# REGULARIZATION IN HIGH-DIMENSIONAL STATISTICS

A DISSERTATION
SUBMITTED TO THE INSTITUTE FOR
COMPUTATIONAL AND MATHEMATICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

YUEKAI SUN
JUNE 2015

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Michael A. Saunders) Co-principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Jonathan E. Taylor) Co-principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Andrea Montanari)

Approved for the University Committee on Graduate Studies.

Dedicated to the loving memory of 付俊.

1925 – 2011

# ABSTRACT

Modern datasets are growing in terms of samples but even more so in terms of variables. We often encounter datasets where samples consists of time series, images, even movies, so that each sample has thousands, even millions of variables. Classical statistical approaches are inadequate for working with such high-dimensional data because they rely on theoretical and computational tools developed without such data in mind. The work in this thesis seeks to close the apparent gap between the growing size of emerging datasets and the capabilities of existing approaches to statistical estimation, inference, and computing.

This thesis focuses on two problems that arise in *learning* from high-dimensional data (versus black-box approaches that do not yield insights into the underlying data-generation process). They are:

1. model selection and post-selection inference: discover the latent low-dimensional structure in high-dimensional data;

2. scalable statistical computing: design scalable estimators and algorithms that avoid communication and minimize "passes" over the data.

The work relies crucially on results from convex analysis and geometry. Many of the algorithms and proofs are inspired by results from this beautiful but dusty corner of mathematics.

# ACKNOWLEDGMENTS

I have always looked forward to writing the acknowledgements section of my thesis. It is a convenient excuse to reminisce about all the good times over the past five years. My experience at Stanford has been abundantly enriched by the Stanford Community. Before delving into the technical part of my thesis, I wish to express my gratitude to some of these people.

To begin, I wish to thank my two advisers: Michael Saunders and Jonathan Taylor. Michael, you are so much more than an adviser to me; you are a good friend, a role model. Thank you for supporting me during my years at Stanford and granting me the freedom to explore my own ideas. Your trust and support gave me the courage to choose my own path at Stanford, and your commitment to mentorship sets an aspirational standard for my own development as a mentor. Thank you!

Jonathan, it has been a tremendous pleasure working with you. I could not have asked for a more admirable and inspiring mentor. Thank you for taking me under your wing over the past three years. Your broad perspective on research has immeasurably enriched my tastes. I am both honored and humbled to call myself your student.

In addition to my advisers, I wish to specially acknowledge my partner in crime: Jason Lee. It is said that one is the average of the five people one spends the most time with. If that is the case, I am exceedingly glad that we spent so much time together over the past five years. I only wish our paths had crossed earlier.

There are so many others I owe credit to that I will most likely neglect to mention some. I ask your forgiveness in advance. To begin, I wish to thank the ICME faculty, especially Margot, and the statistics faculty, especially Andrea and Trevor, for their generous support and guidance. I also wish to thank the ICME staff: Brian, Chris, Antoinette, Emily, and Matt for all their efforts behind the scenes to keep ICME running smoothly. Last but not least, I wish to thank my friends in ICME: Ed, Ernest, Anil, Alex, Austin, Eileen, Nolan, Jiyan, Nick, Milinda, Tania, Tiago, Victor, Brett, Faye, Kari, Ruoxi, Ronan, Ding, Xiaotong; and in the statistics department: Dennis, Edgar, Josh, Will, Weijie, Xiaoying for all the good times over the past five years. It has been a blast!

I also wish to acknowledge my parents. There are no words to express my eternal gratitude. My parents left China soon after it opened its doors

in pursuit of a better life. The journey first took us to Singapore and then to the United States. Along the way, they eschewed their own happiness to create a brighter future for me and my sister time and again. Thank you for taking me on this journey and seeing me all the way through!

Finally, I must thank my wife Hannah for my happiness. Thank you for your unwavering support over the past six years. Thank you for inspiring me to be a better man. It has been a great journey with you so far, and the end is nowhere in sight!

My work was supported by the Department of Energy and the National Institute of General Medical Sciences of the National Institutes of Health (award U01GM102098). The content is solely the responsibility of the author and does not necessarily represent the official views of the funding agencies.

# CONTENTS

Part I

# ESTIMATION

# 1 REGULARIZATION IN HIGH-DIMENSIONAL STATISTICS

Regularization is an old idea. It was first proposed by Tikhonov (1943) in the context of solving ill-posed inverse problems and soon appeared in statistics (e. g. in Stein (1956), James and Stein (1961)). It has since become a standard tool in the well-trained statistician's toolkit. In the contemporary era of high-dimensional statistics, where the sample size is of the same order as the dimension of the samples or substantially smaller, ill-posedness is the norm rather than the exception, and regularization is essential. There is a voluminous literature on regularization in statistics, and a comprehensive survey is beyond the scope of this thesis. Bühlmann and Van De Geer (2011) gives a review of the recent developments spurred by the proliferation of high-dimensional problems. We focus on recent developments spurred by trends in the size and complexity of modern datasets.

Modern datasets are growing in terms of sample size $n$ but even more so in terms of dimension $p$. Often, we come across datasets where $p \sim n$ or even $p \gtrsim n$. In such *high-dimensional* settings, regularization serves two purposes: one statistical and the other computational. Statistically, regularization is essential: it prevents overfitting and allows us to design estimators that exploit latent low-dimensional structure in the data to achieve consistency. From the computational point of view, regularization improves the stability of the problem and often leads to computational gains.

This thesis studies *regularized M-estimators* in the high-dimensional setting. The goal is to estimate a parameter $\theta^* \in \mathbf{R}^p$ by minimizing the sum of a loss function and a regularizer. More precisely, let $\mathcal{Z}^n := \{z_1, \dots, z_n\}$ be a collection of samples with marginal distribution $\mathbf{P}$ and $\theta^* = \theta^*(\mathbf{P})$. An *M-estimator* estimates $\theta^*$ by

$$\hat{\theta} \in \arg\min_{\theta \in \mathbf{R}^p} \ell_n(\theta, \mathcal{Z}^n) := \frac{1}{n} \sum_{i=1}^{n} \ell_n(\theta, z_i), \tag{1.1}$$

where $\ell_n : \mathbf{R}^p \times \mathcal{Z} \to \mathbf{R}$ is a *loss function* that measures the fit of a parameter $\theta$ to a sample $z_i$. The loss function is usually chosen so that the unknown parameter $\theta^*$ minimizes the population risk; i. e.

$$\theta^* = \arg\min_{\theta \in \mathbf{R}^p} \mathbf{E}[\ell_n(\theta, Z)].$$

A regularized M-estimator combines an M-estimator with a *regularizer* or *penalty* $\rho : \mathbf{R}^p \to \mathbf{R}_+$ to induce solutions with some particular structure. It is possible to combine the loss function and regularizer in two ways. The first option is to minimize the loss function subject to a constraint on the regularizer:

$$\hat{\theta} \in \arg\min_{\theta \in \mathbf{R}^p} \ell_n(\theta, \mathcal{Z}^n) \text{ subject to } \rho(\theta) \leq r, \tag{1.2}$$

where $\tau > 0$ is a radius. We focus on the second option: to minimize the Lagrangian form of the constrained problem:

$$\hat{\theta} \in \arg\min_{\theta \in \mathbf{R}^p} \ell_n(\theta, \mathcal{Z}^n) + \lambda \rho(\theta), \tag{1.3}$$

where $\lambda > 0$ is a regularization weight. If $\ell_n$ and $\rho$ are convex in $\theta$, the two options are equivalent: for any choice of the regularization weight $\lambda$, there is a radius $r$ for which the solution set of (1.2) coincides with that of (1.3).

To set the stage for more complicated regularizers, we describe two simple examples. A classical example of a regularized M-estimator is the *ridge regression estimator* by Hoerl and Kennard (1970). Given samples of the form $z_i = (x_i, y_i) \in \mathbf{R}^p \times \mathbf{R}$, the ridge regression estimator minimizes the sum of the least-squares criterion and the squared $\ell_2$ regularizer:

$$\hat{\theta} := \arg\min_{\theta \in \mathbf{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \frac{\lambda}{2} \|\theta\|_2^2 . \tag{1.4}$$

Although linear regression is broadly applicable, some problems require richer, more flexible models. The non-parametric analog of ridge regression in the non-parametric setting is

$$\hat{\theta} := \arg\min_{f \in \mathcal{H}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2 , \tag{1.5}$$

where $\mathcal{H}$ is some Hilbert space of real-valued functions equipped with norm $\|\cdot\|_{\mathcal{H}}$. The Hilbert norm regularizer is usually chosen to induce some kind of smoothness on the solution.

Another M-estimator that has been the subject of intensive study recently is the *lasso* estimator:

$$\underset{\theta \in \mathbf{R}^p}{\text{minimize}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 \text{ subject to } \|\theta\|_1 \leq \sigma.$$

Its Lagrangian form, also known as *basis pursuit denoising*, is

$$\underset{\theta \in \mathbf{R}^p}{\text{minimize}} \; \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda \left\| \theta \right\|_1 . \tag{1.6}$$

In statistics, the Lagrangian form is often called the lasso. The constrained form of the estimator was proposed by Tibshirani (1996), and the Lagrangian form, by Chen et al. (1998). The $\ell_1$ regularizer induces sparse solutions, which is most appropriate when the unknown regression coefficients $\theta^*$ are sparse. There is an extensive literature on the theoretical propoerties of the lasso and other $\ell_1$ regularized M-estimators in the high-dimensional setting, including persistency (Greenshtein et al. (2004), Bunea et al. (2007), Bickel et al. (2009)), consistency (Donoho (2006), Zhang and Huang (2008), Donoho and Tanner (2009), Bickel et al. (2009)), and selection consistency (Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Tropp (2006), Wainwright (2009)). We describe some extensions of the $\ell_1$ regularizer in Section 1.1.

## 1.1 STRUCTURED SPARSITY INDUCING REGULARIZERS

In many problems, we expect the unknown parameters to be sparse in a group-wise, possibly hierarchical way. To induce such structured sparse solutions, many group-sparsity inducing regularizers have been proposed. The simplest example is the *group lasso* regularizer by Kim et al. (2006) and Yuan and Lin (2006):

$$\rho(\theta) := \sum_{g \in \mathcal{G}} \left\| \theta_g \right\|_2 , \tag{1.7}$$

where each $g \in \mathcal{G}$ is a subset of the indices $[p]$. In the original form of the group lasso regularizer, the groups are non-overlapping. A variant of the group lasso penalty that penalizes the sum of the $\ell_\infty$ norms of the groups was proposed by Turlach et al. (2005). More recently, Zhao et al. (2009), Jacob et al. (2009), and Baraniuk et al. (2010) proposed extensions of the group lasso penalty to induce structured sparsity with overlapping groups.

The naive overlapping group lasso regularizer suffers from a subtle drawback. By design, regularizing with (1.7) encourages groups of parameters to be zero. Thus the complement of the support of a solution is usually the union of some subset of the groups. Unfortunately, in many applications, we seek solutions whose support is the union of groups—the opposite effect.

To correct this fault, Jacob et al. (2009) proposed the *latent group lasso* regularizer:

$$\rho(\theta) := \inf_{\theta} \left\{ \textstyle\sum_{g\in\mathcal{G}} \left\|\theta_g\right\|_2 : \theta = \sum_{g\in\mathcal{G}} \theta_g \right\}. \tag{1.8}$$

We emphasize that $\theta_g$ in (1.7) is a point in $\mathbf{R}^{|g|}$, but $\theta_g$ in (1.8) is a point in $\mathbf{R}^p$. The latent group lasso is based on the observation that when the groups overlap, a point $\theta$ has many possible group-sparse decompositions. By minimizing over the decompositions, the latent group lasso ensures the support of a solution is a union of groups.

Another form of structured sparsity is *sparsity in a basis*. That is, $D\theta$ is sparse for some matrix $D \in \mathbf{R}^{m\times p}$. Regularizers that induce sparsity in a basis regularize $D\theta$ instead of $\theta$. In statistics, such regularizers are called *generalized lasso* penalties: $\|D\theta\|_1$. In signal processing, they are known as *analysis regularizers*.

## 1.2 CONVEX RELAXATIONS OF THE RANK

There are many problems in multivariate statistics that boil down to optimizing over the set of low-rank matrices. A compelling application is the matrix completion problem: estimate an unknown matrix given (possibly noisy) observations of a small subset of its entries. The problem arises in collaborative filtering, where the goal is to recommend goods to users based on the users' ratings of a subset of items. The problem as stated is ill-posed; some additional structural assumption is imperative. An empirically justified assumption is that the unknown matrix has small rank. Although an ideal approach is to penalize the rank (or enforce a rank constraint), the rank function is non-convex. Thus the ideal approach is not computationally practical for all but the smallest problems.

