High-dimensional statistics with systematically corrupted data

by

Po-Ling Loh

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

and the Designated Emphasis

in

Communication, Computation, and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martin Wainwright, Chair
Professor Laurent El Ghaoui
Professor Bin Yu

Spring 2014

High-dimensional statistics with systematically corrupted data

Abstract

High-dimensional statistics with systematically corrupted data

by

Po-Ling Loh

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Martin Wainwright, Chair

Noisy and missing data are prevalent in many real-world statistical estimation problems. Popular techniques for handling nonidealities in data, such as imputation and expectation-maximization, are often difficult to analyze theoretically and/or terminate in local optima of nonconvex functions—these problems are only exacerbated in high-dimensional settings. We present new methods for obtaining high-dimensional regression estimators in the presence of corrupted data, and provide theoretical guarantees for the statistical consistency of our methods. Although our estimators also arise as minima of nonconvex functions, we show the rather surprising result that all stationary points are clustered around a global minimum. We describe extensions of our work to nonconvex regularizers, and demonstrate that an adaptation of composite gradient descent may be used to compute a global optimum up to statistical precision in log-linear time. Finally, we show how our corrupted regression methods may be applied to structure estimation for undirected graphical models, even when data are observed with systematic corruption. We derive new relationships between augmented inverse covariance matrices and the edge structure of discrete-valued graphs, and combine our population-level results with corrupted estimation methods to create new algorithms for graph estimation. We close with theoretical results and preliminary simulations in the domain of compressed sensing MRI.

To Hana

Contents

Acknowledgments

The last five years of my PhD have been a journey. My journey has had its ups and downs and some very rough patches, but what has remained constant is the tremendous amount of love and support I have received from people I have encountered along the way. Before I delve into the technical content of my thesis, I want to thank some of those special people.

First, I thank my adviser, Martin Wainwright, for being the best type of adviser any grad student could wish for. Martin was extremely intentional in his mentorship, providing just the right amount of guidance to help me develop from a skittish grad student to an independent researcher. Thank you for believing in my potential and helping me to achieve it. Thank you for being a wonderful role model in your professional and personal relationships with other members of the academic community.

I also thank Bin Yu, who in addition to being a member of my dissertation committee, has provided me with much wisdom and advice during my time as a grad student. You are an inspiration to me. I thank Peter Bickel for taking me under his wing in the spring semester of my second year, when I was feeling ungrounded while Martin was away on sabbatical. I thank Miki Lustig for teaching me everything I know about MRI and proliferating his excitement about how compressed sensing can powerfully impact medical technology.

I have been extremely fortunate to forge deep friendships with people in the statistics, CS, and EE communities. In the statistics department, I especially thank Ngoc Tran and Christine Ho for our "misery parties" that ultimately resulted in much happiness. Thanks to Miki Racz for being my partner in crime during my fourth-year stint as SGSA co-president. In CS, I thank my cubemates from the RAD/AMPLab—Fabian Wauthier, Purna Sarkar, and Andre Wibisono, among others—with whom I have formed some of my fondest memories. I also thank Venkat Chandrasekaran for his mentorship and Steffi Jegelka for her friendship. A special thanks to Andrew Chan for his alacrity in troubleshooting all my technical problems throughout grad school. Finally, I thank my EE friends in Wi-Fo for their incredible kindness and hospitality when I moved to Cory during my last two years of grad school. Special thanks to Nihar Shah, Rashmi Korlakai Vinayak, Vasuki Narasimha Swamy, Venky Ekambaram, and of course Varun Jog, with whom I have shared much gossip and even more laughter. And a sincere thanks to my peers in Martin's group, especially my older academic brothers, Garvesh Raskutti, Nima Noorshams, Alekh Agarwal, and Sahand Negahban, for providing plentiful encouragement and useful advice at critical points in my grad career.

I also thank all the wonderful members of the Seminar für Statistik at ETH Zürich for their warm welcome when I visited Switzerland during the fall semester of my fifth year. I am immensely grateful to Peter Bühlmann for his hospitality and mentorship and for making my time at ETH so enjoyable. He is an incredible role model for everyone in our field.

Lastly, I thank my parents and brothers for being a steady source of advice and support. Thank you for bringing me to the beginning of this journey and seeing me all the way through!

Chapter 1

Introduction

Statistics is entering an exciting new era, as technology continues to propagate and society advances through the Information Age. Whereas scientific studies were previously limited by the cost or time required for data collection, modern technology allows massive datasets to be acquired cheaply and efficiently, shifting the focus of statistics to regimes where the number of measured variables is comparable to or exceeds the number of samples. From a computational perspective, it is important to find low-dimensional representations of high-dimensional data and filter through datasets in a more temporally and spatially efficient manner than directly processing all samples.

This thesis brings together several areas of statistics that involve new challenges arising in the field of high-dimensional settings. Scenarios where the number of parameters exceeds the number of observations involve intrinsic non-identifiability, which is overcome through appropriate assumptions. In the sections that follow, we outline some of the core problems and key contributions that will be developed in the remainder of this thesis.

1.1 High-dimensional inference

Throughout this thesis, we are concerned with statistical problems where the number of parameters *exceeds* the number of observations. Such *high-dimensional* problems differ from their low-dimensional analogs, in which the number of parameters is small (and fixed) and the number of observations grows to infinity. Since classical statistical theory focuses on characterizing the asymptotic behavior of estimators in low-dimensional settings, new theory must be derived to establish nonasymptotic results for high-dimensional estimators. In fact, estimators that are statistically consistent in low-dimensional settings may be ill-defined in high-dimensional problems, giving rise to an entire subspace of solutions rather than a unique estimator. One popular technique is to leverage known structure of the underlying parameter vector (such as sparsity) and incorporate it into a composite objective that trades off the prediction error of the estimate with its deviation from the ideal structure. The goal, from a statistical perspective, is to devise an appropriate estimator and then prove that it

achieves optimal rates of convergence among the class of models under consideration.

1.2 Systematically corrupted data

Another salient characteristic of many traditional statistical algorithms is the underlying assumption that observations are cleanly observed, independent, and identically distributed. What happens when such assumptions do not hold? Intuitively, systematic corruptions lead to systematic biases in inference, which still persist as the number of samples tends to infinity. Some corruption mechanisms of interest include additive noise and missing data, which were previously only studied in the context of low-dimensional problems. It is interesting to ask what can be said about high-dimensional statistical inference in the presence of systematically corrupted data—both in terms of devising natural estimators and establishing rates of convergence. We show that a simple variant of the Lasso for linear regression enjoys provably good behavior when the underlying parameter vector is sparse.

The methods we develop for high-dimensional linear regression have natural applications to compressed sensing, where the goal is to reduce the number of acquisitions and still accurately reconstruct a signal. In compressed sensing MRI, the number of samples is only required to scale as the logarithm of the overall dimensionality times the sparsity of the image in an appropriate wavelet basis. However, it is unrealistic to assume that data are acquired noiselessly: In addition to noise in the readout signal, nonidealities in the magnetic field may lead to systematic noise in the acquisition frequency, thereby creating a garbled image. We propose a variant of the corrected Lasso algorithm that is designed specifically for corrupted acquisitions in compressed sensing MRI and performs well in synthetic experiments.

1.3 Nonconvex optimization

On the algorithmic side, high-dimensional statistical inference gives rise to interesting families of objective functions that do not satisfy the canonical assumptions necessary for efficient optimization. Whereas low-dimensional problems often result in optimizing objective functions that are nicely smooth and strongly convex, their high-dimensional analogs generally only have positive curvature in a restricted set of directions. This necessitates the inclusion of a regularization function, which encourages solutions to lie in a lower-dimensional space within which the composite objective is well-behaved. Although it may still be possible to prove consistency of a global optimum, finding global optima may be exceedingly difficult in practice. This problem is exacerbated when the objective function possesses nonconvexity due to corrupted observations, yielding multiple local optima. Another parallel line of work involves using nonconvex regularization functions to reduce bias in estimated parameters. Although statistical consistency of global optima may again be established in a fairly straightforward manner, optimization algorithms are only guaranteed to locate local optima, for which theoretical guarantees do not exist.

We establish a general framework of sufficient conditions under which composite objective functions formed as a sum of a nonconvex loss and nonconvex regularizer are still tractable to standard optimization procedures. In particular, when the loss function satisfies a condition known as restricted strong convexity (RSC) and the penalty satisfies an upper bound on the level of nonconvexity, all local and global optima are guaranteed to lie within a small ball of the true parameter, where the radius of the ball is on the order of statistical precision. We also described how a variant of the composite gradient descent algorithm, typically only used to locate optima of strongly convex loss functions with convex penalties, may be used to efficiently obtain local optima within a small radius of the truth.

1.4 Graphical models

Graphical models are used to represent conditional independencies between variables in a joint distribution, where nodes represent variables and absent edges indicate conditional independence. In a high-dimensional setting, the goal is to infer the edge structure of a sparse graph based on samples from the joint distribution. Many theoretical results have been derived for consistent edge recovery when variables are jointly Gaussian; in practice, the same algorithms are often applied even when data are *not* Gaussian, and practitioners attempt to extract inferences from the output of the learning algorithm. When do algorithms such as the graphical Lasso yield meaningful results? Do efficient algorithms exist for edge recovery in highly non-Gaussian settings?

We show that when individual variables take states in a finite discrete alphabet, a fundamental connection still exists between generalized (augmented) inverse covariance matrices and the structure of the graph. Our result hinges on the theory of sufficient statistics in an exponential family representation of the graph, and constitutes a significant generalization of results on inverse covariances previously only known for Gaussians. In addition, we propose new methods for estimtating the edge structure of an arbitrary discrete-valued graphical model. Our methods are particularly attractive for graphs with bounded treewidth, such as trees, in which case a (group) graphical Lasso may be applied to the appropriate choice of sufficient statistics to recover the edges of the graph.

Our research has widespread implications in application domains where the theory of graphical models is used to learn relationships between individuals in a network. For instance, the goal of learning in social networks is to infer connections between individuals based on joint observations of their states. In computational biology, scientists wish to reconstruct gene networks based on joint measurements of gene expression levels. In neuroscience, researchers learn neural networks from measured brain activity. Our work on graphical model estimation in discrete graphs demonstrates that there is still hope for these learned networks to be meaningful even when the assumption of multivariate Gaussianity is not strictly satisfied.

1.5 Thesis overview

The remainder of the thesis is organized as follows. We begin in Chapter 2 with basic background material. In Chapter 3, we devise a modified Lasso estimator that may be used for sparse high-dimensional linear regression in the presence of corrupted observations, and derive statistical properties and optimization guarantees for the resulting estimator. We also present lower bounds based on information-theoretic arguments, which show that the modified Lasso estimator is minimax optimal. In Chapter 4, we expand our scope to more general classes of estimators, and establish sufficient conditions under which stationary points of nonconvex M-estimators with (possibly nonconvex) regularizers are statistically consistent. In Chapter 5, we show how our regression-based results, in conjunction with newly established connections between the edge structure of certain discrete-valued graphical models and the inverse covariance matrix of the augmented distribution, may be used to perform structural estimation even in the presence of systematically corrupted observations. Finally, we close in Chapter 6 with remarks and simulations about how our work may be applied in the context of compressed sensing MRI. Proofs of the more technical results are contained in the Appendices.

Chapter 2

Background

We devote this chapter to expository material introducing some of the basic statistical and optimization terminology to be used later in the thesis. Each chapter is self-contained, however, so we invite the reader to examine the introductory material of individual chapters for more detailed descriptions.

2.1 High-dimensional regression

The basic statistical model to be discussed in this thesis is as follows: Data pairs $\{(x_i, y_i)\}_{i=1}^n$ are generated according to a distribution

$$y_i \sim \mathbb{P}_{\beta^*}(\cdot \mid x_i), \tag{2.1}$$

where $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, and $\beta^* \in \mathbb{R}^d$ is an unknown regression vector. In the models we consider, we will generally take $d = p$. In most cases, we will assume that the pairs (x_i, y_i) are independent over $1 \leq i \leq n$, but we will always state explicitly whether or not this is the case.

We are primarily interested in *high-dimensional* models, in which we assume that the number of parameters p exceeds the number of observations n. Consequently, the parametric model (2.1) may be nonidentifiable without introducing further assumptions. We will assume that β^* is a *sparse* vector: Denoting by $\|\beta^*\|_0$ the number of nonzero entries of β^*, we assume $\|\beta^*\|_0 \leq k$ for some $k \leq n$. In many cases, this reduces the parameter space sufficiently in order to perform efficient statistical inference.

2.1.1 Examples

As a first example, we consider the problem of high-dimensional linear regression. The data-generating mechanism (2.1) is given by

$$y_i = x_i^T \beta^* + \epsilon_i,$$

where $\epsilon \perp\!\!\!\perp x_i$ is independent observation noise.

Another example of interest is the generalized linear model (GLM), which includes linear models as a special case, but also includes other classes of regression models such as logistic and Poisson regression. For GLMs, the conditional distribution (2.1) is given by

$$\mathbb{P}_{\beta^*,\sigma}(y_i \mid x_i) = \exp\left\{\frac{y_i x_i^T \beta^* - \psi(x_i^T \beta^*)}{c(\sigma)}\right\},$$

where $\sigma > 0$ is a scale parameter and ψ is the cumulant function [59]. In our settings of interest (e.g., maximum likelihood estimation), β^* may be estimated independent of σ.

2.1.2 Corrupted observations

We also allow for some *corruption* in the data, meaning that we only observe surrogates $z_i \in \mathbb{R}^p$ in place of the covariates x_i, according to some conditional distribution

$$z_i \sim \mathbb{Q}(\cdot \mid x_i). \tag{2.2}$$

Some examples of corruption mechanisms include the following:

Additive noise. Here,

$$z_i = x_i + w_i,$$

where $x_i \perp\!\!\!\perp w_i$ and we assume $\mathrm{Cov}(w_i)$ is known or may be estimated efficiently. This model follows the standard errors-in-variables model of Carroll et al. [18]. Although needing to know $\mathrm{Cov}(w_i)$ a priori is a somewhat restrictive assumption, it is noted in Carroll et al. [18] that $\mathrm{Cov}(w_i)$ may be estimated in settings where repeated noisy measurements of the same covariate are available. Knowledge of $\mathrm{Cov}(w_i)$ is also reasonable in some engineering applications (e.g., compressed sensing), where the noise covariance may correspond to instrument error and may be measured independently.

Missing data. For some fixed fraction $\alpha \in [0, 1)$, and independently for all $1 \le j \le p$, we have

$$z_{ij} = \begin{cases} x_{ij}, & \text{with probability } 1 - \alpha, \\ \text{missing}, & \text{with probability } \alpha. \end{cases}$$

In the statistical literature, this corresponds to the data being missing completely at random (MCAR) [51]. In our algorithms, we do not need to assume that α is known a priori, since a sufficiently good estimate may be obtained simply by taking an empirical average of the number of missing entries in the data matrix.

2.1.3 Regularized M-estimators

In order to estimate the unknown regression vector β^* from observations $\{(z_i, y_i)\}_{i=1}^n$, we will use the technique of M-estimation. Suppose

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^p} \mathcal{L}(\beta),$$

for a function $\mathcal{L}$ which we call the *population risk*. For example, $\mathcal{L}$ may be the expected (conditional) negative log likelihood:

$$\mathcal{L}(\beta) = -\mathbb{E}[\log \mathbb{P}_\beta(y_i \mid x_i)]. \tag{2.3}$$

We denote the *empirical risk* by $\mathcal{L}_n$, where $\mathcal{L}_n$ is a function satisfying $\mathbb{E}[\mathcal{L}_n(\beta)] = \mathcal{L}(\beta)$. For instance, when $\mathcal{L}$ is given by equation (2.3), we may take

$$\mathcal{L}_n(\beta) = -\frac{1}{n}\sum_{i=1}^n \log \mathbb{P}_\beta(y_i \mid x_i). \tag{2.4}$$

In the high-dimensional setting, the minimizer $\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \mathcal{L}_n(\beta)$ may not be unique. Hence, we instead minimize a regularized version, given by

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \mathcal{L}_n(\beta) + \rho_\lambda(\beta) \right\}, \tag{2.5}$$

where ρ_λ is the *regularizer* or *penalty* function, and $\lambda > 0$ is the regularization parameter. Following the terminology of Huber [36], with the addition of a regularizer, we call the estimator (2.5) a *regularized M-estimator* when $\mathcal{L}_n$ may be written as an average of functions of individual observations (e.g., equation (2.4)). We will also allow for an extra side condition in equation (2.5), where the vector β is constrained to lie in a convex set Ω.

A standard choice for the regularizer ρ_λ when β^* is sparse is the ℓ_1-norm,

$$\rho_\lambda(\beta) = \lambda \|\beta\|_1,$$

which may be viewed as a convex relaxation of the nonconvex regularizer

$$\rho_\lambda(\beta) = \lambda \|\beta\|_0.$$

Other nonconvex regularizers of interest, including the smoothly clipped absolute deviation (SCAD) penalty [28] and the minimax concave penalty [104] will be introduced later. Note that the well-known Lasso estimator [85] is an example of a regularized M-estimator (2.5), where $\mathcal{L}_n$ is the least-squares loss for linear regression and ρ_λ is the ℓ_1-penalty.

2.2 Graphical models

We now turn our attention to graphical models. Given a joint probability distribution $q(x_1, \ldots, x_p)$, we study graphical structures $G = (V, E)$, with $V = \{1, \ldots, p\}$ and $E \subseteq V \times V$, which respect certain characteristics of the distribution. In particular, the absence of edges in G indicates conditional independence relations between subsets of variables. While we will focus on undirected graphical models in this thesis, we include a brief overview of directed graphical models, as well.

2.2.1 Undirected graphs

An undirected graph $G = (V, E)$ is a *conditional independence graph* or *Markov random field* for the distribution q if the following property holds: For any disjoint triple $(A, B, S) \subseteq V$ such that S separates A from B, meaning any path from a vertex in A to a vertex in B must pass through a vertex in S, we have $X_A \perp\!\!\!\perp X_B \mid X_S$. Here, $X_C := \{X_j : j \in C\}$ for any subset $C \subseteq V$. We also say that G *represents* the distribution q.

By the well-known Hammersley-Clifford theorem [47], if q is a strictly positive distribution (i.e., $q(x_1, \ldots, x_p) > 0$ for all $(x_1, \ldots, x_p)$), then G represents q if and only if we may write

$$q(x_1, \ldots, x_p) = \prod_{C \in \mathcal{C}} \psi_C(x_C),$$

for some potential functions $\{\psi_C : C \in \mathcal{C}\}$ defined over the set of cliques $\mathcal{C}$ of G. In particular, the complete graph on p vertices always constitutes an undirected graphical model representation for q, but representations with fewer edges may exist.

2.2.2 Directed graphs

We now consider a *directed* graph $G = (V, E)$, where we distinguish between edges (j, k) and (k, j). We say that G is a *directed acyclic graph* (DAG) if there are no directed paths starting and ending at the same node. For each node $j \in V$, let $\mathrm{Pa}(j) := \{k \in V : (k, j) \in E\}$ denote the *parent set* of j. A DAG G *represents* a distribution $q(x_1, \ldots, x_p)$ if q factorizes as

$$q(x_1, \ldots, x_p) \propto \prod_{j=1}^{p} q(x_j \mid x_{\mathrm{Pa}(j)}). \tag{2.6}$$

A permutation π of the vertex set V is a *topological order* for G if $\pi(j) < \pi(k)$ whenever $(j, k) \in E$. The factorization (2.6) implies $X_j \perp\!\!\!\perp X_{\nu(j)} \mid X_{\mathrm{Pa}(j)}$ fo all j, where $\nu(j)$ is the set of nondescendants of j.

2.2.3 Structure estimation

Given joint observations $\{(x_1, \ldots, x_p)\}_{i=1}^n$ from the distribution q, our goal is to infer the unknown edge structure of the graph G. When G is undirected, existing methods for structure

estimation generally fall into two categories: local (nodewise) methods [60, 73] and global methods [100, 30, 24].

For local methods, the procedure is to estimate the neighborhood set $N(j)$ of each node $j \in \{1, \ldots, p\}$ in succession. The edge set E will then be defined using either an AND function (i.e., $(j,k) \in E$ if and only if $j \in N(k)$ AND $k \in N(j)$) or an OR function (i.e., $(j,k) \in E$ if and only if $j \in N(k)$ OR $k \in N(j)$). For global methods, the procedure involves minimizing an empirical loss function defined in terms of an appropriate summary statistic of the graph. For instance, when the underlying distribution is multivariate Gaussian, it is well-known that the support of the inverse covariance matrix $\Theta := (\mathrm{Cov}(X))^{-1}$ coincides with the edge structure of the conditional independence graph [47]. Consequently, popular methods for structure estimation of Gaussian graphical models reduce to performing a maximum likelihood calculation over the space of positive semidefinite matrices.

When G is a directed graph, structure estimation is a significantly harder problem. If a topological order of the vertices is known a priori, one may simply regress each vertex upon its predecessors and select a neighborhood set that maximizes the function fit. However, when a topological order is unknown, existing methods for DAG estimation involve costly search algorithms that scale exponentially with the size of the graph [69, 83].

We will again focus our attention on high-dimensional settings, where $p \gg n$. In order to avoid issues of nonidentifiability, we will assume that the number of edges and/or the maximal degree of the underlying graph G are sparse. This manifests itself in the addition of a regularization term in both nodewise and global estimation methods.

2.3 Optimization algorithms

Finally, we include background two optimization algorithms that we employ and analyze in this thesis. Both are *first-order* methods, meaning they are iterative methods for minimizing a target function based on gradients.

2.3.1 Projected gradient descent

The projected gradient method is used to optimize functions of the form

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t. } x \in \Omega, \end{aligned}$$

where f is differentiable and $\Omega \subseteq \mathbb{R}^p$ is a closed convex set. The algorithm is initialized at a point $x^0 \in \Omega$, and successive iterates take the form

$$x^{t+1} = \arg\min_{x \in \Omega} \left\| x - (x^t - \eta^t \nabla f(x^t)) \right\|_2^2,$$

or equivalently,

$$x^{t+1} = P_\Omega\Big(x^t - \eta^t \nabla f(x^t)\Big),$$

where P_Ω is the projection operator onto the set Ω and η^t is the stepsize at iteration t. Our results will be derived for a fixed stepsize η, but the stepsize may also be chosen adaptively after each successive iteration. For more details on projected gradient methods and convergence guarantees, see Bertsekas [6].

2.3.2 Composite gradient descent

Now consider the case when the function to be optimized is not smooth. The composite gradient descent method is used to optimize functions of the form

$$\begin{aligned} &\min \ \{f(x) + g(x)\} \\ &\text{s.t. } x \in \Omega, \end{aligned}$$

where $\Omega \subseteq \mathbb{R}^p$ is a closed convex set, f is differentiable, and g is convex but not necessarily differentiable. The algorithm is initialized at a point $x^0 \in \Omega$, and successive iterates take the form

$$x^{t+1} = \arg\min_{x \in \Omega} \left\{ f(x^t) + \langle \nabla f(x^t),\, x - x^t \rangle + \frac{L^t}{2} \|x - x^t\|_2^2 + g(x) \right\},$$

or equivalently,

$$x^{t+1} = \arg\min_{x \in \Omega} \left\{ \left\| x - (x^t - \eta^t \nabla f(x^t)) \right\|_2^2 + 2\eta^t g(x) \right\},$$

where $\eta^t = \frac{1}{L^t}$ is the stepsize. Again, the stepsize may be constant or chosen adaptively. For more details and convergence guarantees, see Nesterov [65].

Chapter 3

Modified Lasso algorithm

3.1 Introduction

In standard formulations of prediction problems, it is assumed that the covariates are fully-observed and sampled independently from some underlying distribution. However, these assumptions are not realistic for many applications, in which covariates may be observed only partially, observed subject to corruption, or exhibit some type of dependency. Consider the problem of modeling the voting behavior of politicians: in this setting, votes may be missing due to abstentions, and temporally dependent due to collusion or "tit-for-tat" behavior. Similarly, surveys often suffer from the missing data problem, since users fail to respond to all questions. Sensor network data also tends to be both noisy due to measurement error, and partially missing due to failures or drop-outs of sensors.

There are a variety of methods for dealing with noisy and/or missing data, including various heuristic methods, as well as likelihood-based methods involving the expectation-maximization (EM) algorithm (e.g., see the book [51] and references therein). A challenge in this context is the possible nonconvexity of associated optimization problems. For instance, in applications of EM, problems in which the negative likelihood is a convex function often become nonconvex with missing or noisy data. Consequently, although the EM algorithm will converge to a local minimum, it is difficult to guarantee that the local optimum is close to a global minimum.

In this chapter, we study these issues in the context of high-dimensional sparse linear regression—in particular, in the case when the predictors or covariates are noisy, missing, and/or dependent. Our main contribution is to develop and study simple methods for handling these issues, and to prove theoretical results about both the associated statistical error and the optimization error. Like EM-based approaches, our estimators are based on solving optimization problems that may be nonconvex; however, despite this nonconvexity, we are still able to prove that a simple form of projected gradient descent will produce an output that is "sufficiently close"—as small as the statistical error—to any global optimum. As a second result, we bound the statistical error, showing that it has the same scaling as

the minimax rates for the classical cases of perfectly observed and independently sampled covariates. In this way, we obtain estimators for noisy, missing, and/or dependent data that have the same scaling behavior as the usual fully-observed and independent case. The resulting estimators allow us to solve the problem of high-dimensional Gaussian graphical model selection with missing data.

There is a large body of work on the problem of corrupted covariates or error-in-variables for regression problems (e.g., see the papers and books [39, 18, 41, 95], as well as references therein). Much of the earlier theoretical work is classical in nature, meaning that it requires that the sample size n diverges with the dimension p fixed. Most relevant to this chapter is more recent work that has examined issues of corrupted and/or missing data in the context of high-dimensional sparse linear models, allowing for $n \ll p$. Städler and Bühlmann [84] developed an EM-based method for sparse inverse covariance matrix estimation in the missing data regime, and used this result to derive an algorithm for sparse linear regression with missing data. As mentioned above, however, it is difficult to guarantee that EM will converge to a point close to a global optimum of the likelihood, in contrast to the methods studied here. Rosenbaum and Tsybakov [76] studied the sparse linear model when the covariates are corrupted by noise, and proposed a modified form of the Dantzig selector (see the discussion following our main results for a detailed comparison to this past work, and also to concurrent work [77] by the same authors). For the particular case of multiplicative noise, the type of estimator that we consider here has been studied in past work [95]; however, this theoretical analysis is of the classical type, holding only for $n \gg p$, in contrast to the high-dimensional models that are of interest here.

The remainder of this chapter is organized as follows. We begin in Section 3.2 with background and a precise description of the problem. We then introduce the class of estimators we will consider and the form of the projected gradient descent algorithm. Section 3.3 is devoted to a description of our main results, including a pair of general theorems on the statistical and optimization error, and then a series of corollaries applying our results to the cases of noisy, missing, and dependent data. In Section 3.4, we demonstrate simulations to confirm that our methods work in practice, and verify the theoretically-predicted scaling laws. In Section 3.5, we derive information-theoretic lower bounds establishing the minimax optimality of the modified Lasso for an important subclass of problems. Section 3.6 contains proofs of some of the main results, with the remaining proofs contained in Appendix A.

Notation: For a matrix M, we write $\|M\|_{\max} := \max_{i,j} |m_{ij}|$ to be the elementwise ℓ_∞-norm of M. Furthermore, $|\!|\!| M |\!|\!|_1$ denotes the induced ℓ_1-operator norm (maximum absolute column sum) of M, and $|\!|\!| M |\!|\!|_{\text{op}}$ is the spectral norm of M. We write $\kappa(M) := \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$, the condition number of M. For matrices M_1, M_2, we write $M_1 \odot M_2$ to denote the componentwise Hadamard product, and write $M_1 \oslash M_2$ to denote componentwise division. For functions $f(n)$ and $g(n)$, we write $f(n) \precsim g(n)$ to mean that $f(n) \leq c g(n)$ for a universal constant $c \in (0, \infty)$, and similarly, $f(n) \succsim g(n)$ when $f(n) \geq c' g(n)$ for some universal constant $c' \in (0, \infty)$. Finally, we write $f(n) \asymp g(n)$ when $f(n) \precsim g(n)$ and $f(n) \succsim g(n)$ hold

simultaneously.

3.2 Background and problem setup

In this section, we provide background and a precise description of the problem, and then motivate the class of estimators analyzed in this chapter. We then discuss a simple class of projected gradient descent algorithms that may be used to obtain an estimator.

3.2.1 Observation model and high-dimensional framework

Suppose we observe a response variable $y_i \in \mathbb{R}$ linked to a covariate vector $x_i \in \mathbb{R}^p$ via the linear model

$$y_i = \langle x_i, \beta^* \rangle + \epsilon_i, \qquad \text{for } i = 1, 2, \ldots, n. \tag{3.1}$$

Here, the regression vector $\beta^* \in \mathbb{R}^p$ is unknown, and $\epsilon_i \in \mathbb{R}$ is observation noise, independent of x_i. Rather than directly observing each $x_i \in \mathbb{R}^p$, we observe a vector $z_i \in \mathbb{R}^p$ linked to x_i via some conditional distribution, i.e.,

$$z_i \sim \mathbb{Q}(\cdot \mid x_i), \qquad \text{for } i = 1, 2, \ldots, n. \tag{3.2}$$

This setup applies to various disturbances to the covariates, including:

(a) *Covariates with additive noise:* We observe $z_i = x_i + w_i$, where $w_i \in \mathbb{R}^p$ is a random vector independent of x_i, say zero-mean with known covariance matrix Σ_w.

(b) *Missing data:* For some fraction $\alpha \in [0, 1)$, we observe a random vector $z_i \in \mathbb{R}^p$ such that for each component j, we independently observe $z_{ij} = x_{ij}$ with probability $1 - \alpha$, and $z_{ij} = *$ with probability α. We can also consider the case when the entries in the j^{th} column have a different probability α_j of being missing.

(c) *Covariates with multiplicative noise:* Generalizing the missing data problem, suppose we observe $z_i = x_i \odot u_i$, where $u_i \in \mathbb{R}^p$ is again a random vector independent of x_i, and $\odot$ is the Hadamard product. The problem of missing data is a special case of multiplicative noise, where all u_{ij}'s are independent and $u_{ij} \sim \text{Bernoulli}(1 - \alpha_j)$.

Our first set of results is deterministic, depending on specific instantiations of the observations $\{(y_i, z_i)\}_{i=1}^n$. However, we are also interested in results that hold with high probability when the x_i's and z_i's are drawn at random. We consider both the case when the x_i's are drawn i.i.d. from a fixed distribution; and the case of dependent covariates, when the x_i's are generated according to a stationary vector autoregressive (VAR) process.

We work within a high-dimensional framework that allows the number of predictors p to grow and possibly exceed the sample size n. Of course, consistent estimation when $n \ll p$ is impossible unless the model is endowed with additional structure—for instance, sparsity in the parameter vector β^*. Consequently, we study the class of models where β^* has at most k non-zero parameters, where k is also allowed to increase to infinity with p and n.

3.2.2 M-estimators for noisy and missing covariates

In order to motivate the class of estimators we will consider, let us begin by examining a simple deterministic problem. Let $\Sigma_x \succ 0$ be the covariance matrix of the covariates, and consider the ℓ_1-constrained quadratic program

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \leq R} \big\{ \frac{1}{2} \beta^T \Sigma_x \beta - \langle \Sigma_x \beta^*, \, \beta \rangle \big\}. \tag{3.3}$$

As long as the constraint radius R is at least $\|\beta^*\|_1$, the unique solution to this convex program is $\widehat{\beta} = \beta^*$. Of course, this program is an idealization, since in practice we may not know the covariance matrix Σ_x, and we certainly do not know $\Sigma_x \beta^*$—after all, β^* is the quantity we are trying to estimate!

Nonetheless, this idealization still provides useful intuition, as it suggests various estimators based on the plug-in principle. Given a set of samples, it is natural to form estimates of the quantities Σ_x and $\Sigma_x \beta^*$, which we denote by $\widehat{\Gamma} \in \mathbb{R}^{p \times p}$ and $\widehat{\gamma} \in \mathbb{R}^p$, respectively, and to consider the modified program

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \leq R} \big\{ \frac{1}{2} \beta^T \widehat{\Gamma} \beta - \langle \widehat{\gamma}, \, \beta \rangle \big\}, \tag{3.4}$$

or alternatively, the regularized version

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \big\{ \frac{1}{2} \beta^T \widehat{\Gamma} \beta - \langle \widehat{\gamma}, \, \beta \rangle + \lambda_n \|\beta\|_1 \big\}, \tag{3.5}$$

where $\lambda_n > 0$ is a user-defined regularization parameter. Note that the two problems are equivalent by Lagrangian duality when the objectives are convex, but not in the case of a nonconvex objective. The Lasso [85, 20] is a special case of these programs, obtained by setting

$$\widehat{\Gamma}_{\text{Las}} := \frac{1}{n} X^T X \quad \text{and} \quad \widehat{\gamma}_{\text{Las}} := \frac{1}{n} X^T y, \tag{3.6}$$

where we have introduced the shorthand $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, and $X \in \mathbb{R}^{n \times p}$, with x_i^T as its i^{th} row. A simple calculation shows that $(\widehat{\Gamma}_{\text{Las}}, \widehat{\gamma}_{\text{Las}})$ are unbiased estimators of the pair $(\Sigma_x, \, \Sigma_x \beta^*)$. This unbiasedness and additional concentration inequalities (to be described in the sequel) underlie the well-known analysis of the Lasso in the high-dimensional regime.

In this chapter, we focus on more general instantiations of the programs (3.4) and (3.5), involving different choices of the pair $(\widehat{\Gamma}, \widehat{\gamma})$ that are adapted to the cases of noisy and/or missing data. Note that the matrix $\widehat{\Gamma}_{\text{Las}}$ is positive semidefinite, so the Lasso program is convex. In sharp contrast, for the case of noisy or missing data, the most natural choice of the matrix $\widehat{\Gamma}$ is *not positive semidefinite*, hence the quadratic losses appearing in the problems (3.4) and (3.5) are *nonconvex*. Furthermore, when $\widehat{\Gamma}$ has negative eigenvalues, the

objective in equation (3.5) is unbounded from below. Hence, we make use of the following regularized estimator:

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \leq b_0\sqrt{k}} \{\frac{1}{2}\beta^T\widehat{\Gamma}\beta - \langle\widehat{\gamma}, \beta\rangle + \lambda_n\|\beta\|_1\}, \tag{3.7}$$

for a suitable constant b_0.

In the presence of nonconvexity, it is generally impossible to provide a polynomial-time algorithm that converges to a (near) global optimum, due to the presence of local minima. Remarkably, we are able to prove that this issue is not significant in our setting, and a simple projected gradient descent algorithm applied to the programs (3.4) or (3.7) converges with high probability to a vector extremely close to any global optimum.

Let us illustrate these ideas with some examples. Recall that $(\widehat{\Gamma}, \widehat{\gamma})$ serve as unbiased estimators for $(\Sigma_x, \Sigma_x\beta^*)$.

Example 3.1 (Additive noise)**.** *Suppose we observe $Z = X + W$, where W is a random matrix independent of X, with rows w_i drawn i.i.d. from a zero-mean distribution with known covariance Σ_w. We consider the pair*

$$\widehat{\Gamma}_{add} := \frac{1}{n}Z^TZ - \Sigma_w \quad \textit{and} \quad \widehat{\gamma}_{add} := \frac{1}{n}Z^Ty. \tag{3.8}$$

Note that when $\Sigma_w = 0$ (corresponding to the noiseless case), the estimators reduce to the standard Lasso. However, when $\Sigma_w \neq 0$, the matrix $\widehat{\Gamma}_{add}$ is not *positive semidefinite in the high-dimensional regime ($n \ll p$). Indeed, since the matrix $\frac{1}{n}Z^TZ$ has rank at most n, the subtracted matrix Σ_w may cause $\widehat{\Gamma}_{add}$ to have a large number of negative eigenvalues. For instance, if $\Sigma_w = \sigma_w^2 I$ for $\sigma_w^2 > 0$, then $\widehat{\Gamma}_{add}$ has $p - n$ eigenvalues equal to $-\sigma_w^2$.*

Example 3.2 (Missing data)**.** *We now consider the case where the entries of X are missing at random. Let us first describe an estimator for the special case where each entry is missing at random, independently with some constant probability $\alpha \in [0, 1)$. (In Example 3.3 to follow, we will describe the extension to general missing probabilities.) Consequently, we observe the matrix $Z \in \mathbb{R}^{n\times p}$ with entries*

$$Z_{ij} = \begin{cases} X_{ij} & \textit{with probability } 1 - \alpha, \\ 0 & \textit{otherwise.} \end{cases}$$

Given the observed matrix $Z \in \mathbb{R}^{n\times p}$, we use

$$\widehat{\Gamma}_{mis} := \frac{\widetilde{Z}^T\widetilde{Z}}{n} - \alpha\ \mathrm{diag}\,(\frac{\widetilde{Z}^T\widetilde{Z}}{n}) \quad \textit{and} \quad \widehat{\gamma}_{mis} := \frac{1}{n}\widetilde{Z}^Ty, \tag{3.9}$$

where $\widetilde{Z}_{ij} = Z_{ij}/(1 - \alpha)$. It is easy to see that the pair $(\widehat{\Gamma}_{mis}, \widehat{\gamma}_{mis})$ reduces to the pair $(\widehat{\Gamma}_{Las}, \widehat{\gamma}_{Las})$ for the standard Lasso when $\alpha = 0$, corresponding to no missing data. In the

more interesting case when $\alpha \in (0,1)$, the matrix $\frac{\widetilde{Z}^T\widetilde{Z}}{n}$ in equation (3.9) has rank at most n, so the subtracted diagonal matrix may cause the matrix $\widehat{\Gamma}_{mis}$ to have a large number of negative eigenvalues when $n \ll p$. As a consequence, the matrix $\widehat{\Gamma}_{mis}$ is not (in general) positive semidefinite, so the associated quadratic function is not convex.

Example 3.3 (Multiplicative noise)**.** *As a generalization of the previous example, we now consider the case of multiplicative noise. In particular, suppose we observe the quantity $Z = X \odot U$, where U is a matrix of nonnegative noise variables. In many applications, it is natural to assume that the rows u_i of U are drawn in an i.i.d. manner, say from some distribution in which both the vector $\mathbb{E}[u_1]$ and the matrix $\mathbb{E}[u_1 u_1^T]$ have strictly positive entries. This general family of multiplicative noise models arises in various applications; we refer the reader to the papers [39, 18, 41, 95] for more discussion and examples. A natural choice of the pair $(\widehat{\Gamma}, \widehat{\gamma})$ is given by the quantities*

$$\widehat{\Gamma}_{mul} := \frac{1}{n} Z^T Z \oslash \mathbb{E}(u_1 u_1^T) \quad \text{and} \quad \widehat{\Gamma}_{mul} := \frac{1}{n} Z^T y \oslash \mathbb{E}(u_1), \tag{3.10}$$

where $\oslash$ denotes elementwise division. A small calculation shows that these are unbiased estimators of Σ_x and $\Sigma_x \beta^*$, respectively. The estimators (3.10) have been studied in past work [95], but only under classical scaling ($n \gg p$).

As a special case of the estimators (3.10), suppose the entries u_{ij} of U are independent Bernoulli$(1-\alpha_j)$ random variables. Then the observed matrix $Z = X \odot U$ corresponds to a missing-data matrix, where each element of the j^{th} column has probability α_j of being missing. In this case, the estimators (3.10) become

$$\widehat{\Gamma}_{mis} = \frac{Z^T Z}{n} \oslash M \quad \text{and} \quad \widehat{\gamma}_{mis} = \frac{1}{n} Z^T y \oslash (\mathbf{1} - \boldsymbol{\alpha}), \tag{3.11}$$

where $M := \mathbb{E}(u_1 u_1^T)$ satisfies

$$M_{ij} = \begin{cases} (1-\alpha_i)(1-\alpha_j) & \text{if } i \neq j \\ 1 - \alpha_i & \text{if } i = j, \end{cases}$$

$\boldsymbol{\alpha}$ is the parameter vector containing the α_j's, and $\mathbf{1}$ is the vector of all 1's. In this way, we obtain a generalization of the estimator discussed in Example 3.2.

3.2.3 Restricted eigenvalue conditions

Given an estimate $\widehat{\beta}$, there are various ways to assess its closeness to β^*. In this chapter, we focus on the ℓ_2-norm $\|\widehat{\beta} - \beta^*\|_2$, as well as the closely related ℓ_1-norm $\|\widehat{\beta} - \beta^*\|_1$. When the covariate matrix X is fully observed (so that the Lasso can be applied), it is now well understood that a sufficient condition for ℓ_2-recovery is that the matrix $\widehat{\Gamma}_{\text{Las}} = \frac{1}{n} X^T X$ satisfy a certain type of restricted eigenvalue (RE) condition (e.g., [8, 32]). In this chapter, we make use of the following condition.

Definition 1 (Lower-RE condition)**.** The matrix $\widehat{\Gamma}$ satisfies a lower restricted eigenvalue condition with curvature $\alpha_\ell > 0$ and tolerance $\tau(n,p) > 0$ if

$$\theta^T \widehat{\Gamma} \theta \geq \alpha_\ell \, \|\theta\|_2^2 \; - \; \tau(n,p) \|\theta\|_1^2 \qquad \text{for all } \theta \in \mathbb{R}^p. \tag{3.12}$$

It can be shown that when the Lasso matrix $\widehat{\Gamma}_{\text{Las}} = \frac{1}{n} X^T X$ satisfies this RE condition (3.12), the Lasso estimate has low ℓ_2-error for any vector β^* supported on any subset of size at most $k \lesssim \frac{1}{\tau(n,p)}$. In particular, bound (3.12) implies a sparse RE condition for all k of this magnitude, and conversely, Lemma A.11 in the Appendix shows that a sparse RE condition implies bound (3.12). In this chapter, we work with condition (3.12), since it is especially convenient for analyzing optimization algorithms.

In the standard setting (with uncorrupted and fully observed design matrices), it is known that for many choices of the design matrix X (with rows having covariance Σ), the Lasso matrix $\widehat{\Gamma}_{\text{Las}}$ will satisfy such an RE condition with high probability (e.g., [70, 80]) with $\alpha_\ell = \frac{1}{2}\lambda_{\min}(\Sigma)$ and $\tau(n,p) \asymp \frac{\log p}{n}$. A significant portion of the analysis in this chapter is devoted to proving that different choices of $\widehat{\Gamma}$, such as the matrices $\widehat{\Gamma}_{\text{add}}$ and $\widehat{\Gamma}_{\text{mis}}$ defined earlier, also satisfy condition (3.12) with high probability. This fact is by no means obvious, since as previously discussed, the matrices $\widehat{\Gamma}_{\text{add}}$ and $\widehat{\Gamma}_{\text{mis}}$ generally have large numbers of negative eigenvalues.

Finally, although such upper bounds are not necessary for statistical consistency, our algorithmic results make use of the analogous upper restricted eigenvalue condition, formalized in the following:

Definition 2 (Upper-RE condition)**.** The matrix $\widehat{\Gamma}$ satisfies an upper restricted eigenvalue condition with smoothness $\alpha_u > 0$ and tolerance $\tau(n,p) > 0$ if

$$\theta^T \widehat{\Gamma} \theta \leq \alpha_u \|\theta\|_2^2 + \tau(n,p) \|\theta\|_1^2 \qquad \text{for all } \theta \in \mathbb{R}^p. \tag{3.13}$$

In recent work on high-dimensional projected gradient descent, Agarwal et al. [1] make use of a more general form of the lower and upper bounds (3.12) and (3.13), applicable to non-quadratic losses as well, which are referred to as the restricted strong convexity (RSC) and restricted smoothness (RSM) conditions, respectively. For various class of random design matrices, it can be shown that the Lasso matrix $\widehat{\Gamma}_{\text{Las}}$ satisfies the upper bound (3.13) with $\alpha_u = 2\lambda_{\max}(\Sigma_x)$ and $\tau(n,p) \asymp \frac{\log p}{n}$; see Raskutti et al. [70] for the Gaussian case and Rudelson and Zhou [80] for the sub-Gaussian setting. We will establish similar scaling for our choices of $\widehat{\Gamma}$.

3.2.4 Gradient descent algorithms

In addition to proving results about the global minima of the (possibly nonconvex) programs (3.4) and (3.5), we are also interested in polynomial-time procedures for approximating such optima. In this chapter, we analyze some simple algorithms for solving either

the constrained program (3.4) or the Lagrangian version (3.7). Note that the gradient of the quadratic loss function takes the form $\nabla \mathcal{L}(\beta) = \widehat{\Gamma}\beta - \widehat{\gamma}$. In application to the constrained version, the method of projected gradient descent generates a sequence of iterates $\{\beta^t, t = 0, 1, 2, \ldots\}$ by the recursion

$$\beta^{t+1} = \arg \min_{\|\beta\|_1 \leq R} \big\{ \mathcal{L}(\beta^t) + \langle \nabla \mathcal{L}(\beta^t), \, \beta - \beta^t \rangle + \frac{\eta}{2} \|\beta - \beta^t\|_2^2 \big\}, \tag{3.14}$$

where $\eta > 0$ is a stepsize parameter. Equivalently, this update can be written as $\beta^{t+1} = \Pi\big(\beta^t - \frac{1}{\eta} \nabla \mathcal{L}(\beta^t)\big)$, where Π denotes the ℓ_2-projection onto the ℓ_1-ball of radius R. This projection can be computed rapidly in $\mathcal{O}(p)$ time using a procedure due to Duchi et al. [26]. For the Lagrangian update, we use a slight variant of the projected gradient update (3.14), namely

$$\beta^{t+1} = \arg \min_{\|\beta\|_1 \leq R} \big\{ \mathcal{L}(\beta^t) + \langle \nabla \mathcal{L}(\beta^t), \, \beta - \beta^t \rangle + \frac{\eta}{2} \|\beta - \beta^t\|_2^2 + \lambda_n \|\beta\|_1 \big\}, \tag{3.15}$$

with the only difference being the inclusion of the regularization term. This update can also performed efficiently by performing two projections onto the ℓ_1-ball (see the paper [1] for details).

When the objective function is convex (equivalently, $\widehat{\Gamma}$ is positive semidefinite), the iterates (3.14) or (3.15) are guaranteed to converge to a global minimum of the objective functions (3.4) and (3.7), respectively. In our setting, the matrix $\widehat{\Gamma}$ need not be positive semidefinite, so the best generic guarantee is that the iterates converge to a local optimum. However, our analysis shows that for the family of programs (3.4) or (3.7), under a reasonable set of conditions satisfied by various statistical models, the iterates actually converge to a point extremely close to any global optimum in both ℓ_1-norm and ℓ_2-norm; see Theorem 3.2 to follow for a more detailed statement.

3.3 Main results and consequences

We now state our main results and discuss their consequences for noisy, missing, and dependent data.

3.3.1 General results

We provide theoretical guarantees for both the constrained estimator (3.4) and the Lagrangian version (3.7). Note that we obtain different optimization problems as we vary the choice of the pair $(\widehat{\Gamma}, \widehat{\gamma}) \in \mathbb{R}^{p \times p} \times \mathbb{R}^p$. We begin by stating a pair of general results, applicable to any pair that satisfies certain conditions. Our first result (Theorem 3.1) provides bounds on the statistical error, namely the quantity $\|\widehat{\beta} - \beta^*\|_2$, as well as the corresponding ℓ_1-error, where $\widehat{\beta}$ is any global optimum of the programs (3.4) or (3.7). Since the problem may be

nonconvex in general, it is not immediately obvious that one can obtain a *provably good* approximation to any global optimum without resorting to costly search methods. In order to assuage this concern, our second result (Theorem 3.2) provides rigorous bounds on the optimization error, namely the differences $\|\beta^t - \widehat{\beta}\|_2$ and $\|\beta^t - \widehat{\beta}\|_1$ incurred by the iterate β^t after running t rounds of the projected gradient descent updates (3.14) or (3.15).

3.3.1.1 Statistical error

In controlling the statistical error, we assume that the matrix $\widehat{\Gamma}$ satisfies a lower-RE condition with curvature α_ℓ and tolerance $\tau(n,p)$, as previously defined (3.12). Recall that $\widehat{\Gamma}$ and $\widehat{\gamma}$ serve as surrogates to the deterministic quantities $\Sigma_x \in \mathbb{R}^{p\times p}$ and $\Sigma_x\beta^* \in \mathbb{R}^p$, respectively. Our results also involve a measure of deviation in these surrogates. In particular, we assume that there is some function $\varphi(\mathbb{Q}, \sigma_\epsilon)$, depending on the two sources of noise in our problem: the standard deviation σ_ϵ of the observation noise vector ϵ from equation (3.1), and the conditional distribution $\mathbb{Q}$ from equation (3.2) that links the covariates x_i to the observed versions z_i. With this notation, we consider the deviation condition

$$\|\widehat{\gamma} - \widehat{\Gamma}\beta^*\|_\infty \leq \varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}. \tag{3.16}$$

To aid intuition, note that inequality (3.16) holds whenever the following two deviation conditions are satisfied:

$$\|\widehat{\gamma} - \Sigma_x\beta^*\|_\infty \leq \varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}} \quad \text{and} \quad \|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty \leq \varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}. \tag{3.17}$$

The pair of inequalities (3.17) clearly measures the deviation of the estimators $(\widehat{\Gamma}, \widehat{\gamma})$ from their population versions, and they are sometimes easier to verify theoretically. However, inequality (3.16) may be used directly to derive tighter bounds (e.g., in the additive noise case). Indeed, the bounds established via inequalities (3.17) is not sharp in the limit of low noise on the covariates, due to the second inequality. In the proofs of our corollaries to follow, we will verify the deviation conditions for various forms of noisy, missing, and dependent data, with the quantity $\varphi(\mathbb{Q}, \sigma_\epsilon)$ changing depending on the model. We have the following result, which applies to any global optimum $\widehat{\beta}$ of the regularized version (3.7) with $\lambda_n \geq 4\,\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}$:

Theorem 3.1 (Statistical error). *Suppose the pair* $(\widehat{\Gamma}, \widehat{\gamma})$ *satisfies the deviation bound* (3.16)*, and the matrix* $\widehat{\Gamma}$ *satisfies the lower-RE condition* (3.12) *with parameters* (α_ℓ, τ) *such that*

$$\sqrt{k}\,\tau(n,p) \leq \min\Big\{\frac{\alpha_\ell}{128\sqrt{k}},\ \frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{b_0}\sqrt{\frac{\log p}{n}}\Big\}. \tag{3.18}$$

Then for any vector β^* with sparsity at most k, there is a universal positive constant c_0 such that any global optimum $\widehat{\beta}$ of the Lagrangian program (3.7) with any $b_0 \geq \|\beta^*\|_2$ satisfies the bounds

$$\|\widehat{\beta} - \beta^*\|_2 \leq \frac{c_0\sqrt{k}}{\alpha_\ell} \max\left\{\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}, \lambda_n\right\}, \quad \text{and} \tag{3.19a}$$

$$\|\widehat{\beta} - \beta^*\|_1 \leq \frac{8\, c_0\, k}{\alpha_\ell} \max\left\{\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}, \lambda_n\right\}. \tag{3.19b}$$

The same bounds (without λ_n) also apply to the constrained program (3.4) with radius choice $R = \|\beta^*\|_1$.

Remarks To be clear, all the claims of Theorem 3.1 are deterministic. Probabilistic conditions will enter when we analyze specific statistical models and certify that the RE condition (3.18) and deviation conditions are satisfied by a random pair $(\widehat{\Gamma}, \widehat{\gamma})$ with high probability. We note that for the standard Lasso choice $(\widehat{\Gamma}_{\text{Las}}, \widehat{\gamma}_{\text{Las}})$ of this matrix-vector pair, bounds of the form (3.19) for sub-Gaussian noise are well known from past work (e.g., [8, 102, 61, 64]). The novelty of Theorem 3.1 is in allowing for general pairs of such surrogates, which—as shown by the examples discussed earlier—can lead to nonconvexity in the underlying M-estimator. Moreover, some interesting differences arise due to the term $\varphi(\mathbb{Q}, \sigma_\epsilon)$, which changes depending on the nature of the model (missing, noisy, and/or dependent). As will be clarified in the sequel, proving that the conditions of Theorem 3.1 are satisfied with high probability for noisy/missing data requires some non-trivial analysis, involving both concentration inequalities and random matrix theory.

Note that in the presence of nonconvexity, it is possible in principle for the optimization problems (3.4) and (3.7) to have *many* global optima that are separated by large distances. Interestingly, Theorem 3.1 guarantees that this unpleasant feature does not arise under the stated conditions: given any two global optima $\widehat{\beta}$ and $\widetilde{\beta}$ of the program (3.4), Theorem 3.1 combined with the triangle inequality guarantees that

$$\|\widehat{\beta} - \widetilde{\beta}\|_2 \leq \|\widehat{\beta} - \beta^*\|_2 + \|\widetilde{\beta} - \beta^*\|_2 \leq 2c_0 \frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\alpha_\ell}\sqrt{\frac{k \log p}{n}}$$

and similarly for the program (3.7). Consequently, under any scaling such that $\frac{k \log p}{n} = o(1)$, the set of all global optima must lie within an ℓ_2-ball whose radius shrinks to zero.

In addition, it is worth observing that Theorem 3.1 makes a specific prediction for the scaling behavior of the ℓ_2-error $\|\widehat{\beta} - \beta^*\|_2$. In order to study this scaling prediction, we performed simulations under the additive noise model described in Example 3.1, using the parameter setting $\Sigma_x = I$ and $\Sigma_w = \sigma_w^2 I$ with $\sigma_w = 0.2$. Panel (a) of Figure 3.1 provides plots[1] of

[1] Corollary 3.1, to be stated shortly, guarantees that the conditions of Theorem 3.1 are satisfied with high probability for the additive noise model. In addition, Theorem 3.2 to follow provides an efficient method of obtaining an accurate approximation of the global optimum.

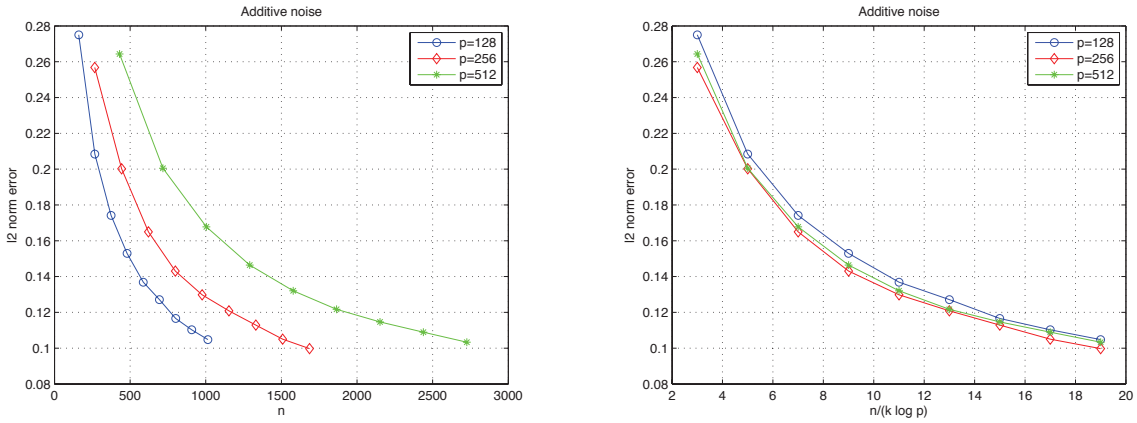


Figure 3.1: Plots of the error $\|\widehat{\beta} - \beta^*\|_2$ after running projected gradient descent on the nonconvex objective, with sparsity $k \approx \sqrt{p}$. Plot (a) is an error plot for i.i.d. data with additive noise, and plot (b) shows ℓ_2-error versus the rescaled sample size $\frac{n}{k \log p}$. As predicted by Theorem 3.1, the curves align for different values of p in the rescaled plot.

the error $\|\widehat{\beta} - \beta^*\|_2$ versus the sample size n, for problem dimensions $p \in \{128, 256, 512\}$. Note that for all three choices of dimensions, the error decreases to zero as the sample size n increases, showing consistency of the method. The curves also shift to the right as the dimension p increases, reflecting the natural intuition that larger problems are harder in a certain sense. Theorem 3.1 makes a specific prediction about this scaling behavior: in particular, if we plot the ℓ_2-error versus the rescaled sample size $n/(k \log p)$, the curves should roughly align for different values of p. Panel (b) shows the same data re-plotted on these rescaled axes, thus verifying the predicted "stacking behavior."

Finally, as noted by a reviewer, the constraint $R = \|\beta^*\|_1$ in the program (3.4) is rather restrictive, since β^* is unknown. Theorem 3.1 merely establishes a heuristic for the scaling expected for this optimal radius. In this regard, the Lagrangian estimator (3.7) is more appealing, since it only requires choosing b_0 to be larger than $\|\beta^*\|_2$, and the conditions on the regularizer λ_n are the standard ones from past work on the Lasso.

3.3.1.2 Optimization error

Although Theorem 3.1 provides guarantees that hold uniformly for any global minimizer, it does not provide guidance on how to approximate such a global minimizer using a polynomial-time algorithm. Indeed, for nonconvex programs in general, gradient-type methods may become trapped in local minima, and it is impossible to guarantee that all such local minima are close to a global optimum. Nonetheless, we are able to show that for the family of programs (3.4), under reasonable conditions on $\widehat{\Gamma}$ satisfied in various settings, simple gradient methods will converge geometrically fast to a very good approximation of

any global optimum. The following theorem supposes that we apply the projected gradient updates (3.14) to the constrained program (3.4), or the composite updates (3.15) to the Lagrangian program (3.7), with stepsize $\eta = 2\alpha_u$. In both cases, we assume that $n \gtrsim k \log p$, as is required for statistical consistency in Theorem 3.1.

Theorem 3.2 (Optimization error). *Under the conditions of Theorem 3.1:*

(a) For any global optimum $\widehat{\beta}$ of the constrained program (3.4), there are universal positive constants (c_1, c_2) and a contraction coefficient $\gamma \in (0,1)$, independent of (n,p,k), such that the gradient descent iterates (3.14) satisfy the bounds

$$\|\beta^t - \widehat{\beta}\|_2^2 \leq \gamma^t \|\beta^0 - \widehat{\beta}\|_2^2 + c_1 \frac{\log p}{n} \|\widehat{\beta} - \beta^*\|_1^2 + c_2 \|\widehat{\beta} - \beta^*\|_2^2, \tag{3.20}$$

$$\|\beta^t - \widehat{\beta}\|_1 \leq 2\sqrt{k}\, \|\beta^t - \widehat{\beta}\|_2 + 2\sqrt{k}\, \|\widehat{\beta} - \beta^*\|_2 + 2\, \|\widehat{\beta} - \beta^*\|_1, \tag{3.21}$$

for all $t \geq 0$.

(b) Letting ϕ denote the objective function of Lagrangian program (3.7) with global optimum $\widehat{\beta}$, and applying composite gradient updates (3.15), there are universal positive constants (c_1, c_2) and a contraction coefficient $\gamma \in (0,1)$, independent of (n,p,k), such that

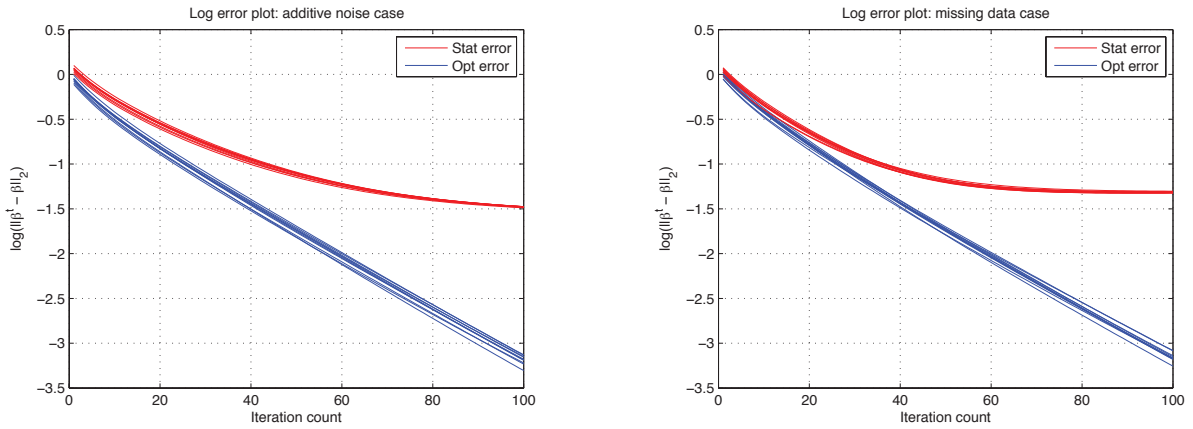
$$\|\beta^t - \widehat{\beta}\|_2^2 \leq \underbrace{c_1 \|\widehat{\beta} - \beta^*\|_2^2}_{\delta^2} \qquad \text{for all iterates } t \geq T, \tag{3.22}$$

where $T := c_2 \log \frac{(\phi(\beta^0) - \phi(\widehat{\beta}))}{\delta^2} \big/ \log(1/\gamma)$.

Remarks As with Theorem 3.1, these claims are deterministic in nature. Probabilistic conditions will enter into the corollaries, which involve proving that the surrogate matrices $\widehat{\Gamma}$ used for noisy, missing, and/or dependent data satisfy the lower- and upper-RE conditions with high probability. The proof of Theorem 3.2 itself is based on an extension of a result due to Agarwal et al. [1] on the convergence of projected gradient descent and composite gradient descent in high dimensions. Their result as originally stated imposed convexity of the loss function, but the proof can be modified so as to apply to the nonconvex loss functions of interest here. As noted following Theorem 3.1, all global minimizers of the nonconvex program (3.4) lie within a small ball. In addition, Theorem 3.2 guarantees that the local minimizers also lie within a ball of the same magnitude. Note that in order to show that Theorem 3.2 can be applied to the specific statistical models of interest in this chapter, a considerable amount of technical analysis remains in order to establish that its conditions hold with high probability.

In order to understand the significance of the bounds (3.20) and (3.22), note that they provide upper bounds for the ℓ_2-distance between the iterate β^t at time t, which is easily computed in polynomial-time, and any global optimum $\widehat{\beta}$ of the program (3.4) or (3.7), which

may be difficult to compute. Focusing on bound (3.20), since $\gamma \in (0, 1)$, the first term in the bound vanishes as t increases. The remaining terms involve the statistical errors $\|\widehat{\beta} - \beta^*\|_q$, for $q = 1, 2$, which are controlled in Theorem 3.1. It can be verified that the two terms involving the statistical error on the right-hand side are bounded as $\mathcal{O}(\frac{k \log p}{n})$, so Theorem 3.2 guarantees that projected gradient descent produce an output that is essentially as good—in terms of statistical error—as any global optimum of the program (3.4). Bound (3.22) provides a similar guarantee for composite gradient descent applied to the Lagrangian version.



(a) (b)

Figure 3.2: Plots of the optimization error $\log(\|\beta^t - \widehat{\beta}\|_2)$ and statistical error $\log(\|\beta^t - \beta^*\|_2)$ versus iteration number t, generated by running projected gradient descent on the nonconvex objective. Each plot shows the solution path for the same problem instance, using 10 different starting points. As predicted by Theorem 3.2, the optimization error decreases geometrically.

Experimentally, we have found that the predictions of Theorem 3.2 are borne out in simulations. Figure 3.2 shows the results of applying the projected gradient descent method to solve the optimization problem (3.4) in the case of additive noise (panel (a)), and missing data (panel (b)). In each case, we generated a random problem instance, and then applied the projected gradient descent method to compute an estimate $\widehat{\beta}$. We then reapplied the projected gradient method to the same problem instance 10 times, each time with a random starting point, and measured the error $\|\beta^t - \widehat{\beta}\|_2$ between the iterates and the first estimate (optimization error), and the error $\|\beta^t - \beta^*\|_2$ between the iterates and the truth (statistical error). Within each panel, the blue traces show the optimization error over 10 trials, and the red traces show the statistical error. On the logarithmic scale given, a geometric rate of convergence corresponds to a straight line. As predicted by Theorem 3.2, regardless of the starting point, the iterates $\{\beta^t\}$ exhibit geometric convergence to the same fixed point.[2]

[2]To be precise, Theorem 3.2 states that the iterates will converge geometrically to a small neighborhood of all the global optima.

The statistical error contracts geometrically up to a certain point, then flattens out.

3.3.2 Some consequences

As discussed previously, both Theorems 3.1 and 3.2 are deterministic results. Applying them to specific statistical models requires some additional work in order to establish that the stated conditions are met. We now turn to the statements of some consequences of these theorems for different cases of noisy, missing, and dependent data. In all the corollaries below, the claims hold with probability greater than $1 - c_1 \exp(-c_2 \log p)$, where (c_1, c_2) are universal positive constants, independent of all other problem parameters. Note that in all corollaries, the triplet (n, p, k) is assumed to satisfy scaling of the form $n \gtrsim k \log p$, as is necessary for ℓ_2-consistent estimation of k-sparse vectors in p dimensions.

Definition 3. We say that a random matrix $X \in \mathbb{R}^{n \times p}$ is sub-Gaussian with parameters (Σ, σ^2) if:

(a) each row $x_i^T \in \mathbb{R}^p$ is sampled independently from a zero-mean distribution with covariance Σ, and

(b) for any unit vector $u \in \mathbb{R}^p$, the random variable $u^T x_i$ is sub-Gaussian with parameter at most σ.

For instance, if we form a random matrix by drawing each row independently from the distribution $N(0, \Sigma)$, then the resulting matrix $X \in \mathbb{R}^{n \times p}$ is a sub-Gaussian matrix with parameters $(\Sigma, |||\Sigma|||_{\text{op}})$.

3.3.2.1 Bounds for additive noise: i.i.d. case

We begin with the case of i.i.d. samples with additive noise, as described in Example 3.1.

Corollary 3.1. *Suppose that we observe $Z = X + W$, where the random matrices $X, W \in \mathbb{R}^{n \times p}$ are sub-Gaussian with parameters (Σ_x, σ_x^2), and let ϵ be an i.i.d. sub-Gaussian vector with parameter σ_ϵ^2. Let $\sigma_z^2 = \sigma_x^2 + \sigma_w^2$. Then under the scaling $n \gtrsim \max\left\{\frac{\sigma_z^4}{\lambda_{\min}^2(\Sigma_x)}, 1\right\} k \log p$, for the M-estimator based on the surrogates $(\widehat{\Gamma}_{add}, \widehat{\gamma}_{add})$, the results of Theorems 3.1 and 3.2 hold with parameters $\alpha_\ell = \frac{1}{2}\lambda_{\min}(\Sigma_x)$ and $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0 \sigma_z(\sigma_w + \sigma_\epsilon)\|\beta^*\|_2$, with probability at least $1 - c_1 \exp(-c_2 \log p)$.*

Remarks

(a) Consequently, the ℓ_2-error of any optimal solution $\widehat{\beta}$ satisfies the bound

$$\|\widehat{\beta} - \beta^*\|_2 \precsim \frac{\sigma_z(\sigma_w + \sigma_\epsilon)}{\lambda_{\min}(\Sigma_x)} \|\beta^*\|_2 \sqrt{\frac{k \log p}{n}}$$

with high probability. The prefactor in this bound has a natural interpretation as an inverse signal-to-noise ratio; for instance, when X and W are zero-mean Gaussian matrices with row covariances $\Sigma_x = \sigma_x^2 I$ and $\Sigma_w = \sigma_w^2 I$, respectively, we have $\lambda_{\min}(\Sigma_x) = \sigma_x^2$, so

$$\frac{(\sigma_w + \sigma_\epsilon)\sqrt{\sigma_x^2 + \sigma_w^2}}{\lambda_{\min}(\Sigma_x)} = \frac{\sigma_w + \sigma_\epsilon}{\sigma_x}\sqrt{1 + \frac{\sigma_w^2}{\sigma_x^2}}.$$

This quantity grows with the ratios σ_w/σ_x and σ_ϵ/σ_x, which measure the SNR of the observed covariates and predictors, respectively. Note that when $\sigma_w = 0$, corresponding to the case of uncorrupted covariates, the bound on ℓ_2-error agrees with known results. See Section 3.4 for simulations and further discussions of the consequences of Corollary 3.1.

(b) We may also compare the results in (a) with bounds from past work on high-dimensional sparse regression with noisy covariates [77]. In this work, Rosenbaum and Tsybakov derive similar concentration bounds on sub-Gaussian matrices. The tolerance parameters are all $\mathcal{O}(\sqrt{\frac{\log p}{n}})$, with prefactors depending on the sub-Gaussian parameters of the matrices. In particular, in their notation,

$$\nu \asymp (\sigma_x\sigma_w + \sigma_w\sigma_\epsilon + \sigma_w^2)\sqrt{\frac{\log p}{n}}\|\beta^*\|_1,$$

leading to the bound (cf. Theorem 2 of Rosenbaum and Tsybakov [77])

$$\|\widehat{\beta} - \beta^*\|_2 \precsim \frac{\nu\sqrt{k}}{\lambda_{\min}(\Sigma_x)} \asymp \frac{\sigma^2}{\lambda_{\min}(\Sigma_x)}\sqrt{\frac{k\log p}{n}}\|\beta^*\|_1.$$

Extensions to unknown noise covariance: Situations may arise where the noise covariance Σ_w is unknown, and must be estimated from the data. One simple method is to assume that Σ_w is estimated from independent observations of the noise. In this case, suppose we independently observe a matrix $W_0 \in \mathbb{R}^{n\times p}$ with n i.i.d. vectors of noise. Then we use $\widehat{\Sigma}_w = \frac{1}{n}W_0^T W_0$ as our estimate of Σ_w. A more sophisticated variant of this method (cf. Chapter 4 of Carroll et al. [18]) assumes that we observe k_i replicate measurements $Z_{i1}, \ldots, Z_{ik}$ for each x_i and form the estimator

$$\widehat{\Sigma}_w = \frac{\sum_{i=1}^n \sum_{j=1}^{k_i}(Z_{ij} - \overline{Z}_{i\cdot})(Z_{ij} - \overline{Z}_{i\cdot})^T}{\sum_{i=1}^n (k_i - 1)}. \tag{3.23}$$

Based on the estimator $\widehat{\Sigma}_w$, we form the pair $(\widetilde{\Gamma}, \widetilde{\gamma})$ such that $\widetilde{\gamma} = \frac{1}{n}Z^T y$ and $\widetilde{\Gamma} = \frac{Z^T Z}{n} - \widehat{\Sigma}_w$. In the proofs of Section 3.6, we will analyze the case where $\widehat{\Sigma}_w = \frac{1}{n}W_0^T W_0$ and show that the result of Corollary 3.1 still holds when Σ_w must be estimated from the data. Note that the estimator in equation (3.23) will also yield the same result, but the analysis is more complicated.

