subspace $\mathcal{M} := \{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j$ so that $\theta^* \in \mathcal{M}$. Recall from Section 6.2.4 that the ℓ_1-norm is decomposable with respect to $\mathcal{M}$ and $\mathcal{M}^\perp$; as a consequence, we may also set $\overline{\mathcal{M}}^\perp = \mathcal{M}$ in the definitions (6.17) and (6.18). By definition (3.21) of the subspace compatibility between with ℓ_1-norm as the regularizer, and ℓ_2-norm as the error norm, we have $\Psi^2(\mathcal{M}) = k$. Using the settings of $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ guaranteed by Lemma 6.6 and substituting into equation (6.17), we obtain a contraction coefficient

$$\kappa(\Sigma) := \left\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\right\}\left\{1 - \chi_n(\Sigma)\right\}^{-1}, \tag{6.57}$$

where $\chi_n(\Sigma) := \frac{c_2\zeta(\Sigma)}{\sigma_{\max}(\Sigma)}\frac{k\log d}{n}$ for some universal constant c_2. A similar calculation shows that the tolerance term takes the form

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\, \phi(\Sigma; k, d, n)\left\{\frac{\|\Delta^*\|_1^2}{k} + \|\Delta^*\|_2^2\right\} \qquad \text{for some constant } c_3.$$

Since $\rho \leq \|\theta^*\|_1$, then Lemma 6.5 (as exploited in the proof of Corollary 6.1) shows that $\|\Delta^*\|_1^2 \leq 4k\|\Delta^*\|_2^2$, and hence that $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\ \chi_n(\Sigma)\ \|\Delta^*\|_2^2$. This completes the proof of the claim (6.31) for $q = 0$.

We now turn to the case $q \in (0, 1]$, for which we bound the term $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ using a slightly different choice of the subspace pair $\mathcal{M}$ and $\overline{\mathcal{M}}^\perp$. For a truncation level $\mu > 0$ to be chosen, define the set $S_\mu := \{j \in \{1, 2, \ldots, d\} \mid |\theta_j^*| > \mu\}$, and define the associated subspaces $\mathcal{M} = \mathcal{M}(S_\mu)$ and $\overline{\mathcal{M}}^\perp = \mathcal{M}^\perp(S_\mu)$. By combining Lemma 6.5 and the definition (6.18) of $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, for any pair $(\mathcal{M}(S_\mu), \mathcal{M}^\perp(S_\mu))$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^\perp) \leq \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{\log d}{n} \big(\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 + \sqrt{|S_\mu|}\, \|\Delta^*\|_2\big)^2,$$

where to simplify notation, we have omitted the dependence of $\mathcal{M}$ and $\mathcal{M}^\perp$ on S_μ. We now choose the threshold μ optimally, so as to trade-off the term $\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1$, which decreases as μ increases, with the term $\sqrt{S_\mu}\|\Delta^*\|_2$, which increases as μ increases.

By definition of $\mathcal{M}^\perp(S_\mu)$, we have

$$\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 = \sum_{j \notin S_\mu} |\theta_j^*| \ = \ \mu \sum_{j \notin S_\mu} \frac{|\theta_j^*|}{\mu} \ \leq \ \mu \sum_{j \notin S_\mu} \left(\frac{|\theta_j^*|}{\mu}\right)^q,$$

where the inequality holds since $|\theta_j^*| \leq \mu$ for all $j \notin S_\mu$. Now since $\theta^* \in \mathbb{B}_q(R_q)$, we conclude that

$$\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 \leq \mu^{1-q} \sum_{j \notin S_\mu} |\theta_j^*|^q \ \leq \mu^{1-q} R_q. \tag{6.58}$$

On the other hand, again using the inclusion $\theta^* \in \mathbb{B}_q(R_q)$, we have $R_q \geq \sum_{j \in S_\mu} |\theta_j^*|^q \ \geq |S_\mu|\, \mu^q$ which implies that $|S_\mu| \leq \mu^{-q} R_q$. By combining this bound with inequality (6.58), we obtain the upper bound

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^\perp) \leq \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{\log d}{n} \big(\mu^{2-2q} R_q^2 + \mu^{-q} R_q \|\Delta^*\|_2^2\big) \ = \ \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{\log d}{n} \mu^{-q} R_q \big(\mu^{2-q} R_q + \|\Delta^*\|_2^2\big).$$

Setting $\mu^2 = \frac{\log d}{n}$ then yields

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^\perp) \leq \varphi_n(\Sigma) \left\{ R_q \big(\frac{\log d}{n}\big)^{1-q/2} + \|\Delta^*\|_2^2 \right\}, \qquad \text{where } \varphi_n(\Sigma) := \tfrac{c\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} R_q \big(\tfrac{\log d}{n}\big)^{1-q/2}.$$

Finally, let us verify the stated form of the contraction coefficient. For the given subspace $\overline{\mathcal{M}}^\perp = \mathcal{M}(S_\mu)$ and choice of μ, we have $\Psi^2(\overline{\mathcal{M}}^\perp) = |S_\mu| \leq \mu^{-q} R_q$. From Lemma 6.6, we have

$$16\Psi^2(\overline{\mathcal{M}}^\perp) \frac{\tau_\ell(\mathcal{L}_n) + \tau_u(\mathcal{L}_n)}{\gamma_u} \leq \varphi_n(\Sigma),$$

and hence, by definition (6.17) of the contraction coefficient,

$$\kappa \leq \left\{1 - \frac{\gamma_\ell}{2\gamma_u} + \varphi_n(\Sigma)\right\}\left\{1 - \varphi_n(\Sigma)\right\}^{-1}.$$

For proving Corollary 6.3, we observe that the stated settings $\overline{\gamma_\ell}$, $\chi_n(\Sigma)$ and κ follow directly from Lemma 6.6. The bound for condition 6.2(a) follows from a standard argument about the suprema of d independent Gaussians with variance ν.

6.5.5 Proof of Corollary 6.4

This proof is analogous to that of Corollary 6.2, but appropriately adapted to the matrix setting. We first state a lemma that allows us to establish appropriate forms of the RSC/RSM conditions. Recall that we are studying an instance of matrix regression with random design, where the vectorized form $\text{vec}(X)$ of each matrix is drawn from a $N(0, \Sigma)$ distribution, where $\Sigma \in \mathbb{R}^{m^2 \times m^2}$ is some covariance matrix. In order to state this result, let us define the quantity

$$\zeta_{\text{mat}}(\Sigma) := \sup_{\|u\|_2=1,\, \|v\|_2=1} \text{var}(u^T X v), \quad \text{where } \text{vec}(X) \sim N(0, \Sigma). \tag{6.59}$$

Lemma 6.7. *Under the conditions of Corollary 6.4, there are universal positive constants* (c_0, c_1) *such that*

$$\frac{\|\mathfrak{X}_n(\Delta)\|_2^2}{n} \geq \frac{1}{2}\,\sigma_{\min}(\Sigma)\,|||\Delta|||_F^2 - c_1\zeta_{mat}(\Sigma)\frac{m}{n}\,|||\Delta|||_{\text{nuc}}^2, \qquad \textit{and} \tag{6.60a}$$

$$\frac{\|\mathfrak{X}_n(\Delta)\|_2^2}{n} \leq 2\,\sigma_{\max}(\Sigma)\,|||\Delta|||_F^2 - c_1\,\zeta_{mat}(\Sigma)\,\frac{m}{n}\,|||\Delta|||_{\text{nuc}}^2, \qquad \textit{for all } \Delta \in \mathbb{R}^{m\times m}. \tag{6.60b}$$

with probability at least $1 - \exp(-c_0\, n)$.

Given the quadratic nature of the least-squares loss, the bound (6.60a) implies that the RSC condition holds with $\gamma_\ell = \frac{1}{2}\sigma_{\min}(\Sigma)$ and $\tau_\ell(\mathcal{L}_n) = c_1\zeta_{\text{mat}}(\Sigma)\frac{m}{n}$, whereas the bound (6.60b) implies that the RSM condition holds with $\gamma_u = 2\sigma_{\max}(\Sigma)$ and $\tau_u(\mathcal{L}_n) = c_1\zeta_{\text{mat}}(\Sigma)\frac{m}{n}$.

We now prove Corollary 6.4 in the special case of exactly low rank matrices ($q = 0$), in which Θ^* has some rank $r \leq m$. Given the singular value decomposition $\Theta^* = UDV^T$, let U^r and V^r be the $m \times r$ matrices whose columns correspond to the r non-zero (left and right, respectively) singular vectors of Θ^*. As in Section 6.2.4, define the subspace of matrices

$$\mathcal{M}(U^r, V^r) := \left\{\Theta \in \mathbb{R}^{m\times m} \mid \text{col}(\Theta) \subseteq U^r \text{ and } \text{row}(\Theta) \subseteq V^r\right\}, \tag{6.61}$$

as well as the associated set $\overline{\mathcal{M}}^\perp(U^r, V^r)$. Note that $\Theta^* \in \mathcal{M}$ by construction, and moreover (as discussed in Section 6.2.4, the nuclear norm is decomposable with respect to the pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$.

By definition (3.21) of the subspace compatibility with nuclear norm as the regularizer and Frobenius norm as the error norm, we have $\Psi^2(\mathcal{M}) = r$. Using the settings of $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ guaranteed by Lemma 6.7 and substituting into equation (6.17), we obtain a contraction coefficient

$$\kappa(\Sigma) := \Big\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\Big\}\Big\{1 - \chi_n(\Sigma)\Big\}^{-1}, \tag{6.62}$$

where $\chi_n(\Sigma) := \frac{c_2 \zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{rm}{n}$ for some universal constant c_2. A similar calculation shows that the tolerance term takes the form

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\, \phi(\Sigma; r, m, n)\Big\{\frac{|||\Delta^*|||_{\mathrm{nuc}}^2}{r} + |||\Delta^*|||_F^2\Big\} \qquad \text{for some constant } c_3.$$

Since $\rho \leq |||\Theta^*|||_{\mathrm{nuc}}$ by assumption, Lemma 6.5 (as exploited in the proof of Corollary 6.1) shows that $|||\Delta^*|||_{\mathrm{nuc}}^2 \leq 4r|||\Delta^*|||_F^2$, and hence that

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\, \chi_n(\Sigma)\, |||\Delta^*|||_F^2,$$

which show the claim (6.36) for $q = 0$.

We now turn to the case $q \in (0, 1]$; as in the proof of this case for Corollary 6.2, we bound $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ using a slightly different choice of the subspace pair. Recall our notation $\sigma_1(\Theta^*) \geq \sigma_2(\Theta^*) \geq \cdots \geq \sigma_m(\Theta^*) \geq 0$ for the ordered singular values of Θ^*. For a threshold μ to be chosen, define $S_\mu = \big\{j \in \{1, 2, \ldots, d\} \mid \sigma_j(\Theta^*) > \mu\big\}$, and $U(S_\mu) \in \mathbb{R}^{m \times |S_\mu|}$ be the matrix of left singular vectors indexed by S_μ, with the matrix $V(S_\mu)$ defined similarly. We then define the subspace $\mathcal{M}(S_\mu) := \mathcal{M}(U(S_\mu), V(S_\mu))$ in an analogous fashion to equation (6.61), as well as the subspace $\overline{\mathcal{M}}^\perp(S_\mu)$.

Now by a combination of Lemma 6.5 and the definition (6.18) of $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, for any pair $(\mathcal{M}(S_\mu), \overline{\mathcal{M}}^\perp(S_\mu))$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^\perp) \leq \frac{c\, \zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{m}{n}\Big(\sum_{j \notin S_\mu} \sigma_j(\Theta^*) + \sqrt{|S_\mu|}\, |||\Delta^*|||_F\Big)^2,$$

where to simplify notation, we have omitted the dependence of $\mathcal{M}$ and $\mathcal{M}^\perp$ on S_μ. As in the proof of Corollary 6.2, we now choose the threshold μ optimally, so as to trade-off the term $\sum_{j \notin S_\mu} \sigma_j(\Theta^*)$ with its competitor $\sqrt{|S_\mu|}\, |||\Delta^*|||_F$. Exploiting the fact that $\Theta^* \in \mathbb{B}_q(R_q)$ and following the same steps as the proof of Corollary 6.2 yields the bound

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^\perp) \leq \frac{c\, \zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{m}{n}\big(\mu^{2-2q} R_q^2 + \mu^{-q} R_q |||\Delta^*|||_F^2\big).$$

Setting $\mu^2 = \frac{m}{n}$ then yields

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^\perp) \leq \varphi_n(\Sigma)\, \Big\{R_q\big(\frac{m}{n}\big)^{1-q/2} + |||\Delta^*|||_F^2\Big\},$$

as claimed. The stated form of the contraction coefficient can be verified by a calculation analogous to the proof of Corollary 6.2.

6.5.6 Proof of Corollary 6.5

In this case, we let $\mathfrak{X}_n : \mathbb{R}^{m \times m} \to \mathbb{R}^n$ be the operator defined by the model of random signed matrix sampling 5.3. As previously argued, establishing the RSM/RSC property amounts to obtaining a form of uniform control over $\frac{\|\mathfrak{X}_n(\Theta)\|_2^2}{n}$. More specifically, from the proof of Theorem 6.1, we see that it suffices to have a form of RSC for the difference $\widehat{\Delta}^t = \Theta^t - \widehat{\Theta}$, and a form of RSM for the difference $\Theta^{t+1} - \Theta^t$. The following two lemmas summarize these claims:

Lemma 6.8. *There is a constant c such that for all iterations $t = 0, 1, 2, \ldots$ and integers $r = 1, 2, \ldots, m-1$, with probability at least $1 - \exp(-m \log m)$,*

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2}\|\widehat{\Delta}^t\|_F^2 - \underbrace{c\alpha\sqrt{\frac{r\, m \log m}{n}}\left\{\frac{\sum_{j=r+1}^m \sigma_j(\Theta^*)}{\sqrt{r}} + \alpha\sqrt{\frac{rm\log m}{n}} + \|\Delta^*\|_F\right\}}_{\delta_\ell(r)}. \tag{6.63}$$

Lemma 6.9. *There is a constant c such that for all iterations $t = 0, 1, 2, \ldots$ and integers $r = 1, 2, \ldots, m-1$, with probability at least $1 - \exp(-m \log m)$, the difference $\Gamma^t := \Theta^{t+1} - \Theta^t$ satisfies the inequality $\frac{\|\mathfrak{X}_n(\Gamma^t)\|_2^2}{n} \leq 2\|\Gamma^t\|_F^2 + \delta_u(r)$, where*

$$\delta_u(r) := c\alpha\sqrt{\frac{rm\log m}{n}}\left\{\frac{\sum_{j=r+1}^m \sigma_j(\Theta^*)}{\sqrt{r}} + \alpha\sqrt{\frac{rm\log m}{n}} + \|\Delta^*\|_F + \|\widehat{\Delta}^t\|_F + \|\widehat{\Delta}^{t+1}\|_F\right\}.$$

We can now complete the proof of Corollary 6.5 by a minor modification of the proof of Theorem 6.1. Recalling the elementary relation (6.47), we have

$$\|\Theta^{t+1} - \widehat{\Theta}\|_F^2 = \|\Theta^t - \widehat{\Theta}\|_F^2 + \|\Theta^t - \Theta^{t+1}\|_F^2 - 2\langle\langle \Theta^t - \widehat{\Theta},\ \Theta^t - \Theta^{t+1}\rangle\rangle.$$

From the proof of Lemma 6.2, we see that the combination of Lemma 6.8 and 6.9 (with $\gamma_\ell = \frac{1}{2}$ and $\gamma_u = 2$) imply that

$$2\langle\langle \Theta^t - \Theta^{t+1},\ \Theta^t - \widehat{\Theta}\rangle\rangle \geq \|\Theta^t - \Theta^{t+1}\|_F^2 + \frac{1}{4}\|\Theta^t - \widehat{\Theta}\|_F^2 - \delta_u(r) - \delta_\ell(r)$$

and hence that

$$\|\widehat{\Delta}^{t+1}\|_F^2 \leq \frac{3}{4}\|\widehat{\Delta}^t\|_F^2 + \delta_\ell(r) + \delta_u(r).$$

We substitute the forms of $\delta_\ell(r)$ and $\delta_u(r)$ given in Lemmas 6.8 and 6.9 respectively; performing some algebra then yields

$$\left\{1 - \frac{c\,\alpha\sqrt{\frac{rm\log m}{n}}}{\|\widehat{\Delta}^{t+1}\|_F}\right\}\|\widehat{\Delta}^{t+1}\|_F^2 \leq \left\{\frac{3}{4} + \frac{c\alpha\sqrt{\frac{rm\log m}{n}}}{\|\widehat{\Delta}^t\|_F}\right\}\|\widehat{\Delta}^t\|_F^2 + c'\,\delta_\ell(r).$$

Consequently, as long as $\min\{\|\widehat{\Delta}^t\|_F^2, \|\widehat{\Delta}^{t+1}\|_F^2\} \geq c_3 \alpha \frac{rm \log m}{n}$ for a sufficiently large constant c_3, we are guaranteed the existence of some $\kappa \in (0,1)$ such that

$$\|\widehat{\Delta}^{t+1}\|_F^2 \leq \kappa \|\widehat{\Delta}^t\|_F^2 + c' \delta_\ell(r). \tag{6.64}$$

Since $\delta_\ell(r) = \Omega(\frac{rm \log m}{n})$, this inequality (6.64) is valid for all $t = 0, 1, 2, \ldots$ as long as c' is sufficiently large. As in the proof of Theorem 6.1, iterating the inequality (6.64) yields

$$\|\widehat{\Delta}^{t+1}\|_F^2 \leq \kappa^t \|\widehat{\Delta}^0\|_F^2 + \frac{c'}{1-\kappa} \delta_\ell(r). \tag{6.65}$$

It remains to choose the cut-off $r \in \{1, 2, \ldots, m-1\}$ so as to minimize the term $\delta_\ell(r)$. In particular, when $\Theta^* \in \mathbb{B}_q(R_q)$, then as shown in Chapter 4, the optimal choice is $r \asymp \alpha^{-q} R_q \big(\frac{n}{m \log m}\big)^{q/2}$. Substituting into the inequality (6.65) and performing some algebra yields that there is a universal constant c_4 such that the bound

$$\|\widehat{\Delta}^{t+1}\|_F^2 \leq \kappa^t \|\widehat{\Delta}^0\|_F^2 + \frac{c_4}{1-\kappa} \Big\{ R_q \big(\frac{\alpha m \log m}{n}\big)^{1-q/2} + \sqrt{R_q \big(\frac{\alpha m \log m}{n}\big)^{1-q/2}} \, \|\Delta^*\|_F \Big\}.$$

holds. Now by the Cauchy-Schwarz inequality we have

$$\sqrt{R_q \big(\frac{\alpha m \log m}{n}\big)^{1-q/2}} \, \|\Delta^*\|_F \leq \frac{1}{2} R_q \big(\frac{\alpha m \log m}{n}\big)^{1-q/2} + \frac{1}{2} \|\Delta^*\|_F^2,$$

and the claimed inequality (6.39) follows.

6.5.7 Proof of Corollary 6.6

Again the main argument in the proof would be to establish the RSM and RSC properties for the decomposition problem. We define $\widehat{\Delta}_\Theta^t = \Theta^t - \widehat{\Theta}$ and $\widehat{\Delta}_\Gamma^t = \Gamma^t - \widehat{\Gamma}$. We start with giving a lemma that establishes RSC for the differences $(\widehat{\Delta}_\Theta^t, \widehat{\Delta}_\Gamma^t)$. We recall that just like noted in the previous section, it suffices to show RSC only for these differences. Showing RSC/RSM in this example amounts to analyzing $\|\widehat{\Delta}_\Theta^t + \widehat{\Delta}_\Gamma^t\|_F^2$. We recall that this section assumes that Γ^* has only k non-zero columns.

Lemma 6.10. *There is a constant c such that for all iterations $t = 0, 1, 2, \ldots,$*

$$\|\widehat{\Delta}_\Theta^t + \widehat{\Delta}_\Gamma^t\|_F^2 \geq \frac{1}{2} \big(\|\widehat{\Delta}_\Theta^t\|_F^2 + \|\widehat{\Delta}_\Gamma^t\|_F^2 \big) - c\alpha \sqrt{\frac{k}{d_2}} \Big(\|\widehat{\Gamma} - \Gamma^*\|_F + \alpha \sqrt{\frac{k}{d_2}} \Big) \tag{6.66}$$

This proof of this lemma follows by a straightforward modification of analogous results in the paper [2].

Matrix decomposition has the interesting property that the RSC condition holds in a deterministic sense (as opposed to with high probability). The same deterministic guarantee holds for the RSM condition; indeed, we have

$$\|\widehat{\Delta}^t_\Delta + \widehat{\Delta}^t_\Gamma\|_F^2 \leq 2\big(\|\widehat{\Delta}^t_\Theta\|_F^2 + \|\widehat{\Delta}^t_\Gamma\|_F^2\big), \tag{6.67}$$

by Cauchy-Schwartz inequality. Now we appeal to the more general form of Theorem 6.1 as stated in Equation 6.43, which gives

$$\|\widehat{\Delta}^{t+1}_\Theta\|_F^2 + \|\widehat{\Delta}^{t+1}_\Gamma\|_F^2 \leq \left(\frac{3}{4}\right)^t \big(\|\widehat{\Delta}^0_\Theta\|_F^2 + \|\widehat{\Delta}^0_\Gamma\|_F^2\big) + c\sqrt{\frac{\alpha k}{d_2}}\left(\|\widehat{\Gamma} - \Gamma^*\|_F + \frac{\alpha k}{d_2}\right).$$

The stated form of the corollary follows by an application of Cauchy-Schwarz inequality.

6.6 Discussion

In this chapter, we have shown that even though high-dimensional M-estimators in statistics are neither strongly convex nor smooth, simple first-order methods can still enjoy global guarantees of geometric convergence. The key insight is that strong convexity and smoothness need only hold in restricted senses, and moreover, these conditions are satisfied with high probability for many statistical models and decomposable regularizers used in practice. Examples include sparse linear regression and ℓ_1-regularization, various statistical models with group-sparse regularization, matrix regression with nuclear norm constraints (including matrix completion and multi-task learning), and matrix decomposition problems. Overall, our results highlight some important connections between computation and statistics: the properties of M-estimators favorable for fast rates in a statistical sense can also be used to establish fast rates for optimization algorithms.