The nuclear norm of a matrix is a natural convex relaxation of the rank function. It is the analog of the $\ell_1$ norm relaxation of the sparsity of a point. For any $\Theta \in \mathbf{R}^{p_1\times p_2}$, the rank of $\Theta$ is the number of non-zero singular values. Based on this observation, Fazel et al. (2001) suggest the nuclear norm, which is given by the $\ell_1$ norm of the singular values, as a convex relaxation of the rank:

$$\|\Theta\|_{\text{nuc}} := \sum_{i\in[p]} \sigma_i(\Theta), \tag{1.9}$$

where $\{\sigma_i(\Theta)\}_{i\in[p]}$ are the singular values of $\Theta$. A rich line of work, beginning with Recht et al. (2010), shows that minimizing the nuclear norm is often an *exact* surrogate for minimizing the rank. The theoretical properties of nuclear norm regularization under various statistical models has since been extensively studied, including matrix completion (e. g. Candès and Recht (2009), Mazumder et al. (2010), Gross (2011), Koltchinskii et al. (2011), Recht (2011), Negahban and Wainwright (2012)) and, more generally, matrix regression (e. g. Bach (2008), Recht et al. (2010), Candes and Plan (2011), Negahban et al. (2011), Rohde et al. (2011)).

The nuclear norm also has a variational characterization:

$$\|\Theta\|_{\text{nuc}} := \inf_{\Theta=UV^T} \|U\|_{\text{F}} \|V\|_{\text{F}} = \inf_{\Theta=UV^T} \tfrac{1}{2}\left(\|U\|_{\text{F}}^2 + \|V\|_{\text{F}}^2\right), \qquad (1.10)$$

which suggests other convex relaxations by replacing the Frobenius norm with other matrix norms. A well-studied example proposed by Srebro et al. (2004) is the *max norm*:

$$\|\Theta\|_{\max} := \inf_{\Theta=UV^T} \tfrac{1}{2}\left(\|U\|_{2,\infty}^2 + \|V\|_{2,\infty}^2\right),$$

where the $\ell_q/\ell_r$ matrix norm is

$$\|A\|_{q,r} := \Big(\sum_{j\in[p]} \|a_j^T\|_q^r\Big)^{\frac{1}{r}}.$$

The variational characterization also leads to alternative approaches to minimizing the nuclear norm. We defer the details to Part II.

## 1.3 REGULARIZERS FOR STRUCTURED MATRIX DECOMPOSITION

It is possible to combine the aforementioned regularizers to obtain regularizers that induce solutions that are sums of components each possessing some particular structure. For example, consider the robust form of the matrix completion problem, where a few entries of the unknown low-rank matrix may be contaminated with (possibly adversarial) noise. Thus the unknown matrix has the form $\Theta = \Theta_1^* + \Theta_2^*$, where $\Theta_1^*$ has low rank and $\Theta_2^*$ is sparse.

To induce solutions that are the sum of structured components, we consider regularizers of the form

$$\rho(\theta) := \inf_{\theta_1,\theta_2} \{\rho_1(\theta_1) + \rho_2(\theta_2) : \theta = \theta_1 + \theta_2\}, \qquad (1.11)$$

where $\rho_1$ and $\rho_2$ are regularizers chosen to induce the correct structure in $\theta_1$ and $\theta_2$. In the robust matrix completion problem, natural choices of the constituent regularizers are the nuclear norm and the (entry-wise) $\ell_1$ norm:

$$\rho(\Theta) := \inf_{\Theta_1,\Theta_2} \{\|\Theta\|_1 + \|\Theta_2\|_{\text{nuc}} : \Theta = \Theta_1 + \Theta_2\}. \tag{1.12}$$

Since being proposed by Candès et al. (2011), the "low-rank plus sparse" regularizer (1.12) has been extensively studied (e. g. Chandrasekaran et al. (2011), Hsu et al. (2011)).

Another example of (1.11) is the combination of the $\ell_1$ norm and the $\ell_1/\ell_q$ norm. It was proposed as an improvement upon $\ell_1/\ell_\infty$ regularization to induce group sparsity. Negahban and Wainwright (2011) showed that the statistical efficiency of pure $\ell_1/\ell_q$ regularization may be worse than that of pure $\ell_1$ regularization when the groups are incorrectly specified. To correct this deficiency, Jalali et al. (2010) propose a regularizer of the form (1.11) that combines the $\ell_1$ norm and the $\ell_1/\ell_q$ norm. They show that the combined regularizer outperforms pure $\ell_1$ or pure $\ell_1/\ell_q$ regularization.

In the first part, we focus on the statistical properties of regularized M-estimators. To begin, we study the consistency of regularized M-estimators in the high-dimensional setting. Our study identifies a key property of the regularizer that enables the estimator to identify latent low-dimensional structure in the data, which in turn enables efficient estimation in high dimensions.

In the second part, we turn our attention to computational issues. We study two ways to evaluate regularized M-estimators efficiently. In the sequential setting we describe a family of methods that interpolate between first- and second-order methods. In the distributed setting, we describe a way to evaluate the estimators with a single round of communication. The work in this thesis was performed jointly with Jason Lee, who contributed equally.

# 2 CONSISTENCY OF REGULARIZED M-ESTIMATORS

We turn our attention to the consistency of regularized M-estimators in the high-dimensional setting. In this chapter, we focus on *geometrically decomposable regularizers*; i. e. regularizers that are sums of support functions. The material in this and the following chapter appears in Lee et al. (2015b). Before delving in, we review some concepts from convex analysis that appear in our study.

## 2.1 CONVEX ANALYSIS BACKGROUND

Let $\mathcal{C} \subset \mathbf{R}^p$ be a closed, convex set. The *polar set* $\mathcal{C}^\circ$ is given by

$$\{x \in \mathbf{R}^n \mid x^T y \leq 1 \text{ for any } y \in \mathcal{C}\}.$$

When $\mathcal{C}$ is a cone, i. e. $\mathcal{C} = \lambda \mathcal{C}$ for any $\lambda > 0$, its polar set is known as its polar cone:

$$\mathcal{C}^\circ := \{x \in \mathbf{R}^n \mid x^T y \leq 0 \text{ for any } y \in \mathcal{C}\}.$$

The notion of polarity is a generalization of the notion of orthogonality. In particular, the polar cone of a halfspace $\mathcal{H} = \{x \in \mathbf{R}^n \mid x^T y \leq 0 \text{ for some } y \neq 0\}$ is the *ray* generated by its (outward) normal

$$\mathcal{H}^\circ = \{\lambda y \mid \lambda \geq 0\},$$

and the polar cone of a subspace is its orthocomplement. Further, given a convex cone $\mathcal{K} \subset \mathbf{R}^n$, any point $x \in \mathbf{R}^n$ has an orthogonal decomposition into its projections onto $\mathcal{K}$ and $\mathcal{K}^\circ$.

**Lemma 2.1.** *Let $\mathcal{K} \subset \mathbf{R}^n$ be a closed convex cone. Any point $x \in \mathbf{R}^n$ has a unique decomposition into its projections onto $\mathcal{K}$ and $\mathcal{K}^\circ$, i. e. $x = P_{\mathcal{K}}(x) + P_{\mathcal{K}^\circ}(x)$. Further, the components $P_{\mathcal{K}}(x)$ and $P_{\mathcal{K}^\circ}(x)$ are orthogonal.*

Recall the *indicator function* of a closed, convex set $\mathcal{C} \subset \mathbf{R}^p$ is

$$I_{\mathcal{C}}(x) := \begin{cases} 0 & x \in \mathcal{C}, \\ \infty & \text{otherwise.} \end{cases} \tag{2.1}$$

Its convex conjugate is the *support function* of $\mathcal{C}$ :

$$h_{\mathcal{C}}(x) := \sup_{y\in\mathcal{C}} x^T y. \tag{2.2}$$

Intuitively, support functions are (semi-)norms. In particular, they are *sublinear*: $h_{\mathcal{C}}(\alpha x) = \alpha h_{\mathcal{C}}(x)$ for any $\alpha > 0$ and $h_{\mathcal{C}}(x+y) \le h_{\mathcal{C}}(x) + h_{\mathcal{C}}(y)$. If $\mathcal{C}$ is symmetric about the origin, i. e. $-x \in \mathcal{C}$ for any $x \in \mathcal{C}$, the first property holds for any $\alpha \in \mathbf{R}$. Support functions (as functions of the set $\mathcal{C}$) are also additive:

$$h_{\mathcal{C}_1+\mathcal{C}_2}(x) = h_{\mathcal{C}_1}(x) + h_{\mathcal{C}_2}(x).$$

Since support functions are supremums of linear functions, their subdifferentials consist of the linear functions that attain the supremum:

$$\partial h_C(x) = \{y \in C \mid y^T x = h_C(x)\}. \tag{2.3}$$

## 2.2 GEOMETRICALLY DECOMPOSABLE PENALTIES

Since support functions are sublinear, they should be thought of as seminorms. In particular, the support function of a norm ball is the dual norm. If $\mathcal{C}$ is symmetric about the origin and contains a neighborhood of the origin, $h_{\mathcal{C}}$ is a norm. This observation leads us to consider regularizers of the form $\rho(\theta) = h_{\mathcal{C}}(\theta)$ for some set $\mathcal{C}$.

**Definition 2.2** (Geometric decomposability)**.** *For any two closed convex sets $\mathcal{A}, \mathcal{I} \subset \mathbf{R}^p$ containing the origin, a regularizer is* geometrically decomposable *with respect to the pair $(\mathcal{A}, \mathcal{I})$ if*

$$\rho(\theta) = h_{\mathcal{A}}(\theta) + h_{\mathcal{I}}(\theta) \textit{ for any } \theta \in \mathbf{R}^p. \tag{2.4}$$

The notation $\mathcal{A}$ and $\mathcal{I}$ should be as read as "active" and "inactive": span($\mathcal{A}$) should contain the unknown parameter and span($\mathcal{I}$) should contain deviations that we wish to penalize.[1] For example, if we know the sparsity pattern of the unknown parameter, then $\mathcal{A}$ should span the subspace of all points with the correct sparsity pattern.

The form (2.4) is general; if $\rho$ is a sum of support functions, i. e.

$$\rho(\theta) = h_{\mathcal{C}_1}(\theta) + \cdots + h_{\mathcal{C}_k}(\theta),$$

[1] More generally, span$(\mathcal{I})^\perp$ should contain the unknown parameter. Often, span$(\mathcal{A})$ = span$(\mathcal{I})^\perp$.

then, by the additivity of support functions, $\rho$ has the form (2.4), where $\mathcal{A}$ and $\mathcal{I}$ are sums of the sets $\mathcal{C}_1, \dots, \mathcal{C}_k$. In many cases of interest, $\mathcal{A} + \mathcal{I}$ is a norm ball and $h_{\mathcal{A}+\mathcal{I}} = h_{\mathcal{A}} + h_{\mathcal{I}}$ is the dual norm. In our study, we further assume

1. the set $\mathcal{A}$ is bounded and contains the origin.

2. the set $\mathcal{I}$ contains a relative neighborhood of the origin, i.e. $0 \in \text{relint}(\mathcal{I})$.

To allow for unregularized parameters, we do not assume $\mathcal{A} + \mathcal{I}$ contains a neighborhood of the origin. Thus $\rho$ is not necessarily a norm.

To build some intuition, consider the sparse linear regression problem: recover a sparse $\theta^* \in \mathbf{R}^p$ given predictors $X \in \mathbf{R}^{n \times p}$ and responses $y = X \in \mathbf{R}^{n \times p} + \epsilon$. The lasso (BPDN) estimates $\theta^*$ by the solution of

$$\underset{\theta \in \mathbf{R}^p}{\text{minimize}} \, \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1. \tag{2.5}$$

Let $\mathcal{S} \subset [p]$ be the support of $\theta$, and $\mathcal{S}^c$ be the complementary subset of $[p]$. It is possible to show that the $\ell_1$ norm is geometrically decomposable with respect to the sets

$$\begin{aligned} \mathcal{B}_{\infty,\mathcal{S}} &= \left\{\theta \in \mathbf{R}^p \mid \|\theta\|_\infty \leq 1, \theta_{\mathcal{S}^c} = 0\right\} \\ \mathcal{B}_{\infty,\mathcal{S}^c} &= \left\{\theta \in \mathbf{R}^p \mid \|\theta\|_\infty \leq 1, \theta_{\mathcal{S}} = 0\right\}. \end{aligned}$$

There is a well-developed theory of the lasso that says, under suitable assumptions on $X$, the lasso estimator is consistent. As we shall see, the geometric decomposability of the $\ell_1$ norm is the key to the performance of the lasso.