3.3.2.2 Bounds for missing data: i.i.d. case

Next, we turn to the case of i.i.d. samples with missing data, as discussed in Example 3.3. For a missing data parameter vector $\boldsymbol{\alpha}$, we define $\alpha_{\max} := \max_j \alpha_j$, and assume $\alpha_{\max} < 1$.

Corollary 3.2. *Let $X \in \mathbb{R}^{n \times p}$ be sub-Gaussian with parameters (Σ_x, σ_x^2), and Z the missing data matrix with parameter $\boldsymbol{\alpha}$. Let ϵ be an i.i.d. sub-Gaussian vector with parameter σ_ϵ^2. If $n \gtrsim \max\big(\frac{1}{(1-\alpha_{\max})^4}\ \frac{\sigma_x^4}{\lambda_{\min}^2(\Sigma_x)}, 1\big) k \log p$, then Theorems 3.1 and 3.2 hold with probability at least $1 - c_1 \exp(-c_2 \log p)$ for $\alpha_\ell = \frac{1}{2}\lambda_{\min}(\Sigma_x)$ and $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0 \frac{\sigma_x}{1-\alpha_{\max}}\big(\sigma_\epsilon + \frac{\sigma_x}{1-\alpha_{\max}}\big)\|\beta^*\|_2$.*

Remarks Suppose X is a Gaussian random matrix and $\alpha_j = \alpha$ for all j. In this case, the ratio $\frac{\sigma_x^2}{\lambda_{\min}(\Sigma_x)} = \frac{\lambda_{\max}(\Sigma_x)}{\lambda_{\min}(\Sigma_x)} = \kappa(\Sigma_x)$ is the condition number of Σ_x. Then

$$\frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\alpha} \asymp \left(\frac{1}{\lambda_{\min}(\Sigma_x)} \frac{\sigma_x \sigma_\epsilon}{1-\alpha} + \frac{\kappa(\Sigma_x)}{(1-\alpha)^2}\right) \|\beta^*\|_2,$$

a quantity that depends on both the conditioning of Σ_x, and the fraction $\alpha \in [0, 1)$ of missing data. We will consider the results of Corollary 3.2 applied to this example in the simulations of Section 3.4.

Extensions to unknown α: As in the additive noise case, we may wish to consider the case when the missing data parameters $\boldsymbol{\alpha}$ are not observed and must be estimated from the data. For each $j = 1, 2, \ldots, p$, we estimate α_j using $\widehat{\alpha}_j$, the empirical average of the number of observed entries per column. Let $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^p$ denote the resulting estimator of $\boldsymbol{\alpha}$. Naturally, we use the pair of estimators $(\widetilde{\Gamma}, \widetilde{\gamma})$ defined by

$$\widetilde{\Gamma} = \frac{Z^T Z}{n} \oslash \widetilde{M} \quad \text{and} \quad \widetilde{\gamma} = \frac{1}{n} Z^T y \oslash (\mathbf{1} - \widehat{\boldsymbol{\alpha}}), \tag{3.24}$$

where

$$\widetilde{M}_{ij} = \begin{cases} (1 - \widehat{\alpha}_i)(1 - \widehat{\alpha}_j) & \text{if } i \neq j \\ 1 - \widehat{\alpha}_i & \text{if } i = j. \end{cases}$$

We will show in Section 3.6 that Corollary 3.2 holds when $\boldsymbol{\alpha}$ is estimated by $\widehat{\boldsymbol{\alpha}}$.

3.3.2.3 Bounds for dependent data

Turning to the case of dependent data, we consider the setting where the rows of X are drawn from a stationary vector autoregressive (VAR) process according to

$$x_{i+1} = A x_i + v_i, \qquad \text{for } i = 1, 2, \ldots, n-1, \tag{3.25}$$

where $v_i \in \mathbb{R}^p$ is a zero-mean noise vector with covariance matrix Σ_v, and $A \in \mathbb{R}^{p \times p}$ is a driving matrix with spectral norm $|||A|||_2 < 1$. We assume the rows of X are drawn from a Gaussian distribution with covariance Σ_x, such that $\Sigma_x = A\Sigma_x A^T + \Sigma_v$. Hence, the rows of X are identically distributed but not independent, with the choice $A = 0$ giving rise to the i.i.d. scenario. Corollaries 3.3 and 3.4 correspond to the case of additive noise and missing data for a Gaussian VAR process.

Corollary 3.3. *Suppose the rows of X are drawn according to a Gaussian VAR process with driving matrix A. Suppose the additive noise matrix W is i.i.d. with Gaussian rows, and let ϵ be an i.i.d. sub-Gaussian vector with parameter σ_ϵ^2. If $n \gtrsim \max\left(\frac{\zeta^4}{\lambda_{\min}^2(\Sigma_x)}, 1\right) k \log p$, with $\zeta^2 = |||\Sigma_w|||_{op} + \frac{2|||\Sigma_x|||_{op}}{1-|||A|||_{op}}$, then the results of Theorems 3.1 and 3.2 hold with probability at least $1 - c_1 \exp(-c_2 \log p)$ for $\alpha_\ell = \frac{1}{2}\lambda_{\min}(\Sigma_x)$ and $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0(\sigma_\epsilon \zeta + \zeta^2)\|\beta^*\|_2$.*

Corollary 3.4. *Suppose the rows of X are drawn according to a Gaussian VAR process with driving matrix A, and Z is the observed matrix subject to missing data, with parameter $\boldsymbol{\alpha}$. Let ϵ be an i.i.d. sub-Gaussian vector with parameter σ_ϵ^2. If $n \gtrsim \max\left(\frac{\zeta'^4}{\lambda_{\min}^2(\Sigma_x)}, 1\right) k \log p$, with $\zeta'^2 = \frac{1}{(1-\alpha_{\max})^2}\frac{2|||\Sigma_x|||_{op}}{1-|||A|||_{op}}$, then the results of Theorems 3.1 and 3.2 hold with probability at least $1 - c_1 \exp(-c_2 \log p)$ for $\alpha_\ell = \frac{1}{2}\lambda_{\min}(\Sigma_x)$ and $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0(\sigma_\epsilon \zeta' + \zeta'^2)\|\beta^*\|_2$.*

Remark 3.1. *Note that the scaling and the form of φ in Corollaries 2-4 are very similar, except with different effective variances $\sigma^2 = \frac{\sigma_x^2}{(1-\alpha_{\max})^2}$, ζ^2, or ζ'^2, depending on the type of corruption in the data. As we will see in Section 3.6, the proofs involve verifying the deviation conditions (3.17) using similar techniques. On the other hand, the proof of Corollary 1 proceeds via deviation condition (3.16), which produces a tighter bound.*

We may also extend the cases of dependent data to situations when Σ_w and $\boldsymbol{\alpha}$ are unknown and must be estimated from the data. The proofs of these extensions are identical to the i.i.d case, so we will omit them.

3.3.3 Application to graphical model inverse covariance estimation

The problem of inverse covariance estimation for a Gaussian graphical model is also related to the Lasso. Meinshausen and Bühlmann [60] prescribed a way to recover the support of the precision matrix Θ when each column of Θ is k-sparse, via linear regression and the Lasso. More recently, Yuan [99] proposed a method for estimating Θ using the Dantzig selector, and obtained error bounds on $|||\widehat{\Theta} - \Theta|||_1$ when the columns of Θ are bounded in ℓ_1. Both of these results assume that X is fully-observed and has i.i.d. rows.

Suppose we are given a matrix $X \in \mathbb{R}^{n \times p}$ of samples from a multivariate Gaussian distribution, where each row is distributed according to $N(0, \Sigma)$. We assume the rows of X are either i.i.d. or sampled from a Gaussian VAR process. Based on the modified Lasso of the previous section, we devise a method to estimate Θ based on a corrupted observation

matrix Z, when Θ is sparse. Our method bears similarity to the method of Yuan [99], but is valid in the case of corrupted data, and does not require an ℓ_1 column bound. Let X^j denote the j^{th} column of X, and let X^{-j} denote the matrix X with j^{th} column removed. By standard results on Gaussian graphical models, there exists a vector $\theta^j \in \mathbb{R}^{p-1}$ such that

$$X^j = X^{-j}\theta^j + \epsilon^j, \tag{3.26}$$

where ϵ^j is a vector of i.i.d. Gaussians and $\epsilon^j \perp\!\!\!\perp X^{-j}$. Setting $a_j := -(\Sigma_{jj} - \Sigma_{j,-j}\theta^j)^{-1}$, we can verify that $\Theta_{j,-j} = a_j\theta^j$. Our algorithm, described below, forms estimates $\widehat{\theta^j}$ and $\widehat{a}_j$ for each j, then combines the estimates to obtain an estimate $\widehat{\Theta}_{j,-j} = \widehat{a}_j\widehat{\theta^j}$.

In the additive noise case, we observe the matrix $Z = X + W$. From the equations (3.26), we obtain $Z^j = X^{-j}\theta^j + (\epsilon^j + W^j)$. Note that $\delta^j = \epsilon^j + W^j$ is a vector of i.i.d. Gaussians, and since $X \perp\!\!\!\perp W$, we have $\delta^j \perp\!\!\!\perp X^{-j}$. Hence, our results on covariates with additive noise allow us to recover θ^j from Z. We can verify that this reduces to solving the program (3.4) or (3.7) with the pair $(\widehat{\Gamma}^{(j)}, \widehat{\gamma}^{(j)}) = (\widehat{\Sigma}_{-j,-j}, \frac{1}{n}Z^{-jT}Z^j)$, where $\widehat{\Sigma} = \frac{1}{n}Z^TZ - \Sigma_w$.

When Z is a missing-data version of X, we similarly estimate the vectors θ^j via equation (3.26), using our results on the Lasso with missing covariates. Here, both covariates and responses are subject to missing data, but this makes no difference in our theoretical results. For each j, we use the pair

$$(\widehat{\Gamma}^{(j)}, \widehat{\gamma}^{(j)}) = \big(\widehat{\Sigma}_{-j,-j},\ \frac{1}{n}Z^{-jT}Z^j \oplus (\mathbf{1} - \boldsymbol{\alpha}^{-j})(1 - \alpha_j)\big),$$

where $\widehat{\Sigma} = \frac{1}{n}Z^TZ \oplus M$, and M is defined as in Example 3.3.

To obtain the estimate $\widehat{\Theta}$, we therefore propose the following procedure, based on the estimators $\{(\widehat{\Gamma}^{(j)}, \widehat{\gamma}^{(j)})\}_{j=1}^p$ and $\widehat{\Sigma}$.

Algorithm 3.1. *(1) Perform p linear regressions of the variables Z^j upon the remaining variables Z^{-j}, using the program (3.4) or (3.7) with the estimators $(\widehat{\Gamma}^{(j)}, \widehat{\gamma}^{(j)})$, to obtain estimates $\widehat{\theta^j}$ of θ^j.*

(2) Estimate the scalars a_j using the quantity $\widehat{a}_j := -(\widehat{\Sigma}_{jj} - \widehat{\Sigma}_{j,-j}\widehat{\theta^j})^{-1}$, based on the estimator $\widehat{\Sigma}$. Form $\widetilde{\Theta}$ with $\widetilde{\Theta}_{j,-j} = \widehat{a}_j\widehat{\theta^j}$ and $\widetilde{\Theta}_{jj} = -\widehat{a}_j$.

(3) Set $\widehat{\Theta} = \arg\min_{\Theta \in S^p} \|\Theta - \widetilde{\Theta}\|_1$, where S^p is the set of symmetric matrices.

Note that the minimization in step (3) is a linear program, so is easily solved with standard methods. We have the following corollary about $\widehat{\Theta}$:

Corollary 3.5. *Suppose the columns of the matrix Θ are k-sparse, and suppose the condition number $\kappa(\Theta)$ is nonzero and finite. Suppose we have*

$$\|\widehat{\gamma}^{(j)} - \widehat{\Gamma}^{(j)}\theta^j\|_\infty \le \varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}, \quad \forall j, \tag{3.27}$$

and suppose we have the following additional deviation condition on $\widehat{\Sigma}$*:*

$$\|\widehat{\Sigma} - \Sigma\|_{\max} \leq c\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}. \tag{3.28}$$

Finally, suppose the lower-RE condition holds uniformly over the matrices $\widehat{\Gamma}^{(j)}$ *with the scaling* (3.18)*. Then under the estimation procedure of Algorithm 3.1, there exists a universal constant* c_0 *such that*

$$|||\widehat{\Theta} - \Theta|||_{op} \leq \frac{c_0\kappa^2(\Sigma)}{\lambda_{\min}(\Sigma)}\Big(\frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\lambda_{\min}(\Sigma)} + \frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\alpha_\ell}\Big)k\sqrt{\frac{\log p}{n}}.$$

Remark 3.2. *Note that Corollary 3.5 is again a deterministic result, with parallel structure to Theorem 3.1. Furthermore, the deviation bounds* (3.27) *and* (3.28) *hold for all scenarios considered in Section 3.3.2 above, using Corollaries 1-4 for the first two inequalities, and a similar bounding technique for* $\|\widehat{\Sigma} - \Sigma\|_{\max}$*; and the lower-RE condition holds over all matrices* $\widehat{\Gamma}^{(j)}$ *by the same technique used to establish the lower-RE condition for* $\widehat{\Gamma}$*. The uniformity of the lower-RE bound over all sub-matrices holds because*

$$0 < \lambda_{\min}(\Sigma) \leq \lambda_{\min}(\Sigma_{-j,-j}) \leq \lambda_{\max}(\Sigma_{-j,-j}) \leq \lambda_{\max}(\Sigma) < \infty.$$

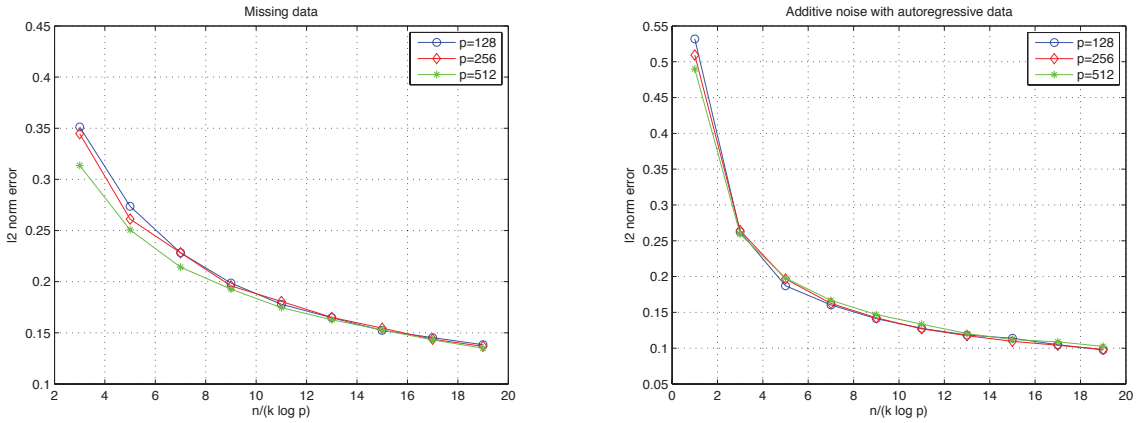
Hence, the error bound in Corollary 3.5 holds with probability at least $1 - c_1\exp(-c_2\log p)$ *when* $n \gtrsim k\log p$*, for the appropriate values of* φ *and* α_ℓ*.*

3.4 Simulations

In this section, we report some additional simulation results to confirm that the scalings predicted by our theory are sharp. In Figure 3.1 following Theorem 3.1, we showed that the error curves align when plotted against a suitably rescaled sample size, in the case of additive noise perturbations. Panel (a) of Figure 3.3 shows these same types of rescaled curves for the case of missing data, with sparsity $k \approx \sqrt{p}$, covariate matrix $\Sigma_x = I$, and missing fraction $\alpha = 0.2$, whereas panel (b) shows the rescaled plots for the vector autoregressive case with additive noise perturbations, using a driving matrix A with $|||A|||_{\text{op}} = 0.2$. Each point corresponds to an average over 100 trials. Once again, we see excellent agreement with the scaling law provided by Theorem 3.1.

We also ran simulations to verify the form of the function $\varphi(\mathbb{Q}, \sigma_\epsilon)$ appearing in Corollaries 3.1 and 3.2. In the additive noise setting for i.i.d. data, we set $\Sigma_x = I$ and ϵ equal to i.i.d. Gaussian noise with $\sigma_\epsilon = 0.5$. For a fixed value of the parameters $p = 256$ and $k \approx \log p$, we ran the projected gradient descent algorithm for different values of $\sigma_w \in (0.1, 0.3)$, such that $\Sigma_w = \sigma_w^2 I$ and $n \approx 60(1 + \sigma_w^2)^2 k\log p$, with $\|\beta^*\|_2 = 1$. According to the theory, $\frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\alpha} \asymp (\sigma_w + 0.5)\sqrt{1 + \sigma_w^2}$, so that

$$\|\widehat{\beta} - \beta^*\|_2 \precsim (\sigma_w + 0.5)\sqrt{1 + \sigma_w^2}\sqrt{\frac{k\log p}{(1 + \sigma_w^2)^2 k\log p}} \asymp \frac{\sigma_w + 0.5}{\sqrt{1 + \sigma_w^2}}.$$



(a) (b)

Figure 3.3: Plots of the error $\|\widehat{\beta} - \beta^*\|_2$ after running projected gradient descent on the nonconvex objective, with sparsity $k \approx \sqrt{p}$. In all cases, we plotted the error versus the rescaled sample size $\frac{n}{k \log p}$. As predicted by Theorems 3.1 and 3.2, the curves align for different values of p when plotted in this rescaled manner. (a) Missing data case with i.i.d. covariates. (b) Vector autoregressive data with additive noise. Each point represents an average over 100 trials.

In order to verify this prediction, we plotted σ_w versus the rescaled error $\frac{\sqrt{1+\sigma_w^2}}{\sigma_w + 0.5}\|\widehat{\beta} - \beta^*\|_2$. As shown by panel (a) of Figure 3.4(a), the curve is roughly constant, as predicted by the theory.

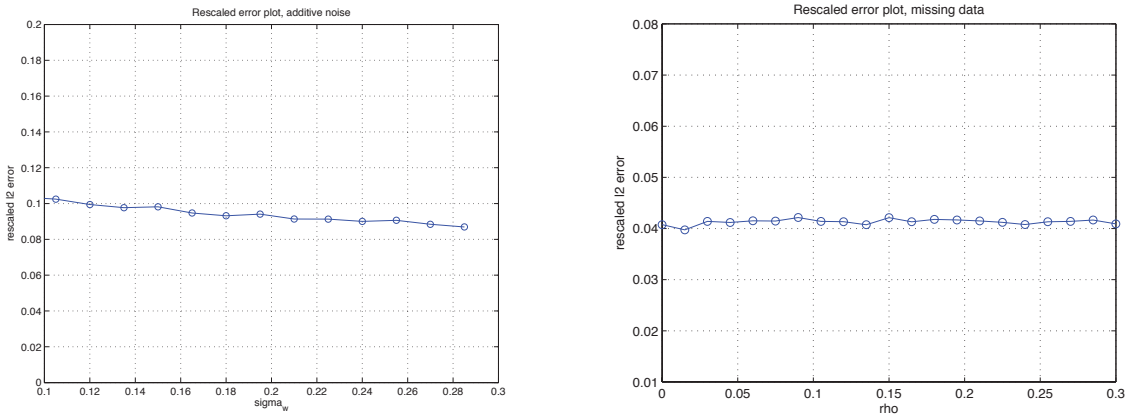
Similarly, in the missing data setting for i.i.d. data, we set $\Sigma_x = I$ and ϵ equal to i.i.d. Gaussian noise with $\sigma_\epsilon = 0.5$. For a fixed value of the parameters $p = 128$ and $k \approx \log p$, we ran simulations for different values of the missing data parameter $\alpha \in (0, 0.3)$, such that $n \approx \frac{60}{(1-\alpha)^4} k \log p$. According to the theory, $\frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\alpha} \asymp \frac{\sigma_\epsilon}{1-\alpha} + \frac{1}{(1-\alpha)^2}$. Consequently, with our specified scalings of (n, p, k), we should expect a bound of the form

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\alpha} \sqrt{\frac{k \log p}{n}} \asymp 1 + 0.5(1 - \alpha).$$

The plot of α versus the rescaled error $\frac{\|\widehat{\beta} - \beta^*\|_2}{1 + 0.5(1-\alpha)}$ is shown in Figure 3.4(b). The curve is again roughly constant, agreeing with theoretical results.

Finally, we studied the behavior of the inverse covariance matrix estimation algorithm on three types of Gaussian graphical models:

(a) *Chain-structured graphs.* In this case, all nodes of the graph are arranged in a linear chain. Hence, each node (except the two end nodes) has degree $k = 2$. The diagonal



(a) Additive noise scaling (b) Missing data scaling

Figure 3.4: (a) Plot of the rescaled ℓ_2-error $\frac{\sqrt{1+\sigma_w^2}}{\sigma_w+0.5}\|\widehat{\beta}-\beta^*\|_2$ versus the additive noise standard deviation σ_w for the i.i.d. model with additive noise. (b) Plot of the rescaled ℓ_2-error $\frac{\|\widehat{\beta}-\beta^*\|_2}{1+0.5(1-\alpha)}$ versus the missing fraction α for the i.i.d. model with missing data. Both curves are roughly constant, showing that our error bounds on $\|\widehat{\beta}-\beta^*\|_2$ exhibit the proper scaling. Each point represents an average over 200 trials.

entries of Θ are set equal to 1, and all entries corresponding to links in the chain are set equal to 0.1. Then Θ is rescaled so $|||\Theta|||_{\text{op}} = 1$.

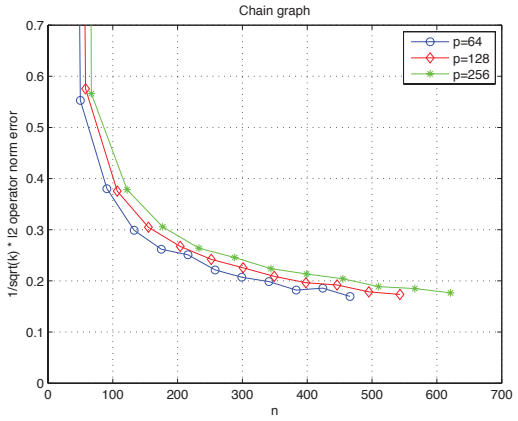
(b) *Star-structured graphs.* In this case, all nodes are connected to a central node, which has degree $k \approx 0.1p$. All other nodes have degree 1. The diagonal entries of Θ are set equal to 1, and all entries corresponding to edges in the graph are set equal to 0.1. Then Θ is rescaled so $|||\Theta|||_{\text{op}} = 1$.

(c) *Erdös-Renyi graphs.* This example comes from Rothman et al. [78]. For a sparsity parameter $k \approx \log p$, we randomly generate the matrix Θ by first generating the matrix B such that the diagonal entries are 0, and all other entries are independently equal to 0.5 with probability k/p, and 0 otherwise. Then δ is chosen so that $\Theta = B + \delta I$ has condition number p. Finally, Θ is rescaled so $|||\Theta|||_{\text{op}} = 1$.

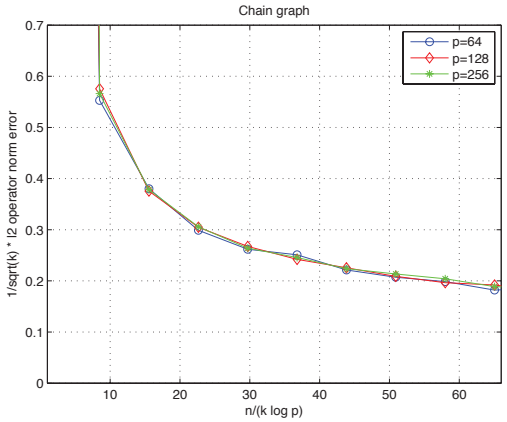
After generating the matrix X of n i.i.d. samples from the appropriate graphical model, with covariance matrix $\Sigma_x = \Theta^{-1}$, we generated the corrupted matrix $Z = X + W$ with $\Sigma_w = (0.2)^2 I$ in the additive noise case, or the missing data matrix Z with $\alpha = 0.2$ in the missing data case.

Panels (a) and (c) in Figure 3.5 show the rescaled ℓ_2-error $\frac{1}{\sqrt{k}}|||\widehat{\Theta} - \Theta|||_{\text{op}}$ plotted against the sample size n for a chain-structured graph. In panels (b) and (d), we have ℓ_2-error plotted against the rescaled sample size, $n/(k \log p)$. Once again, we see good agreement

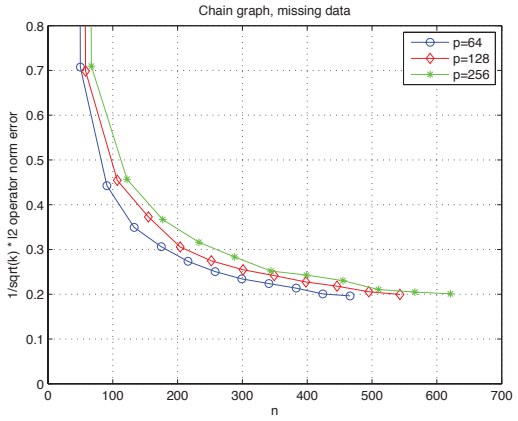
with the theoretical predictions. We have obtained qualitatively similar results for the star and Erdös-Renyi graphs.



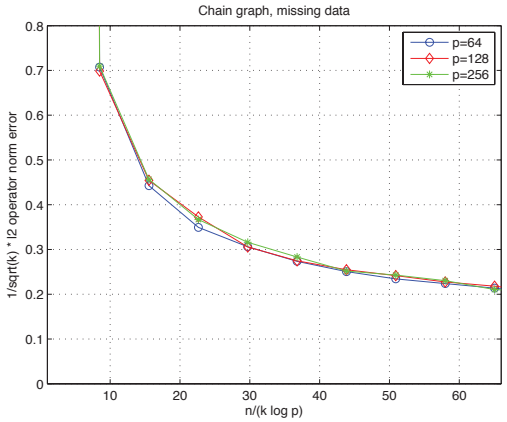
(a) ℓ_2 error plot for chain graph, additive noise



(b) rescaled plot



(c) ℓ_2 error plot for chain graph, missing data



(d) rescaled plot

Figure 3.5: (a) Plots of the error $|||\widehat{\Theta} - \Theta|||_{\mathrm{op}}$ after running projected gradient descent on the nonconvex objective for a chain-structured Gaussian graphical model with additive noise. As predicted by Theorems 3.1 and 3.2, all curves align when the error is rescaled by $\frac{1}{\sqrt{k}}$ and plotted against the ratio $\frac{n}{k \log p}$, as shown in (b). Plots (c) and (d) show the results of simulations on missing data sets. Each point represents the average over 50 trials.

3.5 Lower bounds

We now focus on fundamental information-theoretic limitations of prediction under various forms of corrupted covariates. Our approach consists of a two-pronged attack: On the

statistical side, we demonstrate an efficient estimator for our model and prove upper bounds on ℓ_2-error between the estimator and the population parameter that slightly sharpen the results of Section 3.3; while on the information-theoretic side, we establish lower bounds on ℓ_2-error that hold for *any* estimator derived from the data. Our upper and lower bounds in the additive noise setting agree up to constant factors, demonstrating that our proposed estimator is minimax optimal.

To compare with the upper bounds in Section 3.3, here we improve the asymptotic scaling in the squared ℓ_2-error from $\frac{k \log p}{n}$ to $\frac{k \log(p/k)}{n}$, and tighten the prefactor so it achieves known minimax results in the limit of no corruption. However, whereas the upper bounds in Section 3.3 apply to arbitrary sub-Gaussian variables with nondiagonal covariances, the lower bounds derived in this section only apply when covariates are Gaussian and covariances are multiples of the identity. Our proof techniques for lower bounds closely follow those of Raskutti et al. [71].

3.5.1 Problem setup

We again focus on the linear regression model

$$y_i = \langle x_i, \beta^* \rangle + \epsilon_i, \qquad \text{for } i = 1, 2, \ldots, n,$$

where the x_i's are p-dimensional covariates, the y_i's are response variables, the ϵ_i's are independent noise, and $\beta^* \in \mathbb{R}^p$ is the unknown vector. In matrix form, we write $y = X\beta^* + \epsilon$, where $X \in \mathbb{R}^{n \times p}$ and $y, \epsilon \in \mathbb{R}^n$. Since we are working in a high-dimensional setting ($p \gg n$), we must impose additional structure on β^*. Henceforth, we assume that $\|\beta^*\|_0 \leq k$, meaning β^* has at most k nonzero entries.

In the traditional linear regression framework, one would estimate β^* based on observations (y, X). However, we assume that only the pair (y, Z) is available, where Z is a version of X corrupted by noise. We analyze the following settings:

(a) *Additive noise:* For each i, observe $z_i = x_i + w_i$, where w_i is independent of x_i.

(b) *Missing data:* For each i and each component j, independently observe $z_{ij} = x_{ij}$ with probability $1 - \alpha$, and $z_{ij} = \star$ with probability α, where $\alpha \in [0, 1)$.

In both cases, we assume the x_i's and ϵ_i's are drawn i.i.d. from the distributions $N(0, \sigma_x^2 I)$ and $N(0, \sigma_\epsilon^2 I)$, respectively. We assume the w_i's are drawn i.i.d. from $N(0, \sigma_w^2 I)$ in the additive noise case.

Our analysis focuses on the *minimax squared ℓ_2-error*

$$\mathcal{M}(n, p, k) := \inf_{\widehat{\beta}} \sup_{\beta^* \in \mathbb{B}_0(k) \cap \mathbb{B}_2(1)} \|\widehat{\beta} - \beta^*\|_2^2.$$

Note that the supremum is taken over k-sparse vectors in the ℓ_2 unit ball, whereas the infimum is taken over all measureable functions $\widehat{\beta}$ of the observed data (y, Z). In Theorems 3.3

and 3.5, we derive upper bounds for $\mathcal{M}$ by analyzing a modified version of the Lasso for corrupted covariates. In Theorems 3.4 and 3.6, we derive lower bounds via information-theoretic techniques, where we first reduce the estimation problem to a hypothesis testing problem and then apply Fano's inequality to lower-bound the error probability. This type of reduction is standard in minimax statistical analysis (e.g., [9, 98, 97]).

3.5.2 Main results and consequences

We now state our main results. Following Section 3.3, we define the surrogate $\widehat{\Gamma} \in \mathbb{R}^{p \times p}$ for Σ_x, defined in the additive noise and missing cases as

(a) $\widehat{\Gamma} = \frac{Z^T Z}{n} - \Sigma_w$,

(b) $\widehat{\Gamma} = \frac{\widehat{Z}^T \widehat{Z}}{n} - \alpha \operatorname{diag}\left(\frac{\widehat{Z}^T \widehat{Z}}{n}\right), \quad \widehat{Z} = \frac{Z}{1-\alpha}$,

respectively. We assume $\widehat{\Gamma}$ obeys the lower-RE condition:

Assumption 3.1 (Lower-RE condition). *For some $\alpha_\ell > 0$, we have $\theta^T \widehat{\Gamma} \theta \geq \alpha_\ell \|\theta\|_2^2$ whenever $\|\theta\|_1 \leq c_0 \sqrt{k} \|\theta\|_2$.*

By Lemmas A.1 and A.3 in Appendix A.1, Assumption 3.1 holds w.h.p. for $\alpha_\ell \asymp \sigma_x^2$ in both settings of interest.

3.5.3 Additive noise setting

We begin by stating an upper bound for the additive noise setting, when X and W are Gaussian with covariance $\sigma_x^2 I$ and $\sigma_w^2 I$, respectively. We write $\sigma_z^2 := \sigma_x^2 + \sigma_w^2$ and $\kappa := \frac{\sigma_w^2}{\sigma_x^2}$.

Theorem 3.3. *In the additive noise setting, if $\widehat{\Gamma}$ satisfies Assumption 3.1 and $n \gtrsim k \log(p/k)$, we have*

$$\mathcal{M} \leq \frac{c((1+\kappa)\sigma_x^2 \sigma_w^2 + \sigma_\epsilon^2 \sigma_z^2)}{\alpha_\ell^2} \frac{k \log(p/k)}{n}, \tag{3.29}$$

with probability at least $1 - c_1 \exp(-c_2 k \log(p/k))$.

Note that when $\sigma_w = 0$, corresponding to the classical case of fully-observed covariates, the upper bound reduces to

$$\frac{c \sigma_\epsilon^2 \sigma_x^2}{\alpha_\ell^2} \frac{k \log(p/k)}{n}.$$

Past work has established bounds of this form for the Lasso and related estimators [14, 8], and this rate has been shown to be minimax optimal [71]. In the more general setting with $\sigma_w > 0$, the bound (3.29) has a qualitatively similar form, with the prefactor growing with the magnitude of σ_w.

We now turn to a lower bound that matches the upper bound up to a constant factor. The probability for the lower bound is chosen to be 1/2; it may be replaced by a constant arbitrarily close to 1, by a suitable modification of the universal constants.

Theorem 3.4. *In the additive noise setting, if* $8 \le k \le p/2$ *and* $n \gtrsim k \log(p/k)$*, we have*

$$\mathcal{M} \ge \frac{c'(\sigma_x^2\sigma_w^2 + \sigma_\epsilon^2\sigma_z^2)}{\sigma_x^4} \frac{k \log(p/k)}{n}, \tag{3.30}$$

with probability at least 1/2*.*

Note in particular that when the $\kappa = \frac{\sigma_w^2}{\sigma_x^2}$ is bounded above by a constant and $\alpha_\ell \asymp \sigma_x^2$, the bounds in Theorems 3.3 and 3.4 match up to constant factors, identifying minimax optimal rates for the additive noise setting. The assumption of bounded κ merely requires the SNR to be bounded away from zero.

3.5.4 Missing data setting

In the missing data setting, we assume $x_i \sim N(0, \sigma_x^2 I)$, and $\alpha \in [0, 1)$ is the probability that a given entry is missing. We have the following upper bound:

Theorem 3.5. *In the missing data setting, suppose* $\widehat{\Gamma}$ *satisfies Assumption 3.1 and the sample size satisfies* $n \gtrsim \frac{1}{(1-\alpha)^2} k \log(p/k)$*. Then*

$$\mathcal{M} \le \frac{c\sigma_x^2}{\alpha_\ell^2} \left(\sigma_\epsilon + \frac{\alpha \sigma_x}{1-\alpha} \right)^2 \frac{k \log(p/k)}{n},$$

with probability at least $1 - c_2 \exp(-c_2 k \log(p/k))$*.*

For a lower bound, we have the following:

Theorem 3.6. *In the missing data setting, if* $8 \le k \le p/2$ *and* $n \gtrsim \frac{1}{(1-\alpha)^2} k \log(p/k)$*, we have*

$$\mathcal{M} \ge \frac{c\sigma_\epsilon^2}{\sigma_x^2(1-\alpha)} \frac{\sigma_\epsilon^2}{\sigma_x^2 + \sigma_\epsilon^2} \frac{k \log(p/k)}{n}, \tag{3.31}$$

with probability at least 1/2*.*

Note that when $\alpha = 0$, corresponding to no missing data, Theorem 3.5 again reduces to known results. Furthermore, both the upper and lower bounds grow as the inverse of $(1-\alpha)$, agreeing with intuition—as the proportion of missing entries increases, the estimation problem increases in difficulty. However, a gap of a factor of $(1-\alpha)$ remains between the scaling in Theorems 3.5 and 3.6.

3.6 Proofs

In this section, we turn to the proofs of our main theorems in Sections 3.3 and 3.5. The proofs of corollaries and more technical lemmas are contained in Appendix A.

3.6.1 Proof of Theorem 3.1

Let $\mathcal{L}(\beta) = \frac{1}{2}\beta^T\widehat{\Gamma}\beta - \langle\widehat{\gamma}, \beta\rangle + \lambda_n\|\beta\|_1$ denote the loss function to be minimized. This definition captures both the estimator (3.4) with $\lambda_n = 0$ and the estimator (3.7) with the choice of λ_n given in the theorem statement. For either estimator, we are guaranteed that β^* is feasible and $\widehat{\beta}$ is optimal for the program, so $\mathcal{L}(\widehat{\beta}) \leq \mathcal{L}(\beta^*)$. Indeed, in the regularized case, the k-sparsity of β^* implies that $\|\beta^*\|_1 \leq \sqrt{k}\|\beta^*\|_2 \leq b_0\sqrt{k}$. Defining the error vector $\widehat{\nu} := \widehat{\beta} - \beta^*$ and performing some algebra leads to the equivalent inequality

$$\frac{1}{2}\widehat{\nu}^T\widehat{\Gamma}\widehat{\nu} \leq \langle\widehat{\nu}, \widehat{\gamma} - \widehat{\Gamma}\beta^*\rangle + \lambda_n\big\{\|\beta^*\|_1 - \|\beta^* + \widehat{\nu}\|_1\big\}. \tag{3.32}$$

In the remainder of the proof, we first derive an upper bound for the right-hand side of this inequality. We then use this upper bound and the lower-RE condition to show that the error vector $\widehat{\nu}$ must satisfy the inequality

$$\|\widehat{\nu}\|_1 \leq 8\sqrt{k}\|\widehat{\nu}\|_2. \tag{3.33}$$

Finally, we combine the inequality (3.33) with the lower-RE condition to derive a lower bound on the left-hand side of the basic inequality (3.32). Combined with our earlier upper bound on the right-hand side, some algebra yields the claim.

Upper bound on right-hand side We first upper-bound the RHS of inequality (3.32). Hölder's inequality gives $\langle\widehat{\nu}, \widehat{\gamma} - \widehat{\Gamma}\beta^*\rangle \leq \|\widehat{\nu}\|_1\|\widehat{\gamma} - \widehat{\Gamma}\beta^*\|_\infty$. By the triangle inequality, we have

$$\|\widehat{\gamma} - \widehat{\Gamma}\beta^*\|_\infty \leq \|\widehat{\gamma} - \Sigma_x\beta^*\|_\infty + \|(\Sigma_x - \widehat{\Gamma})\beta^*\|_\infty \overset{(i)}{\leq} 2\,\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}},$$

where inequality (i) follows from the deviation conditions (3.17). Combining the pieces, we conclude that

$$\langle\widehat{\nu}, \widehat{\gamma} - \widehat{\Gamma}\beta^*\rangle \leq 2\|\widehat{\nu}\|_1\,\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}} = \big(\|\widehat{\nu}_S\|_1 + \|\widehat{\nu}_{S^c}\|_1\big)\,2\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}. \tag{3.34}$$

On the other hand, we have

$$\|\beta^* + \widehat{\nu}\|_1 - \|\beta^*\|_1 \geq \big\{\|\beta^*_S\|_1 - \|\widehat{\nu}_S\|_1\big\} + \|\widehat{\nu}_{S^c}\|_1 - \|\beta^*\|_1 = \|\widehat{\nu}_{S^c}\|_1 - \|\widehat{\nu}_S\|_1, \tag{3.35}$$

where we have exploited the sparsity of β^* and applied the triangle inequality. Combining the pieces, we conclude that the right-hand side of inequality (3.32) is upper-bounded by

$$2\,\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}(\|\widehat{\nu}_S\|_1 + \|\widehat{\nu}_{S^c}\|_1) + \lambda_n\{\|\widehat{\nu}_S\|_1 - \|\widehat{\nu}_{S^c}\|_1\}, \tag{3.36}$$

a bound that holds for any nonnegative choice of λ_n.

Proof of inequality (3.33) In the case of the constrained estimator (3.4) with $R = \|\beta^*\|_1$, we have $\|\widehat{\beta}\|_1 = \|\beta^* + \widehat{\nu}\|_1 \leq \|\beta^*\|_1$. Combined with inequality (3.35), we conclude that $\|\widehat{\nu}_{S^c}\|_1 \leq \|\widehat{\nu}_S\|_1$. Consequently, we have the inequality $\|\widehat{\nu}\|_1 \leq 2\|\widehat{\nu}_S\|_1 \leq 2\sqrt{k}\|\widehat{\nu}\|_2$, which is a slightly stronger form of the bound (3.33).

For the regularized estimator (3.7), we first note that our choice of λ_n guarantees that the term (3.36) is at most $\frac{3\lambda_n}{2}\|\widehat{\nu}_S\|_1 - \frac{\lambda_n}{2}\|\widehat{\nu}_{S^c}\|_1$. Returning to the basic inequality, we apply the lower-RE condition to lower-bound the left-hand side, thereby obtaining the inequality

$$-\frac{\tau}{2}\|\widehat{\nu}\|_1^2 \;\leq\; \frac{1}{2}\big(\alpha_\ell\|\widehat{\nu}\|_2^2 - \tau\,\|\widehat{\nu}\|_1^2\big) \leq \frac{3\lambda_n}{2}\|\widehat{\nu}_S\|_1 - \frac{\lambda_n}{2}\|\widehat{\nu}_{S^c}\|_1.$$

By the triangle inequality, we have $\|\widehat{\nu}\|_1 \leq \|\widehat{\beta}\|_1 + \|\beta^*\|_1 \leq 2b_0\sqrt{k}$. Since we have assumed $\sqrt{k}\,\tau(n,p) \leq \frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{b_0}\sqrt{\frac{\log p}{n}}$, we are guaranteed that

$$\frac{\tau(n,p)}{2}\,\|\widehat{\nu}\|_1^2 \leq \varphi(\mathbb{Q},\sigma_\epsilon)\sqrt{\frac{\log p}{n}}\,\|\widehat{\nu}\|_1 \;\leq\; \frac{\lambda_n}{4}\|\widehat{\nu}\|_1,$$

by our choice of λ_n. Combining the pieces, we conclude that

$$0 \leq \frac{3\lambda_n}{2}\|\widehat{\nu}_S\|_1 - \frac{\lambda_n}{2}\|\widehat{\nu}_{S^c}\|_1 + \frac{\lambda_n}{4}\big(\|\widehat{\nu}_S\|_1 + \|\widehat{\nu}_{S^c}\|_1\big) = \frac{7\lambda_n}{4}\|\widehat{\nu}_S\|_1 - \frac{\lambda_n}{4}\|\widehat{\nu}_{S^c}\|_1,$$

and rearranging implies $\|\widehat{\nu}_{S^c}\|_1 \leq 7\|\widehat{\nu}_S\|_1$, from which we conclude that $\|\widehat{\nu}\|_1 \leq 8\sqrt{k}\|\widehat{\nu}\|_2$, as claimed.

Lower bound on left-hand side We now derive a lower bound on the left-hand side of inequality (3.32). Combining inequality (3.33) with the RE condition (3.12) gives

$$\widehat{\nu}^T\widehat{\Gamma}\widehat{\nu} \geq \alpha_\ell\|\widehat{\nu}\|_2^2 - \tau(n,p)\|\widehat{\nu}\|_1^2 \;\geq\; \big\{\alpha_\ell - 64k\tau(n,p)\big\}\|\widehat{\nu}\|_2^2 \;\geq\; \frac{\alpha_\ell}{2}\,\|\widehat{\nu}\|_2^2, \tag{3.37}$$

where the final step uses our assumption that $k\tau(n,p) \leq \frac{\alpha_\ell}{128}$.

Finally, combining bounds (3.36), (3.33), and (3.37) gives

$$\begin{aligned}\frac{\alpha_\ell}{4}\|\widehat{\nu}\|_2^2 &\leq 2\max\Big\{2\,\varphi(\mathbb{Q},\sigma_\epsilon)\sqrt{\frac{\log p}{n}},\;\lambda_n\Big\}\,\|\widehat{\nu}\|_1 \\ &\leq 32\,\sqrt{k}\max\Big\{\varphi(\mathbb{Q},\sigma_\epsilon)\sqrt{\frac{\log p}{n}},\;\lambda_n\Big\}\,\|\widehat{\nu}\|_2,\end{aligned}$$

yielding inequality (3.19a). Using inequality (3.33) again gives inequality (3.19b).

3.6.2 Proof of Theorem 3.2

We begin by proving the claims for the constrained problem, and projected gradient descent. For the ℓ_2-error bound, we make use of Theorem 1 in the pre-print of Agarwal et al. [1]. Their theory, as originally stated, requires that the loss function be convex, but a careful examination of their proof shows that their arguments hinge on restricted strong convexity and smoothness assumptions, corresponding to a more general version of the lower- and upper-RE conditions given here. Apart from these conditions, the proof exploits the fact that the sub-problems defining the gradient updates (3.14) and (3.15) are convex. Since the loss function itself appears only in a linear term, their theory still applies.

In order to apply Theorem 1 in their paper, we first need to compute the tolerance parameter ϵ^2 defined there; since β^* is supported on the set S with $|S| = k$ and the RE conditions hold with $\tau \asymp \frac{\log p}{n}$, we find that

$$\begin{aligned}
\epsilon^2 &\leq c\,\frac{\log p}{\alpha_u n}\big(\sqrt{k}\|\widehat{\beta} - \beta^*\|_2 + 2\|\widehat{\beta} - \beta^*\|_1\big)^2 \\
&\leq c_2'\,\frac{k\log p}{\alpha_u n}\|\widehat{\beta} - \beta^*\|_2^2 + c_1\frac{\log p}{\alpha_u n}\|\widehat{\beta} - \beta^*\|_1^2 \\
&\leq c_2\|\widehat{\beta} - \beta^*\|_2^2 + c_1\frac{\log p}{\alpha_u n}\|\widehat{\beta} - \beta^*\|_1^2,
\end{aligned}$$

where the final inequality makes use of the assumption that $n \gtrsim k\log p$. Similarly, we may compute the contraction coefficient to be

$$\gamma = \left(1 - \frac{\alpha_\ell}{\alpha_u} + \frac{c_1 k\log p}{\alpha_u n}\right)\left(1 - \frac{c_2 k\log p}{\alpha_u n}\right)^{-1}, \tag{3.38}$$

so $\gamma \in (0,1)$ for $n \gtrsim k\log p$.

We now establish the ℓ_1-error bound. First, let $\Delta^t := \beta^t - \beta^*$. Since β^t is feasible and $\widehat{\beta}$ is optimal with an active constraint, we have $\|\beta^t\|_1 \leq \|\widehat{\beta}\|_1$. Applying the triangle inequality gives

$$\begin{aligned}
\|\widehat{\beta}\|_1 \leq \|\beta^*\|_1 + \|\widehat{\beta} - \beta^*\|_1 = \|\beta^*_S\|_1 + \|\widehat{\beta} - \beta^*\|_1, \\
\|\beta^t\|_1 = \|\beta^* + \Delta^t\|_1 \geq \|\beta^*_S + \Delta^t_{S^c}\|_1 - \|\Delta^t_S\|_1 = \|\beta^*_S\|_1 + \|\Delta^t_{S^c}\|_1 - \|\Delta^t_S\|_1;
\end{aligned}$$

combining the bounds yields $\|\Delta^t_{S^c}\|_1 \leq \|\Delta^t_S\|_1 + \|\widehat{\beta} - \beta^*\|_1$. Then

$$\|\Delta^t\|_1 \leq 2\|\Delta^t_S\|_1 + \|\widehat{\beta} - \beta^*\|_1 \leq 2\sqrt{k}\|\Delta^t\|_2 + \|\widehat{\beta} - \beta^*\|_1,$$

so

$$\|\beta^t - \widehat{\beta}\|_1 \leq \|\widehat{\beta} - \beta^*\|_1 + \|\Delta^t\|_1 \leq 2\sqrt{k}(\|\beta^t - \widehat{\beta}\|_2 + \|\widehat{\beta} - \beta^*\|_2) + 2\|\widehat{\beta} - \beta^*\|_1.$$

Turning to the Lagrangian version, we exploit Theorem 2 in Agarwal et al., with $\mathcal{M}$ corresponding to the subspace of all vectors with support contained within the support set

of β^*. With this choice, we have $\psi(\mathcal{M}) = \sqrt{k}$, and the contraction coefficient γ takes the previous form (3.38), so that the assumption $n \gtrsim k \log p$ guarantees that $\gamma \in (0, 1)$. It remains to verify that the requirements are satisfied. From the conditions in our Theorem 2 and using the notation of Agarwal et al., we have $\beta(\mathcal{M}) = \mathcal{O}(\frac{\log p}{n})$ and $\overline{\alpha} = \sqrt{k}$, and the condition $n \gtrsim k \log p$ implies that $\xi(\mathcal{M}) = \mathcal{O}(1)$. Putting together the pieces, we find that the compound tolerance parameter ϵ^2 satisfies the bound $\epsilon^2 = \mathcal{O}(\frac{k \log p}{n} \|\widehat{\beta} - \beta^*\|_2^2) = \mathcal{O}(\|\widehat{\beta} - \beta^*\|_2^2)$, so the claim follows.

3.6.3 Proof of Theorem 3.3

It suffices to demonstrate an estimator for β^* which, w.h.p., has small ℓ_2-norm error. We use the same estimator (3.7) as before, where $(\widehat{\Gamma}, \widehat{\gamma}) = \left(\frac{Z^T Z}{n} - \Sigma_w, \frac{Z^T y}{n}\right)$ are unbiased estimators for $(\Sigma_x, \mathrm{Cov}(x_i, y_i))$, and the regularization parameter $\lambda \asymp \sqrt{\frac{\log(p/k)}{n}}$ is chosen appropriately. We show that if $\beta^* \in \mathbb{B}_0(k) \cap \mathbb{B}_2(1)$, then $\|\widehat{\beta} - \beta^*\|_2^2$ satisfies the proper upper bound.

Since β^* is feasible and $\widehat{\beta}$ is optimal, we have

$$\frac{1}{2} \widehat{\nu}^T \widehat{\Gamma} \widehat{\nu} \leq \langle \widehat{\nu}, \widehat{\gamma} - \widehat{\Gamma} \beta^* \rangle + \lambda \{\|\beta^*\|_1 - \|\beta^* + \widehat{\nu}\|_1\}, \tag{3.39}$$

where $\widehat{\nu} = \widehat{\beta} - \beta^*$. Since $\|\widehat{\nu}\|_1 \leq 8\sqrt{k}\|\widehat{\nu}\|_2$, we may lower-bound the LHS of inequality (3.39) using Assumption 3.1.

To upper-bound the RHS of inequality (3.39), we use the following combinatorial lemma, a slight generalization of Lemma A.11 in Appendix A.2:

Lemma 3.1. *For any constant $c > 0$, we have*

$$\mathbb{B}_1(c\sqrt{k}) \cap \mathbb{B}_2(1) \subseteq (1 + 2c)\, \mathrm{cl}\{\mathrm{conv}\{\mathbb{B}_0(k) \cap \mathbb{B}_2(1)\}\},$$

where $\mathrm{cl}\{\cdot\}$ *and* $\mathrm{conv}\{\cdot\}$ *denote the topological closure and convex hull, respectively.*

Since $\|\widehat{\nu}\|_1 \leq c\sqrt{k}\|\widehat{\nu}\|_2$, we apply Lemma 3.1 to $u = \frac{\widehat{\nu}}{\|\widehat{\nu}\|_2}$ to obtain

$$u \subseteq (1 + 2c)\, \mathrm{cl}\{\mathrm{conv}\{\mathbb{B}_0(k) \cap \mathbb{B}_2(1)\}\},$$

so

$$|\widehat{\nu}^T(\widehat{\gamma} - \widehat{\Gamma}\beta^*)| \leq (1 + 2c)\|\widehat{\nu}\|_2 \sup_{\substack{u' \in \mathrm{cl}\{\mathrm{conv}\{\mathbb{B}_0(k) \\ \cap \mathbb{B}_2(1)\}\}}} |u'^T(\widehat{\gamma} - \widehat{\Gamma}\beta^*)|.$$

Clearly, the sup may be taken over $u' \in \mathbb{B}_0(k) \cap \mathbb{B}_2(1)$. Furthermore, we may use a standard discretization argument for each support set S' with $|S'| \leq k$, followed by a union bound over all choices of S'. Since the discretization gives a factor of c^k and the union bound gives

a factor of $\binom{p}{k} \leq \left(\frac{p}{k}\right)^k$, it suffices to bound the sup w.h.p. for an arbitrary fixed unit vector $\widetilde{u}$ with $\|\widetilde{u}\|_0 \leq k$. This yields a bound of the form

$$|\widehat{\nu}^T(\widehat{\gamma} - \widehat{\Gamma}\beta^*)| \leq C\varphi\|\widehat{\nu}\|_2\sqrt{\frac{k\log(p/k)}{n}},$$

with probability at least $1 - c_1\exp(-c_2 k\log(p/k))$, where φ is a function of the problem parameters, derived below.

Finally, note that

$$\|\beta^*\|_1 - \|\beta^* + \widehat{\nu}\|_1 \leq \|\widehat{\nu}\|_1 \leq c\sqrt{k}\|\widehat{\nu}\|_2.$$

Combining this with inequality (3.39) and the lower-RE bound then implies

$$\alpha_\ell\|\widehat{\nu}\|_2^2 \leq C\varphi\|\widehat{\nu}\|_2\sqrt{\frac{k\log(p/k)}{n}} + c\sqrt{k}\lambda\|\widehat{\nu}\|_2.$$

Dividing through by $\|\widehat{\nu}\|_2$ yields

$$\|\widehat{\nu}\|_2 \leq \frac{c\sqrt{k}}{\alpha_\ell}\max\left\{\varphi\sqrt{\frac{\log(p/k)}{n}}, \lambda\right\}.$$

Hence, choosing $\lambda \asymp \varphi\sqrt{\frac{\log(p/k)}{n}}$, we obtain the bound $\mathcal{M} \leq \frac{c\varphi^2}{\alpha_\ell^2}\frac{k\log(p/k)}{n}$. The remaining component is to find an appropriate choice of the prefactor φ.

Let $\widetilde{u} \in \mathbb{B}_0(k) \cap \mathbb{B}_2(1)$. Then

$$\begin{aligned}
|\widetilde{u}^T(\widehat{\gamma} - \widehat{\Gamma}\beta^*)| &= \left|\widetilde{u}^T\left(\frac{Z^T y}{n} - \left(\frac{Z^T Z}{n} - \Sigma_w\right)\beta^*\right)\right| \\
&= \left|\widetilde{u}^T\left(\frac{Z^T(X\beta^* + \epsilon)}{n} - \left(\frac{Z^T Z}{n} - \Sigma_w\right)\beta^*\right)\right| \\
&\leq \left|\frac{\widetilde{u}^T Z^T\epsilon}{n}\right| + \left|\widetilde{u}^T\left(\Sigma_w - \frac{Z^T W}{n}\right)\beta^*\right| \\
&\leq (\sigma_\epsilon\sigma_z + \sigma_w\sigma_z)t,
\end{aligned}$$

with probability at least $1 - 2\exp(-cnt^2)$, using standard tail bounds for sub-Gaussian matrices. Taking $\varphi = (\sigma_\epsilon\sigma_z + \sigma_w\sigma_z)$ and $t = \sqrt{\frac{k\log(p/k)}{n}}$ gives

$$\mathcal{M} \leq \frac{c\sigma_z^2(\sigma_w + \sigma_\epsilon)^2}{\alpha_\ell^2}\frac{k\log(p/k)}{n},$$

w.h.p. Finally, we bound

$$\sigma_z^2(\sigma_w + \sigma_\epsilon)^2 \leq \sigma_z^2(2\sigma_w^2 + 2\sigma_\epsilon^2) = 2(1+\kappa)\sigma_x^2\sigma_w^2 + 2\sigma_\epsilon^2\sigma_z^2.$$

3.6.4 Proof of Theorem 3.4

For lower bounds, we follow a standard argument [9, 97, 98] to transform the estimation problem into a hypothesis testing problem. Namely, given any δ-packing $\{\beta_1, \ldots, \beta_M\}$ of the target set $\mathbb{B}_0(k) \cap \mathbb{B}_2(1)$, we have the inequality

$$\mathbb{P}\left(\min_{\widehat{\beta}} \max_{\beta^* \in \mathbb{B}_0(k) \cap \mathbb{B}_2(1)} \|\widehat{\beta} - \beta^*\|_2^2 \geq \frac{\delta^2}{4}\right) \geq \min_{\widetilde{\beta}} \mathbb{P}(\widetilde{\beta} \neq B), \tag{3.40}$$

where B is uniformly distributed over $\{\beta_1, \ldots, \beta_M\}$ and $\widetilde{\beta}$ is an estimator. We then lower-bound the RHS using Fano:

$$\mathbb{P}(\widetilde{\beta} \neq B) \geq 1 - \frac{I(y; B) + \log 2}{\log M}. \tag{3.41}$$

In order to upper-bound the mutual information $I(y; B)$, let $\mathbb{P}_\beta$ denote the distribution of y given $B = \beta$ (when Z is observed). For conciseness of notation, denote $\mathbb{P}_j = \mathbb{P}_{\beta_j}$. Since y is distributed as the mixture $\frac{1}{M}\sum_j \mathbb{P}_j$, we have

$$\begin{aligned} I(y; B) \;=\; \mathbb{E}_B[D(\mathbb{P}_{y|B} \| \mathbb{P}_y)] \;&= \frac{1}{M}\sum_j D\left(\mathbb{P}_j \middle\| \frac{1}{M}\sum_\ell \mathbb{P}_\ell\right) \\ &\leq \frac{1}{M^2}\sum_{j,\ell} D(\mathbb{P}_j \| \mathbb{P}_\ell), \end{aligned} \tag{3.42}$$

exploiting the convexity of the KL divergence in the last inequality. Finally, we upper-bound the pairwise KL divergences $D(\mathbb{P}_j \| \mathbb{P}_\ell)$ explicitly, and then choose an appropriate value of δ to ensure that $\mathbb{P}(\widetilde{\beta} \neq B) \geq 1/2$. The key steps therefore involve finding an appropriate δ-packing of the target set and an upper-bound on the mutual information.

The following lemma shows that there exists a $\frac{1}{2}$-packing of the target set with $\log M \geq \frac{k}{2}\log\frac{p-k}{k/2}$:

Lemma 3.2. *There exists a $\frac{1}{2}$-packing of $\mathbb{B}_0(k) \cap \mathbb{B}_2(1)$ in ℓ_2-norm with $\log M \geq \frac{k}{2}\log\frac{p-k}{k/2}$. In particular, if $\delta < \frac{1}{2}$, there exists a 2δ-packing $\{\beta_1, \ldots, \beta_M\}$ of the same set such that $\|\beta_j - \beta_k\|_2 \leq 4\delta$ for all pairs (j, k).*

The proof is based on a modification of a result due to Raskutti et al. [71]. We now derive an explicit expression for $\mathbb{P}_\beta$, which we will use to compute the KL divergences appearing in inequality (3.42). By independence, $\mathbb{P}_\beta$ is a product distribution of $y_i|z_i$, over all i. We claim that for each i,

$$y_i \mid z_i \sim N(\beta^T\Sigma_x\Sigma_z^{-1}z_i, \beta^T(\Sigma_x - \Sigma_x\Sigma_z^{-1}\Sigma_x)\beta + \sigma_\epsilon^2). \tag{3.43}$$

Indeed, (y_i, z_i) is clearly jointly Gaussian with mean 0, and by computing covariances,

$$\begin{bmatrix} y_i \\ z_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \beta^T\Sigma_x\beta + \sigma_\epsilon^2 & \beta^T\Sigma_x \\ \Sigma_x\beta & \Sigma_x + \Sigma_w \end{bmatrix}\right),$$

so equation (3.43) follows immediately by standard results on conditional Gaussians. We now derive the following lemma:

Lemma 3.3. *For any $\beta, \beta' \in \mathbb{B}_0(k)$, we have*

$$D(\mathbb{P}_\beta \| \mathbb{P}_{\beta'}) \leq \frac{cn\sigma_x^4}{\sigma_x^2\sigma_w^2 + \sigma_\epsilon^2\sigma_z^2}\|\beta - \beta'\|_2^2.$$

Proof. Assume σ_ϵ and σ_w are not both 0; otherwise, the theorem is trivially true. By equation (3.43), we can write

$$\begin{aligned} D(\mathbb{P}_\beta \| \mathbb{P}_{\beta'}) &= \mathbb{E}_{\mathbb{P}_\beta}\left[\log\frac{\mathbb{P}_\beta(y)}{\mathbb{P}_{\beta'}(y)}\right] \\ &= \mathbb{E}_{\mathbb{P}_\beta}\left[\frac{n}{2}\log\left(\frac{\sigma_{\beta'}^2}{\sigma_\beta^2}\right) - \frac{\|y - Z\Sigma_z^{-1}\Sigma_x\beta\|_2^2}{2\sigma_\beta^2}\right. \\ &\qquad \left. + \frac{\|y - Z\Sigma_z^{-1}\Sigma_x\beta'\|_2^2}{2\sigma_{\beta'}^2}\right] \\ &= \frac{n}{2}\log\left(\frac{\sigma_{\beta'}^2}{\sigma_\beta^2}\right) + \frac{n}{2}\left(\frac{\sigma_\beta^2}{\sigma_{\beta'}^2} - 1\right) \\ &\qquad + \frac{1}{2\sigma_{\beta'}^2}\|Z\Sigma_x\Sigma_z^{-1}(\beta - \beta')\|_2^2, \end{aligned} \tag{3.44}$$

where $\sigma_\beta^2 = \beta^T(\Sigma_x - \Sigma_x\Sigma_z^{-1}\Sigma_x)\beta + \sigma_\epsilon^2$, and $\sigma_{\beta'}^2$ is defined analogously.

In our setting, since $\Sigma_x = \sigma_x^2 I$, $\Sigma_w = \sigma_w^2 I$, and $\|\beta\|_2 = 1$,

$$\sigma_\beta^2 = \left(\sigma_x^2 - \frac{\sigma_x^4}{\sigma_z^2}\right)\|\beta\|_2^2 + \sigma_\epsilon^2 = \frac{\sigma_x^2\sigma_w^2}{\sigma_z^2} + \sigma_\epsilon^2 := \sigma'^2$$

gives the same value for all β. Then equation (3.44) becomes

$$D(\mathbb{P}_\beta \| \mathbb{P}_{\beta'}) = \frac{1}{2\sigma'^2}\frac{\sigma_x^4}{\sigma_z^4}\|Z(\beta - \beta')\|_2^2 \leq \frac{cn\sigma_x^4}{\sigma'^2\sigma_z^2}\|\beta - \beta'\|_2^2, \tag{3.45}$$

where the last inequality uses Lemma A.16 in Appendix A.3. Expanding σ'^2 in inequality (3.45) then yields the desired result. □

In particular, for β_j, β_ℓ in the δ-packing, Lemmas 3.2 and 3.3 together imply that

$$D(\mathbb{P}_j \| \mathbb{P}_\ell) \leq \frac{cn\delta^2\sigma_x^4}{\sigma_x^2\sigma_w^2 + \sigma_\epsilon^2\sigma_z^2},$$

so by inequality (3.42), we also have

$$I(y; B) \leq \frac{cn\delta^2\sigma_x^4}{\sigma_x^2\sigma_w^2 + \sigma_\epsilon^2\sigma_z^2}.$$

Substituting into Fano's inequality (3.41) gives

$$\mathbb{P}(\widetilde{\beta} \neq B) \geq 1 - \frac{\frac{cn\delta^2\sigma_x^4}{\sigma_x^2\sigma_w^2+\sigma_\epsilon^2\sigma_z^2} + \log 2}{\frac{k}{2}\log\frac{p-k}{k/2}}. \tag{3.46}$$

Note that for $p/k \geq 2$ and $k \geq 8$, we have $\log 2 \leq \frac{k}{8}\log\frac{p-k}{k/2}$, so inequality (3.46) implies that

$$\mathbb{P}(\widetilde{\beta} \neq B) \geq 1 - \left[\frac{\frac{cn\delta^2\sigma_x^4}{\sigma_x^2\sigma_w^2+\sigma_\epsilon^2\sigma_z^2}}{\frac{k}{2}\log\frac{p-k}{k/2}} + \frac{1}{4}\right].$$

Choosing

$$\delta^2 = \frac{c(\sigma_x^2\sigma_w^2 + \sigma_\epsilon^2\sigma_z^2)}{\sigma_x^4}\frac{k}{n}\log\frac{p-k}{k/2},$$

and using inequality (3.40), we conclude that $\mathcal{M} \geq \frac{\delta^2}{4}$ with probability at least $1/2$. Finally, note that when $p/k \geq 2$, we have $\frac{p-k}{k/2} = 2\left(\frac{p}{k} - 1\right) \geq \frac{p}{k}$, so we may replace the quotient $\frac{p-k}{k/2}$ by $\frac{p}{k}$ in the lower bound to obtain the result we seek.

3.6.5 Proof of Theorem 3.5

Again following Section 3.3, we use the estimators $\widehat{\Gamma} = \frac{\widehat{Z}^T\widehat{Z}}{n} - \alpha\,\mathrm{diag}\left(\frac{\widehat{Z}^T\widehat{Z}}{n}\right)$, $\widehat{\gamma} = \frac{\widehat{Z}^T y}{n}$, where $\widehat{Z} = \frac{Z}{1-\alpha}$ is a rescaled version of the missing data matrix. Let $\widetilde{u} \in \mathbb{B}_0(k) \cap \mathbb{B}_2(1)$. Using the fact that $y = X\beta^* + \epsilon$ and expanding, we have the bound

$$\begin{aligned}|\widetilde{u}^T(\widehat{\gamma} - \widehat{\Gamma}\beta^*)| \leq &\left|\widetilde{u}^T\left(\frac{\widehat{Z}^T(\widehat{Z} - X)}{n} - \frac{\alpha}{1-\alpha}\sigma_x^2 I\right)\beta^*\right| \\ &+ \left|\frac{\widetilde{u}^T\widehat{Z}^T\epsilon}{n}\right| + \alpha\left|\widetilde{u}^T\left(\mathrm{diag}\left(\frac{\widehat{Z}^T\widehat{Z}}{n}\right) - \frac{\sigma_x^2}{1-\alpha}I\right)\beta^*\right|.\end{aligned}$$

Note that $\widehat{Z}$ is sub-Gaussian with parameter $\frac{\sigma_x^2}{(1-\alpha)^2}$, so we can bound the second term by $\frac{\sigma_x\sigma_\epsilon}{1-\alpha}t$ with probability at least $1-2\exp(-cnt^2)$. Similarly, we may bound the third term by $\frac{\alpha\sigma_x^2}{(1-\alpha)^2}t$. Finally, note that

$$\frac{\widehat{Z}^T(\widehat{Z}-X)}{n} = \frac{1}{n}\sum_{i=1}^n \hat{z}_i(\hat{z}_i - x_i)^T = \frac{1}{n}\frac{\alpha}{(1-\alpha)^2}\sum_{i=1}^n z_i z_i^T,$$

where z_i is the observed vector with 0's in missing positions. Conditioned on the missing positions, $\widetilde{u}^T\frac{\widehat{Z}^T(\widehat{Z}-X)}{n}\beta^*$ is sub-exponential with parameter $\frac{\alpha\sigma_x^2}{(1-\alpha)^2}$. Since a mixture of sub-exponentials is sub-exponential with the same parameter, we have a bound of the form $\frac{\alpha\sigma_x^2}{(1-\alpha)^2}t$. Then $\varphi = \frac{\sigma_x\sigma_\epsilon}{1-\alpha} + \frac{\alpha\sigma_x^2}{(1-\alpha)^2}$ with $t = (1-\alpha)\sqrt{\frac{k\log(p/k)}{n}}$ yields the bound.