Appendix A

Proofs for Chapter 3

A.1 Proofs related to Theorem 3.1

In this section, we collect the proofs of Lemma 3.1 and our main result. All our arguments in this section are deterministic, and both proofs make use of the function $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}$ given by $\mathcal{F}(\Delta) := \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n\big\{\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*)\big\}$. In addition, we exploit the following fact: since $\mathcal{F}(0) = 0$, the optimal error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ must satisfy $\mathcal{F}(\widehat{\Delta}) \leq 0$.

A.1.1 Proof of Lemma 3.1

Note that the function $\mathcal{F}$ consists of two parts: a difference of loss functions, and a difference of regularizers. In order to control $\mathcal{F}$, we require bounds on these two quantities:

Lemma A.1 (Deviation inequalities)**.** *For any decomposable regularizer and d-dimensional vectors θ^* and Δ, we have*

$$\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \geq \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) - 2\mathcal{R}(\theta^*_{\mathcal{M}^\perp}). \tag{A.1}$$

Moreover, as long as $\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\theta^*))$ and $\mathcal{L}$ is convex, we have

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq -\frac{\lambda_n}{2}\big[\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp})\big]. \tag{A.2}$$

Proof. Since $\mathcal{R}\big(\theta^* + \Delta\big) = \mathcal{R}\big(\theta^*_{\mathcal{M}} + \theta^*_{\mathcal{M}^\perp} + \Delta_{\bar{\mathcal{M}}} + \Delta_{\bar{\mathcal{M}}^\perp}\big)$, triangle inequality implies that

$$\mathcal{R}\big(\theta^* + \Delta\big) \geq \mathcal{R}\big(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}\big) - \mathcal{R}\big(\theta^*_{\mathcal{M}^\perp} + \Delta_{\bar{\mathcal{M}}}\big) \;\geq\; \mathcal{R}\big(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}\big) - \mathcal{R}\big(\theta^*_{\mathcal{M}^\perp}\big) - \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}}\big).$$

By decomposability applied to $\theta^*_{\mathcal{M}}$ and $\Delta_{\bar{\mathcal{M}}^\perp}$, we have $\mathcal{R}\big(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}\big) = \mathcal{R}\big(\theta^*_{\mathcal{M}}\big) + \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}^\perp}\big)$, so that

$$\mathcal{R}\big(\theta^* + \Delta\big) \geq \mathcal{R}\big(\theta^*_{\mathcal{M}}\big) + \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}^\perp}\big) - \mathcal{R}\big(\theta^*_{\mathcal{M}^\perp}\big) - \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}}\big). \tag{A.3}$$

Similarly, by triangle inequality, we have $\mathcal{R}(\theta^*) \leq \mathcal{R}\big(\theta^*_{\bar{\mathcal{M}}}\big) + \mathcal{R}\big(\theta^*_{\bar{\mathcal{M}}^\perp}\big)$. Combining this inequality with the bound (A.3), we obtain

$$\begin{aligned}\mathcal{R}\big(\theta^* + \Delta\big) - \mathcal{R}(\theta^*) &\geq \mathcal{R}\big(\theta^*_{\bar{\mathcal{M}}}\big) + \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}^\perp}\big) - \mathcal{R}\big(\theta^*_{\bar{\mathcal{M}}^\perp}\big) - \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}}\big) - \big\{\mathcal{R}\big(\theta^*_{\bar{\mathcal{M}}}\big) + \mathcal{R}\big(\theta^*_{\bar{\mathcal{M}}^\perp}\big)\big\} \\ &= \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}^\perp}\big) - \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}}\big) - 2\mathcal{R}\big(\theta^*_{\bar{\mathcal{M}}^\perp}\big),\end{aligned}$$

which yields the claim (A.1).

Turning to the loss difference, using the convexity of the loss function $\mathcal{L}$, we have

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle \;\geq\; -|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle|.$$

Applying the (generalized) Cauchy-Schwarz inequality with the regularizer and its dual, we obtain

$$|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \leq \mathcal{R}^*(\nabla \mathcal{L}(\theta^*))\, \mathcal{R}(\Delta) \;\leq\; \frac{\lambda_n}{2}\, \big[\mathcal{R}\big(\Delta_{\bar{\mathcal{M}}}\big) + \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}^\perp}\big)\big],$$

where the final equality uses triangle inequality, and the assumed bound $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\theta^*))$. Consequently, we conclude that $\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq -\frac{\lambda_n}{2}\, \big[\mathcal{R}\big(\Delta_{\bar{\mathcal{M}}}\big) + \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}^\perp}\big)\big]$, as claimed. □

We can now complete the proof of Lemma 3.1. Combining the two lower bounds (A.1) and (A.2), we obtain

$$\begin{aligned}0 \;\geq\; \mathcal{F}(\widehat{\Delta}) &\geq \lambda_n \big\{\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) - 2\mathcal{R}(\theta^*_{\bar{\mathcal{M}}^\perp})\big\} - \frac{\lambda_n}{2}\, \big[\mathcal{R}\big(\Delta_{\bar{\mathcal{M}}}\big) + \mathcal{R}\big(\Delta_{\bar{\mathcal{M}}^\perp}\big)\big] \\ &= \frac{\lambda_n}{2}\big\{\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - 3\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) - 4\mathcal{R}(\theta^*_{\bar{\mathcal{M}}^\perp})\big\},\end{aligned}$$

from which the claim follows.

A.1.2 Proof of Theorem 3.1

Recall the set $\mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^\perp; \theta^*)$ from equation (3.17). Since the subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ and true parameter θ^* remain fixed throughout this proof, we adopt the shorthand notation $\mathbb{C}$. Letting $\epsilon > 0$ be a given error radius, the following lemma shows that it suffices to control the sign of the function $\mathcal{F}$ over the set $\mathbb{S}(\epsilon) := \mathbb{C} \cap \{\|\Delta\| = \epsilon\}$.

Lemma A.2. *If $\mathcal{F}(\Delta) > 0$ for all vectors $\Delta \in \mathbb{S}(\epsilon)$, then $\|\widehat{\Delta}\| \leq \epsilon$.*

Proof. We first claim that $\mathbb{C}$ is star-shaped, meaning that if $\widehat{\Delta} \in \mathbb{C}$, then the entire line $\{t\widehat{\Delta} \mid t \in (0,1)\}$ connecting $\widehat{\Delta}$ with the all-zeroes vector is contained with $\mathbb{C}$. This property is immediate whenever $\theta^* \in \mathcal{M}$, since $\mathbb{C}$ is then a cone, as illustrated in Figure 3.1(a). Now consider the general case, when $\theta^* \notin \mathcal{M}$. We first observe that for any $t \in (0,1)$,

$$\Pi_{\bar{\mathcal{M}}}(t\Delta) = \arg\min_{\gamma \in \bar{\mathcal{M}}} \|t\Delta - \gamma\| \;=\; t\, \arg\min_{\gamma \in \bar{\mathcal{M}}} \|\Delta - \frac{\gamma}{t}\| \;=\; t\, \Pi_{\bar{\mathcal{M}}}(\Delta),$$

using the fact that γ/t also belongs to the subspace $\overline{\mathcal{M}}$. The equality $\Pi_{\bar{\mathcal{M}}^\perp}(t\Delta) = t\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)$ follows similarly. Consequently, for all $\Delta \in \mathbb{C}$, we have

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(t\Delta)) \;=\; \mathcal{R}(t\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)) \stackrel{(i)}{=} t\,\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)) \stackrel{(ii)}{\leq} t\,\big\{3\,\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))\big\}$$

where step (i) uses the fact that any norm is positive homogeneous,[1] and step (ii) uses the inclusion $\Delta \in \mathbb{C}$. We now observe that $3\,t\,\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta)) = 3\,\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(t\Delta))$, and moreover, since $t \in (0,1)$, we have $4t\,\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \leq 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))$. Putting together the pieces, we find that

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(t\Delta)) \leq\; 3\,\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(t\Delta)) + t\,4\Pi_{\mathcal{M}^\perp}(\theta^*) \;\leq\; 3\,\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(t\Delta)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),$$

showing that $t\Delta \in \mathbb{C}$ for all $t \in (0,1)$, and hence that $\mathbb{C}$ is star-shaped.

Turning to the lemma itself, we prove the contrapositive statement: in particular, we show that if for some optimal solution $\widehat{\theta}$, the associated error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ satisfies the inequality $\|\widehat{\Delta}\| > \epsilon$, then there must be some vector $\widetilde{\Delta} \in \mathbb{S}(\epsilon)$ such that $\mathcal{F}(\widetilde{\Delta}) \leq 0$. If $\|\widehat{\Delta}\| > \epsilon$, then the line joining $\widehat{\Delta}$ to 0 must intersect the set $\mathbb{S}(\epsilon)$ at some intermediate point $t^*\widehat{\Delta}$, for some $t^* \in (0,1)$. Since the loss function $\mathcal{L}$ and regularizer $\mathcal{R}$ are convex, the function $\mathcal{F}$ is also convex for any choice of the regularization parameter, so that by Jensen's inequality,

$$\mathcal{F}\big(t^*\Delta\big) \;=\; \mathcal{F}\big(t^*\Delta + (1-t^*)\,0\big) \leq t^*\,\mathcal{F}(\widehat{\Delta}) + (1-t^*)\mathcal{F}(0) \stackrel{(i)}{=} t^*\mathcal{F}(\widehat{\Delta}),$$

where equality (i) uses the fact that $\mathcal{F}(0) = 0$ by construction. But since $\widehat{\Delta}$ is optimal, we must have $\mathcal{F}(\widehat{\Delta}) \leq 0$, and hence $\mathcal{F}(t^*\Delta) \leq 0$ as well. Thus, we have constructed a vector $\widetilde{\Delta} = t^*\Delta$ with the claimed properties, thereby establishing Lemma A.2. □

On the basis of Lemma A.2, the proof of Theorem 3.1 will be complete if we can establish a lower bound on $\mathcal{F}(\Delta)$ over $\mathbb{S}(\epsilon)$ for an appropriately chosen radius $\epsilon > 0$. For an arbitrary $\Delta \in \mathbb{S}(\epsilon)$, we have

$$\begin{aligned}
\mathcal{F}(\Delta) &= \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n\big\{\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*)\big\} \\
&\stackrel{(i)}{\geq} \langle \nabla\mathcal{L}(\theta^*),\, \Delta\rangle + \kappa_{\mathcal{L}}\|\Delta\|^2 - \tau^2_{\mathcal{L}}(\theta^*) + \lambda_n\big\{\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*)\big\} \\
&\stackrel{(ii)}{\geq} \langle \nabla\mathcal{L}(\theta^*),\, \Delta\rangle + \kappa_{\mathcal{L}}\|\Delta\|^2 - \tau^2_{\mathcal{L}}(\theta^*) + \lambda_n\big\{\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) - 2\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\big\},
\end{aligned}$$

where inequality (i) follows from the RSC condition, and inequality (ii) follows from the bound (A.1).

[1] Explicitly, for any norm and non-negative scalar t, we have $\|tx\| = t\|x\|$.

By the Cauchy-Schwarz inequality applied to the regularizer $\mathcal{R}$ and its dual $\mathcal{R}^*$, we have $|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \leq \mathcal{R}^*(\nabla \mathcal{L}(\theta^*))\, \mathcal{R}(\Delta)$. Since $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\theta^*))$ by assumption, we conclude that $|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \leq \frac{\lambda_n}{2}\mathcal{R}(\Delta)$, and hence that

$$\mathcal{F}(\Delta) \geq \kappa_{\mathcal{L}}\|\Delta\|^2 - \tau^2_{\mathcal{L}}(\theta^*) + \lambda_n\big\{\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) - 2\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\big\} - \frac{\lambda_n}{2}\mathcal{R}(\Delta)$$

By triangle inequality, we have $\mathcal{R}(\Delta) = \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp} + \Delta_{\bar{\mathcal{M}}}) \leq \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}})$, and hence, following some algebra

$$\begin{aligned}\mathcal{F}(\Delta) &\geq \kappa_{\mathcal{L}}\|\Delta\|^2 - \tau^2_{\mathcal{L}}(\theta^*) + \lambda_n\big\{\frac{1}{2}\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - \frac{3}{2}\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) - 2\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\big\} \\ &\geq \kappa_{\mathcal{L}}\|\Delta\|^2 - \tau^2_{\mathcal{L}}(\theta^*) - \frac{\lambda_n}{2}\big\{3\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\big\}. \qquad \text{(A.4)}\end{aligned}$$

Now by definition (3.21) of the subspace compatibility, we have the inequality $\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) \leq \Psi(\overline{\mathcal{M}})\|\Delta_{\bar{\mathcal{M}}}\|$. Since the projection $\Delta_{\bar{\mathcal{M}}} = \Pi_{\bar{\mathcal{M}}}(\Delta)$ is defined in terms of the norm $\|\cdot\|$, it is non-expansive. Since $0 \in \overline{\mathcal{M}}$, we have

$$\|\Delta_{\bar{\mathcal{M}}}\| = \|\Pi_{\bar{\mathcal{M}}}(\Delta) - \Pi_{\bar{\mathcal{M}}}(0)\| \overset{(i)}{\leq} \|\Delta - 0\| = \|\Delta\|,$$

where inequality (i) uses non-expansivity of the projection. Combining with the earlier bound, we conclude that $\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) \leq \Psi(\overline{\mathcal{M}})\|\Delta\|$. Substituting into the lower bound (A.4), we obtain $\mathcal{F}(\Delta) \geq \kappa_{\mathcal{L}}\|\Delta\|^2 - \tau^2_{\mathcal{L}}(\theta^*) - \frac{\lambda_n}{2}\big\{3\Psi(\overline{\mathcal{M}})\,\|\Delta\| + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\big\}$. The right-hand side of this inequality is a strictly positive definite quadratic form in $\|\Delta\|$, and so will be positive for $\|\Delta\|$ sufficiently large. In particular, some algebra shows that this is the case as long as

$$\|\Delta\|^2 \geq \epsilon^2 := 9\frac{\lambda_n^2}{\kappa^2_{\mathcal{L}}}\,\Psi^2(\overline{\mathcal{M}}) + \frac{\lambda_n}{\kappa_{\mathcal{L}}}\big\{2\tau^2_{\mathcal{L}}(\theta^*) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\big\},$$

thereby completing the proof of Theorem 3.1.

A.2 Proof of Lemma 3.2

For any Δ in the set $\mathbb{C}(S_\mu)$, we have

$$\|\Delta\|_1 \leq 4\|\Delta_{S_\mu}\|_1 + 4\|\theta^*_{S^c_\mu}\|_1 \leq \sqrt{|S_\mu|}\|\Delta\|_2 + 4R_q\,\mu^{1-q} \leq 4\sqrt{R_q}\,\mu^{-q/2}\,\|\Delta\|_2 + 4R_q\,\mu^{1-q},$$

where we have used the bounds (3.39) and (3.40). Therefore, for any vector $\Delta \in \mathbb{C}(S_\mu)$, the condition (3.31) implies that

$$\begin{aligned}\frac{\|X\Delta\|_2}{\sqrt{n}} &\geq \kappa_1\,\|\Delta\|_2 - \kappa_2\sqrt{\frac{\log d}{n}}\big\{\sqrt{R_q}\,\mu^{-q/2}\,\|\Delta\|_2 + R_q\,\mu^{1-q}\big\} \\ &= \|\Delta\|_2\Big\{\kappa_1 - \kappa_2\sqrt{\frac{R_q\log d}{n}}\mu^{-q/2}\Big\} - \kappa_2\sqrt{\frac{\log d}{n}}\,R_q\,\mu^{1-q}.\end{aligned}$$

By our choices $\mu = \frac{\lambda_n}{\kappa_1}$ and $\lambda_n = 4\,\sigma\sqrt{\frac{\log d}{n}}$, we have $\kappa_2\sqrt{\frac{R_q \log d}{n}}\,\mu^{-q/2} = \frac{\kappa_2}{(8\sigma)^{q/2}}\,\sqrt{R_q}\,\big(\frac{\log d}{n}\big)^{1-\frac{q}{2}}$, which is less than $\kappa_1/2$ under the stated assumptions. Thus, we obtain the lower bound $\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{\kappa_1}{2}\|\Delta\|_2 - 2\kappa_2\,\sqrt{\frac{\log d}{n}}\,R_q\,\mu^{1-q}$, as claimed.

A.3 Proofs for group-sparse norms

In this section, we collect the proofs of results related to the group-sparse norms in Section 3.5.

A.3.1 Proof of Proposition 3.1

The proof of this result follows similar lines to the proof of condition (3.31) given by Raskutti et al. [108], hereafter RWY, who established this result in the special case of the ℓ_1-norm. Furthermore, the result can be viewed as a specific instance of the general Gaussian operator result presented in Appendix D.4 based on the particular choice of regularizer. Here we describe only those portions of the proof that require modification. For a radius $t > 0$, define the set

$$V(t) := \big\{\theta \in \mathbb{R}^d \;\mid\; \|\Sigma^{1/2}\theta\|_2 = 1, \|\theta\|_{\mathcal{G},\alpha} \leq t\big\},$$

as well as the random variable $M(t; X) := 1 - \inf_{\theta \in V(t)} \frac{\|X\theta\|_2}{\sqrt{n}}$. The argument in Section 4.2 of RWY makes use of the Gordon-Slepian comparison inequality in order to upper bound this quantity. Following the same steps, we obtain the modified upper bound

$$\mathbb{E}[M(t; X)] \leq \frac{1}{4} + \frac{1}{\sqrt{n}}\,\mathbb{E}\big[\max_{j=1,\ldots,N_{\mathcal{G}}} \|w_{G_j}\|_{\alpha^*}\big]\; t,$$

where $w \sim N(0, \Sigma)$. The argument in Section 4.3 uses concentration of measure to show that this same bound will hold with high probability for $M(t; X)$ itself; the same reasoning applies here. Finally, the argument in Section 4.4 of RWY uses a peeling argument to make the bound suitably uniform over choices of the radius t. This argument allows us to conclude that

$$\inf_{\theta \in \mathbb{R}^d} \frac{\|X\theta\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2}\theta\|_2 - 9\,\mathbb{E}\big[\max_{j=1,\ldots,N_{\mathcal{G}}} \|w_{G_j}\|_{\alpha^*}\big]\;\|\theta\|_{\mathcal{G},\alpha} \quad \text{for all } \theta \in \mathbb{R}^d$$

with probability greater than $1 - c_1\exp(-c_2 n)$. Recalling the definition of $\rho_{\mathcal{G}}(\alpha^*)$, we see that in the case $\Sigma = I_{d\times d}$, the claim holds with constants $(\kappa_1, \kappa_2) = (\frac{1}{4}, 9)$. Turning to the case of general Σ, let us define the matrix norm $|\!|\!|A|\!|\!|_{\alpha^*} := \max_{\|\beta\|_{\alpha^*}=1} \|A\beta\|_{\alpha^*}$. With this notation, some algebra shows that the claim holds with $\kappa_1 = \frac{1}{4}\sigma_{\min}(\Sigma^{1/2})$ and $\kappa_2 = 9 \max_{t=1,\ldots,N_{\mathcal{G}}} |\!|\!|(\Sigma^{1/2})_{G_t}|\!|\!|_{\alpha^*}$.

A.3.2 Proof of Corollary 3.4

In order to prove this claim, we need to verify that Theorem 3.1 may be applied. Doing so requires defining the appropriate model and perturbation subspaces, computing the compatibility constant, and checking that the specified choice (3.48) of regularization parameter λ_n is valid. For a given subset $S_{\mathcal{G}} \subseteq \{1, 2, \ldots, N_{\mathcal{G}}\}$, define the subspaces

$$\mathcal{M}(S_{\mathcal{G}}) := \big\{\theta \in \mathbb{R}^d \mid \theta_{G_t} = 0 \quad \text{for all } t \notin S_{\mathcal{G}}\big\}, \quad \text{and} \quad \mathcal{M}^{\perp}(S_{\mathcal{G}}) := \big\{\theta \in \mathbb{R}^d \mid \theta_{G_t} = 0 \quad \text{for all } t \in S_{\mathcal{G}}\big\}.$$

As discussed in Example 3.2, the block norm $\|\cdot\|_{\mathcal{G},\alpha}$ is decomposable with respect to these subspaces. Let us compute the regularizer-error compatibility function, as defined in equation (3.21), that relates the regularizer ($\|\cdot\|_{\mathcal{G},\alpha}$ in this case) to the error norm (here the ℓ_2-norm). For any $\Delta \in \mathcal{M}(S_{\mathcal{G}})$, we have

$$\|\Delta\|_{\mathcal{G},\alpha} = \sum_{t \in S_{\mathcal{G}}} \|\Delta_{G_t}\|_{\alpha} \overset{(a)}{\leq} \sum_{t \in S_{\mathcal{G}}} \|\Delta_{G_t}\|_2 \leq \sqrt{k}\, \|\Delta\|_2,$$

where inequality (a) uses the fact that $\alpha \geq 2$.

Finally, let us check that the specified choice of λ_n satisfies the condition (3.23). As in the proof of Corollary 3.2, we have $\nabla \mathcal{L}(\theta^*; Z_1^n) = \frac{1}{n} X^T w$, so that the final step is to compute an upper bound on the quantity $\mathcal{R}^*(\frac{1}{n} X^T w) = \frac{1}{n} \max_{t=1,\ldots,N_{\mathcal{G}}} \|(X^T w)_{G_t}\|_{\alpha^*}$ that holds with high probability.