Before we state the main results, we note that regularizers of the form $\rho(D\theta)$ for some $D \in \mathbf{R}^{m \times p}$ are geometrically decomposable, as long as $\rho$ is geometrically decomposable. Indeed,

$$\begin{aligned} \rho(D\theta) &= h_{\mathcal{A}}(D\theta) + h_{\mathcal{I}}(D\theta) \\ &= h_{D^T\mathcal{A}}(\theta) + h_{D^T\mathcal{I}}(\theta). \end{aligned}$$

Thus $\rho$ is geometrically decomposable with respect to the images of $\mathcal{A}$ and $\mathcal{I}$ under $D^T$. This property makes geometric decomposability amendable to studying analysis regularizers.

## 2.3 CONSISTENCY OF REGULARIZED M-ESTIMATORS WITH GEOMETRICALLY DECOMPOSABLE REGULARIZERS

We begin by recalling the problem setup. We are given a collection of samples $\mathcal{Z}^n := \{z_1, \ldots, z_n\}$ with marginal distribution $\mathbf{P}$. We seek to estimate a parameter $\theta^* \in \mathcal{M} \subset \mathbf{R}^p$, where $\mathcal{M}$ is the *model subspace.* The model subspace is usually low-dimensional and captures the simple structure of the model. For example, $\mathcal{M}$ may be the subspace of vectors with a particular support or a subspace of low-rank matrices.

Let $\ell_n : \mathbf{R}^p \to \mathbf{R}$ be a convex and twice-continuously differentiable loss function. We estimate $\theta^*$ by an M-estimator with a geometrically decomposable regularizer:

$$\underset{\theta \in \mathbf{R}^p}{\text{minimize}}\ \ell_n(\theta) + \lambda(h_{\mathcal{A}}(\theta) + h_{\mathcal{I}}(\theta)), \tag{2.6}$$

where $\mathcal{I} \subset \mathbf{R}^p$ is chosen so that $\mathcal{M} = \text{span}(\mathcal{I})^\perp$. This interplay between $\mathcal{I}$ and $\mathcal{M}$ is crucial to the statistical properties of (2.6). Returning to the sparse regression example, the model subspace is $\left\{\theta \in \mathbf{R}^p : \theta_{\mathcal{S}^c} = 0\right\}$. It is easy to show that $\text{span}(B_{\infty,\mathcal{S}^c})^\perp$ is the model subspace. Thus the lasso is an instance of (2.6).

Before describing our results, we briefly review the voluminous literature on sufficient conditions for consistency of regularized M-estimators. Negahban et al. (2012) proposes a unified framework for establishing consistency and convergence rates for M-estimators with regularizers $\rho$ that are *decomposable* with respect to a pair of subspaces $M, \bar{M}$:

$$\rho(x + y) = \rho(x) + \rho(y), \text{for all } x \in M, y \in \bar{M}^\perp.$$

Many common regularizers such as the lasso, group lasso, and nuclear norm are decomposable in this sense. Negahban et al. (2012) also develop a general notion of restricted strong convexity and prove a general result that establishes the consistency of M-estimators with decomposable regularizers. Using their framework, they establish estimation consistency results for different statistical models including sparse and group sparse linear regression. Our results include a general framework for model selection consistency in a similar setting.

More recently, van de Geer (2012) proposed the notion of *weakly decomposability*. A regularizer $\rho$ is weakly decomposable if there is some norm $\rho_{\mathcal{S}^c}$ on $\mathbf{R}^{p-|\mathcal{S}|}$ such that $\rho$ is superior to the sum of $\rho$ and $\rho_{\mathcal{S}^c}$; i.e.

$$\rho(x) \geq \rho(x_{\mathcal{S}}) + \rho_{\mathcal{S}^c}(x_{\mathcal{S}^c}), \text{for all } x \in \mathbf{R}^p,$$

where $\mathcal{S} \subset [p]$ and $x_{\mathcal{S}} \in \mathbf{R}^{|\mathcal{S}|}, x_{\mathcal{S}^c} \in \mathbf{R}^{p-|\mathcal{S}|}$. Many common sparsity inducing regularizers, including the $\ell_2/\ell_1$-norm (with possibly overlapping groups), are weakly decomposable. van de Geer (2012) shows oracle inequalities for the $\ell_1$ regularizer generalizes to weakly decomposable regularizers.

Given an estimator there are various ways to assess its performance. We consider two notions: consistency and model selection consistency. An estimator $\hat{\theta}$ is *consistent* (in the $\ell_2$ norm) if the error decays to zero in probability:

$$\left\|\hat{\theta} - \theta^*\right\|_2 \xrightarrow{p} 0 \text{ as } n, p \to \infty.$$

An estimator $\hat{\theta}$ is *model selection consistent* if $\hat{\theta}$ is in the *model subspace* with high probability:

$$\mathbf{Pr}(\hat{\theta} \in \mathcal{M}) \to 1. \tag{2.7}$$

First, we state our assumptions on the problem. Our main assumptions are on the sample *Fisher information*: $Q_n = \nabla^2 \ell_n(\theta^*)$ : *restricted strong convexity*, *strong smoothness*, and *irrepresentability*.

**Assumption 2.3** (Restricted strong convexity)**.** *Let $\mathcal{C} \subset \mathbf{R}^p$ be some (a priori) known convex set containing $\theta^*$. The loss function $\ell_n$ is restricted strongly convex (on $\mathcal{C} \cap \mathcal{M}$) with constant $\mu_l > 0$ when*

$$\Delta^T \nabla^2 \ell_n(\theta) \Delta \geq \mu_l \left\|\Delta\right\|_2^2$$

*for any $\theta \in \mathcal{C} \cap \mathcal{M}$ and any $\Delta \in (\mathcal{C} \cap \mathcal{M}) - (\mathcal{C} \cap \mathcal{M})$.*

**Assumption 2.4** (Strong smoothness)**.** *The loss function $\ell_n$ is strongly smooth on $\mathcal{C}$ with constant $\mu_u > 0$ when*

$$\|\nabla^2 \ell_n(\theta) - Q_n\|_2 \leq \mu_u \left\|\theta - \theta^*\right\|_2 \textit{ for any } \theta \in \mathcal{B}.$$

When $\mathcal{C}$ is compact, which it often is, restricted strong smoothness necessarily holds by the continuity of $\nabla^2 \ell_n$. Similar notions of restricted strong convexity/smoothness are common in the literature on high-dimensional statistics. For example, the unified framework by Negahban et al. (2012) requires a (slightly stronger) notion of restricted strong convexity.

For a concrete example, we return to the sparse linear regression problem. When the rows of $X$ are *i.i.d.* Gaussian random vectors, Raskutti et al. (2010) showed there are constants $\mu_1, \mu_2 > 0$ such that

$$\frac{1}{n} \left\|X\Delta\right\|_2^2 \geq \mu_1 \left\|\Delta\right\|_2^2 - \mu_2 \frac{\log p}{n} \left\|\Delta\right\|_1^2 \text{ for any } \Delta \in \mathbf{R}^p$$

with probability at least $1 - c_1 \exp\left(-c_2 n\right)$. Their result implies RSC on $\text{span}(B_{\infty,\mathcal{S}})$ (for any $\mathcal{S} \subset [p]$) with constant $\frac{\mu_1}{2}$ as long as $n > 2\frac{\mu_2}{\mu_1} |\mathcal{S}| \log p$. Thus sparse linear regression with random Gaussian designs satisfies RSC, even when there are dependencies among the predictors.

**Assumption 2.5** (Irrepresentability)**.** *There is* $\delta \in [0,1)$ *such that*

$$\sup_{z \in \partial h_{\mathcal{A}}(\mathcal{M})} h_{\mathcal{I}^\circ}(P_{\mathcal{M}^\perp}(Q_n P_{\mathcal{M}}(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} z - z)) < 1 - \delta,$$

*where* $\partial h_{\mathcal{A}}(\mathcal{M}) := \bigcup_{x \in \mathcal{M}} \partial h_{\mathcal{A}}(x)$.

To interpret the irrepresentable condition, consider again the sparse regression problem. Since $Q_n$ is the sample covariance matrix $\frac{1}{n}X^T X$, irrepresentibility is

$$\left\| X_{\mathcal{S}^c}^T \left(X_{\mathcal{S}}^T\right)^\dagger \text{sign}(\theta_{\mathcal{S}}^*) \right\|_\infty \leq 1 - \delta. \tag{2.8}$$

To ensure (2.8), it is sufficient to assume

$$\left\| X_{\mathcal{S}^c}^T \left(X_{\mathcal{S}}^T\right)^\dagger \right\|_\infty \leq 1 - \delta \text{ for some } \delta \in [0,1). \tag{2.9}$$

The rows of $X_{\mathcal{S}^c}^T \left(X_{\mathcal{S}}^T\right)^\dagger$ are the regression coefficients of $x_j, j \in \mathcal{S}^c$ on $X_{\mathcal{S}}$. Thus (2.9) says the relevant predictors (columns of $X_{\mathcal{S}}$) are not overly well-aligned with the irrelevant predictors. Ideally, we would like the irrelevant predictors to be orthogonal to the relevant predictors: $\left\| X_{\mathcal{S}^c}^T \left(X_{\mathcal{S}}^T\right)^\dagger \right\|_\infty = 0$. Unfortunately, orthogonality is impossible in the high-dimensional setting. The irrepresentable condition relaxes orthogonality to "near orthogonality".

Finally, we require the regularization parameter $\lambda$ be large enough to dominate the "empirical process" part of the problem. More precisely, we require $\lambda \gtrsim \rho^*(\nabla \ell_n^*)$. However, when $\rho$ is not a norm (e. g. when there are unregularized parameters), $\rho^*(\nabla \ell_n^*)$ is usually infinite. To allow for unregularized parameters, we relax the requirement to $\lambda \gtrsim \bar{\rho}^*(\nabla \ell_n^*)$ for a norm $\bar{\rho} : \mathbf{R}^p \to \mathbf{R}_+$ that dominates $\rho : \bar{\rho}(\theta) \geq \rho(\theta)$ for any $\theta \in \mathbf{R}^p$.

Before we state the main consistency result, we define some *compatibility constants* that appear in its statement:

1. $\kappa_\rho \in \mathbf{R}_+$ (resp. $\kappa_{\bar{\rho}}$, $\kappa_{\bar{\rho}^*}$) is the compatibility constant between $\rho$ (resp. $\bar{\rho}$, $\bar{\rho}^*$) and the $\ell_2$ norm on $\mathcal{M}$ :

$$\kappa_\rho := \sup_\theta \left\{\rho(\theta) : \theta \in \mathcal{B}_2 \cap \mathcal{M}\right\}.$$

2. $\kappa_{\text{ir}} \in \mathbf{R}_+$ is the compatibility constant between the irrepresentable term and $\bar{\rho}^*$ :

$$\kappa_{\text{ir}} := \sup_z \left\{ h_{\mathcal{I}^\circ}(P_{\mathcal{M}^\perp}(Q_n P_{\mathcal{M}}(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} z - z)) : \bar{\rho}^*(z) \leq 1 \right\}.$$

The constants are finite because $\mathcal{B}_2 \cap \mathcal{M}$, $\{z \in \mathbf{R}^p \mid \bar{\rho}^*(z) \leq 1\}$ are compact sets, and $\rho$, $\bar{\rho}$, $\bar{\rho}^*$ are locally bounded.

**Theorem 2.6.** *For any M-estimator of the form* (2.6), *suppose*

1. *the loss function $\ell_n$ is strongly convex and strongly smooth on $\mathcal{C} \cap \mathcal{M}$ with constants $\mu_l$ and $\mu_u$,*

2. *the loss function satisfies the irrepresentable condition,*

3. *the regularization parameter $\lambda$ is in the interval*

$$\left[ \tfrac{4\kappa_{\text{ir}}}{\delta} \bar{\rho}^*(\nabla \ell_n^*), \tfrac{\mu_l^2}{2\mu_u} \left( 2\kappa_\rho + \tfrac{\delta \kappa_{\bar{\rho}}}{2\kappa_{\text{ir}}} \right)^{-2} \tfrac{\delta}{\kappa_{\bar{\rho}^*} \kappa_{\text{ir}}} \right]. \tag{2.10}$$

*Then, the estimator is unique,*

1. *consistent: $\left\| \hat{\theta} - \theta^* \right\|_2 \leq \frac{2}{\mu_l}(\kappa_\rho + \frac{\delta \kappa_{\bar{\rho}}}{4\kappa_{\text{ir}}})\lambda$,*

2. *model selection consistent: $\hat{\theta} \in M$.*

Theorem 2.6 makes a *deterministic* statement about the optimal solution to (2.6). To use the result to derive consistency and model selection consistency results for a particular M-estimator under a particular statistical model, we must

1. show the M-estimator has the form given by (2.6),

2. show the loss function and regularizer satisfies restricted strong convexity, restricted strong smoothness and irrepresentability,

3. choose a regularization parameter between (2.10). Since the left side of (2.10) is $O_p(\frac{1}{\sqrt{n}})$ for most statistical models of interest, there is such a $\lambda$ for $n$ large enough.