3.6.6 Proof of Theorem 3.6

Note that when $\sigma_\epsilon = 0$, the theorem is trivially true; hence, we assume $\sigma_\epsilon > 0$. We use the same δ-packing obtained in Lemma 3.2. To compute the KL divergences, we first derive the distribution of $y \mid Z$ for a fixed β, which is a product distribution of $y_i \mid z_i$ over all i. Furthermore, we may write

$$y_i = \langle x_{i,obs},\, \beta_{obs}\rangle + \langle x_{i,mis},\, \beta_{mis}\rangle + \epsilon_i, \tag{3.47}$$

where *obs* denotes the indices of the the observed coordinates and *mis* denotes the indices of the missing coordinates. Note that β_{obs} and β_{mis} vary with i. From equation (3.47), we have

$$y_i \mid z_i \sim N(z_i^T\beta, \beta_{mis}^T\Sigma_{x,mis}\beta_{mis} + \sigma_\epsilon^2).$$

Denote $\sigma_{i,\beta}^2 = \beta_{mis}^T\Sigma_{x,mis}\beta_{mis} = \sigma_\beta^2 = \sigma_x^2\|\beta_{mis}\|_2^2 + \sigma_\epsilon^2$. By a similar computation as before, for $\beta' \neq \beta$, we have

$$\frac{1}{n}D(\mathbb{P}_\beta\|\mathbb{P}_{\beta'}) = \frac{1}{n}\sum_{i=1}^n\left[\frac{1}{2}\log\left(\frac{\sigma_{i,\beta'}^2}{\sigma_{i,\beta}^2}\right) + \frac{1}{2}\left(\frac{\sigma_{i,\beta}^2}{\sigma_{i,\beta'}^2} - 1\right) + \frac{1}{2\sigma_{i,\beta'}^2}(z_i^T(\beta-\beta'))^2\right]. \tag{3.48}$$

By a Taylor expansion, $\log x \le (x-1) + c(x-1)^2$ for x close to 1. Taking $x = \frac{\sigma_{i,\beta}^2}{\sigma_{i,\beta'}^2}$, equation (3.48) becomes

$$\frac{1}{n}D(\mathbb{P}_\beta\|\mathbb{P}_{\beta'}) \le \frac{1}{n}\sum_{i=1}^n\left[\frac{c}{2}\left(\frac{\sigma_{i,\beta}^2}{\sigma_{i,\beta'}^2} - 1\right)^2 + \frac{(z_i^T(\beta-\beta'))^2}{2\sigma_{i,\beta'}^2}\right].$$

Further note that

$$\frac{\sigma_{i,\beta}^2}{\sigma_{i,\beta'}^2} - 1 = \frac{\sigma_x^2(\|\beta_{mis}\|_2^2 - \|\beta'_{mis}\|_2^2)}{\sigma_{i,\beta'}^2} = \frac{\sigma_x^2 u_i^T(\beta^2 - \beta'^2)}{\sigma_{i,\beta'}^2},$$

where $u_i \in \mathbb{R}^p$ is a binary vector with 1's corresponding to missing values in z_i, and β^2 and β'^2 are obtained by componentwise squaring of β and β'. The matrix U with rows u_i^T is i.i.d. Bernoulli, hence sub-Gaussian with parameter 1. Applying Lemma A.16 to U and Z with covariances $\alpha(1-\alpha)I$ and $(1-\alpha)\sigma_x^2 I$, taking $t = (1-\alpha)\sigma^2\sqrt{\frac{k\log(p/k)}{n}}$, we obtain

$$\frac{1}{n}D(\mathbb{P}_\beta\|\mathbb{P}_{\beta'}) \leq \frac{c\sigma_x^4}{\sigma_\epsilon^4}(1-\alpha)\|\beta^2 - \beta'^2\|_2^2 + \frac{c'\sigma_x^2}{\sigma_\epsilon^2}(1-\alpha)\|\beta - \beta'\|_2^2,$$

since $\sigma_{i,\beta'}^2 \geq \sigma_\epsilon^2$. Finally, note that

$$\|\beta^2 - \beta'^2\|_2^2 \leq \|\beta - \beta'\|_2^2\|\beta + \beta'\|_2^2 \leq 2\|\beta - \beta'\|_2^2,$$

by Cauchy-Schwarz and the triangle inequality. When $\{\beta_1, \ldots, \beta_M\}$ is a δ-packing of the set $\mathbb{B}_0(k) \cap \mathbb{B}_2(1)$, we have

$$\begin{aligned} D(\mathbb{P}_j\|\mathbb{P}_\ell) &\leq cn\delta^2(1-\alpha)\left(\frac{\sigma_x^4}{\sigma_\epsilon^4} + \frac{\sigma_x^2}{\sigma_\epsilon^2}\right) \\ &= cn\delta^2(1-\alpha)\frac{\sigma_x^2}{\sigma_\epsilon^2}\frac{\sigma_x^2 + \sigma_\epsilon^2}{\sigma_\epsilon^2} \end{aligned}$$

for all $j \neq \ell$, with probability at least $1 - \exp(-ck\log(p/k))$. Choosing

$$\delta^2 = \frac{c\sigma_\epsilon^2}{\sigma_x^2(1-\alpha)}\frac{\sigma_\epsilon^2}{\sigma_x^2 + \sigma_\epsilon^2}\frac{k}{n}\log\frac{p-k}{k/2}$$

yields the bound.

3.7 Discussion

In this chapter, we formulated an ℓ_1-constrained minimization problem for sparse linear regression on corrupted data. The source of corruption may be additive noise or missing data, and although the resulting objective is not generally convex, we showed that projected gradient descent is guaranteed to converge to a point within statistical precision of the optimum. In addition, we established ℓ_1- and ℓ_2-error bounds that hold with high probability when the data are drawn i.i.d. from a sub-Gaussian distribution, or drawn from a Gaussian vector autoregressive process. In the case when covariates are Gaussian with diagonal covariance matrix, we derived matching lower bounds on rates of estimation, showing that our procedures are minimax optimal. Finally, we used our results on linear regression to

perform sparse inverse covariance estimation for a Gaussian graphical model, where the data are observed subject to corruption. The bounds we obtain for the spectral norm of the error are of the same order as existing bounds for inverse covariance matrix estimation when the data are uncorrupted and i.i.d.

Future directions of research include studying more general types of dependencies or corruption in the covariates of regression, such as more general types of multiplicative noise; and performing sparse linear regression for corrupted data with additive noise when the noise covariance is unknown and replicates of the data may be unavailable. It would also be interesting to study the performance of our algorithms on data that are not sub-Gaussian, or even under model mismatch. In addition, one might consider other loss functions, where it is more difficult to correct the objective for corrupted covariates. Finally, it remains to be seen whether or not our techniques—used here to show that certain nonconvex problems can solved to statistical precision—can be applied more broadly.

Chapter 4

Nonconvex M-estimators

4.1 Introduction

Although recent years have brought about a flurry of work on optimization of convex functions, optimizing nonconvex functions is in general computationally intractable [66, 88]. Nonconvex functions may possess local optima that are not global optima, and iterative methods such as gradient or coordinate descent may terminate undesirably in local optima. Unfortunately, standard statistical results for nonconvex M-estimators often only provide guarantees for *global* optima. This leads to a significant gap in the theory, since computing global optima—or even near-global optima—in an efficient manner may be extremely difficult in practice. Nonetheless, empirical studies have shown that local optima of various nonconvex M-estimators arising in statistical problems appear to be well-behaved [10]. This is the starting point of our work.

A key insight is that nonconvex functions occurring in statistics are not constructed adversarially, so that "good behavior" might be expected in practice. The results of Chapter 3 confirmed this intuition for one specific case: a modified version of the Lasso applicable to errors-in-variables regression. Although the Hessian of the modified objective has many negative eigenvalues in the high-dimensional setting, the objective function resembles a strongly convex function when restricted to a cone set that includes the stationary points of the objective. This allows us to establish bounds on the statistical and optimization error.

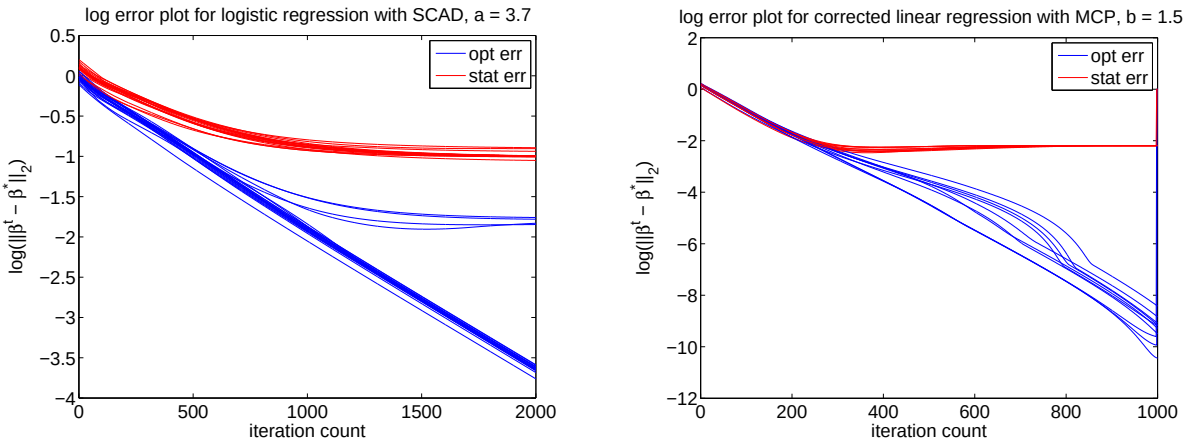
Our current chapter is framed in a more general setting, and we focus on various M-estimators coupled with (nonconvex) regularizers of interest. On the statistical side, we establish bounds on the distance between *any local optimum* of the empirical objective and the unique minimizer of the population risk. Although the nonconvex functions may possess multiple local optima (as demonstrated in simulations), our theoretical results show that all local optima are essentially as good as a global optima from a statistical perspective. The results presented here subsume the results of Chapter 3, and our present proof techniques are much more direct.

Our theory also sheds new light on a recent line of work involving the nonconvex SCAD

and MCP regularizers [28, 10, 101, 103]. Various methods previously proposed for nonconvex optimization include local quadratic approximation (LQA) [28], minorization-maximization (MM) [38], local linear approximation (LLA) [106], and coordinate descent [10, 58]. However, these methods may terminate in local optima, which were not previously known to be well-behaved. In a recent paper, Zhang and Zhang [103] provided statistical guarantees for global optima of least-squares linear regression with nonconvex penalties and showed that gradient descent starting from a Lasso solution would terminate in specific local minima. Fan et al. [29] also showed that if the LLA algorithm is initialized at a Lasso optimum satisfying certain properties, the two-stage procedure produces an oracle solution for various nonconvex penalties. Finally, Chen and Gu [19] showed that specific local optima of nonconvex regularized least-squares problems are stable, so optimization algorithms initialized sufficiently closeby will converge to the same optima. See the survey paper [103] for a more complete overview of related work.

In contrast, our results are the first to establish appropriate regularity conditions under which *all stationary points* (including both local and global optima) lie within a small ball of the population-level minimum. Thus, standard first-order methods such as projected and composite gradient descent [65] will converge to stationary points that lie within statistical error of the truth, eliminating the need for specially designed optimization algorithms that converge to specific local optima. Figure 4.1 provides an illustration of the type of behavior explained by the theory in this chapter. Panel (a) shows the behavior of composite gradient descent for a form of logistic regression with the nonconvex SCAD [28] as a regularizer: the red curve shows the *statistical error*, namely the ℓ_2-norm of the difference between the iterates and the underlying true regression vector. The blue curve shows the *optimization error*, meaning the difference between the iterates and a given stationary point of the objective. As shown by the blue curves, this problem possesses multiple local optima, since the algorithm converges to different final points depending on the initialization. However, as shown by the red curves, the statistical error of each local optimum is very low, so that they are all essentially comparable from a statistical point of view. Panel (b) exhibits the same behavior for a problem in which both the cost function (a corrected form of least-squares suitable for missing data, as described in Chapter 3) and the regularizer (the MCP function [101]) are nonconvex. Nonetheless, as guaranteed by our theory, we still see the same qualitative behavior of the statistical and optimization error. Moreover, our theory also predicts the geometric convergence rates that are apparent in these plots. More precisely, under the same sufficient conditions for statistical consistency, we show that a modified form of composite gradient descent only requires $\log(1/\epsilon_{\text{stat}})$ steps to achieve a solution that is accurate up to the statistical precision ϵ_{stat}, which is the rate expected for *strongly convex* functions. Furthermore, our techniques are more generally applicable than the methods proposed by previous authors, and are not restricted to least-squares or even convex loss functions.

While our paper was under review after its arXiv posting [56], we became aware of an independent line of related work by Wang et al. [93]. Our contributions are substantially different, in that we provide sufficient conditions guaranteeing statistical consistency for *all* local optima, whereas their work is only concerned with establishing good behavior of suc-



(a) (b)

Figure 4.1: Plots of the optimization error (blue curves) and statistical error (red curves) for a modified form of composite gradient descent, applicable to problems that may involve nonconvex cost functions and regularizers. (a) Plots for logistic regression with the nonconvex SCAD regularizer. (b) Plots for a corrected form of least squares (a nonconvex quadratic program) with the nonconvex MCP regularizer.

cessive iterates along a certain path-following algorithm. In addition, our techniques are applicable even to regularizers that do not satisfy smoothness constraints on the entire positive axis (such as capped-ℓ_1). Finally, we provide rigorous proofs showing the applicability of our sufficient condition on the loss function to a broad class of generalized linear models, whereas the applicability of their "sparse eigenvalue" condition to such objectives was not established.

The remainder of the chapter is organized as follows: In Section 2, we establish basic notation and provide background on nonconvex regularizers and loss functions of interest. In Section 3, we provide our main theoretical results, including bounds on ℓ_1-, ℓ_2-, and prediction error, and also state corollaries for special cases. Section 4 contains a modification of composite gradient descent that may be used to obtain near-global optima, and includes theoretical results establishing the linear convergence of our optimization algorithm. Section 5 supplies the results of various simulations. Proofs are contained in Appendix B.

Notation: For functions $f(n)$ and $g(n)$, we write $f(n) \precsim g(n)$ to mean that $f(n) \leq cg(n)$ for some universal constant $c \in (0, \infty)$, and similarly, $f(n) \succsim g(n)$ when $f(n) \geq c'g(n)$ for some universal constant $c' \in (0, \infty)$. We write $f(n) \asymp g(n)$ when $f(n) \precsim g(n)$ and $f(n) \succsim g(n)$ hold simultaneously. For a vector $v \in \mathbb{R}^p$ and a subset $S \subseteq \{1, \ldots, p\}$, we write $v_S \in \mathbb{R}^S$ to denote the vector v restricted to S. For a matrix M, we write $|||M|||_2$ and $|||M|||_F$ to denote the spectral and Frobenius norms, respectively, and write $|||M|||_{\max} := \max_{i,j} |m_{ij}|$

to denote the elementwise ℓ_∞-norm of M. For a function $h : \mathbb{R}^p \to \mathbb{R}$, we write ∇h to denote a gradient or subgradient, if it exists. Finally, for $q, r > 0$, let $\mathbb{B}_q(r)$ denote the ℓ_q-ball of radius r centered around 0.

4.2 Problem formulation

In this section, we develop some general theory for regularized M-estimators. We begin by establishing our notation and basic assumptions, before turning to the class of nonconvex regularizers and nonconvex loss functions to be covered in this chapter.

4.2.1 Background

Given a collection of n samples $Z_1^n = \{Z_1, \ldots, Z_n\}$, drawn from a marginal distribution $\mathbb{P}$ over a space $\mathcal{Z}$, consider a loss function $\mathcal{L}_n : \mathbb{R}^p \times (\mathcal{Z})^n \to \mathbb{R}$. The value $\mathcal{L}_n(\beta; Z_1^n)$ serves as a measure of the "fit" between a parameter vector $\beta \in \mathbb{R}^p$ and the observed data. This empirical loss function should be viewed as a surrogate to the *population risk function* $\mathcal{L} : \mathbb{R}^p \to \mathbb{R}$, given by

$$\mathcal{L}(\beta) := \mathbb{E}_Z\big[\mathcal{L}_n(\beta; Z_1^n)\big].$$

Our goal is to estimate the parameter vector $\beta^* := \arg\min_{\beta \in \mathbb{R}^p} \mathcal{L}(\beta)$ that minimizes the population risk, assumed to be unique.

To this end, we consider a regularized M-estimator of the form

$$\widehat{\beta} \in \arg \min_{g(\beta) \leq R,\ \beta \in \Omega} \left\{ \mathcal{L}_n(\beta; Z_1^n) + \rho_\lambda(\beta) \right\}, \tag{4.1}$$

where $\rho_\lambda : \mathbb{R}^p \to \mathbb{R}$ is a *regularizer*, depending on a tuning parameter $\lambda > 0$, which serves to enforce a certain type of structure on the solution. In all cases, we consider regularizers that are separable across coordinates, and with a slight abuse of notation, we write

$$\rho_\lambda(\beta) = \sum_{j=1}^{p} \rho_\lambda(\beta_j).$$

Our theory allows for possible nonconvexity in *both* the loss function $\mathcal{L}_n$ and the regularizer ρ_λ. Due to this potential nonconvexity, our M-estimator also includes a side constraint $g : \mathbb{R}^p \to \mathbb{R}_+$, which we require to be a convex function satisfying the lower bound $g(\beta) \geq \|\beta\|_1$, for all $\beta \in \mathbb{R}^p$. Consequently, any feasible point for the optimization problem (4.1) satisfies the constraint $\|\beta\|_1 \leq R$, and as long as the empirical loss and regularizer are continuous, the Weierstrass extreme value theorem guarantees that a global minimum $\widehat{\beta}$ exists. Finally, we allow for an additional side constraint $\beta \in \Omega$, where Ω is some convex set containing β^*. For the graphical Lasso considered in Section 4.3.4, we take $\Omega = \mathcal{S}_+$ to be the set of positive semidefinite matrices; in settings where such an additional condition is extraneous, we simply set $\Omega = \mathbb{R}^p$.

4.2.2 Nonconvex regularizers

We now state and discuss conditions on the regularizer, defined in terms of a univariate function $\rho_\lambda : \mathbb{R} \to \mathbb{R}$.

Assumption 4.1.

(i) *The function ρ_λ satisfies $\rho_\lambda(0) = 0$ and is symmetric around zero (i.e., $\rho_\lambda(t) = \rho_\lambda(-t)$ for all $t \in \mathbb{R}$).*

(ii) *On the nonnegative real line, the function ρ_λ is nondecreasing.*

(iii) *For $t > 0$, the function $t \mapsto \frac{\rho_\lambda(t)}{t}$ is nonincreasing in t.*

(iv) *The function ρ_λ is differentiable for all $t \neq 0$ and subdifferentiable at $t = 0$, with nonzero subgradients at $t = 0$ bounded by λL.*

(v) *There exists $\mu > 0$ such that $\rho_{\lambda,\mu}(t) := \rho_\lambda(t) + \mu t^2$ is convex.*

It is instructive to compare the conditions of Assumption 4.1 to similar conditions previously proposed in literature. Conditions (i)–(iii) are the same as those proposed in Zhang and Zhang [103], except we omit the extraneous condition of subadditivity (cf. Lemma 1 of Chen and Gu [19]). Such conditions are relatively mild and are satisfied for a wide variety of regularizers. Condition (iv) restricts the class of penalties by excluding regularizers such as the bridge (ℓ_q-) penalty, which has infinite derivative at 0; and the capped-ℓ_1 penalty, which has points of non-differentiability on the positive real line. However, one may check that if ρ_λ has unbounded derivative at zero, then $\widetilde{\beta} = 0$ is *always* a local optimum of the composite objective (4.1), so there is no hope for $\|\widetilde{\beta} - \beta^*\|_2$ to be vanishingly small. Condition (v), known as *weak convexity* [90], also appears in Chen and Gu [19] and is a type of curvature constraint that controls the level of nonconvexity of ρ_λ. Although this condition is satisfied by many regularizers of interest, it is again not satisfied by capped-ℓ_1 for any $\mu > 0$. For details on how our arguments may be modified to handle the more tricky capped-ℓ_1 penalty, see Appendix B.6.

Nonetheless, many regularizers that are commonly used in practice fully satisfy Assumption 4.1. It is easy to see that the standard ℓ_1-norm $\rho_\lambda(\beta) = \|\beta\|_1$ satisfies these conditions. More exotic functions have been studied in a line of past work on nonconvex regularization, and we provide a few examples here:

SCAD penalty: This penalty, due to Fan and Li [28], takes the form

$$\rho_\lambda(t) := \begin{cases} \lambda|t|, & \text{for } |t| \leq \lambda, \\ -(t^2 - 2a\lambda|t| + \lambda^2)/(2(a-1)), & \text{for } \lambda < |t| \leq a\lambda, \\ (a+1)\lambda^2/2, & \text{for } |t| > a\lambda, \end{cases} \tag{4.2}$$

where $a > 2$ is a fixed parameter. As verified in Lemma B.3 of Appendix B.1.2, the SCAD penalty satisfies the conditions of Assumption 4.1 with $L = 1$ and $\mu = \frac{1}{a-1}$.

MCP regularizer: This penalty, due to Zhang [104], takes the form

$$\rho_\lambda(t) := \operatorname{sign}(t)\,\lambda \cdot \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz, \tag{4.3}$$

where $b > 0$ is a fixed parameter. As verified in Lemma B.4 in Appendix B.1.2, the MCP regularizer satisfies the conditions of Assumption 4.1 with $L = 1$ and $\mu = \frac{1}{b}$.

4.2.3 Nonconvex loss functions and restricted strong convexity

Throughout this chapter, we require the loss function $\mathcal{L}_n$ to be differentiable, but we do not require it to be convex. Instead, we impose a weaker condition known as restricted strong convexity (RSC). Such conditions have been discussed in previous literature [63, 1], and involve a lower bound on the remainder in the first-order Taylor expansion of $\mathcal{L}_n$. In particular, our main statistical result is based on the following RSC condition:

$$\langle \nabla\mathcal{L}_n(\beta^* + \Delta) - \nabla\mathcal{L}_n(\beta^*),\, \Delta\rangle \geq \begin{cases} \alpha_1\|\Delta\|_2^2 - \tau_1\dfrac{\log p}{n}\|\Delta\|_1^2, & \forall \|\Delta\|_2 \leq 1, \quad \text{(4.4a)} \\ \alpha_2\|\Delta\|_2 - \tau_2\sqrt{\dfrac{\log p}{n}}\|\Delta\|_1, & \forall \|\Delta\|_2 \geq 1, \quad \text{(4.4b)} \end{cases}$$

where the α_j's are strictly positive constants and the τ_j's are nonnegative constants.

To understand this condition, note that if $\mathcal{L}_n$ were actually strongly convex, then both these RSC inequalities would hold with $\alpha_1 = \alpha_2 > 0$ and $\tau_1 = \tau_2 = 0$. However, in the high-dimensional setting ($p \gg n$), the empirical loss $\mathcal{L}_n$ can never be strongly convex, but the RSC condition may still hold with strictly positive (α_j, τ_j). On the other hand, if $\mathcal{L}_n$ is convex (but not strongly convex), the left-hand expression in inequality (4.4) is always nonnegative, so inequalities (4.4a) and (4.4b) hold trivially for $\frac{\|\Delta\|_1}{\|\Delta\|_2} \geq \sqrt{\frac{\alpha_1 n}{\tau_1 \log p}}$ and $\frac{\|\Delta\|_1}{\|\Delta\|_2} \geq \frac{\alpha_2}{\tau_2}\sqrt{\frac{n}{\log p}}$, respectively. Hence, the RSC inequalities only enforce a type of strong convexity condition over a cone set of the form $\left\{\frac{\|\Delta\|_1}{\|\Delta\|_2} \leq c\sqrt{\frac{n}{\log p}}\right\}$.

It is important to note that the class of functions satisfying RSC conditions of this type is much larger than the class of convex functions; the results of Chapter 3 contain a large family of nonconvex quadratic functions that satisfy this condition (see Section 4.3.2 below for further discussion). Finally, note that we have stated two separate RSC inequalities (4.4), unlike in past work [63, 1, 55], which only imposes the first condition (4.4a) over the entire range of Δ. As illustrated in the corollaries of Sections 4.3.3 and 4.3.4 below, the first inequality (4.4a) can only hold locally over Δ for more complicated types of functions; in contrast, as proven in Appendix B.2.1, inequality (4.4b) is implied by inequality (4.4a) in cases when $\mathcal{L}_n$ is convex.

4.3 Statistical guarantees and consequences

With this setup, we now turn to the statements and proofs of our main statistical guarantees, as well as some consequences for various statistical models. Our theory applies to any vector $\widetilde{\beta} \in \mathbb{R}^p$ that satisfies the *first-order necessary conditions* to be a local minimum of the program (4.1):

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) + \nabla \rho_\lambda(\widetilde{\beta}),\, \beta - \widetilde{\beta} \rangle \geq 0, \qquad \text{for all feasible } \beta \in \mathbb{R}^p. \tag{4.5}$$

When $\widetilde{\beta}$ lies in the interior of the constraint set, this condition reduces to the usual zero-subgradient condition:

$$\nabla \mathcal{L}_n(\widetilde{\beta}) + \nabla \rho_\lambda(\widetilde{\beta}) = 0.$$

4.3.1 Main statistical results

Our main theorems are deterministic in nature, and specify conditions on the regularizer, loss function, and parameters, which guarantee that any local optimum $\widetilde{\beta}$ lies close to the target vector $\beta^* = \arg\min_{\beta \in \mathbb{R}^p} \mathcal{L}(\beta)$. Corresponding probabilistic results will be derived in subsequent sections, where we establish that for appropriate choices of parameters (λ, R), the required conditions hold with high probability. Applying the theorems to particular models requires bounding the random quantity $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty$ and verifying the RSC conditions (4.4). We begin with a theorem that provides guarantees on the error $\widetilde{\beta} - \beta^*$ as measured in the ℓ_2- and ℓ_1-norms:

Theorem 4.1. *Suppose the regularizer ρ_λ satisfies Assumption 4.1, the empirical loss $\mathcal{L}_n$ satisfies the RSC conditions (4.4) with $\alpha_1 > \mu$, and β^* is feasible for the objective. Consider any choice of λ such that*

$$\frac{2}{L} \cdot \max\left\{ \|\nabla \mathcal{L}_n(\beta^*)\|_\infty,\; \alpha_2 \sqrt{\frac{\log p}{n}} \right\} \;\leq\; \lambda \;\leq\; \frac{\alpha_2}{6RL}, \tag{4.6}$$

and suppose $n \geq \frac{16R^2 \max(\tau_1^2, \tau_2^2)}{\alpha_2^2} \log p$. Then any vector $\widetilde{\beta}$ satisfying the first-order necessary conditions (4.5) satisfies the error bounds

$$\|\widetilde{\beta} - \beta^*\|_2 \leq \frac{7\lambda L \sqrt{k}}{4(\alpha_1 - \mu)}, \qquad \text{and} \qquad \|\widetilde{\beta} - \beta^*\|_1 \leq \frac{56 \lambda L k}{4(\alpha_1 - \mu)}, \tag{4.7}$$

where $k = \|\beta^*\|_0$.

From the bound (4.7), note that the squared ℓ_2-error grows proportionally with k, the number of non-zeros in the target parameter, and with λ^2. As will be clarified in the following sections, choosing λ proportional to $\sqrt{\frac{\log p}{n}}$ and R proportional to $\frac{1}{\lambda}$ will satisfy

the requirements of Theorem 4.1 w.h.p. for many statistical models, in which case we have a squared-ℓ_2 error that scales as $\frac{k \log p}{n}$, as expected.

Our next theorem provides a bound on a measure of the prediction error, as defined by the quantity

$$D(\widetilde{\beta}; \beta^*) := \langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\beta} - \beta^* \rangle. \tag{4.8}$$

When the empirical loss $\mathcal{L}_n$ is a convex function, this measure is always nonnegative, and in various special cases, it has a form that is readily interpretable. For instance, in the case of the least-squares objective function $\mathcal{L}_n(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2$, we have

$$D(\widetilde{\beta}; \beta^*) = \frac{1}{n}\|X(\widetilde{\beta} - \beta^*)\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} \big(\langle x_i, \widetilde{\beta} - \beta^* \rangle\big)^2,$$

corresponding to the usual measure of (fixed design) prediction error for a linear regression problem (cf. Corollary 4.1 below). More generally, when the loss function is the negative log likelihood for a generalized linear model with cumulant function ψ, the error measure (4.8) is equivalent to the symmetrized Bregman divergence defined by ψ. (See Section 4.3.3 for further details.)

Theorem 4.2. *Under the same conditions as Theorem 4.1, the error measure (4.8) is bounded as*

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\beta} - \beta^* \rangle \leq \lambda^2 L^2 k \left(\frac{21}{8(\alpha_1 - \mu)} + \frac{49\mu}{16(\alpha_1 - \mu)^2} \right). \tag{4.9}$$

This result shows that the prediction error (4.8) behaves similarly to the squared Euclidean norm between $\widetilde{\beta}$ and β^*.

Remark 4.1. *It is worthwhile to discuss the quantity $\alpha_1 - \mu$ appearing in the denominator of the bounds in Theorems 4.1 and 4.2. Recall that α_1 measures the level of curvature of the loss function $\mathcal{L}_n$, while μ measures the level of nonconvexity of the penalty ρ_λ. Intuitively, the two quantities should play opposing roles in our result: Larger values of μ correspond to more severe nonconvexity of the penalty, resulting in worse behavior of the overall objective (4.1), whereas larger values of α_1 correspond to more (restricted) curvature of the loss, leading to better behavior. However, while the condition $\alpha_1 > \mu$ is needed for the proof technique employed in Theorem 4.1, it does not seem to be strictly necessary in order to guarantee good behavior of local optima. Indeed, note that the capped-ℓ_1 penalty may be viewed as a limiting version of SCAD when $a \to 1$, or equivalently, $\mu \to \infty$. Viewed in this light, Theorem B.1, to be stated and proved in Appendix B.6, reveals that the condition $\alpha_1 > \mu$ is not necessary, at least in general, for good behavior of local optima. Moreover, Section 4.5 contains empirical studies using linear regression and the SCAD penalty showing that local optima may be well-behaved even when $\alpha_1 < \mu$. Nonetheless, our simulations (see Figure 4.5) also convey a*

cautionary message: In extreme cases, where α_1 is much smaller than μ, the good behavior of local optima (and the optimization algorithms used to find them) appear to degenerate.

Finally, we note that Negahban et al. [63] have shown that for convex M-estimators, the arguments used to analyze ℓ_1-regularizers may be generalized to other types of "decomposable" regularizers, such as norms for group sparsity or the nuclear norm for low-rank matrices. In our present setting, where we allow for nonconvexity in the loss and regularizer, our theorems have straightforward and analogous generalizations.

We return to the proofs of Theorems 4.1 and 4.2 in Section 4.3.5. First, we develop various consequences of these theorems for various nonconvex loss functions and regularizers of interest. The main technical challenge is to establish that the RSC conditions (4.4) hold with high probability for appropriate choices of positive constants $\{(\alpha_j, \tau_j)\}_{j=1}^2$.

4.3.2 Corrected linear regression

We begin by considering the case of high-dimensional linear regression with systematically corrupted observations. Recall that in the framework of ordinary linear regression, we have the linear model

$$y_i = \underbrace{\langle \beta^*, x_i \rangle}_{\sum_{j=1}^p \beta_j^* x_{ij}} + \epsilon_i, \qquad \text{for } i = 1, \ldots, n, \tag{4.10}$$

where $\beta^* \in \mathbb{R}^p$ is the unknown parameter vector and $\{(x_i, y_i)\}_{i=1}^n$ are observations. Following the framework discussed in Chapter 3, assume we instead observe pairs $\{(z_i, y_i)\}_{i=1}^n$, where the z_i's are systematically corrupted versions of the corresponding x_i's. Some examples of corruption mechanisms include the following:

(a) *Additive noise:* We observe $z_i = x_i + w_i$, where $w_i \in \mathbb{R}^p$ is a random vector independent of x_i, say zero-mean with known covariance matrix Σ_w.

(b) *Missing data:* For some fraction $\vartheta \in [0, 1)$, we observe a random vector $z_i \in \mathbb{R}^p$ such that for each component j, we independently observe $z_{ij} = x_{ij}$ with probability $1 - \vartheta$, and $z_{ij} = *$ with probability ϑ.

We use the population and empirical loss functions

$$\mathcal{L}(\beta) = \frac{1}{2}\beta^T \Sigma_x \beta - \beta^{*T} \Sigma_x \beta, \qquad \text{and} \qquad \mathcal{L}_n(\beta) = \frac{1}{2}\beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta, \tag{4.11}$$

where $(\widehat{\Gamma}, \widehat{\gamma})$ are estimators for $(\Sigma_x, \Sigma_x \beta^*)$ depending only on $\{(z_i, y_i)\}_{i=1}^n$. It is easy to see that $\beta^* = \arg\min_\beta \mathcal{L}(\beta)$. From the formulation (4.1), the corrected linear regression estimator is given by

$$\widehat{\beta} \in \arg \min_{g(\beta) \leq R} \left\{ \frac{1}{2}\beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta + \rho_\lambda(\beta) \right\}. \tag{4.12}$$

We now state a concrete corollary in the case of additive noise (model (a) above). In this case, as discussed in Chapter 3, an appropriate choice of the pair $(\widehat{\Gamma}, \widehat{\gamma})$ is given by

$$\widehat{\Gamma} = \frac{Z^T Z}{n} - \Sigma_w, \qquad \text{and} \qquad \widehat{\gamma} = \frac{Z^T y}{n}. \tag{4.13}$$

Here, we assume the noise covariance Σ_w is known or may be estimated from replicates of the data. Such an assumption also appears in canonical errors-in-variables literature [18], but it is an open question how to devise a corrected estimator when an estimate of Σ_w is not readily available.

In the high-dimensional setting ($p \gg n$), the matrix $\widehat{\Gamma}$ in equation (4.13) is always negative-definite: the matrix $\frac{Z^T Z}{n}$ has rank at most n, and then the positive definite matrix Σ_w is subtracted to obtain $\widehat{\Gamma}$. Consequently, the empirical loss function $\mathcal{L}_n$ previously defined (4.11) is nonconvex. Other choices of $\widehat{\Gamma}$ are applicable to missing data (model (b)), and also lead to nonconvex programs (see Chapter 3 for further details).

Corollary 4.1. *Suppose we have i.i.d. observations $\{(z_i, y_i)\}_{i=1}^n$ from a corrupted linear model with additive noise, where the x_i's are sub-Gaussian. Suppose (λ, R) are chosen such that β^* is feasible and*

$$c\sqrt{\frac{\log p}{n}} \leq \lambda \leq \frac{c'}{R}.$$

Then given a sample size $n \geq C \max\{R^2, k\} \log p$, any local optimum $\widetilde{\beta}$ of the nonconvex program (4.12) satisfies the estimation error bounds

$$\|\widetilde{\beta} - \beta^*\|_2 \leq \frac{c_0 \lambda \sqrt{k}}{\lambda_{\min}(\Sigma_x) - 2\mu}, \qquad \text{and} \qquad \|\widetilde{\beta} - \beta^*\|_1 \leq \frac{c_0' \lambda k}{\lambda_{\min}(\Sigma_x) - 2\mu},$$

and the prediction error bound

$$\widetilde{\nu}^T \widehat{\Gamma} \widetilde{\nu} \leq \lambda^2 k \left(\frac{\widetilde{c}_0}{\lambda_{\min}(\Sigma_x) - 2\mu} + \frac{\widetilde{c}_0' \mu}{(\lambda_{\min}(\Sigma_x) - 2\mu)^2} \right),$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$, where $\|\beta^*\|_0 = k$.

Remark 4.2. *When $\rho_\lambda(\beta) = \lambda \|\beta\|_1$ and $g(\beta) = \|\beta\|_1$, then taking $\lambda \asymp \sqrt{\frac{\log p}{n}}$ and $R = b_0 \sqrt{k}$ for some constant $b_0 \geq \|\beta^*\|_2$ yields the required scaling $n \gtrsim k \log p$. Hence, the bounds of Corollary 4.1 agree with bounds previously established in Theorem 3.1 in Chapter 3. Note, however, that those results are stated only for a* global minimum $\widehat{\beta}$ *of the program (4.12), whereas Corollary 4.1 is a much stronger result holding for* any local minimum $\widetilde{\beta}$*. Theorem 3.2 in Chapter 3 provides a rather indirect (algorithmic) route for establishing similar bounds $\|\widetilde{\beta} - \beta^*\|_1$ and $\|\widetilde{\beta} - \beta^*\|_2$, since the proposed projected gradient descent algorithm may become stuck in a local minimum. In contrast, our argument here is much more direct and does not rely on an algorithmic proof. Furthermore, our result is applicable to a more general class of (possibly nonconvex) penalties beyond the usual ℓ_1-norm.*

Corollary 4.1 also has important consequences in the case where pairs $\{(x_i, y_i)\}_{i=1}^n$ from the linear model (4.10) are observed cleanly without corruption and ρ_λ is a nonconvex penalty. In that case, the empirical loss $\mathcal{L}_n$ previously defined (4.11) is equivalent to the least-squares loss, modulo a constant factor. Much existing work, including that of Fan and Li [28] and Zhang and Zhang [103], first establishes statistical consistency results concerning *global* minima of the program (4.12), then provides specialized algorithms such as a local linear approximation (LLA) for obtaining specific local optima that are provably close to global optima. However, our results show that *any* optimization algorithm guaranteed to converge to a local optimum of the program suffices. See Section 4.4 for a more detailed discussion of optimization procedures and fast convergence guarantees for obtaining local minima. In the fully-observed case, we also have $\widehat{\Gamma} = \frac{X^T X}{n}$, so the prediction error bound in Corollary 4.1 agrees with the familiar scaling $\frac{1}{n}\|X(\widetilde{\beta} - \beta^*)\|_2^2 \precsim \frac{k \log p}{n}$ appearing in ℓ_1-theory.

Furthermore, our theory provides a theoretical motivation for why the usual choice of $a = 3.7$ for linear regression with the SCAD penalty [28] is reasonable. Indeed, as discussed in Section 4.2.2, we have

$$\mu = \frac{1}{a-1} \approx 0.37$$

in that case. Since $x_i \sim N(0, I)$ in the SCAD simulations, we have $\lambda_{\min}(\Sigma_x) > 2\mu$ for the choice $a = 3.7$. For further comments regarding the parameter a in the SCAD penalty, see the discussion concerning Figure 4.3 in Section 4.5.

4.3.3 Generalized linear models

Moving beyond linear regression, we now consider the case where observations are drawn from a generalized linear model (GLM). Recall that a GLM is characterized by the conditional distribution

$$\mathbb{P}(y_i \mid x_i, \beta, \sigma) = \exp\left\{\frac{y_i \langle \beta, x_i \rangle - \psi(x_i^T \beta)}{c(\sigma)}\right\},$$

where $\sigma > 0$ is a scale parameter and ψ is the cumulant function, By standard properties of exponential families [59, 50], we have

$$\psi'(x_i^T \beta) = \mathbb{E}[y_i \mid x_i, \beta, \sigma].$$

In our analysis, we assume that there exists $\alpha_u > 0$ such that $\psi''(t) \leq \alpha_u$ for all $t \in \mathbb{R}$. Note that this boundedness assumption holds in various settings, including linear regression, logistic regression, and multinomial regression, but does not hold for Poisson regression. The bound will be necessary to establish both statistical consistency results in the present section and fast global convergence guarantees for our optimization algorithms in Section 4.4.

The population loss corresponding to the negative log likelihood is then given by

$$\mathcal{L}(\beta) = -\mathbb{E}[\log \mathbb{P}(x_i, y_i)] = -\mathbb{E}[\log \mathbb{P}(x_i)] - \frac{1}{c(\sigma)} \cdot \mathbb{E}[y_i \langle \beta, x_i \rangle - \psi(x_i^T \beta)],$$

giving rise to the population-level and empirical gradients

$$\nabla \mathcal{L}(\beta) = \frac{1}{c(\sigma)} \cdot \mathbb{E}[(\psi'(x_i^T\beta) - y_i)x_i], \quad \text{and}$$
$$\nabla \mathcal{L}_n(\beta) = \frac{1}{c(\sigma)} \cdot \frac{1}{n}\sum_{i=1}^n \big(\psi'(x_i^T\beta) - y_i\big)x_i.$$

Since we are optimizing over β, we will rescale the loss functions and assume $c(\sigma) = 1$. We may check that if β^* is the true parameter of the GLM, then $\nabla \mathcal{L}(\beta^*) = 0$; furthermore,

$$\nabla^2 \mathcal{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^n \psi''(x_i^T\beta)x_i x_i^T \succeq 0,$$

so $\mathcal{L}_n$ is convex.

We will assume that β^* is sparse and optimize the penalized maximum likelihood program

$$\widehat{\beta} \in \arg\min_{g(\beta)\le R} \left\{ \frac{1}{n}\sum_{i=1}^n \big(\psi(x_i^T\beta) - y_i x_i^T\beta\big) + \rho_\lambda(\beta) \right\}. \tag{4.14}$$

We then have the following corollary, proved in Appendix B.2.3:

Corollary 4.2. *Suppose we have i.i.d. observations $\{(x_i, y_i)\}_{i=1}^n$ from a GLM, where the x_i's are sub-Gaussian. Suppose (λ, R) are chosen such that β^* is feasible and*

$$c\sqrt{\frac{\log p}{n}} \le \lambda \le \frac{c'}{R}.$$

Then given a sample size $n \ge CR^2 \log p$, any local optimum $\widetilde{\beta}$ of the nonconvex program (4.14) satisfies

$$\|\widetilde{\beta} - \beta^*\|_2 \le \frac{c_0\lambda\sqrt{k}}{\lambda_{\min}(\Sigma_x) - 2\mu}, \qquad \text{and} \qquad \|\widetilde{\beta} - \beta^*\|_1 \le \frac{c_0'\lambda k}{\lambda_{\min}(\Sigma_x) - 2\mu},$$

with probability at least $1 - c_1\exp(-c_2 \log p)$, where $\|\beta^*\|_0 = k$.

Remark 4.3. *Although $\mathcal{L}_n$ is convex in this case, the overall program may* not *be convex if the regularizer ρ_λ is nonconvex, giving rise to multiple local optima. For instance, see the simulations of Figure 4.4 in Section 4.5 for a demonstration of such local optima. In past work, Breheny and Huang [10] studied logistic regression with SCAD and MCP regularizers, but did not provide any theoretical results on the quality of the local optima. In this context, Corollary 4.2 shows that their coordinate descent algorithms are guaranteed to converge to a local optimum $\widetilde{\beta}$ within close proximity of the true parameter β^*.*

4.3.4 Graphical Lasso

Finally, we specialize our results to the case of the graphical Lasso. Given p-dimensional observations $\{x_i\}_{i=1}^n$, the goal is to estimate the structure of the underlying (sparse) graphical model. Recall that the population and empirical losses for the graphical Lasso are given by

$$\mathcal{L}(\Theta) = \operatorname{trace}(\Sigma\Theta) - \log\det(\Theta), \quad \text{and} \quad \mathcal{L}_n(\Theta) = \operatorname{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta),$$

where $\widehat{\Sigma}$ is an empirical estimate for the covariance matrix $\Sigma = \operatorname{Cov}(x_i)$. The objective function for the graphical Lasso is then given by

$$\widehat{\Theta} \in \arg\min_{g(\Theta)\leq R,\ \Theta\succeq 0} \left\{ \operatorname{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \sum_{j,k=1}^{p} \rho_\lambda(\Theta_{jk}) \right\}, \tag{4.15}$$

where we apply the (possibly nonconvex) penalty function ρ_λ to all entries of Θ, and define $\Omega := \left\{\Theta \in \mathbb{R}^{p\times p} \mid \Theta = \Theta^T,\ \Theta \succeq 0\right\}$.

A host of statistical and algorithmic results have been established for the graphical Lasso in the case of Gaussian observations with an ℓ_1-penalty [4, 30, 78, 100], and more recently for discrete-valued observations, as described in Chapter 5. In addition, a version of the graphical Lasso incorporating a nonconvex SCAD penalty has been proposed [27]. Our results subsume previous Frobenius error bounds for the graphical Lasso, and again imply that even in the presence of a nonconvex regularizer, all local optima of the nonconvex program (4.15) remain close to the true inverse covariance matrix Θ^*.

As suggested in Chapter 5, the graphical Lasso easily accommodates systematically corrupted observations, with the only modification being the form of the sample covariance matrix $\widehat{\Sigma}$. Furthermore, the program (4.15) is always useful for obtaining a consistent estimate of a sparse inverse covariance matrix, regardless of whether the x_i's are drawn from a distribution for which Θ^* is relevant in estimating the edges of the underlying graph. Note that other variants of the graphical Lasso exist in which only off-diagonal entries of Θ are penalized, and similar results for statistical consistency hold in that case. Here, we assume all entries are penalized equally in order to simplify our arguments. The same framework is considered by Fan et al. [27].

We have the following result, proved in Appendix B.2.4. The statement of the corollary is purely deterministic, but in cases of interest (say, sub-Gaussian observations), the deviation condition (4.16) holds with probability at least $1 - c_1\exp(-c_2\log p)$, translating into the Frobenius norm bound (4.17) holding with the same probability.

Corollary 4.3. *Suppose we have an estimate $\widehat{\Sigma}$ of the covariance matrix Σ based on (possibly corrupted) observations $\{x_i\}_{i=1}^n$, such that*

$$\left|\left|\left|\widehat{\Sigma} - \Sigma\right|\right|\right|_{\max} \leq c_0\sqrt{\frac{\log p}{n}}. \tag{4.16}$$

Also suppose Θ^* has at most s nonzero entries. Suppose (λ, R) are chosen such that Θ^* is feasible and

$$c\sqrt{\frac{\log p}{n}} \leq \lambda \leq \frac{c'}{R}.$$

Then with a sample size $n > Cs \log p$, for a sufficiently large constant $C > 0$, any local optimum $\widetilde{\Theta}$ of the nonconvex program (4.15) satisfies

$$\left\|\widetilde{\Theta} - \Theta^*\right\|_F \leq \frac{c_0' \lambda \sqrt{s}}{\left(\|\Theta^*\|_2 + 1\right)^{-2} - \mu}. \tag{4.17}$$

Remark 4.4. *When ρ is simply the ℓ_1-penalty, the bound (4.17) from Corollary 4.3 matches the minimax rates for Frobenius norm estimation of an s-sparse inverse covariance matrix [78, 74].*

4.3.5 Proof of Theorems 4.1 and 4.2

Proof of Theorem 4.1 Introducing the shorthand $\widetilde{\nu} := \widetilde{\beta} - \beta^*$, we begin by proving that $\|\widetilde{\nu}\|_2 \leq 1$. If not, then inequality (4.4b) gives the lower bound

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\nu} \rangle \geq \alpha_2 \|\widetilde{\nu}\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\widetilde{\nu}\|_1. \tag{4.18}$$

Since β^* is feasible, we may take $\beta = \beta^*$ in inequality (4.5), and combining with inequality (4.18) yields

$$\langle -\nabla \rho_\lambda(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\nu} \rangle \geq \alpha_2 \|\widetilde{\nu}\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\widetilde{\nu}\|_1. \tag{4.19}$$

By Hölder's inequality, followed by the triangle inequality, we also have

$$\begin{aligned} \langle -\nabla \rho_\lambda(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\nu} \rangle &\leq \left\{ \|\nabla \rho_\lambda(\widetilde{\beta})\|_\infty + \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \right\} \|\widetilde{\nu}\|_1 \\ &\overset{(i)}{\leq} \left\{ \lambda L + \frac{\lambda L}{2} \right\} \|\widetilde{\nu}\|_1, \end{aligned}$$

where inequality (i) follows since $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \leq \frac{\lambda L}{2}$ by the bound (4.6), and the bound $\|\nabla \rho_\lambda(\widetilde{\beta})\|_\infty \leq \lambda L$ holds by Lemma B.1 in Appendix B.1.1. Combining this upper bound with inequality (4.19) and rearranging then yields

$$\|\widetilde{\nu}\|_2 \leq \frac{\|\widetilde{\nu}\|_1}{\alpha_2} \left(\frac{3\lambda L}{2} + \tau_2 \sqrt{\frac{\log p}{n}} \right) \leq \frac{2R}{\alpha_2} \left(\frac{3\lambda L}{2} + \tau_2 \sqrt{\frac{\log p}{n}} \right).$$

By our choice of λ from inequality (4.6) and the assumed lower bound on the sample size n, the right hand side is at most 1, so $\|\widetilde{\nu}\|_2 \leq 1$, as claimed.

Consequently, we may apply inequality (4.4a), yielding the lower bound

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\nu} \rangle \geq \alpha_1 \|\widetilde{\nu}\|_2^2 - \tau_1 \frac{\log p}{n} \|\widetilde{\nu}\|_1^2. \tag{4.20}$$

Since the function $\rho_{\lambda,\mu}(\beta) := \rho_\lambda(\beta) + \mu\|\beta\|_2^2$ is convex by assumption, we have

$$\rho_{\lambda,\mu}(\beta^*) - \rho_{\lambda,\mu}(\widetilde{\beta}) \geq \langle \nabla \rho_{\lambda,\mu}(\widetilde{\beta}), \beta^* - \widetilde{\beta} \rangle \;=\; \langle \nabla \rho_\lambda(\widetilde{\beta}) + 2\mu\widetilde{\beta}, \beta^* - \widetilde{\beta} \rangle,$$

implying that

$$\langle \nabla \rho_\lambda(\widetilde{\beta}), \beta^* - \widetilde{\beta} \rangle \leq \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \mu\|\widetilde{\beta} - \beta^*\|_2^2. \tag{4.21}$$

Combining inequality (4.20) with inequalities (4.5) and (4.21), we obtain

$$\begin{aligned}
\alpha_1 \|\widetilde{\nu}\|_2^2 - \tau_1 \frac{\log p}{n} \|\widetilde{\nu}\|_1^2 &\leq -\langle \nabla \mathcal{L}_n(\beta^*), \widetilde{\nu} \rangle + \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \mu\|\widetilde{\beta} - \beta^*\|_2^2 \\
&\overset{(i)}{\leq} \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \cdot \|\widetilde{\nu}\|_1 + \lambda L \left(\|\widetilde{\nu}_A\|_1 - \|\widetilde{\nu}_{A^c}\|_1 \right) + \mu\|\widetilde{\nu}\|_2^2 \\
&\overset{(ii)}{\leq} \frac{3\lambda L}{2} \|\widetilde{\nu}_A\|_1 - \frac{\lambda L}{2} \|\widetilde{\nu}_{A^c}\|_1 + \mu\|\widetilde{\nu}\|_2^2,
\end{aligned} \tag{4.22}$$

where inequality (i) is obtained by applying Hölder's inequality to the first term and applying Lemma B.2 in Appendix B.1.1 to the middle two terms, and inequality (ii) uses the bound

$$\|\widetilde{\nu}\|_1 \leq \|\widetilde{\nu}_A\|_1 + \|\widetilde{\nu}_{A^c}\|_1.$$

Here, A is defined to be the index set of the k largest elements of $\widetilde{\beta} - \beta^*$ in magnitude, and A^c is the complement. Rearranging inequality (4.22), we find that

$$\begin{aligned}
0 \leq 2(\alpha_1 - \mu)\|\widetilde{\nu}\|_2^2 &\leq 3\lambda L \|\widetilde{\nu}_A\|_1 - \lambda L \|\widetilde{\nu}_{A^c}\|_1 + 4R\tau_1 \frac{\log p}{n} \|\widetilde{\nu}\|_1 \\
&\leq 3\lambda L \|\widetilde{\nu}_A\|_1 - \lambda L \|\widetilde{\nu}_{A^c}\|_1 + \alpha_2 \sqrt{\frac{\log p}{n}} \|\widetilde{\nu}\|_1 \\
&\leq \frac{7\lambda L}{2} \|\widetilde{\nu}_A\|_1 - \frac{\lambda L}{2} \|\widetilde{\nu}_{A^c}\|_1,
\end{aligned} \tag{4.23}$$

implying that $\|\widetilde{\nu}_{A^c}\|_1 \leq 7\|\widetilde{\nu}_A\|_1$. Consequently,

$$\|\widetilde{\nu}\|_1 = \|\widetilde{\nu}_A\|_1 + \|\widetilde{\nu}_{A^c}\|_1 \leq 8\|\widetilde{\nu}_A\|_1 \leq 8\sqrt{k}\|\widetilde{\nu}_A\|_2 \leq 8\sqrt{k}\|\widetilde{\nu}\|_2. \tag{4.24}$$

Furthermore, inequality (4.23) implies that

$$2(\alpha_1 - \mu)\|\widetilde{\nu}\|_2^2 \leq \frac{7\lambda L}{2} \|\widetilde{\nu}_A\|_1 \leq \frac{7\lambda L\sqrt{k}}{2} \|\widetilde{\nu}\|_2.$$

Rearranging yields the ℓ_2-bound, whereas the ℓ_1-bound follows from by combining the ℓ_2-bound with the cone inequality (4.24).

Proof of Theorem 4.2 To establish inequality (4.9), note that combining the first-order condition (4.5) with the upper bounds of inequalities (4.21) and (4.22), we have

$$\begin{aligned}\langle \nabla \mathcal{L}_n(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\nu} \rangle &\leq \langle -\nabla \rho_\lambda(\widetilde{\beta}) - \nabla \mathcal{L}_n(\beta^*), \widetilde{\nu} \rangle \\ &\leq \rho_\lambda(\beta^*) - \rho_\lambda(\widetilde{\beta}) + \mu \|\widetilde{\nu}\|_2^2 + \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \cdot \|\widetilde{\nu}\|_1 \\ &\leq \frac{3\lambda L}{2} \|\widetilde{\nu}_A\|_1 - \frac{\lambda L}{2} \|\widetilde{\nu}_{A^c}\|_1 + \mu \|\widetilde{\nu}\|_2^2 \\ &\leq \frac{3\lambda L}{2} \sqrt{k} \|\widetilde{\nu}\|_2 + \mu \|\widetilde{\nu}\|_2^2,\end{aligned}$$

so substituting in the ℓ_2-bound (4.7) yields the desired result.

4.4 Optimization algorithms

We now describe how a version of composite gradient descent may be applied to efficiently optimize the nonconvex program (4.1), and show that it enjoys a linear rate of convergence under suitable conditions. In this section, we focus exclusively on a version of the optimization problem with the side function

$$g_{\lambda,\mu}(\beta) := \frac{1}{\lambda} \Big\{ \rho_\lambda(\beta) + \mu \|\beta\|_2^2 \Big\}, \tag{4.25}$$

which is convex by Assumption 4.1. We may then write the program (4.1) as

$$\widehat{\beta} \in \arg \min_{g_{\lambda,\mu}(\beta) \leq R,\, \beta \in \Omega} \Big\{ \underbrace{\big(\mathcal{L}_n(\beta) - \mu \|\beta\|_2^2 \big)}_{\bar{\mathcal{L}}_n} + \lambda g_{\lambda,\mu}(\beta) \Big\}. \tag{4.26}$$

In this way, the objective function decomposes nicely into a sum of a differentiable but nonconvex function and a possibly nonsmooth but convex penalty. Applied to the representation (4.26) of the objective function, the composite gradient descent procedure of Nesterov [65] produces a sequence of iterates $\{\beta^t\}_{t=0}^\infty$ via the updates

$$\beta^{t+1} \in \arg \min_{g_{\lambda,\mu}(\beta) \leq R} \left\{ \frac{1}{2} \left\| \beta - \left(\beta^t - \frac{\nabla \bar{\mathcal{L}}_n(\beta^t)}{\eta} \right) \right\|_2^2 + \frac{\lambda}{\eta} g_{\lambda,\mu}(\beta) \right\}, \tag{4.27}$$

where $\frac{1}{\eta}$ is the stepsize. As discussed in Section 4.4.2, these updates may be computed in a relatively straightforward manner.

4.4.1 Fast global convergence

The main result of this section is to establish that the algorithm defined by the iterates (4.27) converges very quickly to a δ-neighborhood of any global optimum, for all tolerances δ that are of the same order (or larger) than the statistical error.

We begin by setting up the notation and assumptions underlying our result. The *Taylor error* around the vector β_2 in the direction $\beta_1 - \beta_2$ is given by

$$\mathcal{T}(\beta_1, \beta_2) := \mathcal{L}_n(\beta_1) - \mathcal{L}_n(\beta_2) - \langle \nabla \mathcal{L}_n(\beta_2), \beta_1 - \beta_2 \rangle. \tag{4.28}$$

We analogously define the Taylor error $\overline{\mathcal{T}}$ for the modified loss function $\overline{\mathcal{L}}_n$, and note that

$$\overline{\mathcal{T}}(\beta_1, \beta_2) = \mathcal{T}(\beta_1, \beta_2) - \mu \|\beta_1 - \beta_2\|_2^2. \tag{4.29}$$

For all vectors $\beta_2 \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R)$, we require the following form of restricted strong convexity:

$$\mathcal{T}(\beta_1, \beta_2) \geq \begin{cases} \alpha_1 \|\beta_1 - \beta_2\|_2^2 - \tau_1 \dfrac{\log p}{n} \|\beta_1 - \beta_2\|_1^2, & \forall \|\beta_1 - \beta_2\|_2 \leq 3, \quad (4.30a) \\ \alpha_2 \|\beta_1 - \beta_2\|_2 - \tau_2 \sqrt{\dfrac{\log p}{n}} \|\beta_1 - \beta_2\|_1, & \forall \|\beta_1 - \beta_2\|_2 \geq 3. \quad (4.30b) \end{cases}$$

The conditions (4.30) are similar but not identical to the earlier RSC conditions (4.4). The main difference is that we now require the Taylor difference to be bounded below uniformly over $\beta_2 \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R)$, as opposed to for a fixed $\beta_2 = \beta^*$. In addition, we assume an analogous upper bound on the Taylor series error:

$$\mathcal{T}(\beta_1, \beta_2) \leq \alpha_3 \|\beta_1 - \beta_2\|_2^2 + \tau_3 \frac{\log p}{n} \|\beta_1 - \beta_2\|_1^2, \qquad \text{for all } \beta_1, \beta_2 \in \Omega, \tag{4.31}$$

a condition referred to as *restricted smoothness* in past work [1]. Throughout this section, we assume $\alpha_i > \mu$ for all i, where μ is the coefficient ensuring the convexity of the function $g_{\lambda,\mu}$ from equation (4.25). Furthermore, we define $\alpha = \min\{\alpha_1, \alpha_2\}$ and $\tau = \max\{\tau_1, \tau_2, \tau_3\}$.

The following theorem applies to any population loss function $\mathcal{L}$ for which the population minimizer β^* is k-sparse and $\|\beta^*\|_2 \leq 1$, and under the scaling $n > Ck \log p$, for a constant C depending on the α_i's and τ_i's. Note that this scaling is reasonable, since no estimator of a k-sparse vector in p dimensions can have low ℓ_2-error unless the condition holds (see Raskutti et al. [71] for minimax rates). We show that the composite gradient updates (4.27) exhibit a type of *globally geometric convergence* in terms of the quantity

$$\kappa := \frac{1 - \frac{\alpha - \mu}{4\eta} + \varphi(n, p, k)}{1 - \varphi(n, p, k)}, \qquad \text{where} \quad \varphi(n, p, k) := \frac{128 \tau k \frac{\log p}{n}}{\alpha - \mu}. \tag{4.32}$$

Under the stated scaling on the sample size, we are guaranteed that $\kappa \in (0, 1)$, so it is a *contraction factor*. Roughly speaking, we show that the squared optimization error will fall below δ^2 within $T \asymp \frac{\log(1/\delta^2)}{\log(1/\kappa)}$ iterations. More precisely, our theorem guarantees δ-accuracy for all iterations larger than

$$T^*(\delta) := \frac{2 \log\left(\frac{\phi(\beta^0) - \phi(\widehat{\beta})}{\delta^2}\right)}{\log(1/\kappa)} + \left(1 + \frac{\log 2}{\log(1/\kappa)}\right) \log \log \left(\frac{\lambda R L}{\delta^2}\right), \tag{4.33}$$

where $\phi(\beta) := \mathcal{L}_n(\beta) + \rho_\lambda(\beta)$ denotes the composite objective function. As clarified in the theorem statement, the squared tolerance δ^2 is not allowed to be arbitrarily small, which would contradict the fact that the composite gradient method may converge to a local optimum. However, our theory allows δ^2 to be of the same order as the squared *statistical error* $\epsilon^2_{\text{stat}} = \|\widehat{\beta} - \beta^*\|_2^2$, the distance between a fixed global optimum and the target parameter β^*. From a statistical perspective, there is no point in optimizing beyond this tolerance.

With this setup, we now turn to a precise statement of our main optimization-theoretic result. As with Theorems 4.1 and 4.2, the statement of Theorem 4.3 is entirely deterministic.

Theorem 4.3. *Suppose the empirical loss $\mathcal{L}_n$ satisfies the RSC/RSM conditions (4.30) and (4.31), and suppose the regularizer ρ_λ satisfies Assumption 4.1. Suppose $\widehat{\beta}$ is any global minimum of the program (4.26), with regularization parameters chosen such that*

$$R\sqrt{\frac{\log p}{n}} \leq c, \qquad \text{and} \qquad \lambda \geq \frac{4}{L} \cdot \max\left\{ \|\nabla \mathcal{L}_n(\beta^*)\|_\infty, \; \tau\sqrt{\frac{\log p}{n}} \right\}.$$

Then for any stepsize parameter $\eta \geq 2 \cdot \max\{\alpha_3 - \mu, \, \mu\}$ and tolerance parameter $\delta^2 \geq \frac{c\epsilon^2_{stat}}{1-\kappa}$, we have

$$\|\beta^t - \widehat{\beta}\|_2^2 \; \leq \; \frac{2}{\alpha - \mu} \left(\delta^2 + \frac{\delta^4}{\tau} + 128\tau \frac{k \log p}{n} \epsilon^2_{stat} \right), \qquad \forall t \geq T^*(\delta). \tag{4.34}$$

Remark 4.5. *Note that for the optimal choice of tolerance parameter $\delta \asymp \epsilon_{stat}$, the error bound appearing in inequality (4.34) takes the form $\frac{c\epsilon^2_{stat}}{\alpha - \mu}$, meaning that successive iterates of the composite gradient descent algorithm are guaranteed to converge to a region within statistical accuracy of the true global optimum $\widehat{\beta}$. More concretely, if the sample size satisfies $n \gtrsim Ck \log p$ and the regularization parameters are chosen appropriately, Theorem 4.1 guarantees that $\epsilon_{stat} = \mathcal{O}\left(\sqrt{\frac{k \log p}{n}}\right)$ with high probability. Combined with Theorem 4.3, we then conclude that*

$$\max\left\{ \|\beta^t - \widehat{\beta}\|_2, \; \|\beta^t - \beta^*\|_2 \right\} = \mathcal{O}\left(\sqrt{\frac{k \log p}{n}} \right),$$

for all iterations $t \geq T(\epsilon_{stat})$.

As would be expected, the (restricted) curvature α of the loss function and nonconvexity parameter μ of the penalty function enter into the bound via the denominator $\alpha - \mu$. Indeed, the bound is tighter when the loss function possesses more curvature or the penalty function is closer to being convex, agreeing with intuition. Similar to our discussion in Remark 4.1, the requirement $\alpha > \mu$ is certainly necessary for our proof technique, but it is possible that composite gradient descent still produces good results when this condition is violated. See Section 4.5 for simulations in scenarios involving mild and severe violations of this condition.

Finally, note that the parameter η must be sufficiently large (or equivalently, the stepsize must be sufficiently small) in order for the composite gradient descent algorithm to be well-behaved. See Nesterov [65] for a discussion of how the stepsize may be chosen via an iterative search when the problem parameters are unknown.

In the case of corrected linear regression (Corollary 4.1), Lemma A.13 in Appendix A.2 establishes the RSC/RSM conditions for various statistical models. The following proposition shows that the conditions (4.30) and (4.31) hold in GLMs when the x_i's are drawn i.i.d. from a zero-mean sub-Gaussian distribution with parameter σ_x and covariance matrix $\Sigma_x = \mathrm{cov}(x_i)$. As usual, we assume a sample size $n \geq c\, k \log p$, for a sufficiently large constant $c > 0$. Recall the definition of the Taylor error $\mathcal{T}(\beta_1, \beta_2)$ from equation (4.28).

Proposition 4.1. *[RSC/RSM conditions for generalized linear models] There exists a constant $\alpha_\ell > 0$, depending only on the GLM and (σ_x, Σ_x), such that for all vectors $\beta_2 \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R)$, we have*

$$\mathcal{T}(\beta_1, \beta_2) \geq \begin{cases} \frac{\alpha_\ell}{2}\|\Delta\|_2^2 - \frac{c^2\sigma_x^2}{2\alpha_\ell}\frac{\log p}{n}\|\Delta\|_1^2, & \text{for all } \|\beta_1 - \beta_2\|_2 \leq 3, \quad (4.35a) \\ \frac{3\alpha_\ell}{2}\|\Delta\|_2 - 3c\sigma_x\sqrt{\frac{\log p}{n}}\|\Delta\|_1, & \text{for all } \|\beta_1 - \beta_2\|_2 \geq 3, \quad (4.35b) \end{cases}$$

with probability at least $1 - c_1 \exp(-c_2 n)$. With the bound $\|\psi''\|_\infty \leq \alpha_u$, we also have

$$\mathcal{T}(\beta_1, \beta_2) \leq \alpha_u \lambda_{max}(\Sigma_x) \left(\frac{3}{2}\|\Delta\|_2^2 + \frac{\log p}{n}\|\Delta\|_1^2\right), \qquad \text{for all } \beta_1, \beta_2 \in \mathbb{R}^p, \quad (4.36)$$

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

For the proof of Proposition 4.1, see Appendix B.4.

4.4.2 Form of updates

In this section, we discuss how the updates (4.27) are readily computable in many cases. We begin with the case $\Omega = \mathbb{R}^p$, so we have no additional constraints apart from $g_{\lambda,\mu}(\beta) \leq R$. In this case, given iterate β^t, the next iterate β^{t+1} may be obtained via the following three-step procedure:

(1) First optimize the unconstrained program

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2}\left\| \beta - \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta}\right)\right\|_2^2 + \frac{\lambda}{\eta} \cdot g_{\lambda,\mu}(\beta) \right\}. \quad (4.37)$$

(2) If $g_{\lambda,\mu}(\widehat{\beta}) \leq R$, define $\beta^{t+1} = \widehat{\beta}$.

(3) Otherwise, if $g_{\lambda,\mu}(\widehat{\beta}) > R$, optimize the constrained program

$$\beta^{t+1} \in \arg\min_{g_{\lambda,\mu}(\beta) \leq R} \left\{ \frac{1}{2} \left\| \beta - \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right) \right\|_2^2 \right\}. \tag{4.38}$$

We derive the correctness of this procedure in Appendix B.3.1. For many nonconvex regularizers ρ_λ of interest, the unconstrained program (4.37) has a convenient closed-form solution: For the SCAD penalty (4.2), the program (4.37) has simple closed-form solution given by

$$\widehat{\beta}_{\text{SCAD}} = \begin{cases} 0 & \text{if } 0 \leq |z| \leq \nu\lambda, \\ z - \text{sign}(z) \cdot \nu\lambda & \text{if } \nu\lambda \leq |z| \leq (\nu+1)\lambda, \\ \frac{z - \text{sign}(z) \cdot \frac{a\nu\lambda}{a-1}}{1 - \frac{\nu}{a-1}} & \text{if } (\nu+1)\lambda \leq |z| \leq a\lambda, \\ z & \text{if } |z| \geq a\lambda. \end{cases} \tag{4.39}$$

For the MCP (4.3), the optimum of the program (4.37) takes the form

$$\widehat{\beta}_{\text{MCP}} = \begin{cases} 0 & \text{if } 0 \leq |z| \leq \nu\lambda, \\ \frac{z - \text{sign}(z) \cdot \nu\lambda}{1 - \nu/b} & \text{if } \nu\lambda \leq |z| \leq b\lambda, \\ z & \text{if } |z| \geq b\lambda. \end{cases} \tag{4.40}$$

In both equations (4.39) and (4.40), we have

$$z := \frac{1}{1 + 2\mu/\eta} \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right), \qquad \text{and} \qquad \nu := \frac{1/\eta}{1 + 2\mu/\eta},$$

and the operations are taken componentwise. See Appendix B.3.2 for the derivation of these closed-form updates.

More generally, when $\Omega \subsetneq \mathbb{R}^p$ (such as in the case of the graphical Lasso), the minimum in the program (4.27) must be taken over Ω, as well. Although the updates are not as simply stated, they still involve solving a convex optimization problem. Despite this more complicated form, however, our results from Section 4.4.1 on fast global convergence under restricted strong convexity and restricted smoothness assumptions carry over without modification, since they only require RSC/RSM conditions holding over a sufficiently small radius together with feasibility of β^*.

4.4.3 Proof of Theorem 4.3

We provide the outline of the proof here, with more technical results deferred to Appendix B.3. In broad terms, our proof is inspired by a result of Agarwal et al. [1], but requires various modifications in order to be applied to the much larger family of nonconvex regularizers considered here.

Our first lemma shows that the optimization error $\beta^t - \widehat{\beta}$ lies in an approximate cone set:

Lemma 4.1. *Under the conditions of Theorem 4.3, suppose there exists a pair $(\bar{\eta}, T)$ such that*

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \bar{\eta}, \qquad \forall t \geq T. \tag{4.41}$$

Then for any iteration $t \geq T$, we have

$$\|\beta^t - \widehat{\beta}\|_1 \leq 4\sqrt{k}\|\beta^t - \widehat{\beta}\|_2 + 8\sqrt{k}\|\widehat{\beta} - \beta^*\|_2 + 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right).$$

Our second lemma shows that as long as the composite gradient descent algorithm is initialized with a solution β^0 within a constant radius of a global optimum $\widehat{\beta}$, all successive iterates also lie within the same ball:

Lemma 4.2. *Under the conditions of Theorem 4.3, and with an initial vector β^0 such that $\|\beta^0 - \widehat{\beta}\|_2 \leq 3$, we have*

$$\|\beta^t - \widehat{\beta}\|_2 \leq 3, \qquad \text{for all } t \geq 0. \tag{4.42}$$

In particular, suppose we initialize the composite gradient procedure with a vector β^0 such that $\|\beta^0\|_2 \leq \frac{3}{2}$. Then by the triangle inequality,

$$\|\beta^0 - \widehat{\beta}\|_2 \leq \|\beta^0\|_2 + \|\widehat{\beta} - \beta^*\|_2 + \|\beta^*\|_2 \leq 3,$$

where we have assumed our scaling of n guarantees $\|\widehat{\beta} - \beta^*\|_2 \leq 1/2$.