Lemma A.3. *Suppose that X satisfies the block column normalization condition (3.47), and the observation noise is sub-Gaussian (3.33). Then we have*

$$\mathbb{P}\Big[\max_{t=1,\ldots,N_{\mathcal{G}}} \|\frac{X_{G_t}^T w}{n}\|_{\alpha^*} \geq 2\sigma\big\{\frac{m^{1-1/\alpha}}{\sqrt{n}} + \sqrt{\frac{\log N_{\mathcal{G}}}{n}}\big\}\Big] \leq 2\exp\big(-2\log N_{\mathcal{G}}\big). \tag{A.5}$$

Proof. Throughout the proof, we assume without loss of generality that $\sigma = 1$, since the general result can be obtained by rescaling. For a fixed group G of size m, consider the submatrix $X_G \in \mathbb{R}^{n \times m}$. We begin by establishing a tail bound for the random variable $\|\frac{X_G^T w}{n}\|_{\alpha^*}$.

Deviations above the mean: For any pair $w, w' \in \mathbb{R}^n$, we have

$$\Big|\|\frac{X_G^T w}{n}\|_{\alpha^*} - \|\frac{X_G^T w'}{n}\|_{\alpha^*}\Big| \leq \frac{1}{n}\|X_G^T(w - w')\|_{\alpha^*} = \frac{1}{n} \max_{\|\theta\|_{\alpha}=1} \langle X_G\, \theta,\, (w - w')\rangle.$$

By definition of the $(\alpha \to 2)$ operator norm, we have

$$\frac{1}{n}\|X_G^T(w - w')\|_{\alpha^*} \leq \frac{1}{n}\,|||X_G|||_{\alpha \to 2}\, \|w - w'\|_2 \overset{(i)}{\leq} \frac{1}{\sqrt{n}}\|w - w'\|_2,$$

where inequality (i) uses the block normalization condition (3.47). We conclude that the function $w \mapsto \|\frac{X_G^T w}{n}\|_{\alpha^*}$ is a Lipschitz with constant $1/\sqrt{n}$, so that by Gaussian concentration of measure for Lipschitz functions (2.18), we have

$$\mathbb{P}\Big[\|\frac{X_G^T w}{n}\|_{\alpha^*} \geq \mathbb{E}\Big[\|\frac{X_G^T w}{n}\|_{\alpha^*}\Big] + \delta\Big] \leq 2\exp\big(-\frac{n\delta^2}{2}\big) \qquad \text{for all } \delta > 0. \tag{A.6}$$

Upper bounding the mean: For any vector $\beta \in \mathbb{R}^m$, define the zero-mean Gaussian random variable $Z_\beta = \frac{1}{n}\langle \beta, X_G^T w\rangle$, and note the relation $\|\frac{X_G^T w}{n}\|_{\alpha^*} = \max_{\|\beta\|_\alpha = 1} Z_\beta$. Thus, the quantity of interest is the supremum of a Gaussian process, and can be upper bounded using Gaussian comparison principles. For any two vectors $\|\beta\|_\alpha \leq 1$ and $\|\beta'\|_\alpha \leq 1$, we have

$$\mathbb{E}\Big[(Z_\beta - Z_{\beta'})^2\Big] \;=\; \frac{1}{n^2}\|X_G(\beta - \beta')\|_2^2 \overset{(a)}{\leq} \frac{2}{n}\,\frac{|||X_G|||_{\alpha\to 2}^2}{n}\,\|\beta - \beta'\|_2^2 \overset{(b)}{\leq} \frac{2}{n}\|\beta - \beta'\|_2^2,$$

where inequality (a) uses the fact that $\|\beta - \beta'\|_\alpha \leq \sqrt{2}$, and inequality (b) uses the block normalization condition (3.47).

Now define a second Gaussian process $Y_\beta = \sqrt{\frac{2}{n}}\,\langle \beta, \varepsilon\rangle$, where $\varepsilon \sim N(0, I_{m\times m})$ is standard Gaussian. By construction, for any pair $\beta, \beta' \in \mathbb{R}^m$, we have $\mathbb{E}\big[(Y_\beta - Y_{\beta'})^2\big] = \frac{2}{n}\|\beta - \beta'\|_2^2 \geq \mathbb{E}[(Z_\beta - Z_{\beta'})^2]$, so that the Sudakov-Fernique comparison principle [78] implies that

$$\mathbb{E}\Big[\|\frac{X_G^T w}{n}\|_{\alpha^*}\Big] = \mathbb{E}\Big[\max_{\|\beta\|_\alpha = 1} Z_\beta\Big] \;\leq\; \mathbb{E}\Big[\max_{\|\beta\|_\alpha = 1} Y_\beta\Big].$$

By definition of Y_β, we have

$$\mathbb{E}\Big[\max_{\|\beta\|_\alpha = 1} Y_\beta\Big] \;=\; \sqrt{\frac{2}{n}}\mathbb{E}\big[\|\varepsilon\|_{\alpha^*}\big] = \sqrt{\frac{2}{n}}\mathbb{E}\Big[\big(\sum_{j=1}^m |\varepsilon_j|^{\alpha^*}\big)^{1/\alpha^*}\Big] \;\leq\; \sqrt{\frac{2}{n}}\, m^{1/\alpha^*}\big(\mathbb{E}[|\varepsilon_1|^{\alpha^*}]\big)^{1/\alpha^*},$$

using Jensen's inequality, and the concavity of the function $f(t) = t^{1/\alpha^*}$ for $\alpha^* \in [1, 2]$. Finally, we have $\big(\mathbb{E}[|\varepsilon_1|^{\alpha^*}]\big)^{1/\alpha^*} \leq \sqrt{\mathbb{E}[\varepsilon_1^2]} = 1$ and $1/\alpha^* = 1 - 1/\alpha$, so that we have shown that $\mathbb{E}\Big[\max_{\|\beta\|_\alpha = 1} Y_\beta\Big] \leq 2\frac{m^{1-1/\alpha}}{\sqrt{n}}$. Combining this bound with the concentration statement (A.6), we obtain $\mathbb{P}\Big[\|\frac{X_G^T w}{n}\|_{\alpha^*} \geq 2\,\frac{m^{1-1/\alpha}}{\sqrt{n}} + \delta\Big] \leq 2\exp\big(-\frac{n\delta^2}{2}\big)$. We now apply the union bound over all groups, and set $\delta^2 = \frac{4\log N_{\mathcal{G}}}{n}$ to conclude that

$$\mathbb{P}\Big[\max_{t=1,\ldots,N_{\mathcal{G}}} \|\frac{X_{G_t}^T w}{n}\|_{\alpha^*} \geq 2\Big\{\frac{m^{1-1/\alpha}}{\sqrt{n}} + \sqrt{\frac{\log N_{\mathcal{G}}}{n}}\Big\}\Big] \leq 2\exp\big(-2\log N_{\mathcal{G}}\big),$$

as claimed. □

Appendix B

Proofs for Chapter 4

B.1 Proof of Lemma 4.1

Part (a) of the claim was proved in Recht et al. [117]; we simply provide a proof here for completeness. We write the SVD as $\Theta^* = UDV^T$, where $U \in \mathbb{R}^{d_1 \times d_1}$ and $V \in \mathbb{R}^{d_2 \times d_2}$ are orthogonal matrices, and D is the matrix formed by the singular values of Θ^*. Note that the matrices U^r and V^r are given by the first r columns of U and V respectively. We then define the matrix $\Gamma = U^T \Delta V \in \mathbb{R}^{d_1 \times d_2}$, and write it in block form as

$$\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}, \qquad \text{where } \Gamma_{11} \in \mathbb{R}^{r \times r}, \text{ and } \Gamma_{22} \in \mathbb{R}^{(d_1 - r) \times (d_2 - r)}.$$

We now define the matrices

$$\Delta'' = U \begin{bmatrix} 0 & 0 \\ 0 & \Gamma_{22} \end{bmatrix} V^T, \qquad \text{and } \Delta' = \Delta - \Delta''.$$

Note that we have

$$\operatorname{rank}(\Delta') = \operatorname{rank} \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & 0 \end{bmatrix} \leq \operatorname{rank} \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ 0 & 0 \end{bmatrix} + \operatorname{rank} \begin{bmatrix} \Gamma_{11} & 0 \\ \Gamma_{21} & 0 \end{bmatrix} \leq 2r,$$

which establishes Lemma 4.1(a). Moreover, we note for future reference that by construction of Δ'', the nuclear norm satisfies the decomposition

$$|||\Pi_{\mathcal{M}}(\Theta^*) + \Delta''|||_{\text{nuc}} = |||\Pi_{\mathcal{M}}(\Theta^*)|||_{\text{nuc}} + |||\Delta''|||_{\text{nuc}}. \tag{B.1}$$

We now turn to the proof of Lemma 4.1(b). Recall that the error $\Delta = \widehat{\Theta} - \Theta^*$ associated with any optimal solution must satisfy the inequality (4.30), which implies that

$$0 \leq \frac{1}{N} \langle \vec{\varepsilon}, \mathfrak{X}(\Delta) \rangle + \lambda_N \big\{ |||\Theta^*|||_{\text{nuc}} - |||\widehat{\Theta}|||_{\text{nuc}} \big\} \leq |||\frac{1}{N} \mathfrak{X}^*(\vec{\varepsilon})|||_2 \, |||\Delta|||_{\text{nuc}} + \lambda_N \big\{ |||\Theta^*|||_{\text{nuc}} - |||\widehat{\Theta}|||_{\text{nuc}} \big\}, \tag{B.2}$$

where we have used the bound (4.31).

Note that we have the decomposition $\Theta^* = \Pi_{\mathcal{M}}(\Theta^*) + \Pi_{\mathcal{M}}^{\perp}(\Theta^*)$. Using this decomposition, the triangle inequality and the relation (B.1), we have

$$\begin{aligned}
|||\widehat{\Theta}|||_{\text{nuc}} &= |||(\Pi_{\mathcal{M}}(\Theta^*) + \Delta'') + (\Pi_{\mathcal{M}}^{\perp}(\Theta^*) + \Delta')|||_{\text{nuc}} \\
&\geq |||(\Pi_{\mathcal{M}}(\Theta^*) + \Delta'')|||_{\text{nuc}} - |||(\Pi_{\mathcal{M}}^{\perp}(\Theta^*) + \Delta')|||_{\text{nuc}} \\
&\geq |||\Pi_{\mathcal{M}}(\Theta^*)|||_{\text{nuc}} + |||\Delta''|||_{\text{nuc}} - \big\{|||(\Pi_{\mathcal{M}}^{\perp}(\Theta^*)|||_{\text{nuc}} + |||\Delta'|||_{\text{nuc}}\big\}.
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
|||\Theta^*|||_{\text{nuc}} - |||\widehat{\Theta}|||_{\text{nuc}} &\leq |||\Theta^*|||_{\text{nuc}} - \big\{|||\Pi_{\mathcal{M}}(\Theta^*)|||_{\text{nuc}} + |||\Delta''|||_{\text{nuc}}\big\} + \big\{|||(\Pi_{\mathcal{M}}^{\perp}(\Theta^*)|||_{\text{nuc}} + |||\Delta'|||_{\text{nuc}}\big\} \\
&= 2|||\Pi_{\mathcal{M}}^{\perp}(\Theta^*)|||_{\text{nuc}} + |||\Delta'|||_{\text{nuc}} - |||\Delta''|||_{\text{nuc}}.
\end{aligned}$$

Substituting this inequality into the bound (B.2), we obtain

$$0 \leq |||\frac{1}{N}\mathfrak{X}^*(\vec{\varepsilon})|||_2 \, |||\Delta|||_{\text{nuc}} + \lambda_N\big\{2|||\Pi_{\mathcal{M}}^{\perp}(\Theta^*)|||_{\text{nuc}} + |||\Delta'|||_{\text{nuc}} - |||\Delta''|||_{\text{nuc}}\big\}.$$

Finally, since $|||\frac{1}{N}\mathfrak{X}^*(\vec{\varepsilon})|||_2 \leq \lambda_N/2$ by assumption, we conclude that

$$0 \leq \lambda_N\big\{2|||\Pi_{\mathcal{M}}^{\perp}(\Theta^*)|||_{\text{nuc}} + \frac{3}{2}|||\Delta'|||_{\text{nuc}} - \frac{1}{2}|||\Delta''|||_{\text{nuc}}\big\}.$$

Since $|||\Pi_{\mathcal{M}}^{\perp}(\Theta^*)|||_{\text{nuc}} = \sum_{j=r+1}^{m}\sigma_j(\Theta^*)$, the bound (4.32) follows.

B.2 Consistency in operator norm

In this appendix, we derive a bound on the operator norm error for both the low-rank multivariate regression and auto-regressive model estimation problems. In this statement, it is convenient to specify these models in the form $Y = X\Theta^* + W$, where $Y \in \mathbb{R}^{n\times d_2}$ is a matrix of observations.

Proposition B.1 (Operator norm consistency)**.** *Consider the multivariate regression problem and the SDP under the conditions of Corollary 4.3. Then any solution $\widehat{\Theta}$ to the SDP satisfies the bound*

$$|||\widehat{\Theta} - \theta^*|||_2 \leq c' \, \frac{\nu\,\sqrt{\sigma_{\max}(\Sigma)}}{\sigma_{\min}(\Sigma)}\sqrt{\frac{d_1+d_2}{n}}. \tag{B.3}$$

We note that a similar bound applies to the auto-regressive model treated in Corollary 4.4.

Proof. For any subgradient matrix $Z \in \partial |||\widehat{\Theta}|||_{\text{nuc}}$, we are guaranteed $|||Z|||_2 \leq 1$. Furthermore, by the KKT conditions [21] for the nuclear norm SDP, any solution $\widehat{\Theta}$ must satisfy the condition

$$\frac{1}{n} X^T X \widehat{\Theta} - \frac{X^T Y}{n} + \lambda_n Z \;=\; 0.$$

Hence, simple algebra and the triangle inequality yield that

$$|||\widehat{\Theta}|||_2 \;\leq\; |||\big(\frac{1}{n} X^T X\big)^{-1}|||_2 \Big[|||X^T W/n|||_2 + \lambda_n \Big].$$

Lemma 4.2 yields that $|||\big(\frac{1}{n} X^T X\big)^{-1}|||_2 \leq \frac{9}{\sigma_{\min}(\Sigma)}$ with high probability. Combining these inequalities yields

$$|||\widehat{\Theta}|||_2 \;\leq\; c_1 \frac{\lambda_n}{\sigma_{\min}(\Sigma)}.$$

We require that $\lambda_n \geq 2|||X^T W|||_2/n$. From Lemma 4.3, it suffices to set $\lambda_n \geq c_0 \sqrt{\sigma_{\max}(\Sigma)} \nu \sqrt{\frac{d_1+d_2}{n}}$. Combining the pieces yields the claim.

□

B.3 Proof of Lemma 4.3

Let $S^{m-1} = \{u \in \mathbb{R}^m \mid \|u\|_2 = 1\}$ denote the Euclidean sphere in m-dimensions. The operator norm of interest has the variational representation

$$\frac{1}{n} |||X^T W|||_2 = \frac{1}{n} \sup_{u \in S^{d_1-1}} \sup_{v \in S^{d_2-1}} v^T X^T W u$$

For positive scalars a and b, define the (random) quantity

$$\Psi(a,b) := \sup_{u \in a\, S^{d_1-1}} \sup_{v \in b\, S^{d_2-1}} \langle Xv, Wu \rangle.$$

and note that our goal is to upper bound $\Psi(1,1)$. Note moreover that $\Psi(a,b) = a\, b\, \Psi(1,1)$, a relation which will be useful in the analysis.

Let $\mathcal{A} = \{u^1, \ldots, u^A\}$ and $\mathcal{B} = \{v^1, \ldots, v^B\}$ denote $1/4$ coverings of S^{d_1-1} and S^{d_2-1}, respectively. We now claim that we have the upper bound

$$\Psi(1,1) \;\leq\; 4 \max_{u^a \in \mathcal{A}, v^b \in \mathcal{B}} \langle Xv^b, Wu^a \rangle \tag{B.4}$$

To establish this claim, we note that since the sets $\mathcal{A}$ and $\mathcal{B}$ are $1/4$-covers, for any pair $(u,v) \in S^{m-1} \times S^{m-1}$, there exists a pair $(u^a, v^b) \in \mathcal{A} \times \mathcal{B}$ such that $u = u^a + \Delta u$ and $v = v^b + \Delta v$, with $\max\{\|\Delta u\|_2, \|\Delta v\|_2\} \leq 1/4$. Consequently, we can write

$$\langle Xv, Wu \rangle = \langle Xv^b, Wu^a \rangle + \langle Xv^b, W\Delta u \rangle + \langle X\Delta v, Wu^a \rangle + \langle X\Delta v, W\Delta u \rangle. \tag{B.5}$$

By construction, we have the bound $|\langle Xv^b, W\Delta u\rangle| \leq \Psi(1, 1/4) = \frac{1}{4}\Psi(1, 1)$, and similarly $|\langle X\Delta v, Wu^a\rangle| \leq \frac{1}{4}\Psi(1, 1)$ as well as $|\langle X\Delta v, W\Delta u\rangle| \leq \frac{1}{16}\Psi(1, 1)$. Substituting these bounds into the decomposition (B.5) and taking suprema over the left and right-hand sides, we conclude that

$$\Psi(1, 1) \leq \max_{u^a \in \mathcal{A}, v^b \in \mathcal{B}} \langle Xv^b, Wu^a\rangle + \frac{9}{16}\Psi(1, 1),$$

from which the bound (B.4) follows.

We now apply the union bound to control the discrete maximum. It is known (e.g., [78, 88]) that there exists a $1/4$ covering of S^{d_1-1} and S^{d_2-1} with at most $A \leq 8^{d_1}$ and $B \leq 8^{d_2}$ elements respectively. Consequently, we have

$$\mathbb{P}\big[|\Psi(1, 1)| \geq 4\delta\, n\big] \leq 8^{d_1+d_2} \max_{u^a, v^b} \mathbb{P}\left[\frac{|\langle Xv^b, Wu^a\rangle|}{n} \geq \delta\right]. \tag{B.6}$$

It remains to obtain a good bound on the quantity $\frac{1}{n}\langle Xv, Wu\rangle = \frac{1}{n}\sum_{i=1}^n \langle v, X_i\rangle\langle u, W_i\rangle$, where $(u, v) \in S^{d_1-1} \times S^{d_2-1}$ are arbitrary but fixed. Since $W_i \in \mathbb{R}^{d_1}$ has i.i.d. $N(0, \nu^2)$ elements and u is fixed, we have $Z_i := \langle u, W_i\rangle \sim N(0, \nu^2)$ for each $i = 1, \ldots, n$. These variables are independent of one another, and of the random matrix X. Therefore, conditioned on X, the sum $Z := \frac{1}{n}\sum_{i=1}^n \langle v, X_i\rangle\langle u, W_i\rangle$ is zero-mean Gaussian with variance

$$\alpha^2 := \frac{\nu^2}{n}\left(\frac{1}{n}\|Xv\|_2^2\right) \;\leq\; \frac{\nu^2}{n}|||X^TX/n|||_2.$$

Define the event $\mathcal{T} = \{\alpha^2 \leq \frac{9\nu^2|||\Sigma|||_2}{n}\}$. Using Lemma 4.2, we have $|||X^TX/n|||_2 \leq 9\sigma_{\max}(\Sigma)$ with probability at least $1 - 2\exp(-n/2)$, which implies that $\mathbb{P}[\mathcal{T}^c] \leq 2\exp(-n/2)$. Therefore, conditioning on the event $\mathcal{T}$ and its complement $\mathcal{T}^c$, we obtain

$$\begin{aligned}\mathbb{P}[|Z| \geq t] &\leq \mathbb{P}\big[|Z| \geq t \mid \mathcal{T}\big] + \mathbb{P}[\mathcal{T}^c] \\ &\leq \exp\left(-n\frac{t^2}{2\nu^2\,(4 + |||\Sigma|||_2)}\right) + 2\exp(-n/2).\end{aligned}$$

Combining this tail bound with the upper bound (B.6), we have

$$\mathbb{P}\big[|\psi(1, 1)| \geq 4\delta\, n\big] \leq 8^{d_1+d_2}\left\{\exp\left(-n\frac{t^2}{18\nu^2|||\Sigma|||_2}\right) + 2\exp(-n/2)\right\}.$$

Setting $t^2 = 20\nu^2|||\Sigma|||_2\frac{d_1+d_2}{n}$, this probability vanishes as long as $n > 16(d_1 + d_2)$.

B.4 Technical details for Corollary 4.4

In this appendix, we collect the proofs of Lemmas 4.4 and 4.5.

B.4.1 Proof of Lemma 4.4

Recalling that S^{m-1} denotes the unit-norm Euclidean sphere in m-dimensions, we first observe that $|||X|||_2 = \sup_{u \in S^{m-1}} \|Xu\|_2$. Our next step is to reduce the supremum to a maximization over a finite set, using a standard covering argument. Let $\mathcal{A} = \{u^1, \ldots, u^A\}$ denote a $1/2$-cover of it. By definition, for any $u \in S^{m-1}$, there is some $u^a \in \mathcal{A}$ such that $u = u^a + \Delta u$, where $\|\Delta u\|_2 \leq 1/2$. Consequently, for any $u \in S^{m-1}$, the triangle inequality implies that

$$\|Xu\|_2 \leq \|Xu^a\|_2 + \|X\Delta u\|_2,$$

and hence that $|||X|||_2 \leq \max_{u^a \in \mathcal{A}} \|Xu^a\|_2 + \frac{1}{2}|||X|||_2$. Re-arranging yields the useful inequality

$$|||X|||_2 \leq 2 \max_{u^a \in \mathcal{A}} \|Xu^a\|_2. \tag{B.7}$$

Using inequality (B.7), we have

$$\begin{aligned} \mathbb{P}\left[\frac{1}{n}|||X^T X|||_2 > t\right] &\leq \mathbb{P}\left[\max_{u^a \in \mathcal{A}} \frac{1}{n}\sum_{i=1}^n (\langle u^a, X_i\rangle)^2 > \frac{t}{2}\right] \\ &\leq 4^m \max_{u^a \in \mathcal{A}} \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n (\langle u^a, X_i\rangle)^2 > \frac{t}{2}\right]. \end{aligned} \tag{B.8}$$

where the last inequality follows from the union bound, and the fact [78, 88] that there exists a $1/2$-covering of S^{m-1} with at most 4^m elements.