*Proof.* The proof of Theorem 2.6 consists of three main steps:

1. Show the solution to a restricted problem (2.11) is unique and consistent (Lemma 2.7).

2. Establish a *primal-dual witness (PDW) condition* that ensures all solutions to the original problem are also solutions to the restricted problem (Lemma 2.8).

3. Construct a primal-dual pair for the original problem from the solution to the restricted problem that satisfies the dual certificate condition.

Let $(\tilde{\theta}, \tilde{z}_{\mathcal{A}}, \tilde{z}_{\mathcal{M}^\perp})$ be a primal-dual pair to the restricted problem:

$$\underset{\theta \in \mathbf{R}^p}{\text{minimize}}\ \ell_n(\theta) + \lambda(h_{\mathcal{A}}(\theta) + h_{\mathcal{M}^\perp}(\theta)). \tag{2.11}$$

Since $\mathcal{M}^\perp$ is a subspace, $h_{\mathcal{M}^\perp}(\theta)$ is $I_{\mathcal{M}}$. The restricted primal-dual pair satisfies the first-order optimality condition

$$\begin{aligned} &\nabla \ell_n(\tilde{\theta}) + \lambda \tilde{z}_{\mathcal{A}} + \lambda \tilde{z}_{\mathcal{M}^\perp} = 0 \\ &\tilde{z}_{\mathcal{A}} \in \partial h_{\mathcal{A}}(\tilde{\theta}), \quad \tilde{z}_{\mathcal{M}^\perp} \in \mathcal{M}^\perp. \end{aligned} \tag{2.12}$$

First, we show the solution to the restricted problem is consistent.

**Lemma 2.7.** *If $\ell_n$ is strongly convex on $\mathcal{C} \cap \mathcal{M}$ and $\lambda$ is between* (2.10)*, the optimal solution to the restricted problem is unique and consistent:*

$$\left\|\tilde{\theta} - \theta^*\right\|_2 \leq \tfrac{2}{\mu_l}\left(\kappa_\rho + \tfrac{\delta \kappa_{\bar{\rho}}}{4\kappa_{\text{ir}}}\right)\lambda.$$

Next, we establish the PDW condition that ensures all solutions to the original problem are also solutions to the restricted problem.

**Lemma 2.8.** *Suppose $\hat{\theta}$ is a primal solution to* (2.6)*, and $\hat{z}_{\mathcal{A}}, \hat{z}_{\mathcal{I}}$ are dual solutions; i. e. $(\hat{\theta}, \hat{z}_{\mathcal{A}}, \hat{z}_{\mathcal{I}})$ satisfy*

$$\begin{aligned} &\nabla \ell_n(\hat{\theta}) + \lambda(\hat{z}_{\mathcal{A}} + \hat{z}_{\mathcal{I}}) = 0 \\ &\hat{z}_{\mathcal{A}} \in \partial h_{\mathcal{A}}(\hat{\theta}), \quad \hat{z}_{\mathcal{I}} \in \partial h_I(\hat{\theta}). \end{aligned}$$

*If $\hat{z}_{\mathcal{I}} \in \operatorname{relint}(I)$, then all primal solutions to* (2.6) *satisfy $h_{\mathcal{I}}(\theta) = 0$.*

Finally, we use the restricted primal-dual pair to construct a feasible primal-dual pair for the original problem (2.6). The optimality conditions of the original problem are

$$\begin{aligned} &\nabla \ell_n(\hat{\theta}) + \lambda(\hat{z}_{\mathcal{A}} + \hat{z}_{\mathcal{I}}) = 0 \\ &\hat{z}_{\mathcal{A}} \in \partial h_{\mathcal{A}}(\hat{\theta}), \quad \hat{z}_{\mathcal{I}} \in \partial h_I(\hat{\theta}). \end{aligned} \tag{2.13}$$

By construction, the pair $(\tilde{\theta}, \tilde{z}_{\mathcal{A}}, \hat{z}_{\mathcal{M}^\perp})$ satisfies

$$\nabla \ell_n(\hat{\theta}) + \lambda(\hat{z}_{\mathcal{A}} + \hat{z}_{\mathcal{I}}) = 0, \quad \hat{z}_{\mathcal{A}} \in \partial h_{\mathcal{A}}(\hat{\theta}).$$

To show $\tilde{\theta}$ is the *unique* solution to the original problem, it suffices to show $\hat{z}_{\mathcal{M}}$ is PDW feasible: $\hat{z}_{\mathcal{M}^\perp} \in \operatorname{relint}(I)$.

The restricted primal-dual pair $(\tilde{\theta}, \tilde{z}_{\mathcal{A}}, \tilde{z}_{\mathcal{M}^\perp})$ satisfies (2.12) and thus the zero reduced gradient condition:

$$P_{\mathcal{M}} \nabla \ell_n(\tilde{\theta}) + \lambda P_{\mathcal{M}} \tilde{z}_{\mathcal{A}} = 0.$$

We expand $\nabla \ell_n$ around $\theta^*$ (component-wise) to obtain

$$P_{\mathcal{M}} \nabla \ell_n^* + P_{\mathcal{M}} Q_n P_{\mathcal{M}} (\tilde{\theta} - \theta^*) + P_{\mathcal{M}} R_n + \lambda P_{\mathcal{M}} \tilde{z}_{\mathcal{A}} = 0,$$

where $\nabla \ell_n^*$ is shorthand for $\nabla \ell_n(\theta^*)$ and

$$R_n = \nabla \ell(\tilde{\theta}) - \nabla \ell(\theta^*) - Q_n(\tilde{\theta} - \theta^*)$$

is the Taylor remainder term. Since $P_{\mathcal{M}} Q_n P_{\mathcal{M}}$ is invertible on $M$, we solve for the error to obtain

$$\tilde{\theta} - \theta^* = -(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} (\nabla \ell_n^* + \lambda \tilde{z}_{\mathcal{A}} + R_n).$$

We return to (2.12) and expand $\nabla \ell_n$ around $\theta^*$ to obtain

$$\nabla \ell_n^* + Q_n(\tilde{\theta} - \theta^*) + R_n + \lambda(\tilde{z}_{\mathcal{A}} + \tilde{z}_{\mathcal{M}^\perp}) = 0.$$

We substitute in the expression for the error to obtain

$$0 = \nabla \ell_n^* - Q_n (P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} (\nabla \ell_n^* + \lambda \tilde{z}_{\mathcal{A}} + R_n) + R_n + \lambda(\tilde{z}_{\mathcal{A}} + \tilde{z}_{\mathcal{M}^\perp}).$$

Rearranging, we obtain

$$\begin{aligned}
\tilde{z}_{\mathcal{M}^\perp} &= \frac{1}{\lambda}\big(Q_n (P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} (\nabla \ell_n^* + \lambda \tilde{z}_{\mathcal{A}} + R_n) - \nabla \ell_n^* - R_n - \lambda \tilde{z}_{\mathcal{A}})\big) \\
&= Q_n P_{\mathcal{M}} (P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} \tilde{z}_{\mathcal{A}} - \tilde{z}_{\mathcal{A}} \\
&\quad + \frac{1}{\lambda}\big(Q_n P_{\mathcal{M}} (P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} (\nabla \ell_n^* + R_n) - \nabla \ell_n^* + R_n\big).
\end{aligned}$$

Finally, we take $h_{\mathcal{I}^\circ}$ to obtain

$$\begin{aligned}h_{\mathcal{I}^\circ}(\tilde{z}_{\mathcal{M}^\perp}) &\leq h_{\mathcal{I}^\circ}(P_{\mathcal{M}^\perp}(Q_n P_{\mathcal{M}}(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} \tilde{z}_{\mathcal{A}} - \tilde{z}_{\mathcal{A}})) \\ &\quad + \frac{1}{\lambda} h_{\mathcal{I}^\circ}(P_{\mathcal{M}^\perp}(Q_n P_{\mathcal{M}}(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger \nabla \ell_n^* - \nabla \ell_n^*)) \\ &\quad + \frac{1}{\lambda} h_{\mathcal{I}^\circ}(P_{\mathcal{M}^\perp}(Q_n P_{\mathcal{M}}(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} R_n - R_n)).\end{aligned}$$

The irrepresentable condition implies the first term is small:

$$h_{\mathcal{I}^\circ}(P_{\mathcal{M}^\perp}(Q_n P_{\mathcal{M}}(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} \tilde{z}_{\mathcal{A}} - \tilde{z}_{\mathcal{A}})) \leq 1 - \delta.$$

Thus

$$h_{\mathcal{I}^\circ}(\tilde{z}_{\mathcal{M}^\perp}) \leq 1 - \delta + \kappa_{\text{ir}} \Big( \frac{\bar{\rho}^*(\nabla \ell_n^*)}{\lambda} + \frac{\bar{\rho}^*(R_n)}{\lambda} \Big).$$

If $\lambda$ is between (2.10), then $\frac{\kappa_{\text{ir}}}{\lambda} \bar{\rho}^*(\nabla \ell_n^*) \leq \frac{\delta}{4}$ and

$$h_{\mathcal{I}^\circ}(\tilde{z}_{\mathcal{M}^\perp}) < 1 - \delta + \frac{\delta}{4} + \frac{\kappa_{\text{ir}}}{\lambda} \bar{\rho}^*(R_n). \tag{2.14}$$

**Lemma 2.9.** *Under the conditions of Lemma 2.7, if $\ell_n$ is also strongly smooth on $\mathcal{C} \cap \mathcal{M}$ and $\lambda$ is between (2.10), $\frac{\kappa_{\text{ir}}}{\lambda} \bar{\rho}^*(R_n) < \frac{\delta}{4}$.*

We substitute the bound into (2.14) to obtain

$$h_{\mathcal{I}^\circ}(\tilde{z}_{\mathcal{M}^\perp}) < 1 - \delta + \frac{\delta}{4} + \frac{\delta}{4} \leq 1 - \frac{\delta}{2} < 1.$$

Thus $\tilde{z}_{\mathcal{M}^\perp}$ is PDW feasible. By Lemma 2.8 and the uniquenss of the solution to the restricted problem, $\tilde{\theta}$ is the unique solution to the original problem. □

## 2.4 THE NECESSITY OF IRREPRESENTABILITY

Although the irrepresentable condition seems cryptic and hard to verify, Zhao and Yu (2006) and Wainwright (2009) showed it is necessary for sign consistency of the lasso.[2] In this section, we give necessary conditions for an M-estimator with a geometrically decomposable penalty to be both consistent and model selection consistent.

[2] Zhao and Yu (2006) and Wainwright (2009) refer to the (slightly) stronger condition (2.9) as irrepresentability. Thus their results are often summarized as irrepresentability is "almost" necessary for model selection consistency.