Finally, recalling our earlier definition (4.32) of κ, the third lemma combines the results of Lemmas 4.1 and 4.2 to establish a bound on the value of the objective function that decays exponentially with t:

Lemma 4.3. *Under the same conditions of Lemma 4.2, suppose in addition that inequality (4.41) holds and $\frac{32k\tau \log p}{n} \leq \frac{\alpha - \mu}{2}$. Then for any $t \geq T$, we have*

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \kappa^{t-T}(\phi(\beta^T) - \phi(\widehat{\beta})) + \frac{\xi}{1-\kappa}(\epsilon^2 + \bar{\epsilon}^2),$$

where $\bar{\epsilon} := 8\sqrt{k}\epsilon_{stat}$, $\epsilon := 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right)$, the quantities κ and φ are defined according to equations (4.32), and

$$\xi := \frac{1}{1 - \varphi(n,p,k)} \cdot \frac{2\tau \log p}{n} \cdot \left(\frac{\alpha - \mu}{4\eta} + 2\varphi(n,p,k) + 5\right). \tag{4.43}$$

The remainder of the proof follows an argument used in Agarwal et al. [1], so we only provide a high-level sketch. We first prove the following inequality:

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \delta^2, \qquad \text{for all } t \geq T^*(\delta), \tag{4.44}$$

as follows. We divide the iterations $t \geq 0$ into a series of epochs $[T_\ell, T_{\ell+1})$ and define tolerances $\bar{\eta}_0 > \bar{\eta}_1 > \cdots$ such that

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \bar{\eta}_\ell, \qquad \forall t \geq T_\ell.$$

In the first iteration, we apply Lemma 4.3 with $\bar{\eta}_0 = \phi(\beta^0) - \phi(\widehat{\beta})$ to obtain

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \kappa^t \left(\phi(\beta^0) - \phi(\widehat{\beta})\right) + \frac{\xi}{1-\kappa}(4R^2 + \bar{\epsilon}^2), \qquad \forall t \geq 0.$$

Let $\bar{\eta}_1 := \frac{2\xi}{1-\kappa}(4R^2 + \bar{\epsilon}^2)$, and note that for $T_1 := \left\lceil \frac{\log(2\bar{\eta}_0/\bar{\eta}_1)}{\log(1/\kappa)} \right\rceil$, we have

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \bar{\eta}_1 \leq \frac{4\xi}{1-\kappa} \max\{4R^2, \bar{\epsilon}^2\}, \qquad \text{for all } t \geq T_1.$$

For $\ell \geq 1$, we now define

$$\bar{\eta}_{\ell+1} := \frac{2\xi}{1-\kappa}(\epsilon_\ell^2 + \bar{\epsilon}^2), \quad \text{and} \quad T_{\ell+1} := \left\lceil \frac{\log(2\bar{\eta}_\ell/\bar{\eta}_{\ell+1})}{\log(1/\kappa)} \right\rceil + T_\ell,$$

where $\epsilon_\ell := 2 \min\left\{\frac{\bar{\eta}_\ell}{\lambda L}, R\right\}$. From Lemma 4.3, we have

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \kappa^{t-T_\ell} \left(\phi(\beta^{T_\ell}) - \phi(\widehat{\beta})\right) + \frac{\xi}{1-\kappa}(\epsilon_\ell^2 + \bar{\epsilon}^2), \qquad \text{for all } t \geq T_\ell,$$

implying by our choice of $\{(\eta_\ell, T_\ell)\}_{\ell \geq 1}$ that

$$\phi(\beta^t) - \phi(\widehat{\beta}) \leq \bar{\eta}_{\ell+1} \leq \frac{4\xi}{1-\kappa} \max\{\epsilon_\ell^2, \bar{\epsilon}^2\}, \qquad \forall t \geq T_{\ell+1}.$$

Finally, we use the recursion

$$\bar{\eta}_{\ell+1} \leq \frac{4\xi}{1-\kappa} \max\{\epsilon_\ell^2, \bar{\epsilon}^2\}, \qquad T_\ell \leq \ell + \frac{\log(2^\ell \bar{\eta}_0/\bar{\eta}_\ell)}{\log(1/\kappa)}, \tag{4.45}$$

to establish the recursion

$$\bar{\eta}_{\ell+1} \leq \frac{\bar{\eta}_\ell}{4^{2^{\ell-1}}}, \qquad \frac{\bar{\eta}_{\ell+1}}{\lambda L} \leq \frac{R}{4^{2^\ell}}. \tag{4.46}$$

Inequality (4.44) then follows from computing the number of epochs and timesteps necessary to obtain $\frac{\lambda R L}{4^{2^{\ell-1}}} \leq \delta^2$. For the remaining steps used to obtain inequalities (4.46) from inequalities (4.45), we refer the reader to Agarwal et al. [1].

Finally, by inequality (B.29b) in the proof of Lemma 4.3 in Appendix B.3.5 and the relative scaling of (n, p, k), we have

$$\begin{aligned}\frac{\alpha-\mu}{2}\|\beta^t-\widehat{\beta}\|_2^2 &\leq \phi(\beta^t)-\phi(\widehat{\beta})+2\tau\frac{\log p}{n}\left(\frac{2\delta^2}{\lambda L}+\bar{\epsilon}\right)^2 \\ &\leq \delta^2+2\tau\frac{\log p}{n}\left(\frac{2\delta^2}{\lambda L}+\bar{\epsilon}\right)^2,\end{aligned}$$

where we have set $\epsilon = \frac{2\delta^2}{\lambda L}$. Rearranging and performing some algebra with our choice of λ gives the ℓ_2-bound.

4.5 Simulations

In this section, we report the results of simulations we performed to validate our theoretical results. In particular, we present results for two versions of the loss function $\mathcal{L}_n$, corresponding to linear and logistic regression, and three penalty functions, namely the ℓ_1-norm (Lasso), the SCAD penalty, and the MCP, as detailed in Section 4.2.2. In all cases, we chose regularization parameters $R = \frac{1.1}{\lambda}\cdot\rho_\lambda(\beta^*)$, to ensure feasibility of β^*, and $\lambda = \sqrt{\frac{\log p}{n}}$.

Linear regression: In the case of linear regression, we simulated covariates corrupted by additive noise according to the mechanism described in Section 4.3.2, giving the estimator

$$\widehat{\beta} \in \arg\min_{g_{\lambda,\mu}(\beta)\leq R}\left\{\frac{1}{2}\beta^T\left(\frac{X^TX}{n}-\Sigma_w\right)\beta-\frac{y^TZ}{n}\beta+\rho_\lambda(\beta)\right\}. \tag{4.47}$$

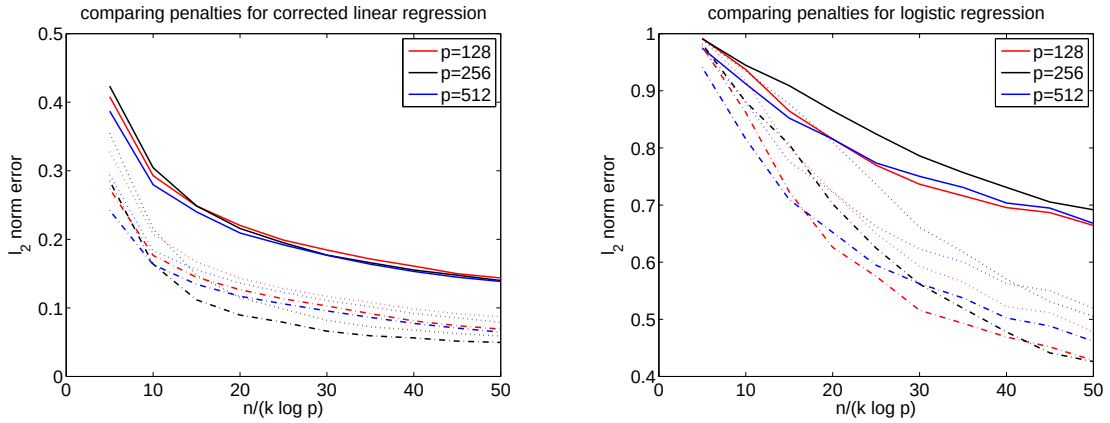
We generated i.i.d. samples $x_i \sim N(0, I)$ and set $\Sigma_w = (0.2)^2 I$, and generated additive noise $\epsilon_i \sim N(0, (0.1)^2)$.

Logistic regression: In the case of logistic regression, we also generated i.i.d. samples $x_i \sim N(0, I)$. Since $\psi(t) = \log(1+\exp(t))$, the program (4.14) becomes

$$\widehat{\beta} \in \arg\min_{g_{\lambda,\mu}(\beta)\leq R}\left\{\frac{1}{n}\sum_{i=1}^n\{\log(1+\exp(\langle\beta,\, x_i\rangle)-y_i\langle\beta,\, x_i\rangle\}+\rho_\lambda(\beta)\right\}. \tag{4.48}$$

We optimized the programs (4.47) and (4.48) using the composite gradient updates (4.27). In order to compute the updates, we used the three-step procedure described in Section 4.4.2, together with the updates for SCAD and MCP given by equations (4.39) and (4.40). Note that the updates for the Lasso penalty may be generated more simply and efficiently as discussed in Agarwal et al. [1].

Figure 4.2 shows the results of corrected linear regression with Lasso, SCAD, and MCP regularizers for three different problem sizes p. In each case, β^* is a k-sparse vector with $k = \lfloor \sqrt{p} \rfloor$, where the nonzero entries were generated from a normal distribution and the vector was then rescaled so $\|\beta^*\|_2 = 1$. As predicted by Theorem 4.1, the three curves corresponding to the same penalty function stack up nicely when the estimation error $\|\widehat{\beta} - \beta^*\|_2$ is plotted against the rescaled sample size $\frac{n}{k \log p}$, and the ℓ_2-error decreases to zero as the number of samples increases, showing that the estimators (4.47) and (4.48) are statistically consistent. The Lasso, SCAD, and MCP regularizers are depicted by solid, dotted, and dashed lines, respectively. We chose the parameter $a = 3.7$ for the SCAD penalty, suggested by Fan and Li [28] to be "optimal" based on cross-validated empirical studies, and chose $b = 3.5$ for the MCP. Each point represents an average over 20 trials.



(a) (b)

Figure 4.2: Plots showing statistical consistency of linear and logistic regression with Lasso, SCAD, and MCP regularizers, at sparsity level $k = \lfloor \sqrt{p} \rfloor$. Panel (a) shows results for corrected linear regression, where covariates are subject to additive noise with $SNR = 5$. Panel (b) shows similar results for logistic regression. Each point represents an average over 20 trials. In both cases, the estimation error $\|\widehat{\beta} - \beta^*\|_2$ is plotted against the rescaled sample size $\frac{n}{k \log p}$. Lasso, SCAD, and MCP results are represented by solid, dotted, and dashed lines, respectively. As predicted by Theorem 4.1 and Corollaries 4.1 and 4.2, the curves for each of the three types stack up for different problem sizes p, and the error decreases to zero as the number of samples increases, showing that our methods are statistically consistent.

The simulations in Figure 4.3 depict the optimization-theoretic conclusions of Theorem 4.3. Each panel shows two different families of curves, corresponding to statistical error (red) and optimization error (blue). Here, the vertical axis measures the ℓ_2-error on a logarithmic scale, while the horizontal axis tracks the iteration number. Within each block, the curves were obtained by running the composite gradient descent algorithm from 10 different initial starting points chosen at random. In all cases, we used the parameter settings $p = 128$,

$k = \lfloor \sqrt{p} \rfloor$, and $n = \lfloor 20k \log p \rfloor$. As predicted by our theory, the optimization error decreases at a linear rate (on the log scale) until it falls to the level of statistical error. Furthermore, it is interesting to compare the plots in panels (c) and (d), which provide simulation results for two different values of the SCAD parameter a. We see that the choice $a = 3.7$ leads to a tighter cluster of local optima, providing further evidence that this setting suggested by Fan and Li [28] is in some sense optimal.

Figure 4.4 provides analogous results to Figure 4.3 in the case of logistic regression, using $p = 64, k = \lfloor \sqrt{p} \rfloor$, and $n = \lfloor 20k \log p \rfloor$. The plot shows solution trajectories for 20 different initializations of composite gradient descent. Again, we see that the log optimization error decreases at a linear rate up to the level of statistical error, as predicted by Theorem 4.3. Furthermore, the Lasso penalty yields a unique local/optimum $\widehat{\beta}$, since the program (4.48) is convex, as we observe in panel (a). In contrast, the nonconvex program based on the SCAD penalty produces multiple local optima, whereas the MCP yields a relatively large number of local optima, albeit all guaranteed to lie within a small ball of β^* by Theorem 4.1.

Finally, Figure 4.5 explores the behavior of our algorithm when the condition $\alpha_1 > \mu$ from Theorem 4.1 is *not* satisfied. We generated i.i.d. samples $x_i \sim N(0, \Sigma)$, with Σ taken to be a Toeplitz matrix with entries $\Sigma_{ij} = \zeta^{|i-j|}$, for some parameter $\zeta \in [0, 1)$, so that $\lambda_{\min}(\Sigma) \geq (1 - \zeta)^2$. We chose $\zeta \in \{0.5, 0.9\}$, resulting in $\alpha_1 \approx \{0.25, 0.01\}$. Panel (a) shows the expected good behavior of ℓ_1-regularization, even for $\alpha_1 = 0.01$; although convergence is slow and the overall statistical error is greater than for $\Sigma = I$ (cf. Figure 4.3(a)), composite gradient descent still converges at a linear rate. Panel (b) shows that for SCAD parameter $a = 2.5$ (corresponding to $\mu \approx 0.67$), local optima still seem to be well-behaved even for $\alpha_1 = 0.25 < \mu$. However, for much smaller values of α_1, the good behavior breaks down, as seen in panels (c) and (d). Note that in the latter two panels, the composite gradient descent algorithm does not appear to be converging, even as the iteration number increases. Comparing (c) and (d) also illustrates the interplay between the curvature parameter α_1 of $\mathcal{L}_n$ and the nonconvexity parameter μ of ρ_λ. Indeed, the plot in panel (d) is slightly "better" than the plot in panel (c), in the sense that initial iterates at least demonstrate some pattern of convergence. This could be attributed to the fact that the SCAD parameter is larger, corresponding to a smaller value of μ.

4.6 Discussion

We have analyzed theoretical properties of local optima of regularized M-estimators, where both the loss and penalty function are allowed to be nonconvex. Our results are the first to establish that *all local optima* of such nonconvex problems are close to the truth, implying that any optimization method guaranteed to converge to a local optimum will provide statistically consistent solutions. We show concretely that a variant of composite gradient descent may be used to obtain near-global optima in linear time, and verify our theoretical results with simulations.

Future directions of research include further generalizing our statistical consistency results

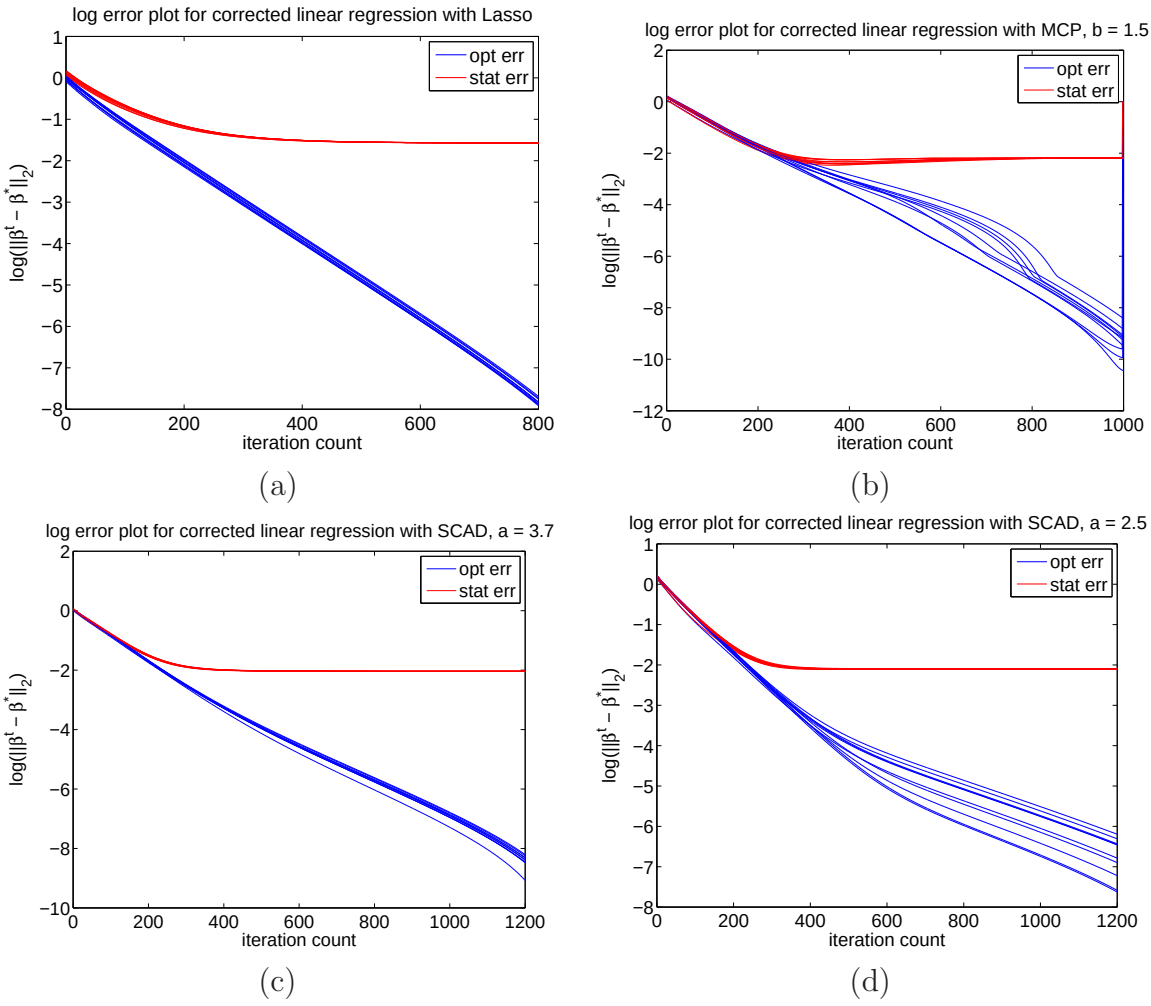


Figure 4.3: Plots illustrating linear rates of convergence on a log scale for corrected linear regression with Lasso, MCP, and SCAD regularizers, with $p = 128$, $k = \lfloor \sqrt{p} \rfloor$, and $n = \lfloor 20k \log p \rfloor$, where covariates are corrupted by additive noise with $SNR = 5$. Red lines depict statistical error $\log\left(\|\widehat{\beta} - \beta^*\|_2\right)$ and blue lines depict optimization error $\log\left(\|\beta^t - \widehat{\beta}\|_2\right)$. As predicted by Theorem 4.3, the optimization error decreases linearly when plotted against the iteration number on a log scale, up to statistical accuracy. Each plot shows the solution trajectory for 10 different initializations of the composite gradient descent algorithm. Panels (a) and (b) show the results for Lasso and MCP regularizers, respectively; panels (c) and (d) show results for the SCAD penalty with two different parameter values. Note that the empirically optimal choice $a = 3.7$ proposed by Fan and Li [28] generates local optima that exhibit a smaller spread than the local optima generated for a smaller setting of the parameter a.

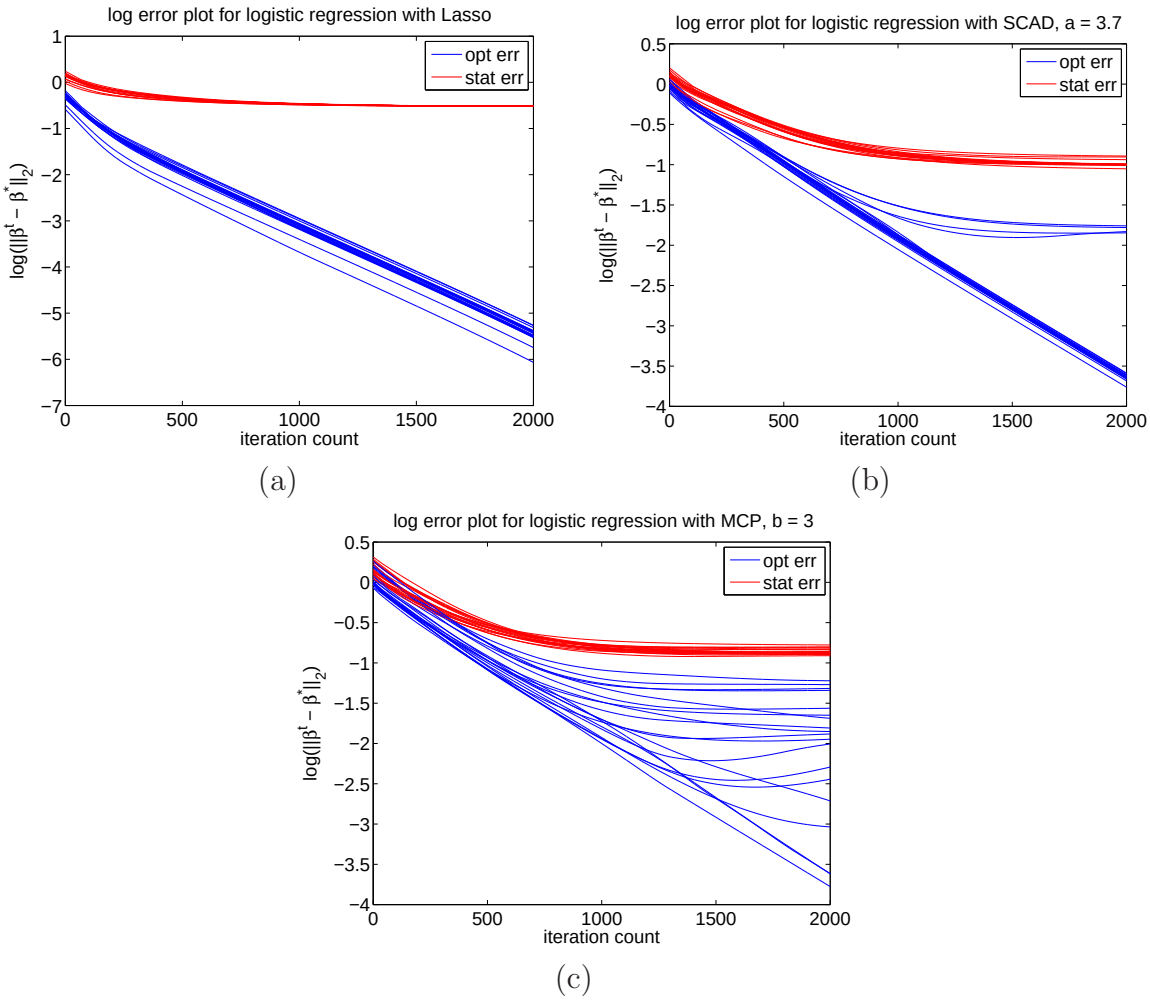


Figure 4.4: Plots that demonstrate linear rates of convergence on a log scale for logistic regression with $p = 64, k = \sqrt{p}$, and $n = \lfloor 20k \log p \rfloor$. Red lines depict statistical error $\log\left(\|\widehat{\beta} - \beta^*\|_2\right)$ and blue lines depict optimization error $\log\left(\|\beta^t - \widehat{\beta}\|_2\right)$. (a) Lasso penalty. (b) SCAD penalty. (c) MCP. As predicted by Theorem 4.3, the optimization error decreases linearly when plotted against the iteration number on a log scale, up to statistical accuracy. Each plot shows the solution trajectory for 20 different initializations of the composite gradient descent algorithm.

to other nonconvex regularizers not covered by our present theory, such as bridge penalties or regularizers that do not decompose across coordinates. In addition, it would be interesting to expand our theory to nonsmooth loss functions such as the hinge loss. For both nonsmooth losses and nonsmooth penalties (including capped-ℓ_1), it remains an open question whether a modified version of composite gradient descent may be used to obtain near-global optima in

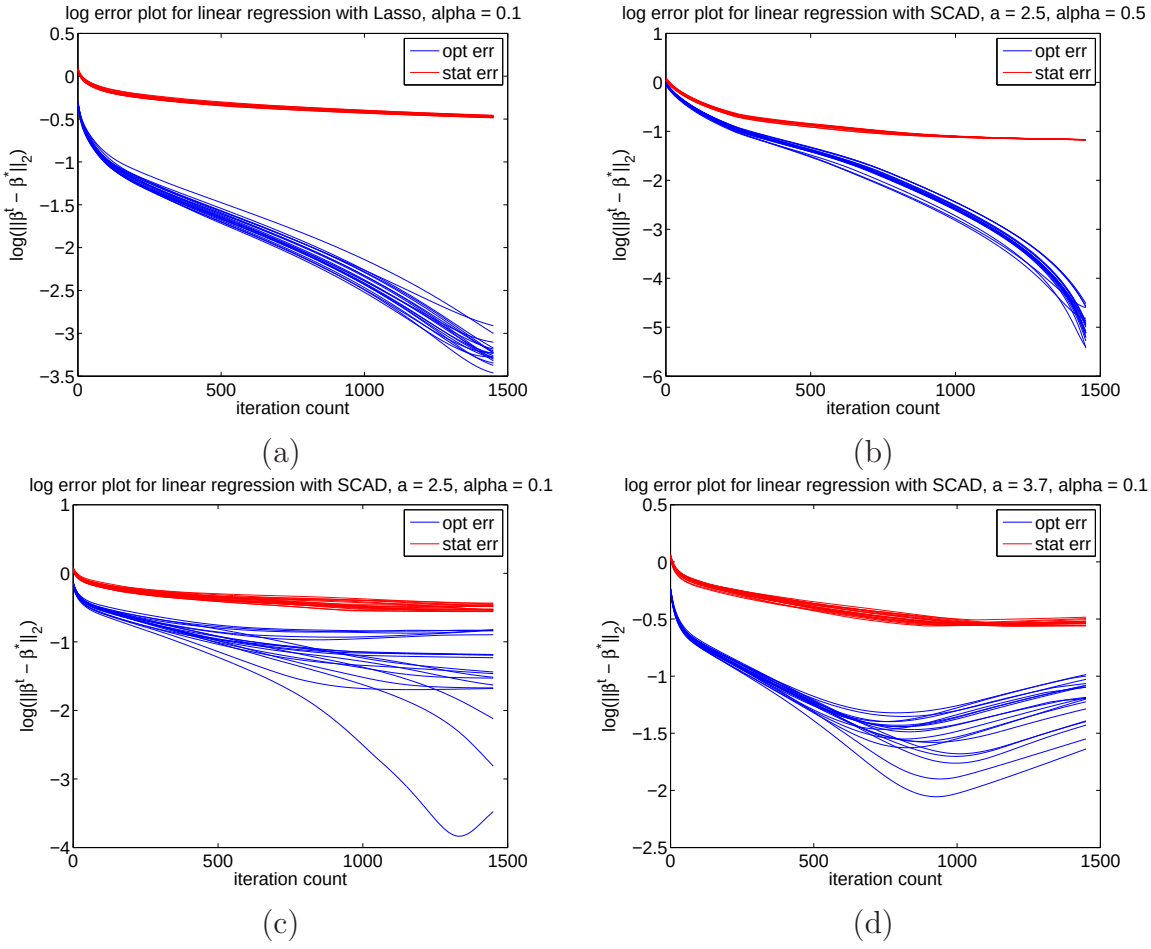


Figure 4.5: Plots showing breakdown points as a function of the curvature parameter α_1 of the loss function and the nonconvexity parameter μ of the penalty function. The loss comes from ordinary least squares linear regression, where covariates are fully-observed and sampled from a Gaussian distribution with covariance equal to a Toeplitz matrix. Panel (a) depicts the good behavior of Lasso-based linear regression. Panel (b) shows that local optima may still be well-behaved even when $\alpha_1 < \mu$, although this situation is not covered by our theory. Panels (c) and (d) show that the good behavior nonetheless disintegrates for very small values of α_1 when the regularizer is nonconvex.

polynomial time. Finally, it would be interesting to develop a general method for establishing RSC and RSM conditions, beyond the specialized methods used for studying GLMs in this chapter.

Chapter 5

Graphical model estimation

5.1 Introduction

Graphical models are used in many application domains, running the gamut from computer vision and civil engineering to political science and epidemiology. In many applications, estimating the edge structure of an underlying graphical model is of significant interest. For instance, a graphical model may be used to represent friendships between people in a social network [4] or links between organisms with the propensity to spread an infectious disease [67]. It is a classical corollary of the Hammersley-Clifford theorem [33, 7, 47] that zeros in the inverse covariance matrix of a multivariate Gaussian distribution indicate absent edges in the corresponding graphical model. This fact, combined with various types of statistical estimators suited to high dimensions, has been leveraged by many authors to recover the structure of a Gaussian graphical model when the edge set is sparse (see the papers [13, 60, 74, 99] and references therein). Recently, Liu et al. [53, 52] introduced the notion of a nonparanormal distribution, which generalizes the Gaussian distribution by allowing for monotonic univariate transformations, and argued that the same structural properties of the inverse covariance matrix carry over to the nonparanormal; see also related work of Xue and Zou [96] on copula transformations.

However, for non-Gaussian graphical models, the question of whether a general relationship exists between conditional independence and the structure of the inverse covariance matrix remains unresolved. In this chapter, we establish a number of interesting links between covariance matrices and the edge structure of an underlying graph in the case of discrete-valued random variables. (Although we specialize our treatment to multinomial random variables due to their widespread applicability, several of our results have straightforward generalizations to other types of exponential families.) Instead of only analyzing the standard covariance matrix, we show that it is often fruitful to augment the usual covariance matrix with higher-order interaction terms. Our main result has an interesting corollary for tree-structured graphs: for such models, the inverse of a generalized covariance matrix is always (block) graph-structured. In particular, for binary variables, the inverse of the

usual covariance matrix may be used to recover the edge structure of the tree. We also establish more general results that apply to arbitrary (non-tree) graphs, specified in terms of graph triangulations. This more general correspondence exploits ideas from the geometry of exponential families [12, 91], as well as the junction tree framework [46, 47].

As we illustrate, these population-level results have a number of corollaries for graph selection methods. Graph selection methods for Gaussian data include neighborhood regression [60, 105] and the graphical Lasso [30, 74, 78, 24], which corresponds to maximizing an ℓ_1-regularized version of the Gaussian likelihood. Alternative methods for selection of discrete graphical models include the classical Chow-Liu algorithm for trees [21]; techniques based on conditional entropy or mutual information [3, 11]; and nodewise logistic regression for discrete graphical models with pairwise interactions [43, 73]. Our population-level results imply that minor variants of the graphical Lasso and neighborhood regression methods, though originally developed for Gaussian data, remain consistent for trees and the broader class of graphical models with singleton separator sets. They also convey a cautionary message, in that these methods will be inconsistent (generically) for other types of graphs. We also describe a new method for neighborhood selection in an arbitrary sparse graph, based on linear regression over subsets of variables. This method is most useful for bounded-degree graphs with correlation decay, but less computationally tractable for larger graphs.

In addition, we show that our methods for graph selection may be adapted to handle noisy or missing data in a seamless manner. Naively applying nodewise logistic regression when observations are systematically corrupted yields estimates that are biased even in the limit of infinite data. There are various corrections available, such as multiple imputation [79] and the expectation-maximization (EM) algorithm [25], but in general, these methods are not guaranteed to be statistically consistent due to local optima. To the best of our knowledge, our work provides the first method that is provably consistent under high-dimensional scaling for estimating the structure of discrete graphical models with corrupted observations. Further background on corrupted data methods for low-dimensional logistic regression may be found in Carroll et al. [18] and Ibrahim et al. [40].

The remainder of this chapter is organized as follows: In Section 5.2, we provide brief background and notation on graphical models and describe the classes of augmented covariance matrices we will consider. In Section 5.3, we state our main population-level result (Theorem 5.1) on the relationship between the support of generalized inverse covariance matrices and the edge structure of a discrete graphical model, and then develop a number of corollaries. The proof of Theorem 5.1 is provided in Section 5.3.4, with proofs of corollaries and more technical results deferred to the appendices. In Section 5.4, we develop consequences of our population-level results in the context of specific methods for graphical model selection. We provide simulation results in Section 5.4.4 in order to confirm the accuracy of our theoretically-predicted scaling laws, dictating how many samples are required (as a function of graph size and maximum degree) to recover the graph correctly.

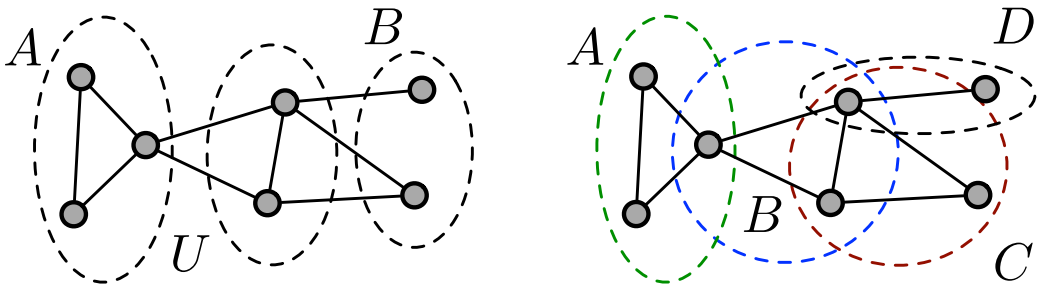
5.2 Background and problem setup

In this section, we provide background on graphical models and exponential families. We then present a simple example illustrating the phenomena and methodology underlying this chapter.

5.2.1 Undirected graphical models

An *undirected graphical model* or *Markov random field* (MRF) is a family of probability distributions respecting the structure of a fixed graph. We begin with some basic graph-theoretic terminology. An undirected graph $G = (V, E)$ consists of a collection of vertices $V = \{1, 2, \ldots, p\}$ and a collection of unordered[1] vertex pairs $E \subseteq V \times V$. A *vertex cutset* is a subset U of vertices whose removal breaks the graph into two or more nonempty components (see Figure 5.1(a)). A *clique* is a subset $C \subseteq V$ such that $(s, t) \in E$ for all distinct $s, t \in C$. The cliques in Figure 5.1(b) are all *maximal*, meaning they are not properly contained within any other clique. For $s \in V$, we define the neighborhood $N(s) := \{t \in V \mid (s, t) \in E\}$ to be the set of vertices connected to s by an edge.

For an undirected graph G, we associate to each vertex $s \in V$ a random variable X_s taking values in a space $\mathcal{X}$. For any subset $A \subseteq V$, we define $X_A := \{X_s, s \in A\}$, and for three subsets of vertices, A, B and U, we write $X_A \perp\!\!\!\perp X_B \mid X_U$ to mean that the random vector X_A is conditionally independent of X_B given X_U. The notion of a Markov random field may be defined in terms of certain *Markov properties* indexed by vertex cutsets, or in terms of a *factorization property* described by the graph cliques.



(a) (b)

Figure 5.1: (a) Illustration of a vertex cutset: when the set U is removed, the graph breaks into two disjoint subsets of vertices A and B. (b) Illustration of maximal cliques, corresponding to fully-connected subsets of vertices.

[1]No distinction is made between the edge (s, t) and the edge (t, s). In this chapter, we forbid graphs with self-loops, meaning $(s, s) \notin E$ for all $s \in V$.

Definition 4 (Markov property)**.** The random vector $X := (X_1, \ldots, X_p)$ is *Markov with respect to the graph* G if $X_A \perp\!\!\!\perp X_B \mid X_U$ whenever U is a vertex cutset that breaks the graph into disjoint subsets A and B.

Note that the neighborhood set $N(s)$ is always a vertex cutset for the sets $A = \{s\}$ and $B = V \backslash \{s \cup N(s)\}$. Consequently, $X_s \perp\!\!\!\perp X_{V\backslash\{s\cup N(s)\}} \mid X_{N(s)}$. This property is important for nodewise methods for graphical model selection to be discussed later.

The factorization property is defined directly in terms of the probability distribution q of the random vector X. For each clique C, a *clique compatibility function* ψ_C is a mapping from configurations $x_C = \{x_s, s \in V\}$ of variables to the positive reals. Let $\mathcal{C}$ denote the set of all cliques in G.

Definition 5 (Factorization property)**.** The distribution of X *factorizes according to* G if it may be written as a product of clique functions:

$$q(x_1, \ldots, x_p) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C). \tag{5.1}$$

The factorization may always be restricted to maximal cliques of the graph, but it is sometimes convenient to include terms for non-maximal cliques.

5.2.2 Graphical models and exponential families

By the Hammersley-Clifford theorem [7, 33, 47], the Markov and factorization properties are equivalent for any strictly positive distribution. We focus on such strictly positive distributions, in which case the factorization (5.1) may alternatively be represented in terms of an *exponential family* associated with the clique structure of G. We begin by defining this exponential family representation for the special case of binary variables ($\mathcal{X} = \{0, 1\}$), before discussing a natural generalization to m-ary discrete random variables.

Binary variables For a binary random vector $X \in \{0, 1\}^p$, we associate with each clique C—both maximal and non-maximal—a sufficient statistic $\mathbb{I}_C(x_C) := \prod_{s \in C} x_s$. Note that $\mathbb{I}_C(x_C) = 1$ if and only if $x_s = 1$ for all $s \in C$, so it is an indicator function for the event $\{x_s = 1, \ \forall s \in C\}$. In the exponential family, this sufficient statistic is weighted by a natural parameter $\theta_C \in \mathbb{R}$, and we rewrite the factorization (5.1) as

$$q_\theta(x_1, \ldots, x_p) = \exp\Big\{ \sum_{C \in \mathcal{C}} \theta_C \, \mathbb{I}_C(x_C) - \Phi(\theta) \Big\}, \tag{5.2}$$

where $\Phi(\theta) := \log \sum_{x \in \{0,1\}^p} \exp(\sum_{C \in \mathcal{C}} \theta_C \mathbb{I}_C(x_C))$ is the log normalization constant. It may be verified (cf. Proposition 4.3 of Darroch and Speed [23]) that the factorization (5.2) defines a minimal exponential family; i.e., the statistics $\{\mathbb{I}_C(x_C), \ C \in \mathcal{C}\}$ are affinely independent.

In the special case of pairwise interactions, equation (5.2) reduces to the classical *Ising model*:

$$q_\theta(x_1, \ldots, x_p) = \exp\Big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - \Phi(\theta) \Big\}. \tag{5.3}$$

The model (5.3) is a particular instance of a pairwise Markov random field.

Multinomial variables In order to generalize the Ising model to non-binary variables, say $\mathcal{X} = \{0, 1, \ldots, m-1\}$, we introduce a larger set of sufficient statistics. We first illustrate this extension for a pairwise Markov random field. For each node $s \in V$ and configuration $j \in \mathcal{X}_0 := \mathcal{X} \backslash \{0\} = \{1, 2, \ldots, m-1\}$, we introduce the binary-valued indicator function

$$\mathbb{I}_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j, \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

We also introduce a vector $\theta_s = \{\theta_{s;j}, j \in \mathcal{X}_0\}$ of natural parameters associated with these sufficient statistics. Similarly, for each edge $(s,t) \in E$ and configuration $(j,k) \in \mathcal{X}_0^2 := \mathcal{X}_0 \times \mathcal{X}_0$, we introduce the binary-valued indicator function $\mathbb{I}_{st;jk}$ for the event $\{x_s = j, x_t = k\}$, as well as the collection $\theta_{st} := \{\theta_{st;jk}, (j,k) \in \mathcal{X}_0^2\}$ of natural parameters. Then any pairwise Markov random field over m-ary random variables may be written in the form

$$q_\theta(x_1, \ldots, x_p) = \exp\Big\{ \sum_{s \in V} \langle \theta_s, \mathbb{I}_s(x_s) \rangle + \sum_{(s,t) \in E} \langle \theta_{st}, \mathbb{I}_{st}(x_s, x_t) \rangle - \Phi(\theta) \Big\}, \tag{5.5}$$

where we have used the shorthand $\langle \theta_s, \mathbb{I}_s(x_s) \rangle := \sum_{j=1}^{m-1} \theta_{s;j} \mathbb{I}_{s;j}(x_s)$ and

$$\langle \theta_{st}, \mathbb{I}_{st}(x_s, x_t) \rangle := \sum_{j,k=1}^{m-1} \theta_{st;jk} \mathbb{I}_{st;jk}(x_s, x_t).$$

Note that equation (5.5) defines a minimal exponential family, where the dimension is $|V|(m-1) + |E|(m-1)^2$ [23]. Furthermore, the family (5.5) is a natural generalization of the Ising model (5.3); in particular, when $m = 2$, we have a single sufficient statistic $\mathbb{I}_{s;1}(x_s) = x_s$ for each vertex, and a single sufficient statistic $\mathbb{I}_{st;11}(x_s, x_t) = x_s x_t$ for each edge. (We have omitted the additional subscripts 1 or 11 in our earlier notation for the Ising model, since they are superfluous in that case.)

Finally, for a graphical model involving higher-order interactions, we require additional sufficient statistics. For each clique $C \in \mathcal{C}$, we define the subset of configurations

$$\mathcal{X}_0^{|C|} := \underbrace{\mathcal{X}_0 \times \cdots \times \mathcal{X}_0}_{C \text{ times}} = \big\{ (j_s, s \in C) \in \mathcal{X}^{|C|} : j_s \neq 0 \quad \forall s \in C \big\},$$

a set of cardinality $(m-1)^{|C|}$. As before, $\mathcal{C}$ is the set of all maximal and non-maximal cliques. For any configuration $J = \{j_s, s \in C\} \in \mathcal{X}_0^{|C|}$, we define the corresponding indicator function

$$\mathbb{I}_{C;J}(x_C) = \begin{cases} 1 & \text{if} \quad x_C = J, \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

We then consider the general multinomial exponential family

$$q_\theta(x_1, \ldots, x_p) = \exp\Big\{ \sum_{C \in \mathcal{C}} \langle \theta_C, \mathbb{I}_C \rangle - \Phi(\theta) \Big\}, \text{ for } x_s \in \mathcal{X} = \{0, 1, \ldots, m-1\}, \tag{5.7}$$

with $\langle \theta_C, \mathbb{I}_C(x_C) \rangle = \sum_{J \in \mathcal{X}_0^{|C|}} \theta_{C;J} \mathbb{I}_{C;J}(x_C)$. Note that our previous models—namely, the binary models (5.2) and (5.3), as well as the pairwise multinomial model (5.5)—are special cases of this general factorization.

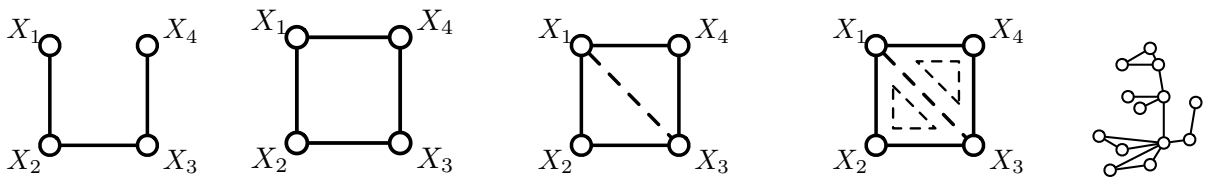
Recall that an exponential family is *minimal* if no nontrivial linear combination of sufficient statistics is almost surely equal to a constant. The family is *regular* if $\{\theta : \Phi(\theta) < \infty\}$ is an open set. As will be relevant later, the exponential families described in this section are all minimal and regular [23].

5.2.3 Covariance matrices and beyond

We now turn to a discussion of the phenomena that motivate the analysis of this chapter. Consider the usual covariance matrix $\Sigma = \mathrm{cov}(X_1, \ldots, X_p)$. When X is jointly Gaussian, it is an immediate consequence of the Hammersley-Clifford theorem that the sparsity pattern of the precision matrix $\Gamma = \Sigma^{-1}$ reflects the graph structure—that is, $\Gamma_{st} = 0$ whenever $(s,t) \notin E$. More precisely, Γ_{st} is a scalar multiple of the correlation of X_s and X_t conditioned on $X_{\setminus\{s,t\}}$ (cf. Lauritzen [47]). For non-Gaussian distributions, however, the conditional correlation will be a function of $X_{\setminus\{s,t\}}$, and it is unknown whether the entries of Γ have any relationship with the strengths of correlations along edges in the graph.

Nonetheless, it is tempting to conjecture that inverse covariance matrices might be related to graph structure in the non-Gaussian case. We explore this possibility by considering a simple case of the binary Ising model (5.3).

Example 5.1. *Consider a simple chain graph on four nodes, as illustrated in Figure 5.2(a). In terms of the factorization (5.3), let the node potentials be $\theta_s = 0.1$ for all $s \in V$ and the edge potentials be $\theta_{st} = 2$ for all $(s,t) \in E$. For a multivariate Gaussian graphical model defined on G, standard theory predicts that the inverse covariance matrix $\Gamma = \Sigma^{-1}$ of the distribution is graph-structured: $\Gamma_{st} = 0$ if and only if $(s,t) \notin E$. Surprisingly, this is also the case for the chain graph with binary variables (see panel (f)). However, this statement is* not *true for the single-cycle graph shown in panel (b). Indeed, as shown in panel (g), the inverse covariance matrix has no nonzero entries at all. Curiously, for the more complicated graph in (e), we again observe a graph-structured inverse covariance matrix.*



(a) Chain (b) Single cycle (c) Edge augmented (d) With 3-cliques (e) Dino

$$\Gamma_{\text{chain}} = \begin{bmatrix} 9.80 & -3.59 & 0 & 0 \\ -3.59 & 34.30 & -4.77 & 0 \\ 0 & -4.77 & 34.30 & -3.59 \\ 0 & 0 & -3.59 & 9.80 \end{bmatrix} \qquad \Gamma_{\text{loop}} = \begin{bmatrix} 51.37 & -5.37 & -0.17 & -5.37 \\ -5.37 & 51.37 & -5.37 & -0.17 \\ -0.17 & -5.37 & 51.37 & -5.37 \\ -5.37 & -0.17 & -5.37 & 51.37 \end{bmatrix}$$

(f) (g)

Figure 5.2: (a)–(e) Different examples of graphical models. (f) Inverse covariance for chain graph in (a). (g) Inverse covariance for single-cycle graph in (b).

Still focusing on the single-cycle graph in panel (b), suppose that instead of considering the ordinary covariance matrix, we compute the covariance matrix of the augmented *random vector* $(X_1, X_2, X_3, X_4, X_1X_3)$*, where the extra term* X_1X_3 *is represented by the dotted edge shown in panel (c). The* 5×5 *inverse of this generalized covariance matrix takes the form*

$$\Gamma_{aug} = 10^3 \times \begin{bmatrix} 1.15 & -0.02 & 1.09 & -0.02 & -1.14 \\ -0.02 & 0.05 & -0.02 & 0 & 0.01 \\ 1.09 & -0.02 & 1.14 & -0.02 & -1.14 \\ -0.02 & 0 & -0.02 & 0.05 & 0.01 \\ -1.14 & 0.01 & -1.14 & 0.01 & 1.19 \end{bmatrix}. \tag{5.8}$$

This matrix safely separates nodes 1 *and* 4*, but the entry corresponding to the non-edge* $(1, 3)$ *is* not *equal to zero. Indeed, we would observe a similar phenomenon if we chose to augment the graph by including the edge* $(2, 4)$ *rather than* $(1, 3)$*. This example shows that the usual inverse covariance matrix is not always graph-structured, but inverses of augmented matrices involving higher-order interaction terms may reveal graph structure.*

Now let us consider a more general graphical model that adds the 3-clique interaction terms shown in panel (d) to the usual Ising terms. We compute the covariance matrix of the augmented vector

$$\Psi(X) = \big\{X_1, X_2, X_3, X_4, X_1X_2,\ X_2X_3,\ X_3X_4, \\ X_1X_4,\ X_1X_3,\ X_1X_2X_3,\ X_1X_3X_4\big\} \in \{0, 1\}^{11}.$$

Empirically, one may show that the 11×11 *inverse* $(\text{cov}[\Psi(X)])^{-1}$ *respects aspects of the graph structure: there are zeros in position* (α, β)*, corresponding to the associated functions*

$X_\alpha = \prod_{s\in\alpha} X_s$ and $X_\beta = \prod_{s\in\beta} X_\beta$, whenever α and β do not lie within the same maximal clique. (For instance, this applies to the pairs $(\alpha,\beta) = (\{2\},\{4\})$ and $(\alpha,\beta) = (\{2\},\{1,4\})$.)

The goal of this chapter is to understand when certain inverse covariances do (and *do not*) capture the structure of a graphical model. At its root is the principle that the augmented inverse covariance matrix $\Gamma = \Sigma^{-1}$, suitably defined, is *always* graph-structured with respect to a graph triangulation. In some cases (e.g., the dino graph in Figure 5.2(e)), we may leverage the block-matrix inversion formula [35], namely

$$\Sigma^{-1}_{A,A} = \Gamma_{A,A} - \Gamma_{A,B}\Gamma^{-1}_{B,B}\Gamma_{B,A}, \tag{5.9}$$

to conclude that the inverse of a sub-block of the augmented matrix (e.g., the ordinary covariance matrix) is still graph-structured. This relation holds whenever A and B are chosen in such a way that the second term in equation (5.9) continues to respect the edge structure of the graph. These ideas will be made rigorous in Theorem 5.1 and its corollaries in the next section.

5.3 Generalized covariance matrices and graph structure

We now state our main results on the relationship between the zero pattern of generalized (augmented) inverse covariance matrices and graph structure. In Section 5.4 to follow, we develop some consequences of these results for data-dependent estimators used in structure estimation.

We begin with some notation for defining generalized covariance matrices, stated in terms of the sufficient statistics previously defined (5.6). Recall that a clique $C \in \mathcal{C}$ is associated with the collection $\{\mathbb{I}_{C;J}, J \in \mathcal{X}_0^{|C|}\}$ of binary-valued sufficient statistics. Let $\mathcal{S} \subseteq \mathcal{C}$, and define the random vector

$$\Psi(X;\mathcal{S}) = \big\{\mathbb{I}_{C;J}, J \in \mathcal{X}_0^{|C|}, \quad C \in \mathcal{S}\big\}, \tag{5.10}$$

consisting of all the sufficient statistics indexed by elements of $\mathcal{S}$. As in the previous section, $\mathcal{C}$ contains both maximal and non-maximal cliques.

We will often be interested in situations where $\mathcal{S}$ contains all subsets of a given set. For a subset $A \subseteq V$, let $\mathrm{pow}(A)$ denote the collection of all $2^{|A|} - 1$ nonempty subsets of A. We extend this notation to $\mathcal{S}$ by defining

$$\mathrm{pow}(\mathcal{S}) := \bigcup_{C\in\mathcal{S}} \mathrm{pow}(C).$$

5.3.1 Triangulation and block structure

Our first main result concerns a connection between the inverses of generalized inverse covariance matrices associated with the model (5.7) and any triangulation of the underlying graph G. The notion of a triangulation is defined in terms of chordless cycles, which are sequences of distinct vertices $\{s_1, \ldots, s_\ell\}$ such that:

- $(s_i, s_{i+1}) \in E$ for all $1 \leq i \leq \ell - 1$, and also $(s_\ell, s_1) \in E$;
- no other nodes in the cycle are connected by an edge.

As an illustration, the 4-cycle in Figure 5.2(b) is a chordless cycle.

Definition 6 (Triangulation)**.** Given an undirected graph $G = (V, E)$, a *triangulation* is an augmented graph $\widetilde{G} = (V, \widetilde{E})$ that contains no chordless cycles of length greater than 3.

Note that a tree is trivially triangulated, since it contains no cycles. On the other hand, the chordless 4-cycle in Figure 5.2(b) is the simplest example of a non-triangulated graph. By adding the single edge $(1, 3)$ to form the augmented edge set $\widetilde{E} = E \cup \{(1, 3)\}$, we obtain the triangulated graph $\widetilde{G} = (V, \widetilde{E})$ shown in panel (c). One may check that the more complicated graph shown in Figure 5.2(e) is triangulated, as well.

Our first result concerns the inverse Γ of the matrix $\mathrm{cov}(\Psi(X; \widetilde{\mathcal{C}}))$, where $\widetilde{\mathcal{C}}$ is the set of all cliques arising from some triangulation $\widetilde{G}$ of G. For any two subsets $A, B \in \widetilde{\mathcal{C}}$, we write $\Gamma(A, B)$ to denote the sub-block of Γ indexed by all indicator statistics on A and B, respectively. (Note that we are working with respect to the exponential family representation over the triangulated graph $\widetilde{G}$.) Given our previously-defined sufficient statistics (5.6), the sub-block $\Gamma(A, B)$ has dimensions $d_A \times d_B$, where

$$d_A := (m-1)^{|A|}, \quad \text{and} \quad d_B := (m-1)^{|B|}.$$

For example, when $A = \{s\}$ and $B = \{t\}$, the submatrix $\Gamma(A, B)$ has dimension $(m-1) \times (m-1)$. With this notation, we have the following result:

Theorem 5.1. *[Triangulation and block graph-structure.] Consider an arbitrary discrete graphical model of the form (5.7), and let $\widetilde{\mathcal{C}}$ be the set of all cliques in any triangulation of G. Then the generalized covariance matrix $\mathrm{cov}(\Psi(X; \widetilde{\mathcal{C}}))$ is invertible, and its inverse Γ is* block graph-structured*:*

(a) For any two subsets $A, B \in \widetilde{\mathcal{C}}$ that are not subsets of the same maximal clique, the block $\Gamma(A, B)$ is identically zero.

(b) For almost all parameters θ, the entire block $\Gamma(A, B)$ is nonzero whenever A and B belong to a common maximal clique.

In part (b), "almost all" refers to all parameters θ apart from a set of Lebesgue measure zero. The proof of Theorem 5.1, which we provide in Section 5.3.4, relies on the geometry of exponential families [12, 91] and certain aspects of convex analysis [75], involving the log partition function Φ and its Fenchel-Legendre dual Φ^*. Although we have stated Theorem 5.1 for discrete variables, it easily generalizes to other classes of random variables. The only difference is the specific choices of sufficient statistics used to define the generalized covariance matrix. This generality becomes apparent in the proof.

To provide intuition for Theorem 5.1, we consider its consequences for specific graphs. When the original graph is a tree (such as the graph in Figure 5.2(a)), it is already triangulated, so the set $\widetilde{\mathcal{C}}$ is equal to the edge set E, together with singleton nodes. Hence, Theorem 5.1 implies that the inverse Γ of the matrix of sufficient statistics for vertices and edges is graph-structured, and blocks of nonzeros in Γ correspond to edges in the graph. In particular, we may apply Theorem 5.1(a) to the subsets $A = \{s\}$ and $B = \{t\}$, where s and t are distinct vertices with $(s,t) \notin E$, and conclude that the $(m-1) \times (m-1)$ sub-block $\Gamma(A, B)$ is equal to zero.

When G is not triangulated, however, we may need to invert a larger augmented covariance matrix and include sufficient statistics over pairs $(s,t) \notin E$, as well. For instance, the augmented graph shown in Figure 5.2(c) is a triangulation of the chordless 4-cycle in panel (b). The associated set of maximal cliques is given by $\bar{\mathcal{C}} = \{(1,2), (2,3), (3,4), (1,4), (1,3)\}$; among other predictions, our theory guarantees that the generalized inverse covariance Γ will have zeros in the sub-block $\Gamma(\{2\}, \{4\})$.

5.3.2 Separator sets and graph structure

In fact, it is not necessary to take sufficient statistics over *all* maximal cliques, and we may consider a slightly smaller augmented covariance matrix. (This simpler type of augmented covariance matrix explains the calculations given in Section 5.2.3.)

By classical graph theory, any triangulation $\widetilde{G}$ gives rise to a *junction tree* representation of G. Nodes in the junction tree are subsets of V corresponding to maximal cliques of $\widetilde{G}$, and the intersection of any two adjacent cliques C_1 and C_2 is referred to as a *separator set* $S = C_1 \cap C_2$. Furthermore, any junction tree must satisfy the *running intersection property,* meaning that for any two nodes of the junction tree—say corresponding to cliques C and D—the intersection $C \cap D$ must belong to every separator set on the unique path between C and D. The following result shows that it suffices to construct generalized covariance matrices augmented by separator sets:

Corollary 5.1. *Let $\mathcal{S}$ be the set of separator sets in any triangulation of G, and let Γ be the inverse of* $\mathrm{cov}(\Psi(X; V \cup \mathit{pow}(\mathcal{S})))$. *Then $\Gamma(\{s\}, \{t\}) = 0$ whenever $(s,t) \notin \widetilde{E}$.*

Note that $V \cup \mathrm{pow}(\mathcal{S}) \subseteq \widetilde{\mathcal{C}}$, and the set of sufficient statistics considered in Corollary 5.1 is generally much smaller than the set of sufficient statistics considered in Theorem 5.1.

Hence, the generalized covariance matrix of Corollary 5.1 has a smaller dimension than the generalized covariance matrix of Theorem 5.1, which becomes significant when we consider exploiting these population-level results for statistical estimation.

The graph in Figure 5.2(c) of Example 5.1 and the associated matrix in equation (5.8) provide a concrete example of Corollary 5.1 in action. In this case, the single separator set in the triangulation is $\{1, 3\}$, so when $\mathcal{X} = \{0, 1\}$, augmenting the usual covariance matrix with the additional sufficient statistic $\mathbb{I}_{13;11}(x_1, x_3) = x_1 x_3$ and taking the inverse yields a graph-structured matrix. Indeed, since $(2, 4) \notin \widetilde{E}$, we observe that $\Gamma_{\text{aug}}(2, 4) = 0$ in equation (5.8), consistent with the result of Corollary 5.1.

Although Theorem 5.1 and Corollary 5.1 are clean population-level results, however, forming an appropriate augmented covariance matrix requires prior knowledge of the graph—namely, which edges are involved in a suitable triangulation. This is infeasible in settings where the goal is to recover the edge structure of the graph. Corollary 5.1 is most useful for edge recovery when G admits a triangulation with only singleton separator sets, since then $V \cup \text{pow}(\mathcal{S}) = V$. In particular, this condition holds when G is a tree. The following corollary summarizes our result:

Corollary 5.2. *For any graph with singleton separator sets, the inverse* Γ *of the covariance matrix* $\text{cov}(\Psi(X; V))$ *of vertex statistics is graph-structured. (This class includes trees as a special case.)*

In the special case of binary variables, we have $\Psi(X; V) = (X_1, \ldots, X_p)$, so Corollary 5.2 implies that the inverse of the ordinary covariance matrix $\text{cov}(X)$ is graph-structured. For m-ary variables, $\text{cov}(\Psi(X; V))$ is a matrix of dimensions $(m-1)p \times (m-1)p$ involving indicator functions for each variable. Again, we may relate this corollary to Example 5.1—the inverse covariance matrices for the tree graph in panel (a) and the dino graph in panel (e) are exactly graph-structured. Although the dino graph is not a tree, it possesses the nice property that the only separator sets in its junction tree are singletons.

Corollary 5.1 also guarantees that inverse covariances may be partially graph-structured, in the sense that $\Gamma(\{s\}, \{t\}) = 0$ for any pair of vertices (s, t) separable by a singleton separator set, where $\Gamma = (\text{cov}(\Psi(X; V)))^{-1}$. This is because for any such pair (s, t), we may form a junction tree with two nodes, one containing s and one containing t, and apply Corollary 5.1. Indeed, the matrix Γ defined over singleton vertices is agnostic to which triangulation we choose for the graph.

In settings where there exists a junction tree representation of the graph with only singleton separator sets, Corollary 5.2 has a number of useful implications for the consistency of methods that have traditionally only been applied for edge recovery in Gaussian graphical models: for tree-structured discrete graphs, it suffices to estimate the support of $(\text{cov}(\Psi(X; V)))^{-1}$ from the data. We will review methods for Gaussian graphical model selection and describe their analogs for discrete tree graphs in Sections 5.4.1 and 5.4.2.

5.3.3 Generalized covariances and neighborhood structure

Theorem 5.1 also has a corollary that is relevant for nodewise neighborhood selection approaches to graph selection [60, 74], which are applicable to graphs with arbitrary topologies. Nodewise methods use the basic observation that recovering the edge structure of G is equivalent to recovering the neighborhood set $N(s) = \{t \in V : (s,t) \in E\}$ for each vertex $s \in V$. For a given node $s \in V$ and positive integer d, consider the collection of subsets

$$\mathcal{S}(s;d) := \big\{U \subseteq V \backslash \{s\}, \quad |U| = d\big\}.$$

The following corollary provides an avenue for recovering $N(s)$ based on the inverse of a certain generalized covariance matrix:

Corollary 5.3. *[Neighborhood selection] For any graph and node $s \in V$ with $\deg(s) \leq d$, the inverse Γ of the matrix $\operatorname{cov}(\Psi(X; \{s\} \cup pow(\mathcal{S}(s;d))))$ is s-block graph-structured; i.e., $\Gamma(\{s\}, B) = 0$ whenever $\{s\} \neq B \subsetneq N(s)$. In particular, $\Gamma(\{s\}, \{t\}) = 0$ for all vertices $t \notin N(s)$.*

Note that $\operatorname{pow}(\mathcal{S}(s;d))$ is the set of subsets of all candidate neighborhoods of s of size d. This result follows from Theorem 5.1 (and the related Corollary 5.1) by constructing a particular junction tree for the graph, in which s is separated from the rest of the graph by $N(s)$. Due to the well-known relationship between the rows of an inverse covariance matrix and linear regression coefficients [60], Corollary 5.3 motivates the following neighborhood-based approach to graph selection: For a fixed vertex $s \in V$, perform a single *linear regression* of $\Psi(X; \{s\})$ on the vector $\Psi(X; \operatorname{pow}(\mathcal{S}(s;d)))$. Via elementary algebra and an application of Corollary 5.3, the resulting regression vector will expose the neighborhood $N(s)$ in an arbitrary discrete graphical model; i.e., the indicators $\Psi(X; \{t\})$ corresponding to X_t will have a nonzero weight only if $t \in N(s)$. We elaborate on this connection in Section 5.4.2.

5.3.4 Proof of Theorem 5.1

We now turn to the proof of Theorem 5.1, which is based on certain fundamental correspondences arising from the theory of exponential families [5, 12, 91]. Recall that our exponential family (5.7) has binary-valued indicator functions (5.6) as its sufficient statistics. Let D denote the cardinality of this set and let $\mathbb{I} : \mathcal{X}^p \to \{0,1\}^D$ denote the multivariate function that maps each configuration $x \in \mathcal{X}^p$ to the vector $\mathbb{I}(x)$ obtained by evaluating the D indicator functions on x. Using this notation, our exponential family may be written in the compact form $q_\theta(x) = \exp\{\langle \theta, \mathbb{I}(x) \rangle - \Phi(\theta)\}$, where

$$\langle \theta, \mathbb{I}(x) \rangle = \sum_{C \in \mathcal{C}} \langle \theta_C, \mathbb{I}_C(x) \rangle \; = \sum_{C \in \mathcal{C}} \sum_{J \in \mathcal{X}_0^{|C|}} \theta_{C;J} \mathbb{I}_{C;J}(x_C).$$

Since this exponential family is known to be minimal, we are guaranteed [23] that

$$\nabla \Phi(\theta) = \mathbb{E}_\theta[\mathbb{I}(X)], \quad \text{and} \quad \nabla^2 \Phi(\theta) = \operatorname{cov}_\theta[\mathbb{I}(X)],$$

where $\mathbb{E}_\theta$ and cov_θ denote (respectively) the expectation and covariance taken under the density q_θ [12, 91]. The conjugate dual [75] of the cumulant function is given by

$$\Phi^*(\mu) := \sup_{\theta \in \mathbb{R}^D} \{\langle \mu, \theta \rangle - \Phi(\theta)\}.$$

The function Φ^* is always convex and takes values in $\mathbb{R} \cup \{+\infty\}$. From known results [91], the dual function Φ^* is finite only for $\mu \in \mathbb{R}^D$ belonging to the marginal polytope

$$\mathcal{M} := \{\mu \in \mathbb{R}^p \mid \exists \text{ some density } q \text{ s.t. } \sum_x q(x)\mathbb{I}(x) = \mu\}. \tag{5.11}$$

The following lemma, proved in Appendix C.1.1, provides a connection between the covariance matrix and the Hessian of Φ^*:

Lemma 5.1. *Consider a regular, minimal exponential family, and define $\mu = \mathbb{E}_\theta[\mathbb{I}(X)]$ for any fixed $\theta \in \Omega = \{\theta : \Phi(\theta) < \infty\}$. Then*

$$\big(\text{cov}_\theta[\mathbb{I}(X)]\big)^{-1} = \nabla^2 \Phi^*(\mu). \tag{5.12}$$

Note that the minimality and regularity of the family implies that $\text{cov}_\theta[\mathbb{I}(X)]$ is strictly positive definite, so the matrix is invertible.

For any $\mu \in \text{int}(\mathcal{M})$, let $\theta(\mu) \in \mathbb{R}^D$ denote the unique natural parameter θ such that $\nabla\Phi(\theta) = \mu$. It is known [91] that the negative dual function $-\Phi^*$ is linked to the Shannon entropy via the relation

$$-\Phi^*(\mu) = H(q_{\theta(\mu)}(x)) \;=\; -\sum_{x \in \mathcal{X}^p} q_{\theta(\mu)}(x) \log q_{\theta(\mu)}(x). \tag{5.13}$$

In general, expression (5.13) does *not* provide a straightforward way to compute $\nabla^2\Phi^*$, since the mapping $\mu \mapsto \theta(\mu)$ may be extremely complicated. However, when the exponential family is defined with respect to a triangulated graph, Φ^* has an explicit closed-form representation in terms of the mean parameters μ. Consider a junction tree triangulation of the graph, and let $(\bar{\mathcal{C}}, \mathcal{S})$, be the collection of maximal cliques and separator sets, respectively. By the junction tree theorem [46, 91, 44], we have the factorization

$$q(x_1, \ldots, x_p) = \frac{\prod_{C \in \bar{\mathcal{C}}} q_C(x_C)}{\prod_{S \in \mathcal{S}} q_S(x_S)}, \tag{5.14}$$

where q_C and q_S are the marginal distributions over maximal clique C and separator set S. Consequently, the entropy may be decomposed into the sum

$$H(q) = -\sum_{x \in \mathcal{X}^p} q(x) \log q(x) \;=\; \sum_{C \in \bar{\mathcal{C}}} H_C(q_C) - \sum_{S \in \mathcal{S}} H_S(q_S), \tag{5.15}$$

where we have introduced the clique- and separator-based entropies

$$H_S(q_S) := -\sum_{x_S \in \mathcal{X}^{|S|}} q_S(x_S) \log q_S(x_S), \quad \text{and}$$
$$H_C(q_C) := -\sum_{x_C \in \mathcal{X}^{|C|}} q_C(x_C) \log q_C(x_C).$$

Given our choice of sufficient statistics (5.6), we show that q_C and q_S may be written explicitly as "local" functions of mean parameters associated with C and S. For each subset $A \subseteq V$, let $\mu_A \in (m-1)^{|A|}$ be the associated collection of mean parameters, and let

$$\mu_{\text{pow}(A)} := \{\mu_B \mid \emptyset \neq B \subseteq A\}$$

be the set of mean parameters associated with all nonempty subsets of A. Note that $\mu_{\text{pow}(A)}$ contains a total of $\sum_{k=1}^{|A|} \binom{|A|}{k}(m-1)^k = m^{|A|} - 1$ parameters, corresponding to the number of degrees of freedom involved in specifying a marginal distribution over the random vector x_A. Moreover, $\mu_{\text{pow}(A)}$ uniquely determines the marginal distribution q_A:

Lemma 5.2. *For any marginal distribution q_A in the $m^{|A|}$-dimensional probability simplex, there is a unique mean parameter vector $\mu_{pow(A)}$ and matrix M_A such that $q_A = M_A \cdot \mu_{pow(A)}$.*

For the proof, see Appendix C.1.2.

We now combine the dual representation (5.13) with the decomposition (5.15), along with the matrices $\{M_C, M_S\}$ from Lemma 5.2, to conclude that

$$-\Phi^*(\mu) = \sum_{C \in \overline{\mathcal{C}}} H_C(M_C(\mu_{\text{pow}(C)})) - \sum_{S \in \mathcal{S}} H_S(M_S(\mu_{\text{pow}(S)})). \tag{5.16}$$

Now consider two subsets $A, B \in \widetilde{\mathcal{C}}$ that are not contained in the same maximal clique. Suppose A is contained within maximal clique C. Differentiating expression (5.16) with respect to μ_A preserves only terms involving q_C and q_S, where S is any separator set such that $A \subseteq S \subseteq C$. Since $B \subsetneq C$, we clearly cannot have $B \subseteq S$. Consequently, all cross-terms arising from the clique C and its associated separator sets vanish when we take a second derivative with respect to μ_B. Repeating this argument for any other maximal clique C' containing A but not B, we have $\frac{\partial^2 \Phi^*}{\partial \mu_A \partial \mu_B}(\mu) = 0$. This proves part (a).

Turning to part (b), note that if A and B are in the same maximal clique, the expression obtained by taking second derivatives of the entropy results in an algebraic expression with only finitely many solutions in the parameters μ (consequently, also θ). Hence, assuming the θ's are drawn from a continuous distribution, the corresponding values of the block $\Gamma(A, B)$ are a.s. nonzero.

5.4 Consequences for graph structure estimation

Moving beyond the population level, we now state and prove several results concerning the statistical consistency of different methods—both known and some novel—for graph selection in discrete graphical models, based on i.i.d. draws from a discrete graph. For sparse Gaussian models, existing methods that exploit sparsity of the inverse covariance matrix fall into two main categories: global graph selection methods (e.g., [24, 30, 78, 74]) and local (nodewise) neighborhood selection methods [60, 105]. We divide our discussion accordingly.

5.4.1 Graphical Lasso for singleton separator graphs

We begin by describing how a combination of our population-level results and some concentration inequalities may be leveraged to analyze the statistical behavior of log-determinant methods for discrete graphical models with singleton separator sets, and suggest extensions of these methods when observations are systematically corrupted by noise or missing data. Given a p-dimensional random vector $(X_1, \ldots, X_p)$ with covariance Σ^*, consider the estimator

$$\widehat{\Theta} \in \arg\min_{\Theta \succeq 0} \{\operatorname{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta) + \lambda_n \sum_{s \neq t} |\Theta_{st}|\}, \tag{5.17}$$

where $\widehat{\Sigma}$ is an estimator for Σ^*. For multivariate Gaussian data, this program is an ℓ_1-regularized maximum likelihood estimate known as the *graphical Lasso* and is a well-studied method for recovering the edge structure in a Gaussian graphical model [4, 30, 100, 78]. Although the program (5.17) has no relation to the MLE in the case of a discrete graphical model, it may still be useful for estimating $\Theta^* := (\Sigma^*)^{-1}$. Indeed, as shown in Ravikumar et al. [74], existing analyses of the estimator (5.17) require only tail conditions such as sub-Gaussianity in order to guarantee that the sample minimizer is close to the population minimizer. The analysis of this chapter completes the missing link by guaranteeing that the population-level inverse covariance is in fact graph-structured. Consequently, we obtain the interesting result that the program (5.17)—even though it is ostensibly derived from Gaussian considerations—is a consistent method for recovering the structure of any binary graphical model with singleton separator sets.

In order to state our conclusion precisely, we introduce additional notation. Consider a general estimate $\widehat{\Sigma}$ of the covariance matrix Σ such that

$$\mathbb{P}\big[\|\widehat{\Sigma} - \Sigma^*\|_{\max} \geq \varphi(\Sigma^*)\sqrt{\frac{\log p}{n}}\big] \leq c\exp(-\psi(n,p)) \tag{5.18}$$

for functions φ and ψ, where $\|\cdot\|_{\max}$ denotes the elementwise ℓ_∞-norm. In the case of fully-observed i.i.d. data with sub-Gaussian parameter σ^2, where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T - \overline{x}\overline{x}^T$ is the usual sample covariance, this bound holds with $\varphi(\Sigma^*) = \sigma^2$ and $\psi(n,p) = c'\log p$.