In order to complete the proof, we need to obtain a sharp upper bound on the quantity $\mathbb{P}\big[\frac{1}{n}\sum_{i=1}^n (\langle u, X_i\rangle)^2 > \frac{t}{2}\big]$, valid for any fixed $u \in S^{m-1}$. Define the random vector $Y \in \mathbb{R}^n$ with elements $Y_i = \langle u, X_i \rangle$. Note that Y is zero mean, and its covariance matrix R has elements $R_{ij} = \mathbb{E}[Y_i Y_j] = u^T \Sigma (\Theta^*)^{|j-i|} u$. In order to bound the spectral norm of R, we note that since it is symmetric, we have $|||R|||_2 \leq \max_{i=1,\ldots,m} \sum_{j=1}^m |R_{ij}|$, and moreover

$$|R_{ij}| = |u^T \Sigma (\Theta^*)^{|j-i|} u| \;\leq\; (|||\Theta^*|||_2)^{|j-i|}\, \Sigma \;\leq\; \gamma^{|j-i|}\, |||\Sigma|||_2.$$

Combining the pieces, we obtain

$$|||R|||_2 \leq \max_i \sum_{j=1}^m |\gamma|^{|i-j|} |||\Sigma|||_2 \;\leq\; 2|||\Sigma|||_2 \sum_{j=0}^{\infty} |\gamma|^j \;\leq\; \frac{2|||\Sigma|||_2}{1-\gamma}. \tag{B.9}$$

Moreover, we have $\mathrm{trace}(R)/n = u^T \Sigma u \leq |||\Sigma|||_2$. Applying Lemma B.2 with $t = 5\sqrt{\frac{m}{n}}$, we conclude that

$$\mathbb{P}\left[\frac{1}{n}\|Y\|_2^2 \;>\; |||\Sigma|||_2 + 5\sqrt{\frac{m}{n}}\, |||R|||_2\right] \;\leq\; 2\exp\big(-5m\big) + 2\exp -n/2)..$$

Combined with the bound (B.8), we obtain

$$\|\frac{1}{n}X^TX\|_2 \leq \|\Sigma\|_2\left\{2+\frac{20}{(1-\gamma)}\sqrt{\frac{m}{n}}\right\} \leq \frac{24\|\Sigma\|_2}{(1-\gamma)}, \tag{B.10}$$

with probability at least $1-c_1\exp(-c_2\,m)$, which establishes the upper bound (4.35)(a).

Turning to the lower bound (4.35)(b), we let $\mathcal{B}=\{v^1,\ldots,v^B\}$ be an ϵ-cover of S^{m-1} for some $\epsilon\in(0,1)$ to be chosen. Thus, for any $v\in\mathbb{R}^m$, there exists some v^b such that $v=v^b+\Delta v$, and $\|\Delta v\|_2\leq\epsilon$. Define the function $\Psi:\mathbb{R}^m\times\mathbb{R}^m\to\mathbb{R}$ via $\Psi(u,v)=u^T\Big(\frac{1}{n}X^TX\Big)v$, and note that $\Psi(u,v)=\Psi(v,u)$. With this notation, we have

$$\begin{aligned}v^T\Big(\frac{1}{n}X^TX\Big)v=\Psi(v,v)&=\Psi(v^k,v^k)+2\Psi(\Delta v,v)+\Psi(\Delta v,\Delta v)\\&\geq\Psi(v^k,v^k)+2\Psi(\Delta v,v),\end{aligned}$$

since $\Psi(\Delta v,\Delta v)\geq 0$. Since $|\Psi(\Delta v,v)|\leq\epsilon\,\|\Big(\frac{1}{n}X^TX\Big)\|_2$, we obtain the lower bound

$$\sigma_{\min}\left(\Big(\frac{1}{n}X^TX\Big)\right)=\inf_{v\in S^{m-1}}v^T\Big(\frac{1}{n}X^TX\Big)v\;\geq\;\min_{v^b\in\mathcal{B}}\Psi(v^b,v^b)-2\epsilon\|\frac{1}{n}X^TX\|_2.$$

By the previously established upper bound(4.35)(a), have $\|\frac{1}{n}X^TX\|_2\leq\frac{24\|\Sigma\|_2}{(1-\gamma)}$ with high probability. Hence, choosing $\epsilon=\frac{(1-\gamma)\sigma_{\min}(\Sigma)}{200\|\Sigma\|_2}$ ensures that $2\epsilon\|\frac{1}{n}X^TX\|_2\leq\sigma_{\min}(\Sigma)/4$.

Consequently, it suffices to lower bound the minimum over the covering set. We first establish a concentration result for the function $\Psi(v,v)$ that holds for any fixed $v\in S^{m-1}$. Note that we can write

$$\Psi(v,v)=\frac{1}{n}\sum_{i=1}^{n}(\langle v,\,X_i\rangle)^2,$$

As before, if we define the random vector $Y\in\mathbb{R}^n$ with elements $Y_i=\langle v,\,X_i\rangle$, then $Y\sim N(0,R)$ with $\|R\|_2\leq\frac{2\|\Sigma\|_2}{1-\gamma}$. Moreover, we have $\mathrm{trace}(R)/n=v^T\Sigma v\geq\sigma_{\min}(\Sigma)$. Consequently, applying Lemma B.2 yields

$$\mathbb{P}\left[\frac{1}{n}\,\|Y\|_2^2\;<\;\sigma_{\min}(\Sigma)-\frac{8t\|\Sigma\|_2}{1-\gamma}\right]\leq 2\exp\big(-n(t-2/\sqrt{n})^2/2\big)+2\exp(-\frac{n}{2}),$$

Note that this bound holds for any fixed $v\in S^{m-1}$. Setting $t^*=\frac{(1-\gamma)\,\sigma_{\min}(\Sigma)}{16\|\Sigma\|_2}$ and applying the union bound yields that

$$\mathbb{P}\big[\min_{v^b\in\mathcal{B}}\Psi(v^b,v^b)\;<\;\sigma_{\min}(\Sigma)/2\big]\leq\;(\frac{4}{\epsilon})^m\;\left\{2\exp\big(-n(t^*-2/\sqrt{n})^2/2\big)+2\exp(-\frac{n}{2})\right\},$$

which vanishes as long as $n>\frac{4\log(4/\epsilon)}{(t^*)^2}m$.

B.4.2 Proof of Lemma 4.5

Let $S^{m-1} = \{u \in \mathbb{R}^m \mid \|u\|_2 = 1\}$ denote the Euclidean sphere in m-dimensions, and for positive scalars a and b, define the random variable

$$\Psi(a, b) := \sup_{u \in a\, S^{m-1}} \sup_{v \in b\, S^{m-1}} \langle Xv, Wu \rangle.$$

Note that our goal is to upper bound $\Psi(1, 1)$. Let $\mathcal{A} = \{u^1, \ldots, u^A\}$ and $\mathcal{B} = \{v^1, \ldots, v^B\}$ denote $1/4$ coverings of S^{m-1} and S^{m-1}, respectively. Following the same argument as in the proof of Lemma 4.3, we obtain the upper bound

$$\Psi(1, 1) \;\leq\; 4 \max_{u^a \in \mathcal{A}, v^b \in \mathcal{B}} \langle Xv^b, Wu^a \rangle \tag{B.11}$$

We now apply the union bound to control the discrete maximum. It is known (e.g., [78, 88]) that there exists a $1/4$ covering of S^{m-1} with at most 8^m elements. Consequently, we have

$$\mathbb{P}\big[|\psi(1, 1)| \geq 4\delta\, n\big] \leq 8^{2m} \max_{u^a, v^b} \mathbb{P}\big[\frac{|\langle Xv^b, Wu^a \rangle|}{n} \geq \delta\big]. \tag{B.12}$$

It remains to obtain a tail bound on the quantity $\mathbb{P}\big[\frac{|\langle Xv, Wu \rangle|}{n} \geq \delta\big]$, for any fixed pair $(u, v) \in \mathcal{A} \times \mathcal{B}$.

For each $i = 1, \ldots, n$, let X_i and W_i denote the i^{th} row of X and W. Following some simple algebra, we have the decomposition $\frac{\langle Xv, Wu \rangle}{n} = T_1 - T_2 - T_3$, where

$$\begin{aligned}
T_1 &= \frac{1}{2n} \sum_{i=1}^{n} \big(\langle u, W_i \rangle + \langle v, X_i \rangle\big)^2 - \frac{1}{2}(u^T C u + v^T \Sigma v) \\
T_2 &= \frac{1}{2n} \sum_{i=1}^{n} \big(\langle u, W_i \rangle\big)^2 - \frac{1}{2} u^T C u \\
T_3 &= \frac{1}{2n} \sum_{i=1}^{n} \big(\langle v, X_i \rangle\big)^2 - \frac{1}{2} v^T \Sigma v
\end{aligned}$$

We may now bound each T_j for $j = 1, 2, 3$ in turn; in doing so, we make repeated use of Lemma B.2, which provides concentration bounds for a random variable of the form $\|Y\|_2^2$, where $Y \sim N(0, Q)$ for some matrix $Q \succeq 0$.

Bound on T_3: We can write the term T_3 as a deviation of $\|Y\|_2^2/n$ from its mean, where in this case the covariance matrix Q is no longer the identity. In concrete terms, let us define a random vector $Y \in \mathbb{R}^n$ with elements $Y_i := \langle v, X_i \rangle$. As seen in the proof of Lemma 4.4 from Appendix B.4.1, the vector Y is zero-mean Gaussian with covariance matrix R such that

$|||R|||_2 \leq \frac{2|||\Sigma|||_2}{1-\gamma}$ (see equation (B.9)). Since we have $\text{trace}(R)/n = v^T R v$, applying Lemma B.2 yields that

$$\mathbb{P}\big[|T_3| \geq \frac{8|||\Sigma|||_2}{1-\gamma} t\big] \leq 2\exp\Big(-\frac{n\,(t - 2/\sqrt{n})^2}{2}\Big) + 2\exp(-n/2). \tag{B.13}$$

Bound on T_2: We control the term T_2 in a similar way. Define the random vector $Y' \in \mathbb{R}^n$ with elements $Y'_i := \langle u, W_i\rangle$. Then Y is a sample from the distribution $N\big(0, (u^T C u) I_{n\times n}\big)$, so that $\frac{2}{u^T C u} T_2$ is the difference between a rescaled χ^2 variable and its mean. Applying Lemma B.2 with $Q = (u^T C u) I$, we obtain

$$\mathbb{P}\big[|T_2| > 4(u^T C u)\, t\big] \leq 2\exp\Big(-\frac{n\,(t - 2/\sqrt{n})^2}{2}\Big) + 2\exp(-n/2). \tag{B.14}$$

Bound on T_1: To control this quantity, let us define a zero-mean Gaussian random vector $Z \in \mathbb{R}^n$ with elements $Z_i = \langle v, X_i\rangle + \langle u, W_i\rangle$. This random vector has covariance matrix S with elements

$$S_{ij} = \mathbb{E}[Z_i Z_j] = (u^T C u)\delta_{ij} + (1-\delta_{ij})(u^T C u) v^T (\theta^*)^{|i-j|-1} u + v^T (\theta^*)^{|i-j|}\Sigma v,$$

where δ_{ij} is the Kronecker delta for the event $\{i = j\}$. As before, by symmetry of S, we have $|||S|||_2 \leq \max_{i=1,\ldots,n} \sum_{j=1}^n |S_{ij}|$, and hence

$$|||S|||_2 \leq (u^T C u) + |||\Sigma|||_2 + \sum_{j=1}^{i-1} |(u^T C u)\; v^T (\theta^*)^{|i-j|-1} u + v^T (\theta^*)^{|i-j|}\Sigma v| \\ + \sum_{j=i+1}^{n} |(u^T C u)\; v^T (\theta^*)^{|i-j|-1} u + v^T (\theta^*)^{|i-j|}\Sigma v|.$$

Since $|||\theta^*|||_2 \leq \gamma < 1$, and $(u^T C u) \leq |||C|||_2 \leq |||\Sigma|||_2$, we have

$$\begin{aligned} |||S|||_2 &\leq |||C|||_2 + |||\Sigma|||_2 + 2\sum_{j=1}^{\infty} |||C|||_2 \gamma^{j-1} + 2\sum_{j=1}^{\infty} |||\Sigma|||_2 \gamma^j \\ &\leq 4\,|||\Sigma|||_2 \big(1 + \frac{1}{1-\gamma}\big) \end{aligned}$$

Moreover, we have $\frac{\text{trace}(S)}{n} = (u^T C u) + v^T \Sigma v \leq 2|||\Sigma|||_2$, so that by applying Lemma B.2, we conclude that

$$\mathbb{P}\Big[|T_1| > \big(\frac{24|||\Sigma|||_2}{1-\gamma}\big)\, t\Big] \leq 2\exp\Big(-\frac{n\,(t - 2/\sqrt{n})^2}{2}\Big) + 2\exp(-n/2), \tag{B.15}$$

which completes the analysis of this term.

Combining the bounds (B.13), (B.14) and (B.15), we conclude that for all $t > 0$,

$$\mathbb{P}\big[\frac{|\langle Xv,\, Wu\rangle|}{n} \geq \frac{40(|\!|\!|\Sigma|\!|\!|_2\, t)}{1-\gamma}\big] \leq 6\exp\Big(-\frac{n\,(t\,-\,2/\sqrt{n})^2}{2}\Big) + 6\exp(-n/2). \tag{B.16}$$

Setting $t = 10\sqrt{m/n}$ and combining with the bound (B.12), we conclude that

$$\mathbb{P}\big[|\psi(1,1)| \geq \frac{1600|\!|\!|\Sigma|\!|\!|_2}{1-\gamma}\sqrt{\frac{m}{n}}\big] \leq 8^{2m}\left\{6\exp(-16m) + 6\exp(-n/2)\right\} \leq 12\exp(-m)$$

as long as $n > ((4\log 8) + 1)m$.

B.5 Proof of Proposition 4.1

We begin by stating and proving a useful lemma and invite the reader to compare the following result to the discussion on general Gaussian operators presented in Appendix D.4. Recall the definition (4.22) of $\zeta_{\text{mat}}(\Sigma)$.

Lemma B.1. *Let $X \in \mathbb{R}^{d_1 \times d_2}$ be a random sample from the Σ-ensemble. Then we have*

$$\mathbb{E}[|\!|\!|X|\!|\!|_2] \leq 12\sqrt{\zeta_{mat}(\Sigma)}\,\big[\sqrt{d_1} + \sqrt{d_2}\big] \tag{B.17}$$

and moreover

$$\mathbb{P}\big[|\!|\!|X|\!|\!|_2 \geq \mathbb{E}[|\!|\!|X|\!|\!|_2] + t\big] \leq \exp\big(-\frac{t^2}{2\rho^2(\Sigma)}\big). \tag{B.18}$$

Proof. We begin by making note of the variational representation

$$|\!|\!|X|\!|\!|_2 = \sup_{(u,v)\in S^{d_1-1}\times S^{d_2-1}} u^T X v.$$

Since each variable $u^T X v$ is zero-mean Gaussian, we thus recognize $|\!|\!|X|\!|\!|_2$ as the supremum of a Gaussian process. The bound (B.18) thus follows from Theorem 7.1 in Ledoux [77].

We now use a simple covering argument establish the upper bound (B.17). Let $\{v^1, \ldots, v^{M_2}\}$ be a $1/4$ covering of the sphere S^{d_2-1}. For an arbitrary $v \in S^{d_2-1}$, there exists some v^j in the cover such that $\|v - v^j\|_2 \leq 1/4$, whence

$$\|Xv\|_2 \leq \|Xv^j\|_2 + \|X(v - v_j)\|_2.$$

Taking suprema over both sides, we obtain that $|\!|\!|X|\!|\!|_2 \leq \max_{j=1,\ldots,M_2}\|Xv^j\|_2 + \frac{1}{4}|\!|\!|X|\!|\!|_2$. A similar argument using a $1/4$-covering $\{u^1, \ldots, u^{M_1}\}$ of S^{d_1-1} yields that

$$\|Xv^j\|_2 \leq \max_{i=1,\ldots,M_1}\langle u^i,\, Xv^j\rangle + \frac{1}{4}|\!|\!|X|\!|\!|_2.$$

Combining the pieces, we conclude that

$$|||X|||_2 \leq 2 \max_{\substack{i=1,\ldots,M_1 \\ j=1,\ldots,M_2}} \langle u^i, Xv^j \rangle.$$

By construction, each variable $\langle u^i, Xv^j \rangle$ is zero-mean Gaussian with variance at most $\rho(\Sigma)$, so that by standard bounds on Gaussian maxima, we obtain

$$\mathbb{E}[|||X|||_2] \leq 4\sqrt{\zeta_{\text{mat}}(\Sigma)}\sqrt{\log(M_1 M_2)} \;\leq\; 4\sqrt{\zeta_{\text{mat}}(\Sigma)}\big[\sqrt{\log M_1} + \sqrt{\log M_2}\big].$$

There exist $1/4$-coverings of S^{d_1-1} and S^{d_2-1} with $\log M_1 \leq d_1 \log 8$ and $\log M_2 \leq d_2 \log 8$, from which the bound (B.17) follows.

□

We now return to the proof of Proposition 4.1. To simplify the proof, let us define an operator $T_\Sigma : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ such that $\text{vec}(T_\Sigma(\Theta)) = \sqrt{\Sigma}\,\text{vec}(\Theta)$. Let $\mathfrak{X}' : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^N$ be a random Gaussian operator formed with X_i' sampled with i.i.d. $N(0,1)$ entries. By construction, we then have $\mathfrak{X}(\Theta) = \mathfrak{X}'(T_\Sigma(\Theta))$ for all $\Theta \in \mathbb{R}^{d_1 \times d_2}$. Now by the variational characterization of the ℓ_2-norm, we have

$$\|\mathfrak{X}'(T_\Sigma(\Theta))\|_2 = \sup_{u \in S^{N-1}} \langle u, \mathfrak{X}'(T_\Sigma(\Theta)) \rangle.$$

Since the original claim (4.25) is invariant to rescaling, it suffices to prove it for matrices such that $|||T_\Sigma(\Theta)|||_F = 1$. Letting $t \geq 1$ be a given radius, we seek lower bounds on the quantity

$$Z^*(t) := \inf_{\Theta \in \mathcal{R}(t)} \sup_{u \in S^{N-1}} \langle u, \mathfrak{X}'(T_\Sigma(\Theta))) \rangle, \quad \text{where } \mathcal{R}(t) = \{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid |||T_\Sigma(\Theta)|||_F = 1,\ |||\Theta|||_{\text{nuc}} \leq t\}.$$

In particular, our goal is to prove that for any $t \geq 1$, the lower bound

$$\frac{Z^*(t)}{\sqrt{N}} \geq c_1 - c_2 \sqrt{\zeta_{\text{mat}}(\Sigma)}\Big[\frac{d_1 + d_2}{N}\Big]^{1/2} t \tag{B.19}$$

holds with probability at least $1 - c_1 \exp(-c_2 N)$. By a standard peeling argument (see Raskutti et al. [109] for details), this lower bound implies the claim (4.25).

We establish the lower bound (B.19) using Gaussian comparison inequalities [78] and concentration of measure (see Lemma 2.3). For each pair $(u, \Theta) \in S^{N-1} \times \mathcal{R}(t)$, consider the random variable $Z_{u,\Theta} = \langle u, \mathfrak{X}'(T_\Sigma(\Theta)) \rangle$, and note that it is Gaussian with zero mean. For any two pairs (u, Θ) and (u', Θ'), some calculation yields

$$\mathbb{E}\big[(Z_{u,\Theta} - Z_{u',\Theta'})^2\big] = |||u \otimes T_\Sigma(\Theta) - u' \otimes T_\Sigma(\Theta')|||_F^2. \tag{B.20}$$

We now define a second Gaussian process $\{Y_{u,\Theta} \mid (u,\Theta) \in S^{N-1} \times \mathcal{R}(t)\}$ via

$$Y_{u,\Theta} := \langle g,\, u\rangle + \langle\!\langle G,\, T_\Sigma(\Theta)\rangle\!\rangle,$$

where $g \in \mathbb{R}^N$ and $G \in \mathbb{R}^{d_1 \times d_2}$ are independent with i.i.d. $N(0,1)$ entries. By construction, $Y_{u,\Theta}$ is zero-mean, and moreover, for any two pairs (u,Θ) and (u',Θ'), we have

$$\mathbb{E}\big[(Y_{u,\Theta} - Y_{u',\Theta'})^2\big] = \|u-u'\|_2^2 + |\!|\!| T_\Sigma(\Theta) - T_\Sigma(\Theta') |\!|\!|_F^2. \tag{B.21}$$

For all pairs $(u,\Theta),(u',\Theta') \in S^{N-1} \times \mathcal{R}(t)$, we have $\|u\|_2 = \|u'\|_2 = 1$, and moreover $|\!|\!| T_\Sigma(\Theta) |\!|\!|_F = |\!|\!| T_\Sigma(\Theta') |\!|\!|_F = 1$. Using this fact, some algebra yields that

$$|\!|\!| u \otimes T_\Sigma(\Theta) - u' \otimes T_\Sigma(\Theta') |\!|\!|_F^2 \le \|u-u'\|_2^2 + |\!|\!| T_\Sigma(\Theta) - T_\Sigma(\Theta') |\!|\!|_F^2. \tag{B.22}$$

Moreover, equality holds whenever $\Theta = \Theta'$. The conditions of the Gordon-Slepian inequality [78] are satisfied, so that we are guaranteed that

$$\mathbb{E}[\inf_{\Theta\in\mathcal{R}(t)} \|\mathfrak{X}'(T_\Sigma(\Theta))\|_2] \;=\; \mathbb{E}\Big[\inf_{\Theta\in\mathcal{R}(t)} \sup_{u\in S^{N-1}} Z_{u,\Theta}\Big] \ge \mathbb{E}\Big[\inf_{\Theta\in\mathcal{R}(t)} \sup_{u\in S^{N-1}} Y_{u,\Theta}\Big] \tag{B.23}$$

We compute

$$\begin{aligned}
\mathbb{E}\Big[\inf_{\Theta\in\mathcal{R}(t)} \sup_{u\in S^{N-1}} Y_{u,\Theta}\Big] &= \mathbb{E}\Big[\sup_{u\in S^{N-1}} \langle g,\, u\rangle\Big] + \mathbb{E}\Big[\inf_{\Theta\in\mathcal{R}(t)} \langle\!\langle G,\, T_\Sigma(\Theta)\rangle\!\rangle\Big] \\
&= \mathbb{E}[\|g\|_2] - \mathbb{E}[\sup_{\Theta\in\mathcal{R}(t)} \langle\!\langle G,\, T_\Sigma(\Theta)\rangle\!\rangle] \\
&\ge \frac{1}{2}\sqrt{N} - t\,\mathbb{E}[|\!|\!| T_\Sigma(G) |\!|\!|_2],
\end{aligned}$$

where we have used the fact that T_Σ is self-adjoint, and Hölder's inequality (involving the operator and nuclear norms). Since $T_\Sigma(G)$ is a random matrix from the Σ-ensemble, Lemma B.1 yields the upper bound $\mathbb{E}[|\!|\!| T_\Sigma(G) |\!|\!|_2] \le 12\sqrt{\zeta_{\mathrm{mat}}(\Sigma)}\,(\sqrt{d_1}+\sqrt{d_2})$. Putting together the pieces, we conclude that

$$\mathbb{E}\Big[\inf_{\Theta\in\mathcal{R}(t)} \frac{\|\mathfrak{X}'(T_\Sigma(\Theta))\|_2}{\sqrt{N}}\Big] \ge \frac{1}{2} - 12\,\sqrt{\zeta_{\mathrm{mat}}(\Sigma)}\,\Big(\frac{\sqrt{d_1}+\sqrt{d_2}}{\sqrt{N}}\Big)\,t.$$

Finally, we need to establish sharp concentration around the mean. Since $|\!|\!| T_\Sigma(\Theta) |\!|\!|_F = 1$ for all $\Theta \in \mathcal{R}(t)$, the function $f(\mathfrak{X}) := \inf_{\Theta\in\mathcal{R}(t)} \|\mathfrak{X}'(T_\Sigma(\Theta))\|_2/\sqrt{N}$ is Lipschitz with constant $1/\sqrt{N}$, so that Proposition 2.3 implies that

$$\mathbb{P}\Big[\inf_{\Theta\in\mathcal{R}(t)} \frac{\|\mathfrak{X}(\Theta)\|_2}{\sqrt{N}} \le \frac{1}{2} - 12\,\sqrt{\zeta_{\mathrm{mat}}(\Sigma)}\,\Big(\frac{\sqrt{d_1}+\sqrt{d_2}}{\sqrt{N}}\Big)\,t - \delta\Big] \le 2\exp(-N\delta^2/2) \qquad \text{for all } \delta > 0.$$

Setting $\delta = 1/4$ yields the claim.