**Theorem 2.10.** *Suppose*

1. *the loss function $\ell_n$ is strongly convex on $\mathcal{C} \cap \mathcal{M}$,*
2. *the loss function satisfies the irrepresentable condition,*
3. *the optimal solution to (2.6) is unique, consistent, and model selection consistent, i. e. $\hat{\theta} \in (\theta^* + r\mathcal{B}_2) \cap \mathcal{M}$.*

*Then*

$$\begin{aligned} &P_{\mathcal{M}^\perp} Q_n P_{\mathcal{M}} (P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger (\nabla \ell_n(\theta^*) + \lambda \hat{z}_{\mathcal{A}} + R_n) \\ &\quad \in P_{\mathcal{M}^\perp}(\nabla \ell_n(\theta^*) + R_n + \lambda(\hat{z}_{\mathcal{A}} + \mathcal{I})) \end{aligned}$$

*for some $\hat{z}_{\mathcal{A}} \in \partial h_{\mathcal{A}}((\theta^* + r\mathcal{B}_2) \cap \mathcal{M})$, where $R_n = \nabla \ell_n(\hat{\theta}) - \nabla \ell_n(\theta^*) - Q_n(\hat{\theta} - \theta^*)$ is the Taylor remainder term.*

*Proof.* The proof proceeds like the proof of Theorem 2.6. The optimal solution to (2.6) satisfies (2.13). By assumption, $\hat{\theta} \in (\theta^* + rB_2) \cap \mathcal{M}$. We solve for the error to obtain

$$\hat{\theta} - \theta^* = -(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} (\nabla \ell_n(\theta^*) + \lambda \hat{z}_A + R_n).$$

Substituting in the expression for the error into (2.13),

$$0 = \nabla \ell_n(\theta^*) - Q(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} (\nabla \ell_n(\theta^*) + \lambda \hat{z}_A + R_n) + R_n + \lambda(\hat{z}_A + \hat{z}_{\mathcal{I}} + \hat{z}_{E^\perp}).$$

We project onto $M^\perp$ to obtain the stated result. □

Theorem 2.10 is a deterministic statement concerning the optimal solution of (2.6). It says

$$P_{\mathcal{M}^\perp}(\nabla \ell_n(\theta^*) + R_n) - P_{\mathcal{M}^\perp} Q_n P_{\mathcal{M}} (P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger (\nabla \ell_n(\theta^*) + R_n) \tag{2.15}$$

falls in the set

$$\begin{aligned} &P_{\mathcal{M}^\perp}(\partial h_{\mathcal{A}}((\theta^* + r\mathcal{B}_2) \cap \mathcal{M}) + \mathcal{I}) \\ &\quad - P_{\mathcal{M}^\perp} Q_n P_{\mathcal{M}} (P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger \partial h_{\mathcal{A}}((\theta^* + r\mathcal{B}_2) \cap \mathcal{M}). \end{aligned} \tag{2.16}$$

To deduce the necessity of irrepresentability, it suffices to show the claims of Lemma 2.10 are invalid with non-zero probability when irrepresentability is violated. Although the distribution of (2.15) is generally hard to characterize, it suffices to show the distribution is symmetric, i. e.

$$\mathbf{Pr}((2.15) \in \mathcal{C}) = \mathbf{Pr}((2.15) \in -\mathcal{C}) \text{ for any measurable set } \mathcal{C}.$$

**Corollary 2.11.** *Under the conditions of Theorem 2.10, if we also assume*

1. *the set $\mathcal{A}$ is a convex polytope,*
2. *the distribution of (2.15) is symmetric,*
3. *the unknown parameter $\theta^*$ is in $\bigcup_{\theta \in \text{ext}(\mathcal{A})} \text{relint}(N_{\mathcal{A}}(\theta))$.*

*When irrepresentability is violated—say*

$$\inf_{z \in \partial h_{\mathcal{A}}(\theta^*)} h_{\mathcal{I}^\circ}(P_{\mathcal{M}^\perp}(Q_n P_{\mathcal{M}}(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger P_{\mathcal{M}} z - z)) \geq 1,$$

$$\mathbf{Pr}(\hat{\theta} \in (\theta^* + r\mathcal{B}_2) \cap \mathcal{M}) \leq \tfrac{1}{2}$$

*for any $r$ small enough such that $\theta^* + r\mathcal{B}_2 \subset \bigcup_{x \in \text{ext}(\mathcal{A})} \text{relint}(N_{\mathcal{A}}(x))$.*

*Proof.* Since $\theta^* \in \bigcup_{x \in \text{ext}(A)} \text{relint}(N_{\mathcal{A}}(x))$, $\partial h_{\mathcal{A}}(\theta^*)$ is a point. Call the point $\hat{z}^*_{\mathcal{A}}$. For any $r$ small enough such that

$$\theta^* + r\mathcal{B}_2 \subset \bigcup_{x \in \text{ext}(\mathcal{A})} \text{relint}(N_{\mathcal{A}}(x)),$$

$\partial h_{\mathcal{A}}((\theta^* + r\mathcal{B}_2) \cap \mathcal{M})$ is also the point $\hat{z}^*_{\mathcal{A}}$. Thus (2.16) is given by

$$P_{\mathcal{M}^\perp}(\partial h_{\mathcal{A}}(\theta^*) + \mathcal{I}) - P_{\mathcal{M}^\perp} Q_n P_{\mathcal{M}}(P_{\mathcal{M}} Q_n P_{\mathcal{M}})^\dagger \hat{z}^*_{\mathcal{A}}. \tag{2.17}$$

When irrepresentability is violated, (2.17) is a convex set that does not contain a relative neighborhood of the origin. Thus there is a halfspace (through the origin) that contains (2.17). Since the distribution of (2.15) is symmetric, $\mathbf{Pr}((2.15) \in (2.16)) \leq \frac{1}{2}$. □

Although Corollary 2.11 "justifies" the irrepresentable condition, the necessity of the condition offers little comfort to practitioners whose predictors are often correlated. Jia and Rohe (2012) propose *preconditioning* regression problems to conform to the irrepresentable condition. They show their technique improves the performance of a broad class of model selection techniques in linear regression.

# 3 RANK CONSISTENCY OF NUCLEAR NORM MINIMIZATION

Geometric decomposability, although general, excludes some widely used regularizers. In this chapter, we turn our attention to *weakly decomposable* regularizers; i. e. regularizers that are well-approximated by sums of support functions. The example we have in mind is low-rank multivariate regression. Consider the (multivariate) linear model

$$Y = X\Theta^* + W, \tag{3.1}$$

where the rows of $Y \in \mathbf{R}^{n \times p_2}$ are (multivariate) responses. When $\Theta^* \in \mathbf{R}^{p_1 \times p_2}$ is low-rank, a natural approach to estimating $\Theta^*$ is *nuclear norm minimization:*

$$\underset{\Theta \in \mathbf{R}^{p_1 \times p_2}}{\text{minimize}} \ \frac{1}{2n} \|Y - X\Theta\|_F^2 + \lambda \|\Theta\|_{\text{nuc}}, \tag{3.2}$$

where the nuclear norm is given by (1.9). Bach (2008) showed that nuclear norm minimization is *rank consistent*, i. e.

$$\mathbf{Pr}\big(\text{rank}(\hat{\Theta}) = \text{rank}(\Theta^*)\big) \to 1, \tag{3.3}$$

subject to irrepresentability. Although rank consistency is not an instance of our notion of model selection consistency because the set of rank $r$ matrices is not a subspace, our results may be used to derive a non-asymptotic form of Bach's rank consistency result.

## 3.1 WEAKLY DECOMPOSABLE REGULARIZERS

To study the rank consistency of nuclear norm minimization, we consider a weaker notion of decomposability: *weak decomposability.* Our notion of weak decomposability was inspired by the notion proposed by van de Geer (2012). Although similar in spirit, our notion is more general. In particular, it does not depend on the component-wise separability of the regularizer.

**Definition 3.1** (Weakly decomposability)**.** *For any two closed convex sets $\mathcal{A}, \mathcal{I} \subset \mathbf{R}^p$ containing the origin, a regularizer is* weakly decomposable *with respect to the pair $(\mathcal{A}, \mathcal{I})$ at $\theta^* \in \mathbf{R}^p$ if*

$$\partial\rho(\theta^*) = \partial h_{\mathcal{A}}(\theta^*) + \partial h_{\mathcal{I}}(\theta^*).$$

*We assume $\mathcal{A}$ is bounded and $0 \in \operatorname{relint}(\mathcal{I})$.*

Weak decomposability is more general than geometric decomposability. However, the structure of the subdifferential of a weakly decomposable penalty at $\theta^*$ is very similar to that of a geometrically decomposable penalty. Consequently, the directional derivative of $\rho$ at $\theta^*$ along any $\Delta$ is geometrically decomposable:

$$\rho'(\theta^*, \Delta) = h_{\partial h_{\mathcal{A}}(\theta^*)}(\Delta) + h_{\partial h_{\mathcal{I}}(\theta^*)}(\Delta).$$

As we shall see, the geometric decomposability of $\rho'(\theta^*, \Delta)$ is the key to the model selection properties of weakly decomposable penalties.

The setup is similar to the setup in Chapter 2. We are given a samples $\mathcal{Z}^n := \{z_1, \dots, z_n\}$ with marginal distribution $\mathbf{P}$. We seek to estimate a parameter $\theta^* \in \mathcal{M} \subset \mathbf{R}^p$, where $\mathcal{M}$ is the *model manifold*. To keep things simple, we focus on regularized least squares:

$$\underset{\theta \in \mathbf{R}^p}{\text{minimize}} \ \frac{1}{2}\theta^T Q_n \theta - c_n^T \theta + \lambda\rho(\theta), \tag{3.4}$$

where $\rho$ is weakly decomposable with respect to sets $\mathcal{A}, \mathcal{I} \subset \mathbf{R}^p$ at $\theta^*$. The set $\mathcal{I}$ is chosen so that the *tangent space* of $\mathcal{M}$ at $\theta^*$ is $\operatorname{span}(\mathcal{I})^\perp$. That is, $\operatorname{span}(\mathcal{I})$ contains deviations from $\theta^*$ that we wish to kill.

## 3.2 DUAL CONSISTENCY OF REGULARIZED M-ESTIMATORS

To study the model selection properties of (3.4), we compare its optimum to that of a linearized problem

$$\underset{\theta \in \mathbf{R}^p}{\text{minimize}} \ \frac{1}{2}\theta^T Q_n \theta - c_n^T \theta + \lambda(\rho(\theta^*) + \rho'(\theta^*, \theta - \theta^*)). \tag{3.5}$$

Since the objective functions of (3.5) and (3.4) are similar, we expect the (optimal) solutions of be close. Unfortunately, due to the lack of strong convexity, we cannot conclude the solutions are close. However, as we shall see, the dual solutions are close.

After a change of variables, the linearized problem is

$$\underset{\Delta \in \mathbf{R}^p}{\text{minimize}} \ \frac{1}{2}\Delta^T Q_n \Delta + (Q_n\theta^* - c_n)^T\Delta + \lambda(h_{\partial h_{\mathcal{A}}(\theta^*)}(\Delta) + h_{\mathcal{I}}(\Delta)). \tag{3.6}$$

We recognize (3.6) is an M-estimator with a geometrically decomposable regularizer. Under the conditions of Theorem 2.6, there is a unique primal-dual pair $(\tilde{\Delta}, \tilde{z}_{\mathcal{A}}, \tilde{z}_{\mathcal{I}})$ that satisfies

$$\begin{gathered} Q_n(\theta^* + \tilde{\Delta}) - c_n + \lambda(\tilde{z}_{\mathcal{A}} + \tilde{z}_{\mathcal{I}}) = 0 \\ \tilde{z}_{\mathcal{A}} \in \partial h_{\mathcal{A}}(\theta^*), \quad \tilde{z}_{\mathcal{I}} \in \mathcal{I}. \end{gathered} \tag{3.7}$$

Further, $\tilde{\Delta}$ is consistent and $\tilde{z}_{\mathcal{I}}$ is PDW feasible. We summarize the properties of $(\tilde{\Delta}, \tilde{z}_{\mathcal{A}}, \tilde{z}_{\mathcal{I}})$ in a lemma.

**Lemma 3.2.** *When the restricted eigenvalues of $Q_n$ on $\operatorname{span}(\mathcal{I})^\perp$ are at least $\mu_l$, $Q_n$ satisfies the irrepresentable condition, and the regularization parameter $\lambda$ is at least $\frac{4\kappa_{\text{ir}}}{\delta}\rho'^*(Q_n\theta^* - c_n)$, the unique primal-dual pair of (3.6) $(\tilde{\Delta}, \tilde{z}_{\mathcal{A}}, \tilde{z}_{\mathcal{I}})$ is*

1. *consistent: $\|\tilde{\Delta}\|_2 \le \frac{2}{\mu_l}(\kappa_{\rho'} + \frac{\delta\kappa_{\rho'}}{4\kappa_{\text{ir}}})\lambda$;*
2. *PDW feasible: $h_{\mathcal{I}^\circ}(\tilde{z}_{\mathcal{I}}) \le 1 - \frac{\delta}{2}$.*

The main result of this chapter shows the optimal dual solutions of (3.4) and (3.6) are close.