As in past analysis of the graphical Lasso [74], we require a certain *mutual incoherence* condition on the true covariance matrix Σ^* to control the correlation of non-edge variables

with edge variables in the graph. Let $\Gamma^* = \Sigma^* \otimes \Sigma^*$, where $\otimes$ denotes the Kronecker product. Then Γ^* is a $p^2 \times p^2$ matrix indexed by vertex pairs. The incoherence condition is given by

$$\max_{e \in S^c} \|\Gamma^*_{eS}(\Gamma^*_{SS})^{-1}\|_1 \leq 1 - \alpha, \qquad \alpha \in (0, 1], \tag{5.19}$$

where $S := \{(s,t) : \Theta^*_{st} \neq 0\}$ is the set of vertex pairs corresponding to nonzero entries of the precision matrix Θ^*—equivalently, the edge set of the graph, by our theory on tree-structured discrete graphs. For more intuition on the mutual incoherence condition, see Ravikumar et al. [74].

With this notation, our global edge recovery algorithm proceeds as follows:

Algorithm 5.1 (Graphical Lasso)**.**

1. *Form a suitable estimate* $\widehat{\Sigma}$ *of the true covariance matrix* Σ*.*
2. *Optimize the graphical Lasso program* (5.17) *with parameter* λ_n*, and denote the solution by* $\widehat{\Theta}$*.*
3. *Threshold the entries of* $\widehat{\Theta}$ *at level* τ_n *to obtain an estimate of* Θ^**.*

It remains to choose the parameters (λ_n, τ_n). In the following corollary, we will establish statistical consistency of $\widehat{\Theta}$ under the following settings:

$$\lambda_n \geq \frac{c_1}{\alpha}\sqrt{\frac{\log p}{n}}, \qquad \tau_n = c_2\left\{\frac{c_1}{\alpha}\sqrt{\frac{\log p}{n}} + \lambda_n\right\}, \tag{5.20}$$

where α is the incoherence parameter in inequality (5.19) and c_1, c_2 are universal positive constants. The following result applies to Algorithm 5.1 when $\widehat{\Sigma}$ is the sample covariance matrix and (λ_n, τ_n) are chosen as in equations (5.20):

Corollary 5.4. *Consider an Ising model* (5.3) *defined by an undirected graph with singleton separator sets and with degree at most* d*, and suppose that the mutual incoherence condition* (5.19) *holds. With* $n \gtrsim d^2 \log p$ *samples, there are universal constants* (c, c') *such that with probability at least* $1 - c\exp(-c' \log p)$*, Algorithm 5.1 recovers all edges* (s,t) *with* $|\Theta^*_{st}| > \tau/2$*.*

The proof is contained in Appendix C.5.1; it is a relatively straightforward consequence of Corollary 5.1 and known concentration properties of $\widehat{\Sigma}$ as an estimate of the population covariance matrix. Hence, if $|\Theta^*_{st}| > \tau/2$ for all edges $(s,t) \in E$, Corollary 5.4 guarantees that the log-determinant method plus thresholding recovers the full graph exactly.

In the case of the standard sample covariance matrix, a variant of the graphical Lasso has been implemented by Banerjee et al. [4]. Our analysis establishes consistency of the

graphical Lasso for Ising models on single separator graphs using $n \gtrsim d^2 \log p$ samples. This lower bound on the sample size is unavoidable, as shown by information-theoretic analysis [81], and also appears in other past work on Ising models [73, 43, 3]. Our analysis also has a *cautionary message*: the proof of Corollary 5.4 relies heavily on the population-level result in Corollary 5.2, which ensures that Θ^* is graph-structured when G has only singleton separators. For a general graph, we have no guarantees that Θ^* will be graph-structured (e.g., see panel (b) in Figure 5.2), so the graphical Lasso (5.17) is *inconsistent in general*.

On the positive side, if we restrict ourselves to tree-structured graphs, the estimator (5.17) is attractive, since it relies only on an estimate $\widehat{\Sigma}$ of the population covariance Σ^* that satisfies the deviation condition (5.18). In particular, even when the samples $\{x_i\}_{i=1}^n$ are contaminated by noise or missing data, we may form a good estimate $\widehat{\Sigma}$ of Σ^*. Furthermore, the program (5.17) is always convex regardless of whether $\widehat{\Sigma}$ is positive semidefinite.

As a concrete example of how we may correct the program (5.17) to handle corrupted data, consider the case when each entry of x_i is missing independently with probability α, and the corresponding observations z_i are zero-filled for missing entries. A natural estimator is

$$\widehat{\Sigma} = \left(\frac{1}{n} \sum_{i=1}^{n} z_i z_i^T \right) \div M - \frac{1}{(1-\alpha)^2} \overline{z}\overline{z}^T, \tag{5.21}$$

where $\div$ denotes elementwise division by the matrix M with diagonal entries $(1-\alpha)$ and off-diagonal entries $(1-\alpha)^2$, correcting for the bias in both the mean and second moment terms. As in the results of Chapter 3, the deviation condition (5.18) may be shown to hold w.h.p., where $\varphi(\Sigma^*)$ scales with $(1-\alpha)$. Similarly, we may derive an appropriate estimator $\widehat{\Sigma}$ for other forms of additive or multiplicative corruption.

Generalizing to the case of m-ary discrete graphical models with $m > 2$, we may easily modify the program (5.17) by replacing the elementwise ℓ_1-penalty by the corresponding group ℓ_1-penalty, where the groups are the indicator variables for a given vertex. Precise theoretical guarantees follow from results on the group graphical Lasso [42].

5.4.2 Consequences for nodewise regression in trees

Turning to local neighborhood selection methods, recall the neighborhood-based method due to Meinshausen and Bühlmann [60]. In a Gaussian graphical model, the column corresponding to node s in the inverse covariance matrix $\Gamma = \Sigma^{-1}$ is a scalar multiple of $\widetilde{\beta} = \Sigma^{-1}_{\backslash s, \backslash s} \Sigma_{\backslash s, s}$, the limit of the linear regression vector for X_s upon $X_{\backslash s}$. Based on n i.i.d. samples from a p-dimensional multivariate Gaussian distribution, the support of the graph may then be estimated consistently under the usual Lasso scaling $n \gtrsim d \log p$, where $d = |N(s)|$.

Motivated by our population-level results on the graph structure of the inverse covariance matrix (Corollary 5.2), we now propose a method for neighborhood selection in a tree-structured graph. Although the method works for arbitrary m-ary trees, we state explicit results only in the case of the binary Ising model to avoid cluttering our presentation.

The method is based on the following steps. For each node $s \in V$, we first perform ℓ_1-regularized linear regression of X_s against $X_{\setminus s}$ by solving the modified Lasso program

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \leq b_0\sqrt{k}} \{\frac{1}{2}\beta^T \widehat{\Gamma}\beta - \widehat{\gamma}^T\beta + \lambda_n\|\beta\|_1\}, \tag{5.22}$$

where $b_0 > \|\widetilde{\beta}\|_1$ is a constant, $(\widehat{\Gamma}, \widehat{\gamma})$ are suitable estimators for $(\Sigma_{\setminus s,\setminus s}, \Sigma_{\setminus s,s})$, and λ_n is an appropriate parameter. We then combine the neighborhood estimates over all nodes via an AND operation (edge (s,t) is present if both s and t are inferred to be neighbors of each other) or an OR operation (at least one of s or t is inferred to be a neighbor of the other).

Note that the program (5.22) differs from the standard Lasso in the form of the ℓ_1-constraint. Indeed, the normal setting of the Lasso assumes a linear model where the predictor and response variables are linked by independent sub-Gaussian noise, but this is not the case for X_s and $X_{\setminus s}$ in a discrete graphical model. Furthermore, the generality of the program (5.22) allows it to be easily modified to handle corrupted variables via an appropriate choice of $(\widehat{\Gamma}, \widehat{\gamma})$, as in Chapter 3.

The following algorithm summarizes our nodewise regression procedure for recovering the neighborhood set $N(s)$ of a given node s:

Algorithm 5.2 (Nodewise method for trees)**.**

1. *Form a suitable pair of estimators* $(\widehat{\Gamma}, \widehat{\gamma})$ *for covariance submatrices* $(\Sigma_{\setminus s,\setminus s}, \Sigma_{\setminus s,s})$.
2. *Optimize the modified Lasso program* (5.22) *with parameter* λ_n*, and denote the solution by* $\widehat{\beta}$.
3. *Threshold the entries of* $\widehat{\beta}$ *at level* τ_n*, and define the estimated neighborhood set* $\widehat{N(s)}$ *as the support of the thresholded vector.*

In the case of fully-observed i.i.d. observations, we choose $(\widehat{\Gamma}, \widehat{\gamma})$ to be the recentered estimators

$$(\widehat{\Gamma}, \widehat{\gamma}) = \left(\frac{X_{\setminus s}^T X_{\setminus s}}{n} - \bar{x}_{\setminus s}\bar{x}_{\setminus s}^T, \; \frac{X_{\setminus s}^T X_s}{n} - \bar{x}_s\bar{x}_{\setminus s} \right), \tag{5.23}$$

and assign the parameters (λ_n, τ_n) according to the scaling

$$\lambda_n \gtrsim \varphi\|\widetilde{\beta}\|_2\sqrt{\frac{\log p}{n}}, \qquad \tau_n \asymp \varphi\|\widetilde{\beta}\|_2\sqrt{\frac{\log p}{n}}, \tag{5.24}$$

where $\widetilde{\beta} := \Sigma_{\setminus s,\setminus s}^{-1}\Sigma_{\setminus s,s}$ and φ is some parameter such that $\langle x_i, u\rangle$ is sub-Gaussian with parameter $\varphi^2\|u\|_2^2$ for any d-sparse vector u, and φ is independent of u. The following result applies to Algorithm 5.2 using the pairs $(\widehat{\Gamma}, \widehat{\gamma})$ and (λ_n, τ_n) defined as in equations (5.23) and (5.24), respectively.

Proposition 5.1. *Suppose we have i.i.d. observations* $\{x_i\}_{i=1}^n$ *from an Ising model and that* $n \gtrsim \varphi^2 \max\left\{\frac{1}{\lambda_{\min}(\Sigma_x)}, |||\Sigma_x^{-1}|||_\infty^2\right\} d^2 \log p$. *Then there are universal constants* (c, c', c'') *such that with probability greater than* $1 - c\exp(-c' \log p)$, *for any node* $s \in V$, *Algorithm 5.2 recovers all neighbors* $t \in N(s)$ *for which* $|\widetilde{\beta}_t| \geq c'' \varphi \|\widetilde{\beta}\|_2 \sqrt{\frac{\log p}{n}}$.

We prove this proposition in Appendix C.3, as a corollary of a more general theorem on the ℓ_∞-consistency of the program (5.22) for estimating $\widetilde{\beta}$, allowing for corrupted observations. The theorem builds upon the analysis of Chapter 3, introducing techniques for ℓ_∞-bounds and departing from the framework of a linear model with independent sub-Gaussian noise.

Remark 5.1. *Regarding the sub-Gaussian parameter* φ *appearing in Proposition 5.1, note that we may always take* $\varphi = \sqrt{d}$, *since* $|x_i^T u| \leq \|u\|_1 \leq \sqrt{d}\|u\|_2$ *when* u *is* d*-sparse and* x_i *is a binary vector. This leads to a sample complexity requirement of* $n \gtrsim d^3 \log p$. *We suspect that a tighter analysis, possibly combined with assumptions about the correlation decay of the graph, would reduce the sample complexity to the scaling* $n \gtrsim d^2 \log p$, *as required by other methods with fully-observed data [43, 3, 73]. See the simulations in Section 5.4.4 for further discussion.*

For corrupted observations, the strength and type of corruption enters into the factors (φ_1, φ_2) *appearing in the deviation bounds (C.6a) and (C.6b) below, and Proposition 5.1 has natural extensions to the corrupted case. We emphasize that although analogs of Proposition 5.1 exist for other methods of graph selection based on logistic regression and/or mutual information, the theoretical analysis of those methods does not handle corrupted data, whereas our results extend easily with the appropriate scaling.*

In the case of m-ary tree-structured graphical models with $m > 2$, we may perform multivariate regression with the multivariate group Lasso [68] for neighborhood selection, where groups are defined (as in the log-determinant method) as sets of indicators for each node. The general relationship between the best linear predictor and the block structure of the inverse covariance matrix follows from block matrix inversion, and from a population-level perspective, it suffices to perform multivariate linear regression of all indicators corresponding to a given node against all indicators corresponding to other nodes in the graph. The resulting vector of regression coefficients has nonzero blocks corresponding to edges in the graph. We may also combine these ideas with the group Lasso for multivariate regression [68] to reduce the complexity of the algorithm.

5.4.3 Consequences for nodewise regression in general graphs

Moving on from tree-structured graphical models, our method suggests a graph recovery method based on nodewise linear regression for general discrete graphs. Note that by Corollary 5.3, the inverse of $\mathrm{cov}(\Psi(X; \mathrm{pow}(\mathcal{S}(s; d))))$ is s-block graph-structured, where d is such

that $|N(s)| \leq d$. It suffices to perform a single multivariate regression of the indicators $\Psi(X; \{s\})$ corresponding to node s upon the other indicators in $\Psi(X; V \cup \text{pow}(\mathcal{S}(s; d)))$.

We again make precise statements for the binary Ising model ($m = 2$). In this case, the indicators $\Psi(X; \text{pow}(U))$ corresponding to a subset of vertices U of size d' are all $2^{d'} - 1$ distinct products of variables X_u, for $u \in U$. Hence, to recover the d neighbors of node s, we use the following algorithm. Note that knowledge of an upper bound d is necessary for applying the algorithm.

Algorithm 5.3 (Nodewise method for general graphs)**.**

1. *Use the modified Lasso program* (5.22) *with a suitable choice of* $(\widehat{\Gamma}, \widehat{\gamma})$ *and regularization parameter* λ_n *to perform a linear regression of* X_s *upon all products of subsets of variables of* $X_{\setminus s}$ *of size at most* d*. Denote the solution by* $\widehat{\beta}$*.*

2. *Threshold the entries of* $\widehat{\beta}$ *at level* τ_n*, and define the estimated neighborhood set* $\widehat{N(s)}$ *as the support of the thresholded vector.*

Our theory states that at the population level, nonzeros in the regression vector correspond exactly to subsets of $N(s)$. Hence, the statistical consistency result of Proposition 5.1 carries over with minor modifications. Since Algorithm 5.3 is essentially a version of Algorithm 5.4 with the first two steps omitted, we refer the reader to the statement and proof of Corollary 5.5 below for precise mathematical statements. Note here that since the regression vector has $\mathcal{O}(p^d)$ components, $2^d - 1$ of which are nonzero, the sample complexity of Lasso regression in step (1) of Algorithm 5.3 is $\mathcal{O}(2^d \log(p^d)) = \mathcal{O}(2^d \log p)$.

For graphs exhibiting correlation decay [11], we may reduce the computational complexity of the nodewise selection algorithm by prescreening the nodes of $V \setminus s$ before performing a Lasso-based linear regression. We define the nodewise correlation according to

$$r_C(s, t) := \sum_{x_s, x_t} |\mathbb{P}(X_s = x_s, X_t = x_t) - \mathbb{P}(X_s = x_s)\mathbb{P}(X_t = x_t)|,$$

and say that the graph exhibits *correlation decay* if there exist constants $\zeta, \kappa > 0$ such that

$$r_C(s, t) > \kappa \quad \forall (s, t) \in E, \quad \text{and} \quad r_C(s, t) \leq \exp(-\zeta r(s, t)) \tag{5.25}$$

for all $(s, t) \in V \times V$, where $r(s, t)$ is the length of the shortest path between s and t. With this notation, we then have the following algorithm for neighborhood recovery of a fixed node s in a graph with correlation decay:

Algorithm 5.4 (Nodewise method with correlation decay)**.**

1. *Compute the empirical correlations*

$$\widehat{r}_C(s, t) := \sum_{x_s, x_t} |\widehat{\mathbb{P}}(X_s = x_s, X_t = x_t) - \widehat{\mathbb{P}}(X_s = x_s)\widehat{\mathbb{P}}(X_t = x_t)|$$

between s *and all other nodes* $t \in V$*, where* $\widehat{\mathbb{P}}$ *denotes the empirical distribution.*

2. *Let* $\mathcal{C} := \{t \in V : \widehat{r}_C(s,t) > \kappa/2\}$ *be the candidate set of nodes with sufficiently high correlation. (Note that* $\mathcal{C}$ *is a function of both* s *and* κ*, and by convention,* $s \notin \mathcal{C}$*.)*

3. *Use the modified Lasso program (5.22) with parameter* λ_n *to perform a linear regression of* X_s *against* $\mathcal{C}_d := \Psi(X; V \cup pow(\mathcal{C}(s;d)))\backslash\{X_s\}$*, the set of all products of subsets of variables* $\{X_c : c \in \mathcal{C}\}$ *of size at most* d*, together with singleton variables. Denote the solution by* $\widehat{\beta}$*.*

4. *Threshold the entries of* $\widehat{\beta}$ *at level* τ_n*, and define the estimated neighborhood set* $\widehat{N(s)}$ *as the support of the thresholded vector.*

Note that Algorithm 5.3 is a version of Algorithm 5.4 with $\mathcal{C} = V\backslash s$, indicating the absence of a prescreening step. Hence, the statistical consistency result below applies easily to Algorithm 5.3 for graphs with no correlation decay.

For fully-observed i.i.d. observations, we choose $(\widehat{\Gamma}, \widehat{\gamma})$ according to

$$(\widehat{\Gamma}, \widehat{\gamma}) = \left(\frac{X_{\mathcal{C}}^T X_{\mathcal{C}}}{n} - \overline{x_{\mathcal{C}}}\,\overline{x_{\mathcal{C}}}^T, \; \frac{X_{\mathcal{C}}^T X_s}{n} - \overline{x}_s \overline{x_{\mathcal{C}}} \right), \tag{5.26}$$

and parameters (λ_n, τ_n) as follows: For a candidate set $\mathcal{C}$, let $x_{\mathcal{C},i} \in \{0,1\}^{|\mathcal{C}_d|}$ denote the augmented vector corresponding to the observation x_i, and define $\Sigma_{\mathcal{C}} := \mathrm{Cov}(x_{\mathcal{C},i}, x_{\mathcal{C},i})$. Let $\widetilde{\beta} := \Sigma_{\mathcal{C}}^{-1} \mathrm{Cov}(x_{\mathcal{C},i}, x_{s,i})$. Then set

$$\lambda_n \gtrsim \varphi \|\widetilde{\beta}\|_2 \sqrt{\frac{\log |\mathcal{C}_d|}{n}}, \qquad \tau_n \asymp \varphi \|\widetilde{\beta}\|_2 \sqrt{\frac{\log |\mathcal{C}_d|}{n}}, \tag{5.27}$$

where φ is some function such that $\langle x_{\mathcal{C},i}, u \rangle$ is sub-Gaussian with parameter $\varphi^2 \|u\|_2^2$ for any $(2^d - 1)$-sparse vector u, and φ does not depend on u. We have the following consistency result, the analog of Proposition 5.1 for the augmented set of vectors. It applies to Algorithm 5.4 with the pairs $(\widehat{\Gamma}, \widehat{\gamma})$ and (λ_n, τ_n) chosen as in equations (5.26) and (5.27).

Corollary 5.5. *Consider i.i.d. observations* $\{x_i\}_{i=1}^n$ *generated from an Ising model satisfying the correlation decay condition (5.25), and suppose*

$$n \gtrsim \left(\kappa^2 + \varphi^2 \max\left\{ \frac{1}{\lambda_{\min}(\Sigma_{\mathcal{C}})}, \left\| \Sigma_{\mathcal{C}}^{-1} \right\|_\infty^2 \right\} 2^{2d} \right) \log |\mathcal{C}_d|. \tag{5.28}$$

Then there are universal constants (c, c', c'') *such that with probability at least* $1 - c\exp(-c' \log p)$*, and for any* $s \in V$*:*

(i) The set $\mathcal{C}$ *from step (2) of Algorithm 5.4 satisfies* $|\mathcal{C}| \leq d^{\frac{\log(4/\kappa)}{\zeta}}$.

(ii) Algorithm 5.4 recovers all neighbors $t \in N(s)$ *such that*

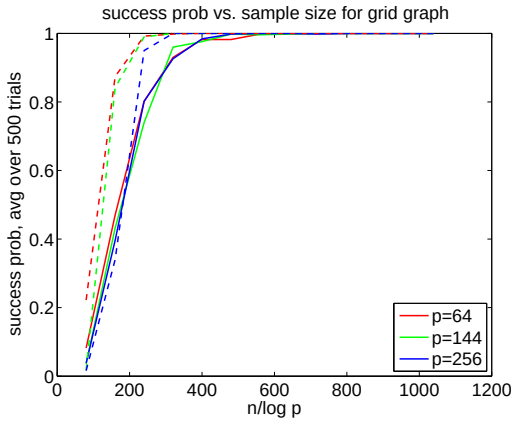
$$|\widetilde{\beta}_t| \geq c'' \varphi \|\widetilde{\beta}\|_2 \sqrt{\frac{\log |\mathcal{C}_d|}{n}}.$$

The proof of Corollary 5.5 is contained in Appendix C.5.2. Due to the exponential factor 2^d appearing in the lower bound (5.28) on the sample size, this method is suitable only for bounded-degree graphs. However, for reasonable sizes of d, the dimension of the linear regression problem decreases from $\mathcal{O}(p^d)$ to $|\mathcal{C}_d| = \mathcal{O}(|\mathcal{C}|^d) = \mathcal{O}\left(d^{\frac{d\log(4/\kappa)}{\zeta}}\right)$, which has a significant impact on the runtime of the algorithm. We explore two classes of bounded-degree graphs with correlation decay in the simulations of Section 5.4.4, where we generate Erdös-Renyi graphs with edge probability c/p and square grid graphs in order to test the behavior of our recovery algorithm on non-trees. When $m > 2$, corresponding to non-binary states, we may combine these ideas with the overlapping group Lasso [42] to obtain similar algorithms for nodewise recovery of non-tree graphs. However, the details are more complicated, and we do not include them here. Note that our method for nodewise recovery in non-tree graphical models are again easily adapted to handle noisy and missing data, which is a clear advantage over other existing methods.

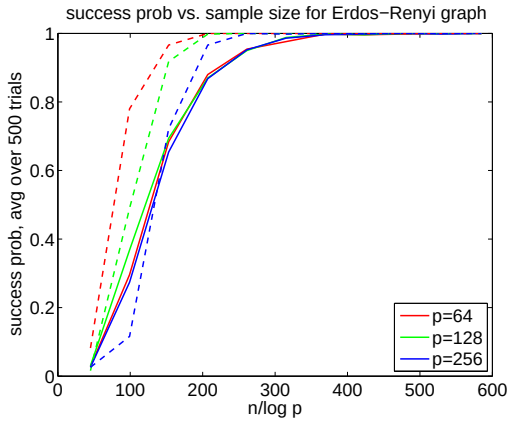
5.4.4 Simulations

In this section, we report the results of various simulations we performed to illustrate the sharpness of our theoretical claims. In all cases, we generated data from binary Ising models. We first applied the nodewise linear regression method (Algorithm 5.2 for trees; Algorithm 5.3 in the general case) to the method of ℓ_1-regularized logistic regression, analyzed in past work for Ising model selection by Ravikumar et al. [73]. Their main result was to establish that, under certain incoherence conditions of the Fisher information matrix, performing ℓ_1-regularized logistic regression with a sample size $n \gtrsim d^3 \log p$ is guaranteed to select the correct graph w.h.p. Thus, for any bounded-degree graph, the sample size n need grow only logarithmically in the number of nodes p. Under this scaling, our theory also guarantees that nodewise *linear regression* with ℓ_1-regularization will succeed in recovering the true graph w.h.p.

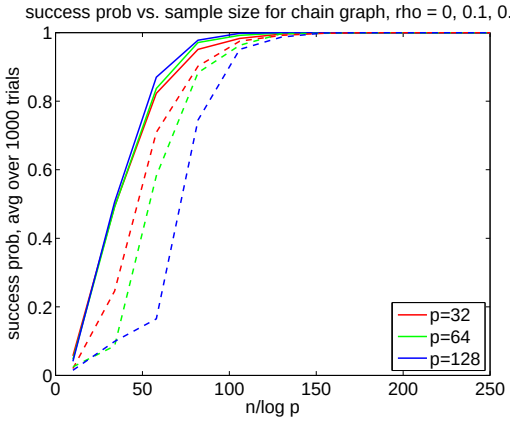
In Figure 5.3, we present the results of simulations with two goals: (i) test the scaling $n \approx \log p$ of the required sample size; and (ii) compare ℓ_1-regularized nodewise linear regression (Algorithms 5.3 and 5.4) to ℓ_1-regularized nodewise logistic regression [73]. We ran simulations for the two methods on both tree-structured and non-tree graphs with data generated from a binary Ising model, with node weights $\theta_s = 0.1$ and edge weights $\theta_{st} = 0.3$. To save on computation, we employed the neighborhood screening method described in Section 5.4.3 to prune the candidate neighborhood set before performing linear regression. We selected a candidate neighborhood set of size $\lfloor 2.5d \rfloor$ with highest empirical correlations, then performed a single regression against all singleton nodes and products of subsets of the candidate neighborhood set of size at most d, via the modified Lasso program (5.22). The size of the candidate neighborhood set was tuned through repeated runs of the algorithm. For both methods, the optimal choice of regularization parameter λ_n scales as $\sqrt{\frac{\log p}{n}}$, and we used the same value of λ_n in comparing logistic to linear regression. In each panel, we plot the probability of successful graph recovery versus the rescaled sample size $\frac{n}{\log p}$, with curves of



(a) Grid graph



(b) Erdös-Renyi graph



(c) Chain graph

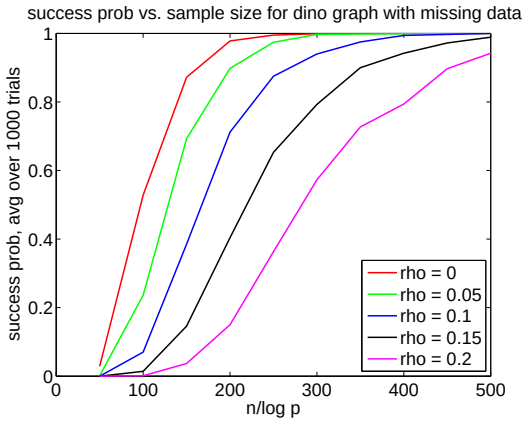
Figure 5.3: Comparison between ℓ_1-regularized logistic vs. linear regression methods for graph recovery. Each panel plots of the probability of correct graph recovery vs. the rescaled sample size $n/\log p$; solid curves correspond to linear regression (method in this chapter), whereas dotted curves correspond to logistic regression [73]. Curves are based on average performance over 500 trials. (a) Simulation results for two-dimensional grids with $d = 4$ neighbors, and number of nodes p varying over $\{64, 144, 256\}$. Consistent with theory, when plotted vs. the rescaled sample size $n/\log p$, all three curves (red, blue, green) are well-aligned with one another. Both linear and logistic regression transition from failure to success at a similar point. (b) Analogous results for an Erdös-Renyi graph with edge probability $3/p$. (c) Analogous results for a chain-structured graph with maximum degree $d = 2$.

different colors corresponding to graphs (from the same family) of different sizes. Solid lines correspond to linear regression, whereas dotted lines correspond to logistic regression; panels (a), (b), and (c) correspond to grid graphs, Erdös-Renyi random graphs, and chain graphs, respectively. For all these graphs, the three solid/dotted curves for different problem sizes

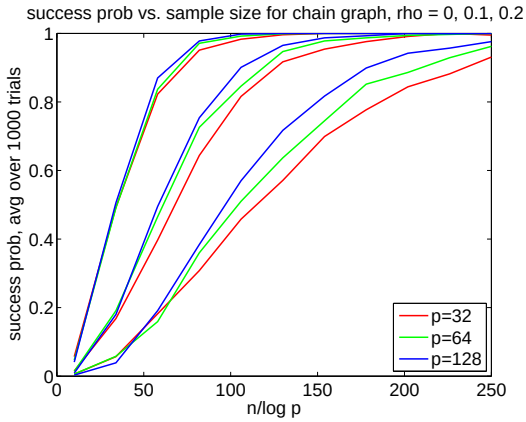
are well-aligned, showing that the method undergoes a transition from failure to success as a function of the ratio $\frac{n}{\log p}$. In addition, both linear and logistic regression are comparable in terms of statistical efficiency (the number of samples n required for correct graph selection to be achieved).

The main advantage of nodewise linear regression and the graphical Lasso over nodewise logistic regression is that they are straightforward to correct for corrupted or missing data. Figure 5.4 shows the results of simulations designed to test the behavior of these corrected estimators in the presence of missing data. Panel (a) shows the results of applying the graphical Lasso method, as described in Section 5.4.1, to the dino graph of Figure 5.2(e). We again generated data from an Ising model with node weights 0.1 and edge weights 0.3. The curves show the probability of success in recovering the 15 edges of the graph, as a function of the rescaled sample size $\frac{n}{\log p}$ for $p = 13$. In addition, we performed simulations for different levels of missing data, specified by the parameter $\alpha \in \{0, 0.05, 0.1, 0.15, 0.2\}$, using the corrected estimator (5.21). Note that all five runs display a transition from success probability 0 to success probability 1 in roughly the same range, as predicted by our theory. Indeed, since the dinosaur graph has only singleton separators, Corollary 5.2 ensures that the inverse covariance matrix is exactly graph-structured, so our global recovery method is consistent at the population level. Further note that the curves shift right as the fraction α of missing data increases, since the recovery problem becomes incrementally harder.

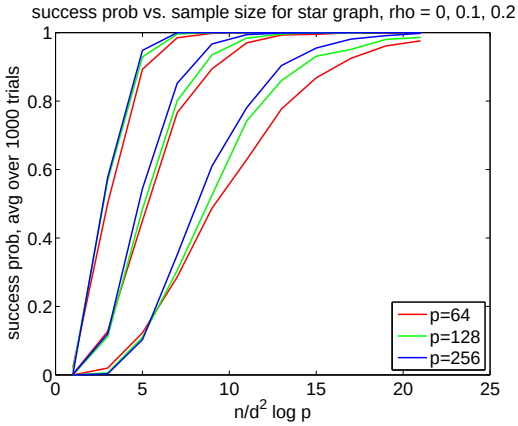
Panels (b) and (c) of Figure 5.4 show the results of the nodewise regression method of Section 5.4.2 applied to chain and star graphs, with increasing numbers of nodes $p \in \{32, 64, 128\}$ and $p \in \{64, 128, 256\}$, respectively. For the chain graphs in panel (b), we set node weights of the Ising model equal to 0.1 and edge weights equal to 0.3. For the varying-degree star graph in panel (c), we set node weights equal to 0.1 and edge weights equal to $\frac{1.2}{d}$, where the degree d of the central hub grows with the size of the graph as $\lfloor \log p \rfloor$. Again, we show curves for different levels of missing data, $\alpha \in \{0, 0.1, 0.2\}$. The modified Lasso program (5.22) was optimized using a form of composite gradient descent due to Agarwal et al. [1], guaranteed to converge to a small neighborhood of the optimum even when the problem is nonconvex (cf. Chapter 3). In both the chain and star graphs, the three curves corresponding to different problem sizes p at each value of the missing data parameter α stack up when plotted against the rescaled sample size. Note that the curves for the star graph stack up nicely with the scaling $\frac{n}{d^2 \log p}$, rather than the worst-case scaling $n \asymp d^3 \log p$, corroborating the remark following Proposition 5.1. Since $d = 2$ is fixed for the chain graph, we use the rescaled sample size $\frac{n}{\log p}$ in our plots, as in the plots in Figure 5.3. Once again, these simulations corroborate our theoretical predictions: the corrected linear regression estimator remains consistent even in the presence of missing data, although the sample size required for consistency grows as the fraction of missing data α increases.



(a) Dino graph with missing data



(b) Chain graph with missing data



(c) Star graph with missing data, $d \approx \log p$

Figure 5.4: Simulation results for global and nodewise recovery methods on binary Ising models, allowing for missing data in the observations. Each point represents an average over 1000 trials. Panel (a) shows simulation results for the graphical Lasso method applied to the dinosaur graph with the fraction α of missing data varying in $\{0, 0.05, 0.1, 0.15, 0.2\}$. Panel (b) shows simulation results for nodewise regression applied to chain graphs for varying p and α. Panel (c) shows simulation results for nodewise regression applied to star graphs with maximal node degree $d = \log p$ and varying α.

5.5 Discussion

The correspondence between the inverse covariance matrix and graph structure of a Gauss-Markov random field is a classical fact with numerous consequences for estimation of Gaussian graphical models. It has been an open question as to whether similar properties extend to a broader class of graphical models. In this chapter, we have provided a partial affirmative answer to this question and developed theoretical results extending such relationships

to discrete undirected graphical models.

As shown by our results, the inverse of the ordinary covariance matrix is graph-structured for special subclasses of graphs with singleton separator sets. More generally, we have considered inverses of *generalized covariance matrices*, formed by introducing indicator functions for larger subsets of variables. When these subsets are chosen to reflect the structure of an underlying junction tree, the edge structure is reflected in the inverse covariance matrix. Our population-level results have a number of statistical consequences for graphical model selection. We have shown that our results may be used to establish consistency (or inconsistency) of standard methods for discrete graph selection, and have proposed new methods for neighborhood recovery which, unlike existing methods, may be applied even when observations are systematically corrupted by mechanisms such as additive noise and missing data. Furthermore, our methods are attractive in their simplicity, in that they only involve simple optimization problems.

Although the methods considered in our chapter are limited to structure estimation in undirected graphs, recent work [54] shows that connections exist between inverse covariance matrices and graph structure for directed graphs, as well, provided the underlying distribution follows a linear structural equation model. Whereas the problem of learning existence and orientation of edges in a directed graph is reasonably tractable given a topological ordering of the nodes, inferring a topological order based on joint observations of the variables appears to be a very difficult problem, and current state-of-the-art approaches involve expensive searches based on conditional independence tests that scale exponentially with the size of the network. Motivated by our work on inverse covariance matrices in undirected graphs and the fact that estimation of inverse covariances is far more tractable than exponential search, it would be fruitful to explore whether similar connections could be leveraged in much broader settings than linear causal networks.

Chapter 6

Application to MRI

6.1 Introduction

Compressed sensing MRI is a relatively new technique used to reconstruct an image from undersampled k-space data [15, 57]. From a statistical perspective, one cannot expect to reconstruct an image of p pixels with fewer than p measurements. However, if the image possesses a sparse representation with respect to some basis (e.g., wavelet or Fourier basis), existing theory predicts that the image may be reconstructed with $n = \mathcal{O}(k \log p)$ measurements, where k is the number of non-zeros in the sparse representation, provided the design matrix is sufficiently incoherent [16]. More precisely, compressed sensing MRI admits the following mathematical formulation:

$$\min \|\Psi m\|_1 \qquad \text{s.t.} \qquad \|\mathcal{F}_u m - y\|_2 < \epsilon, \tag{6.1}$$

where $m \in \mathbb{C}^p$ is the image vector, $\Psi \in \mathbb{C}^{p \times p}$ is the sparsifying transform, $\mathcal{F}_u \in \mathbb{C}^{n \times p}$ is the undersampled 2DFT matrix, and $y \in \mathbb{C}^n$ is the vector of k-space measurements [15].

Current theoretical results establish that when the effective design matrix $\mathcal{F}_u \Psi^*$ is constructed by sampling points in 2D k-space uniformly at random, the desired incoherence property holds w.h.p., so the program (6.1) results in exact recovery [17, 86]. However, such random sampling schemes are infeasible in practice, due to physical limitations of the MRI scanner. As a result, alternate methods for undersampling k-space have been developed, including random subsampling of phase encode and/or readout directions [57, 92] and altering the B1 field through random receiver coil sensitivities [82]. These methods have been shown to perform well on synthetic and real MRI data. A final sampling method for compressed sensing MRI involves designing random spatially-selective RF pulses so the design matrix of the resulting optimization problem has an i.i.d. Gaussian distribution [34]. This method is fascinating, because many theoretical results in the statistical literature assume the design matrix is i.i.d. Gaussian, or at least real.

In this chapter, we explore the challenges presented by corrupted samples of k-space data (perhaps introduced by machine miscalibration or gradient imperfections). We develop

a novel estimator for reconstructing an MRI image based on corrupted k-space samples, and prove theoretical results showing that $n = \mathcal{O}(k \log p)$ measurements are still sufficient for recovery in the noisy setting. In addition, we verify the theoretical predictions through simulations with synthetic data.

6.2 Problem setup

Suppose we have a linear regression model

$$y_i = \langle x_i, \beta^* \rangle + \epsilon_i, \qquad i = 1, \ldots, n, \tag{6.2}$$

where $\beta^* \in \mathbb{C}^p$ is the unknown signal, $x_i \in \mathbb{C}^p$ are the sensing directions, $y_i \in \mathbb{C}$ are the corresponding measurements, and $\epsilon_i \in \mathbb{C}^p$ is i.i.d. sub-Gaussian noise. Based on observation pairs $\{(x_i, y_i)\}_{i=1}^n$, the usual goal is to recover a k-sparse vector β^* when $n \ll p$.

We will adopt a slightly different setup, where we again assume a linear model (6.2), but we wish to perform inference based on observation pairs $\{(z_i, y_i)\}_{i=1}^n$, where the z_i's are controlled by the experimenter and x_i is a noisy version of z_i.

More concretely, in the framework of Fourier analysis, we assume the z_i's are Fourier sensing vectors taken at frequency $\omega_i \in [0, 2\pi)$. Then $z_{ij} = e^{(j-1)\omega_i \mathbf{i}}$. Letting ξ_i denote the frequency perturbations due to measurement error, with $\xi_i \perp\!\!\!\perp \omega_i$, and writing $\psi_i = \omega_i + \xi_i$, we have $x_{ij} = e^{(j-1)\psi_i \mathbf{i}}$. We may write $x_i = z_i \odot u_i$, where $u_{ij} = e^{(j-1)\xi_i \mathbf{i}}$.

In compressed sensing MRI, the frequencies ω_i are chosen randomly from a predetermined distribution. In the discrete case, ω_i is drawn uniformly at random from the set $\{0, \frac{2\pi}{p}, \ldots, \frac{2\pi(p-1)}{p}\}$. In the continuous case, ω_i is a uniform variable in the interval $[0, 2\pi)$. We will focus on the latter case. As established in Section 4.1 in Rauhut [72], we have $\mathbb{E}(z_i \bar{z}_i^T) = I$ in both the discrete and continuous cases. Since $\mathbb{E}(x_i \bar{x}_i^T) = \mathbb{E}(z_i \bar{z}_i^T) \odot \mathbb{E}(u_i \bar{u}_i^T)$ and the diagonals of $\mathbb{E}(u_i \bar{u}_i^T)$ are clearly all 1's, it follows that $\mathbb{E}(x_i \bar{x}_i^T) = I$ whenever $\mathbb{E}(z_i \bar{z}_i^T) = I$.

6.3 Derivation of objective

In this section, we show how to derive a new compressed sensing objective that is applicable when the design matrix is corrupted. Our development parallels the analysis of Section 3.2.2, except we need to tweak the expressions slightly to accommodate complex-valued vectors.

Note that

$$(x_i^T \bar{\beta} - x_i^T \bar{\beta}^*)(\bar{x}_i^T \beta - \bar{x}_i^T \beta^*) \geq 0.$$

Expanding and taking expectations of both sides, we obtain the inequality

$$\bar{\beta}^T \mathbb{E}(x_i \bar{x}_i^T) \beta - \big(\langle \mathbb{E}(x_i^T \bar{x}_i) \beta^*, \beta \rangle + \langle \mathbb{E}(\bar{x}_i^T x_i) \bar{\beta}^*, \bar{\beta} \rangle \big) \geq \bar{\beta}^{*T} \mathbb{E}(x_i \bar{x}_i^T) \beta^*.$$

Hence, if $R \geq \|\beta^*\|_1$, we have

$$\beta^* \in \arg \min_{\|\beta\|_1 \leq R} \{ \bar{\beta}^T \mathbb{E}(x_i \bar{x}_i^T) \beta - (\langle \mathbb{E}(x_i \bar{x}_i^T) \beta^*, \beta \rangle + \langle \mathbb{E}(\bar{x}_i x_i^T) \bar{\beta}^*, \bar{\beta} \rangle) \},$$

where the ℓ_1-norm encourages sparsity in β. Hence, we formulate the constrained quadratic program

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \{\widehat{\beta}^T \mathbb{E}(x_i \overline{x}_i^T)\beta - (\langle \widehat{\eta},\, \beta\rangle + \overline{\langle \widehat{\eta},\, \beta\rangle})\}, \tag{6.3}$$

where $\widehat{\eta}$ is a surrogate for $\mathbb{E}(x_i \overline{x}_i^T)\beta^*$ based on corrupted observations $\{(z_i, y_i)\}_{i=1}^n$.

We also have the following Lagrangian version:

$$\widehat{\beta} \in \arg\min_{\beta} \{\bar{\beta}^{*T} \mathbb{E}(x_i \overline{x}_i^T)\beta - (\langle \widehat{\eta},\, \beta\rangle + \overline{\langle \widehat{\eta},\, \beta\rangle}) + \lambda\|\beta\|_1\}. \tag{6.4}$$

In order to find an appropriate choice for $\widehat{\eta}$, note that

$$\mathbb{E}(\bar{y}_i z_i) = \mathbb{E}(z_i \overline{x}_i^T \beta^*) = \mathbb{E}(z_i(\bar{z}_i \odot \bar{u}_i)^T \beta^*) = \mathbb{E}(z_i \bar{z}_i^T \operatorname{diag}(\bar{u}_i)\beta^*) = \mathbb{E}(z_i \bar{z}_i^T) \operatorname{diag}(\mathbb{E}(\bar{u}_i))\beta^*,$$

where $\operatorname{diag}(v)$ is the $p \times p$ diagonal matrix with entries equal to $v \in \mathbb{C}^p$. Hence, assuming $\mathbb{E}(\bar{u}_i) \neq 0$, we use

$$\widehat{\eta} = \mathbb{E}(x_i \overline{x}_i^T) \operatorname{diag}^{-1}(\mathbb{E}(\bar{u}_i))(\mathbb{E}(z_i \bar{z}_i^T))^{-1} \frac{Z^T \bar{y}}{n},$$

which is an unbiased estimator for $\mathbb{E}(x_i \overline{x}_i^T)\beta^*$. Note that in practice, $\mathbb{E}(z_i \bar{z}_i^T)$ is known by design, and $\mathbb{E}(\bar{u}_i)$ and $\mathbb{E}(x_i \overline{x}_i^T) = \mathbb{E}(z_i \bar{z}_i^T) \odot \mathbb{E}(u_i \bar{u}_i^T)$ may be calculated based on the known distribution of the u_i's.

6.4 Theoretical contributions

In this section, we present theoretical guarantees concerning the consistency of the estimator $\widehat{\beta}$ arising from equations (6.3) and (6.4).

6.4.1 Statistical error

We have a following main result, the analog of Theorem 3.1. We write $\alpha_\ell := \lambda_{\min}(\mathbb{E}(x_i \overline{x}_i^T))$ and $\alpha_u := \lambda_{\max}(\mathbb{E}(x_i \overline{x}_i^T))$.

Theorem 6.1. *Suppose the frequencies ω_i are either drawn uniformly at random from the discrete set $\left\{0, \frac{2\pi}{p}, \ldots, \frac{2\pi(p-1)}{p}\right\}$, or uniformly at random from the continuous interval $[0, 2\pi)$. Suppose the error ϵ_i is i.i.d. sub-Gaussian with parameter σ_ϵ. Also suppose $n \gtrsim k \log p$. When $R = \|\beta^*\|_1$, the solution $\widehat{\beta}$ to the program (6.3) satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \leq \frac{c\alpha_u + c'\sigma_\epsilon}{\alpha_\ell} \sqrt{\frac{k \log p}{n}} \cdot \left\| \mathbb{E}(x_i \overline{x}_i^T) \operatorname{diag}^{-1}(\mathbb{E}(\bar{u}_i))(\mathbb{E}(z_i \bar{z}_i^T))^{-1} \right\|_1, \tag{6.5}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. When

$$\lambda \geq (c\alpha_u + c'\sigma_\epsilon) \sqrt{\frac{\log p}{n}} \cdot \left\| \mathbb{E}(x_i \overline{x}_i^T) \operatorname{diag}^{-1}(\mathbb{E}(\bar{u}_i))(\mathbb{E}(z_i \bar{z}_i^T))^{-1} \right\|_1,$$

the solution $\widehat{\beta}$ *to the program* (6.4) *satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \leq \frac{c\lambda\sqrt{k}}{\alpha_\ell},$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$.

The proof of Theorem 6.1, which is provided in Section 6.5.1, resembles the proof of Theorem 3.1 in that it proceeds via a basic inequality. However, the more technical steps involve matrix concentration results for complex-valued matrices, which draw upon results from Kunis and Rauhut [45]. Similar bounds on the ℓ_1-error follow easily from the ℓ_2-bounds and the cone condition.

6.4.2 Optimization

Now consider the case when $\mathbb{E}(x_i \overline{x}_i^T) = I$ (which occurs when the frequencies ω_i of the z_i's are chosen uniformly on $[0, 2\pi)$). Then the Lagrangian program (6.4) simplifies to

$$\widehat{\beta} \in \arg\min_{\beta} \{\|\beta\|_2^2 - (\langle \widehat{\eta},\, \beta\rangle + \overline{\langle \widehat{\eta},\, \beta\rangle}) + \lambda\|\beta\|_1\}. \tag{6.6}$$

Note that each coordinate of β may be optimized separately, yielding the soft-thresholding solution

$$\widehat{\beta}_i = \begin{cases} 0 & \text{if} \quad |\widehat{\eta}_i| \leq \lambda \\ \frac{|\widehat{\eta}_i| - \lambda}{|\widehat{\eta}_i|}\widehat{\eta}_i & \text{if} \quad |\widehat{\eta}_i| > \lambda \end{cases} \tag{6.7}$$

(cf. [94]). One may check that when $\widehat{\eta}_i \in \mathbb{R}$, the soft-thresholding operator reduces to the usual definition of the operator,

$$\text{SoftThresh}(\widehat{\eta}_i) = \begin{cases} 0 & \text{if} \quad |\widehat{\eta}_i| \leq \lambda \\ \widehat{\eta}_i - \lambda & \text{if} \quad \widehat{\eta}_i > \lambda \\ \widehat{\eta}_i + \lambda & \text{if} \quad \widehat{\eta}_i < -\lambda. \end{cases}$$

In fact, the soft-thresholding operator simply soft-thresholds the amplitude of a complex number while keeping the same phase.

When $\mathbb{E}(x_i \overline{x}_i^T) \neq I$, we may use iterative methods such as projected or composite gradient descent to obtain $\widehat{\beta}$.

6.4.3 Special case: Identity covariance

As noted in the previous section, the case when $\mathbb{E}(z_i \overline{z}_i^T) = I$ lends itself to a nice soft-thresholding solution to the Lagrangian program (6.4). In this case, $\mathbb{E}(x_i \overline{x}_i^T) = I$, and we have

$$\widehat{\eta} = \text{diag}^{-1}(\mathbb{E}(\overline{u}_i))\frac{Z^T \overline{y}}{n},$$

leading to the error bound

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k},$$

w.h.p., when $\lambda \geq \frac{c+c'\sigma_\epsilon}{\min_j |\mathbb{E}(\overline{u}_{ij})|}\sqrt{\frac{\log p}{n}}$.

In particular, we consider two special forms of the noise frequency ψ_i:

(i) $\psi_i \in (-\delta, \delta)$ is chosen uniformly at random.

(ii) $\psi_i \sim N(0, \sigma^2)$.

Since ψ_i has a symmetric distribution in both cases, $\mathbb{E}(\overline{u}_i)$ is real-valued, with component j equal to the characteristic function of ψ_i evaluated at $(j-1)$. For the two cases, we then have

(i) $\mathbb{E}(\overline{u}_{ij}) = \frac{\sin((j-1)\delta)}{\delta(j-1)}$

(ii) $\mathbb{E}(\overline{u}_{ij}) = \exp\left(-\frac{(j-1)^2\sigma^2}{2}\right)$.

We can see the tradeoff due to noise corruptions in the vectors in the magnitude of $\min_j |\mathbb{E}(\overline{u}_{ij})|$, which in the first case is on the order of $\frac{1}{\delta p}$, and in the second case is on the order of $\exp(-p^2\sigma^2/2)$.

6.4.4 Sparsity in another basis

In compressed sensing MRI, it is beneficial to consider sparsity of the image with respect to another basis. Hence, we form the alternative convex program

$$\widehat{\beta} \in \arg\min_{\|\Psi\beta\|_1 \leq R} \{\overline{\beta}^T \mathbb{E}(x_i \overline{x}_i^T)\beta - (\langle\widehat{\eta}, \beta\rangle + \overline{\langle\widehat{\eta}, \beta\rangle}\}, \tag{6.8}$$

where $\Psi \in \mathbb{C}^{p\times p}$ is an orthonormal change-of-basis matrix such that $\Psi\beta^*$ is sparse in the new coordinates.

We then have the following theorem, with an analogous result for the Lagrangian variant of the program (6.8):

Theorem 6.2. *Under the same conditions as Theorem 6.1, the solution $\widehat{\beta}$ to the program (6.8):*

$$\|\widehat{\beta} - \beta^*\|_2 \leq \frac{c\alpha_u + c'\sigma_\epsilon}{\lambda_{\min}(\mathbb{E}(x_i\overline{x}_i^T))} \frac{|||\Psi|||_1}{\min_j |\mathbb{E}(u_{ij})|}\sqrt{\frac{k\log p}{n}},$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

The proof of Theorem 6.2 is contained in Section 6.5.2.

6.5 Proofs

In this section, we include the more technical proofs of our main results.

6.5.1 Proof of Theorem 6.1

We begin by analyzing the program (6.3) with $R = \|\beta^*\|_1$. Since $\widehat{\beta}$ is optimal and β^* is feasible, we may form the usual basic inequality

$$\bar{\widehat{\beta}}^T \mathbb{E}(x_i \bar{x}_i^T)\widehat{\beta} - \langle \widehat{\eta},\, \widehat{\beta}\rangle - \overline{\langle \widehat{\eta},\, \widehat{\beta}\rangle} \leq \bar{\beta}^{*T}\mathbb{E}(x_i\bar{x}_i^T)\beta^* - \langle \widehat{\eta},\, \beta^*\rangle - \overline{\langle \widehat{\eta},\, \beta^*\rangle},$$

which (following some algebra and denoting $\widehat{\nu} = \widehat{\beta} - \beta^*$) is equivalent to

$$\bar{\widehat{\nu}}^T \mathbb{E}(x_i\bar{x}_i^T)\widehat{\nu} \leq \langle \widehat{\eta} - \mathbb{E}(x_i\bar{x}_i^T)\beta^*,\, \widehat{\nu}\rangle + \overline{\langle \widehat{\eta} - \mathbb{E}(x_i\bar{x}_i^T)\beta^*,\, \widehat{\nu}\rangle} = 2\Re(\langle \widehat{\eta} - \mathbb{E}(x_i\bar{x}_i^T)\beta^*,\, \widehat{\nu}\rangle).$$

Hence,

$$\frac{\lambda_{\min}(\mathbb{E}(x_i\bar{x}_i^T))}{2}\|\widehat{\nu}\|_2^2 \leq \|\widehat{\nu}\|_1 \left\|\widehat{\eta} - \mathbb{E}(x_i\bar{x}_i^T)\beta^*\right\|_\infty.$$

Furthermore, it follows via standard arguments that $\|\widehat{\nu}\|_1 \leq 2\sqrt{k}\|\widehat{\nu}\|_2$. Hence, we conclude that

$$\|\widehat{\nu}\|_2 \leq \frac{4\sqrt{k}}{\alpha_\ell}\left\|\widehat{\eta} - \mathbb{E}(x_i\bar{x}_i^T)\beta^*\right\|_\infty, \tag{6.9}$$

where the latter term in the product measures the accuracy of the estimator $\widehat{\eta}$.

We now write

$$\begin{aligned}\left\|\widehat{\eta} - \mathbb{E}(x_i\bar{x}_i^T)\beta^*\right\|_\infty &= \left\|\mathbb{E}(x_i\bar{x}_i^T)\,\mathrm{diag}^{-1}(\mathbb{E}(\bar{u}_i))(\mathbb{E}(z_i\bar{z}_i^T))^{-1}\left(\frac{Z^T\bar{y}}{n} - \mathbb{E}(\bar{y}_i z_i)\right)\right\|_\infty \\ &\leq \left|\!\left|\!\left|\mathbb{E}(x_i\bar{x}_i^T)\,\mathrm{diag}^{-1}(\mathbb{E}(\bar{u}_i))(\mathbb{E}(z_i\bar{z}_i^T))^{-1}\right|\!\right|\!\right|_1 \left\|\frac{Z^T\bar{y}}{n} - \mathbb{E}(\bar{y}_i z_i)\right\|_\infty, \end{aligned} \tag{6.10}$$

where the ℓ_∞-norm is the max modulus of the coordinates of a complex vector, and the ℓ_1-operator norm is the max absolute column sum of a complex matrix. Furthermore, we have

$$\begin{aligned}\left\|\frac{Z^T\bar{y}}{n} - \mathbb{E}(\bar{y}_i z_i)\right\|_\infty &= \left\|\frac{Z^T(\overline{X}\beta^* + \bar{\epsilon})}{n} - \mathbb{E}(z_i\bar{x}_i^T)\beta^*\right\|_\infty \\ &\leq \left\|\left(\frac{Z^T\overline{X}}{n} - \mathbb{E}(z_i\bar{x}_i^T)\right)\beta^*\right\|_\infty + \left\|\frac{Z^T\bar{\epsilon}}{n}\right\|_\infty. \end{aligned} \tag{6.11}$$

To bound the first term on the RHS, we have the following lemma, using ideas from Lemma 3.2 in Kunis and Rauhut [45]:

Lemma 6.1. *Let $v \in \mathbb{C}^p$. For $t > 0$, and for each $1 \le \ell \le p$, we have*

$$\mathbb{P}\left(\left|e_\ell^T\left(\frac{Z^T\overline{X}}{n} - \mathbb{E}(z_i\overline{x}_i^T)\right)v\right| \ge t\right) \le 4\exp\left(\frac{-nt^2}{4\alpha_u\|v\|_2^2 + 8\|v\|_1 t/3\sqrt{2}}\right).$$

Proof. Recall the canonical Bernstein inequality (cf. Theorem 3.1 in Kunis and Rauhut [45]):

Lemma 6.2. *Let Y_i be independent real-valued random variables with $\mathbb{E}(Y_i) = 0$, $\mathbb{E}(Y_i^2) \le b$, and $|Y_i| \le B$ for each i. Then*

$$\mathbb{P}\left(\left|\sum_{i=1}^n Y_i\right| \ge t\right) \le 2\exp\left(-\frac{1}{2}\frac{t^2}{nb + Bt/3}\right).$$

We write

$$e_\ell^T Z^T\overline{X}v - ne_\ell^T\mathbb{E}(z_i\overline{x}_i^T)v = \sum_{j=1}^p\sum_{i=1}^n v_j e^{\ell\omega_i\mathbf{i} - j\psi_i\mathbf{i}} - ne_\ell^T\mathbb{E}(z_i\overline{x}_i^T)v = \sum_{i=1}^n(\widetilde{Y}_i - \mathbb{E}(\widetilde{Y}_i)),$$

where $\widetilde{Y}_i := \sum_{j=1}^p v_j e^{(\ell\omega_i - j\psi_i)\mathbf{i}}$. Let $Y_i := \widetilde{Y}_i - \mathbb{E}(\widetilde{Y}_i)$. Clearly, $\mathbb{E}(Y_i) = 0$ and

$$|\widetilde{Y}_i| \le \sum_{j=1}^p |v_j| = \|v\|_1,$$

so $|Y_i| \le 2\|v\|_1$. Furthermore,

$$\mathbb{E}(|Y_i|^2) = \mathbb{E}(|\widetilde{Y}_i|^2) - |\mathbb{E}(\widetilde{Y}_i)|^2 \le \mathbb{E}(|\widetilde{Y}_i|^2) = \mathbb{E}\left(\sum_{j,j'=1}^p v_j\overline{v}_{j'}e^{(j'-j)\psi_i\mathbf{i}}\right) = v^T\mathbb{E}(x_i\overline{x}_i^T)\overline{v} \le \alpha_u\|v\|_2^2.$$

Note that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i\right| \ge t\right) \le \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \Re(Y_i)\right| \ge \frac{t}{\sqrt{2}}\right) + \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \Im(Y_i)\right| \ge \frac{t}{\sqrt{2}}\right).$$

Hence, applying Lemma 6.2 (with $b = 2\|v\|_1$, $B = \alpha_u\|v\|_2^2$, $t = t/\sqrt{2}$) to $\Re(Y_i)$ and $\Im(Y_i)$, and using the fact that $|\Re(Y_i)|, |\Im(Y_i)| \le |Y_i|$, we obtain the desired result. □

Returning to inequality (6.10), note that

$$\left\|\left(\frac{Z^T\overline{X}}{n} - \mathbb{E}(z_i\overline{x}_i^T)\right)\beta^*\right\|_\infty = \max_\ell\left|e_\ell^T\left(\frac{Z^T\overline{X}}{n} - \mathbb{E}(z_i\overline{x}_i^T)\right)\beta^*\right|.$$

Applying Lemma 6.1, together with a union bound, we conclude that

$$\mathbb{P}\left(\left\|\left(\frac{Z^T\overline{X}}{n} - \mathbb{E}(z_i\overline{x}_i^T)\right)\beta^*\right\|_\infty \ge t\right) \le 4p\exp\left(\frac{-nt^2}{4\alpha_u\|\beta^*\|_2^2 + 8\|\beta^*\|_1 t/2\sqrt{2}}\right).$$

Let $t = c\alpha_u \|\beta^*\|_2 \sqrt{\frac{\log p}{n}}$. Then for $n \gtrsim k \log p$, we have

$$\|\beta^*\|_1 t \leq 2\sqrt{k}\|\beta^*\|_2 t \leq c'\alpha_u \|\beta^*\|_2^2,$$

so

$$\mathbb{P}\left(\left\| \left(\frac{Z^T \overline{X}}{n} - \mathbb{E}(z_i \overline{x}_i^T) \right) \beta^* \right\|_\infty \geq c\alpha_u \sqrt{\frac{\log p}{n}} \right) \leq 4p \exp(-C \log p) \leq 4 \exp(-C' \log p).$$

Turning to the second term in inequality (6.10), note that for each ℓ, we have

$$e_\ell^T \frac{Z^T \bar{\epsilon}}{n} = \frac{1}{n} \sum_{i=1}^n e^{\ell \omega_i \mathbf{i}} \bar{\epsilon}_i.$$

Assuming ϵ is an i.i.d. sub-Gaussian vector (meaning the real and imaginary parts are separately sub-Gaussian), we see that the above quantity is an i.i.d. average of sub-Gaussians with parameter σ_ϵ^2, since $|e^{\ell \omega_i \mathbf{i}}| \leq 1$. Hence, applying standard arguments and a union bound, we have

$$\mathbb{P}\left(\left\| \frac{Z^T \bar{\epsilon}}{n} \right\|_\infty \geq c\sigma_\epsilon \sqrt{\frac{\log p}{n}} \right) \leq c_1 \exp(-c_2 \log p).$$

Putting everything together and using inequalities (6.9) and (6.10), we arrive at the bound (6.5).

A very similar argument shows that the when $\widehat{\beta}$ optimizes the Lagrangian program (6.4), the corresponding error bound is satisfied.

6.5.2 Proof of Theorem 6.2

Let $\alpha^* = \Psi \beta^*$, and consider the program

$$\widehat{\alpha} \in \arg \min_{\|\alpha\|_1 \leq R} \{ \overline{\alpha}^T \Psi \mathbb{E}(x_i \overline{x}_i^T) \Psi^* \alpha - (\langle \Psi \widehat{\eta}, \alpha \rangle + \overline{\langle \Psi \widehat{\eta}, \alpha \rangle}) \}. \tag{6.12}$$

Clearly, the problems (6.8) and (6.12) are related via $\widehat{\alpha} = \Psi \widehat{\beta}$.

Proceeding as in the proof of Theorem 6.1, and assuming $R = \|\alpha^*\|_1$, we obtain the bound

$$\begin{aligned} \frac{\|\widehat{\nu}\|_2^2}{2} \lambda_{\min}(\Psi \mathbb{E}(x_i \overline{x}_i^T) \Psi^*) &\leq \|\Psi \widehat{\eta} - \Psi \mathbb{E}(x_i \overline{x}_i^T) \Psi^* \alpha^*\|_\infty \|\widehat{\nu}\|_1, \\ &\leq |\!|\!| \Psi |\!|\!|_1 \, \|\widehat{\eta} - \Psi \mathbb{E}(x_i \overline{x}_i^T) \beta^*\|_\infty \|\widehat{\nu}\|_1, \end{aligned}$$

where $\widehat{\nu} = \widehat{\alpha} - \alpha^* = \Psi(\widehat{\beta} - \beta^*)$, so

$$\|\Psi(\widehat{\beta} - \beta^*)\|_2 \leq \frac{c\alpha_u + c'\sigma_\epsilon}{\lambda_{\min}(\Psi \mathbb{E}(x_i \overline{x}_i^T) \Psi^*)} |\!|\!| \Psi |\!|\!|_1 \sqrt{\frac{k \log p}{n}}$$

w.h.p., using the same concentration bound on $\widehat{\eta}$ as before. Finally, noting that

$$\begin{aligned}\|\Psi(\widehat{\beta}-\beta^*)\|_2 &= \sqrt{\lambda_{\min}(\Psi^T\Psi)}\|\widehat{\beta}-\beta^*\|_2,\\ \lambda_{\min}(\Psi\mathbb{E}(x_i\overline{x}_i^T)\Psi^*) &\geq \lambda_{\min}(\mathbb{E}(x_i\overline{x}_i^T))\lambda_{\min}(\Psi\Psi^*)\\ &= \lambda_{\min}(\mathbb{E}(x_i\overline{x}_i^T)),\end{aligned}$$

the desired result follows.

6.6 Simulations

We test our theoretical results empirically through simulations. In order to simplify computations, we generate our data such that $\mathbb{E}(z_i\overline{z}_i^T) = I$, leading to the convex program (6.6) and the soft-thresholding solution (6.7). We compare the output of our algorithm for $\widehat{\eta} = \widehat{\eta}_{\text{corr}}$ and $\widehat{\eta}_{\text{naive}}$, where

$$\begin{aligned}\widehat{\eta}_{\text{corr}} &= \text{diag}^{-1}(\mathbb{E}(\overline{u}_i))\frac{Z^T\overline{y}}{n},\\ \widehat{\eta}_{\text{naive}} &= \frac{Z^T\overline{y}}{n}\end{aligned}$$

are the corrected and uncorrected estimators, respectively. Note that $\widehat{\eta}_{\text{naive}}$ is the reconstructed output assuming Z is an uncorrupted sensing matrix. It is interesting to note that although $\widehat{\eta}_{\text{naive}}$ leads to a biased estimator of β^* (even as $n, p, k \to \infty$), this biased estimator has the same support as β^*; this phenomenon is verified through our simulations. (However, this behavior is specific to the case when $\mathbb{E}(z_i\overline{z}_i^T) = I$, and will not happen for general frequency distributions.)

In Figure 6.6, we show the ℓ_2-norm error $\|\widehat{\beta}-\beta^*\|_2$ for the naive and noise-corrected estimators. For $p = 128$, $k = \lfloor\sqrt{p}\rfloor$, we generated a unit vector β^* with $\frac{1}{\sqrt{k}}$ in each of k random components. We then chose $n = \alpha k \log p$, for $\alpha \in [100, 2400]$, and generated the sensing frequencies ω_i uniformly at random in $[0, 2\pi)$, and the error frequencies $\psi_i \sim Unif(-\delta, \delta)$ with $\delta = \frac{\pi/2}{p}$. We then generated the measurement vector $y \in \mathbb{C}^n$, assuming the noise $\epsilon = 0$. Finally, we chose the regularization parameter $\lambda = \frac{2}{\min_j |\mathbb{E}(u_{ij})|}\sqrt{\frac{\log p}{n}}$. The plot depicts the sample size n versus the ℓ_2-norm error. As expected, the ℓ_2-error decreases to 0 for the corrected estimator, but not the naive estimator (which converges to a different vector as $n \to \infty$). In statistical terminology, the corrected estimator is consistent, but not the naive estimator.

In Figure 6.6, we reran the experiments with $p = 128, 256$, and 512, and $\alpha \in [100, 1000]$, in order to test the relative scaling of n, p, and k. Panels (a) and (b) show plots of the corrected and naive ℓ_2-error for the three problem sizes. In panels (c) and (d), we plotted the same data with the horizontal axis rescaled as $n/k \log p$. According to theory, the curves for different problem sizes should roughly stack up in the case of the corrected estimator.

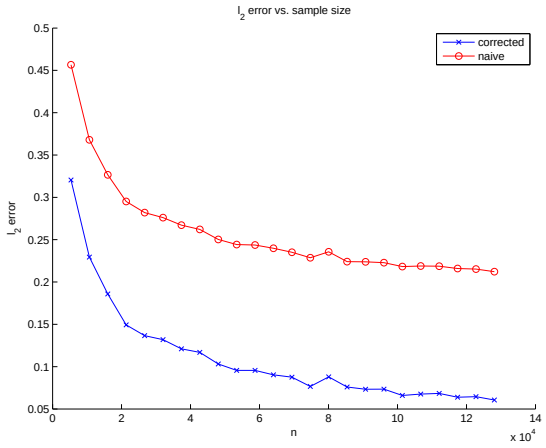


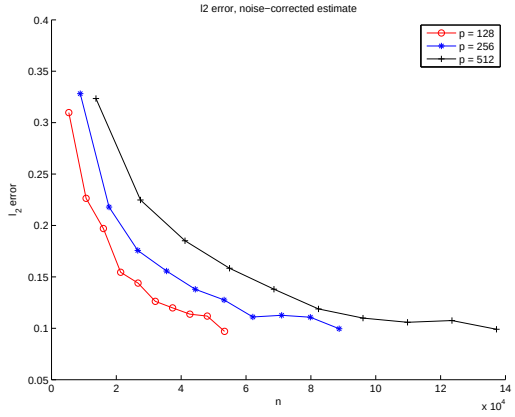
Figure 6.1: ℓ_2-error versus sample size for $p = 128$ and $k \approx \sqrt{p}$

We see this in panel (c); in contrast, the curves do not stack up for the naive estimator in panel (d).

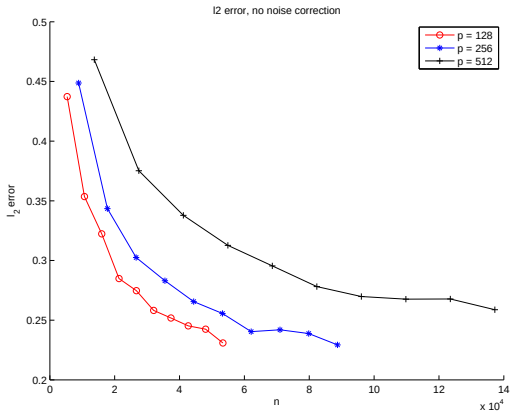
Finally, we simulated our algorithm on real image data. We began with a 512×512 axial T_2-weighted image of the brain, with each pixel taking on complex values arising from transverse magnetization. In order to speed up computation, we downsampled the original image by summing blocks of neighboring pixels, and then renormalizing to make the maximal signal intensity equal to 1. We then sparsified the image by taking the top $\frac{1}{\sqrt{256}} = 6.25\%$ wavelet coefficients according to the Daubechies basis. In our earlier notation, the parameter values are $p = 256^2$, $k = 4096$. The original image (plotted by amplitudes) is shown in Figure 6.6(a); the downsampled, sparsified image is shown in Figure 6.6(b).

Next, we sampled k-space frequencies (ω_i, ω_i') independently and uniformly at random on the interval $[0, 2\pi)$, and sampled noise perturbations (ψ_i, ψ_i') independently according to $N(0, \sigma^2)$, with $\sigma = \frac{1}{256}$. We modified the algorithm appropriately to handle a sensing matrix corresponding to a 2DFT rather than 1DFT. The reconstruction based on corrected and noisy estimators, for $n = 1,000,000$ and $\lambda = 0.1$, is given in Figure 6.6. Panels (a) and (b) show the reconstructed images, while (c) and (d) show the pixel-wise difference between the reconstructed images and the original downsampled image.

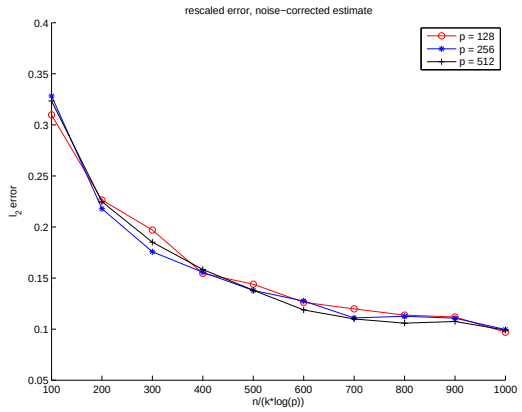
A few comments are in order. First, although the image difference shows a clear systematic bias in the naive reconstruction, overall the naive reconstruction looks fairly close to the original. Indeed, as noted at the beginning of this section, the naive estimator provides an estimator which has the same support as the corrected estimator, and differs only in amplitude. It is well-known that phases are much more important than amplitudes in reconstructing an image, so this phenomenon is not too surprising. However, if the frequencies ω_i were chosen in such a way that $\mathbb{E}(z_i \bar{z}_i^T) \neq I$, we would expect the naive reconstruction to fare much worse than the noise-corrected reconstruction.