B.6 Some useful concentration results

We recall Proposition 2.3, which states that a Lipschitz function of Gaussian random variables concentrates around its mean. By exploiting this proposition, we can prove the following result, which yields concentration of the squared ℓ_2-norm of an arbitrary Gaussian vector:

Lemma B.2. *Given a Gaussian random vector $Y \sim N(0, Q)$, for all $t > 2/\sqrt{n}$, we have*

$$\mathbb{P}\left[\frac{1}{n}\big|\|Y\|_2^2 - \operatorname{trace} Q\big| > 4\,t\,|||Q|||_2\right] \;\leq\; 2\exp\left(-\frac{n(t-\frac{2}{\sqrt{n}})^2}{2}\right) + 2\exp\left(-n/2\right). \tag{B.24}$$

Proof. Let $\sqrt{Q}$ be the symmetric matrix square root, and consider the function $f(x) = \|\sqrt{Q}x\|_2/\sqrt{n}$. Since it is Lipschitz with constant $|||\sqrt{Q}|||_2/\sqrt{n}$, Lemma 2.3 implies that

$$\mathbb{P}\big[\,\big|\,\|\sqrt{Q}X\|_2 - \mathbb{E}\|\sqrt{Q}X\|_2\,\big| > \sqrt{n}\delta\big] \leq 2\exp\left(-\frac{n\delta^2}{2|||Q|||_2}\right) \qquad \text{for all } \delta > 0. \tag{B.25}$$

By integrating this tail bound, we find that the variable $Z = \|\sqrt{Q}X\|_2/\sqrt{n}$ satisfies the bound $\operatorname{var}(Z) \leq 4|||Q|||_2/n$, and hence conclude that

$$\big|\sqrt{\mathbb{E}[Z^2]} - |\mathbb{E}[Z]|\big| = \big|\sqrt{\operatorname{trace}(Q)/n} - \mathbb{E}[\|\sqrt{Q}X\|_2/\sqrt{n}]\big| \;\leq\; \frac{2\sqrt{|||Q|||_2}}{\sqrt{n}}. \tag{B.26}$$

Combining this bound with the tail bound (B.25), we conclude that

$$\mathbb{P}\Big[\frac{1}{\sqrt{n}}\big|\|\sqrt{Q}X\|_2 - \sqrt{\operatorname{trace}(Q)}\,\big| > \delta + 2\sqrt{\frac{|||Q|||_2}{n}}\,\Big] \leq 2\exp\left(-\frac{n\delta^2}{2|||Q|||_2}\right) \qquad \text{for all } \delta > 0. \tag{B.27}$$

Setting $\delta = (t - 2/\sqrt{n})\sqrt{|||Q|||_2}$ in the bound (B.27) yields that

$$\mathbb{P}\Big[\frac{1}{\sqrt{n}}\big|\|\sqrt{Q}X\|_2 - \sqrt{\operatorname{trace}(Q)}\,\big| > t\sqrt{|||Q|||_2}\,\Big] \leq 2\exp\left(-\frac{n(t-2/\sqrt{n})^2}{2}\right). \tag{B.28}$$

Similarly, setting $\delta = \sqrt{|||Q|||_2}$ in the tail bound (B.27) yields that with probability greater than $1 - 2\exp(-n/2)$, we have

$$\left|\frac{\|Y\|_2}{\sqrt{n}} + \sqrt{\frac{\operatorname{trace}(Q)}{n}}\right| \leq \sqrt{\frac{\operatorname{trace}(Q)}{n}} + 3\sqrt{|||Q|||_2} \;\leq\; 4\sqrt{|||Q|||_2}. \tag{B.29}$$

Using these two bounds, we obtain

$$\left|\frac{\|Y\|_2^2}{n} - \frac{\operatorname{trace}(Q)}{n}\right| = \left|\frac{\|Y\|_2}{\sqrt{n}} - \sqrt{\frac{\operatorname{trace}(Q)}{n}}\right| \left|\frac{\|Y\|_2}{\sqrt{n}} + \sqrt{\frac{\operatorname{trace}(Q)}{n}}\right| \;\leq\; 4\,t\,|||Q|||_2$$

with the claimed probability. □

Appendix C

Proofs for Chapter 5

C.1 Proof of Lemma 5.1

We proceed via the probabilistic method, in particular by showing that a random procedure succeeds in generating such a set with probability at least 1/2. Let $M' = \exp\left(\frac{rm}{128}\right)$, and for each $\ell = 1, \ldots, M'$, we draw a random matrix $\widetilde{\Theta}^\ell \in \mathbb{R}^{m \times m}$ according to the following procedure:

(a) For rows $i = 1, \ldots, r$ and for each column $j = 1, \ldots, m$, choose each $\widetilde{\Theta}^\ell_{ij} \in \{-1, +1\}$ uniformly at random, independently across (i, j).

(b) For rows $i = r+1, \ldots, m$, set $\widetilde{\Theta}^\ell_{ij} = 0$.

We then let $Q \in \mathbb{R}^{m \times m}$ be a random unitary matrix, and define $\Theta^\ell = \frac{\delta}{\sqrt{rm}} Q\widetilde{\Theta}^\ell$ for all $\ell = 1, \ldots, M'$. The remainder of the proof analyzes the random set $\{\Theta^1, \ldots, \Theta^{M'}\}$, and shows that it contains a subset of size at least $M = M'/4$ that has properties (a) through (d) with probability at least 1/2.

By construction, each matrix $\widetilde{\Theta}^\ell$ has rank at most r, and Frobenius norm $|||\widetilde{\Theta}^\ell|||_F = \sqrt{rm}$. Since Q is unitary, the rescaled matrices Θ^ℓ have Frobenius norm $|||\Theta^\ell|||_F = \delta$. We now prove that

$$|||\Theta^\ell - \Theta^k|||_F \geq \delta \qquad \text{for all } \ell \neq k$$

with probability at least 7/8. Again, since Q is unitary, it suffices to show that $|||\widetilde{\Theta}^\ell - \widetilde{\Theta}^k|||_F \geq \sqrt{rm}$ for any pair $\ell \neq k$. We have

$$\frac{1}{rm}|||\widetilde{\Theta}^k - \widetilde{\Theta}^\ell|||_F^2 = \frac{1}{rm}\sum_{i=1}^{r}\sum_{j=1}^{m}\left(\widetilde{\Theta}^\ell_{ij} - \widetilde{\Theta}^k_{ij}\right)^2.$$

This is a sum of rm i.i.d. variables, each bounded by 4. The mean of the sum is 2, so that the Hoeffding bound implies that

$$\mathbb{P}\big[\frac{1}{rm}\|\widetilde{\Theta}^k - \widetilde{\Theta}^\ell\|_F^2 \leq 2 - t\big] \leq 2\exp\big(-rm\,t^2/32\big).$$

Since there are less than $(M')^2$ pairs of matrices in total, setting $t = 1$ yields

$$\mathbb{P}\big[\min_{\ell,k=1,\ldots,M'} \frac{\|\widetilde{\Theta}^\ell - \widetilde{\Theta}^k\|_F^2}{rm} \geq 1\big] \geq 1 - 2\exp\big(-\frac{rm}{32} + 2\log M'\big) \geq \frac{7}{8},$$

where we have used the facts $\log M' = \frac{rm}{128}$ and $m \geq 10$. Recalling the definition of Θ^ℓ, we conclude that

$$\mathbb{P}\big[\min_{\ell,k=1,\ldots,M'} \|\Theta^\ell - \Theta^k\|_F^2 \geq \delta^2\big] \geq \frac{7}{8}. \tag{C.1}$$

We now establish bounds on $\alpha_{\mathrm{sp}}(\Theta^\ell)$ and $\|\Theta^\ell\|_2$. We first prove that for any fixed index $\ell \in \{1, 2, \ldots, M'\}$, our construction satisfies

$$\mathbb{P}\Big[\alpha_{\mathrm{sp}}(\Theta^\ell) \leq \sqrt{32\log m}\Big] \geq \frac{3}{4}. \tag{C.2}$$

Indeed, for any pair of indices (i, j), we have $|\Theta^\ell_{ij}| = |\langle q_i, v_j\rangle|$, where $q_i \in \mathbb{R}^m$ is drawn from the uniform distribution over the m-dimensional sphere, and $\|v_j\|_2 = \sqrt{r}\,\frac{\delta}{\sqrt{rm}} = \frac{\|\Theta^\ell\|_F}{\sqrt{m}}$. By Levy's theorem for concentration on the sphere [77], we have

$$\mathbb{P}\big[|\langle q_i, v_j\rangle| \geq t\big] \leq 2\exp\big(-\frac{m^2}{8\,\|\Theta^\ell\|_F^2}t^2\big).$$

Setting $t = s/m$ and taking the union bound over all m^2 indices, we obtain

$$\mathbb{P}\big[m\,\|\Theta^\ell\|_\infty \geq s\big] \leq 2\exp\Big(-\frac{1}{8\,\|\Theta^\ell\|_F^2}s^2 + 2\log m\Big).$$

This probability is less than $1/2$ for $s = \|\Theta^\ell\|_F\sqrt{32\log m}$ and $m \geq 2$, which establishes the intermediate claim (C.2).

Finally, we turn to property (d). For each fixed ℓ, by definition of Θ^ℓ and the unitary nature of Q, we have $\|\Theta^\ell\|_2 = \frac{\delta}{\sqrt{rm}}\|U\|_2$, where $U \in \{-1, +1\}^{r\times m}$ is a random matrix with i.i.d. Rademacher (and hence sub-Gaussian) entries. Known results on sub-Gaussian matrices [42] yield

$$\mathbb{P}\Big[\frac{\delta}{\sqrt{rm}}\|U\|_2 \leq \frac{2\delta}{\sqrt{rm}}\big(\sqrt{r} + \sqrt{m}\big)\Big] \geq 1 - 2\exp\big(-\frac{1}{4}(\sqrt{r} + \sqrt{m})^2\big) \geq \frac{3}{4}$$

for $m \geq 10$. Since $r \leq m$, we conclude that

$$\mathbb{P}\Big[|||\Theta^\ell|||_2 \leq \frac{4\delta}{\sqrt{r}}\Big] \geq \frac{3}{4}. \tag{C.3}$$

By combining the bounds (C.2) and (C.3), we find that for each fixed $\ell = 1, \ldots, M'$, we have

$$\mathbb{P}\left[|||\Theta^\ell|||_2 \leq \frac{4\delta}{\sqrt{r}}, \ \frac{\alpha_{\mathrm{sp}}(\Theta^\ell)}{|||\Theta|||_F} \leq \sqrt{32 \log m}\right] \geq \frac{1}{2} \tag{C.4}$$

Consider the event $\mathcal{E}$ that there exists a subset $S \subset \{1, \ldots, M'\}$ of cardinality $M = \frac{1}{4}M'$ such that

$$|||\Theta^\ell|||_2 \leq 4\frac{\delta}{\sqrt{n}}, \quad \text{and} \quad \frac{\alpha_{\mathrm{sp}}(\Theta^\ell)}{|||\Theta|||_F} \leq \sqrt{32 \log m} \qquad \text{for all } \ell \in S.$$

By the bound (C.4), we have

$$\mathbb{P}[\mathcal{E}] \geq \sum_{k=M}^{M'} \binom{M'}{k} (1/2)^k.$$

Since we have chosen $M < M'/2$, we are guaranteed that $\mathbb{P}[\mathcal{E}] \geq 1/2$, thereby completing the proof.

C.2 Proof of Lemma 5.2

We first observe that for any $\Gamma \in \mathfrak{C}'(n; c_0)$ with $\|\Gamma\|_\infty = \frac{1}{m}$, we have

$$|||\Gamma|||_F^2 \geq c_0 \, |||\Gamma|||_{\mathrm{nuc}} \sqrt{\frac{m \log m}{n}} \ \geq c_0 |||\Gamma|||_F \sqrt{\frac{m \log m}{n}},$$

whence $|||\Gamma|||_F \geq c_0 \sqrt{\frac{m \log m}{n}}$. Accordingly, recalling the definition (5.40), it suffices to restrict our attention to sets $\mathcal{R}(R_{\mathrm{P}})$ with $R_{\mathrm{P}} \geq \mu := c_0 \sqrt{\frac{m \log m}{n}}$. For $\ell = 1, 2, \ldots$ and $\alpha = \frac{7}{6}$, define the sets

$$\mathbb{S}_\ell := \big\{\Gamma \in \mathfrak{C}'(n; c_0) \mid \|\Gamma\|_\infty = \frac{1}{m}, \quad \alpha^{\ell-1}\mu \leq |||\Gamma|||_F \leq \alpha^\ell \mu, \text{ and } |||\Gamma|||_{\mathrm{nuc}} \leq \rho(\alpha^\ell \mu)\big\}. \tag{C.5}$$

From the definition (5.40), note that by construction, we have $\mathbb{S}_\ell \subset \mathcal{R}(\alpha^\ell \mu)$.

Now if the event $\mathcal{E}(\mathfrak{X}')$ holds for some matrix Γ, then this matrix Γ must belong to some set $\mathbb{S}_\ell$. When $\Gamma \in \mathbb{S}_\ell$, then we are guaranteed the existence of a matrix $\Gamma \in \mathcal{R}(\alpha^\ell \mu)$ such

that

$$\begin{aligned}\Big|\frac{\|\mathfrak{X}'(\Gamma)\|_2}{\sqrt{n}} - |||\Gamma|||_F\Big| &> \frac{7}{8}|||\Gamma|||_F + \frac{48L}{\sqrt{n}} \\ &\geq \frac{7}{8}\alpha^{\ell-1}\mu + \frac{48L}{\sqrt{n}} \\ &= \frac{3}{4}\alpha^{\ell}\mu + \frac{48L}{\sqrt{n}},\end{aligned}$$

where the final equality follows since $\alpha = 7/6$. Thus, we have shown that when the violating matrix $\Gamma \in \mathbb{S}_\ell$, then event $\mathcal{E}(\mathfrak{X}'; \alpha^\ell\mu)$ must hold. Since any violating matrix must fall into some set $\mathbb{S}_\ell$, the union bound implies that

$$\begin{aligned}\mathbb{P}[\mathcal{E}(\mathfrak{X}')] &\leq \sum_{\ell=1}^{\infty} \mathbb{P}[\mathcal{E}(\mathfrak{X}'; \alpha^\ell\mu)] \\ &\leq c_1 \sum_{\ell=1}^{\infty} \exp\big(-c_2 n\alpha^{2\ell}\mu^2\big) \\ &\leq c_1 \sum_{\ell=1}^{\infty} \exp\big(-2c_2\log(\alpha)\,\ell\, n\mu^2\big) \\ &\leq c_1 \frac{\exp(-c_2' n\mu^2)}{1-\exp(-c_2' n\mu^2)}\end{aligned}$$

Since $n\mu^2 = \Omega(m\log m)$, the claim follows.

C.3 Proof of Lemma 5.3

For a fixed matrix Γ, define the function $F_\Gamma(\mathfrak{X}') = \frac{1}{\sqrt{n}}\|\mathfrak{X}'(\Gamma)\|_2$. We prove the lemma in two parts: first, we establish that for any fixed Γ, the function F_Γ satisfies the tail bound

$$\mathbb{P}\big[|F_\Gamma(\mathfrak{X}') - |||\Gamma|||_F| \geq \delta + \frac{48L}{\sqrt{n}}\big] \leq 4\exp\big(-\frac{n\delta^2}{4L^2}\big). \tag{C.6}$$

We then show that there exists a δ-covering of $\overline{\mathcal{R}}(R_{\mathrm{P}})$ such that

$$\log N(\delta) \leq 36\big(\Upsilon(R_{\mathrm{P}})/\delta\big)^2\ m. \tag{C.7}$$

Combining the tail bound (C.6) with the union bound, we obtain

$$\begin{aligned}\mathbb{P}\big[\max_{k=1,\ldots,N(\delta)} |F_\Gamma(\mathfrak{X}') - |||\Gamma^k|||_F| \geq \delta + \frac{16L}{\sqrt{n}}\big] &\leq 4\exp\big(-\frac{n\delta^2}{4L^2} + \log N(\delta)\big) \\ &\leq 4\exp\Big\{-\frac{n\delta^2}{4L^2} + 36\big(\Upsilon(R_{\mathrm{P}})/\delta\big)^2\, m\Big\}\end{aligned}$$

where the final inequality follows uses the bound (C.7). Since Lemma 5.3 is based on the choice $\delta = R_{\mathrm{P}}/8$, it suffices to show that

$$\begin{aligned}\frac{nR_{\mathrm{P}}^2}{512\,L^2} &\geq 36\left(\Upsilon(R_{\mathrm{P}})/(R_{\mathrm{P}}/8)\right)^2 m \\ &= 36\left(\frac{8\,R_{\mathrm{P}}}{c_0\,L}\sqrt{\frac{n}{m\log m}}\right)^2 m \\ &= \frac{2304\,R_{\mathrm{P}}^2}{c_0^2\,L^2}\,\frac{n}{\log m}.\end{aligned}$$

Noting that the terms involving R_{P}^2, L^2, and n both cancel out, we see that for any fixed c_0, this inequality holds once $\log m$ is sufficiently large. By choosing c_0 sufficiently large, we can ensure that it holds for all $m \geq 2$.

It remains to establish the two intermediate claims (C.6) and (C.7).

Upper bounding the covering number (C.7): We start by proving the upper bound (C.7) on the covering number. To begin, let $\widetilde{N}(\delta)$ denote the δ-covering number (in Frobenius norm) of the nuclear norm ball $\mathbb{B}_1(\Upsilon(R_{\mathrm{P}})) = \{\Delta \in \mathbb{R}^{m\times m} \mid |\!|\!|\Delta|\!|\!|_{\mathrm{nuc}} \leq \Upsilon(R_{\mathrm{P}})\}$, and let $N(\delta)$ be the covering number of the set $\overline{\mathcal{R}}(R_{\mathrm{P}})$. We first claim that $N(\delta) \leq \widetilde{N}(\delta)$. Let $\{\Gamma^1, \ldots, \Gamma^{\widetilde{N}(\delta)}\}$ be a δ-cover of $\mathbb{B}_1(\Upsilon(R_{\mathrm{P}}))$, From equation (5.45), note that the set $\overline{\mathcal{R}}(R_{\mathrm{P}})$ is contained within $\mathbb{B}_1(\Upsilon(R_{\mathrm{P}}))$; in particular, it is obtained by intersecting the latter set with the set

$$\mathcal{S} := \{\Delta \in \mathbb{R}^{m\times m} \mid \|\Delta\|_\infty \leq \frac{1}{m},\ |\!|\!|\Delta|\!|\!|_F \leq R_{\mathrm{P}}\}.$$

Letting $\Pi_{\mathcal{S}}$ denote the projection operator under Frobenius norm onto this set, we claim that $\{\Pi_{\mathcal{S}}(\Gamma^j), j = 1, \ldots, \widetilde{N}(\delta)\}$ is a δ-cover of $\overline{\mathcal{R}}(R_{\mathrm{P}})$. Indeed, since $\mathcal{S}$ is non-empty, closed and convex, the projection operator is non-expansive [15], and thus for any $\Gamma \in \overline{\mathcal{R}}(R_{\mathrm{P}}) \subset \mathcal{S}$, we have

$$|\!|\!|\Pi_{\mathcal{S}}(\Gamma^j) - \Gamma|\!|\!|_F \;=\; |\!|\!|\Pi_{\mathcal{S}}(\Gamma^j) - \Pi_{\mathcal{S}}(\Gamma)|\!|\!|_F \leq |\!|\!|\Gamma^j - \Gamma|\!|\!|_F,$$

which establishes the claim.