**Theorem 3.3.** *Under the conditions of Lemma 3.2, the optimal dual solutions of (3.5) and (3.4) satisfy*

$$\|\tilde{z}_{\mathcal{A}} + \tilde{z}_{\mathcal{I}} - \hat{z}\|_2^2 \le \tfrac{\|Q_n\|_2}{\lambda}(R_\rho(\tilde{\Delta}) - R_\rho(\hat{\Delta})),$$

*where $R_\rho(\Delta) = \rho(\theta^* + \Delta) - \rho(\theta^*) - \rho'(\theta^*, \Delta)$.*

*Proof.* After a change of variables, the original problem is

$$\underset{\Delta \in \mathbf{R}^p}{\text{minimize}} \ \frac{1}{2}\Delta^T Q \Delta + (Q_n\theta^* - c_n)^T\Delta + \lambda\rho(\theta^* + \Delta). \tag{3.8}$$

Its optimality conditions are

$$\begin{gathered} Q_n(\theta^* + \hat{\Delta}) - \gamma + \lambda\hat{z} = 0 \\ \hat{z} \in \partial\rho(\theta^* + \hat{\Delta}). \end{gathered} \tag{3.9}$$

Let $\tilde{\Delta}$ and $\hat{\Delta}$ be the optimums of (3.6) and (3.8). By Fermat's rule, $\tilde{z}_{\mathcal{A}} + \tilde{z}_{\mathcal{I}}$ and $\hat{z}_{\mathcal{A}} + \hat{z}_{\mathcal{I}}$ are also the optimal dual solutions of (3.5) and (3.4). We subtract (3.9) from (3.7) to obtain

$$Q_n(\hat{\Delta} - \tilde{\Delta}) = \lambda(\tilde{z}_{\mathcal{A}} + \tilde{z}_{\mathcal{I}} - \hat{z}). \tag{3.10}$$

To complete the proof, we show $\left\|Q_n(\hat{\Delta} - \tilde{\Delta})\right\|_2^2$ is small. By inspection of the optimality conditions (3.9) and (3.7), $\tilde{\Delta}$ and $\hat{\Delta}$ are also the solutions of

$$\begin{aligned}
&\underset{\Delta \in \mathbf{R}^p}{\text{minimize}}\, \tilde{\Delta}^T Q_n \Delta + (Q_n\theta^* - c_n)^T\Delta + \lambda\rho'(\theta^*, \Delta), \\
&\underset{\Delta \in \mathbf{R}^p}{\text{minimize}}\, \hat{\Delta}^T Q_n \Delta + (Q_n\theta^* - c_n)^T\Delta + \lambda(\rho(\theta^* + \Delta) - \rho(\theta^*)).
\end{aligned}$$

Since $\tilde{\Delta}$ and $\hat{\Delta}$ are their respective optimums, we know

$$\begin{aligned}
&\tilde{\Delta}^T Q_n \tilde{\Delta} + (Q_n\theta^* - c_n)^T\tilde{\Delta} + \lambda\rho'(\theta^*, \tilde{\Delta}) \\
&\quad \le \tilde{\Delta}^T Q_n \hat{\Delta} + (Q_n\theta^* - c_n)^T\hat{\Delta} + \lambda\rho'(\theta^*, \hat{\Delta}), \\
&\hat{\Delta}^T Q_n \hat{\Delta} + (Q_n\theta^* - c_n)^T\hat{\Delta} + \lambda(\rho(\theta^* + \hat{\Delta}) - \rho(\theta^*)) \\
&\quad \le \hat{\Delta}^T Q_n \tilde{\Delta} + (Q_n\theta^* - c_n)^T\tilde{\Delta} + \lambda\left(\rho(\theta^* + \tilde{\Delta}) - \rho(\theta^*)\right).
\end{aligned}$$

We add the inequalities and rearrange to obtain

$$(\tilde{\Delta} - \hat{\Delta})^T Q_n (\tilde{\Delta} - \hat{\Delta}) = \|\Delta\|_Q^2 \le \lambda\big(R_\rho(\tilde{\Delta}) - R_\rho(\hat{\Delta})\big),$$

where $R_\rho(\Delta) = \rho(\theta^* + \Delta) - \rho(\theta^*) - \rho'(\theta^*, \Delta)$. Since $\|Q_n\Delta\|_2^2 \le \|Q_n\|_2 \|\Delta\|_{Q_n}^2$,

$$\left\|Q_n(\hat{\Delta} - \tilde{\Delta})\right\|_2^2 \le \|Q_n\|_2 \left\|\hat{\Delta} - \tilde{\Delta}\right\|_{Q_n}^2 \le \|Q_n\|_2\, \lambda\big(R_\rho(\tilde{\Delta}) - R_\rho(\hat{\Delta})\big).$$

We substitute in (3.10) to obtain the stated conclusion. □

## 3.3 RANK CONSISTENCY OF LOW-RANK MULTIVARIATE REGRESSION

We return to the low-rank multivariate regression problem. The nuclear norm is weakly decomposable. Let $\Theta^* = U\Sigma V^T$ be the (full) SVD of $\Theta^*$ and define the sets

$$\begin{aligned}
\mathcal{A} &= \left\{\Theta \in \mathcal{B}_2 \subset \mathbf{R}^{p_1 \times p_2} \mid \Theta = U_r D V_r^T \text{ for some diagonal } D\right\}, \\
\mathcal{I} &= \left\{\Theta \in \mathcal{B}_2 \subset \mathbf{R}^{p_1 \times p_2} \mid \Theta = U_{p_1 - r} D V_{p_2 - r}^T \text{ for some diagonal } D\right\},
\end{aligned}$$

where $r = \operatorname{rank}(\Theta^*)$ and $U_r, U_{p_1-r}$ (resp. $V_r, V_{p_2-r}$) are the submatrices of $U$ (resp. $V$) consisting of the first $r$ and last $p_1 - r$ left (resp. $p_2 - r$ right) singular vectors of $\Theta^*$. It is not hard to check that the nuclear norm is weakly decomposable at $\Theta^*$ in terms of $\mathcal{A}, \mathcal{I}$. Since $\mathcal{A} + \mathcal{I} \subset \mathcal{B}_2$,

$$\|\Theta\|_{\text{nuc}} = h_{\mathcal{B}_2}(\Theta) \geq h_{\mathcal{A}}(\Theta) + h_{\mathcal{I}}(\Theta).$$

Before we delve into the rank consistency of low-rank multivariate regression, we state the assumptions on the problem. Let $\vec{X} \in \mathbf{R}^{p_1 p_2}$ be the vectorized form of $X \in \mathbf{R}^{p_1 \times p_2}$. In vector notation, the model is

$$\vec{Y} = X(\Theta^*) + \vec{W}, \tag{3.11}$$

where $X : \mathbf{R}^{p_1 \times p_2} \to \mathbf{R}^n$ is a linear mapping. Since $X$ is a linear map, we abuse notation by writing $\vec{Y} = X\vec{\Theta} + \vec{W}$. The Fisher information $Q_n : \mathbf{R}^{p_1 \times p_2} \to \mathbf{R}^{p_1 \times p_2}$ is given by $\frac{1}{n} X^* X$. We assume

1. the restricted eigenvalues of $Q_n$ on $\operatorname{span}(\mathcal{I})^\perp$ are at least $\mu_l$,
2. the predictors satisfy the strong irrepresentable condition:

   $$\sup_{Z \in \mathcal{B}_2} \big\| U_{p_1-r}^T \big[ P_{\mathcal{I}} Q_n P_{\mathcal{I}^\perp} (P_{\mathcal{I}^\perp} Q_n P_{\mathcal{I}^\perp})^\dagger Z \big] V_{p_2-r} \big\|_2 \leq 1 - \delta, \tag{3.12}$$

   where $P_{\mathcal{I}} : \mathbf{R}^{p_1 \times p_2} \to \mathbf{R}^{p_1 \times p_2}$ (resp. $P_{\mathcal{I}^\perp}$) is the projector onto $\operatorname{span}(\mathcal{I})$ (resp. $\operatorname{span}(\mathcal{I})^\perp$).
3. the entries of $W$ are *i.i.d.* subgaussian random variables with mean zero and subgaussian norm $\sigma$.

As its name suggests, assumption (3.12) is stronger than irrepresentability. It implies irrepresentability:

$$\begin{aligned} &\big\| U_{p_1-r}^T \big[ P_{\mathcal{I}} \big( Q_n P_{I^\perp} (P_{I^\perp} Q_n P_{I^\perp})^\dagger U_r V_r^T - U_r V_r^T \big) \big] V_{p_2-r} \big\|_2 \\ &\quad = \big\| U_{p_1-r}^T \big[ P_{\mathcal{I}} \big( Q_n P_{I^\perp} (P_{I^\perp} Q_n P_{I^\perp})^\dagger U_r V_r^T \big) \big] V_{p_2-r} \big\|_2 \\ &\quad \leq \sup_{Z \in \mathcal{B}_2} \big\| U_{p_1-r}^T \big[ P_{\mathcal{I}} Q_n P_{I^\perp} (P_{I^\perp} Q_n P_{I^\perp})^\dagger Z \big] V_{p_2-r} \big\|_2. \end{aligned} \tag{3.13}$$

We make the stronger assumption to obtain an explicit expression for the constant $\kappa_{\text{ir}}$ (in terms of the constant $\delta$).

The final ingredient we require is a "Taylor's theorem" for the nuclear norm that says the nuclear norm is well-approximated by its linearization.

**Lemma 3.4.** *For any $\Delta \in \operatorname{span}(\mathcal{I})^\perp$, $\|\Delta\|_2 < \frac{\sigma_r^*}{2}$, we have*

$$\|\Theta^* + \Delta\|_{\text{nuc}} - \|\Theta^*\|_{\text{nuc}} - \operatorname{tr}\big(U_r^T \Delta V_r\big) \leq \frac{4}{3\sigma_r^*} \|\Delta\|_F^2,$$

*where $\sigma_r^*$ is the smallest non-zero singular value of $\Theta^*$.*

We put the pieces togther to deduce the rank-consistency of low-rank multivariate regression.

**Corollary 3.5.** *Under the aforementioned conditions, the optimum of (3.2) with regularization parameter $\lambda = \frac{8(2-\delta)}{\delta}\nu\big(\frac{p_1+p_2}{n}\big)^{\frac{1}{2}}$ is unique and rank consistent when*

$$n > \max\Big\{\frac{128^2M^2(\sqrt{2}+\delta')^4}{9\sigma_r^{*2}\mu_l^4\delta^4\delta'^2}r^2, \frac{16(\sqrt{2}+\delta')^2}{\mu_l^2}r\Big\}\nu^2(p_1+p_2)$$

*with probability at least $1 - c_1e^{-c_2(p_1+p_2)}$. The constants $M$ and $\delta'$ are given by $\sup_{\Delta\in\mathcal{B}_{\mathrm{F}}}\|Q_n\Delta\|_{\mathrm{F}}$ and $\frac{4(2-\delta)}{\delta}$.*

*Proof.* To show $\hat{\Theta}$ has rank at most $r$, it suffices to show the optimal dual solution $\hat{U}\hat{V}^T$ has no more than $r$ non-zero singular values. At a high level, the proof consists of three steps:

1. Show that the unique primal-dual pair to a linearized problem $\big(\tilde{\Delta}, U_rV_r^T, \tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T\big)$ is consistent and PDW feasible.
2. By Theorem 3.3, $\hat{U}\hat{V}^T$ is close to the optimal dual solution of the linearized problem $U_rV_r^T + \tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T$. Since $\tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T$ is PDW feasible, its singular values are bounded away from one.
3. Apply a singular value perturbation result to deduce $\hat{U}\hat{V}^T$ has (no more than) $r$ unit singular values.

Consider the linearized problem

$$\underset{\Delta\in\mathbf{R}^{p_1\times p_2}}{\text{minimize}}\ \frac{1}{2n}\|Y - X(\Theta^*+\Delta)\|_{\mathrm{F}}^2 + \lambda\big(\mathrm{tr}\big(U_r^T\Delta V_r\big) + \|U_{p_1-r}^T\Delta V_{p_2-r}\|_*\big). \tag{3.14}$$

By Lemma 3.2, a primal-dual pair $(\tilde{\Delta}, U_rV_r^T, \tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T)$ that satisfies

$$\begin{gathered} Q_n(\Theta^*+\tilde{\Delta}) - c_n + \lambda(U_rV^T + \tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T) = 0, \\ \tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T \in I \end{gathered}$$

is unique, consistent, and PDW feasible.

**Lemma 3.6.** *Under the aforementioned conditions, the unique primal-dual pair of (3.14) with regularization parameter $\lambda = \frac{8(2-\delta)}{\delta}\sigma\big(\frac{p_1+p_2}{n}\big)^{\frac{1}{2}}$ is*

1. *consistent: $\|\tilde{\Delta}\|_{\mathrm{F}} \le \frac{4}{m}\Big(\sqrt{2}+\frac{4(2-\delta)}{\delta}\Big)\nu\big(\frac{r(p_1+p_2)}{n}\big)^{\frac{1}{2}}$.*

2. *PDW feasible:* $\left\|\tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T\right\|_2 \le 1 - \frac{\delta}{2}$.