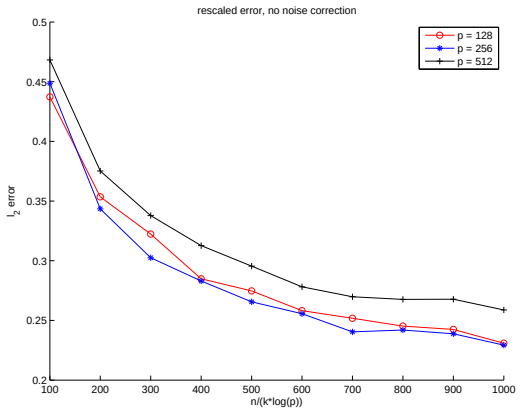
(a) ℓ_2-error vs. n for corrected estimator



(b) ℓ_2-error vs. n for naive estimator



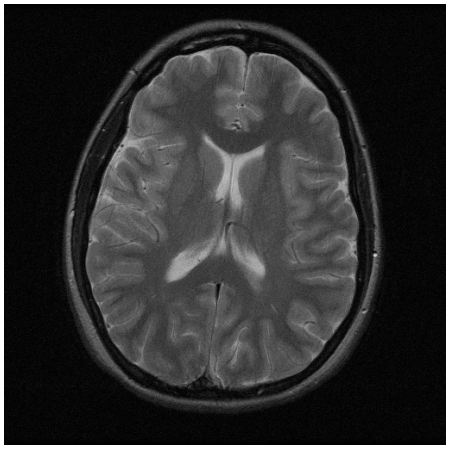
(c) ℓ_2-error vs. $\frac{n}{k \log p}$ for corrected estimator



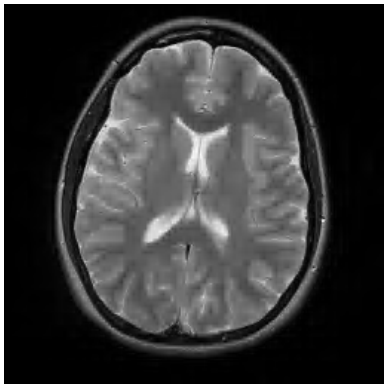
(d) ℓ_2-error vs. $\frac{n}{k \log p}$ for naive estimator

Second, our choice of $n = 10^6$ may seem excessive, especially since we are working in the context of compressed sensing, when we normally assume $n \ll p$. However, we wanted to choose n large enough such that the noise-corrected image would be fairly close to the original, and noticeably better than the naive reconstruction. In light of the discussion in the previous paragraph, we can imagine that when $\mathbb{E}(z_i \bar{z}_i^T) \neq I$, a smaller value of n would show the desired differences. Furthermore, although n is fairly large for a 512×512, our theory predicts that for larger image sizes, the required n would be relatively smaller than p, since it scales as $k \log p$.

Finally, we chose (ω_i, ω_i') independently and uniformly at random to simplify the coding for the simulation. Due to physical limitations of MRI, it is reasonable to expect that the ω_i's would be chosen independently and uniformly at random (corresponding to the phase encode frequency), but the ω_i''s would be equally spaced as multiples of $\frac{2\pi}{p}$ (corresponding to readout frequencies). This model could similarly be analyzed by our theory, and it would be interesting to simulate image reconstruction for this k-space sampling pattern.



(a) Original 512×512 image



(b) Downsampled, sparsified 256×256 image

6.7 Discussion

We have analyzed a model for compressed sensing with measurement error, where the error enters as quantifiable uncertainty in the sensing frequencies. We have provided theoretical results establishing statistical consistency of our reconstruction algorithm, and preliminary simulations show that our method may indeed lead to cleaner reconstruction when covariates are corrupted by additive noise.

Future directions of research are plentiful, and of primary interest is applying our reconstruction techniques to real data acquired in the lab. In addition, it would be interesting to explore efficient methods for solving the program (6.3) when $\mathbb{E}(x_i \overline{x}_i^T) \neq I$, and to see if the recovery algorithm performs better with non-uniformly sampled frequencies. One might also be interested in studying the behavior caused by adding other penalties in place of (or in addition to) the ℓ_1-penalty, such as the total variation penalty, which is used to encourage sparsity in the canonical basis, as exhibited by certain medical images (e.g., angiograms).

(a) Reconstructed image, corrected estimator

(b) Reconstructed image, naive estimator

(c) Image difference, corrected estimator

(d) Image difference, naive estimator

Chapter 7

Future directions

Many open problems remain in the domain of corrupted data. A very natural question is whether our methods designed for correcting systematic errors in linear regression may be extended to other regression settings such as generalized linear models. Some approaches involving conditional scores or reweighted estimating equations seem promising, but more theory needs to be developed to justify the statistical consistency of these methods. It is also interesting to ask what one might do in more general settings where the data are not corrupted completely at random. For instance, although our results cover scenarios such as missing survey data, where a respondent decides with a certain probability not to answer a particular question, they do not cover scenarios involving censored data, where the probability that an entry is missing depends on its unobserved value. Another question is whether information about the inferred regression function could be used to impute the true uncorrupted values. Finally, it would be interesting to interface our algorithmic ideas for handling corrupted MRI data with practitioners in medical imaging to find ways to apply the proposed algorithms to advance existing technology.

Turning to nonconvex M-estimators, although we have explicitly proven that the RSC condition holds for a variety of loss functions, it is unclear how one might establish RSC for an arbitrary nonconvex function. This is an important and relevant area for future research. For instance, the expectation-maximization (EM) algorithm is observed to perform well empirically on various nonconvex objectives, hinting that local optima of such functions may also be well-behaved. Other types of alternating minimization algorithms are used, for instance, in low-rank matrix completion, and the solution path is shown to have provably good behavior. It would be interesting to establish connections between broader families of nonconvex problems arising from statistical estimation, which might involve devising a more general measure of nonconvexity subsuming RSC and developing a more direct method for verifying the condition. Perhaps for certain nonconvex functions, it is only possible to establish good behavior of specific local optima, but such a unifying analysis is still nonexistent in the literature.

The newfound connections between inverse covariance matrices and the edge structure of an undirected graphical model show promise for Gaussian-based learning techniques to

be applied rigorously in non-Gaussian settings. However, there is still a wide gap between traditional Gaussian distributions and the multinomial distributions covered by our work. One question is whether it is possible to quantify "approximate Gaussianity," since approximately Gaussian distributions should still give rise to inverse covariances that approximately reflect the edge structure of the graph and could still be useful for graph estimation. Current theory on the inverse covariance structure of Gaussian distributions is completely non-robust to distributional assumptions. From a more philosophical perspective, it would be interesting to develop a deeper understanding of whether undirected graphical models are fundamentally the correct structures to infer in applications such as genetics, neuroscience, or social networks. Although the statistical theory of graphical models gives rise to many elegant mathematical results, practitioners do not seem to concur on the precise meaning of "connections" in a gene network. In addition to pushing the frontiers of inference in standard graphical models, it would be fascinating to understand the connections between different mathematical representations of network structures and the statistical methods that may be imported from one domain to another.

Appendix A

Proofs for Chapter 3

A.1 Proofs of corollaries

In this section, we include proofs of the corollaries appearing in Chapter 3.

A.1.1 Proof of Corollary 3.1

The proof of this corollary is based on two technical lemmas, one establishing that the lower- and upper-RE conditions hold with high probability, and the other proving a form of the deviation bounds (3.17).

Lemma A.1 (RE conditions, i.i.d. with additive noise)**.** *Under the conditions of Corollary 3.1, there are universal positive constants c_i such that the matrix $\widehat{\Gamma}_{add}$ satisfies the lower- and upper-RE conditions with parameters $\alpha_\ell = \frac{\lambda_{\min}(\Sigma_x)}{2}, \alpha_u = \frac{3}{2}\lambda_{\max}(\Sigma_x)$, and*

$$\tau(n,p) = c_0\, \lambda_{\min}(\Sigma_x) \max\big(\frac{(\sigma_x^2+\sigma_w^2)^2}{\lambda^2_{\min}(\Sigma_x)}, 1\big)\frac{\log p}{n},$$

with probability at least $1 - c_1 \exp\big(-c_2 n \min\big(\frac{\lambda^2_{\min}(\Sigma_x)}{(\sigma_x^2+\sigma_w^2)^2}, 1\big)\big)$.

Proof. Using Lemma A.13 in Appendix A.2, together with the substitutions

$$\widehat{\Gamma} - \Sigma_x = \frac{Z^T Z}{n} - \Sigma_z, \quad \text{and} \quad s := \frac{1}{c}\frac{n}{\log p}\min\left\{\frac{\lambda^2_{\min}(\Sigma_x)}{\sigma^4}, 1\right\}, \tag{A.1}$$

where $\sigma^2 = \sigma_x^2 + \sigma_w^2$ and c is chosen sufficiently small so $s \geq 1$, we see that it suffices to show that

$$\underbrace{\sup_{v \in \mathbb{K}(2s)} \big|v^T\big(\frac{Z^T Z}{n} - \Sigma_z\big)v\big|}_{D(s)} \leq \frac{\lambda_{\min}(\Sigma_x)}{54}$$

with high probability.

Note that the matrix Z is sub-Gaussian with parameters $(\Sigma_x + \Sigma_w, \sigma^2)$. Consequently, by Lemma A.15 in Appendix A.3, we have

$$\mathbb{P}\big[D(s) \geq t\big] \leq 2\exp\big(-c'n\min\big(\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\big) + 2s\log p\big),$$

for some universal constant $c' > 0$. Setting $t = \frac{\lambda_{\min}(\Sigma_x)}{54}$, we see that as long as the constant c in the definition (A.1) is chosen sufficiently small, we are guaranteed that

$$\mathbb{P}\big[D(s) \geq \frac{\lambda_{\min}(\Sigma_x)}{54}\big] \leq 2\exp\big(-c_2 n\min\big(\frac{\lambda^2_{\min}(\Sigma_x)}{(\sigma_x^2+\sigma_w^2)^2}, 1\big)\big), \tag{A.2}$$

which establishes the result.

□

Lemma A.2 (Deviation conditions, additive noise)**.** *Under the conditions of Corollary 3.1, there are universal positive constants c_i such the deviation bound (3.16) holds with parameter*

$$\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0\sigma_z(\sigma_w + \sigma_\epsilon)\|\beta^*\|_2,$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

Proof. Using the fact that $y = X\beta^* + \epsilon$, we may write

$$\begin{aligned}
\|\widehat{\gamma} - \widehat{\Gamma}\beta^*\|_\infty &= \big\|\frac{Z^T y}{n} - \big(\frac{Z^T Z}{n} - \Sigma_w\big)\beta^*\big\|_\infty \\
&= \big\|\frac{Z^T(X\beta^* + \epsilon)}{n} - \big(\frac{Z^T Z}{n} - \Sigma_w\big)\beta^*\big\|_\infty \\
&\leq \big\|\frac{Z^T\epsilon}{n}\big\|_\infty + \big\|\big(\Sigma_w - \frac{Z^T W}{n}\big)\beta^*\big\|_\infty.
\end{aligned}$$

Hence, the conclusion follows easily from Lemma A.14 in Appendix A.3.

□

Extension to unknown Σ_w In the case when Σ_w is unknown, we first verify the deviation bound (3.16). Note that the form of $\widehat{\gamma}$ is the same as in the case when Σ_w is known, so it suffices to bound the quantity $\|(\widetilde{\Gamma} - \Sigma_x)\beta^*\|_\infty$ w.h.p. Furthermore,

$$\begin{aligned}
\|(\widetilde{\Gamma} - \Sigma_x)\beta^*\|_\infty &\leq \|(\widetilde{\Gamma} - \widehat{\Gamma})\beta^*\|_\infty + \|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty \\
&= \|(\widehat{\Sigma}_w - \Sigma_w)\beta^*\|_\infty + \|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty,
\end{aligned}$$

and the second term is bounded by $c\sigma_z^2\sqrt{\frac{\log p}{n}}$ w.h.p., by Lemma A.14 in Appendix A.3. If we use the estimator $\widehat{\Sigma}_w = \frac{1}{n}W_0^T W_0$, then

$$\mathbb{P}\big(\|(\widehat{\Sigma}_w - \Sigma_w)\beta^*\|_\infty \leq c\sigma_w^2\sqrt{\frac{\log p}{n}}\big) \geq 1 - c_1\exp(-c_2\log p)$$

by the same sub-Gaussian tail bounds. Since $\sigma_w^2 \leq \sigma_z^2$, we conclude that

$$\|(\widetilde{\Gamma} - \Sigma_x)\beta^*\|_\infty \leq c\sigma_z^2\sqrt{\frac{\log p}{n}}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$, as wanted.

Turning to the RE conditions, we similarly write

$$\begin{aligned}|v^T(\widetilde{\Gamma} - \Sigma_x)v| &\leq |v^T(\widetilde{\Gamma} - \widehat{\Gamma})v| + |v^T(\widehat{\Gamma} - \Sigma_x)v| \\ &= |v^T(\widehat{\Sigma}_w - \Sigma_w)v| + |v^T(\widehat{\Gamma} - \Sigma_x)v|.\end{aligned}$$

Then applying Lemma A.15 to both terms, followed by Lemma A.13, yields the required bounds.

A.1.2 Proof of Corollary 3.2

We now turn to the proof of Corollary 3.2, which applies to the case of missing data, based on the general M-estimator using the pair $(\widehat{\Gamma}_{\text{mis}}, \widehat{\gamma}_{\text{mis}})$ defined in equation (3.11). We will establish that the RE conditions and deviation conditions (3.17) hold with high probability.

Lemma A.3 (RE conditions, i.i.d. with missing data)**.** *Under the conditions of Corollary 3.2, there are universal positive constants c_i such that $\widehat{\Gamma}_{mis}$ satisfies the lower- and upper-RE conditions with parameters $\alpha_\ell = \frac{\lambda_{\min}(\Sigma_x)}{2}$, $\alpha_u = \frac{3}{2}\lambda_{\max}(\Sigma_x)$, and*

$$\tau(n,p) = c_0\lambda_{\min}(\Sigma_x)\max\Big(\frac{1}{(1-\alpha_{\max})^2}\frac{\sigma_x^4}{\lambda_{\min}^2(\Sigma_x)}, 1\Big)\frac{\log p}{n},$$

with probability at least $1 - c_1\exp\big(-c_2 n \min\big((1-\alpha_{\max})^4 \cdot \frac{\lambda_{\min}^2(\Sigma_x)}{\sigma_x^4}, 1\big)\big)$.

Proof. This proof parallels the proof of Lemma A.1 for the additive noise case. We make use of Lemma A.13. This time, we have

$$\widehat{\Gamma} - \Sigma_x = \frac{Z^TZ}{n} \circledcirc M - \Sigma_x = \Big(\frac{Z^TZ}{n} - \Sigma_z\Big) \circledcirc M,$$

with the parameter s defined as in equation (A.1), with $\sigma^2 = \frac{\sigma_x^2}{(1-\alpha_{\max})^2}$. Note that for a vector $v \in \mathbb{R}^p$, we have

$$\begin{aligned}\big|v^T(\widehat{\Gamma} - \Sigma_x)v\big| &= \big|v^T\big(\big(\frac{Z^TZ}{n} - \Sigma_z\big) \circledcirc M\big)v\big| \\ &\leq \frac{1}{\|M\|_{\min}}\big|v^T\big(\frac{Z^TZ}{n} - \Sigma_z\big)v\big| \\ &\leq \frac{1}{(1-\alpha_{\max})^2}\big|v^T\big(\frac{Z^TZ}{n} - \Sigma_z\big)v\big|. \qquad \text{(A.3)}\end{aligned}$$

Furthermore, Z is a sub-Gaussian matrix with parameters (Σ_z, σ_x^2), so applying Lemma A.15 in Appendix A.3 with $t = (1-\alpha_{\max})^2 \frac{\lambda_{\min}(\Sigma_x)}{54}$ to the right-hand expression, we obtain the bound

$$\mathbb{P}\big[D(s) \geq \frac{\lambda_{\min}(\Sigma_x)}{54}\big] \leq 2\exp\big(-c_2 n \min\big((1-\alpha_{\max})^4 \cdot \frac{\lambda_{\min}^2(\Sigma_x)}{\sigma_x^4}, 1\big)\big).$$

□

Lemma A.4 (Deviation conditions, missing data)**.** *Under the conditions of Corollary 3.2, there are universal positive constants c_i such the deviation bounds (3.17) hold with parameter*

$$\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0 \, \frac{\sigma_x}{1-\alpha_{\max}} \, \big(\sigma_\epsilon + \frac{\sigma_x}{1-\alpha_{\max}}\big),$$

with probability at least $1 - c_1 \exp\big(-c_2 \log p\big)$.

Proof. The key idea is to note that the observed matrix Z is a sub-Gaussian matrix with parameter σ_x^2. Indeed, recalling that the hidden matrix X is sub-Gaussian with parameter σ_x^2, we see that for any unit vector $v \in \mathbb{R}^p$, and any missing value pattern of X_i, we have

$$\mathbb{E}\big[\exp(\lambda Z v) \;\big|\; \text{missing values}\big] = \mathbb{E}(\exp(\lambda X_i u)) \leq \exp\big(\frac{\sigma_x^2 \lambda^2}{2}\big), \tag{A.4}$$

where the vector $u \in \mathbb{R}^p$ has entries $u_i = v_i$ when entry i is observed, and $u_i = 0$ otherwise. By the tower property of conditional expectation, it follows that the moment generating function of Zv is upper-bounded by the same quantity, so Z is also a sub-Gaussian matrix with parameter at most σ_x^2.

Observe that

$$\begin{aligned}
\|\widehat{\gamma} - \Sigma_x \beta^*\|_\infty &= \big\|\big(\frac{1}{n}\big(Z^T y - \operatorname{cov}(Z_i, y)\big) \oslash (\mathbf{1} - \boldsymbol{\alpha})\big)\beta^*\big\| \\
&\leq \frac{1}{1-\alpha_{\max}} \big\|\frac{1}{n}\big(Z^T y - \operatorname{cov}(z_i, y)\big)\beta^*\big\|_\infty \\
&\leq \frac{1}{1-\alpha_{\max}} \Big(\underbrace{\big\|\frac{1}{n}\big(Z^T X - \operatorname{cov}(z_i, x_i)\big)\beta^*\big\|_\infty}_{T_1} + \underbrace{\big\|\frac{Z^T \epsilon}{n}\big\|_\infty}_{T_2}\Big).
\end{aligned} \tag{A.5}$$

Using the sub-Gaussianity of the matrices X, Z, and ϵ, and Lemma A.14, the two terms may be bounded as

$$\mathbb{P}\Big[T_1 \geq c_0 \frac{\sigma_x^2}{(1-\alpha_{\max})^2}\sqrt{\frac{\log p}{n}}\Big] \leq c_1 \exp(-c_2 \log p), \tag{A.6a}$$

$$\mathbb{P}\Big[T_2 \geq c_0 \frac{\sigma_x \sigma_\epsilon}{(1-\alpha_{\max})}\sqrt{\frac{\log p}{n}}\Big] \leq c_1 \exp(-c_2 \log p). \tag{A.6b}$$

Now consider the quantity $\|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty$. By a similar manipulation, we have

$$\|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty = \big\|\big((\frac{Z^T Z}{n} - \Sigma_z) \circledcirc M\big)\beta^*\big\|_\infty \tag{A.7}$$

$$\leq \frac{1}{(1-\alpha_{\max})^2}\big\|(\frac{Z^T Z}{n} - \Sigma_z)\beta^*\big\|_\infty, \tag{A.8}$$

so using Lemma A.14 yields

$$\big\|\big(\widehat{\Gamma} - \Sigma_x\big)\beta^*\big\|_\infty \leq c_0 \frac{\sigma_x^2}{(1-\alpha_{\max})^2}\sqrt{\frac{\log p}{n}}, \tag{A.9}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. Combining bounds (A.6) and (A.9), we conclude that the deviation conditions (3.17) both hold with parameter

$$\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0 \, \frac{\sigma_x}{1-\alpha_{\max}}\big(\frac{\sigma_x}{1-\alpha_{\max}} + \sigma_\epsilon\big),$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$, as claimed. □

Extension to unknown α_j We now consider the more challenging case when the missing probabilities α_j are unknown. Note that the estimates $\widehat{\alpha}_j$ satisfy the deviation bound

$$\mathbb{P}\big(\max_j |\widehat{\alpha}_j - \alpha_j| \geq t\big) \leq c_1 \exp(-c_2 n t^2 + \log p), \tag{A.10}$$

by a Hoeffding bound for Bernoulli random variables, together with a union bound. In particular, taking $t = c_0\sqrt{\frac{\log p}{n}}$, we have

$$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_\infty \leq c_0\sqrt{\frac{\log p}{n}}, \tag{A.11}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$.

As long as $n \gtrsim \frac{\log p}{(1-\alpha_{\max})}$, as required by our results, the deviation condition (A.11) implies that $|\widehat{\alpha}_j - \alpha_j| \leq \frac{1-\alpha_{\max}}{2}$ for each j, so

$$(1 - \widehat{\alpha}_j) \geq (1 - \alpha_j) - \frac{1-\alpha_{\max}}{2} \geq \frac{1}{2}(1 - \alpha_j). \tag{A.12}$$

In particular, we obtain the bound

$$\max_j \left|\frac{1-\alpha_j}{1-\widehat{\alpha}_j} - 1\right| \leq \max_j \frac{|\alpha_j - \widehat{\alpha}_j|}{(1-\alpha_{\max})\,(1-\widehat{\alpha}_j)} \;\leq\; \frac{2}{(1-\alpha_{\max})^2}\;\max_j |\alpha_j - \widehat{\alpha}_j|, \tag{A.13}$$

and since $\left|\frac{1-\alpha_j}{1-\widehat{\alpha}_j} - 1\right| \leq 1$ by inequality (A.12), we also have

$$\begin{aligned} &\max_{i,j} \left| \frac{(1-\alpha_i)(1-\alpha_j)}{(1-\widehat{\alpha}_i)(1-\widehat{\alpha}_j)} - 1 \right| \qquad \text{(A.14)} \\ &= \left| \big(\frac{1-\alpha_i}{1-\widehat{\alpha}_i} - 1\big)\big(\frac{1-\alpha_j}{1-\widehat{\alpha}_j} - 1\big) + \big(\frac{1-\alpha_i}{1-\widehat{\alpha}_i} - 1\big) + \big(\frac{1-\alpha_j}{1-\widehat{\alpha}_j} - 1\big) \right| \\ &\leq 3 \max_j \left| \frac{1-\alpha_j}{1-\widehat{\alpha}_j} - 1 \right| \\ &\leq \frac{6}{(1-\alpha_{\max})^2} \max_j |\alpha_j - \widehat{\alpha}_j|, \qquad \text{(A.15)} \end{aligned}$$

using the triangle inequality and inequality (A.13).

We will use these bounds to verify the results of Lemmas A.3 and A.4 for the estimators (3.24). For the deviation bounds, we begin by writing

$$\|(\widetilde{\Gamma} - \Sigma_x)\beta^*\|_\infty \leq \|(\widetilde{\Gamma} - \widehat{\Gamma})\beta^*\|_\infty + \|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty.$$

Note that we have already bounded $\|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty$ in inequality (A.9). Furthermore,

$$\begin{aligned} \|(\widetilde{\Gamma} - \widehat{\Gamma})\beta^*\|_\infty &= \big\|\big(\frac{Z^T Z}{n} \circledcirc \widetilde{M} - \frac{Z^T Z}{n} \circledcirc M\big)\beta^*\|_\infty \\ &= \big\|\big(\frac{Z^T Z}{n} \circledcirc M\big) \odot \big(M \circledcirc \widetilde{M} - \mathbf{1}\mathbf{1}^T\big)\beta^*\big\|_\infty \\ &\leq \big\|M \circledcirc \widetilde{M} - \mathbf{1}\mathbf{1}^T\big\|_{\max} \big\|\big(\frac{Z^T Z}{n} \circledcirc M\big)\beta^*\big\|_\infty \\ &\leq \frac{2}{(1-\alpha_{\max})^2} \max_j |\alpha_j - \widehat{\alpha}_j| \, \big(\|\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty + \|\Sigma_x \beta^*\|_\infty\big), \end{aligned}$$

where we have used inequality (A.13) and the triangle inequality in the last inequality above. Noting that $\|\Sigma_x \beta^*\|_\infty \leq \lambda_{\max}(\Sigma_x)\|\beta^*\|_2 \leq c\sigma_x^2$ and using the bounds (A.9) and (A.11), we obtain

$$\|(\widetilde{\Gamma} - \Sigma_x)\beta^*\|_\infty \leq \frac{c\sigma_x^2}{(1-\alpha_{\max})^2}\sqrt{\frac{\log p}{n}}.$$

Combining the pieces, we conclude that the deviation conditions (3.17) are satisfied with $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0 \frac{\sigma_x}{1-\alpha_{\max}} \big(\frac{\sigma_x}{1-\alpha_{\max}} + \sigma_\epsilon\big)$, as claimed.

For the RE conditions, we use a similar argument. Note that by Lemma A.13, we need to show that $|v^T(\widetilde{\Gamma} - \Sigma_x)v| \leq \frac{\lambda_{\min}(\Sigma_x)}{54}$ for all $v \in \mathbb{K}(2s)$, with high probability. We write

$$|v^T(\widetilde{\Gamma} - \Sigma_x)v| \leq |v^T(\widetilde{\Gamma} - \widehat{\Gamma})v| + |v^T(\widehat{\Gamma} - \Sigma_x)v|,$$

and note that we have already shown how to upper-bound $|v^T(\widetilde{\Gamma} - \widehat{\Gamma})v|$ by $c\lambda_{\min}(\Sigma_x)$ with high probability. Furthermore,

$$\begin{aligned}
\big|v^T(\widetilde{\Gamma} - \widehat{\Gamma})v\big| &= \big|v^T\big(\frac{Z^TZ}{n} \widetilde{\oslash} \widetilde{M} - \frac{Z^TZ}{n} \widetilde{\oslash} M\big)v\big| \\
&= \big|v^T\big(\frac{Z^TZ}{n} \widetilde{\oslash} M\big) \odot \big(\widetilde{M} \widetilde{\oslash} M - \mathbf{1}\mathbf{1}^T\big)v\big| \\
&\leq \big\|\widetilde{M} \widetilde{\oslash} M - \mathbf{1}\mathbf{1}^T\big\|_{\max}\big|v^T\big(\frac{Z^TZ}{n} \widetilde{\oslash} M\big)v\big| \\
&\leq \big\|\widetilde{M} \widetilde{\oslash} M - \mathbf{1}\mathbf{1}^T\big\|_{\max}\big(|v^T(\widehat{\Gamma} - \Sigma_x)v| + |v^T\Sigma_x v|\big).
\end{aligned}$$

Making use of inequality (A.14) and the concentration bound (A.10) with the assignment $t = c\frac{\lambda_{\min}(\Sigma_x)}{\lambda_{\max}(\Sigma_x)}(1 - \alpha_{\max})^2$, we obtain

$$\|\widetilde{M} \widetilde{\oslash} M - \mathbf{1}\mathbf{1}^T\|_{\max}|v^T\Sigma_x v| \leq c\lambda_{\min}(\Sigma_x)$$

with probability at least

$$1 - c_1 \exp\big[-c_2 n(1 - \alpha_{\max})^4 \frac{\lambda^2_{\min}(\Sigma_x)}{\lambda^2_{\max}(\Sigma_x)}\big] \geq 1 - c_1 \exp\big[-c_2 n(1 - \alpha_{\max})^4 \frac{\lambda^2_{\min}(\Sigma_x)}{\sigma_x^4}\big].$$

Note that $t \leq c'$, so the earlier upper bound on $|v^T(\widehat{\Gamma} - \Sigma_x)v|$ is sufficient to ensure that $|v^T(\widetilde{\Gamma} - \Sigma_x)v| \leq \frac{\lambda_{\min}(\Sigma_x)}{54}$ with the required probability.

A.1.3 Proof of Corollary 3.3

We now need to establish the RE conditions and deviation bounds (3.17) for the Gaussian VAR case, which we summarize in the following:

Lemma A.5 (RE conditions, dependent case with missing data)**.** *Under the conditions of Corollary 3.3, there are universal positive constants c_i such that $\widehat{\Gamma}_{mis}$ satisfies the lower- and upper-RE conditions with $\alpha_\ell = \frac{\lambda_{\min}(\Sigma_x)}{2}$, $\alpha_u = \frac{3}{2}\lambda_{\max}(\Sigma_x)$, and*

$$\tau(n, p) = c_0\lambda_{\min}(\Sigma_x)\max\big(\frac{\zeta^4}{\lambda^2_{\min}(\Sigma_x)}1\big)\frac{\log p}{n},$$

with probability at least $1 - c_1\exp\big(-c_2 n \min\big(\frac{\lambda^2_{\min}(\Sigma_x)}{\zeta^4}, 1\big)\big)$.

Proof. The proof is identical to the proof of Lemma A.1, except we use Lemma A.19 instead of Lemma A.15 in Appendix A.3.

□

Lemma A.6 (Deviation conditions, VAR with additive noise)**.** *Under the conditions of Corollary 3.3, there are universal positive constants c_i such the deviation bounds (3.17) hold with parameter*

$$\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0 \zeta \left(\zeta + \sigma_\epsilon\right),$$

with probability at least $1 - c_1 \exp\big(-c_2 \log p\big)$.

Proof. We begin by bounding the term

$$\|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty = \max_{1 \le j \le p} \big|e_j^T \big(\frac{1}{n} Z^T Z - \Sigma_z\big)\beta^*\big|.$$

Define the function $\Phi(u, v) := u^T \big(\frac{1}{n} Z^T Z - \Sigma_z\big) v$ and rewrite the term as $\max_{1 \le j \le p} |\Phi(e_j, \beta^*)|$. For each fixed j, some simple algebra shows that

$$\Phi(e_j, \beta^*) = \frac{1}{2}\big\{\Phi(e_j + \beta^*, e_j + \beta^*) - \Phi(e_j, e_j) - \Phi(\beta^*, \beta^*)\big\}, \tag{A.16}$$

so it suffices to have a high-probability upper bound on the quantity $\Phi(v, v)$ for each fixed unit vector v. In particular, combining inequality (A.41) from Lemma A.18 (see Appendix A.3) with the union bound and the relation (A.16), we conclude that

$$\mathbb{P}\big[\|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty \ge c_0 \zeta^2 \sqrt{\frac{\log p}{n}}\big] \le c_1 \exp(-c_2 \log p). \tag{A.17}$$

We now turn to the quantity $\|\widehat{\gamma} - \Sigma_x \beta^*\|_\infty$, which by the triangle inequality may be upper-bounded as

$$\begin{aligned} \|\widehat{\gamma} - \Sigma_x \beta^*\|_\infty \le & \big\|\big(\frac{1}{n} Z^T Z - \Sigma_z\big)\beta^*\big\|_\infty + \big\|\big(\frac{1}{n} W^T W - \Sigma_w\big)\beta^*\big\|_\infty \\ & + \big\|\frac{1}{n} Z^T \epsilon\big\|_\infty + \big\|\frac{1}{n} X^T W \beta^*\big\|_\infty. \end{aligned} \tag{A.18}$$

We have already bounded the first term in inequality (A.17) above. As for the second term, the matrix W is sub-Gaussian, so that Lemma A.14 can be used to control it (as in previous arguments). In order to upper-bound the third term on the RHS of equation (A.18), we first condition on Z. Under this conditioning, the third term may be written as $\max_{\ell=1,\ldots,p} v_\ell$, where $v_\ell := \frac{1}{n}\langle Z e_\ell, \epsilon\rangle$ is a zero-mean Gaussian variable with variance at most $\frac{\sigma_\epsilon^2}{n}\big(\frac{\|Z e_\ell\|_2}{\sqrt{n}}\big)^2$. Combining the union bound and the deviation bound (A.41) with $t = 1$, we conclude that as long as $n \gtrsim \log p$, then

$$\mathbb{P}\big[\max_{\ell=1,\ldots,p} \big(\frac{\|Z e_\ell\|_2}{\sqrt{n}}\big)^2 \ge c\zeta^2\big] \le c_1 \exp\big(-c_2 \log p\big).$$

Conditioning on this event and applying standard tail bounds to control $\{v_\ell\}$, we conclude that $\mathbb{P}\big[|v_\ell| \ge t\big] \le \exp\big(-\frac{c_2}{\zeta^2 \sigma_\epsilon^2} n t^2\big)$. Setting $t = c_0 \sigma_\epsilon \zeta \sqrt{\frac{\log p}{n}}$ and then taking a union bound over $\ell \in \{1, \ldots, p\}$ yields the desired result. A similar analysis can be used to bound the fourth term, since the matrices X and W are independent. Combining the pieces yields the claim. □

A.1.4 Proof of Corollary 3.4

Lemma A.7 (RE conditions, dependent case with missing data). *Under the conditions of Corollary 3.4, there are universal positive constants c_i such that $\widehat{\Gamma}_{mis}$ satisfies the lower- and upper-RE conditions with $\alpha_\ell = \frac{\lambda_{\min}(\Sigma_x)}{2}$, $\alpha_u = \frac{3}{2}\lambda_{\max}(\Sigma_x)$, and*

$$\tau(n,p) = c_0 \lambda_{\min}(\Sigma_x) \max\big(\frac{\zeta'^4}{\lambda^2_{\min}(\Sigma_x)}, 1\big)\frac{\log p}{n},$$

with probability at least $1 - c_1 \exp\big(-c_2 n \min\big(\frac{\lambda^2_{\min}(\Sigma_x)}{\zeta'^4}, 1\big)\big)$.

Proof. Again, we simply substitute the bound of Lemma A.19 for the bound of Lemma A.15 in the proof of Lemma A.3.

□

The final step is verify the deviation bounds (3.17).

Lemma A.8 (Deviation conditions, VAR with missing data). *Under the conditions of Corollary 3.4, there are universal positive constants c_i such the deviation bounds (3.17) hold with parameter*

$$\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0 \zeta'\big(\zeta' + \sigma_\epsilon\big),$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$.

Proof. To control the term $\|(\widehat{\Gamma} - \Sigma_x)\beta^*\|_\infty$, we use the same argument as in Lemma A.6 to obtain inequality (A.17). For the term $\|\widehat{\gamma} - \Sigma_x\beta^*\|_\infty$, we use the expansion (A.5) from the i.i.d. case. We show how to bound the terms T_1 and T_2 appearing in the expansion.

For a vector $v \in \mathbb{R}^p$, write $\Psi(v) = \frac{\|v\|_2^2}{n} - \mathbb{E}\big(\frac{\|v\|_2^2}{n}\big)$, and note that

$$T_1 = \max_j \frac{1}{2}\big[\Psi(Ze_j + X\beta^*) - \Psi(Ze_j) - \Psi(X\beta^*)\big]. \tag{A.19}$$

By Lemma A.18, we may upper-bound the last term in equation (A.19) by $C\zeta'^2(1-\alpha_{\max})^2\sqrt{\frac{\log p}{n}}$, with probability at least $1 - c_1\exp(-c_2 \log p)$. In order to bound the other two terms, we again use Lemma A.18. Note that Ze_j is a mixture of Gaussians $N(0, Q_j)$, each with $\|Q_j\|_{\text{op}} \leq (1-\alpha_{\max})^2\zeta'^2$. Then $Ze_j + X\beta^*$ is also a mixture of Gaussians $N(0, Q'_j)$, and we have the bound $\|Q'_i\|_{\text{op}} \leq 4\zeta'^2(1-\alpha_{\max})^2$. Hence, by Lemma A.18 and a union bound, we conclude that $T_1 \leq c\zeta'^2(1-\alpha_{\max})^2\sqrt{\frac{\log p}{n}}$, with probability at least $1 - c_1\exp(-c_2\log p)$. Turning to term T_2 in the expansion (A.5), we condition on Z. Repeating the argument in Lemma A.6, we obtain the bound

$$T_2 \leq c_0\sigma_\epsilon\zeta'(1-\alpha_{\max})\sqrt{\frac{\log p}{n}}.$$

Finally, plugging back into inequality (A.5), we arrive at the bound

$$\|\widehat{\gamma} - \Sigma_x \beta^*\|_\infty \leq c(\zeta'^2(1-\alpha_{\max}) + \sigma_\epsilon \zeta')\sqrt{\frac{\log p}{n}}.$$

Altogether, we have the form of φ given by $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0(\sigma_\epsilon \zeta' + \zeta'^2)$, as claimed. □

A.1.5 Proof of Corollary 3.5

First note that by Theorem 3.1, we have the bounds

$$\|\widehat{\theta}^j - \theta^j\|_1 \leq \frac{c\varphi(\mathbb{Q}, \sigma_\epsilon)}{\alpha_\ell} k\sqrt{\frac{\log p}{n}}, \tag{A.20}$$

$$\|\widehat{\theta}^j - \theta^j\|_2 \leq \frac{c\varphi(\mathbb{Q}, \sigma_\epsilon)}{\alpha_\ell} \sqrt{\frac{k \log p}{n}}. \tag{A.21}$$

We now establish the following lemma, which we will use to prove the theorem.

Lemma A.9. *For each $1 \leq j \leq p$, we have*

$$\frac{1}{\lambda_{\max}(\Sigma)} \leq |a_j| \leq \frac{1}{\lambda_{\min}(\Sigma)} \quad \text{and} \quad \|\theta^j\|_2 \leq \kappa(\Sigma). \tag{A.22}$$

Proof. Observe that $|a_j| \leq \max_j |\Theta_{jj}| \leq \lambda_{\max}(\Theta) = \frac{1}{\lambda_{\min}(\Sigma)}$, and similarly, $|a_j| \geq \frac{1}{\lambda_{\max}(\Sigma)}$, which establishes the first inequality (A.22). Next, note that the rows (and also columns) of Θ are bounded in ℓ_2-norm according to the inequality

$$\|\Theta_{j\cdot}\|_2 = \|\Theta e_j\|_2 \leq \lambda_{\max}(\Theta) = \frac{1}{\lambda_{\min}(\Sigma)},$$

which implies that $\|\theta^j\|_2 = \|\Theta_{\cdot j}/a_j\|_2 = \|\Theta_{j\cdot}\|_2/|a_j| \leq \kappa(\Sigma)$, as claimed. □

Moving forward, we establish the following deviation inequalities between a_j and $\Theta_{\cdot j}$ and their respective estimators.

Lemma A.10. *For all j, we have the following deviation inequalities:*

$$|\widehat{a}_j - a_j| \leq \frac{c\kappa(\Sigma)}{\lambda_{\min}(\Sigma)}\Big(\frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\lambda_{\min}(\Sigma)} + \frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\alpha_\ell}\Big)\sqrt{\frac{k \log p}{n}}, \tag{A.23}$$

$$\|\widetilde{\Theta}_{\cdot j} - \Theta_{\cdot j}\|_1 \leq \frac{c\kappa^2(\Sigma)}{\lambda_{\min}(\Sigma)}\Big(\frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\lambda_{\min}(\Sigma)} + \frac{\varphi(\mathbb{Q}, \sigma_\epsilon)}{\alpha_\ell}\Big)k\sqrt{\frac{\log p}{n}}. \tag{A.24}$$

Proof. We first derive inequality (A.23). Since the columns of Θ are k-sparse, we have the bound $\|\theta^j\|_1 \leq \sqrt{k}\|\theta^j\|_2$. Then

$$\begin{aligned}\|\widehat{\theta}^j\|_1 &\leq \|\theta^j\|_1 + \|\widehat{\theta}^j - \theta^j\|_1 \leq c\sqrt{k}\big(\|\theta^j\|_2 + \frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\alpha_\ell}\sqrt{\frac{k\log p}{n}}\big) \\ &\leq c\sqrt{k}\big(\kappa(\Sigma) + \frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\alpha_\ell}\sqrt{\frac{\log p}{n}}\big),\end{aligned}$$

where we have used Lemma A.9 and inequality (A.20). By the assumed sample size scaling $n \gtrsim k\log p$, this simplifies to the inequality

$$\|\widehat{\theta}^j\|_1 \leq c\kappa(\Sigma)\sqrt{k}. \tag{A.25}$$

We now have

$$\begin{aligned}|\widehat{a}_j^{-1} - a_j^{-1}| &= \big|(\widehat{\Sigma}_{jj} - \widehat{\Sigma}_{j,-j}\widehat{\theta}^j) - (\Sigma_{jj} - \Sigma_{j,-j}\theta^j)\big| \\ &\leq \underbrace{|\widehat{\Sigma}_{jj} - \Sigma_{jj}|}_{T_1} + \underbrace{|\widehat{\Sigma}_{j,-j}\widehat{\theta}^j - \Sigma_{j,-j}\theta^j|}_{T_2}.\end{aligned} \tag{A.26}$$

Using inequality (3.28), we have $T_1 \leq c\varphi(\mathbb{Q},\sigma_\epsilon)\sqrt{\frac{\log p}{n}}$. Furthermore,

$$\begin{aligned}T_2 &\leq |(\widehat{\Sigma}_{j,-j} - \Sigma_{j,-j})\widehat{\theta}^j| + |\Sigma_{j,-j}(\widehat{\theta}^j - \theta^j)| \\ &\leq \|\widehat{\Sigma} - \Sigma\|_{\max}\|\widehat{\theta}^j\|_1 + \|\Sigma_{j,-j}\|_2\|\widehat{\theta}^j - \theta^j\|_2 \\ &\leq c\big(\varphi(\mathbb{Q},\sigma_\epsilon)\kappa(\Sigma) + \lambda_{\max}(\Sigma_x)\frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\alpha_\ell}\big)\sqrt{\frac{k\log p}{n}},\end{aligned}$$

using inequality (A.25) and inequality (A.21). Substituting back into inequality (A.26), we obtain

$$|\widehat{a}_j^{-1} - a_j^{-1}| \leq c\big(\varphi(\mathbb{Q},\sigma_\epsilon)\kappa(\Sigma) + \lambda_{\max}(\Sigma_x)\frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\alpha_\ell}\big)\sqrt{\frac{k\log p}{n}}$$

for all j. Hence,

$$\left|\frac{a_i}{\widehat{a}_i} - 1\right| = |a_i||\widehat{a}_i^{-1} - a_i^{-1}| \leq c\kappa(\Sigma)\big(\frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\lambda_{\min}(\Sigma)} + \frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\alpha_\ell}\big)\sqrt{\frac{k\log p}{n}},$$

using Lemma A.9, so $|\widehat{a}_j| \leq 2a_j$ for $n \gtrsim k\log p$, and

$$|\widehat{a}_j - a_j| = |\widehat{a}_j|\left|\frac{a_j}{\widehat{a}_j} - 1\right| \leq \frac{c\kappa(\Sigma)}{\lambda_{\min}(\Sigma)}\big(\frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\lambda_{\min}(\Sigma)} + \frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\alpha_\ell}\big)\sqrt{\frac{k\log p}{n}}, \tag{A.27}$$

which establishes inequality (A.23).

Turning to inequality (A.24), we have

$$\begin{aligned}\|\widehat{\Theta}_{\cdot j} - \Theta_{\cdot j}\|_1 &= |\widehat{a}_j - a_j| + \|\widehat{a}_j\widehat{\theta}^j - a_j\theta^j\|_1 \\ &\leq |\widehat{a}_j - a_j| + |a_j|\|\widehat{\theta}^j - \theta^j\|_1 + |\widehat{a}_j - a_j|\|\widehat{\theta}^j\|_1 \\ &\leq c\Big(\frac{1}{\lambda_{\min}(\Sigma)}\frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\alpha_\ell} + \frac{\kappa^2(\Sigma)}{\lambda_{\min}(\Sigma)}\Big(\frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\lambda_{\min}(\Sigma)} + \frac{\varphi(\mathbb{Q},\sigma_\epsilon)}{\alpha_\ell}\Big)\Big)k\sqrt{\frac{\log p}{n}},\end{aligned}$$

by a combination of inequalities (A.20), (A.23), (A.25), and (A.27). Noting that $\kappa(\Sigma) > 1$, we arrive at inequality (A.24).

□

Returning to the proof of Corollary 3.5, observe that since $\widehat{\Theta}$ and Θ are symmetric, we have $|||\widehat{\Theta} - \Theta|||_{\text{op}} \leq |||\widehat{\Theta} - \Theta|||_1$. Furthermore, by the triangle inequality and the definition of $\widehat{\Theta}$,

$$|||\widehat{\Theta} - \Theta|||_1 \leq |||\widehat{\Theta} - \widetilde{\Theta}|||_1 + |||\widetilde{\Theta} - \Theta|||_1 \leq 2|||\widetilde{\Theta} - \Theta|||_1 = 2\max_j \|\widetilde{\Theta}_{\cdot j} - \Theta_{\cdot j}\|_1,$$

so that the union bound and inequality (A.24) yield the claim.

A.2 Restricted eigenvalue conditions

In this appendix, we provide the proofs for various lemmas used to establish restricted eigenvalue conditions for different classes of random matrices, depending on the observation model. We begin by establishing two auxiliary lemmas, and then proceed to the main lemma used directly in the proofs of the corollaries. Our first result shows how to bound the intersection of the ℓ_1-ball with the ℓ_2-ball in terms of a simpler set.

Lemma A.11. *For any constant $s \geq 1$, we have*

$$\mathbb{B}_1(\sqrt{s}) \cap \mathbb{B}_2(1) \subseteq 3\ \text{cl}\{\text{conv}\,\{\mathbb{B}_0(s) \cap \mathbb{B}_2(1)\}\}, \tag{A.28}$$

where the balls are taken in p-dimensional space, and $\text{cl}\{\cdot\}$ *and* $\text{conv}\{\cdot\}$ *denote the topological closure and convex hull, respectively.*

Proof. Note that when $s > p$, the containment is trivial, since the right-hand set equals $\mathbb{B}_2(3)$, and the left-hand set is contained in $\mathbb{B}_2(1)$. Hence, we will assume $1 \leq s \leq p$.

Let $A, B \subseteq \mathbb{R}^p$ be closed convex sets, with support function given by $\phi_A(z) = \sup_{\theta\in A}\langle\theta,\, z\rangle$ and ϕ_B similarly defined. It is a well-known fact that $\phi_A(z) \leq \phi_B(z)$ if and only if $A \subseteq B$ (cf. Theorem 2.3.1 of Hug and Weil [37]). We now check this condition for the pair of sets $A = \mathbb{B}_1(\sqrt{s}) \cap \mathbb{B}_2(1)$ and $B = 3\ \text{cl}\big\{\,\text{conv}\,\big\{\mathbb{B}_0(s) \cap \mathbb{B}_2(1)\big\}\big\}$.

For any $z \in \mathbb{R}^p$, let $S \subseteq \{1, 2, \ldots, p\}$ be the subset that indexes the top $\lfloor s \rfloor$ elements of z in absolute value. Then $\|z_{S^c}\|_\infty \leq |z_j|$ for all $j \in S$, whence

$$\|z_{S^c}\|_\infty \ \leq\ \frac{1}{\lfloor s \rfloor}\|z_S\|_1 \leq \frac{1}{\sqrt{\lfloor s \rfloor}}\|z_S\|_2. \tag{A.29}$$

We now split the supremum over A into two parts, corresponding to the elements indexed by S and its complement S^c, thereby obtaining

$$\begin{aligned}\phi_A(z) \;=\; \sup_{\theta \in A} \langle \theta,\, z \rangle &\leq \sup_{\|\theta_S\|_2 \leq 1} \langle \theta_S,\, z_S \rangle + \sup_{\|\theta_{S^c}\|_1 \leq \sqrt{s}} \langle \theta_{S^c},\, z_{S^c} \rangle \\ &\leq \|z_S\|_2 + \sqrt{s}\|z_{S^c}\|_\infty \\ &\overset{(i)}{\leq} \big(1 + \sqrt{\frac{s}{\lfloor s \rfloor}}\big)\|z_S\|_2 \\ &\leq 3\|z_S\|_2,\end{aligned}$$

where step (i) makes use of inequality (A.29). Finally, we recognize that

$$\phi_B(z) \;=\; \sup_{\theta \in B} \langle \theta,\, z \rangle = 3 \max_{|U| = \lfloor s \rfloor} \sup_{\|\theta_U\|_2 \leq 1} \langle \theta_U,\, z_U \rangle \;=\; 3\|z_S\|_2,$$

from which the claim follows. □

For ease of notation, define the sparse set $\mathbb{K}(s) := \mathbb{B}_0(s) \cap \mathbb{B}_2(1)$ and the cone set

$$\mathbb{C}(s) := \{v : \|v\|_1 \leq \sqrt{s}\|v\|_2\}.$$

Our next result builds on Lemma A.11, showing how to control deviations uniformly over vectors in $\mathbb{R}^p$.

Lemma A.12. *For a fixed matrix $\Gamma \in \mathbb{R}^{p \times p}$, parameter $s \geq 1$, and tolerance $\delta > 0$, suppose we have the deviation condition*

$$|v^T \Gamma v| \leq \delta \qquad \forall v \in \mathbb{K}(2s). \tag{A.30}$$

Then

$$\big|v^T \Gamma v\big| \leq 27\,\delta\big(\|v\|_2^2 + \frac{1}{s}\|v\|_1^2\big) \qquad \forall v \in \mathbb{R}^p. \tag{A.31}$$

Proof. We begin by establishing the inequalities

$$|v^T \Gamma v| \leq 27\,\delta\,\|v\|_2^2 \qquad \forall v \in \mathbb{C}(s), \tag{A.32a}$$

$$|v^T \Gamma v| \leq \frac{27\,\delta}{s}\,\|v\|_1^2 \qquad \forall v \notin \mathbb{C}(s). \tag{A.32b}$$

Inequality (A.31) then follows immediately.

By rescaling, inequality (A.32a) follows if we can show that

$$|v^T \Gamma v| \leq 27\delta \quad \text{for all } v \text{ such that } \|v\|_2 = 1 \text{ and } \|v\|_1 \leq \sqrt{s}. \tag{A.33}$$

By Lemma A.11 and continuity, we further reduce the problem to proving the bound (A.33) for all vectors $v \in 3\,\mathrm{conv}\{\mathbb{K}(s)\} = \mathrm{conv}\{\mathbb{B}_0(s) \cap \mathbb{B}_2(3)\}$. Consider a weighted linear combination of the form $v = \sum_i \alpha_i v_i$, with weights $\alpha_i \geq 0$ such that $\sum_i \alpha_i = 1$, and $\|v_i\|_0 \leq s$ and $\|v_i\|_2 \leq 3$ for each i. Expanding, we can write

$$v^T \Gamma v = \big(\sum \alpha_i v_i\big)^T \Gamma \big(\sum \alpha_i v_i\big) = \sum_{i,j} \alpha_i \alpha_j \big(v_i^T \Gamma v_j\big).$$

Applying inequality (A.30) to the vectors $\frac{1}{3}v_i$, $\frac{1}{3}v_j$, and $\frac{1}{6}(v_i + v_j)$, we have

$$|v_i^T \Gamma v_j| = \frac{1}{2}|(v_i + v_j)^T \Gamma (v_i + v_j) - v_i^T \Gamma v_i - v_j^T \Gamma v_j| \leq \frac{1}{2}(36\delta + 9\delta + 9\delta) = 27\delta$$

for all i, j, and hence $|v^T \Gamma v| \leq \sum_{i,j} \alpha_i \alpha_j (27\delta) = 27\delta \|\alpha\|_2^2 = 27\delta$, establishing inequality (A.32a).

Turning to inequality (A.32b), first note that for $v \notin \mathbb{C}(s)$, we have

$$\frac{|v^T \Gamma v|}{\|v\|_1^2} \leq \frac{1}{s} \sup_{\substack{\|u\|_1 \leq \sqrt{s} \\ \|u\|_2 \leq 1}} |u^T \Gamma u| \leq \frac{27\,\delta}{s}, \tag{A.34}$$

where the first inequality follows by the substitution $u = \sqrt{s}\frac{v}{\|v\|_1}$, and the second follows by the same argument used to establish inequality (A.32a), since u is in the set appearing in Lemma A.11. Rearranging inequality (A.34) yields inequality (A.32b). □

Lemma A.13 (RE conditions)**.** *Suppose $s \geq 1$ and $\widehat{\Gamma}$ is an estimator of Σ_x satisfying the deviation condition*

$$|v^T(\widehat{\Gamma} - \Sigma_x)v| \leq \frac{\lambda_{\min}(\Sigma_x)}{54} \qquad \forall v \in \mathbb{K}(2s).$$

Then we have the lower-RE condition

$$v^T \widehat{\Gamma} v \geq \frac{\lambda_{\min}(\Sigma_x)}{2}\|v\|_2^2 - \frac{\lambda_{\min}(\Sigma_x)}{2s}\|v\|_1^2 \tag{A.35}$$

and the upper-RE condition

$$v^T \widehat{\Gamma} v \leq \frac{3}{2}\lambda_{\max}(\Sigma_x)\|v\|_2^2 + \frac{\lambda_{\min}(\Sigma_x)}{2s}\|v\|_1^2. \tag{A.36}$$

Proof. This result follows easily from Lemma A.12. Setting $\Gamma = \widehat{\Gamma} - \Sigma_x$ and $\delta = \frac{\lambda_{\min}(\Sigma_x)}{54}$, we have the bound

$$|v^T(\widehat{\Gamma} - \Sigma_x)v| \leq \frac{\lambda_{\min}(\Sigma_x)}{2}\big(\|v\|_2^2 + \frac{1}{s}\|v\|_1^2\big).$$

Then

$$v^T \widehat{\Gamma} v \geq v^T \Sigma_x v - \frac{\lambda_{\min}(\Sigma_x)}{2}\big(\|v\|_2^2 + \frac{1}{s}\|v\|_1^2\big) \quad \text{and}$$
$$v^T \widehat{\Gamma} v \leq v^T \Sigma_x v + \frac{\lambda_{\min}(\Sigma_x)}{2}\big(\|v\|_2^2 + \frac{1}{s}\|v\|_1^2\big),$$

so the inequalities follow from $\lambda_{\min}(\Sigma_x)\|v\|_2^2 \leq v^T \Sigma_x v \leq \lambda_{\max}(\Sigma_x)\|v\|_2^2$.

□

A.3 Deviation bounds

In this appendix, we state and prove some deviation bounds for various types of random matrices.

A.3.1 Bounds in the i.i.d. setting

Given a zero-mean random variable Y, we refer to the quantity $\|Y\|_{\psi_1} := \sup_{\ell \geq 1} \ell^{-1}(\mathbb{E}|Y|^{\ell})^{1/\ell}$ as its sub-exponential parameter. The finiteness of this quantity guarantees existence of all moments, and hence large-deviation bounds of the Bernstein type.

By Lemma 14 of Vershynin [89], if X is a zero-mean sub-Gaussian random variable with parameter σ, then the random variable $Y = X^2 - \mathbb{E}(X^2)$ is sub-exponential with $\|Y\|_{\psi_1} \leq 2\sigma^2$. It then follows that if $X_1, \ldots, X_n$ are zero-mean i.i.d. sub-Gaussian variables, we have the deviation inequality

$$\mathbb{P}\Big[\big|\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mathbb{E}[X_i^2]\big| \geq t\Big] \leq 2\exp\big(-c\min\big(\frac{nt^2}{4\sigma^4}, \frac{nt}{2\sigma^2}\big)\big) \quad \text{for all } t > 0,$$

where $c > 0$ is a universal constant (see Proposition 16 in Vershynin [89]). This deviation bound may be used to establish the following useful result:

Lemma A.14. *If $X \in \mathbb{R}^{n \times p_1}$ is a zero-mean sub-Gaussian matrix with parameters (Σ_x, σ_x^2), then for any fixed (unit) vector $v \in \mathbb{R}^{p_1}$, we have*

$$\mathbb{P}\big[\big|\|Xv\|_2^2 - \mathbb{E}\big[\|Xv\|_2^2\big]\big| \geq nt\big] \leq 2\exp\big(-cn\min\big(\frac{t^2}{\sigma_x^4}, \frac{t}{\sigma_x^2}\big)\big). \tag{A.37}$$

Moreover, if $Y \in \mathbb{R}^{n \times p_2}$ is a zero-mean sub-Gaussian matrix with parameters (Σ_y, σ_y^2), then

$$\mathbb{P}\big(\big\|\frac{Y^T X}{n} - \mathrm{cov}(y_i, x_i)\big\|_{\max} \geq t\big) \leq 6p_1p_2\exp\big(-cn\min\big(\frac{t^2}{(\sigma_x\sigma_y)^2}, \frac{t}{\sigma_x\sigma_y}\big)\big), \tag{A.38}$$

where X_i and Y_i are the i^{th} rows of X and Y, respectively. In particular, if $n \gtrsim \log p$, then

$$\mathbb{P}\big(\big\|\frac{Y^T X}{n} - \mathrm{cov}(y_i, x_i)\big\|_{\max} \geq c_0\sigma_x\,\sigma_y\sqrt{\frac{\log p}{n}}\big) \leq c_1\exp(-c_2\log p). \tag{A.39}$$

Proof. Inequality (A.37) follows from the above discussion and the fact that Xv is a vector of i.i.d. sub-Gaussians with parameter σ. In order to prove inequality (A.38), we first note that if Z is a zero-mean sub-Gaussian variable with parameter σ_z, then the rescaled variable Z/σ_z is sub-Gaussian with parameter 1. Consequently, we may assume that $\sigma_x = \sigma_y = 1$ without loss of generality, rescaling as necessary. We then observe that

$$e_i^T\Big\{\frac{Y^T X}{n} - \mathrm{cov}(y_i, x_i)\Big\}e_j = \frac{1}{2}\big[\Phi(Xe_j + Ye_i) - \Phi(Xe_j) - \Phi(Ye_i)\big],$$

where we have defined $\Phi(v) := \frac{\|v\|_2^2}{n} - \mathbb{E}\big(\frac{\|v\|_2^2}{n}\big)$. Since $Xe_j + Ye_i$ is sub-Gaussian with parameter at most 4, we may apply inequality (A.37) to each of the three terms, to obtain

$$\big|e_i^T\big(\frac{Y^TX}{n} - \mathrm{cov}(y_i, x_i)\big)e_j\big| \leq \frac{3t}{2}$$

with probability at least $1 - 6\exp\big(-cn\min\big\{t^2, t\big\}\big)$. Taking a union bound over all $1 \leq i \leq p_1$ and $1 \leq j \leq p_2$ yields inequality (A.38). Finally, setting $t = c_0\,\sigma_x\,\sigma_y\sqrt{\frac{\log p}{n}}$ and using the assumption $n \gtrsim \log p$ yields inequality (A.39).

□

We combine this lemma with a discretization argument and union bound to obtain the next result. For a parameter $s \geq 1$, recall the notation $\mathbb{K}(s) := \big\{v \in \mathbb{R}^p \mid \|v\|_2 \leq 1, \|v\|_0 \leq s\big\}$.

Lemma A.15. *If $X \in \mathbb{R}^{n\times p}$ is a zero-mean sub-Gaussian matrix with parameters (Σ, σ^2), then there is a universal constant $c > 0$ such that*

$$\mathbb{P}\Big[\sup_{v\in\mathbb{K}(2s)}\big|\frac{\|Xv\|_2^2}{n} - \mathbb{E}\big[\frac{\|Xv\|_2^2}{n}\big]\big| \geq t\Big] \leq 2\exp\big(-cn\min\big(\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\big) + 2s\log p\big). \tag{A.40}$$

Proof. Given $U \subseteq \{1, \ldots, p\}$, define $S_U = \{v \in \mathbb{R}^p : \|v\|_2 \leq 1, \mathrm{supp}(v) \subseteq U\}$, and note that $\mathbb{K}(2s) = \bigcup_{|U|\leq 2s} S_U$. If $\mathcal{A} = \{u_1, \ldots, u_m\}$ is a 1/3-cover of S_U, then for every $v \in S_U$, there is some $u_i \in \mathcal{A}$ such that $\|\Delta v\|_2 \leq \frac{1}{3}$, where $\Delta v = v - u_i$. It is known [49] that we can construct $\mathcal{A}$ with $|\mathcal{A}| \leq 9^{2s}$. If we define $\Phi(v_1, v_2) = v_1^T\big(\frac{X^TX}{n} - \Sigma\big)v_2$, we have

$$\sup_{v\in S_U}|\Phi(v,v)| \leq \max_i|\Phi(u_i,u_i)| + 2\sup_{v\in S_U}|\max_i\Phi(\Delta v, u_i)| + \sup_{v\in S_U}|\Phi(\Delta v, \Delta v)|.$$

Since $3\Delta v \in S_U$, it follows that

$$\sup_{v\in S_U}|\Phi(v,v)| \leq \max_i|\Phi(u_i,u_i)| + \sup_{v\in S_U}\big(\frac{2}{3}|\Phi(v,v)| + \frac{1}{9}|\Phi(v,v)|\big),$$

hence $\sup_{v\in S_U}|\Phi(v,v)| \leq \frac{9}{2}\max_i|\Phi(u_i,u_i)|$. By Lemma A.14 and a union bound, we obtain

$$\mathbb{P}\big(\sup_{v\in S_U}\big|\frac{\|Xv\|_2^2}{n} - \mathbb{E}\big(\frac{\|Xv\|_2^2}{n}\big)\big| \geq t\big) \leq 9^{2s}\cdot 2\exp\big(-cn\min\big(\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\big)\big).$$

Finally, taking a union bound over the $\binom{p}{\lfloor 2s\rfloor} \leq p^{2s}$ choices of U yields

$$\mathbb{P}\big(\sup_{v\in\mathbb{K}(2s)}\big|\frac{\|Xv\|_2^2}{n} - \mathbb{E}\big(\frac{\|Xv\|_2^2}{n}\big)\big| \geq t\big) \leq 2\exp\big(-cn\min\big(\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\big) + 2s\log p\big),$$

as claimed. □

We also have the following lemma, a slight variant of Lemma A.15 that employs the tighter bound $\binom{p}{2k} \leq (p/k)^k$:

Lemma A.16. *Suppose $X \in \mathbb{R}^{n \times p}$ is a sub-Gaussian matrix with parameter σ_x^2. For $t \leq \sigma_x^2$, we have*

$$\left| \frac{1}{n} \|Xv\|_2^2 - v^T \Sigma_x v \right| \geq t \qquad \forall v \in \mathbb{B}_0(2k) \cap \mathbb{B}_2(1),$$

with probability at most $c_1 \exp(-c_2 n t^2/\sigma_x^4 + 2k \log(p/k))$.

A.3.2 Bounds for autoregressive processes

We base our analysis of Gaussian autoregressive matrices on the following lemma:

Lemma A.17. *Suppose $Y \in \mathbb{R}^m$ is a mixture of multivariate Gaussians $Y_j \sim N(0, Q_j)$, and let $\sigma^2 = \sup_j |||Q_j|||_{op}$. Then for all $t > \frac{2}{\sqrt{m}}$, we have*

$$\mathbb{P}\left[\frac{1}{n} \left| \|Y\|_2^2 - \mathbb{E}(\|Y\|_2^2) \right| > 4t\sigma^2 \right] \leq 2 \exp\Big(- \frac{m\big(t - \frac{2}{\sqrt{m}}\big)^2}{2} \Big) + 2 \exp(-m/2).$$

Proof. This result is a generalization of Lemma I.2 in the paper [62]. By definition, the random vector Y is a mixture of random vectors of the form $\sqrt{Q_j} X_j$, where $X_j \sim N(0, I_m)$. For each index j, the function $f_j(x) = \|\sqrt{Q_j} x\|_2 / \sqrt{m}$ is Lipschitz with constant $|||\sqrt{Q}|||_{\mathrm{op}}/\sqrt{m}$. Since each X_j is Gaussian, it follows from the concentration for Lipschitz functions of Gaussians [48] that $f_j(X_j)$ is a sub-Gaussian random variable with parameter $\sigma_j^2 = |||Q_j|||_{\mathrm{op}}/m$. Therefore, the mixture $\|Y\|_2/\sqrt{n}$ is sub-Gaussian with parameter $\sigma^2 = \frac{1}{m} \sup_j |||Q_j|||_{\mathrm{op}}$. The remainder of the proof proceeds as in the paper [62]. □

We now specialize the preceding lemma to the cases of additive noise and missing data appearing in our paper.

Lemma A.18. *Let $X \in \mathbb{R}^{n \times p}$ be a Gaussian random matrix, with rows x_i generated according to a vector autoregression (3.25) with driving matrix A. Let $v \in \mathbb{R}^p$ be a fixed vector with unit norm. Then for all $t > \frac{2}{\sqrt{n}}$,*

$$\mathbb{P}\left[\left| v^T \big(\widehat{\Gamma} - \Sigma_x \big) v \right| \geq 4t\zeta^2 \right] \leq 2 \exp\Big(- \frac{n\big(t - \frac{2}{\sqrt{n}}\big)^2}{2} \Big) + 2 \exp(-n/2), \tag{A.41}$$

where

$$\zeta^2 := \begin{cases} |||\Sigma_w|||_{op} + \frac{2|||\Sigma_x|||_{op}}{1 - |||A|||_{op}} & \text{(additive noise case)}, \\ \frac{1}{(1-\alpha_{\max})^2} \frac{2|||\Sigma_x|||_{op}}{1-|||A|||_{op}} & \text{(missing data case)}. \end{cases}$$

Proof. First consider the additive noise case, where $\widehat{\Gamma} - \Sigma_x = \frac{Z^T Z}{n} - \Sigma_z$. For any fixed vector with $\|v\|_2 = 1$, the variable $Zv \in \mathbb{R}^n$ is a zero-mean Gaussian random variable with covariance matrix, say $Q \succeq 0$. In order to apply Lemma A.17, we need to upper-bound the spectral norm of Q, which we do using the elementary upper bound $|||Q|||_{\text{op}} \leq \max_{1 \leq i \leq n} \sum_{\ell=1}^n |Q_{i\ell}|$. For each pair $i, \ell \in \{1, 2, \ldots, n\}$, we have

$$|Q_{i\ell}| = |\operatorname{cov}(e_i^T Zv, e_\ell^T Zv)| = |v^T \operatorname{cov}(Z_i, Z_\ell) v|,$$

where Z_i and Z_ℓ are the i^{th} and ℓ^{th} rows of Z, and $\|v\|_2 = 1$. For $i \neq \ell$, we have

$$|v^T \operatorname{cov}(Z_i, Z_\ell) v| = |v^T \operatorname{cov}(X_i, X_\ell) v| = |v^T A^{|i-\ell|} \Sigma_x v| \leq |||\Sigma_x|||_{\text{op}} |||A|||_{\text{op}}^{|i-\ell|},$$

and for $i = \ell$, we have $|v^T \operatorname{cov}(Z^i, Z^i) v| \leq |||\Sigma_z|||_{\text{op}} \leq |||\Sigma_w|||_{\text{op}} + |||\Sigma_x|||_{\text{op}}$. Putting together the pieces, we conclude that $|||Q|||_{\text{op}} \leq \zeta^2$, with ζ as defined in the lemma statement. Consequently, the bound (A.41) follows from Lemma A.17.

In the missing data case, the variable Zv is a zero-mean mixture of Gaussians, conditioned on the positions of the missing data. Suppose Z' is the random matrix Z corresponding to a given positioning scheme (with 0's in the missing positions). We claim that

$$|||Q_j|||_{\text{op}} \leq \frac{2|||\Sigma_x|||_{\text{op}}}{1 - |||A|||_{\text{op}}}, \tag{A.42}$$

where $Q_j = \operatorname{Cov}(Z'v)$. Indeed, we write $|||Q_j|||_{\text{op}} \leq \max_i \sum_{\ell=1}^n |Q_{j,i\ell}|$, and for each pair (i, ℓ),

$$|Q_{j,i\ell}| = |\operatorname{cov}(e_i^T Z'v, e_\ell^T Z'v)| = |\operatorname{cov}(Z'^i v, Z'^\ell v)| = |\operatorname{cov}(Z^i v_1, Z^\ell v_2)|,$$

where v_1 and v_2 are the vector v with 0's in the positions corresponding to the 0's of Z'^i and Z'^ℓ, respectively. Since $|\operatorname{cov}(Z^i v_1, Z^\ell v_2)| \leq |||\Sigma_x||| \, |||A|||^{|i-\ell|}$ for $i \neq \ell$ and

$$|\operatorname{cov}(Z^i v_1, Z^i v_2)| \leq |||\Sigma_w|||_{\text{op}} + |||\Sigma_x|||_{\text{op}}$$

by a similar argument as before, the claim (A.42) follows. By the bounding technique (A.3) earlier in the paper, together with Lemma A.17, we arrive at inequality (A.41).

□

Lemma A.19. *Let X be a Gaussian matrix with rows generated from a vector autoregression with driving matrix A. Let $s \geq 1$. Then for all $t > \frac{2}{\sqrt{n}}$,*

$$\mathbb{P}\Big[\sup_{v \in \mathbb{K}(2s)} \big|v^T\big(\widehat{\Gamma} - \Sigma_x\big)v\big| \geq 4t\zeta^2\Big] \leq 4\exp\big(-cn\min\big(\big(t - \frac{2}{\sqrt{n}}\big)^2, 1\big) + 2s\log p\big), \tag{A.43}$$

with ζ as defined in Lemma A.18.

Proof. We use the single-deviation bounds from Lemma A.18, together with a discretization argument identical to that of Lemma A.15.

□

Appendix B

Proofs for Chapter 4

B.1 Properties of regularizers

In this section, we establish properties of some nonconvex regularizers covered by our theory (Section B.1.1) and verify that specific regularizers satisfy Assumption 4.1 (Section B.1.2). The properties given in Section B.1.1 are used in the proof of Theorem 4.1.

B.1.1 General properties

We begin with some general properties of regularizers that satisfy Assumption 4.1.