We now upper bound $\widetilde{N}(\delta)$. Let $G \in \mathbb{R}^{m\times m}$ be a random matrix with i.i.d. $N(0,1)$ entries. By Sudakov minoration (cf. Theorem 5.6 in Pisier [107]), we have

$$\begin{aligned}\sqrt{\log \widetilde{N}(\delta)} &\leq \frac{3}{\delta}\,\mathbb{E}\big[\sup_{|\!|\!|\Delta|\!|\!|_{\mathrm{nuc}}\leq \Upsilon(R_{\mathrm{P}})} \langle\!\langle G,\ \Delta\rangle\!\rangle\big] \\ &\leq \frac{3\Upsilon(R_{\mathrm{P}})}{\delta}\,\mathbb{E}\big[|\!|\!|G|\!|\!|_2\big],\end{aligned}$$

where the second inequality follows from the duality between the nuclear and operator norms. From known results on the operator norms Gaussian random matrices [42], we have the upper bound $\mathbb{E}[|||G|||_2] \leq 2\sqrt{m}$, so that

$$\sqrt{\log \widetilde{N}(\delta)} \;\leq\; \frac{6\Upsilon(R_{\mathrm{P}})}{\delta}\,\sqrt{m},$$

thereby establishing the bound (C.7).

Establishing the tail bound (C.6): Recalling the definition of the operator $\mathfrak{X}'$, we have

$$\begin{aligned} F_\Gamma(\mathfrak{X}') &= \frac{1}{\sqrt{n}}\Big\{\sum_{i=1}^{n} \langle\!\langle \widetilde{X}^{(i)},\, \Gamma\rangle\!\rangle^2\Big\}^{1/2} \\ &= \frac{1}{\sqrt{n}} \sup_{\|u\|_2=1} \sum_{i=1}^{n} u_i \langle\!\langle \widetilde{X}^{(i)},\, \Gamma\rangle\!\rangle \\ &= \frac{1}{\sqrt{n}} \sup_{\|u\|_2=1} \sum_{i=1}^{n} u_i Y_i \end{aligned}$$

where we have defined the random variables $Y_i := \langle\!\langle \widetilde{X}^{(i)},\, \Gamma\rangle\!\rangle$. Note that each Y_i is zero-mean, and bounded by $2L$ since

$$\begin{aligned} |Y_i| &= |\langle\!\langle \widetilde{X}^{(i)},\, \Gamma\rangle\!\rangle| \\ &\leq \Big(\sum_{a,b} |\widetilde{X}^{(i)}|_{ab}\Big)\, \|\Gamma\|_\infty \;\leq\; 2L. \end{aligned}$$

where we have used the facts that $\|\Gamma\|_\infty \leq 2/m$, and $\sum_{a,b} |\widetilde{X}^{(i)}|_{ab} \leq L\, m$, by definition of the matrices $\widetilde{X}^{(i)}$.

Therefore, applying Corollary 4.8 from Ledoux [77], we conclude that

$$\mathbb{P}\big[|F_\Gamma(\mathfrak{X}') - \mathbb{E}[F_\Gamma(\mathfrak{X}')]| \geq \delta + \frac{32L}{\sqrt{n}}\big] \leq 4\exp\big(-\frac{n\delta^2}{4L^2}\big).$$

The same corollary implies that

$$\big|\sqrt{\mathbb{E}[F_\Gamma^2(\mathfrak{X}')]} - \mathbb{E}[F_\Gamma(\mathfrak{X}')]\big| \leq \frac{16L}{\sqrt{n}}.$$

Since $\mathbb{E}[F_\Gamma^2(\mathfrak{X}')] = |||\Gamma|||_F^2$, the tail bound (C.6) follows.

C.4 Proof of Lemma 5.4

From the proof of Lemma 5.3, recall the definition $F_\Gamma(\mathfrak{X}') = \frac{1}{\sqrt{n}}\|\mathfrak{X}'(\Gamma)\|_2$ where $\mathfrak{X}'$ is the random sampling operator defined by the n matrices $(\widetilde{X}^{(1)}, \ldots, \widetilde{X}^{(n)})$. Using this notation, our goal is to bound the function

$$G(\mathfrak{X}') := \sup_{\Delta \in \mathfrak{D}(\delta, R)} F_\Delta(\mathfrak{X}'),$$

where we recall that $\mathfrak{D}(\delta, R) := \big\{\Delta \in \mathbb{R}^{m_r \times m_c} \mid |\!|\!|\Delta|\!|\!|_F \leq \delta,\ |\!|\!|\Delta|\!|\!|_{\text{nuc}} \leq 2\Upsilon(R_{\text{P}}),\ \|\Delta\|_\infty \leq \frac{2}{m}\big\}$. Ultimately, we will set $\delta = \frac{R_{\text{P}}}{8}$, but we use δ until the end of the proof for compactness in notation.

Our approach is a standard one: first show concentration of G around its expectation $\mathbb{E}[G(\mathfrak{X}')]$, and then upper bound the expectation. We show concentration via a bounded difference inequality; since G is a symmetric function of its arguments, it suffices to establish the bounded difference property with respect to the first co-ordinate. In order to do so, consider a second operator $\widetilde{\mathfrak{X}}'$ defined by the matrices $(Z^{(1)}, \widetilde{X}^{(2)}, \ldots, \widetilde{X}^{(n)})$, differing from $\mathfrak{X}'$ only in the first matrix. Given the pair $(\mathfrak{X}', \widetilde{\mathfrak{X}}')$, we have

$$\begin{aligned}
G(\mathfrak{X}') - G(\widetilde{\mathfrak{X}}') &= \sup_{\Delta \in \mathfrak{D}(\delta, R)} F_\Delta(\mathfrak{X}') - \sup_{\Theta \in \mathfrak{D}(\delta, R)} F_\Theta(\widetilde{\mathfrak{X}}') \\
&\leq \sup_{\Delta \in \mathfrak{D}(\delta, R)} \big[F_\Delta(\mathfrak{X}') - F_\Delta(\widetilde{\mathfrak{X}}')\big] \\
&\leq \sup_{\Delta \in \mathfrak{D}(\delta, R)} \frac{1}{\sqrt{n}} \|\mathfrak{X}'(\Delta) - \widetilde{\mathfrak{X}}'(\Delta)\|_2 \\
&= \sup_{\Delta \in \mathfrak{D}(\delta, R)} \frac{1}{\sqrt{n}} \big|\langle\!\langle \widetilde{X}^{(1)} - Z^{(1)},\ \Delta \rangle\!\rangle\big|.
\end{aligned}$$

For any fixed $\Delta \in \mathfrak{D}(\delta, R)$, we have

$$\big|\langle\!\langle \widetilde{X}^{(1)} - Z^{(1)},\ \Delta \rangle\!\rangle\big| \leq 2Lm\, \|\Delta\|_\infty \ \leq\ 4L,$$

where we have used the fact that the matrix $\widetilde{X}^{(1)} - Z^{(1)}$ is non-zero in at most two entries with values upper bounded by $2Lm$. Combining the pieces yields $G(\mathfrak{X}') - G(\widetilde{\mathfrak{X}}') \leq \frac{4L}{\sqrt{n}}$. Since the same argument can be applied with the roles of $\mathfrak{X}'$ and $\widetilde{\mathfrak{X}}'$ interchanged, we conclude that $|G(\mathfrak{X}') - G(\widetilde{\mathfrak{X}}')| \leq \frac{4L}{\sqrt{n}}$. Therefore, by the bounded differences variant of the Azuma-Hoeffding inequality [77], we have

$$\mathbb{P}\big[|G(\mathfrak{X}') - \mathbb{E}[G(\mathfrak{X}')]| \geq t\big] \leq 2\exp\big(-\frac{nt^2}{32L^2}\big). \tag{C.8}$$

Next we bound the expectation. First applying Jensen's inequality, we have

$$\begin{aligned}(\mathbb{E}[G(\mathfrak{X}')])^2 &\leq \mathbb{E}[G^2(\mathfrak{X}')] \\ &= \mathbb{E}\big[\sup_{\Delta\in\mathfrak{D}(\delta,R)} \frac{1}{n}\sum_{i=1}^n \langle\!\langle \widetilde{X}^{(i)},\, \Delta\rangle\!\rangle^2\big] \\ &= \mathbb{E}\Big[\sup_{\Delta\in\mathfrak{D}(\delta,R)} \Big\{\frac{1}{n}\sum_{i=1}^n \big[\langle\!\langle \widetilde{X}^{(i)},\, \Delta\rangle\!\rangle^2 - \mathbb{E}[\langle\!\langle \widetilde{X}^{(i)},\, \Delta\rangle\!\rangle^2]\big] + |\!|\!|\Delta|\!|\!|_F^2\Big\}\Big] \\ &\leq \mathbb{E}\Big[\sup_{\Delta\in\mathfrak{D}(\delta,R)} \Big\{\frac{1}{n}\sum_{i=1}^n \big[\langle\!\langle \widetilde{X}^{(i)},\, \Delta\rangle\!\rangle^2 - \mathbb{E}[\langle\!\langle \widetilde{X}^{(i)},\, \Delta\rangle\!\rangle^2]\big]\Big\}\Big] + \delta^2,\end{aligned}$$

where we have used the fact that $\mathbb{E}[\langle\!\langle \widetilde{X}^{(i)},\, \Delta\rangle\!\rangle^2 = |\!|\!|\Delta|\!|\!|_F^2 \leq \delta^2$. Now a standard symmetrization argument [78] yields

$$\mathbb{E}_{\mathfrak{X}'}[G^2(\mathfrak{X}')] \;\leq 2\,\mathbb{E}_{\mathfrak{X}',\varepsilon}\big[\sup_{\Delta\in\mathfrak{D}(\delta,R)} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle\!\langle \widetilde{X}^{(i)},\, \Delta\rangle\!\rangle^2\big] + \delta^2,$$

where $\{\varepsilon_i\}_{i=1}^n$ is an i.i.d. Rademacher sequence. Since $|\langle\!\langle \widetilde{X}^{(i)},\, \Delta\rangle\!\rangle| \leq 2L$ for all i, the Ledoux-Talagrand contraction inequality (p. 112, Ledoux and Talagrand [78]) implies that

$$\mathbb{E}[G^2(\mathfrak{X}')] \;\leq 16L\;\mathbb{E}\big[\sup_{\Delta\in\mathfrak{D}(\delta,R)} \big\{\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle\!\langle \widetilde{X}^{(i)},\, \Delta\rangle\!\rangle\big\}\big] + \delta^2.$$

By the duality between operator and nuclear norms, we have

$$\big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle\!\langle \widetilde{X}^{(i)},\, \Delta\rangle\!\rangle\big| \leq |\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \widetilde{X}^{(i)}|\!|\!|_2\; |\!|\!|\Delta|\!|\!|_{\mathrm{nuc}},$$

and hence, since $|\!|\!|\Delta|\!|\!|_{\mathrm{nuc}} \leq \rho(R_{\mathrm{P}})$ for all $\Delta \in \mathfrak{D}(\delta, R)$, we have

$$\mathbb{E}[G^2(\mathfrak{X}')] \leq\; 16\,L\,\rho(R_{\mathrm{P}})\;\mathbb{E}\big[|\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \widetilde{X}^{(i)}|\!|\!|_2\big] + \delta^2. \tag{C.9}$$

It remains to bound the operator norm $\mathbb{E}\big[|\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \widetilde{X}^{(i)}|\!|\!|_2\big]$. The following lemma, proved in Appendix C.5, provides a suitable upper bound:

Lemma C.1. *We have the upper bound*

$$\mathbb{E}\big[|\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \widetilde{X}^{(i)}|\!|\!|_2\big] \leq 10\;\max\Big\{\sqrt{\frac{m\log m}{n}},\; \frac{L\,m\log m}{n}\Big\}. \tag{C.10}$$

Thus, as long as $n = \Omega(L\, m \log m)$, combined with the earlier bound (C.9), we conclude that

$$\mathbb{E}[G(\mathfrak{X}')] \leq \sqrt{\mathbb{E}[G^2(\mathfrak{X}')]} \leq \big[160\, L\, \Upsilon(R_{\mathrm{P}}) \sqrt{\frac{m \log m}{n}} + \delta^2\big]^{1/2},$$

where we have used the fact that $L \geq 1$. By definition of $\Upsilon(R_{\mathrm{P}})$, we have

$$160\, L\, \Upsilon(R_{\mathrm{P}}) \sqrt{\frac{m \log m}{n}} = \frac{160}{c_0} R_{\mathrm{P}}^2 \ \leq \big(\frac{5R_{\mathrm{P}}}{16}\big)^2,$$

where the final inequality can be guaranteed by choosing c_0 sufficiently large.

Consequently, recalling our choice $\delta = R_{\mathrm{P}}/8$ and using the inequality $\sqrt{a^2 + b^2} \leq |a| + |b|$, we obtain

$$\mathbb{E}[G(\mathfrak{X}')] \leq \frac{5}{16} R_{\mathrm{P}} + \frac{R_{\mathrm{P}}}{8} = \frac{7}{16} R_{\mathrm{P}}.$$

Finally, setting $t = \frac{R_{\mathrm{P}}}{16}$ in the concentration bound (C.8) yields

$$G(\mathfrak{X}') \leq \frac{R_{\mathrm{P}}}{16} + \frac{7}{16} R_{\mathrm{P}} = \frac{R_{\mathrm{P}}}{2}$$

with probability at least $1 - 2\exp\big(-c' \frac{nR_{\mathrm{P}}^2}{L^2}\big)$ as claimed.

C.5 Proof of Lemma C.1

We prove this lemma by applying a form of Ahlswede-Winter matrix bound [3], as stated in Appendix C.6, to the matrix $Y^{(i)} := \varepsilon_i \widetilde{X}^{(i)}$. We first compute the quantities involved in Lemma C.2. Note that $Y^{(i)}$ is a zero-mean random matrix, and satisfies the bound

$$\|Y^{(i)}\|_2 = m \frac{1}{\sqrt{R_{j(i)}}\sqrt{C_{k(i)}}} \|\varepsilon_i\, e_{j(i)}\, e_{k(i)}^T\|_2 \ \leq \ L\, m.$$

Let us now compute the quantities σ_i in Lemma C.2. We have

$$\mathbb{E}\big[(Y^{(i)^T}) Y^{(i)}\big] = \mathbb{E}\left[\frac{m^2}{R_{j(i)}\, C_{k(i)}} e_{k(i)} e_{k(i)}^T\right] \ = \ m I_{m\times m}$$

and similarly, $\mathbb{E}\big[Y^{(i)}\,(Y^{(i)})^T\big] = m I_{m\times m}$, so that

$$\sigma_i^2 = \max\Big\{\|\mathbb{E}\big[Y^{(i)}\,(Y^{(i)})^T\big]\|_2,\ \|\mathbb{E}\big[(Y^{(i)})^T\, Y^{(i)}\big]\|_2\Big\} \ = \ m.$$

Thus, applying Lemma C.2 yields the tail bound

$$\mathbb{P}\big[\|\sum_{i=1}^{n} \varepsilon_i \widetilde{X}^{(i)}\|_2 \geq t\big] \leq 2\,m\,\max\big\{\exp(-\frac{t^2}{4nm}),\ \exp(-\frac{t}{2Lm})\big\}.$$

Setting $t = n\delta$, we obtain

$$\mathbb{P}\big[\|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \widetilde{X}^{(i)}\|_2 \geq \delta\big] \leq 2m\,\max\big\{\exp(-\frac{n\delta^2}{4m}), \exp(-\frac{n\delta}{2Lm})\big\}.$$

Recall that for any non-negative random variable T, we have $\mathbb{E}[T] = \int_0^\infty \mathbb{P}[T \geq s]ds$. Applying this fact to $T := \|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \widetilde{X}^{(i)}\|_2$ and integrating the tail bound, we obtain

$$\mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \widetilde{X}^{(i)}\|_2\big] \leq 10\,\max\big\{\sqrt{\frac{m\log m}{n}},\ \frac{Lm\log m}{n}\big\}.$$

C.6 Ahlswede-Winter matrix bound

Here we state a Bernstein version of the Ahlswede-Winter tail bound [3] for the operator norm of a sum of random matrices. The version here is a slight weakening (but sufficient for our purposes) of a result due to Recht [115]; we also refer the reader to the notes of Vershynin [143], and the strengthened results provided by Tropp [133].

Let $Y^{(i)}$ be independent $m_r \times m_c$ zero-mean random matrices such that $\|Y^{(i)}\|_2 \leq M$, and define $\sigma_i^2 := \max\big\{\|\mathbb{E}[(Y^{(i)})^T Y^{(i)}]\|_2,\quad \|\mathbb{E}[Y^{(i)}(Y^{(i)})^T]\|_2\big\}$ as well as $\sigma^2 := \sum_{i=1}^{n} \sigma_i^2$.

Lemma C.2. *We have*

$$\mathbb{P}\big[\|\sum_{i=1}^{n} Y^{(i)}\|_2 \geq t\big] \leq (m_r \times m_c)\,\max\big\{\exp(-t^2/(4\sigma^2)),\ \exp(-\frac{t}{2M})\big\} \tag{C.11}$$

As noted by Vershynin [143], the same bound also holds under the assumption that each $Y^{(i)}$ is sub-exponential with parameter $M = \|Y^{(i)}\|_{\psi_1}$. Here we are using the Orlicz norm

$$\|Z\|_{\psi_1} := \inf\{t > 0 \mid \mathbb{E}[\psi(|Z|/t)] < \infty\},$$

defined by the function $\psi_1(x) = \exp(x) - 1$, as is appropriate for sub-exponential variables (e.g., see the book [78]).

Appendix D

Proofs for Chapter 6

D.1 Auxiliary results for Theorem 6.1

In this appendix, we provide the proofs of various auxiliary lemmas required in the proof of Theorem 6.1.

D.1.1 Proof of Lemma 6.1

Since θ^t and $\widehat{\theta}$ are both feasible and $\widehat{\theta}$ lies on the constraint boundary, we have $\mathcal{R}(\theta^t) \leq \mathcal{R}(\widehat{\theta})$. Since $\mathcal{R}(\widehat{\theta}) \leq \mathcal{R}(\theta^*) + \mathcal{R}(\widehat{\theta} - \theta^*)$ by triangle inequality, we conclude that

$$\mathcal{R}(\theta^t) \leq \mathcal{R}(\theta^*) + \mathcal{R}(\Delta^*).$$

Since $\theta^* = \Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*)$, a second application of triangle inequality yields

$$\mathcal{R}(\theta^t) \leq \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*). \tag{D.1}$$

Now define the difference $\Delta^t := \theta^t - \theta^*$. (Note that this is slightly different from $\widehat{\Delta}^t$, which is measured relative to the optimum $\widehat{\theta}$.) With this notation, we have

$$\begin{aligned}
\mathcal{R}(\theta^t) &= \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\bar{\mathcal{M}}}(\Delta^t) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) \\
&\overset{(i)}{\geq} \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) - \mathcal{R}\big(\Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\bar{\mathcal{M}}}(\Delta^t)\big) \\
&\overset{(ii)}{\geq} \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)),
\end{aligned}$$

where steps (i) and (ii) each use the triangle inequality. Now by the decomposability condition, we have $\mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) = \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t))$, so that we have shown that

$$\mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) \leq \mathcal{R}(\theta^t).$$

Combining this inequality with the earlier bound (D.1) yields

$$\mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) \leq \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*).$$

Re-arranging yields the inequality

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) \leq \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*). \tag{D.2}$$

The final step is to translate this inequality into one that applies to the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$. Recalling that $\Delta^* = \widehat{\theta} - \theta^*$, we have $\widehat{\Delta}^t = \Delta^t - \Delta^*$, and hence

$$\mathcal{R}(\widehat{\Delta}^t) \leq \mathcal{R}(\Delta^t) + \mathcal{R}(\Delta^*), \qquad \text{by triangle inequality.} \tag{D.3}$$

In addition, we have

$$\begin{aligned}\mathcal{R}(\Delta^t) \;\leq\; \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) &\overset{(i)}{\leq} 2\,\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*) \\ &\overset{(ii)}{\leq} 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*),\end{aligned}$$

where inequality (i) uses the bound (D.2), and inequality (ii) uses the definition (3.3) of the subspace compatibility Ψ. Combining with the inequality (D.3) yields

$$\mathcal{R}(\widehat{\Delta}^t) \leq 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\mathcal{R}(\Delta^*).$$

Since projection onto a subspace is non-expansive, we have $\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| \leq \|\Delta^t\|$, and hence

$$\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| \leq \|\widehat{\Delta}^t + \Delta^*\| \;\leq\; \|\widehat{\Delta}^t\| + \|\Delta^*\|.$$

Combining the pieces, we obtain the claim (6.44).

D.1.2 Proof of Lemma 6.2

We start by applying the RSC assumption to the pair $\widehat{\theta}$ and θ^t, thereby obtaining the lower bound

$$\begin{aligned}\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 &\geq \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t), \widehat{\theta} - \theta^t\rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\ &= \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t), \theta^{t+1} - \theta^t\rangle + \langle \nabla\mathcal{L}_n(\theta^t), \widehat{\theta} - \theta^{t+1}\rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}).\end{aligned} \tag{D.4}$$

Here the second inequality follows by adding and subtracting terms.