By Theorem 3.3 (and the convexity of the nuclear norm),

$$\begin{aligned}&\left\|\hat{U}\hat{V}^T - U_rV_r^T - \tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T\right\|_2^2 \\ &\quad\le \|\hat{U}\hat{V}^T - U_rV_r^T - \tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T\|_{\mathrm{F}}^2 \\ &\quad\le \frac{M}{\lambda}\big(R(\tilde{\Delta}) - R(\hat{\Delta})\big) \le \frac{M}{\lambda}R(\tilde{\Delta}),\end{aligned}$$

where $M := \sup_{\|\Delta\|_{\mathrm{F}}\le 1}\|Q\Delta\|_{\mathrm{F}}$. Since $\tilde{\Delta} \in \mathrm{span}(I)^{\perp}$, by Lemma 3.4,

$$\left\|\hat{U}\hat{V}^T - U_rV_r^T - \tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T\right\|_2^2 \le \frac{4}{3\sigma_r^*}\frac{M}{\lambda}\|\tilde{\Delta}\|_{\mathrm{F}}^2$$

as long as $\|\tilde{\Delta}\|_2 \le \frac{\sigma_r^*}{2}$. By the consistency of the linearized problem,

$$\|\tilde{\Delta}\|_2 \le \|\tilde{\Delta}\|_{\mathrm{F}} \le \frac{4}{\mu_l}(\sqrt{2}+\delta')\nu\Big(\frac{r(p_1+p_2)}{n}\Big)^{\frac{1}{2}},$$

where $\delta' = \frac{4(2-\delta)}{\delta}$. We put the pieces together to obtain

$$\begin{aligned}&\left\|\hat{U}\hat{V}^T - U_rV_r^T - \tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T\right\|_2^2 \\ &\quad\le \frac{32}{3\sigma_r^*}\frac{M}{\mu_l^2}\frac{(\sqrt{2}+\delta')^2}{\delta'}\nu r\Big(\frac{(p_1+p_2)}{n}\Big)^{\frac{1}{2}},\end{aligned} \tag{3.15}$$

when $n > \frac{16}{\mu_l^2}\frac{\nu^2}{\sigma_r^{*2}}(\sqrt{2}+\delta')^2 r(p_1+p_2)$.

By Lemma 3.2, $\tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T$ is PDW feasible. Thus it has at most $r$ unit singular values. Its $p-r$ remaining singular values are smaller than $1-\frac{\delta}{2}$. By Weyl's inequality, it suffices to ensure

$$\left\|\hat{U}\hat{V}^T - U_rV_r^T - \tilde{U}_{p_1-r}\tilde{V}_{p_2-r}^T\right\|_2 \le \frac{\delta}{2} \tag{3.16}$$

to ensure $\tilde{U}\tilde{V}^T$ has no more than $r$ unit singular values. We combine (3.15) and (3.16) to deduce the requirement on $n$. □

# Part II

# COMPUTING

# 4 PROXIMAL NEWTON-TYPE METHODS

In the second part of the thesis, we turn our attention to evaluating regularized M-estimators, which require minimizing *composite functions* of the form

$$\operatorname*{minimize}_{x\in\mathbf{R}^p} \phi(x) := \phi_{\text{sm}}(x) + \phi_{\text{ns}}(x) \tag{4.1}$$

where $\phi_{\text{sm}} : \mathbf{R}^p \to \mathbf{R}$ convex, twice continuously differentiable, and $\phi_{\text{ns}} : \mathbf{R}^p \to \mathbf{R}$ is convex but not necessarily differentiable. The material in this and the subsequent chapter appears in Lee et al. (2014).

Optimization methods are broadly classified into first- and second-order methods, depending on whether they incorporate second-order (curvature) information to guide the optimization. Second-order methods usually take fewer iterations to converge, with the caveat that each iteration is more expensive. First-order methods tend to scale better to the large-scale problems that arise in modern statistics, making them especially appealing to practitioners.

Most first-order methods for minimizing composite functions form the next iterate $x_{t+1}$ from the current iterate $x_t$ by forming a simple quadratic approximation (the quadratic term is a multiple of $I$) to the smooth part:

$$\hat{\phi}_{\text{sm},t}(x) = \hat{\phi}_{\text{sm}}(x_t) + \nabla\hat{\phi}_{\text{sm}}(x_t)^T(x - x_t) + \frac{1}{2\alpha_t}\|x - x_t\|_2^2,$$

where $\alpha_t > 0$ is a step size, and setting

$$x_{t+1} \leftarrow \arg\min_x \hat{\phi}_{\text{sm},t}(x) + \phi_{\text{ns}}(x).$$

The efficiency of first-order methods depends on the cost of evaluating $\nabla\hat{\phi}_{\text{sm}}$ and that of minimizing $\hat{\phi}_{\text{sm},t} + \phi_{\text{ns}}$. For many regularizers of interest, it is possible to solve the subproblem in closed form.

In this chapter, we describe a family of methods that "interpolate" between first- and second-order methods. The methods can be interpreted as generalizations of first-order proximal methods that incorporate curvature information in the subproblem. To set the stage, we give an overview of proximal methods.

## 4.1 BACKGROUND ON PROXIMAL METHODS

The proximal mapping of a convex function $\phi$ at $x$ is

$$\operatorname{prox}_{\phi}(x) := \arg\min_{y \in \mathbf{R}^p} \phi(y) + \frac{1}{2} \|y - x\|_2^2 . \tag{4.2}$$

Proximal mappings can be interpreted as generalized projections because if $\phi$ is the indicator function of a convex set, $\operatorname{prox}_{\phi}(x)$ is the projection of $x$ onto the set.

The *proximal gradient method* alternates between taking gradient descent steps to optimize the smooth part and taking a proximal step to optimize the nonsmooth part. More precisely, the proximal gradient iteration is $x_{t+1} \leftarrow \operatorname{prox}_{\alpha_t \phi_{\text{ns}}}(x_t - \alpha_t \nabla \phi_{\text{sm}}(x_t))$, where $\alpha_t > 0$ is a step size. Equivalently,

$$\begin{aligned} x_{t+1} &\leftarrow x_t - \alpha_t g_{\alpha_t}(x_t) \\ g_{\alpha_t}(x_t) &:= \frac{1}{\alpha_t}\left(x_t - \operatorname{prox}_{\alpha_t \phi_{\text{ns}}}(x_t - \alpha_t \nabla \phi_{\text{sm}}(x_t))\right), \end{aligned} \tag{4.3}$$

where $g_{\alpha_t}(x_t)$ is a *composite gradient step*. The composite gradient step at $x$ is zero if and only if $x$ is an optimum of $\phi$; i. e. $g(x) = 0$, where $g(x) := x - \operatorname{prox}_{\phi_{\text{ns}}}(x - \nabla \phi_{\text{sm}}(x))$ generalizes the familiar zero gradient optimality condition to composite functions.

Most first-order methods are variants of the proximal gradient method. A popular method is SpaRSA by Wright et al. (2009), which combines a *spectral step size* with a *nonmonotone line search* to improve convergence. It is also possible to accelerate the convergence rate of first-order methods using ideas in Nesterov (2003). The resulting methods, aptly called *accelerated first-order methods*, achieve $\epsilon$-suboptimality within $O(1/\sqrt{\epsilon})$ iterations. The most popular methods in this family are Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) by Nesterov (2007).

Behind the scenes, the proximal gradient algorithm forms simple quadratic models of $\phi_{\text{sm}}$ near the current iterate:

$$\hat{\phi}_{\text{sm},t}(x) := \phi_{\text{sm}}(x_t) + \nabla \phi_{\text{sm}}(x_t)^T (x - x_t) + \frac{1}{2\alpha_t} \|x - x_t\|_2^2 .$$

The composite gradient step moves to the optimum of $\hat{\phi}_{\text{sm},t} + \phi_{\text{ns}}$ :

$$\begin{aligned} x_{t+1} &= \operatorname{prox}_{\alpha_t \phi_{\text{ns}}} (x_t - \alpha_t \nabla \phi_{\text{sm}}(x_t)) \\ &= \arg\min_x \alpha_t \phi_{\text{ns}}(x) + \frac{1}{2} \|x - x_t + \alpha_t \nabla \phi_{\text{sm}}(x_t)\|_2^2 \\ &= \arg\min_x \nabla \phi_{\text{sm}}(x_t)^T (x - x_t) + \frac{1}{2\alpha_t} \|x - x_t\|_2^2 + \phi_{\text{ns}}(x). \end{aligned} \tag{4.4}$$

By the optimality of the composite gradient step, it is possible to see the composite gradient step is neither a gradient nor a subgradient of $\phi$ at any point; rather it is the sum of an explicit gradient (at $x_t$) and an implicit subgradient (at $x_{t+1}$). By rearranging the optimality conditions of (4.4), we have

$$g_{\alpha_t}(x_t) \in \nabla \phi_{\text{sm},t}(x_t) + \partial \phi_{\text{ns}}(x_{t+1}).$$

## 4.2 PROXIMAL NEWTON-TYPE METHODS

Proximal Newton-type methods replace the simple quadratic form with a general quadratic form to incorporate curvature information (of $\phi_{\text{sm}}$) into the choice of search direction:

$$\hat{\phi}_{\text{sm},t}(x) = \phi_{\text{sm}}(x_t) + \nabla \phi_{\text{sm}}(x_t)^T (x - x_t) + \frac{1}{2}(y - x_t)^T H_t (x - x_t).$$

The basic idea can be traced back to the projected Newton method by *generalized proximal point method* by Fukushima and Mine (1981). A proximal Newton-type search direction $\Delta x_t$ solves the subproblem

$$\Delta x_t \leftarrow \arg\min_d \hat{\phi}_t(x_t + d) := \hat{\phi}_{\text{sm},t}(x_t + d) + \phi_{\text{ns}}(x_t + d). \tag{4.5}$$

There are many strategies for choosing $H_t$. If we set $H_t = \nabla^2 \phi_{\text{sm}}(x_t)$, we obtain the *proximal Newton method*. If we form $H_t$ according to a quasi-Newton strategy, we obtain a *proximal quasi-Newton method*. If the problem is large, we can use limited memory quasi-Newton updates to reduce memory usage. Generally speaking, most strategies for choosing Hessian approximations in Newton-type methods (for minimizing smooth functions) can be adapted to forming $H_t$ in proximal Newton-type methods.

When $H_t$ is not positive definite, we can also adapt strategies for handling indefinite Hessian approximations in Newton-type methods. The most simple strategy is Hessian modification: we add a multiple of the identity to $H_t$ when $H_t$ is indefinite. This makes the subproblem strongly convex and damps the search direction. In a proximal quasi-Newton method, we can also do update skipping: if an update causes $H_t$ to become indefinite, simply skip the update.

Many popular methods for minimizing composite functions are special cases of proximal Newton-type methods. Methods tailored to a specific problem include `glmnet` by Friedman et al. (2007), `newglmnet` by Yuan et al. (2012), QUIC by Hsieh et al. (2011), and the Newton-LASSO method by Olsen et al. (2012). Generic methods include *projected Newton-type methods* by Schmidt et al. (2009, 2011), proximal quasi-Newton methods by Schmidt (2010), Becker and Fadili (2012), and the method by Tseng and Yun (2009).

To highlight the connection between a proximal Newton-type search direction and the composite gradient step, we express the search direction in terms of *scaled proximal mappings*. This allows us to interpret the search direction as a "composite Newton step".

**Definition 4.1.** *Let $\phi$ be a convex function and $H$ be a positive definite matrix. The scaled proximal mapping of $\phi$ at $x$ is*

$$\operatorname{prox}_{\phi}^{H}(x) := \arg\min_{y\in\mathbf{R}^n} \phi(y) + \frac{1}{2}\left\|y-x\right\|_H^2.$$

Scaled proximal mappings share many properties with (unscaled) proximal mappings:

1. The scaled proximal point $\operatorname{prox}_{\phi}^{H}(x)$ exists and is unique for $x \in \operatorname{dom}\phi$ because the proximity function is strongly convex if $H$ is positive definite.

2. Let $\partial\phi(x)$ be the subdifferential of $\phi$ at $x$. Then $\operatorname{prox}_{\phi}^{H}(x)$ satisfies

$$H\big(x - \operatorname{prox}_{\phi}^{H}(x)\big) \in \partial\phi\big(\operatorname{prox}_{\phi}^{H}(x)\big).$$

3. Scaled proximal mappings are *firmly nonexpansive* in the $H$-norm. That is, if $u = \operatorname{prox}_{\phi}^{H}(x)$ and $v = \operatorname{prox}_{\phi}^{H}(y)$, then

$$(u-v)^T H(x-y) \geq \left\|u-v\right\|_H^2,$$

and the Cauchy-Schwarz inequality implies $\left\|u-v\right\|_H \leq \left\|x-y\right\|_H$.