Lemma B.1. *Under conditions (i)–(ii) of Assumption 4.1, conditions (iii) and (iv) together imply that ρ_λ is λL-Lipschitz as a function of t. In particular, all subgradients and derivatives of ρ_λ are bounded in magnitude by λL.*

Proof. Suppose $0 \le t_1 \le t_2$. Then

$$\frac{\rho_\lambda(t_2) - \rho_\lambda(t_1)}{t_2 - t_1} \le \frac{\rho_\lambda(t_1)}{t_1},$$

by condition (iii). Applying (iii) once more, we have

$$\frac{\rho_\lambda(t_1)}{t_1} \le \lim_{t \to 0^+} \frac{\rho_\lambda(t)}{t} \le \lambda L,$$

where the last inequality comes from condition (iv). Hence,

$$0 \le \rho_\lambda(t_2) - \rho_\lambda(t_1) \le \lambda L(t_2 - t_1).$$

A similar argument applies to the cases when one (or both) of t_1 and t_2 are negative. □

Lemma B.2. *For any vector $v \in \mathbb{R}^p$, let A denote the index set of its k largest elements in magnitude. Under Assumption 4.1, we have*

$$\rho_\lambda(v_A) - \rho_\lambda(v_{A^c}) \leq \lambda L(\|v_A\|_1 - \|v_{A^c}\|_1). \tag{B.1}$$

Moreover, for an arbitrary vector $\beta \in \mathbb{R}^p$, we have

$$\rho_\lambda(\beta^*) - \rho_\lambda(\beta) \leq \lambda L(\|\nu_A\|_1 - \|\nu_{A^c}\|_1), \tag{B.2}$$

where $\nu := \beta - \beta^*$ and β^* is k-sparse.

Proof. We first establish inequality (B.1). Define $f(t) := \frac{t}{\rho_\lambda(t)}$ for $t > 0$. By our assumptions on ρ_λ, the function f is nondecreasing in $|t|$, so

$$\|v_{A^c}\|_1 = \sum_{j \in A^c} \rho_\lambda(v_j) \cdot f(|v_j|) \leq \sum_{j \in A^c} \rho_\lambda(v_j) \cdot f(\|v_{A^c}\|_\infty) = \rho_\lambda(v_{A^c}) \cdot f\left(\|v_{A^c}\|_\infty\right). \tag{B.3}$$

Again using the nondecreasing property of f, we have

$$\rho_\lambda(v_A) \cdot f(\|v_{A^c}\|_\infty) = \sum_{j \in A} \rho_\lambda(v_j) \cdot f(\|v_{A^c}\|_\infty) \leq \sum_{j \in A} \rho_\lambda(v_j) \cdot f(|v_j|) = \|v_A\|_1. \tag{B.4}$$

Note that for $t > 0$, we have

$$f(t) \geq \lim_{s \to 0^+} f(s) = \lim_{s \to 0^+} \frac{s - 0}{\rho_\lambda(s) - \rho_\lambda(0)} \geq \frac{1}{\lambda L},$$

where the last inequality follows from the bounds on the subgradients of ρ_λ from Lemma B.1. Combining this result with inequalities (B.3) and (B.4) yields

$$\rho_\lambda(v_A) - \rho_\lambda(v_{A^c}) \leq \frac{1}{f(\|v_{A^c}\|_\infty)} \cdot (\|v_A\|_1 - \|v_{A^c}\|_1) \leq \lambda L(\|v_A\|_1 - \|v_{A^c}\|_1),$$

as claimed.

We now turn to the proof of the bound (B.2). Letting $S := \operatorname{supp}(\beta^*)$ denote the support of β^*, the triangle inequality and subadditivity of ρ imply that

$$\begin{aligned}
\rho_\lambda(\beta^*) - \rho_\lambda(\beta) &= \rho_\lambda(\beta^*_S) - \rho_\lambda(\beta_S) - \rho_\lambda(\beta_{S^c}) \\
&\leq \rho_\lambda(\nu_S) - \rho_\lambda(\beta_{S^c}) \\
&= \rho_\lambda(\nu_S) - \rho_\lambda(\nu_{S^c}) \\
&\leq \rho_\lambda(\nu_A) - \rho_\lambda(\nu_{A^c}) \\
&\leq \lambda L(\|\nu_A\|_1 - \|\nu_{A^c}\|_1),
\end{aligned}$$

thereby completing the proof. □

B.1.2 Verification for specific regularizers

We now verify that Assumption 4.1 is satisfied by the SCAD and MCP regularizers. (The properties are trivial to verify for the Lasso penalty.)

Lemma B.3. *The SCAD regularizer (4.2) with parameter a satisfies the conditions of Assumption 4.1 with $L = 1$ and $\mu = \frac{1}{a-1}$.*

Proof. Conditions (i)–(iii) were already verified in Zhang and Zhang [103]. Furthermore, we may easily compute the derivative of the SCAD regularizer to be

$$\frac{\partial}{\partial t}\rho_\lambda(t) = \text{sign}(t) \cdot \left(\lambda \cdot \mathbb{I}\left\{|t| \leq \lambda\right\} + \frac{(a\lambda - |t|)_+}{a-1} \cdot \mathbb{I}\left\{|t| > \lambda\right\}\right), \qquad t \neq 0, \tag{B.5}$$

and any point in the interval $[-\lambda, \lambda]$ is a valid subgradient at $t = 0$, so condition (iv) is satisfied for any $L \geq 1$. Furthermore, we have $\frac{\partial^2}{\partial t^2}\rho_\lambda(t) \geq \frac{-1}{a-1}$, so $\rho_{\lambda,\mu}$ is convex whenever $\mu \geq \frac{1}{a-1}$, giving condition (v). □

Lemma B.4. *The MCP regularizer (4.3) with parameter b satisfies the conditions of Assumption 4.1 with1 $L = 1$ and $\mu = \frac{1}{b}$.*

Proof. Again, the conditions (i)–(iii) are already verified in Zhang and Zhang [103]. We may compute the derivative of the MCP regularizer to be

$$\frac{\partial}{\partial t}\rho_\lambda(t) = \lambda \cdot \text{sign}(t) \cdot \left(1 - \frac{|t|}{\lambda b}\right)_+, \qquad t \neq 0, \tag{B.6}$$

with subgradient $\lambda[-1, +1]$ at $t = 0$, so condition (iv) is again satisfied for any $L \geq 1$. Taking another derivative, we have $\frac{\partial^2}{\partial t^2}\rho_\lambda(t) \geq \frac{-1}{b}$, so condition (v) of Assumption 4.1 holds with $\mu = \frac{1}{b}$. □

B.2 Proofs of corollaries in Section 4.3

In this section, we provide proofs of the corollaries to Theorem 4.1 stated in Section 4.3. Throughout this section, we use the convenient shorthand notation

$$\mathcal{E}_n(\Delta) := \langle \nabla \mathcal{L}_n(\beta^* + \Delta) - \nabla \mathcal{L}_n(\beta^*), \Delta \rangle. \tag{B.7}$$

B.2.1 General results for verifying RSC

We begin with two lemmas that will be useful for establishing the RSC conditions (4.4) in the special case where $\mathcal{L}_n$ is convex. We assume throughout that $\|\Delta\|_1 \leq 2R$, since β^* and $\beta^* + \Delta$ lie in the feasible set.

Lemma B.5. *Suppose $\mathcal{L}_n$ is convex. If condition (4.4a) holds and $n \geq 4R^2\tau_1^2 \log p$, then*

$$\mathcal{E}_n(\Delta) \geq \alpha_1\|\Delta\|_2 - \sqrt{\frac{\log p}{n}}\|\Delta\|_1, \qquad \text{for all } |\Delta\|_2 \geq 1. \tag{B.8}$$

Proof. We fix an arbitrary $\Delta \in \mathbb{R}^p$ with $\|\Delta\|_2 \geq 1$. Since $\mathcal{L}_n$ is convex, the function $f : [0,1] \to \mathbb{R}$ given by $f(t) := \mathcal{L}_n(\beta^* + t\Delta)$ is also convex, so $f'(1) - f'(0) \geq f'(t) - f'(0)$ for all $t \in [0,1]$. Computing the derivatives of f yields the inequality

$$\mathcal{E}_n(\Delta) \;=\; \langle \nabla\mathcal{L}_n(\beta^* + \Delta) - \nabla\mathcal{L}_n(\beta^*),\, \Delta\rangle \geq \frac{1}{t}\,\langle \nabla\mathcal{L}_n(\beta^* + t\Delta) - \nabla\mathcal{L}_n(\beta^*),\, t\Delta\rangle.$$

Taking $t = \frac{1}{\|\Delta\|_2} \in (0,1]$ and applying condition (4.4a) to the rescaled vector $\frac{\Delta}{\|\Delta\|_2}$ then yields

$$\begin{aligned}
\mathcal{E}_n(\Delta) &\geq \|\Delta\|_2 \left(\alpha_1 - \tau_1 \frac{\log p}{n}\frac{\|\Delta\|_1^2}{\|\Delta\|_2^2}\right) \\
&\geq \|\Delta\|_2 \left(\alpha_1 - \frac{2R\tau_1 \log p}{n}\frac{\|\Delta\|_1}{\|\Delta\|_2^2}\right) \\
&\geq \|\Delta\|_2 \left(\alpha_1 - \sqrt{\frac{\log p}{n}}\frac{\|\Delta\|_1}{\|\Delta\|_2}\right) \\
&= \alpha_1\|\Delta\|_2 - \sqrt{\frac{\log p}{n}}\|\Delta\|_1,
\end{aligned}$$

where the third inequality uses the assumption on the relative scaling of (n,p) and the fact that $\|\Delta\|_2 \geq 1$. □

On the other hand, if inequality (4.4a) holds globally over $\Delta \in \mathbb{R}^p$, we obtain inequality (4.4b) for free:

Lemma B.6. *If inequality (4.4a) holds for all $\Delta \in \mathbb{R}^p$ and $n \geq 4R^2\tau_1^2 \log p$, then inequality (4.4b) holds, as well.*

Proof. Suppose $\|\Delta\|_2 \geq 1$. Then

$$\alpha_1\|\Delta\|_2^2 - \tau_1\frac{\log p}{n}\|\Delta\|_1^2 \geq \alpha_1\|\Delta\|_2 - 2R\tau_1\frac{\log p}{n}\|\Delta\|_1 \geq \alpha_1\|\Delta\|_2 - \sqrt{\frac{\log p}{n}}\|\Delta\|_1,$$

again using the assumption on the scaling of (n,p). □

B.2.2 Proof of Corollary 4.1

Note that $\mathcal{E}_n(\Delta) = \Delta^T\widehat{\Gamma}\Delta$, so in particular,

$$\mathcal{E}_n(\Delta) \geq \Delta^T\Sigma_x\Delta - |\Delta^T(\Sigma_x - \widehat{\Gamma})\Delta|.$$

Applying Lemma A.12 in Appendix A.2 with $s = \frac{n}{\log p}$ to bound the second term, we have

$$\begin{aligned}\mathcal{E}_n(\Delta) &\geq \lambda_{\min}(\Sigma_x)\|\Delta\|_2^2 - \left(\frac{\lambda_{\min}(\Sigma_x)}{2}\|\Delta\|_2^2 + \frac{c\log p}{n}\|\Delta\|_1^2\right) \\ &= \frac{\lambda_{\min}(\Sigma_x)}{2}\|\Delta\|_2^2 - \frac{c\log p}{n}\|\Delta\|_1^2,\end{aligned}$$

a bound which holds for all $\Delta \in \mathbb{R}^p$ with probability at least $1 - c_1\exp(-c_2 n)$ whenever $n \gtrsim k\log p$. Then Lemma B.6 in Appendix B.2.1 implies that the RSC condition (4.4a) holds. It remains to verify the validity of the specified choice of λ. We have

$$\begin{aligned}\|\nabla\mathcal{L}_n(\beta^*)\|_\infty = \|\widehat{\Gamma}\beta^* - \widehat{\gamma}\|_\infty &= \|(\widehat{\gamma} - \Sigma_x\beta^*) + (\Sigma_x - \widehat{\Gamma})\beta^*\|_\infty \\ &\leq \|(\widehat{\gamma} - \Sigma_x\beta^*)\|_\infty + \|(\Sigma_x - \widehat{\Gamma})\beta^*\|_\infty.\end{aligned}$$

As derived in the proofs of Chapter 3, both of these terms are upper-bounded by $c'\sqrt{\frac{\log p}{n}}$ with high probability. Consequently, the claim in the corollary follows by applying Theorem 4.1.

B.2.3 Proof of Corollary 4.2

In the case of GLMs, we have

$$\mathcal{E}_n(\Delta) = \frac{1}{n}\sum_{i=1}^n (\psi'(\langle x_i,\, \beta^* + \Delta\rangle) - \psi'(\langle x_i,\, \beta^*\rangle))\, x_i^T\Delta.$$

Applying the mean value theorem, we find that

$$\mathcal{E}_n(\Delta) = \frac{1}{n}\sum_{i=1}^n \psi''(\langle x_i,\, \beta^*\rangle + t_i\,\langle x_i,\, \Delta\rangle)\,\big(\langle x_i,\, \Delta\rangle\big)^2,$$

where $t_i \in [0,1]$. From (the proof of) Proposition 2 in Negahban et al. [63], we then have

$$\mathcal{E}_n(\Delta) \geq \alpha_1\|\Delta\|_2^2 - \tau_1\sqrt{\frac{\log p}{n}}\|\Delta\|_1\|\Delta\|_2, \qquad \forall\|\Delta\|_2 \leq 1, \tag{B.9}$$

with probability at least $1 - c_1\exp(-c_2 n)$, where $\alpha_1 \asymp \lambda_{\min}(\Sigma_x)$. Note that by the arithmetic mean-geometric mean inequality,

$$\tau_1\sqrt{\frac{\log p}{n}}\|\Delta\|_1\|\Delta\|_2 \leq \frac{\alpha_1}{2}\|\Delta\|_2^2 + \frac{\tau_1^2}{2\alpha_1}\frac{\log p}{n}\|\Delta\|_1^2,$$

and consequently,

$$\mathcal{E}_n(\Delta) \geq \frac{\alpha_1}{2}\|\Delta\|_2^2 - \frac{\tau_1^2}{2\alpha_1}\frac{\log p}{n}\|\Delta\|_1^2,$$

which establishes inequality (4.4a). Inequality (4.4b) then follows via Lemma B.5 in Appendix B.2.1.

It remains to show that there are universal constants (c, c_1, c_2) such that

$$\mathbb{P}\left(\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \geq c\sqrt{\frac{\log p}{n}}\right) \leq c_1 \exp(-c_2 \log p). \tag{B.10}$$

For each $1 \leq i \leq n$ and $1 \leq j \leq p$, define the random variable $V_{ij} := (\psi'(x_i^T \beta^*) - y_i)x_{ij}$. Our goal is to bound $\max_{j=1,\ldots,p} |\frac{1}{n}\sum_{i=1}^n V_{ij}|$. Note that

$$\mathbb{P}\left[\max_{j=1,\ldots,p} \left|\frac{1}{n}\sum_{i=1}^n V_{ij}\right| \geq \delta\right] \leq \mathbb{P}[\mathcal{A}^c] + \mathbb{P}\left[\max_{j=1,\ldots,p} \left|\frac{1}{n}\sum_{i=1}^n V_{ij}\right| \geq \delta \;\middle|\; \mathcal{A}\right], \tag{B.11}$$

where

$$\mathcal{A} := \left\{\max_{j=1,\ldots,p} \left\{\frac{1}{n}\sum_{i=1}^n x_{ij}^2\right\} \leq 2\mathbb{E}[x_{ij}^2]\right\}.$$

Since the x_{ij}'s are sub-Gaussian and $n \gtrsim \log p$, there exist universal constants (c_1, c_2) such that $\mathbb{P}[\mathcal{A}^c] \leq c_1 \exp(-c_2 n)$. The last step is to bound the second term on the right side of inequality (B.11). For any $t \in \mathbb{R}$, we have

$$\begin{aligned}\log \mathbb{E}[\exp(tV_{ij}) \mid x_i] &= \log\left[\exp(tx_{ij}\psi'(x_i^T\beta^*)\right] \cdot \mathbb{E}[\exp(-tx_{ij}y_i)] \\ &= tx_{ij}\psi'(x_i^T\beta^*) + \left(\psi(-tx_{ij} + x_i^T\beta^*) - \psi(x_i^T\beta^*)\right),\end{aligned}$$

using the fact that ψ is the cumulant generating function for the underlying exponential family. Thus, by a Taylor series expansion, there is some $v_i \in [0,1]$ such that

$$\log \mathbb{E}[\exp(tV_{ij}) \mid x_i] = \frac{t^2 x_{ij}^2}{2}\psi''(x_i^T\beta^* - v_i\, tx_{ij}) \;\leq\; \frac{\alpha_u t^2 x_{ij}^2}{2}, \tag{B.12}$$

where the inequality uses the boundedness of ψ''. Consequently, conditioned on the event $\mathcal{A}$, the variable $\frac{1}{n}\sum_{i=1}^n V_{ij}$ is sub-Gaussian with parameter at most $\kappa = \alpha_u \cdot \max_{j=1,\ldots,p} \mathbb{E}[x_{ij}^2]$, for each $j = 1, \ldots, p$. By a union bound, we then have

$$\mathbb{P}\left[\max_{j=1,\ldots,p} \left|\frac{1}{n}\sum_{i=1}^n V_{ij}\right| \geq \delta \;\middle|\; \mathcal{A}\right] \leq p\, \exp\left(-\frac{n\delta^2}{2\kappa^2}\right).$$

The claimed ℓ_1- and ℓ_2-bounds then follow directly from Theorem 4.1.

B.2.4 Proof of Corollary 4.3

We first verify condition (4.4a) in the case where $|||\Delta|||_F \leq 1$. A straightforward calculation yields

$$\nabla^2 \mathcal{L}_n(\Theta) = \Theta^{-1} \otimes \Theta^{-1} = (\Theta \otimes \Theta)^{-1}.$$

Moreover, letting $\text{vec}(\Delta) \in \mathbb{R}^{p^2}$ denote the vectorized form of the matrix Δ, applying the mean value theorem yields

$$\mathcal{E}_n(\Delta) = \text{vec}(\Delta)^T \left(\nabla^2 \mathcal{L}_n(\Theta^* + t\Delta)\right) \text{vec}(\Delta) \geq \lambda_{\min}(\nabla^2 \mathcal{L}_n(\Theta^* + t\Delta)) \, |||\Theta|||_F^2, \tag{B.13}$$

for some $t \in [0, 1]$. By standard properties of the Kronecker product [35], we have

$$\begin{aligned}\lambda_{\min}(\nabla^2 \mathcal{L}_n(\Theta^* + t\Delta)) &= |||\Theta^* + t\Delta|||_2^{-2} \geq (|||\Theta^*|||_2 + t\,|||\Delta|||_2)^{-2} \\ &\geq (|||\Theta^*|||_2 + 1)^{-2},\end{aligned}$$

using the fact that $|||\Delta|||_2 \leq |||\Delta|||_F \leq 1$. Plugging back into inequality (B.13) yields

$$\mathcal{E}_n(\Delta) \geq (|||\Theta^*|||_2 + 1)^{-2} \, |||\Theta|||_F^2,$$

so inequality (4.4a) holds with $\alpha_1 = (|||\Theta^*|||_2 + 1)^{-2}$ and $\tau_1 = 0$. Lemma B.6 then implies inequality (4.4b) with $\alpha_2 = (|||\Theta^*|||_2 + 1)^{-2}$. Finally, we need to establish that the given choice of λ satisfies the requirement (4.6) of Theorem 4.1. By the assumed deviation condition (4.16), we have

$$|||\nabla \mathcal{L}_n(\Theta^*)|||_{\max} = \left|\left|\left|\widehat{\Sigma} - (\Theta^*)^{-1}\right|\right|\right|_{\max} = |||\widehat{\Sigma} - \Sigma|||_{\max} \leq c_0 \sqrt{\frac{\log p}{n}}.$$

Applying Theorem 4.1 then implies the desired result.

B.3 Auxiliary optimization-theoretic results

In this section, we provide proofs of the supporting lemmas used in Section 4.4.

B.3.1 Derivation of three-step procedure

We begin by deriving the correctness of the three-step procedure given in Section 4.4.2. Let $\widehat{\beta}$ be the unconstrained optimum of the program (4.37). If $g_{\lambda,\mu}(\widehat{\beta}) \leq R$, we clearly have the update given in step (2). Suppose instead that $g_{\lambda,\mu}(\widehat{\beta}) > R$. Then since the program (4.27) is convex, the iterate β^{t+1} must lie on the boundary of the feasible set; i.e.,

$$g_{\lambda,\mu}(\beta^{t+1}) = R. \tag{B.14}$$

By Lagrangian duality, the program (4.27) is also equivalent to

$$\beta^{t+1} \in \arg \min_{g_{\lambda,\mu}(\beta) \leq R'} \left\{ \frac{1}{2} \left\| \beta - \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right) \right\|_2^2 \right\},$$

for some choice of constraint parameter R'. Note that this is projection of $\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta}$ onto the set $\{\beta \in \mathbb{R}^p \mid g_{\lambda,\mu}(\beta) \leq R'\}$. Since projection decreases the value of $g_{\lambda,\mu}$, equation (B.14) implies that

$$g_{\lambda,\mu}\left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta}\right) \geq R.$$

In fact, since the projection will shrink the vector to the boundary of the constraint set, equation (B.14) forces $R' = R$. This yields the update (4.38) appearing in step (3).

B.3.2 Derivation of updates for SCAD and MCP

We now derive the explicit form of the updates (4.39) and (4.40) for the SCAD and MCP regularizers, respectively. We may rewrite the unconstrained program (4.37) as

$$\begin{aligned}
\beta^{t+1} &\in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \left\| \beta - \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right) \right\|_2^2 + \frac{1}{\eta} \cdot \rho_\lambda(\beta) + \frac{\mu}{\eta} \|\beta\|_2^2 \right\} \\
&= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \left(\frac{1}{2} + \frac{\mu}{\eta} \right) \|\beta\|_2^2 - \beta^T \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right) + \frac{1}{\eta} \cdot \rho_\lambda(\beta) \right\} \\
&= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \left\| \beta - \frac{1}{1 + 2\mu/\eta} \left(\beta^t - \frac{\nabla \overline{\mathcal{L}}_n(\beta^t)}{\eta} \right) \right\|_2^2 + \frac{1/\eta}{1 + 2\mu/\eta} \cdot \rho_\lambda(\beta) \right\}. \qquad \text{(B.15)}
\end{aligned}$$

Since the program in the last line of equation (B.15) decomposes by coordinate, it suffices to solve the scalar optimization problem

$$\widehat{x} \in \arg\min_x \left\{ \frac{1}{2}(x - z)^2 + \nu \rho(x; \lambda) \right\}, \qquad \text{(B.16)}$$

for general $z \in \mathbb{R}$ and $\nu > 0$.

We first consider the case when ρ is the SCAD penalty. The solution $\widehat{x}$ of the program (B.16) in the case when $\nu = 1$ is given in Fan and Li [28]; the expression (4.39) for the more general case comes from writing out the subgradient of the objective as

$$(x - z) + \nu \rho'(x; \lambda) = \begin{cases} (x - z) + \nu\lambda[-1, +1] & \text{if } x = 0, \\ (x - z) + \nu\lambda & \text{if } 0 < x \leq \lambda, \\ (x - z) + \frac{\nu(a\lambda - x)}{a - 1} & \text{if } \lambda \leq x \leq a\lambda, \\ x - z & \text{if } x \geq a\lambda, \end{cases}$$

using the equation for the SCAD derivative (B.5), and setting the subgradient equal to zero.

Similarly, when ρ is the MCP parametrized by (b, λ), the subgradient of the objective takes the form

$$(x - z) + \nu\rho'(x; \lambda) = \begin{cases} (x - z) + \nu\lambda[-1, +1] & \text{if } x = 0, \\ (x - z) + \nu\lambda\left(1 - \frac{x}{b\lambda}\right) & \text{if } 0 < x \le b\lambda, \\ x - z & \text{if } x \ge b\lambda, \end{cases}$$

using the expression for the MCP derivative (B.6), leading to the closed-form solution given in equation (4.40). This agrees with the expression provided in Breheny and Huang [10] for the special case when $\nu = 1$.

B.3.3 Proof of Lemma 4.1

We first show that if $\lambda \ge \frac{4}{L} \cdot \|\nabla\mathcal{L}_n(\beta^*)\|_\infty$, then for any feasible β such that

$$\phi(\beta) \le \phi(\beta^*) + \bar{\eta}, \tag{B.17}$$

we have

$$\|\beta - \beta^*\|_1 \le 4\sqrt{k}\|\beta - \beta^*\|_2 + 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right). \tag{B.18}$$

Defining the error vector $\Delta := \beta - \beta^*$, inequality (B.17) implies

$$\mathcal{L}_n(\beta^* + \Delta) + \rho_\lambda(\beta^* + \Delta) \le \mathcal{L}_n(\beta^*) + \rho_\lambda(\beta^*) + \bar{\eta},$$

so subtracting $\langle\nabla\mathcal{L}_n(\beta^*), \Delta\rangle$ from both sides gives

$$\mathcal{T}(\beta^* + \Delta, \beta^*) + \rho_\lambda(\beta^* + \Delta) - \rho_\lambda(\beta^*) \le -\langle\nabla\mathcal{L}_n(\beta^*), \Delta\rangle + \bar{\eta}. \tag{B.19}$$

We claim that

$$\rho_\lambda(\beta^* + \Delta) - \rho_\lambda(\beta^*) \le \frac{\lambda L}{2}\|\Delta\|_1 + \bar{\eta}. \tag{B.20}$$

We divide the argument into two cases. First suppose $\|\Delta\|_2 \le 3$. Since $\mathcal{L}_n$ satisfies the RSC condition (4.30a), we may lower-bound the left side of inequality (B.19) and apply Hölder's inequality to obtain

$$\begin{aligned} \alpha_1\|\Delta\|_2^2 - \tau_1\frac{\log p}{n}\|\Delta\|_1^2 + \rho(\beta^* + \Delta) - \rho(\beta^*) &\le \|\nabla\mathcal{L}_n(\beta^*)\|_\infty \cdot \|\Delta\|_1 + \bar{\eta} \\ &\le \frac{\lambda L}{4}\|\Delta\|_1 + \bar{\eta}. \end{aligned} \tag{B.21}$$

Since $\|\Delta\|_1 \le 2R$ by the feasibility of β^* and $\beta^* + \Delta$, we see that inequality (B.21) together with the condition $\lambda L \ge \frac{4R\tau_1 \log p}{n}$ gives inequality (B.20). On the other hand, when $\|\Delta\|_2 \ge 3$, the RSC condition (4.30b) gives

$$\alpha_2\|\Delta\|_2 - \tau_2\sqrt{\frac{\log p}{n}}\|\Delta\|_1 + \rho_\lambda(\beta^* + \Delta) - \rho_\lambda(\beta^*) \le \frac{\lambda L}{4}\|\Delta\|_1 + \bar{\eta},$$

so for $\lambda L \geq 4\tau_2\sqrt{\frac{\log p}{n}}$, we also arrive at inequality (B.20).

By Lemma B.2 in Appendix B.1.1, we have

$$\rho_\lambda(\beta^*) - \rho_\lambda(\beta) \leq \lambda L(\|\Delta_A\|_1 - \|\Delta_{A^c}\|_1),$$

where A indexes the top k components of Δ in magnitude. Combining this bound with inequality (B.20) then implies that

$$\|\Delta_{A^c}\|_1 - \|\Delta_A\|_1 \leq \frac{1}{2}\|\Delta\|_1 + \frac{\bar{\eta}}{\lambda L} = \frac{1}{2}\|\Delta_{A^c}\|_1 + \frac{1}{2}\|\Delta_A\|_1 + \frac{\bar{\eta}}{\lambda L},$$

and consequently,

$$\|\Delta_{A^c}\|_1 \leq 3\|\Delta_A\|_1 + \frac{2\bar{\eta}}{\lambda L}.$$

Putting together the pieces, we have

$$\|\Delta\|_1 \leq 4\|\Delta_A\|_1 + \frac{2\bar{\eta}}{\lambda L} \leq 4\sqrt{k}\|\Delta\|_2 + \frac{2\bar{\eta}}{\lambda L}.$$

Using the bound $\|\Delta\|_1 \leq 2R$ once more, we obtain inequality (B.18).

We now apply the implication (B.17) to the vectors $\widehat{\beta}$ and β^t. Note that by optimality of $\widehat{\beta}$, we have

$$\phi(\widehat{\beta}) \leq \phi(\beta^*),$$

and by the assumption (4.41), we also have

$$\phi(\beta^t) \leq \phi(\widehat{\beta}) + \bar{\eta} \leq \phi(\beta^*) + \bar{\eta}.$$

Hence,

$$\begin{aligned}
\|\widehat{\beta} - \beta^*\|_1 &\leq 4\sqrt{k}\|\widehat{\beta} - \beta^*\|_2, \qquad \text{and} \\
\|\beta^t - \beta^*\|_1 &\leq 4\sqrt{k}\|\beta^t - \beta^*\|_2 + 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right).
\end{aligned}$$

By the triangle inequality, we then have

$$\begin{aligned}
\|\beta^t - \widehat{\beta}\|_1 &\leq \|\widehat{\beta} - \beta^*\|_1 + \|\beta^t - \beta^*\|_1 \\
&\leq 4\sqrt{k} \cdot \left(\|\widehat{\beta} - \beta^*\|_2 + \|\beta^t - \beta^*\|_2\right) + 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right) \\
&\leq 4\sqrt{k} \cdot \left(2\|\widehat{\beta} - \beta^*\|_2 + \|\beta^t - \widehat{\beta}\|_2\right) + 2 \cdot \min\left(\frac{\bar{\eta}}{\lambda L}, R\right),
\end{aligned}$$

as claimed.

B.3.4 Proof of Lemma 4.2

Our proof proceeds via induction on the iteration number t. Note that the base case $t = 0$ holds by assumption. Hence, it remains to show that if $\|\beta^t - \widehat{\beta}\|_2 \leq 3$ for some integer $t \geq 1$, then $\|\beta^{t+1} - \widehat{\beta}\|_2 \leq 3$, as well.

We assume for the sake of a contradiction that $\|\beta^{t+1} - \widehat{\beta}\|_2 > 3$. By the RSC condition (4.30b) and the relation (4.29), we have

$$\overline{\mathcal{T}}(\beta^{t+1}, \widehat{\beta}) \geq \alpha\|\widehat{\beta} - \beta^{t+1}\|_2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 - \mu\|\widehat{\beta} - \beta^{t+1}\|_2^2. \tag{B.22}$$

Furthermore, by convexity of $g := g_{\lambda,\mu}$, we have

$$g(\beta^{t+1}) - g(\widehat{\beta}) - \langle \nabla g(\widehat{\beta}),\, \beta^{t+1} - \widehat{\beta}\rangle \geq 0. \tag{B.23}$$

Multiplying by λ and summing with inequality (B.22) then yields

$$\begin{aligned}
&\phi(\beta^{t+1}) - \phi(\widehat{\beta}) - \langle \nabla\phi(\widehat{\beta}),\, \beta^{t+1} - \widehat{\beta}\rangle \\
&\quad \geq \alpha\|\widehat{\beta} - \beta^{t+1}\|_2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 - \mu\|\widehat{\beta} - \beta^{t+1}\|_2^2.
\end{aligned}$$

Together with the first-order optimality condition $\langle \nabla\phi(\widehat{\beta}),\, \beta^{t+1} - \widehat{\beta}\rangle \geq 0$, we then have

$$\phi(\beta^{t+1}) - \phi(\widehat{\beta}) \geq \alpha\|\widehat{\beta} - \beta^{t+1}\|_2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 - \mu\|\widehat{\beta} - \beta^{t+1}\|_2^2. \tag{B.24}$$

Since $\|\widehat{\beta} - \beta^t\|_2 \leq 3$ by the induction hypothesis, applying the RSC condition (4.30a) to the pair $(\widehat{\beta}, \beta^t)$ also gives

$$\overline{\mathcal{L}}_n(\widehat{\beta}) \geq \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla\overline{\mathcal{L}}_n(\beta^t),\, \widehat{\beta} - \beta^t\rangle + (\alpha - \mu)\cdot\|\beta^t - \widehat{\beta}\|_2^2 - \tau\frac{\log p}{n}\|\beta^t - \widehat{\beta}\|_1^2.$$

Combining with the inequality

$$g(\widehat{\beta}) \geq g(\beta^{t+1}) + \langle \nabla g(\beta^{t+1}),\, \widehat{\beta} - \beta^{t+1}\rangle,$$

we then have

$$\begin{aligned}
\phi(\widehat{\beta}) &\geq \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla\overline{\mathcal{L}}_n(\beta^t),\, \widehat{\beta} - \beta^t\rangle + \lambda g(\beta^{t+1}) + \lambda\langle \nabla g(\beta^t),\, \widehat{\beta} - \beta^{t+1}\rangle \\
&\qquad + (\alpha - \mu)\cdot\|\beta^t - \widehat{\beta}\|_2^2 - \tau\frac{\log p}{n}\|\beta^t - \widehat{\beta}\|_1^2 \\
&\geq \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla\overline{\mathcal{L}}_n(\beta^t),\, \widehat{\beta} - \beta^t\rangle + \lambda g(\beta^{t+1}) \\
&\qquad + \lambda\langle \nabla g(\beta^{t+1}),\, \widehat{\beta} - \beta^{t+1}\rangle - \tau\frac{\log p}{n}\|\beta^t - \widehat{\beta}\|_1^2.
\end{aligned} \tag{B.25}$$

Finally, the RSM condition (4.31) on the pair (β^{t+1}, β^t) gives

$$\begin{aligned}\phi(\beta^{t+1}) &\leq \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla \overline{\mathcal{L}}_n(\beta^t),\, \beta^{t+1} - \beta^t \rangle + \lambda g(\beta^{t+1}) \\ &\qquad + (\alpha_3 - \mu)\|\beta^{t+1} - \beta^t\|_2^2 + \tau \frac{\log p}{n} \|\beta^{t+1} - \beta^t\|_1^2 \qquad \text{(B.26)} \\ &\leq \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla \overline{\mathcal{L}}_n(\beta^t),\, \beta^{t+1} - \beta^t \rangle + \lambda g(\beta^{t+1}) \\ &\qquad + \frac{\eta}{2}\|\beta^{t+1} - \beta^t\|_2^2 + \frac{4R^2 \tau \log p}{n}, \qquad \text{(B.27)}\end{aligned}$$

since $\frac{\eta}{2} \geq \alpha_3 - \mu$ by assumption, and $\|\beta^{t+1} - \beta^t\|_1 \leq 2R$. It is easy to check that the update (4.27) may be written equivalently as

$$\beta^{t+1} \in \arg \min_{g(\beta) \leq R,\, \beta \in \Omega} \left\{ \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla \overline{\mathcal{L}}_n(\beta^t),\, \beta - \beta^t \rangle + \frac{\eta}{2}\|\beta - \beta^t\|_2^2 + \lambda g(\beta) \right\},$$

and the optimality of β^{t+1} then yields

$$\langle \nabla \overline{\mathcal{L}}_n(\beta^t) + \eta(\beta^{t+1} - \beta^t) + \lambda \nabla g(\beta^{t+1}),\, \beta^{t+1} - \widehat{\beta} \rangle \leq 0. \qquad \text{(B.28)}$$

Summing up inequalities (B.25), (B.26), and (B.28), we then have

$$\begin{aligned}\phi(\beta^{t+1}) - \phi(\widehat{\beta}) &\leq \frac{\eta}{2}\|\beta^{t+1} - \beta^t\|_2^2 + \eta \langle \beta^t - \beta^{t+1},\, \beta^{t+1} - \widehat{\beta} \rangle + \tau \frac{\log p}{n} \|\beta^t - \widehat{\beta}\|_1^2 \\ &\qquad + \frac{4R^2 \tau \log p}{n} \\ &= \frac{\eta}{2}\|\beta^t - \widehat{\beta}\|_2^2 - \frac{\eta}{2}\|\beta^{t+1} - \widehat{\beta}\|_2^2 + \tau \frac{\log p}{n} \|\beta^t - \widehat{\beta}\|_1^2 + \frac{4R^2 \tau \log p}{n}.\end{aligned}$$

Combining this last inequality with inequality (B.24), we have

$$\begin{aligned}\alpha\|\widehat{\beta} - \beta^{t+1}\|_2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 & \\ \leq \frac{\eta}{2}\|\beta^t - \widehat{\beta}\|_2^2 - \left(\frac{\eta}{2} - \mu\right)\|\beta^{t+1} - \widehat{\beta}\|_2^2 + \frac{8R^2 \tau \log p}{n} & \\ \leq \frac{9\eta}{2} - 3\left(\frac{\eta}{2} - \mu\right)\|\beta^{t+1} - \widehat{\beta}\|_2 + \frac{8R^2 \tau \log p}{n}, &\end{aligned}$$

since $\|\beta^t - \widehat{\beta}\|_2 \leq 3$ by the induction hypothesis and $\|\beta^{t+1} - \widehat{\beta}\|_2 > 3$ by assumption, and using the fact that $\eta \geq 2\mu$. It follows that

$$\begin{aligned}\left(\alpha - 3\mu + \frac{3\eta}{2}\right) \cdot \|\widehat{\beta} - \beta^{t+1}\|_2 &\leq \frac{9\eta}{2} + \tau\sqrt{\frac{\log p}{n}}\|\widehat{\beta} - \beta^{t+1}\|_1 + \frac{8R^2 \tau \log p}{n} \\ &\leq \frac{9\eta}{2} + 2R\tau\sqrt{\frac{\log p}{n}} + \frac{8R^2 \tau \log p}{n} \\ &\leq 3\left(\alpha - 3\mu + \frac{3\eta}{2}\right),\end{aligned}$$

where the final inequality holds whenever $2R\tau\sqrt{\frac{\log p}{n}} + \frac{8R^2\tau\log p}{n} \leq 3(\alpha - 3\mu)$. Rearranging gives $\|\beta^{t+1} - \widehat{\beta}\|_2 \leq 3$, providing the desired contradiction.

B.3.5 Proof of Lemma 4.3

We begin with an auxiliary lemma:

Lemma B.7. *Under the conditions of Lemma 4.3, we have*

$$\overline{\mathcal{T}}(\beta^t, \widehat{\beta}) \geq -2\tau\frac{\log p}{n}(\epsilon + \bar{\epsilon})^2, \quad \text{and} \tag{B.29a}$$

$$\phi(\beta^t) - \phi(\widehat{\beta}) \geq \frac{\alpha - \mu}{2}\|\widehat{\beta} - \beta^t\|_2^2 - \frac{2\tau\log p}{n}(\epsilon + \bar{\epsilon})^2. \tag{B.29b}$$

We prove this result later, taking it as given for the moment.

Define

$$\phi_t(\beta) := \overline{\mathcal{L}}_n(\beta^t) + \langle\nabla\overline{\mathcal{L}}_n(\beta^t),\, \beta - \beta^t\rangle + \frac{\eta}{2}\|\beta - \beta^t\|_2^2 + \lambda g(\beta),$$

the objective function minimized over the constraint set $\{g(\beta) \leq R\}$ at iteration t. For any $\gamma \in [0,1]$, the vector $\beta_\gamma := \gamma\widehat{\beta} + (1-\gamma)\beta^t$ belongs to the constraint set, as well. Consequently, by the optimality of β^{t+1} and feasibility of β_γ, we have

$$\phi_t(\beta^{t+1}) \leq \phi_t(\beta_\gamma) = \overline{\mathcal{L}}_n(\beta^t) + \langle\nabla\overline{\mathcal{L}}_n(\beta^t),\, \gamma\widehat{\beta} - \gamma\beta^t\rangle + \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2 + \lambda g(\beta_\gamma).$$

Appealing to inequality (B.29a), we then have

$$\begin{aligned}
\phi_t(\beta^{t+1}) &\leq (1-\gamma)\overline{\mathcal{L}}_n(\beta^t) + \gamma\overline{\mathcal{L}}_n(\widehat{\beta}) + 2\gamma\tau\frac{\log p}{n}(\epsilon + \epsilon_{\text{stat}})^2 \\
&\qquad + \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2 + \lambda g(\beta_\gamma) \\
&\overset{(i)}{\leq} \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + 2\gamma\tau\frac{\log p}{n}(\epsilon + \epsilon_{\text{stat}})^2 + \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2 \\
&\leq \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + 2\tau\frac{\log p}{n}(\epsilon + \epsilon_{\text{stat}})^2 + \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2,
\end{aligned} \tag{B.30}$$

where inequality (i) incorporates the fact that

$$g(\beta_\gamma) \leq \gamma g(\widehat{\beta}) + (1-\gamma)g(\beta^t),$$

by the convexity of g.

By the RSM condition (4.31), we also have

$$\overline{\mathcal{T}}(\beta^{t+1}, \beta^t) \leq \frac{\eta}{2}\|\beta^{t+1} - \beta^t\|_2^2 + \tau\frac{\log p}{n}\|\beta^{t+1} - \beta^t\|_1^2,$$

since $\alpha_3 - \mu \leq \frac{\eta}{2}$ by assumption, and adding $\lambda g(\beta^{t+1})$ to both sides gives

$$\begin{aligned}\phi(\beta^{t+1}) &\leq \overline{\mathcal{L}}_n(\beta^t) + \langle \nabla\overline{\mathcal{L}}_n(\beta^t),\, \beta^{t+1} - \beta^t\rangle + \frac{\eta}{2}\|\beta^{t+1} - \beta^t\|_2^2 \\ &\qquad + \tau\frac{\log p}{n}\|\beta^{t+1} - \beta^t\|_1^2 + \lambda g(\beta^{t+1}) \\ &= \phi_t(\beta^{t+1}) + \tau\frac{\log p}{n}\|\beta^{t+1} - \beta^t\|_1^2.\end{aligned}$$

Combining with inequality (B.30) then yields

$$\begin{aligned}\phi(\beta^{t+1}) \leq \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \frac{\eta\gamma^2}{2}\|\widehat{\beta} - \beta^t\|_2^2 \\ + \tau\frac{\log p}{n}\|\beta^{t+1} - \beta^t\|_1^2 + 2\tau\frac{\log p}{n}(\epsilon + \bar{\epsilon})^2.\end{aligned} \tag{B.31}$$

By the triangle inequality, we have

$$\|\beta^{t+1} - \beta^t\|_1^2 \leq \big(\|\Delta^{t+1}\|_1 + \|\Delta^t\|_1\big)^2 \;\leq\; 2\|\Delta^{t+1}\|_1^2 + 2\|\Delta^t\|_1^2,$$

where we have defined $\Delta^t := \beta^t - \widehat{\beta}$. Combined with inequality (B.31), we therefore have

$$\begin{aligned}\phi(\beta^{t+1}) \leq \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \frac{\eta\gamma^2}{2}\|\Delta^t\|_2^2 \\ + 2\tau\frac{\log p}{n}(\|\Delta^{t+1}\|_1^2 + \|\Delta^t\|_1^2) + 2\psi(n,p,\epsilon),\end{aligned}$$

where $\psi(n,p,\epsilon) := \tau\frac{\log p}{n}(\epsilon + \bar{\epsilon})^2$. Then applying Lemma 4.1 to bound the ℓ_1-norms, we have

$$\begin{aligned}\phi(\beta^{t+1}) &\leq \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \frac{\eta\gamma^2}{2}\|\Delta^t\|_2^2 \\ &\qquad + 64k\tau\frac{\log p}{n}(\|\Delta^{t+1}\|_2^2 + \|\Delta^t\|_2^2) + 10\psi(n,p,\epsilon) \\ &= \phi(\beta^t) - \gamma(\phi(\beta^t) - \phi(\widehat{\beta})) + \left(\frac{\eta\gamma^2}{2} + 64k\tau\frac{\log p}{n}\right)\|\Delta^t\|_2^2 \\ &\qquad + 64k\tau\frac{\log p}{n}\|\Delta^{t+1}\|_2^2 + 10\psi(n,p,\epsilon).\end{aligned} \tag{B.32}$$

Now introduce the shorthand $\delta_t := \phi(\beta^t) - \phi(\widehat{\beta})$ and $\upsilon(k,p,n) = k\tau\frac{\log p}{n}$. By applying inequality (B.29b) and subtracting $\phi(\widehat{\beta})$ from both sides of inequality (B.32), we have

$$\begin{aligned}\delta_{t+1} \leq \big(1-\gamma\big)\delta_t + \frac{\eta\gamma^2 + 128\upsilon(k,p,n)}{\alpha - \mu}\left(\delta_t + 2\psi(n,p,\epsilon)\right) \\ + \frac{128\upsilon(k,p,n)}{\alpha - \mu}\left(\delta_{t+1} + 2\psi(n,p,\epsilon)\right) + 10\psi(n,p,\epsilon).\end{aligned}$$

Choosing $\gamma = \frac{\alpha-\mu}{2\eta} \in (0,1)$ yields

$$\left(1 - \frac{128\upsilon(k,p,n)}{\alpha-\mu}\right)\delta_{t+1} \leq \left(1 - \frac{\alpha-\mu}{4\eta} + \frac{128\upsilon(k,p,n)}{\alpha-\mu}\right)\delta_t + 2\left(\frac{\alpha-\mu}{4\eta} + \frac{256\upsilon(k,p,n)}{\alpha-\mu} + 5\right)\psi(n,p,\epsilon),$$

or $\delta_{t+1} \leq \kappa\delta_t + \xi(\epsilon+\bar{\epsilon})^2$, where κ and ξ were previously defined in equations (4.32) and (4.43), respectively. Finally, iterating the procedure yields

$$\delta_t \leq \kappa^{t-T}\delta_T + \xi(\epsilon+\bar{\epsilon})^2(1+\kappa+\kappa^2+\cdots+\kappa^{t-T-1}) \leq \kappa^{t-T}\delta_T + \frac{\xi(\epsilon+\bar{\epsilon})^2}{1-\kappa}, \tag{B.33}$$

as claimed.

The only remaining step is to prove the auxiliary lemma.

Proof of Lemma B.7: By the RSC condition (4.30a) and the assumption (4.42), we have

$$\overline{\mathcal{T}}(\beta^t, \widehat{\beta}) \geq (\alpha-\mu)\ \|\widehat{\beta}-\beta^t\|_2^2 - \tau\frac{\log p}{n}\|\widehat{\beta}-\beta^t\|_1^2. \tag{B.34}$$

Furthermore, by convexity of g, we have

$$\lambda\Big(g(\beta^t) - g(\widehat{\beta}) - \langle\nabla g(\widehat{\beta}),\ \beta^t - \widehat{\beta}\rangle\Big) \geq 0, \tag{B.35}$$

and the first-order optimality condition for $\widehat{\beta}$ gives

$$\langle\nabla\phi(\widehat{\beta}),\ \beta^t - \widehat{\beta}\rangle \geq 0. \tag{B.36}$$

Summing inequalities (B.34), (B.35), and (B.36) then yields

$$\phi(\beta^t) - \phi(\widehat{\beta}) \geq (\alpha-\mu)\ \|\widehat{\beta}-\beta^t\|_2^2 - \tau\frac{\log p}{n}\|\widehat{\beta}-\beta^t\|_1^2.$$

Applying Lemma 4.1 to bound the term $\|\widehat{\beta}-\beta^t\|_1^2$ and using the assumption $\frac{32k\tau\log p}{n} \leq \frac{\alpha-\mu}{2}$ yields the bound (B.29b). On the other hand, applying Lemma 4.1 directly to inequality (B.34) with β^t and $\widehat{\beta}$ switched gives

$$\begin{aligned}\overline{\mathcal{T}}(\widehat{\beta}, \beta^t) &\geq (\alpha-\mu)\|\widehat{\beta}-\beta^t\|_2^2 - \tau\frac{\log p}{n}\left(32k\|\widehat{\beta}-\beta^t\|_2^2 + 2(\epsilon+\bar{\epsilon})^2\right) \\ &\geq -2\tau\frac{\log p}{n}(\epsilon+\bar{\epsilon})^2.\end{aligned}$$

This establishes inequality (B.29a).

B.4 Verifying RSC/RSM conditions

In this Appendix, we provide a proof of Proposition 4.1, which verifies the RSC (4.30) and RSM (4.31) conditions for GLMs.

B.4.1 Main argument

Using the notation for GLMs in Section 4.3.3, we introduce the shorthand $\Delta := \beta_1 - \beta_2$ and observe that, by the mean value theorem, we have

$$\mathcal{T}(\beta_1, \beta_2) = \frac{1}{n} \sum_{i=1}^{n} \psi''\big(\langle \beta_1, x_i \rangle) + t_i \langle \Delta, x_i \rangle\big) \, (\langle \Delta, x_i \rangle)^2, \tag{B.37}$$

for some $t_i \in [0, 1]$. The t_i's are i.i.d. random variables, with each t_i depending only on the random vector x_i.

Proof of bound (4.36) The proof of this upper bound is relatively straightforward given the results in Chapter 3. From the Taylor series expansion (B.37) and the boundedness assumption $\|\psi''\|_\infty \leq \alpha_u$, we have

$$\mathcal{T}(\beta_1, \beta_2) \leq \alpha_u \cdot \frac{1}{n} \sum_{i=1}^{n} \big(\langle \Delta, x_i \rangle\big)^2.$$

By known results on restricted eigenvalues for ordinary linear regression (cf. Lemma A.13 in Appendix A.2), we also have

$$\frac{1}{n} \sum_{i=1}^{n} (\langle \Delta, x_i \rangle)^2 \leq \lambda_{\max}(\Sigma_x) \left(\frac{3}{2} \|\Delta\|_2^2 + \frac{\log p}{n} \|\Delta\|_1^2 \right),$$

with probability at least $1 - c_1 \exp(-c_2 n)$. Combining the two inequalities yields the desired result.

Proof of bounds (4.35) The proof of the RSC bound is much more involved, and we provide only high-level details here, deferring the bulk of the technical analysis to the appendix. We define

$$\alpha_\ell := \left(\inf_{|t| \leq 2T} \psi''(t) \right) \frac{\lambda_{\min}(\Sigma_x)}{8},$$

where T is a suitably chosen constant depending only on $\lambda_{\min}(\Sigma_x)$ and the sub-Gaussian parameter σ_x. (In particular, see equation (B.43) below, and take $T = 3\tau$). The core of the proof is based on the following lemma, proved in Section B.4.2:

Lemma B.8. *With probability at least* $1 - c_1 \exp(-c_2 n)$, *we have*

$$\mathcal{T}(\beta_1, \beta_2) \geq \alpha_\ell \|\Delta\|_2^2 - c\sigma_x \|\Delta\|_1 \|\Delta\|_2 \sqrt{\frac{\log p}{n}},$$

uniformly over all pairs (β_1, β_2) *such that* $\beta_2 \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R)$, $\|\beta_1 - \beta_2\|_2 \leq 3$, *and*

$$\frac{\|\Delta\|_1}{\|\Delta\|_2} \leq \frac{\alpha_\ell}{c\sigma_x} \sqrt{\frac{n}{\log p}}. \tag{B.38}$$

Taking Lemma B.8 as given, we now complete the proof of the RSC condition (4.35). By the arithmetic mean-geometric mean inequality, we have

$$c\sigma_x \|\Delta\|_1 \|\Delta\|_2 \sqrt{\frac{\log p}{n}} \leq \frac{\alpha_\ell}{2} \|\Delta\|_2^2 + \frac{c^2 \sigma_x^2}{2\alpha_\ell} \frac{\log p}{n} \|\Delta\|_1^2,$$

so Lemma B.8 implies that inequality (4.35a) holds uniformly over all pairs (β_1, β_2) such that $\beta_2 \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R)$ and $\|\beta_1 - \beta_2\|_2 \leq 3$, whenever the bound (B.38) holds. On the other hand, if the bound (B.38) does not hold, then the lower bound in inequality (4.35a) is negative. By convexity of $\mathcal{L}_n$, we have $\mathcal{T}(\beta_1, \beta_2) \geq 0$, so inequality (4.35a) holds trivially in that case.

We now show that inequality (4.35b) holds: in particular, consider a pair (β_1, β_2) with $\beta_2 \in \mathbb{B}_2(3)$ and $\|\beta_1 - \beta_2\|_2 \geq 3$. For any $t \in [0, 1]$, the convexity of $\mathcal{L}_n$ implies that

$$\mathcal{L}_n(\beta_2 + t\Delta) \leq t\mathcal{L}_n(\beta_2 + \Delta) + (1 - t)\mathcal{L}_n(\beta_2),$$

where $\Delta := \beta_1 - \beta_2$. Rearranging yields

$$\mathcal{L}_n(\beta_2 + \Delta) - \mathcal{L}_n(\beta_2) \geq \frac{\mathcal{L}_n(\beta_2 + t\Delta) - \mathcal{L}_n(\beta_2)}{t},$$

so

$$\mathcal{T}(\beta_2 + \Delta, \beta_2) \geq \frac{\mathcal{T}(\beta_2 + t\Delta, \beta_2)}{t}. \tag{B.39}$$

Now choose $t = \frac{3}{\|\Delta\|_2} \in [0, 1]$ so that $\|t\Delta\|_2 = 1$. Introducing the shorthand $\alpha_1 := \frac{\alpha_\ell}{2}$ and $\tau_1 := \frac{c^2\sigma_x^2}{2\alpha_\ell}$, we may apply inequality (4.35a) to obtain

$$\begin{aligned}
\frac{\mathcal{T}(\beta_2 + t\Delta, \beta_2)}{t} &\geq \frac{\|\Delta\|_2}{3} \left(\alpha_1 \left(\frac{3\|\Delta\|_2}{\|\Delta\|_2} \right)^2 - \tau_1 \frac{\log p}{n} \left(\frac{3\|\Delta\|_1}{\|\Delta\|_2} \right)^2 \right) \\
&= 3\alpha_1 \|\Delta\|_2 - 9\tau_1 \frac{\log p}{n} \frac{\|\Delta\|_1^2}{\|\Delta\|_2}.
\end{aligned} \tag{B.40}$$

Note that inequality (4.35b) holds trivially unless $\frac{\|\Delta\|_1}{\|\Delta\|_2} \leq \frac{\alpha_\ell}{2c\sigma_x} \sqrt{\frac{n}{\log p}}$, due to the convexity of $\mathcal{L}_n$. In that case, inequalities (B.39) and (B.40) together imply

$$\mathcal{T}(\beta_2 + \Delta, \beta_2) \geq 3\alpha_1 \|\Delta\|_2 - \frac{9\tau_1 \, \alpha_\ell}{2c\sigma_x} \sqrt{\frac{\log p}{n}} \|\Delta\|_1,$$

which is exactly the bound (4.35b).

B.4.2 Proof of Lemma B.8

For a truncation level $\tau > 0$ to be chosen, define the functions

$$\varphi_\tau(u) = \begin{cases} u^2, & \text{if } |u| \le \frac{\tau}{2}, \\ (\tau - u)^2, & \text{if } \frac{\tau}{2} \le |u| \le \tau, \\ 0, & \text{if } |u| \ge \tau, \end{cases} \qquad \text{and} \qquad \alpha_\tau(u) = \begin{cases} u, & \text{if } |u| \le \tau, \\ 0, & \text{if } |u| \ge \tau. \end{cases}$$

By construction, φ_τ is τ-Lipschitz and

$$\varphi_\tau(u) \le u^2 \cdot \mathbb{I}\{|u| \le \tau\}, \quad \text{for all } u \in \mathbb{R}. \tag{B.41}$$

In addition, we define the trapezoidal function

$$\gamma_\tau(u) = \begin{cases} 1, & \text{if } |u| \le \frac{\tau}{2}, \\ 2 - \frac{2}{\tau}|u|, & \text{if } \frac{\tau}{2} \le |u| \le \tau, \\ 0, & \text{if } |u| \ge \tau, \end{cases}$$

and note that γ_τ is $\frac{2}{\tau}$-Lipschitz and $\gamma_\tau(u) \le \mathbb{I}\{|u| \le \tau\}$.

Taking $T \ge 3\tau$ so that $T \ge \tau\|\Delta\|_2$ (since $\|\Delta\|_2 \le 3$ by assumption), and defining

$$L_\psi(T) := \inf_{|u| \le 2T} \psi''(u),$$

we have the following inequality:

$$\begin{aligned} \mathcal{T}(\beta + \Delta, \beta) &= \frac{1}{n} \sum_{i=1}^n \psi''(x_i^T\beta + t_i \cdot x_i^T\Delta) \cdot (x_i^T\Delta)^2 \\ &\ge L_\psi(T) \cdot \sum_{i=1}^n (x_i^T\Delta)^2 \cdot \mathbb{I}\{|x_i^T\Delta| \le \tau\|\Delta\|_2\} \cdot \mathbb{I}\{|x_i^T\beta| \le T\} \\ &\ge L_\psi(T) \cdot \frac{1}{n} \sum_{i=1}^n \varphi_{\tau\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta), \end{aligned} \tag{B.42}$$

where the first equality is the expansion (B.37) and the second inequality uses the bound (B.41).

Now define the subset of $\mathbb{R}^p \times \mathbb{R}^p$ via

$$\mathbb{A}_\delta := \left\{ (\beta, \Delta) : \beta \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R),\ \Delta \in \mathbb{B}_2(3),\ \frac{\|\Delta\|_1}{\|\Delta\|_2} \le \delta \right\},$$

as well as the random variable

$$Z(\delta) := \sup_{(\beta,\Delta) \in \mathbb{A}_\delta} \frac{1}{\|\Delta\|_2^2} \left| \frac{1}{n} \sum_{i=1}^n \varphi_{\tau\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta) - \mathbb{E}\left[\varphi_{\tau\|\Delta\|_2}(x_i^T\Delta)\, \gamma_T(x_i^T\beta)\right] \right|.$$

For any pair $(\beta, \Delta) \in \mathbb{A}_\delta$, we have

$$\begin{aligned}
&\mathbb{E}\big[(x_i^T\Delta)^2 - \varphi_{\tau\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta)\big] \\
&\qquad \le \mathbb{E}\left[(x_i^T\Delta)^2\mathbb{I}\left\{|x_i^T\Delta| \ge \frac{\tau\|\Delta\|_2}{2}\right\}\right] + \mathbb{E}\left[(x_i^T\Delta)^2\mathbb{I}\left\{|x_i^T\beta| \ge \frac{T}{2}\right\}\right] \\
&\qquad \le \sqrt{\mathbb{E}\left[(x_i^T\Delta)^4\right]}\cdot\left(\sqrt{\mathbb{P}\left(|x_i^T\Delta| \ge \frac{\tau\|\Delta\|_2}{2}\right)} + \sqrt{\mathbb{P}\left(|x_i^T\beta| \ge \frac{T}{2}\right)}\right) \\
&\qquad \le \sigma_x^2\|\Delta\|_2^2\cdot c\exp\left(-\frac{c'\tau^2}{\sigma_x^2}\right),
\end{aligned}$$

where we have used Cauchy-Schwarz and a tail bound for sub-Gaussians, assuming $\beta \in \mathbb{B}_2(3)$. It follows that for τ chosen such that

$$c\sigma_x^2\exp\left(-\frac{c'\tau^2}{\sigma_x^2}\right) = \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{2}, \tag{B.43}$$

we have the lower bound

$$\mathbb{E}\left[\varphi_{\tau\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta)\right] \ge \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{2}\cdot\|\Delta\|_2^2. \tag{B.44}$$

By construction of φ, each summand in the expression for $Z(\delta)$ is sandwiched as

$$0 \le \frac{1}{\|\Delta\|_2}\cdot\varphi_{\tau\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta) \le \frac{\tau^2}{4}.$$

Consequently, applying the bounded differences inequality yields

$$\mathbb{P}\left(Z(\delta) \ge \mathbb{E}[Z(\delta)] + \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{4}\right) \le c_1\exp(-c_2 n). \tag{B.45}$$

Furthermore, by Lemmas B.9 and B.10 in Appendix B.5, we have

$$\mathbb{E}[Z(\delta)] \le 2\sqrt{\frac{\pi}{2}}\cdot\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta}\frac{1}{\|\Delta\|_2^2}\left|\frac{1}{n}\sum_{i=1}^n g_i\Big(\varphi_{\tau\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta)\Big)\right|\right], \tag{B.46}$$

where the g_i's are i.i.d. standard Gaussians. Conditioned on $\{x_i\}_{i=1}^n$, define the Gaussian processes

$$Z_{\beta,\Delta} := \frac{1}{\|\Delta\|_2^2}\cdot\frac{1}{n}\sum_{i=1}^n g_i\Big(\varphi_{\tau\|\Delta\|_2}(x_i^T\Delta)\cdot\gamma_T(x_i^T\beta)\Big),$$

and note that for pairs (β, Δ) and $(\widetilde{\beta}, \widetilde{\Delta})$, we have

$$\operatorname{var}\left(Z_{\beta,\Delta} - Z_{\widetilde{\beta},\widetilde{\Delta}}\right) \le 2\operatorname{var}\left(Z_{\beta,\Delta} - Z_{\widetilde{\beta},\Delta}\right) + 2\operatorname{var}\left(Z_{\widetilde{\beta},\Delta} - Z_{\widetilde{\beta},\widetilde{\Delta}}\right),$$

with

$$\begin{aligned}\operatorname{var}\left(Z_{\beta,\Delta}-Z_{\widetilde{\beta},\Delta}\right) &= \frac{1}{\|\Delta\|_2^4}\cdot\frac{1}{n^2}\sum_{i=1}^n \varphi^2_{\tau\|\Delta\|_2}(x_i^T\Delta)\cdot\left(\gamma_T(x_i^T\beta)-\gamma_T(x_i^T\widetilde{\beta})\right)^2 \\ &\leq \frac{1}{n^2}\sum_{i=1}^n \frac{\tau^4}{16}\cdot\frac{4}{T^2}\left(x_i^T(\beta-\widetilde{\beta})\right)^2,\end{aligned}$$

since $\varphi_{\tau\|\Delta\|_2}\leq\frac{\tau^2\|\Delta\|_2^2}{4}$ and γ_T is $\frac{2}{T}$-Lipschitz. Similarly,

$$\begin{aligned}\operatorname{var}\left(Z_{\widetilde{\beta},\Delta}-Z_{\widetilde{\beta},\widetilde{\Delta}}\right) &\leq \frac{1}{\|\Delta\|_2^4}\cdot\frac{1}{n^2}\sum_{i=1}^n \gamma_T^2(x_i^T\widetilde{\beta})\left(\varphi_{\tau\|\Delta\|_2}(x_i^T\Delta)-\varphi_{\tau\|\Delta\|_2}(x_i^T\widetilde{\Delta})\right)^2 \\ &\leq \frac{1}{\|\Delta\|_2^2}\cdot\frac{1}{n^2}\sum_{i=1}^n \tau^2\left(x_i^T(\Delta-\widetilde{\Delta})\right)^2.\end{aligned}$$

Defining the centered Gaussian process

$$Y_{\beta,\Delta}:=\frac{\tau^2}{2T}\cdot\frac{1}{n}\sum_{i=1}^n\widehat{g}_i\cdot x_i^T\beta+\frac{\tau}{\|\Delta\|_2}\cdot\frac{1}{n}\sum_{i=1}^n\widetilde{g}_i\cdot x_i^T\Delta,$$

where the $\widehat{g}_i$'s and $\widetilde{g}_i$'s are independent standard Gaussians, it follows that

$$\operatorname{var}\left(Z_{\beta,\Delta}-Z_{\widetilde{\beta},\widetilde{\Delta}}\right)\leq\operatorname{var}\left(Y_{\beta,\Delta}-Y_{\widetilde{\beta},\widetilde{\Delta}}\right).$$

Applying Lemma B.11 in Appendix B.5, we then have

$$\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta} Z_{\beta,\Delta}\right]\leq 2\cdot\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta} Y_{\beta,\Delta}\right]. \tag{B.47}$$

Note further (cf. p.77 of Ledoux and Talagrand [49]) that

$$\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta}|Z_{\beta,\Delta}|\right]\leq\mathbb{E}\left[|Z_{\beta_0,\Delta_0}|\right]+2\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta} Z_{\beta,\Delta}\right], \tag{B.48}$$

for any $(\beta_0,\Delta_0)\in\mathbb{A}_\delta$, and furthermore,

$$\mathbb{E}\left[|Z_{\beta_0,\Delta_0}|\right]\leq\sqrt{\frac{2}{\pi}}\cdot\sqrt{\operatorname{var}\left(Z_{\beta_0,\Delta_0}\right)}\leq\frac{1}{\|\Delta\|_2}\cdot\sqrt{\frac{2}{\pi}}\cdot\sqrt{\frac{\tau^2}{4n}}. \tag{B.49}$$

Finally,

$$\begin{aligned}\mathbb{E}\left[\sup_{(\beta,\Delta)\in\mathbb{A}_\delta} Y_{\beta,\Delta}\right] &\leq \frac{\tau^2 R}{2T}\cdot\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\widehat{g}_i x_i\right\|_\infty\right]+\tau\delta\cdot\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\widetilde{g}_i x_i\right\|_\infty\right] \\ &\leq \frac{c\tau^2 R\sigma_x}{2T}\sqrt{\frac{\log p}{n}}+c\tau\delta\sigma_x\cdot\sqrt{\frac{\log p}{n}},\end{aligned} \tag{B.50}$$

by Lemma B.13 in Appendix B.5. Combining inequalities (B.46), (B.47), (B.48), (B.49), and (B.50), we then obtain

$$\mathbb{E}[Z(\delta)] \leq \frac{c'\tau^2 R\sigma_x}{2T}\sqrt{\frac{\log p}{n}} + c'\tau\delta\sigma_x \cdot \sqrt{\frac{\log p}{n}}. \tag{B.51}$$

Finally, combining inequalities (B.44), (B.45), and (B.51), we see that under the scaling $R\sqrt{\frac{\log p}{n}} \precsim 1$, we have

$$\begin{aligned}
&\frac{1}{\|\Delta\|_2^2} \cdot \frac{1}{n}\sum_{i=1}^{n} \varphi_{\tau\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta) \\
&\qquad \geq \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{4} - \left(\frac{c'\tau^2 R\sigma_x}{2T}\sqrt{\frac{\log p}{n}} + c'\tau\delta\sigma_x\sqrt{\frac{\log p}{n}}\right) \\
&\qquad \geq \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{8} - c'\tau\delta\sigma_x\sqrt{\frac{\log p}{n}},
\end{aligned} \tag{B.52}$$

uniformly over all $(\beta, \Delta) \in \mathbb{A}_\delta$, with probability at least $1 - c_1\exp(-c_2 n)$.

It remains to extend this bound to one that is uniform in the ratio $\frac{\|\Delta\|_1}{\|\Delta\|_2}$, which we do via a peeling argument [2, 31]. Consider the inequality

$$\frac{1}{\|\Delta\|_2^2} \cdot \frac{1}{n}\sum_{i=1}^{n} \varphi_{\tau\|\Delta\|_2}(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta) \geq \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{8} - 2c'\tau\sigma_x\frac{\|\Delta\|_1}{\|\Delta\|_2}\sqrt{\frac{\log p}{n}}, \tag{B.53}$$

as well as the event

$$\mathcal{E} := \left\{\text{ineqality (B.53) holds } \forall \|\beta\|_2 \leq 3 \text{ and } \tfrac{\|\Delta\|_1}{\|\Delta\|_2} \leq \tfrac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{16c'\tau\sigma_x}\sqrt{\tfrac{n}{\log p}}\right\}.$$

Define the function

$$f(\beta, \Delta; X) := \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{8} - \frac{1}{\|\Delta\|_2^2} \cdot \frac{1}{n}\sum_{i=1}^{n} \varphi_\tau(x_i^T\Delta) \cdot \gamma_T(x_i^T\beta), \tag{B.54}$$

along with

$$g(\delta) := c'\tau\sigma_x\delta\sqrt{\frac{\log p}{n}}, \qquad \text{and} \qquad h(\beta, \Delta) := \frac{\|\Delta\|_1}{\|\Delta\|_2}.$$

Note that inequality (B.52) implies

$$\mathbb{P}\left(\sup_{h(\beta,\Delta)\leq\delta} f(\beta, \Delta; X) \geq g(\delta)\right) \leq c_1\exp(-c_2 n), \quad \text{for any } \delta > 0, \tag{B.55}$$

where the sup is also restricted to $\{(\beta, \Delta) : \beta \in \mathbb{B}_2(3) \cap \mathbb{B}_1(R),\ \Delta \in \mathbb{B}_2(3)\}$.

Since $\frac{\|\Delta\|_1}{\|\Delta\|_2} \geq 1$, we have

$$1 \leq\ h(\beta, \Delta)\ \leq \frac{\lambda_{\min}\left(\mathbb{E}[x_i x_i^T]\right)}{16c'\tau\sigma_x}\sqrt{\frac{n}{\log p}}, \tag{B.56}$$

over the region of interest. For each integer $m \geq 1$, define the set

$$\mathbb{V}_m := \left\{(\beta, \Delta)\ \mid\ 2^{m-1}\mu \leq g(h(\beta, \Delta)) \leq 2^m\mu\right\},$$

where $\mu = c'\tau\sigma_x\sqrt{\frac{\log p}{n}}$. By a union bound, we then have

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{m=1}^{M} \mathbb{P}\left(\exists(\beta, \Delta) \in \mathbb{V}_m \text{ s.t. } f(\beta, \Delta; X) \geq 2g(h(\beta, \Delta))\right),$$

where the index m ranges up to $M := \left\lceil \log\left(c\sqrt{\frac{n}{\log p}}\right)\right\rceil$ over the relevant region (B.56). By the definition (B.54) of f, we have

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{m=1}^{M} \mathbb{P}\left(\sup_{h(\beta,\Delta)\leq g^{-1}(2^m\mu)} f(\beta, \Delta; X) \geq 2^m\mu\right) \overset{(i)}{\leq}\ M \cdot 2\exp(-c_2 n),$$

where inequality (i) applies the tail bound (B.55). It follows that

$$\mathbb{P}(\mathcal{E}^c) \leq c_1 \exp\left(-c_2 n + \log\log\left(\frac{n}{\log p}\right)\right)\ \leq c_1' \exp\left(-c_2' n\right).$$

Multiplying through by $\|\Delta\|_2^2$ then yields the desired result.

B.5 Auxiliary results

In this section, we provide some auxiliary results that are useful for our proofs. The first lemma concerns symmetrization and desymmetrization of empirical processes via Rademacher random variables:

Lemma B.9 (Lemma 2.3.6 in van der Vaart and Wellner [87]). *Let $\{Z_i\}_{i=1}^n$ be independent zero-mean stochastic processes. Then*

$$\frac{1}{2}\mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^n \epsilon_i Z_i(t_i)\right|\right] \leq \mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^n Z_i(t_i)\right|\right] \leq 2\mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^n \epsilon_i(Z_i(t_i) - \mu_i)\right|\right],$$

where the ϵ_i's are independent Rademacher variables and the functions $\mu_i : \mathcal{F} \to \mathbb{R}$ are arbitrary.