Now for compactness in notation, define $\varphi_t(\theta) := \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t), \theta - \theta^t\rangle + \frac{\gamma_u}{2}\|\theta - \theta^t\|^2$, and note that by definition of the algorithm, the iterate θ^{t+1} minimizes $\varphi_t(\theta)$ over the ball $\mathbb{B}_{\mathcal{R}}(\rho)$. Moreover, since $\widehat{\theta}$ is feasible, the first-order conditions for optimality imply that

$\langle \nabla \varphi_t(\theta^{t+1}), \widehat{\theta} - \theta^{t+1} \rangle \geq 0$, or equivalently that $\langle \nabla \mathcal{L}_n(\theta^t) + \gamma_u(\theta^{t+1} - \theta^t), \widehat{\theta} - \theta^{t+1} \rangle \geq 0$. Applying this inequality to the lower bound (D.4), we find that

$$\begin{aligned}
\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 &\geq \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t), \theta^{t+1} - \theta^t \rangle + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^{t+1} \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\
&= \varphi_t(\theta^{t+1}) - \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^{t+1} \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\
&= \varphi_t(\theta^{t+1}) + \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^t \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}),
\end{aligned} \tag{D.5}$$

where the last step follows from adding and subtracting θ^{t+1} in the inner product.

Now by the RSM condition, we have

$$\varphi_t(\theta^{t+1}) \geq \mathcal{L}_n(\theta^{t+1}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t) \overset{(a)}{\geq} \mathcal{L}_n(\widehat{\theta}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t), \tag{D.6}$$

where inequality (a) follows by the optimality of $\widehat{\theta}$, and feasibility of θ^{t+1}. Combining this inequality with the previous bound (D.5) yields that $\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2$ is lower bounded by

$$\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^t \rangle - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\theta^t - \widehat{\theta}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t),$$

and the claim (6.46) follows after some simple algebraic manipulations.

D.2 Auxiliary results for Theorem 6.2

In this appendix, we prove the two auxiliary lemmas required in the proof of Theorem 6.2.

D.2.1 Proof of Lemma 6.3

This result is a generalization of Lemma 3.17, with some changes required so as to adapt the statement to the optimization setting. Let θ be any vector, feasible for the problem (6.2), that satisfies the bound

$$\phi(\theta) \leq \phi(\theta^*) + \bar{\eta}, \tag{D.7}$$

and assume that $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}_n(\theta^*))$. We then claim that the error vector $\Delta := \theta - \theta^*$ satisfies the inequality

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)) \;\leq\; 3\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\Big\{\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\Big\}. \tag{D.8}$$

For the moment, we take this claim as given, returning later to verify its validity.

By applying this intermediate claim (D.8) in two different ways, we can complete the proof of Lemma 6.3. First, we observe that when $\theta = \widehat{\theta}$, the optimality of $\widehat{\theta}$ and feasibility of θ^* imply that assumption (D.7) holds with $\bar{\eta} = 0$, and hence the intermediate claim (D.8) implies that the statistical error $\Delta^* = \theta^* - \widehat{\theta}$ satisfies the bound

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \;\leq\; 3\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)). \tag{D.9}$$

Since $\Delta^* = \Pi_{\bar{\mathcal{M}}}(\Delta^*)\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)$, we can write

$$\mathcal{R}(\Delta^*) = \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \;\leq\; 4\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)), \tag{D.10}$$

using the triangle inequality in conjunction with our earlier bound (D.9). Similarly, when $\theta = \theta^t$ for some $t \geq T$, then the given assumptions imply that condition (D.7) holds with $\bar{\eta} > 0$, so that the intermediate claim (followed by the same argument with triangle inequality) implies that the error $\Delta^t = \theta^t - \theta^*$ satisfies the bound

$$\mathcal{R}(\Delta^t) \;\leq\; 4\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\big\}. \tag{D.11}$$

Now let $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ be the optimization error at time t, and observe that we have the decomposition $\widehat{\Delta}^t = \Delta^t + \Delta^*$. Consequently, by triangle inequality

$$\begin{aligned}
\mathcal{R}(\widehat{\Delta}^t) &\leq \mathcal{R}(\Delta^t) + \mathcal{R}(\Delta^*) \\
&\overset{(i)}{\leq} 4\Big\{\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*))\Big\} + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n},\, \bar{\rho}\big\} \\
&\overset{(ii)}{\leq} 4\Psi(\overline{\mathcal{M}})\,\Big\{\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| + \|\Pi_{\bar{\mathcal{M}}}(\Delta^*)\|\Big\} + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n},\, \bar{\rho}\big\} \\
&\overset{(iii)}{\leq} 4\Psi(\overline{\mathcal{M}})\,\Big\{\|\Delta^t\| + \|\Delta^*\|\Big\} + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n},\, \bar{\rho}\big\},
\end{aligned} \tag{D.12}$$

where step (i) follows by applying both equation (D.10) and (D.11); step (ii) follows from the definition (3.3) of the subspace compatibility that relates the regularizer to the norm $\|\cdot\|$; and step (iii) follows from the fact that projection onto a subspace is non-expansive. Finally, since $\Delta^t = \widehat{\Delta}^t - \Delta^*$, the triangle inequality implies that $\|\Delta^t\| \leq \|\widehat{\Delta}^t\| + \|\Delta^*\|$. Substituting this upper bound into inequality (D.12) completes the proof of Lemma 6.3.

It remains to prove the intermediate claim (D.8). Letting θ be any vector, feasible for the program (6.2), and satisfying the condition (D.7), and let $\Delta = \theta - \theta^*$ be the associated error vector. Re-writing the condition (D.7), we have

$$\mathcal{L}_n(\theta^* + \Delta) + \lambda_n\mathcal{R}(\theta^* + \Delta) \;\leq\; \mathcal{L}_n(\theta^*) + \lambda_n\mathcal{R}(\theta^*) + \bar{\eta}.$$

Subtracting $\langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle$ from each side and then re-arranging yields the inequality

$$\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle + \lambda_n \Big\{ \mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \Big\} \leq -\langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle + \bar{\eta}.$$

The convexity of $\mathcal{L}_n$ then implies that $\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle \geq 0$, and hence that

$$\lambda_n \Big\{ \mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \Big\} \leq -\langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle + \bar{\eta}.$$

Applying Hölder's inequality to $\langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle$, as expressed in terms of the dual norms $\mathcal{R}$ and $\mathcal{R}^*$, yields the upper bound

$$\lambda_n \Big\{ \mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \Big\} \leq \mathcal{R}^*(\nabla \mathcal{L}_n(\theta^*)) \, \mathcal{R}(\Delta) + \bar{\eta} \overset{(i)}{\leq} \frac{\lambda_n}{2} \mathcal{R}(\Delta) + \bar{\eta},$$

where step (i) uses the fact that $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}_n(\theta^*))$ by assumption.

For the remainder of the proof, let us introduce the convenient shorthand $\Delta_{\bar{\mathcal{M}}} := \Pi_{\bar{\mathcal{M}}}(\Delta)$ and $\Delta_{\bar{\mathcal{M}}^\perp} := \Pi_{\bar{\mathcal{M}}^\perp}(\Delta)$, with similar shorthand for projections involving θ^*. Making note of the decomposition $\Delta = \Delta_{\bar{\mathcal{M}}} + \Delta_{\bar{\mathcal{M}}^\perp}$, an application of triangle inequality then yields the upper bound

$$\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \leq \frac{1}{2} \Big\{ \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \Big\} + \frac{\bar{\eta}}{\lambda_n}, \tag{D.13}$$

where we have rescaled both sides by $\lambda_n > 0$.

It remains to further lower bound the left-hand side (D.13). By triangle inequality, we have

$$-\mathcal{R}(\theta^*) \geq -\mathcal{R}(\theta^*_{\mathcal{M}}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp}). \tag{D.14}$$

Let us now write $\theta^* + \Delta = \theta^*_{\mathcal{M}} + \theta^*_{\mathcal{M}^\perp} + \Delta_{\bar{\mathcal{M}}} + \Delta_{\bar{\mathcal{M}}^\perp}$. Using this representation and triangle inequality, we have

$$\mathcal{R}(\theta^* + \Delta) \geq \mathcal{R}(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp} + \Delta_{\bar{\mathcal{M}}}) \geq \mathcal{R}(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}).$$

Finally, since $\theta^*_{\mathcal{M}} \in \mathcal{M}$ and $\Delta_{\bar{\mathcal{M}}^\perp} \in \overline{\mathcal{M}}^\perp$, the decomposability of $\mathcal{R}$ implies that $\mathcal{R}(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) = \mathcal{R}(\theta^*_{\mathcal{M}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp})$, and hence that

$$\mathcal{R}(\theta^* + \Delta) \geq \mathcal{R}(\theta^*_{\mathcal{M}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}). \tag{D.15}$$

Adding together equations (D.14) and (D.15), we obtain the lower bound

$$\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \geq \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - 2\mathcal{R}(\theta^*_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}). \tag{D.16}$$

Combining this lower bound with the earlier inequality (D.13), some algebra yields the bound

$$\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \leq 3\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp}) + 2\frac{\eta}{\lambda_n},$$

corresponding to the bound (D.8) when η/λ_n achieves the final minimum. To obtain the final term involving $\bar{\rho}$ in the bound (D.8), two applications of triangle inequality yields

$$\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \leq \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + \mathcal{R}(\Delta) \;\leq\; \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + 2\bar{\rho},$$

where we have used the fact that $\mathcal{R}(\Delta) \leq \mathcal{R}(\theta) + \mathcal{R}(\theta^*) \leq 2\bar{\rho}$, since both θ and θ^* are feasible for the program (6.2).

D.2.2 Proof of Lemma 6.4

The proof of this result follows lines similar to the proof of convergence by Nesterov [102]. Recall our notation $\phi(\theta) = \mathcal{L}_n(\theta) + \lambda_n\mathcal{R}(\theta)$, $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$, and that $\eta^t_\phi = \phi(\theta^t) - \phi(\widehat{\theta})$. We begin by proving that under the stated conditions, a useful version of restricted strong convexity (6.42) is in force:

Lemma D.1. *Under the assumptions of Lemma 6.4, we are guaranteed that*

$$\Big\{\frac{\gamma_\ell}{2} - 32\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\Big\}\|\widehat{\Delta}^t\|^2 \leq 2\tau_\ell(\mathcal{L}_n)\, v^2 \;+\; \phi(\theta^t) - \phi(\widehat{\theta}), \quad \text{and} \tag{D.17a}$$

$$\Big\{\frac{\gamma_\ell}{2} - 32\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\Big\}\|\widehat{\Delta}^t\|^2 \leq 2\,\tau_\ell(\mathcal{L}_n)\, v^2 \;+\; \mathcal{T}_{\mathcal{L}}(\widehat{\theta}; \theta^t), \tag{D.17b}$$

where $v := \bar{\epsilon}_{stat} + 2\min(\frac{\bar{\eta}}{\lambda_n}, \bar{\rho})$.

See Appendix D.2.3 for the proof of this claim. So as to ease notation in the remainder of the proof, let us introduce the shorthand

$$\phi_t(\theta) := \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \theta - \theta^t\big\rangle + \frac{\gamma_u}{2}\|\theta - \theta^t\|^2 + \lambda_n\mathcal{R}(\theta), \tag{D.18}$$

corresponding to the approximation to the regularized loss function ϕ that is minimized at iteration t of the update (6.4). Since θ^{t+1} minimizes ϕ_t over the set $\mathbb{B}_{\mathcal{R}}(\bar{\rho})$, we are guaranteed that $\phi_t(\theta^{t+1}) \leq \phi_t(\theta)$ for all $\theta \in \mathbb{B}_{\mathcal{R}}(\bar{\rho})$. In particular, for any $\alpha \in (0,1)$, the vector $\theta_\alpha = \alpha\widehat{\theta} + (1-\alpha)\theta^t$ lies in the convex set $\mathbb{B}_{\mathcal{R}}(\bar{\rho})$, so that

$$\begin{aligned}
\phi_t(\theta^{t+1}) \;\leq\; \phi_t(\theta_\alpha) &= \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \theta_\alpha - \theta^t\big\rangle + \frac{\gamma_u}{2}\|\theta_\alpha - \theta^t\|^2 + \lambda_n\mathcal{R}(\theta_\alpha) \\
&\overset{(i)}{=} \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \alpha\widehat{\theta} - \alpha\theta^t\big\rangle + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2 + \lambda_n\mathcal{R}(\theta_\alpha) \\
&\overset{(ii)}{\leq} \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \alpha\widehat{\theta} - \alpha\theta^t\big\rangle + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2 + \lambda_n\alpha\mathcal{R}(\widehat{\theta}) + \lambda_n(1-\alpha)\mathcal{R}(\theta^t),
\end{aligned}$$

where step (i) follows from substituting the definition of θ_α, and step (ii) uses the convexity of the regularizer $\mathcal{R}$.

Now, the stated conditions of the lemma ensure that $\gamma_\ell/2 - 32\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}}) \geq 0$, so that by equation (D.17b), we have $\mathcal{L}_n(\widehat{\theta}) + 2\tau_\ell(\mathcal{L}_n)v^2 \geq \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t), \widehat{\theta} - \theta^t \rangle$. Substituting back into our earlier bound yields

$$\begin{aligned}\phi_t(\theta^{t+1}) &\leq (1-\alpha)\mathcal{L}_n(\theta^t) + \alpha\mathcal{L}_n(\widehat{\theta}) + 2\alpha\tau_\ell(\mathcal{L}_n)v^2 + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2 + \alpha\lambda_n\mathcal{R}(\widehat{\theta}) + (1-\alpha)\lambda_n\mathcal{R}(\theta^t) \\ &\overset{(iii)}{=} \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + 2\tau_\ell(\mathcal{L}_n)v^2 + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2, \end{aligned} \tag{D.19}$$

where we have used the definition of ϕ and $\alpha \leq 1$ in step (iii).

In order to complete the proof, it remains to relate $\phi_t(\theta^{t+1})$ to $\phi(\theta^{t+1})$, which can be performed by exploiting restricted smoothness. In particular, applying the RSM condition at the iterate θ^{t+1} in the direction θ^t yields the upper bound

$$\mathcal{L}_n(\theta^{t+1}) \leq \mathcal{L}_n(\theta^t) + \langle \mathcal{L}_n(\theta^t), \theta^{t+1} - \theta^t \rangle + \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t),$$

so that

$$\begin{aligned}\phi(\theta^{t+1}) &\leq \mathcal{L}_n(\theta^t) + \langle \mathcal{L}_n(\theta^t), \theta^{t+1} - \theta^t \rangle + \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t) + \lambda_n\mathcal{R}(\theta^{t+1}) \\ &= \phi_t(\theta^{t+1}) + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t).\end{aligned}$$

Combining the above bound with the inequality (D.19) and recalling the notation $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$, we obtain

$$\begin{aligned}\phi(\theta^{t+1}) &\leq \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t) + 2\tau_\ell(\mathcal{L}_n)v^2 \\ &\overset{(iv)}{\leq} \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u\alpha^2}{2}\|\widehat{\Delta}^t\|^2 + \tau_u(\mathcal{L}_n)[\mathcal{R}(\widehat{\Delta}^{t+1}) + \mathcal{R}(\widehat{\Delta}^t)]^2 + 2\tau_\ell(\mathcal{L}_n)v^2 \\ &\overset{(v)}{\leq} \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u\alpha^2}{2}\|\widehat{\Delta}^t\|^2 + 2\tau_u(\mathcal{L}_n)(\mathcal{R}^2(\widehat{\Delta}^{t+1}) + \mathcal{R}^2(\widehat{\Delta}^t)) + 2\tau_\ell(\mathcal{L}_n)v^2.\end{aligned} \tag{D.20}$$

Here step (iv) uses the fact that $\theta^t - \theta^{t+1} = \widehat{\Delta}^t - \widehat{\Delta}^{t+1}$ and applies triangle inequality to the norm $\mathcal{R}$, whereas step (v) follows from Cauchy-Schwarz inequality.

Next, combining Lemma 6.3 with the Cauchy-Schwarz inequality inequality yields the upper bound

$$\mathcal{R}^2(\widehat{\Delta}^t) \ \leq 32\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^t\|^2 + 2v^2 \tag{D.21}$$

where $v = \bar{\epsilon}_{\text{stat}}(\mathcal{M}, \overline{\mathcal{M}}) + 2\min(\frac{\bar{\eta}}{\lambda_n}, \bar{\rho})$, is a constant independent of θ^t and $\bar{\epsilon}_{\text{stat}}(\mathcal{M}, \overline{\mathcal{M}})$ was previously defined in the lemma statement. Substituting the above bound into inequal-

ity (D.20) yields that $\phi(\theta^{t+1})$ is at most

$$\phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u \alpha^2}{2}\|\widehat{\Delta}^t\|^2 + 64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^{t+1}\|^2 + 64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^t\|^2 + 8\tau_u(\mathcal{L}_n)v^2 + 2\tau_\ell(\mathcal{L}_n)v^2. \quad \text{(D.22)}$$

The final step is to translate quantities involving $\widehat{\Delta}^t$ to functional values, which may be done using the RSC condition (D.17a) from Lemma D.1. In particular, combining the RSC condition (D.17a) with the inequality (D.22) yields

$$\phi(\theta^{t+1}) \leq \phi(\theta^t) - \alpha\eta_\phi^t + \frac{\left(\gamma_u\alpha^2 + 64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\right)}{\overline{\gamma_\ell}}(\eta_\phi^t + 2\tau_\ell(\mathcal{L}_n)v^2) + \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}(\eta_\phi^{t+1} + 2\tau_\ell(\mathcal{L}_n)v^2) + 8\tau_u(\mathcal{L}_n)v^2 + 2\tau_\ell(\mathcal{L}_n)v^2.$$

where we have introduced the shorthand $\overline{\gamma_\ell} := \gamma_\ell - 64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})$. Recalling the definition of β, adding and subtracting $\phi(\widehat{\theta})$ from both sides, and choosing $\alpha = \frac{\overline{\gamma_\ell}}{2\gamma_u} \in (0,1)$, we obtain

$$\left(1 - \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\eta_\phi^{t+1} \leq \left(1 - \frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\eta_\phi^t + \beta(\overline{\mathcal{M}})v^2.$$

Recalling the definition of the contraction factor κ from the statement of Theorem 6.2, the above expression can be rewritten as

$$\eta_\phi^{t+1} \leq \kappa\eta_\phi^t + \beta(\overline{\mathcal{M}})\xi(\overline{\mathcal{M}})v^2, \quad \text{where } \xi(\mathcal{M}) = \left\{1 - \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right\}^{-1}.$$

Finally, iterating the above expression yields $\eta_\phi^t \leq \kappa^{t-T}\eta_\phi^T + \frac{\xi(\overline{\mathcal{M}})\beta(\overline{\mathcal{M}})v^2}{1-\kappa}$, where we have used the condition $\kappa \in (0,1)$ in order to sum the geometric series, thereby completing the proof.

D.2.3 Proof of Lemma D.1

The key idea to prove the lemma is to use the definition of RSC along with the iterated cone bound of Lemma 6.3 for simplifying the error terms in RSC.

Let us first show that condition (D.17a) holds. From the RSC condition assumed in the lemma statement, we have

$$\mathcal{L}_n(\theta^t) - \mathcal{L}_n(\widehat{\theta}) - \langle \nabla\mathcal{L}_n(\widehat{\theta}), \theta^t - \widehat{\theta}\rangle \geq \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\widehat{\theta} - \theta^t). \quad \text{(D.23)}$$

From the convexity of $\mathcal{R}$ and definition of the subdifferential $\partial\mathcal{R}(\theta)$, we obtain

$$\mathcal{R}(\theta^t) - \mathcal{R}(\widehat{\theta}) - \left\langle \partial\mathcal{R}(\widehat{\theta}), \theta^t - \widehat{\theta}\right\rangle \geq 0.$$

Adding this lower bound with the inequality (D.23) yields

$$\phi(\theta^t) - \phi(\widehat{\theta}) - \langle \nabla\phi(\widehat{\theta}), \theta^t - \widehat{\theta}\rangle \geq \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\widehat{\theta} - \theta^t),$$

where we recall that $\phi(\theta) = \mathcal{L}_n(\theta) + \lambda_n\mathcal{R}(\theta)$ is our objective function. By the optimality of $\widehat{\theta}$ and feasibility of θ^t, we are guaranteed that $\langle \nabla\phi(\widehat{\theta}), \theta^t - \widehat{\theta}\rangle \geq 0$, and hence

$$\begin{aligned}\phi(\theta^t) - \phi(\widehat{\theta}) &\geq \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\widehat{\theta} - \theta^t) \\ &\overset{(i)}{\geq} \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\;\left\{32\Psi^2(\overline{\mathcal{M}})\|\widehat{\theta} - \theta^t\|^2 + 2v^2\right\}\end{aligned}$$

where step (i) follows by applying Lemma 6.3. Some algebra then yields the claim (D.17a).

Finally, let us verify the claim (D.17b). Using the RSC condition, we have

$$\mathcal{L}_n(\widehat{\theta}) - \mathcal{L}_n(\theta^t) - \langle \nabla\mathcal{L}_n(\theta^t), \widehat{\theta} - \theta^t\rangle \geq \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\widehat{\theta} - \theta^t). \tag{D.24}$$

As before, applying Lemma 6.3 yields

$$\underbrace{\mathcal{L}_n(\widehat{\theta}) - \mathcal{L}_n(\theta^t) - \langle \nabla\mathcal{L}_n(\theta^t), \widehat{\theta} - \theta^t\rangle}_{\mathcal{T}_{\mathcal{L}}(\widehat{\theta};\theta^t)} \geq \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\;\left(32\Psi^2(\overline{\mathcal{M}})\|\widehat{\theta} - \theta^t\|^2 + 2v^2\right),$$

and rearranging the terms and establishes the claim (D.17b).

D.3 Proof of Lemma 6.5

Given the condition $\mathcal{R}(\widehat{\theta}) \leq \rho \leq \mathcal{R}(\theta^*)$, we have $\mathcal{R}(\widehat{\theta}) = \mathcal{R}(\theta^* + \Delta^*) \leq \mathcal{R}(\theta^*)$. By triangle inequality, we have

$$\mathcal{R}(\theta^*) = \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*)) \;\leq\; \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)).$$

We then write

$$\begin{aligned}\mathcal{R}(\theta^* + \Delta^*) &= \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\bar{\mathcal{M}}}(\Delta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \\ &\overset{(i)}{\geq} \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(ii)}{=} \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),\end{aligned}$$

where the bound (i) follows by triangle inequality, and step (ii) uses the decomposability of $\mathcal{R}$ over the pair $\mathcal{M}$ and $\overline{\mathcal{M}}^\perp$. By combining this lower bound with the previously established upper bound

$$\mathcal{R}(\theta^* + \Delta^*) \leq \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),$$

we conclude that $\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \leq \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))$. Finally, by triangle inequality, we have $\mathcal{R}(\Delta^*) \leq \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*))$, and hence

$$\begin{aligned}\mathcal{R}(\Delta^*) &\leq 2\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(i)}{\leq} 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Pi_{\bar{\mathcal{M}}}(\Delta^*)\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(ii)}{\leq} 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Delta^*\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),\end{aligned}$$

where inequality (i) follows from Definition 3.3 of the subspace compatibility Ψ, and the bound (ii) follows from non-expansivity of projection onto a subspace.

D.4 A general result on Gaussian observation operators

In this appendix, we state a general result about a Gaussian random matrices, and show how it can be adapted to prove Lemmas 6.6 and 6.7. Let $X \in \mathbb{R}^{n \times d}$ be a Gaussian random matrix with i.i.d. rows $x_i \sim N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$ is a covariance matrix. We refer to X as a sample from the Σ-Gaussian ensemble. In order to state the result, we use $\Sigma^{1/2}$ to denote the symmetric matrix square root.

Proposition D.1. *Given a random matrix X drawn from the Σ-Gaussian ensemble, there are universal constants c_i, $i = 0, 1$ such that*

$$\frac{\|X\theta\|_2^2}{n} \geq \frac{1}{2}\|\Sigma^{1/2}\theta\|_2^2 - c_1\frac{(\mathbb{E}[\mathcal{R}^*(x_i)])^2}{n}\mathcal{R}^2(\theta) \qquad \text{and} \tag{D.25a}$$

$$\frac{\|X\theta\|_2^2}{n} \leq 2\|\Sigma^{1/2}\theta\|_2^2 + c_1\frac{(\mathbb{E}[\mathcal{R}^*(x_i)])^2}{n}\mathcal{R}^2(\theta) \qquad \text{for all } \theta \in \mathbb{R}^d \tag{D.25b}$$

with probability greater than $1 - \exp(-c_0\, n)$.

We omit the proof of this result. The two special instances proved in Lemma 6.6 and 6.7 have been stated in the paper [109] and in Proposition 4.1 respectively. We now show how Proposition D.1 can be used to recover various lemmas required in our proofs.

Proof of Lemma 6.6: We begin by establishing this auxiliary result required in the proof of Corollary 6.2. When $\mathcal{R}(\cdot) = \|\cdot\|_1$, we have $\mathcal{R}^*(\cdot) = \|\cdot\|_\infty$. Moreover, the random vector $x_i \sim N(0, \Sigma)$ can be written as $x_i = \Sigma^{1/2}w$, where $w \sim N(0, I_{d\times d})$ is standard normal. Consequently, using properties of Gaussian maxima [78] and defining $\zeta(\Sigma) = \max_{j=1,2,\ldots,d}\Sigma_{jj}$, we have the bound

$$(\mathbb{E}[\|x_i\|_\infty])^2 \;\leq\; \zeta(\Sigma)\;(\mathbb{E}[\|w\|_\infty])^2 \;\leq\; 3\zeta(\Sigma)\;\sqrt{\log d}.$$

Substituting into Proposition D.1 yields the claims (6.56a) and (6.56b).

Proof of Lemma 6.7: In order to prove this claim, we view each random observation matrix $X_i \in \mathbb{R}^{m \times m}$ as a $d = m^2$ vector (namely the quantity $\text{vec}(X_i)$), and apply Proposition D.1 in this vectorized setting. Given the standard Gaussian vector $w \in \mathbb{R}^{m^2}$, we let $W \in \mathbb{R}^{m \times m}$ be the random matrix such that $\text{vec}(W) = w$. With this notation, the term $\mathcal{R}^*(\text{vec}(X_i))$ is equivalent to the operator norm $|||X_i|||_2$. As shown in Chapter 4, $\mathbb{E}[|||X_i|||_2] \leq 24\zeta_{\text{mat}}(\Sigma)\sqrt{m}$, where ζ_{mat} was previously defined (6.59).

D.5 Auxiliary results for Corollary 6.5

In this section, we provide the proofs of Lemmas 6.8 and 6.9 that play a central role in the proof of Corollary 6.5. In order to do so, we require the following result, which is a re-statement of Theorem 5.1:

Proposition D.2. *For the matrix completion operator $\mathfrak{X}_n$, there are universal positive constants (c_1, c_2) such that*

$$\left| \frac{\|\mathfrak{X}_n(\Theta)\|_2^2}{n} - |||\Theta|||_F^2 \right| \leq c_1 \, m \|\Theta\|_\infty \, |||\Theta|||_{\text{nuc}} \sqrt{\frac{m \log m}{n}} \; + \; c_2 \left(m \|\Theta\|_\infty \sqrt{\frac{m \log m}{n}} \right)^2 \qquad \text{for all } \Theta \in \mathbb{R}^{m \times m} \tag{D.26}$$

with probability at least $1 - \exp(-m \log m)$.

D.5.1 Proof of Lemma 6.8

Applying Proposition D.2 to $\widehat{\Delta}^t$ and using the fact that $m\|\widehat{\Delta}^t\|_\infty \leq 2\alpha$ yields

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq |||\widehat{\Delta}^t|||_F^2 - c_1 \alpha |||\widehat{\Delta}^t|||_{\text{nuc}} \sqrt{\frac{m \log m}{n}} - c_2 \, \alpha^2 \frac{m \log m}{n}, \tag{D.27}$$

where we recall our convention of allowing the constants to change from line to line. From Lemma 6.1,

$$|||\widehat{\Delta}^t|||_{\text{nuc}} \leq 2\,\Psi(\overline{\mathcal{M}}^\perp)\, |||\widehat{\Delta}^t|||_F + 2|||\Pi_{\mathcal{M}^\perp}(\theta^*)|||_{\text{nuc}} + 2|||\Delta^*|||_{\text{nuc}} + \Psi(\overline{\mathcal{M}}^\perp)|||\Delta^*|||_F.$$

Since $\rho \leq |||\Theta^*|||_{\text{nuc}}$, Lemma 6.5 implies that $|||\Delta^*|||_{\text{nuc}} \leq 2\Psi(\overline{\mathcal{M}}^\perp)|||\Delta^*|||_F + |||\Pi_{\mathcal{M}^\perp}(\theta^*)|||_{\text{nuc}}$, and hence that

$$|||\widehat{\Delta}^t|||_{\text{nuc}} \leq 2\,\Psi(\overline{\mathcal{M}}^\perp)\, |||\widehat{\Delta}^t|||_F + 4|||\Pi_{\mathcal{M}^\perp}(\theta^*)|||_{\text{nuc}} + 5\Psi(\overline{\mathcal{M}}^\perp)|||\Delta^*|||_F. \tag{D.28}$$

Combined with the lower bound, we obtain that $\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n}$ is lower bounded by

$$|||\widehat{\Delta}^t|||_F^2 \left\{ 1 - \frac{2c_1 \, \alpha \Psi(\overline{\mathcal{M}}^\perp)\sqrt{\frac{m \log m}{n}}}{|||\widehat{\Delta}^t|||_F} \right\} - 2c_1 \, \alpha \sqrt{\frac{m \log m}{n}} \left\{ 4|||\Pi_{\mathcal{M}^\perp}(\theta^*)|||_{\text{nuc}} + 5\Psi(\overline{\mathcal{M}}^\perp)|||\Delta^*|||_F \right\} - c_2 \, \alpha^2 \frac{m \log m}{n}$$

Consequently, for all iterations such that $\|\widehat{\Delta}^t\|_F \geq 4c_1\Psi(\overline{\mathcal{M}}^{\perp})\sqrt{\frac{m\log m}{n}}$, we have

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2}\|\widehat{\Delta}^t\|_F^2 - 2c_1\,\alpha\sqrt{\frac{m\log m}{n}}\Big\{4\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_{\text{nuc}} + 5\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|_F\Big\} - c_2\,\alpha^2\frac{m\log m}{n}.$$

By subtracting off an additional term, the bound is valid for all $\widehat{\Delta}^t$—viz.

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2}\|\widehat{\Delta}^t\|_F^2 - 2c_1\,\alpha\sqrt{\frac{m\log m}{n}}\Big\{4\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_{\text{nuc}} + 5\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|_F\Big\} - c_2\,\alpha^2\frac{m\log m}{n} - 16c_1^2\alpha^2\Psi^2(.$$

D.5.2 Proof of Lemma 6.9

Applying Proposition D.2 to Γ^t and using the fact that $m\|\Gamma^t\|_\infty \leq 2\alpha$ yields

$$\frac{\|\mathfrak{X}_n(\Gamma^t)\|_2^2}{n} \leq \|\Gamma^t\|_F^2 + c_1\alpha\|\Gamma^t\|_{\text{nuc}}\sqrt{\frac{m\log m}{n}} + c_2\,\alpha^2\frac{m\log m}{n}, \tag{D.29}$$

where we recall our convention of allowing the constants to change from line to line. By triangle inequality, we have $\|\Gamma^t\|_{\text{nuc}} \leq \|\Theta^t - \widehat{\Theta}\|_{\text{nuc}} + \|\Theta^{t+1} - \widehat{\Theta}\|_{\text{nuc}} = \|\widehat{\Delta}^t\|_{\text{nuc}} + \|\widehat{\Delta}^{t+1}\|_{\text{nuc}}$. Equation D.28 gives us bounds on $\|\widehat{\Delta}^t\|_{\text{nuc}}$ and $\|\widehat{\Delta}^{t+1}\|_{\text{nuc}}$. Substituting them into the upper bound (D.29) yields the claim.

Bibliography

[1] J. Abernethy, F. Bach, T. Evgeniou, and J. Stein. Low-rank matrix factorization with attributes. Technical Report Technical Report N-24/06/MM, Ecole des mines de Paris, France, September 2006.

[2] A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Annals of Statistics*, 2011. To appear.

[3] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, March 2002.

[4] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semdefinite relaxations for sparse principal component analysis. *Annals of Statistics*, 5B:2877–2921, 2009.

[5] C. W. Anderson, E. A. Stolz, and S. Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalogram during mental tasks. *IEEE Trans. on bio-medical engineering*, 45(3):277, 1998.

[6] T. W. Anderson. *The statistical analysis of time series*. Wiley Classics Library. John Wiley and Sons, New York, 1971.

[7] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2006.

[8] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[9] F. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, June 2008.

[10] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[11] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. Technical report, Rice University, 2008. Available at arxiv:0808.3572.

[12] P. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 2005. To appear.

[13] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[14] S. Becker, J. Bobin, and E. J. Candes. Nesta: a fast and accurate first-order method for sparse recovery. Technical report, Stanford University, 2009.

[15] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.

[16] P. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2008.

[17] P. Bickel and B. Li. Regularization in statistics. *TEST*, 15(2):271–344, 2006.

[18] P. J. Bickel and K. A. Doksum. *Mathematical statistics: basic ideas and selected topics*. Prentice Hall, Upper Saddle River, N.J., 2001.

[19] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

[20] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[21] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[22] K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14:813–837, 2008.

[23] E. N. Brown, R. E. Kass, and P. P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7(5), May 2004.

[24] F. Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

[25] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, pages 169–194, 2007.

[26] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for gaussian regression. *Annals of Statistics*, 35(4):1674–1697, 2007.

[27] F. Bunea, Y. She, and M. Wegkamp. Adaptive rank penalized estimators in multivariate regression. Technical report, Florida State, 2010. available at arXiv:1004.2995.

[28] T. Cai and H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. Technical report, Wharton School of Business, University of Pennsylvania, 2010. available at http://www-stat.wharton.upenn.edu/ tcai/paper/html/Sparse-Covariance-Matrix.html.

[29] E. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98 (6):925–936, 2010.

[30] E. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. Technical Report arXiv:1001.0339v1, Stanford, January 2010.

[31] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Info Theory*, 51(12):4203–4215, December 2005.

[32] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[33] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[34] E. J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[35] E. J. Candes, Y. Ma X. Li, and J. Wright. Stable principal component pursuit. In *International Symposium on Information Theory*, June 2010.

[36] E. J. Candés, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? *Journal of the ACM*, 58(3), May 2011.

[37] Jose M Carmena, Mikhail A Lebedev, Roy E Crist, Joseph E O'Doherty, David M Santucci, Dragan F Dimitrov, Parag G Patil, Craig S Henriquez, and Miguel A. L Nicolelis. Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biol*, 1(2):e42, 10 2003.

[38] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. Technical report, MIT, June 2009. Available at `arXiv:0906.2220v1`.

[39] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.

[40] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. of. American Mathematical Society*, 22(1):211–231, July 2008.

[41] M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Processing*, 46:886–902, April 1998.

[42] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pages 317–336. Elsevier, Amsterdam, NL, 2001.

[43] M. Deza and M. Laurent. *Geometry of Cuts and Metric Embeddings*. Springer-Verlag, New York, 1997.

[44] D. L. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, April 2006.

[45] D. L. Donoho and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, December 1995.

[46] D. L. Donoho and J. M. Tanner. Neighborliness of randomly-projected simplices in high dimensions. *Proceedings of the National Academy of Sciences*, 102(27):9452–9457, 2005.

[47] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International Conference on Machine Learning (ICML)*, 2008.

[48] R. Durrett. *Probability: Theory and Examples*. Duxbury Press, New York, NY, 1995.

[49] R. L. Dykstra. An iterative procedure for obtaining i-projections onto the intersection of convex sets. *Annals of Probability*, 13(3):975–984, 1985.

[50] J. Fan and R. Li. Variable selection via non-concave penalized likelihood and its oracle properties. *Jour. Amer. Stat. Ass.*, 96(456):1348–1360, December 2001.

[51] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford, 2002. Available online: http://faculty.washington.edu/mfazel/thesis-final.pdf.

[52] J. Fisher and M. J. Black. Motor cortical decoding using an autoregressive moving average model,. *IEEE Engineering in Medicine and Biology Society*, pages 1469–1472, September 2005.

[53] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 2007.

[54] R. Garg and R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, New York, NY, USA, 2009. ACM.

[55] V. L. Girko. *Statistical analysis of observations of increasing dimension*. Kluwer Academic, New York, 1995.

[56] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.

[57] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, March 2011.

[58] E. T. Hale, Y. Wotao, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM J. on Optimization*, 19(3):1107–1130, 2008.

[59] L. Harrison, W. D. Penny, and K. Friston. Multivariate autoregressive modeling of fmri time series. *NeuroImage*, 19:1477–1491, 2003.

[60] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[61] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991.

[62] D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, November 2011.

[63] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38 (4):1978–2004, 2010.

[64] L. Jacob, G. Obozinski, and J. P. Vert. Group Lasso with Overlap and Graph Lasso. In *International Conference on Machine Learning (ICML)*, pages 433–440, 2009.

[65] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. Technical report, HAL-Inria, 2010. available at inria-00516723.

[66] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *International Conference on Machine Learning (ICML)*, New York, NY, USA, 2009. ACM.

[67] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, April 2001.

[68] S. M. Kakade, O. Shamir, K. Sridharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *AISTATS*, 2010.

[69] N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics*, 36(6):2717–2756, 2008.

[70] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, June 2010.

[71] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, July 2010.

[72] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of COLT*, 2008.

[73] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Annals of Statistics*, 38:3660–3695, 2010.

[74] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39:2302–2329, 2011.

[75] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37:4254–4278, 2009.

[76] M. Laurent. Matrix completion problems. In *The Encyclopedia of Optimization*, pages 221—229. Kluwer Academic, 2001.

[77] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

[78] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

[79] K. Lee and Y. Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. Technical report, UIUC, 2009. Available at arXiv:0903.4742.

[80] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. Technical Report UILU-ENG-09-2214, Univ. Illinois, Urbana-Champaign, July 2009.

[81] H. Liu, J. Lafferty, and L. Wasserman. Nonparametric regression and classification with joint sparsity constraints. In *Neural Info. Proc. Systems (NIPS) 22*, Vancouver, Canada, December 2008.

[82] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm optimization with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

[83] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. Technical Report arXiv:0903.1468, ETH Zurich, March 2009.

[84] Z. Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46-47:157–178, 1993.

[85] M. Lustig, D. Donoho, J. Santos, and J. Pauly. Compressed sensing MRI. *IEEE Signal Processing Magazine*, 27:72–82, March 2008.

[86] H. Lütkepolhl. *New introduction to multiple time series analysis*. Springer, New York, 2006.

[87] P. Massart. *Concentration Inequalties and Model Selection*. Ecole d'Eté de Probabilités, Saint-Flour. Springer, New York, 2003.

[88] J. Matousek. *Lectures on discrete geometry*. Springer-Verlag, New York, 2002.

[89] R. Mazumber, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, August 2010.

[90] M. McCoy and J. Tropp. Two Proposals for Robust PCA using Semidefinite Programming. Technical report, California Institute of Technology, 2010. URL http://arxiv.org/pdf/1012.1086v3.

[91] M. L. Mehta. *Random matrices*. Academic Press, New York, NY, 1991.

[92] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71, 2008.

[93] L. Meier, S. van de Geer, and P. Buhlmann. High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821, 2009.

[94] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[95] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.

[96] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[97] S. Negahban and M. J. Wainwright. Simultaneous support recovery in high-dimensional regression: Benefits and perils of $\ell_{1,\infty}$-regularization. *IEEE Transactions on Information Theory*, 57(6):3481–3863, June 2011.

[98] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

[99] S. Negahban and M. J. Wainwright. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 2012. To appear; posted at http://arxiv.org/abs/1009.2118.

[100] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *NIPS Conference*, 2009. Full length version at http://arxiv.org/abs/1010.2731v1.

[101] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, New York, 2004.

[102] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 2007.

[103] H. V. Ngai and J. P. Penot. Paraconvex functions and paraconvex sets. *Studia Mathematica*, 184:1–29, 2008.

[104] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Union support recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39(1):1–47, January 2011.

[105] L. A. Pastur. On the spectrum of random matrices. *Theoretical and Mathematical Physics*, 10:67–74, 1972.

[106] D. Paul and I. Johnstone. Augmented sparse principal component analysis for high-dimensional data. Technical report, UC Davis, January 2008.

[107] G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, UK, 1989.

[108] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue conditions for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, August 2010.

[109] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Information Theory*, 57(10): 6976—6994, October 2011.

[110] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 12:389–427, March 2012.

[111] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation: Convergence rates of ℓ_1-regularized log-determinant divergence. Technical report, Department of Statistics, UC Berkeley, September 2008.

[112] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: sparse additive models. *Journal of the Royal Statistical Society, Series B*, 71(5):1009–1030, 2009.

[113] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

[114] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.*, 5:935–980, 2011.

[115] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[116] B. Recht, W. Xu, and B. Hassibi. Null space conditions and thresholds for rank minimization. Technical report, U. Madison, 2009. Available at http://pages.cs.wisc.edu/ brecht/papers/10.RecXuHas.Thresholds.pdf.

[117] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[118] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[119] A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. Technical Report arXiv:0912.5338v2, Universite de Paris, January 2010.

[120] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[121] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. Technical report, University of Michigan, July 2011.

[122] R. Salakhutdinov and N. Srebro. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. Technical Report abs/1002.2780v1, Toyota Institute of Technology, 2010.

[123] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, and W.R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, March 2002.

[124] N. Srebro. *Learning with Matrix Factorizations*. PhD thesis, MIT, 2004. Available online: http://ttic.uchicago.edu/ nati/Publications/thesis.pdf.

[125] N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2004.

[126] N. Srebro, N. Alon, and T. S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2005.

[127] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. In *Proceedings of the NIPS Conference*, Vancouver, Canada, 2005.

[128] J. L. Starck, E. J. Candès, and D. L. Donoho. Astronomical image representation by the curvelet transform. *Astronomy and Astrophysics*, 398:785–800, 2003.

[129] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Transactions on Signal Processing*, 57(8):3075–3085, 2009.

[130] S. J. Szarek. The finite dimensional basis problem with an appendix on nets of grassmann manifolds. *Acta Mathematica*, 151:153–179, 1983.

[131] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[132] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *J. R. Statistical Soc. B*, 67:91–108, 2005.

[133] J. Tropp. User-friendly tail bounds for matrix martingales. Technical report, Caltech, April 2010.

[134] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, December 2007.

[135] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. *Signal Processing*, 86:572–602, April 2006. Special issue on "Sparse approximations in signal and image processing".

[136] D. Tse and P. Viswanath. *Fundamentals of Wireless Communications*. Cambridge University Press, 2005.

[137] B. Turlach, W.N. Venables, and S.J. Wright. Simultaneous variable selection. *Technometrics*, 27:349–363, 2005.

[138] S. van de Geer. The deterministic lasso. In *Proc. of Joint Statistical Meeting*, 2007.

[139] S. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[140] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36:614–645, 2008.

[141] A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK, 1998.

[142] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38:49–95, 1996.

[143] R. Vershynin. A note on sums of independent random matrices after Ahlswede-Winter. Technical report, Univ. Michigan, December 2009.

[144] V. Q. Vu, Pradeep Ravikumar, Thomas Naselaris, Kendrick N. Kay, Jack L. Gallant, and Bin Yu. Encoding and decoding v1 fmri responses to natural images with sparse nonparametric models. *Annals of Applied Statistics*, 5(2B):1159–1182, 2011.

[145] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009.

[146] M. J. Wainwright. Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Info. Theory*, 55:5728–5741, December 2009.

[147] E. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62:548–564, 1955.

[148] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via Outlier Pursuit. Technical report, University of Texas, Austin, 2010. URL http://arxiv.org/pdf/1010.4237v2. available at arXiv:1010.4237.

[149] Kim Y., Kim J., and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16(2), 2006.

[150] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[151] B. Yu. Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, Berlin, 1997.

[152] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):49–67, 2006.

[153] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[154] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal Of The Royal Statistical Society Series B*, 69(3):329–346, 2007.

[155] C. H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[156] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

[157] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.

[158] S. Zhou, J. Lafferty, and L. Wasserman. Time-varying undirected graphs. In *21st Annual Conference on Learning Theory (COLT)*, Helsinki, Finland, July 2008.