We also have a useful lemma that bounds the Gaussian complexity in terms of the Rademacher complexity:

Lemma B.10 (Lemma 4.5 in Ledoux and Talagrand [49]). *Let $Z_1, \ldots, Z_n$ be independent stochastic processes. Then*

$$\mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^{n}\epsilon_i Z_i(t_i)\right|\right] \leq \sqrt{\frac{\pi}{2}}\cdot\mathbb{E}\left[\sup_{t\in T}\left|\sum_{i=1}^{n} g_i Z_i(t_i)\right|\right],$$

where the ϵ_i's are Rademacher variables and the g_i's are standard normal.

We next state a version of the Sudakov-Fernique comparison inequality:

Lemma B.11 (Corollary 3.14 in Ledoux and Talagrand [49]). *Given a countable index set T, let $X(t)$ and $Y(t)$ be centered Gaussian processes such that*

$$\operatorname{var}\left(Y(s)-Y(t)\right) \leq \operatorname{var}\left(X(s)-X(t)\right), \qquad \forall (s,t)\in T\times T.$$

Then

$$\mathbb{E}\left[\sup_{t\in T} Y(t)\right] \leq 2\cdot\mathbb{E}\left[\sup_{\in T} X(t)\right].$$

A zero-mean random variable Z is sub-Gaussian with parameter σ if $\mathbb{P}(Z > t) \leq \exp(-\frac{t^2}{2\sigma^2})$ for all $t \geq 0$. The next lemma provides a standard bound on the expected maximum of N such variables (cf. equation (3.6) in Ledoux and Talagrand [49]):

Lemma B.12. *Suppose $X_1, \ldots, X_N$ are zero-mean sub-Gaussian random variables such that $\max_{j=1,\ldots,N}\|X_j\|_{\psi_2} \leq \sigma$. Then $\mathbb{E}\left[\max_{j=1,\ldots,p}|X_j|\right] \leq c_0\,\sigma\sqrt{\log N}$, where $c_0 > 0$ is a universal constant.*

We also have a lemma about maxima of products of sub-Gaussian variables:

Lemma B.13. *Suppose $\{g_i\}_{i=1}^n$ are i.i.d. standard Gaussians and $\{X_i\}_{i=1}^n \subseteq \mathbb{R}^p$ are i.i.d. sub-Gaussian vectors with parameter bounded by σ_x. Then as long as $n \geq c\sqrt{\log p}$ for some constant $c > 0$, we have*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} g_i X_i\right\|_\infty\right] \leq c'\sigma_x\sqrt{\frac{\log p}{n}}.$$

Proof. Conditioned on $\{X_i\}_{i=1}^n$, for each $j = 1, \ldots, p$, the variable $\left|\frac{1}{n}\sum_{i=1}^n g_i X_{ij}\right|$ is zero-mean and sub-Gaussian with parameter bounded by $\frac{\sigma_x}{n}\sqrt{\sum_{i=1}^n X_{ij}^2}$. Hence, by Lemma B.12, we have

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} g_i X_i\right\|_\infty \middle| X\right] \leq \frac{c_0\sigma_x}{n}\cdot\max_{j=1,\ldots,p}\sqrt{\sum_{i=1}^{n} X_{ij}^2}\cdot\sqrt{\log p},$$

implying that

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} g_i X_i\right\|_\infty\right] \leq c_0\sigma_x\sqrt{\frac{\log p}{n}} \cdot \mathbb{E}\left[\max_j \sqrt{\frac{\sum_{i=1}^n X_{ij}^2}{n}}\right]. \tag{B.57}$$

Furthermore, $Z_j := \frac{\sum_{i=1}^n X_{ij}^2}{n}$ is an i.i.d. average of subexponential variables, each with parameter bounded by $c\sigma_x$. Since $\mathbb{E}[Z_j] \leq 2\sigma_x^2$, we have

$$\mathbb{P}\left(Z_j - \mathbb{E}[Z_j] \geq u + 2\sigma_x^2\right) \leq c_1 \exp\left(-\frac{c_2 n u}{\sigma_x}\right), \qquad \forall u \geq 0 \text{ and } 1 \leq j \leq p. \tag{B.58}$$

Now fix some $t \geq \sqrt{2\sigma_x^2}$. Since the $\{Z_j\}_{j=1}^p$ are all nonnegative, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{j=1,\ldots,p}\sqrt{Z_j}\right] &\leq t + \int_t^\infty \mathbb{P}\left(\max_{j=1,\ldots,p}\sqrt{Z_j} > s\right) ds \\
&\leq t + \sum_{j=1}^p \int_t^\infty \mathbb{P}\left(\sqrt{Z_j} > s\right) ds \\
&\leq t + c_1 p \int_t^\infty \exp\left(-\frac{c_2 n(s^2 - 2\sigma_x^2)}{\sigma_x}\right) ds
\end{aligned}$$

where the final inequality follows from the bound (B.58) with $u = s^2 - 2\sigma_x^2$, valid as long as $s^2 \geq t^2 \geq 2\sigma_x^2$. Integrating, we have the bound

$$\mathbb{E}\left[\max_{j=1,\ldots,p}\sqrt{Z_j}\right] \leq t + c_1' p\sigma_x \exp\left(-\frac{c_2' n(t^2 - 2\sigma_x^2)}{\sigma_x^2}\right).$$

Since $n \gtrsim \sqrt{\log p}$ by assumption, setting t equal to a constant implies $\mathbb{E}\left[\max_j \sqrt{Z_j}\right] = \mathcal{O}(1)$, which combined with inequality (B.57) gives the desired result. □

B.6 Capped-ℓ_1 penalty

In this section, we show how our results on nonconvex but subdifferentiable regularizers may be extended to include certain types of more complicated regularizers that do not possess (sub)gradients everywhere, such as the capped-ℓ_1 penalty.

In order to handle the case when ρ_λ has points where neither a gradient nor subderivative exists, we assume the existence of a function $\widetilde{\rho}_\lambda$ (possibly defined according to the particular local optimum $\widetilde{\beta}$ of interest), such that the following conditions hold:

Assumption B.1.

(i) The function $\widetilde{\rho}_\lambda$ is differentiable/subdifferentiable everywhere, and $\|\nabla\widetilde{\rho}_\lambda(\widetilde{\beta})\|_\infty \leq \lambda L$.

(ii) *For all* $\beta \in \mathbb{R}^p$, *we have* $\widetilde{\rho}_\lambda(\beta) \geq \rho_\lambda(\beta)$.

(iii) *The equality* $\widetilde{\rho}_\lambda(\widetilde{\beta}) = \rho_\lambda(\widetilde{\beta})$ *holds.*

(iv) *There exists* $\mu_1 \geq 0$ *such that* $\widetilde{\rho}_\lambda(\beta) + \mu_1\|\beta\|_2^2$ *is convex.*

(v) *For some index set* A *with* $|A| \leq k$ *and some parameter* $\mu_2 \geq 0$, *we have*

$$\widetilde{\rho}_\lambda(\beta^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}) \leq \lambda L\|\widetilde{\beta}_A - \beta^*_A\|_1 - \lambda L\|\widetilde{\beta}_{A^c} - \beta^*_{A^c}\|_1 + \mu_2\|\widetilde{\beta} - \beta^*\|_2^2.$$

In addition, we assume conditions (i)–(iii) of Assumption 4.1 in Section 4.2.2 above.

Remark B.1. *When* $\rho_\lambda(\beta) + \mu_1\|\beta\|_2^2$ *is convex for some* $\mu_1 \geq 0$ *(as in the case of SCAD or MCP), we may take* $\widetilde{\rho}_\lambda = \rho_\lambda$ *and* $\mu_2 = 0$. *(See Lemma B.2 in Appendix B.1.1.) When no such convexification of* ρ_λ *exists (as in the case of the capped-*ℓ_1 *penalty), we instead construct a separate convex function* $\widetilde{\rho}_\lambda$ *to upper-bound* ρ_λ *and take* $\mu_1 = 0$.

Under the conditions of Assumption B.1, we have the following variation of Theorem 4.1:

Theorem B.1. *Suppose* $\mathcal{L}_n$ *satisfies the RSC conditions* (4.4), *and the functions* ρ_λ *and* $\widetilde{\rho}_\lambda$ *satisfy Assumption 4.1 and Assumption B.1, respectively. With* λ *is chosen according to the bound* (4.6) *and* $n \geq \frac{16R^2\max(\tau_1^2,\tau_2^2)}{\alpha_2^2}\log p$, *we have*

$$\|\widetilde{\beta} - \beta^*\|_2 \leq \frac{7\lambda L\sqrt{k}}{4(\alpha_1 - \mu_1 - \mu_2)}, \qquad \textit{and} \qquad \|\widetilde{\beta} - \beta^*\|_1 \leq \frac{56\lambda Lk}{4(\alpha_1 - \mu_1 - \mu_2)},$$

along with the prediction error bound

$$\langle \nabla\mathcal{L}_n(\widetilde{\beta}) - \nabla\mathcal{L}_n(\beta^*), \widetilde{\nu}\rangle \leq \lambda^2 L^2 k\left(\frac{21}{8(\alpha_1 - \mu_1 - \mu_2)} + \frac{49(\mu_1 + \mu_2)}{16(\alpha_1 - \mu_1 - \mu_2)^2}\right).$$

Proof. The proof is essentially the same as the proofs of Theorems 4.1 and 4.2, so we only mention a few key modifications here. First note that any local minimum $\widetilde{\beta}$ of the program (4.1) is a local minimum of $\mathcal{L}_n + \widetilde{\rho}_\lambda$, since

$$\mathcal{L}_n(\widetilde{\beta}) + \widetilde{\rho}_\lambda(\widetilde{\beta}) = \mathcal{L}_n(\widetilde{\beta}) + \rho_\lambda(\widetilde{\beta}) \leq \mathcal{L}_n(\beta) + \rho_\lambda(\beta) \leq \mathcal{L}_n(\beta) + \widetilde{\rho}_\lambda(\beta),$$

locally for all β in the constraint set, where the first inequality comes from the fact that $\widetilde{\beta}$ is a local minimum of $\mathcal{L}_n + \rho_\lambda$, and the second inequality holds because $\widetilde{\rho}_\lambda$ upper-bounds ρ_λ. Hence, the first-order condition (4.5) still holds with ρ_λ replaced by $\widetilde{\rho}_\lambda$. Consequently, inequality (4.19) holds, as well.

Next, note that inequality (4.21) holds as before, with ρ_λ replaced by $\widetilde{\rho}_\lambda$ and μ replaced by μ_1. By condition (v) on $\widetilde{\rho}_\lambda$, we then have inequality (4.22) with μ replaced by $\mu_1 + \mu_2$. The remainder of the proof is exactly as before. □

Specializing now to the case of the capped-ℓ_1 penalty, we have the following lemma. For a fixed parameter $c \geq 1$, the capped-ℓ_1 penalty [103] is given by

$$\rho_\lambda(t) := \min\left\{\frac{\lambda^2 c}{2},\ \lambda|t|\right\}. \tag{B.59}$$

Lemma B.14. *The capped-ℓ_1 regularizer (B.59) with parameter c satisfies the conditions of Assumption B.1, with $\mu_1 = 0$, $\mu_2 = \frac{1}{c}$, and $L = 1$.*

Proof. We will show how to construct an appropriate choice of $\widetilde{\rho}_\lambda$. Note that ρ_λ is piecewise linear and locally equal to $|t|$ in the range $\left[-\frac{\lambda c}{2}, \frac{\lambda c}{2}\right]$, and takes on a constant value outside that region. However, ρ_λ does not have either a gradient or subgradient at $t = \pm\frac{\lambda c}{2}$, hence is not "convexifiable" by adding a squared-ℓ_2 term.

We begin by defining the function $\widetilde{\rho} : \mathbb{R} \to \mathbb{R}$ via

$$\widetilde{\rho}_\lambda(t) = \begin{cases} \lambda|t|, & \text{if} \quad |t| \leq \frac{\lambda c}{2}, \\ \frac{\lambda^2 c}{2}, & \text{if} \quad |t| > \frac{\lambda c}{2}. \end{cases}$$

For a fixed local optimum $\widetilde{\beta}$, note that we have $\widetilde{\rho}_\lambda(\beta) = \sum_{j \in T} \lambda|\widetilde{\beta}_j| + \sum_{j \in T^c} \frac{\lambda^2 c}{2}$, where $T := \left\{j \mid |\widetilde{\beta}_j| \leq \frac{\lambda c}{2}\right\}$. Clearly, $\widetilde{\rho}_\lambda$ is a convex upper bound on ρ_λ, with $\widetilde{\rho}_\lambda(\widetilde{\beta}) = \rho_\lambda(\widetilde{\beta})$. Furthermore, by the convexity of $\widetilde{\rho}_\lambda$, we have

$$\langle \nabla\widetilde{\rho}_\lambda(\widetilde{\beta}),\, \beta^* - \widetilde{\beta}\rangle \leq \widetilde{\rho}_\lambda(\beta^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}) \ = \ \sum_{j \in S}\Big(\widetilde{\rho}_\lambda(\beta_j^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}_j)\Big) - \sum_{j \notin S}\widetilde{\rho}_\lambda(\widetilde{\beta}_j), \tag{B.60}$$

using decomposability of $\widetilde{\rho}$. For $j \in T$, we have

$$\widetilde{\rho}_\lambda(\beta_j^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}_j) \leq \lambda|\beta_j^*| - \lambda|\widetilde{\beta}_j| \leq \lambda|\widetilde{\nu}_j|,$$

whereas for $j \notin T$, we have $\widetilde{\rho}_\lambda(\beta_j^*) - \widetilde{\rho}_\lambda(\widetilde{\beta}_j) = 0 \leq \lambda|\widetilde{\nu}_j|$. Combined with the bound (B.60), we obtain

$$\begin{aligned} \langle \nabla\widetilde{\rho}_\lambda(\widetilde{\beta}),\, \beta^* - \widetilde{\beta}\rangle &\leq \sum_{j \in S}\lambda|\widetilde{\nu}_j| - \sum_{j \notin S}\widetilde{\rho}_\lambda(\widetilde{\beta}_j) \\ &= \lambda\|\widetilde{\nu}_S\|_1 - \sum_{j \notin S}\rho_\lambda(\widetilde{\beta}_j) \\ &= \lambda\|\widetilde{\nu}_S\|_1 - \lambda\|\widetilde{\nu}_{S^c}\|_1 + \sum_{j \notin S}\Big(\lambda|\widetilde{\beta}_j| - \rho_\lambda(\widetilde{\beta}_j)\Big). \end{aligned} \tag{B.61}$$

Now observe that

$$\lambda|t| - \rho_\lambda(t) = \begin{cases} 0, & \text{if} \quad |t| \leq \frac{\lambda c}{2}, \\ \lambda|t| - \frac{\lambda^2 c}{2}, & \text{if} \quad |t| > \frac{\lambda c}{2}, \end{cases}$$

and moreover, the derivative of $\frac{t^2}{c}$ always exceeds λ for $|t| > \frac{\lambda c}{2}$. Consequently, we have $\lambda|t| - \rho_\lambda(t) \leq \frac{t^2}{c}$ for all $t \in \mathbb{R}$. Substituting this bound into inequality (B.61) yields

$$\langle \nabla \widetilde{\rho}_\lambda(\widetilde{\beta}),\, \beta^* - \widetilde{\beta} \rangle \leq \lambda \|\widetilde{\nu}_S\|_1 - \lambda \|\widetilde{\nu}_{S^c}\|_1 + \frac{1}{c} \|\widetilde{\nu}_{S^c}\|_2^2,$$

which is condition (v) of Assumption B.1 on $\widetilde{\rho}_\lambda$ with $L = 1$, $A = S$, and $\mu_2 = \frac{1}{c}$. The remaining conditions are easy to verify (see also Zhang and Zhang [103]). □

Appendix C

Proofs for Chapter 5

C.1 Proofs of supporting lemmas for Theorem 5.1

In this section, we supply the proofs of Lemmas 5.1 and 5.2, which are used in the proof of Theorem 5.1.

C.1.1 Proof of Lemma 5.1

By Proposition B.2 of Wainwright and Jordan [91] (cf. Theorems 23.5 and 26.3 of Rockafellar [75]), we know that the dual function Φ^* is differentiable on the interior of the marginal polytope $\mathcal{M}$ defined in equation (5.11), in particular with

$$\nabla\Phi^*(\mu) = (\nabla\Phi)^{-1}(\mu) \quad \text{for all } \mu \in \text{int}(\mathcal{M}). \tag{C.1}$$

Also, by Theorem 3.4 of Wainwright and Jordan [91], for any $\mu \in \text{int}(\mathcal{M})$, the negative dual function takes the form $\Phi^*(\mu) = -H(q_\theta(\mu))$, where $\theta(\mu) = (\nabla\Phi)^{-1}(\mu)$.

By relation (C.1), we have

$$(\nabla\Phi)(\nabla\Phi^*(\mu)) = \mu \qquad \text{for all } \mu \in \mathcal{M}.$$

Since this equation holds on an open set, we may take derivatives; employing the chain rule yields

$$(\nabla^2\Phi)(\nabla\Phi^*(\mu)) \cdot (\nabla^2\Phi^*(\mu)) = I_{D\times D}.$$

Rearranging yields the relation $\nabla^2\Phi^*(\mu) = (\nabla^2\Phi(\theta))^{-1}\,|_{\theta=\theta(\mu)}$, as claimed.

C.1.2 Proof of Lemma 5.2

We induct on the subset size. For sets of size 1, the claim is obvious. Now suppose the claim holds for all subsets up to some size $k > 1$, and consider a subset of size $k+1$, which we write as $C = \{1, \ldots, k+1\}$, without loss of generality. For any configuration $J \in \mathcal{X}_0^{|C|}$, the

marginal probability $q_C(x_C = J)$ is equal to $\mu_{C;J}$, by construction. Consequently, we need only specify how to determine the probabilities $q_C(x_C = J)$ for configurations $J \in \mathcal{X}^{|C|} \backslash \mathcal{X}_0^{|C|}$. By the definition of $\mathcal{X}_0^{|C|}$, each $j \in J$ has $j_s = 0$ for at least one $s \in \{1, \ldots, k+1\}$.

We show how to express the remaining marginal probabilities sequentially, inducting on the number of positions s for which $j_s = 0$. Starting with the base case in which there is a single zero, suppose without loss of generality that $j_{k+1} = 0$. For each $\ell \in \{1, 2, \ldots, m-1\}$, let J^ℓ be the configuration such that $J_i^\ell = J_i$ for all $i \neq k+1$ and $J_{k+1}^\ell = \ell$. Defining $D := C \backslash \{k+1\}$, we then have

$$q_C(x_C = J) = q_D(x_D = J') - \sum_{\ell=1}^{m-1} q_C(x_C = J^\ell), \tag{C.2}$$

where $J' \in \mathcal{X}^k$ is the configuration defined by $J'_i = J_i$ for all $i = 1, 2, \ldots, k$. Since $|D| = k$, our induction hypothesis implies that $q_D(x_D = J')$ is a linear function of the specified mean parameters. Moreover, our starting assumption implies that $J^\ell \in \mathcal{X}_0^{|C|}$ for all indices $\ell = \{1, 2, \ldots, m-1\}$, so we have $q_C(x_C = J^\ell) = \mu_{C;J^\ell}$. This establishes the base case.

Now suppose the sub-claim holds for all configurations with at most t nonzeros, for some $t > 1$. Consider a configuration J with $t+1$ zero entries. Again without loss of generality, we may assume $j_{k+1} = 0$, so equation (C.2) may be derived as before. This time, the configurations J^ℓ are not in $\mathcal{X}_0^{|C|}$ (since they still have $t \geq 1$ zero entries); however, our induction hypothesis implies that the corresponding probabilities may be written as functions of the given mean parameters. This completes the inductive proof of the inner claim, thereby completing the outer induction, as well.

C.2 Proofs of population-level corollaries

In this Appendix, we prove Corollaries 5.1 and 5.3. (As previously noted, Corollary 5.2 is an immediate consequence of Corollary 5.1.)

C.2.1 Proof of Corollary 5.1

Recall that $\widetilde{\mathcal{C}}$ denotes the set of all cliques in the triangulation $\widetilde{G}$. The covariance matrix in Theorem 5.1 is indexed by $\widetilde{\mathcal{C}}$, and our goal is to define appropriate blocks of the matrix and then apply the matrix inversion lemma [35]. Consider the collection $\text{pow}(\mathcal{S})$. We define the collection of singleton subsets $V = \{\{1\}, \{2\}, \ldots, \{p\}\}$, and introduce the disjoint partition

$$\widetilde{\mathcal{C}} = \underbrace{\Big(\text{pow}(\mathcal{S}) \cup V\Big)}_{\mathcal{U}} \cup \underbrace{\Big(\widetilde{\mathcal{C}} \backslash \big\{\text{pow}(\mathcal{S}) \cup V\big\}\Big)}_{\mathcal{W}}.$$

The following property of the collection $\mathcal{W}$ is important:

Lemma C.1. *For each maximal clique $C \in \bar{\mathcal{C}}$, define the set collection $\mathcal{F}(C) = pow(C)\backslash\mathcal{U}$. For any $A \in \mathcal{W}$, we have $A \in \mathcal{F}(C)$ for exactly one C.*

Proof. We first establish existence. Since $\mathcal{W} \subseteq \widetilde{\mathcal{C}}$, any set $A \in \mathcal{W}$ is contained in some maximal clique C_A. Since $A \notin \mathcal{U}$, we clearly have $A \in \mathcal{F}(C_A)$.

To establish uniqueness, consider a set A belonging to the intersection $C_1 \cap C_2$ of two maximal cliques. If these cliques are adjacent in the junction tree, then A belongs to the separator set $C_1 \cap C_2$, so A cannot belong to $\mathcal{W}$, by definition. Even when C_1 and C_2 are not adjacent, the running intersection property of the junction tree implies that $C_1 \cap C_2$ must belong to every separator set on the unique path between C_1 and C_2 in the junction tree, implying that $A \notin \mathcal{W}$, as before. This is a contradiction, implying that the maximal clique C_A is unique. □

Define $\Gamma = \big(\operatorname{cov}(\Psi(X;\widetilde{\mathcal{C}}))\big)^{-1}$. By the block-matrix inversion formula [35], we may write

$$\Theta := (\operatorname{cov}\big(\Psi(X;\mathcal{U})\big))^{-1} = \Gamma(\mathcal{U},\mathcal{U}) - \Gamma(\mathcal{U},\mathcal{W})\big(\Gamma(\mathcal{W},\mathcal{W})\big)^{-1}\Gamma(\mathcal{W},\mathcal{U}). \tag{C.3}$$

We need to show that $\Theta(A,B) = 0$ for any members $A, B \in \mathcal{U}$ that do not belong to the same maximal clique. By Theorem 5.1(a), we have $\Gamma(A,B) = 0$ whenever A and B do not belong to the same maximal clique, so it remains to show that $\Gamma(A,\mathcal{W})\big(\Gamma(\mathcal{W},\mathcal{W})\big)^{-1}\Gamma(\mathcal{W},B) = 0$.

We begin by observing that the matrix $\Gamma(\mathcal{W},\mathcal{W})$ is block-diagonal with respect to the partition $\{\mathcal{F}(C) : C \in \mathcal{C}\}$ previously defined in Lemma C.1. (Indeed, consider two sets $D, E \in \mathcal{W}$ with $D \in \mathcal{F}(C)$ and $E \in \mathcal{F}(C')$ for distinct maximal cliques $C \neq C'$. Two such sets cannot belong to the same maximal clique, so Theorem 5.1(a) implies that $\Gamma(D,E) = 0$.) Since block-diagonal structure is preserved by matrix inversion, the inverse $\Upsilon = (\Gamma(\mathcal{W},\mathcal{W}))^{-1}$ shares this property, so for any two members $A, B \in \mathcal{U}$, we may write

$$\begin{aligned}\Gamma(A,\mathcal{W})\big(\Gamma(\mathcal{W},\mathcal{W})\big)^{-1}\Gamma(\mathcal{W},B) \\ = \sum_{\mathcal{F}(C), C\in\bar{\mathcal{C}}} \Gamma(A,\mathcal{F}(C))\Upsilon(\mathcal{F}(C),\mathcal{F}(C))\Gamma(\mathcal{F}(C),B).\end{aligned} \tag{C.4}$$

We claim that each of these terms vanishes. For a given maximal clique C', suppose A is not contained within C'; we first claim that $\Gamma(A,\mathcal{F}(C')) = 0$, or equivalently, for any set $D \in \mathcal{F}(C')$, we have $\Gamma(A,D) = 0$. From Theorem 5.1(a), it suffices to show that A and D cannot be contained within the same maximal clique. From Lemma C.1, we know that A belongs to a unique maximal clique C. Any set $D \in \mathcal{F}(C')$ is contained within C'; if it were also contained within C, then D would be contained in $C \cap C'$. But as argued in the proof of Lemma C.1, this implies that D is contained within some separator set, whence it cannot belong to $\mathcal{F}(C')$. We thus conclude that $\Gamma(A,D) = 0$, as claimed.

Taking any two subsets A and B that are not contained in the same maximal clique, we see that for any clique C, we must either have $\Gamma(A,\mathcal{F}(C)) = 0$ or $\Gamma(\mathcal{F}(C),B) = 0$. Hence, each term in the sum (C.4) indeed vanishes, completing the proof.

C.2.2 Proof of Corollary 5.3

This corollary follows by a similar argument as in the proof of Corollary 5.1. As before, let $\mathcal{C}$ denote the set of all cliques in the triangulation $\widetilde{G}$, and let $V = \{\{1\}, \{2\}, \ldots, \{p\}\}$. Define $\mathcal{U} = \text{pow}(\mathcal{S}(s;d)) \cup V$ and $\mathcal{W} = \widetilde{\mathcal{C}} \backslash \mathcal{U}$.

Let $C_s := s \cup N(s)$, and consider a disjoint partition of $\mathcal{W}$ defined by $\mathcal{F}_1 := \text{pow}(C_s) \backslash \mathcal{U}$ and $\mathcal{F}_2 := \mathcal{W} \backslash \mathcal{F}_1$. Note that C_s is the unique maximal clique in $\widetilde{\mathcal{C}}$ containing s. By construction, every clique in $\mathcal{F}_2$ does not contain s and has more than d elements, whereas every clique in $\mathcal{F}_1$ is contained in C_s, with $|C_s| \leq d+1$. It follows that no two cliques $A \in \mathcal{F}_1$ and $B \in \mathcal{F}_2$ can be contained in the same maximal clique. Denoting $\Gamma := (\text{cov}(\Psi(X; \widetilde{\mathcal{C}})))^{-1}$, we conclude via Theorem 5.1(a) that $\Gamma(\mathcal{W}, \mathcal{W})$ is block-diagonal.

We now use the block matrix-equation formula (C.3). As before, Theorem 5.1(a) implies that $\Gamma(\mathcal{U}, \mathcal{U})$ is graph-structured according to $\widetilde{G}$. In particular, for any $B \in \mathcal{U}$ with $B \subsetneq C_s$, we have $\Gamma(\{s\}, B) = 0$. (The elements of $\mathcal{U}$ that are subsets of C_s are exactly $\{s\}$ and the nonempty subsets of $N(s)$.) Hence, it remains to show that

$$\Gamma(\{s\}, \mathcal{W})(\Gamma(\mathcal{W}, \mathcal{W}))^{-1}\Gamma(\mathcal{W}, B) = 0.$$

Analogous to equation (C.4), we may write

$$\Gamma(\{s\}, \mathcal{W})(\Gamma(\mathcal{W}, \mathcal{W}))^{-1}\Gamma(\mathcal{W}, B) = \sum_{i=1}^{2} \Gamma(\{s\}, \mathcal{F}_i)\Upsilon(\mathcal{F}_i, \mathcal{F}_i)\Gamma(\mathcal{F}_i, B),$$

where $\Upsilon := (\Gamma(\mathcal{W}, \mathcal{W}))^{-1}$. Applying Theorem 5.1(a) once more, we see that $\Gamma(\mathcal{F}_1, B) = 0$, since $B \subsetneq C_s$ and $\Gamma(\{s\}, \mathcal{F}_2) = 0$. Therefore, the matrix $\Theta = (\text{cov}(\Psi(X; \mathcal{U})))^{-1}$ appearing in equation (C.3) is indeed s-block graph-structured.

C.3 Proof of Proposition 5.1

In this section, we provide a proof of our main nodewise recovery result, Proposition 5.1. For proofs of supporting technical lemmas and all corollaries appearing in the text, see Appendix C.4.

C.3.1 Main argument

We derive Proposition 5.1 as a consequence of a more general theorem. Suppose we have i.i.d. observations $\{(x_i, y_i)\}_{i=1}^n$, with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, and we wish to estimate the best linear predictor $\widetilde{\beta} = \Sigma_x^{-1} \text{Cov}(x_i, y_i)$, when $\widetilde{\beta}$ is k-sparse. In Chapter 3, we analyze a modified version of the Lasso based on possibly corrupted observations; however, we assume the linear regression model

$$y_i = x_i^T \widetilde{\beta} + \epsilon_i, \tag{C.5}$$

where ϵ_i is sub-Gaussian noise and $\epsilon_i \perp\!\!\!\perp x_i$. Although the model (4.10) holds in the case where y_i is a sample from a single node and x_i is a sample from all other nodes in a Gaussian graphical model, the model (C.5) does *not* hold in a general discrete graphical model. Nonetheless, we show that essentially the same Lasso estimator provides an estimator for $\widetilde{\beta}$ that is consistent for support recovery. Suppose the pair $(\widehat{\Gamma}, \widehat{\gamma})$ in the Lasso program (5.22) satisfies the following deviation bounds:

$$\|\widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma}\|_\infty \leq \varphi_1 \sqrt{\frac{\log p}{n}}, \tag{C.6a}$$

$$\|(\widehat{\Gamma} - \Sigma_x)v\|_\infty \leq \varphi_2 \|v\|_\infty \sqrt{\frac{k \log p}{n}} \quad \forall v \in \mathbb{B}_1(8k) \cap \mathbb{B}_\infty(1), \tag{C.6b}$$

for some φ_1, φ_2. Also suppose $\widehat{\Gamma}$ satisfies the lower-restricted eigenvalue (RE) condition:

$$v^T \widehat{\Gamma} v \geq \alpha \|v\|_2^2 \quad \forall v \text{ s.t. } \|v\|_1 \leq \sqrt{k}\|v\|_2. \tag{C.7}$$

Then we have the following technical result:

Theorem C.1. *Suppose the pair $(\widehat{\Gamma}, \widehat{\gamma})$ satisfies the deviation conditions (C.6a) and (C.6b), as well as the lower-RE condition (C.7). Also suppose the sample size satisfies the scaling $n \gtrsim \max\left\{\frac{\varphi_1^2}{\alpha^2(b_0 - \|\widetilde{\beta}\|_2)^2}, \varphi_2^2 \left|\!\left|\!\left|\Sigma_x^{-1}\right|\!\right|\!\right|_\infty^2\right\} k \log p$ and $\lambda_n \gtrsim \varphi_1 \sqrt{\frac{\log p}{n}}$. Then any optimum $\widehat{\beta}$ of the Lasso program (5.22) satisfies*

$$\|\widehat{\beta} - \widetilde{\beta}\|_\infty \leq 4\lambda_n \left|\!\left|\!\left|\Sigma_x^{-1}\right|\!\right|\!\right|_\infty.$$

The proof of Theorem C.1 is provided in Appendix C.3.2. In order to prove Proposition 5.1, we first establish that the deviation conditions (C.6a) and (C.6b) of Theorem C.1 hold w.h.p. with $(\varphi_1, \varphi_2) = (\varphi\|\widetilde{\beta}\|_2, \varphi)$, and the lower-RE condition holds with $\alpha = \frac{1}{2}\lambda_{\min}(\Sigma_x)$.

Note that

$$\|\widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma}\|_\infty \leq \|(\widehat{\Gamma} - \Sigma_x)\widetilde{\beta}\|_\infty + \|\operatorname{Cov}(x_i, y_i) - \widehat{\gamma}\|_\infty. \tag{C.8}$$

Furthermore,

$$\|(\widehat{\Gamma} - \Sigma_x)\widetilde{\beta}\|_\infty \leq \left\|\left(\frac{X^T X}{n} - \mathbb{E}(x_i x_i^T)\right)\widetilde{\beta}\right\|_\infty + \|(\overline{x}\overline{x}^T - \Sigma_x)\widetilde{\beta}\|_\infty$$

and

$$\|\operatorname{Cov}(x_i, y_i) - \widehat{\gamma}\|_\infty \leq \left\|\frac{X^T y}{n} - \mathbb{E}(y_i x_i)\right\|_\infty + \|\bar{y}\overline{x} - \mathbb{E}(y_i)\mathbb{E}(x_i)\|_\infty.$$

As in the analysis of inequality (C.17) below, we may disregard the two second terms involving empirical means, since they concentrate at a fast rate. Since $x_i^T \widetilde{\beta}$ is sub-Gaussian with

parameter $\varphi^2\|\widetilde{\beta}\|_2^2$ by assumption, and $e_j^T x_i$ and y_i are clearly sub-Gaussian with parameter 1, the deviation condition (C.6a) follows with $\varphi_1 = \varphi\|\widetilde{\beta}\|_2$ by standard concentration bounds on an i.i.d. average of products of sub-Gaussians (cf. Lemma A.14 in Appendix A.3).

For the second deviation bound, we will verify the bound over a more tractable set via the following lemma:

Lemma C.2. *For any constant $c_0 > 0$, we have*

$$\mathbb{B}_1(c_0 k) \cap \mathbb{B}_\infty(1) \subseteq (1 + c_0)\,\mathrm{cl}\{\mathrm{conv}\{\mathbb{B}_0(k) \cap \mathbb{B}_\infty(1)\}\}.$$

Hence, it suffices to establish the deviation inequality (C.6b) over the set $\mathbb{B}_0(k) \cap \mathbb{B}_\infty(1)$. We proceed via a discretization argument. Suppose $\{v_1, \ldots, v_M\}$ is a $\frac{1}{2}$-covering of the unit ℓ_∞-ball in $\mathbb{R}^k$ in its own metric. By standard results on metric entropy, we know that such a covering exists with $M \leq c^k$. Writing $\psi(v) = \|(\widehat{\Gamma} - \Sigma_x)v\|_\infty$, we know that there exists v_j such that $\|v - v_j\|_\infty \leq \frac{1}{2}$. Let $\Delta v = v - v_j$. Then

$$\psi(v) = \|(\widehat{\Gamma} - \Sigma_x)(v_j + \Delta v)\|_\infty \leq \psi(v_j) + \psi(\Delta v) \leq \sup_{1 \leq j \leq M} \psi(v_j) + \frac{1}{2} \sup_{\|v\|_\infty \leq 1} \psi(v),$$

simply by rescaling. Taking the sup over $\{\|v\|_\infty \leq 1\}$ on the LHS and rearranging then yields

$$\sup_{\|v\|_\infty \leq 1} \leq 2 \sup_{1 \leq j \leq M} \psi(v_j).$$

Hence, it suffices to establish the bound for a given $v \in \mathbb{B}_1(c_0 k) \cap \mathbb{B}_\infty(1)$, then take a union bound over the $M \leq c^k$ elements in the discretization and the $\binom{p}{k} \leq p^k$ choices of the support set.

For a given k-sparse v, note that $x_i^T v$ has sub-Gaussian parameter $\varphi^2\|v\|_2^2$ by assumption, and

$$\|v\|_2^2 \leq \|v\|_1\|v\|_\infty \leq \sqrt{k}\|v\|_2\|v\|_\infty,$$

so $x_i^T v$ is sub-Gaussian with parameter $\varphi^2 k\|v\|_\infty^2$. Since $e_\ell^T x_i$ is sub-Gaussian with parameter 1, it follows from the same recentering techniques as in inequality (C.17) that

$$\|(\widehat{\Gamma} - \Sigma_x)v\|_\infty = \max_\ell |e_\ell^T(\widehat{\Gamma} - \Sigma_x)v| \leq t,$$

with probability at least $1 - c_1 \exp\left(\frac{-c_2 n t^2}{\varphi^2 k\|v\|_\infty^2}\right)$. Taking a union bound over the discretization and setting $t = c\varphi\sqrt{k}\|v\|_\infty\sqrt{\frac{k \log p}{n}}$ then implies the deviation bound (C.6b) with $\varphi_2 = \varphi$, under the scaling $n \gtrsim \varphi^2 k^2 \log p$.

The lower-RE condition (C.7) may be verified analogously to the results in Appendix A.2. The only difference is to use the fact that $x_i^T v$ is sub-Gaussian with parameter $\varphi^2\|v\|_2^2$ in all the deviation bounds. Then the lower-RE condition holds with probability at least $1 - c_1 \exp(-c_2 k \log p)$, under the scaling $n \gtrsim \varphi^2 k \log p$.

We may take $\lambda_n \asymp \varphi \|\widetilde{\beta}\|_2 \sqrt{\frac{\log p}{n}}$ in Theorem C.1 to conclude that w.h.p.,

$$\|\widehat{\beta} - \widetilde{\beta}\|_\infty \precsim \varphi |\widetilde{\beta}\|_2 \sqrt{\frac{\log p}{n}}.$$

Finally, note that the vector $\widetilde{\beta}$ is a scalar multiple of column s of the inverse covariance matrix Γ, as a straightforward consequence of block matrix inversion. Hence, combining Corollary 5.1 and Theorem C.1 implies that thresholding succeeds w.h.p. for neighborhood recovery in a tree graph.

C.3.2 Proof of Theorem C.1

We begin by establishing ℓ_1- and ℓ_2- error bounds, which will be used in the sequel:

Lemma C.3. *Suppose the deviation condition* (C.6a) *holds and* $\widehat{\Gamma}$ *satisfies the lower-RE condition* (C.7). *Also suppose* $\lambda_n \gtrsim \varphi_1 \sqrt{\frac{\log p}{n}}$. *Then any global optimum* $\widehat{\beta}$ *of the Lasso program* (5.22) *satisfies the bounds*

$$\|\widehat{\beta} - \widetilde{\beta}\|_2 \leq \frac{c_0 \sqrt{k}}{\alpha_\ell} \max\{\varphi_1 \sqrt{\frac{\log p}{n}}, \lambda_n\}, \tag{C.9}$$

$$\|\widehat{\beta} - \widetilde{\beta}\|_1 \leq \frac{8 c_0 k}{\alpha_\ell} \max\{\varphi_1 \sqrt{\frac{\log p}{n}}, \lambda_n\}. \tag{C.10}$$

We now argue that for suitable scaling $n \gtrsim k \log p$, any optimum $\widehat{\beta}$ lies in the interior of $\mathbb{B}_1(b_0\sqrt{k})$:

Lemma C.4. *Suppose* $\widehat{\beta}$ *is an optimum of the Lasso program* (5.22). *Then under the scaling* $n \gtrsim \left(\frac{\varphi_1}{\alpha(b_0 - \|\widetilde{\beta}\|_2)}\right)^2 k \log p$, *we have*

$$\widehat{\beta} \notin \partial \mathbb{B}_1(b_0\sqrt{k}).$$

By Lemma C.4, we are guaranteed that $\widehat{\beta}$ is an interior point of the feasible set. Consequently, by Proposition 2.3.2 of Clarke [22], we are guaranteed that 0 is a generalized gradient of the objective function at $\widehat{\beta}$. By Proposition 2.3.3 of Clarke [22], there must exist a vector $\hat{z} \in \partial \|\beta\|_1 \,|_{\beta = \widehat{\beta}}$ such that

$$\widehat{\Gamma}\widehat{\beta} - \widehat{\gamma} + \lambda_n \hat{z} = 0.$$

Denoting the loss function $\mathcal{L}(\beta) = \frac{1}{2}\beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta$, we have $\nabla \mathcal{L}(\beta) = \widehat{\Gamma}\beta - \widehat{\gamma}$, so

$$\nabla \mathcal{L}(\widetilde{\beta}) - \nabla \mathcal{L}(\widehat{\beta}) = \nabla \mathcal{L}(\widetilde{\beta}) + \lambda_n \hat{z} = \widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma} + \lambda_n \hat{z}.$$

Then

$$\|\nabla\mathcal{L}(\widetilde{\beta}) - \nabla\mathcal{L}(\widehat{\beta})\|_\infty \leq \|\widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma}\|_\infty + \lambda_n\|\hat{z}\|_\infty \leq \|\widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma}\|_\infty + \lambda_n. \tag{C.11}$$

Using the deviation bound (C.6a) again, we have

$$\|\widehat{\Gamma}\widetilde{\beta} - \widehat{\gamma}\|_\infty \leq \varphi_1\sqrt{\frac{\log p}{n}}.$$

It follows from equation (C.11) that if $\lambda_n \geq \varphi_1\sqrt{\frac{\log p}{n}}$, then

$$\|\nabla\mathcal{L}(\widetilde{\beta}) - \nabla\mathcal{L}(\widehat{\beta})\|_\infty \leq 2\lambda_n. \tag{C.12}$$

Finally, we lower-bound

$$\begin{aligned}\|\nabla\mathcal{L}(\widetilde{\beta}) - \nabla\mathcal{L}(\widehat{\beta})\|_\infty &= \|\widehat{\Gamma}\widehat{\nu}\|_\infty \\ &\geq \|\Sigma_x\widehat{\nu}\|_\infty - \|(\widehat{\Gamma} - \Sigma_x)\widehat{\nu}\|_\infty \\ &\geq \left|\!\left|\!\left|\Sigma_x^{-1}\right|\!\right|\!\right|_\infty^{-1}\|\widehat{\nu}\|_\infty - \|(\widehat{\Gamma} - \Sigma_x)\widehat{\nu}\|_\infty. \end{aligned} \tag{C.13}$$

Now note that $\|\widehat{\nu}\|_1 \leq 8\sqrt{k}\|\widehat{\nu}\|_2$, as shown in the proofs of Chapter 3, so we have

$$\|\widehat{\nu}\|_2^2 \leq \|\widehat{\nu}\|_\infty\|\widehat{\nu}\|_1 \leq 8\sqrt{k}\|\widehat{\nu}\|_\infty\|\widehat{\nu}\|_2,$$

so $\|\widehat{\nu}\|_2 \leq 8\sqrt{k}\|\widehat{\nu}\|_\infty$. In particular, $\|\widehat{\nu}\|_1 \leq 8k\|\widehat{\nu}\|_\infty$. Applying inequality (C.6b) to $v = \frac{\widehat{\nu}}{\|\widehat{\nu}\|_\infty}$ then gives

$$\|(\widehat{\Gamma} - \Sigma_x)\widehat{\nu}\|_\infty \leq c\varphi_2\|\widehat{\nu}\|_\infty\sqrt{\frac{k\log p}{n}}.$$

Combining this with inequality (C.13), we have

$$\|\widehat{\Gamma}\widehat{\nu}\|_\infty \geq \|\widehat{\nu}\|_\infty\left(\frac{1}{\left|\!\left|\!\left|\Sigma_x^{-1}\right|\!\right|\!\right|_\infty} - c\varphi_2\sqrt{\frac{k\log p}{n}}\right),$$

so when $n \gtrsim \varphi_2^2\left|\!\left|\!\left|\Sigma_x^{-1}\right|\!\right|\!\right|_\infty^2 k\log p$, we have

$$\|\widehat{\Gamma}\widehat{\nu}\|_\infty \geq \frac{1}{2}\frac{\|\widehat{\nu}\|_\infty}{\left|\!\left|\!\left|\Sigma_x^{-1}\right|\!\right|\!\right|_\infty}.$$

Finally, combining with inequality (C.12) yields the result of the theorem.

C.4 Proof of supporting lemmas to Proposition 5.1

In this Appendix, we derive the proofs of technical lemmas used in the proof of Proposition 5.1.

C.4.1 Proof of Lemma C.2

We denote the left-hand set by A and the right-hand set by B. It suffices to show that $\varphi_A(z) \leq \varphi_B(z)$ for all z, where φ is the support function.

For a given z, let S be the set of indices of coordinates of z with highest absolute value. We may write

$$\begin{aligned} \varphi_A(z) &= \sup_{\theta \in A} \langle \theta, z \rangle \\ &= \sup_{\theta \in A} \langle \theta_S, z_S \rangle + \langle \theta_{S^c}, z_{S^c} \rangle \\ &\leq \|z_S\|_1 + c_0 k \|z_{S^c}\|_\infty, \end{aligned} \tag{C.14}$$

since

$$\langle \theta_S, z_S \rangle \leq \|\theta_S\|_\infty \|z_S\|_1 \leq \|\theta\|_\infty \|z_S\|_1 \leq \|z_S\|_1$$

and

$$\langle \theta_{S^c}, z_{S^c} \rangle \leq \|\theta_{S^c}\|_1 \|z_{S^c}\|_\infty \leq c_0 k \|z_{S^c}\|_\infty$$

for $\theta \in A$. Furthermore, $k\|z_{S^c}\|_\infty \leq \|z_S\|_1$. Hence, inequality (C.14) becomes

$$\varphi_A(z) \leq (1 + c_0)\|z_S\|_1.$$

Finally, note that

$$\varphi_B(z) = (1 + c_0) \max_{|U| \leq k} \sup_{\|\theta_U\|_\infty \leq 1} \langle \theta_U, z_U \rangle = (1 + c_0)\|z_S\|_1,$$

establishing the desired result.

C.4.2 Proof of Lemma C.3

The proof is essentially the same as in the case of a standard linear model analyzed in Chapter 3. From the fact that $\widetilde{\beta}$ is feasible and $\widehat{\beta}$ is optimal, we obtain a basic inequality. Furthermore, defining $\widehat{\nu} = \widehat{\beta} - \widetilde{\beta}$, we may verify the cone condition $\|\widehat{\nu}\|_1 \leq c\sqrt{k}\|\widehat{\nu}\|_2$. We will not repeat the arguments here.

C.4.3 Proof of Lemma C.4

Note that

$$\|\widehat{\beta} - \widetilde{\beta}\|_1 \geq \|\widehat{\beta}\|_1 - \|\widetilde{\beta}\|_1 \geq \|\widehat{\beta}\|_1 - \sqrt{k}\|\widetilde{\beta}\|_2.$$

Hence, if $\widehat{\beta} \in \partial \mathbb{B}_1(b_0\sqrt{k})$, we have

$$\|\widehat{\beta} - \widetilde{\beta}\|_1 \geq b_0\sqrt{k} - \|\widetilde{\beta}\|_2\sqrt{k} = (b_0 - \|\widetilde{\beta}\|_2)\sqrt{k}. \tag{C.15}$$

On the other hand, Theorem 3.1 in Chapter 3 guarantees that under deviation condition (C.6a) and the lower-RE condition (C.7), we have the ℓ_1-bound

$$\|\widehat{\beta} - \widetilde{\beta}\|_1 \leq \frac{c\varphi_1 k}{\alpha}\sqrt{\frac{\log p}{n}}. \tag{C.16}$$

Combining inequalities (C.15) and (C.16) gives

$$(b_0 - \|\widetilde{\beta}\|_2)\sqrt{k} \leq \frac{c\varphi_1 k}{\alpha}\sqrt{\frac{\log p}{n}},$$

contradicting the assumption that $n > \left(\frac{c\varphi_1}{\alpha(b_0 - \|\widetilde{\beta}\|_2)}\right)^2 k \log p$.

C.5 Proofs of sample-based corollaries

Here, we provide proofs for the remaining corollaries involved in sample-based approaches to graph selection.

C.5.1 Proof of Corollary 5.4

As noted by Liu et al. [53], the proof of this corollary hinges only on the deviation condition condition (5.18) being satisfied w.h.p.; the rest of the proof follows from the analysis of Ravikumar et al. [74]. We verify inequality (5.18) with $\varphi(\Sigma^*) = c_1$ and $\psi(n,p) = c' \log p$.

Note that

$$\begin{aligned}
\|\widehat{\Sigma} - \Sigma\|_{\max} &= \left\| \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^T - \overline{x}\overline{x}^T \right) - \Sigma \right\|_{\max} \\
&\leq \left\| \frac{1}{n}\sum_{i=1}^n x_i x_i^T - \mathbb{E}(x_i x_i^T) \right\|_{\max} + \left\| \overline{x}\overline{x}^T - \mathbb{E}(x_i)\mathbb{E}(x_i)^T \right\|_{\max},
\end{aligned} \tag{C.17}$$

where we have used the triangle inequality and the fact that $\Sigma = \mathbb{E}(x_i x_i^T) - \mathbb{E}(x_i)\mathbb{E}(x_i)^T$ in the second line. Noting that $\|Y\|_{\max} = \max_{j,k} |e_j^T Y e_k|$ for a matrix Y, and the random variables $e_j^T x_i$ are i.i.d. Bernoulli (sub-Gaussian parameter 1) for each fixed j, we conclude by standard sub-Gaussian tail bounds (cf. Lemma A.14 in Appendix A.3) that the first term is bounded by $\sqrt{\frac{\log p}{n}}$, with probability at least $1 - c\exp(-c' \log p)$. For the second term, we may further bound

$$\begin{aligned}
\|\overline{x}\overline{x}^T - \mathbb{E}(x_i)\mathbb{E}(x_i)^T\|_{\max} \quad &\leq \quad \|(\overline{x} - \mathbb{E}(x_i))(\overline{x} - \mathbb{E}(x_i))^T\|_{\max} \\
&\qquad + 2\|\mathbb{E}(x_i)\|_\infty \|\overline{x} - \mathbb{E}(x_i)\|_\infty,
\end{aligned}$$

by way of the triangle inequality. Note that $e_j^T(\overline{x} - \mathbb{E}(x_i))$ is an average of i.i.d. sub-Gaussian variables with parameter 1, hence has sub-Gaussian parameter $\frac{1}{n}$. Therefore, we have the

even tighter bound $\frac{1}{n}\sqrt{\frac{\log p}{n}}$ for this term. Combining the bounds for the two terms in inequality (C.17) establishes the deviation condition (5.18).

By the machinery of Ravikumar et al. [74], we then have the elementwise bound

$$\mathbb{P}\big[\|\widehat{\Theta} - \Theta^*\|_{\max} \geq \tau_n\big] \leq c \exp(-c' \log p).$$

The statement about thresholding $\widehat{\Theta}$ to obtain a consistent estimate of Θ^* follows immediately.

C.5.2 Proof of Corollary 5.5

The analysis borrows techniques from the paper [11]. We first prove that under the scaling $n \gtrsim \kappa^2 \log p$, we have $|r_C(s,t) - \widehat{r}_C(s,t)| \leq \frac{\kappa}{4}$ for all $(s,t) \in V \times V$, with probability at least $1 - c_1 \exp(-c_2 \log p)$. First fix a pair (s,t) and a corresponding pair of values (x_s, x_t). By a simple application of Hoeffding's inequality, we have

$$\mathbb{P}\left(|\mathbb{P}(X_s = x_s, X_t = x_t) - \widehat{\mathbb{P}}(X_s = x_s, X_t = x_t)| \geq \epsilon\right) \leq c \exp(-c' n \epsilon^2),$$

and similar bounds hold for the marginal deviation terms $|\mathbb{P}(X_s = x_s) - \widehat{\mathbb{P}}(X_s = x_s)|$ and $|\mathbb{P}(X_t = x_t) - \widehat{\mathbb{P}}(X_t = x_t)|$. Note that

$$\begin{aligned} |r_C(s,t) - \widehat{r}_C(s,t)| \leq \sum_{x_s, x_t} \Big(& |\mathbb{P}(X_s = x_s, X_t = x_t) - \widehat{\mathbb{P}}(X_s = x_s, X_t = x_t)| \\ & + |\mathbb{P}(X_s = x_s)\mathbb{P}(X_t = x_t) - \widehat{\mathbb{P}}(X_s = x_s)\widehat{\mathbb{P}}(X_t = x_t)| \Big). \end{aligned}$$

Furthermore,

$$\begin{aligned} |\mathbb{P}(X_s = x_s)\mathbb{P}(X_t = x_t) - & \widehat{\mathbb{P}}(X_s = x_s)\widehat{\mathbb{P}}(X_t = x_t)| \\ & \leq |\mathbb{P}(X_s = x_s) - \widehat{\mathbb{P}}(X_s = x_s)| \cdot \mathbb{P}(X_t = x_t) \\ & \quad + |\mathbb{P}(X_t = x_t) - \widehat{\mathbb{P}}(X_t - x_t)| \cdot \widehat{\mathbb{P}}(X_s = x_s) \\ & \leq 2\epsilon, \end{aligned}$$

so taking a union bound over all pairs (s,t) and all values (x_s, x_t), we have

$$|r_C(s,t) - \widehat{r}_C(s,t)| \leq 3m^2 \epsilon$$

for all $(s,t) \in V \times V$, with probability at least $1 - cm^2 p^2 \exp(-c' n \epsilon^2)$. Finally, taking $\epsilon = \frac{\kappa}{12m^2}$ and using the fact that $n \gtrsim \kappa^2 \log p$ gives the desired bound, with probability at least $1 - c_1 \exp(-c_2 \log p)$.

In particular, it follows that

$$N(s) \subseteq \mathcal{C} \subseteq \left\{ t \in V : r_C(s,t) \geq \frac{\kappa}{4} \right\},$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. Since the last subset has cardinality at most $d^{\frac{\log(4/\kappa)}{\zeta}}$ by the correlation decay condition, we also have $|\mathcal{C}| \leq d^{\frac{\log(4/\kappa)}{\zeta}}$, as claimed.

The remainder of the proof is identical to the proof of Proposition 5.1, and is a consequence of Theorem C.1.

Bibliography

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. "Fast global convergence of gradient methods for high-dimensional statistical recovery". In: *Annals of Statistics* 40.5 (2012), pp. 2452–2482.

[2] K. S. Alexander. "Rates of growth and sample moduli for weighted empirical processes indexed by sets". In: *Probability Theory and Related Fields* 75 (1987), pp. 379–423.

[3] A. Anandkumar, V.Y.F. Tan, and A.S. Willsky. "High-Dimensional Structure Learning of Ising Models: Local Separation Criterion". In: *Annals of Statistics* 40.3 (2012), pp. 1346–1375.

[4] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data". In: *Journal of Machine Learning Research* 9 (2008), pp. 485–516.

[5] O. E. Barndorff-Nielson. *Information and Exponential Families*. Chichester: Wiley, 1978.

[6] D. P. Bertsekas. *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.

[7] J. Besag. "Spatial interaction and the statistical analysis of lattice systems". In: *Journal of the Royal Statistical Society Series B - Statistical Methodology* 36 (1974), pp. 192–236.

[8] P. J. Bickel, Y. Ritov, and A. Tsybakov. "Simultaneous Analysis of Lasso and Dantzig Selector". In: *Annals of Statistics* 37.4 (2009), pp. 1705–1732.

[9] L. Birgé. "Approximation dans les espaces metriques et theorie de l'estimation". In: *Z. Wahrsch. verw. Gebiete* 65 (1983), pp. 181–327.

[10] P. Breheny and J. Huang. "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection". In: *Annals of Applied Statistics* 5.1 (2011), pp. 232–253.

[11] G. Bresler, E. Mossel, and A. Sly. "Reconstruction of Markov Random Fields from Samples: Some Observations and Algorithms". In: *APPROX-RANDOM*. 2008, pp. 343–356.

[12] L. D. Brown. *Fundamentals of statistical exponential families*. Hayward, CA: Institute of Mathematical Statistics, 1986.

[13] T. Cai, W. Liu, and X. Luo. "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation". In: *Journal of the American Statistical Association* 106 (2011), pp. 594–607.

[14] E. Candes and T. Tao. "The Dantzig Selector: Statistical estimation when p is much larger than n". In: *Annals of Statistics* 35.6 (2007), pp. 2313–2351.

[15] E. J. Candes and J. Romberg. "Practical signal recovery from random projections". In: *Proc. SPIE Computational Imaging*. Vol. 5674. 2005, pp. 76–86.

[16] E. J. Candes, J. Romberg, and T. Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 489–509.

[17] E. J. Candes and T. Tao. "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?" In: *IEEE Transactions on Information Theory* 52.12 (2006), pp. 5406–5425.

[18] R. J. Carroll, D. Ruppert, and L. A. Stefanski. *Measurement Error in Nonlinear Models*. Chapman and Hall, 1995.

[19] L. Chen and Y. Gu. "A non-convex approach for sparse recovery with convergence guarantee". In: *arXiv e-prints* (Mar. 2013). Available at http://arxiv.org/abs/1211.7089. arXiv:1211.7089 [cs.IT].

[20] S. Chen, D. L. Donoho, and M. A. Saunders. "Atomic decomposition by basis pursuit". In: *SIAM Journal on Scientific Computing* 20.1 (1998), pp. 33–61.

[21] C.I. Chow and C.N. Liu. "Approximating discrete probability distributions with dependence trees". In: *IEEE Transactions on Information Theory* 14 (1968), pp. 462–467.

[22] F. H. Clarke. *Optimization and Nonsmooth Analysis*. New York: Wiley-Interscience, 1983.

[23] J. N. Darroch and T. P. Speed. "Additive and multiplicative models and interactions". In: *Ann. Statist.* 11.3 (1983), pp. 724–738.

[24] A. d'Aspremont, O. Banerjee, and L. El Ghaoui. "First order methods for sparse covariance selection". In: *SIAM Journal on Matrix Analysis and its Applications* 30.1 (2008), pp. 55–66.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 39.1 (1977), pp. 1–38.

[26] J. Duchi et al. "Efficient projections onto the ℓ_1-ball for learning in high dimensions". In: *International Conference on Machine Learning*. 2008, pp. 272–279.

[27] J. Fan, Y. Feng, and Y. Wu. "Network exploration via the adaptive LASSO and SCAD penalties". In: *Annals of Applied Statistics* (2009), pp. 521–541.

[28] J. Fan and R. Li. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties". In: *Journal of the American Statistical Association* 96 (2001), pp. 1348–1360.

[29] J. Fan, L. Xue, and H. Zou. "Strong oracle optimality of folded concave penalized estimation". In: *arXiv e-prints* (Oct. 2013). Available at http://arxiv.org/abs/1210.5992. arXiv:1210.5992 [math.ST].

[30] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical Lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441.

[31] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[32] S. van de Geer and P. Buhlmann. "On the conditions used to prove oracle results for the Lasso". In: *Electronic Journal of Statistics* 3 (2009), pp. 1360–1392.

[33] G. R. Grimmett. "A theorem about random fields". In: *Bulletin of the London Mathematical Society* 5 (1973), pp. 81–84.

[34] J. P. Haldar, D. Hernando, and Z.-P. Liang. "Compressed-Sensing MRI With Random Encoding." In: *IEEE Trans. Med. Imaging* 30.4 (2011), pp. 893–903.

[35] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[36] P. J. Huber. "Robust Estimation of a Location Parameter". In: *The Annals of Mathematical Statistics* 35.1 (Mar. 1964), pp. 73–101.

[37] D. Hug and W. Weil. *A Course on Convex Geometry*. Lecture notes available at www.math.kit.edu/iag4/lehre/convgeo2010w/media/cg.pdf. 2010.

[38] D. R. Hunter and R. Li. "Variable selection using MM algorithms". In: *Annals of Statistics* 33.4 (2005), pp. 1617–1642.

[39] J. T. Hwang. "Multiplicative Errors-in-Variables Models with Applications to Recent Data Released by the U.S. Department of Energy". In: *Journal of the American Statistical Association* 81.395 (1986), pp. 680–688.

[40] Joseph G. Ibrahim et al. "Missing-Data Methods for Generalized Linear Models: A Comparative Review". In: *Journal of the American Statistical Association* 100.469 (Mar. 2005), pp. 332–346.

[41] S. J. Iturria, R. J. Carroll, and D. Firth. "Polynomial Regression and Estimating Functions in the Presence of Multiplicative Measurement Error". In: *Journal of the Royal Statistical Society Series B - Statistical Methodology* 61 (3 1999), pp. 547–561.

[42] L. Jacob, G. Obozinski, and J. P. Vert. "Group Lasso with Overlap and Graph Lasso". In: *International Conference on Machine Learning (ICML)*. 2009, pp. 433–440.

[43] A. Jalali et al. "On Learning Discrete Graphical Models using Group-Sparse Regularization". In: *Journal of Machine Learning Research - Proceedings Track* 15 (2011), pp. 378–387.

[44] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[45] S. Kunis and H. Rauhut. "Random Sampling of Sparse Trigonometric Polynomials II – Orthogonal Matching Pursuit versus Basis Pursuit". In: *Found. Comput. Math.* 8.6 (Nov. 2008), pp. 737–763. ISSN: 1615-3375.

[46] S. L. Lauritzen and D. J. Spiegelhalter. "Local computations with probabilities on graphical structures and their application to expert systems (with discussion)". In: *Journal of the Royal Statistical Society B* 50 (1988), pp. 155–224.

[47] S.L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[48] M. Ledoux. *The concentration of measure phenomenon*. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society, 2001.

[49] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. New York, NY: Springer-Verlag, 1991.

[50] E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Verlag, 1998.

[51] R. Little and D. B. Rubin. *Statistical analysis with missing data*. New York: Wiley, 1987.

[52] H. Liu, J.D. Lafferty, and L.A. Wasserman. "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs". In: *Journal of Machine Learning Research* 10 (2009), pp. 2295–2328.

[53] H. Liu et al. "High-dimensional semiparametric Gaussian copula graphical models". In: *The Annals of Statistics* 40.4 (Aug. 2012), pp. 2293–2326. DOI: 10.1214/12-AOS1037.

[54] P. Loh and P. Bühlmann. "High-dimensional learning of linear causal networks via inverse covariance estimation". In: *Journal of Machine Learning Research* (To appear). Available at http://arxiv.org/abs/1311.3492.

[55] P. Loh and M.J. Wainwright. "High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity". In: *Annals of Statistics* 40.3 (2012), pp. 1637–1664.

[56] P. Loh and M.J. Wainwright. "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima". In: *arXiv e-prints* (May 2013). Available at http://arxiv.org/abs/1305.2436. arXiv:1305.2436 [math.ST].

[57] M. Lustig, D. Donoho, and J. M. Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging." In: *Magnetic Resonance in Medicine* 58.6 (2007), pp. 1182–95.

[58] R. Mazumder, J.H. Friedman, and T. Hastie. "SparseNet: Coordinate Descent With Nonconvex Penalties". In: *Journal of the American Statistical Association* 106.495 (2011), pp. 1125–1138.

[59] P. McCullagh and J. A. Nelder. *Generalized Linear Models (Second Edition)*. London: Chapman & Hall, 1989.

[60] N. Meinshausen and P. Bühlmann. "High-dimensional graphs and variable selection with the Lasso". In: *Annals of Statistics* 34 (2006), pp. 1436–1462.

[61] N. Meinshausen and B. Yu. "Lasso-type recovery of sparse representations for high-dimensional data". In: *Annals of Statistics* 37.1 (2009), pp. 246–270.

[62] S. Negahban and M. J. Wainwright. "Estimation of (near) low-rank matrices with noise and high-dimensional scaling". In: *Annals of Statistics* 39.2 (2011), pp. 1069–1097.

[63] S. Negahban et al. "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers". In: *Statistical Science* 27.4 (2012). See arXiv version for lemma/propositions cited here, pp. 538–557.

[64] S. Negahban et al. "A unified framework for the analysis of regularized M-estimators". In: *Advances in Neural Information Processing Systems*. 2009.

[65] Y. Nesterov. *Gradient methods for minimizing composite objective function*. CORE Discussion Papers 2007076. Universit Catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007.

[66] Y. Nesterov and A. Nemirovskii. *Interior Point Polynomial Algorithms in Convex Programming*. SIAM studies in applied and numerical mathematics. Society for Industrial and Applied Mathematics, 1987.

[67] M.E.J. Newman and D.J. Watts. "Scaling and percolation in the small-world network model". In: *Phys. Rev. E* 60.6 (Dec. 1999), pp. 7332–7342.

[68] G. Obozinski, M.J. Wainwright, and M.I. Jordan. "Support union recovery in high-dimensional multivariate regression". In: *Annals of Statistics* 39 (2011), pp. 1–47.

[69] J. Pearl. *Causality: Models, Reasoning and Inference*. 2nd. Cambridge University Press, 2009.

[70] G. Raskutti, M. J. Wainwright, and B. Yu. "Restricted Eigenvalue Properties for Correlated Gaussian Designs". In: *Journal of Machine Learning Research* 11 (2010), pp. 2241–2259.

[71] G. Raskutti, M.J. Wainwright, and B. Yu. "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls". In: *IEEE Transactions on Information Theory* 57.10 (2011), pp. 6976–6994.

[72] Holger Rauhut. "Compressive Sensing and Structured Random Matrices". In: *Theoretical Foundations and Numerical Methods for Sparse Recovery*. Ed. by M. Fornasier. Vol. 9. Radon Series Comp. Appl. Math. deGruyter, 2010, pp. 1–92.

[73] P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. "High-dimensional Ising model selection using ℓ_1-regularized logistic regression". In: *Annals of Statistics* 38 (2010), p. 1287.

[74] P. Ravikumar et al. "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence". In: *Electronic Journal of Statistics* 4 (2011), pp. 935–980.

[75] R. T. Rockafellar. *Convex Analysis*. Princeton: Princeton University Press, 1970.

[76] M. Rosenbaum and A. B. Tsybakov. "Sparse recovery under matrix uncertainty". In: *Annals of Statistics* 38 (2010), pp. 2620–2651.

[77] M. Rosenbaum and A.B. Tsybakov. "Improved matrix uncertainty selector". In: *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*. Ed. by M. Banerjee et al. Vol. Volume 9. Collections. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2013, pp. 276–290.

[78] A. J. Rothman et al. "Sparse permutation invariant covariance estimation". In: *Electronic Journal of Statistics* 2 (2008), pp. 494–515.

[79] D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

[80] M. Rudelson and S. Zhou. *Reconstruction from anisotropic random measurements*. Tech. rep. University of Michigan, 2011.

[81] N. P. Santhanam and M. J. Wainwright. "Information-Theoretic Limits of Selecting Binary Graphical Models in High Dimensions". In: *IEEE Transactions on Information Theory* 58.7 (2012), pp. 4117–4134.

[82] F. M. Sebert et al. "Compressed sensing MRI with random B1 field". In: *International Society of Magnetic Resonance in Medicine Scientific Meeting*. 2008, p. 3151.

[83] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Vol. 81. The MIT Press, 2000.

[84] N. Städler and P. Bühlmann. "Missing values: Sparse inverse covariance estimation and an extension to sparse regression". In: *Statistics and Computing* (2010), pp. 1–17.

[85] R. Tibshirani. "Regression shrinkage and selection via the Lasso". In: *Journal of the Royal Statistical Society, Series B* 58.1 (1996), pp. 267–288.

[86] J. Tropp. "On the conditioning of random subdictionaries". In: *Applied and Computational Harmonic Analysis* 25.1 (2008), pp. 1–24.

[87] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. With applications to statistics. New York: Springer-Verlag, 1996.

[88] S. A. Vavasis. "Complexity Issues In Global Optimization: A Survey". In: *Handbook of Global Optimization*. Kluwer, 1995, pp. 27–41.

[89] R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices*. Ed. by Yonina C. Eldar and GittaEditors Kutyniok. Cambridge University Press, 2012, pp. 210–268. ISBN: 9780511794308.

[90] J.-P. Vial. "Strong convexity of sets and functions". In: *Journal of Mathematical Economics* 9.1-2 (1982), pp. 187–205.

[91] M. J. Wainwright and M. I. Jordan. "Graphical Models, Exponential Families, and Variational Inference". In: *Found. Trends Mach. Learn.* 1.1-2 (Jan. 2008), pp. 1–305. ISSN: 1935-8237.

[92] H. Wang, D. Liang, and L. Ying. "Pseudo 2D random sampling for compressed sensing MRI". In: *Proc. IEEE Eng. Med. Biol. Soc.* 2009, pp. 2672–2675.

[93] Z. Wang, H. Liu, and T. Zhang. "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems". In: *arXiv e-prints* (Dec. 2013). Available at http://arxiv.org/abs/1306.4960v3. arXiv:1306.4960v3 [stat.ML].

[94] S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo. "Sparse reconstruction by separable approximation". In: *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. 2008, pp. 3373 –3376.

[95] Q. Xu and J. You. "Covariate Selection for Linear Errors-in-Variables Regression Models". In: *Communications in Statistics - Theory and Methods* 36.2 (2007), pp. 375–386.

[96] L. Xue and H. Zou. "Regularized rank-based estimation of high-dimensional nonparanormal graphical models". In: *Annals of Statistics* 40.5 (2012), pp. 2541–2571.

[97] Y. Yang and A. Barron. "Information-theoretic determination of minimax rates of convergence". In: *Annals of Statistics* 27.5 (1999), pp. 1564–1599.

[98] B. Yu. "Assouad, Fano, and Le Cam". In: *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam*. 1996, pp. 423–435.

[99] M. Yuan. "High-Dimensional Inverse Covariance Matrix Estimation via Linear Programming". In: *Journal of Machine Learning Research* 99 (2010), pp. 2261–2286. ISSN: 1532-4435.

[100] M. Yuan and Y. Lin. "Model selection and estimation in the Gaussian graphical model". In: *Biometrika* 94.1 (2007), pp. 19–35.

[101] C.-H. Zhang. "Nearly unbiased variable selection under minimax concave penalty". In: *Annals of Statistics* 38.2 (2010), pp. 894–942.

[102] C. H. Zhang and J. Huang. "The sparsity and bias of the Lasso selection in high-dimensional linear regression". In: *Annals of Statistics* 36.4 (2008), pp. 1567–1594.

[103] C.-H. Zhang and T. Zhang. "A general theory of concave regularization for high-dimensional sparse estimation problems". In: *Statistical Science* 27.4 (2012), pp. 576–593.

[104] T. Zhang. "Analysis of Multi-stage Convex Relaxation for Sparse Regularization". In: *Journal of Machine Learning Research* 11 (2010), pp. 1081–1107.

[105] P. Zhao and B. Yu. "On model selection consistency of Lasso". In: *Journal of Machine Learning Research* 7 (2006), pp. 2541–2567.

[106] H. Zou and R. Li. "One-step sparse estimates in nonconcave penalized likelihood models". In: *Annals of Statistics* 36.4 (2008), pp. 1509–1533.