We can express proximal Newton-type search directions as "composite Newton steps" using scaled proximal mappings:

$$\Delta x = \operatorname{prox}_{\phi_{\text{ns}}}^{H}\big(x - H^{-1}\nabla\phi_{\text{sm}}(x)\big) - x.$$

We use the second property of scaled proximal mappings to deduce that proximal Newton search directions satisfy

$$H\big(H^{-1}\nabla\phi_{\mathrm{sm}}(x) - \Delta x\big) \in \partial\phi_{\mathrm{ns}}(x + \Delta x).$$

We simplify to obtain

$$H\Delta x \in -\nabla\phi_{\mathrm{sm}}(x) - \partial\phi_{\mathrm{ns}}(x + \Delta x). \tag{4.6}$$

Thus proximal Newton-type search directions, like composite gradient steps, combine an explicit gradient with an implicit subgradient. This expression reduces to the Newton system when $\phi_{\mathrm{ns}} = 0$.

**Lemma 4.2.** *If $H$ is positive definite, then $\Delta x_t$ given by (4.5) satisfies*

$$\phi(x_{t+1}) \le \phi(x_t) + \alpha\big(\nabla\phi_{\mathrm{sm}}(x_t)^T\Delta x + \phi_{\mathrm{ns}}(x_t + \Delta x_t) - \phi_{\mathrm{ns}}(x_t)\big) + O(\alpha_t^2), \tag{4.7}$$

$$\nabla\phi_{\mathrm{sm}}(x_t)^T\Delta x + \phi_{\mathrm{ns}}(x_t + \Delta x_t) - \phi_{\mathrm{ns}}(x_t) \le -\Delta x_t^T H_t \Delta x_t. \tag{4.8}$$

Lemma 4.2 implies the search direction is a descent direction for $\phi$ because we can substitute (4.8) into (4.7) to obtain

$$\phi(x_{t+1}) \le \phi(x_t) - \alpha\Delta x_t^T H_t\Delta x_t + O(\alpha_t^2). \tag{4.9}$$

In a few special cases we can derive a closed-form expression for the proximal Newton search direction, but usually we must resort to an iterative method to solve the subproblem. The user should choose an iterative method that exploits the properties of $\phi_{\mathrm{ns}}$. For instance, if $\phi_{\mathrm{ns}}$ is the $\ell_1$ norm, coordinate descent methods combined with an active set strategy are known to be very efficient.

We suggest a line search procedure to select a step size $\alpha_t$ that satisfies a sufficient descent condition: the next iterate $x_{t+1}$ satisfies $\phi(x_{t+1}) \le \phi(x_t) + \frac{\alpha_t}{2}\delta_t$, where

$$\delta_t := \nabla\phi_{\mathrm{sm}}(x_t)^T\Delta x + \phi_{\mathrm{ns}}(x + \Delta x) - \phi_{\mathrm{ns}}(x_t). \tag{4.10}$$

A simple option is a *backtracking line search* that shortens the step until sufficient descent is achieved. Although simple, backtracking performs admirably in practice.

An alternative strategy is to search along the *proximal arc*, i. e., the arc/curve

$$\Delta x_t(\alpha) := \tilde{\phi}_{\mathrm{sm},t}(x) + \phi_{\mathrm{ns}}(x),$$

where

$$\tilde{\phi}_{\text{sm},t}(x) := \arg\min_x \nabla\phi_{\text{sm}}(x_t)^T(x - x_t) + \frac{1}{2\alpha}(x - x_t)^T H_t(x - x_t).$$

Arc search procedures have some benefits over line search procedures. When the optimal solution lies on a low-dimensional manifold of $\mathbf{R}^p$, an arc search strategy is likely to identify the manifold. The main drawback is the cost of obtaining trial points: a subproblem must be solved at each trial point.

**Lemma 4.3.** *When $H \succeq \mu_l I$ for some $\mu_l > 0$ and $\nabla\phi_{\text{sm}}$ is Lipschitz continuous with constant $\mu_u$, the sufficient descent condition is satisfied by any $\alpha \leq \min\{1, \frac{\mu_l}{\mu_u}\}$.*

**Algorithm 1** Proximal Newton-type method

```
Require: initial point $x_0 \in \operatorname{dom}\phi$
repeat
    choose $H_t$, a positive definite approximation to the Hessian
    solve the subproblem for a search direction:
      $\Delta x_t \leftarrow \arg\min_d \nabla\phi_{\text{sm}}(x_t)^T d + \frac{1}{2}d^T H_t d + \phi_{\text{ns}}(x_t + d)$
    select $\alpha_t$ with a line search
    update: $x_{t+1} \leftarrow x_t + \alpha_t \Delta x_t$
until stopping conditions are satisfied
```

## 4.3 CONVERGENCE OF THE PROXIMAL NEWTON AND PROXIMAL QUASI-NEWTON METHODS

We analyze the convergence behavior of proximal Newton-type methods when the subproblems are solved exactly. We show that proximal Newton-type methods and proximal quasi-Newton methods converge quadratically and superlinearly subject to standard assumptions on the smooth part $\phi_{\text{sm}}$.

To begin, we show proximal Newton-type methods converge globally to some optimal solution $x^*$. There are many similar results; e.g., those in (Patriksson, 1999, section 4), and Theorem 4.4 is neither the first nor the most general. We include the result because the proof is simple and intuitive.

We assume

1. the function $\phi$ is closed and the minimum is attained;

2. the $H_t$'s are (uniformly) positive definite; i. e. $H_t \succeq \mu_l I$ for some $\mu_l > 0$.

The second assumption ensures the methods are executable, i. e. by Lemma 4.3, there are step sizes that satisfy the sufficient descent condition.

**Theorem 4.4.** *Under the aforementioned assumptions, the sequence $\{x_t\}$ converges to an optimum of $\phi$ from any initial point $x_0 \in \operatorname{dom} \phi$.*

*Proof.* By (4.9) and Lemma 4.3, the sequence $\{\phi(x_t)\}$ is decreasing:

$$\phi(x_t) - \phi(x_{t+1}) \leq \frac{\alpha_t}{2}\delta_t \leq 0.$$

The sequence $\{\phi(x_t)\}$ must converge to some limit because $\phi$ is closed and the optimal value is attained. Thus $|\alpha_t \delta_t|$ must decay to zero. The step sizes $\alpha_t$ are bounded away from zero because sufficiently small step sizes satisfy the sufficient descent condition. Thus it is $\delta_t$ that decays to zero. By (4.8), we deduce that $\Delta x_t$ also converges to zero:

$$\|\Delta x_t\|_2^2 \leq \frac{1}{\mu_l}\Delta x_t^T H_t \Delta x_t \leq -\frac{\delta_t}{\mu_l}.$$

It is possible to show $\Delta x_t$ is zero if and only if $x$ is an optimum of (4.1). Thus the sequence $\{x_t\}$ converges to an optimum. □

We turn our attention to the convergence rate of the proximal Newton and quasi-Newton methods. We assume

1. the smooth part $\phi_{\text{sm}}$ is twice-continuously differentiable and its gradient $\nabla \phi_{\text{sm}}$ and Hessian $\nabla^2 \phi_{\text{sm}}$ are Lipschitz continuous with constants $\mu_u$ and $\mu_u'$;

2. $\phi_{\text{sm}}$ is strongly convex with constant $\mu_l > 0$. Since $\phi_{\text{sm}}$ is twice differentiable, strong convexity is equivalent to $\nabla^2 \phi(x) \succeq \mu_l I$ for any $x$.

Both assumptions are standard in the analysis of Newton-type methods for minimizing smooth functions. For our purposes, both assumptions can be relaxed to a local assumption in a neighborhood of $x^*$.

The proximal Newton method incorporates the Hessian $\nabla^2 \phi_{\text{sm}}(x_t)$ in the local quadratic model of $\phi$. It converges $q$-quadratically:

$$\|x_{t+1} - x^*\|_2 = O\big(\|x_t - x^*\|_2^2\big).$$

First, we show that the unit step size satisfies the sufficient descent condition after sufficiently many iterations.

**Lemma 4.5.** *Under the aforementioned conditions, the unit step size satisfies the sufficient decrease condition after sufficiently many iterations.*

**Theorem 4.6.** *Under the aforementioned conditions, the proximal Newton method converges quadratically to $x^*$.*

*Proof.* Since the assumptions of Lemma 4.5 are satisfied, the unit step size satisfies the sufficient descent condition:

$$x_{t+1} = x_t + \Delta x_t = \operatorname{prox}_{\phi_{\text{ns}}}^{\nabla^2 \phi_{\text{sm}}(x_t)} \big( x_t - \nabla^2 \phi_{\text{sm}}(x_t)^{-1} \nabla \phi_{\text{sm}}(x_t) \big).$$

Since scaled proximal mappings are firmly non-expansive in the scaled norm, we have

$$\begin{aligned}
&\|x_{t+1} - x^*\|_{\nabla^2 \phi_{\text{sm}}(x_t)} \\
&\quad = \big\| \operatorname{prox}_{\phi_{\text{ns}}}^{\nabla^2 \phi_{\text{sm}}(x_t)} (x_t - \nabla^2 \phi_{\text{sm}}(x_t)^{-1} \nabla \phi_{\text{sm}}(x_t)) \\
&\qquad - \operatorname{prox}_{\phi_{\text{ns}}}^{\nabla^2 \phi_{\text{sm}}(x_t)} (x^* - \nabla^2 \phi_{\text{sm}}(x_t)^{-1} \nabla \phi_{\text{sm}}(x^*)) \big\|_{\nabla^2 \phi_{\text{sm}}(x_t)} \\
&\quad \le \big\| x_t - x^* + \nabla^2 \phi_{\text{sm}}(x_t)^{-1} (\nabla \phi_{\text{sm}}(x^*) - \nabla \phi_{\text{sm}}(x_t)) \big\|_{\nabla^2 \phi_{\text{sm}}(x_t)} \\
&\quad \le \frac{1}{\sqrt{\mu_l}} \big\| \nabla^2 \phi_{\text{sm}}(x_t)(x_t - x^*) - \nabla \phi_{\text{sm}}(x_t) + \nabla \phi_{\text{sm}}(x^*) \big\|_2 .
\end{aligned}$$

Since $\nabla^2 \phi_{\text{sm}}$ is Lipschitz continuous,

$$\big\| \nabla^2 \phi_{\text{sm}}(x_t)(x_t - x^*) - \nabla \phi_{\text{sm}}(x_t) + \nabla \phi_{\text{sm}}(x^*) \big\|_2 \le \frac{\mu_u'}{2} \|x_t - x^*\|_2^2 .$$

We conclude that $x_t$ converges to $x^*$ quadratically:

$$\|x_{t+1} - x^*\|_2 \le \frac{1}{\sqrt{\mu_l}} \|x_{t+1} - x^*\|_{\nabla^2 \phi_{\text{sm}}(x_t)} \le \frac{\mu_u'}{2\mu_l} \|x_t - x^*\|_2^2 .$$

□

Proximal quasi-Newton methods avoid evaluating $\nabla^2 \phi_{\text{sm}}$ by forming a sequence of Hessian approximations $\{H_t\}$. If the sequence $\{H_t\}$ satisfies the *Dennis-Moré criterion*

$$\frac{\big\| \left(H_t - \nabla^2 \phi_{\text{sm}}(x^*)\right) (x_{t+1} - x_t) \big\|_2}{\|x_{t+1} - x_t\|_2} \to 0, \tag{4.11}$$

it is possible to show that a proximal quasi-Newton method converges superlinearly:

$$\|x_{t+1} - x^*\|_2 \le o(\|x_t - x^*\|_2).$$

Again, we assume that $\phi_{\text{sm}}$ is twice-continuously differentiable and strongly convex with constant $m$, and $\nabla\phi_{\text{sm}}$ and $\nabla^2\phi_{\text{sm}}$ are Lipschitz continuous with constants $\mu_u$ and $\mu_u'$. These are the assumptions required to prove that quasi-Newton methods for minimizing smooth functions converge superlinearly.

First, we state two auxiliary results: (i) the unit step size satisfies the sufficient descent condition after sufficiently many iterations; (ii) the proximal quasi-Newton search direction is close to the proximal Newton search direction.

**Lemma 4.7.** *Under the conditions of Theorem 4.6, if $\{H_t\}$ also has bounded eigenvalues and satisfies the Dennis-Moré criterion (4.11), the unit step satisfies the sufficient descent condition after sufficiently many iterations.*

**Lemma 4.8.** *Let $H_1, H_2$ be positive definite matrices with bounded eigenvalues and $\Delta x_1, \Delta x_2$ be search directions generated by $H_1, H_2$ :*

$$\begin{aligned}\Delta x_1 &= \operatorname{prox}_{\phi_{\text{ns}}}^{H_1}\left(x - H_1^{-1}\nabla\phi_{\text{sm}}(x)\right) - x,\\ \Delta x_2 &= \operatorname{prox}_{\phi_{\text{ns}}}^{H_2}\left(x - H_2^{-1}\nabla\phi_{\text{sm}}(x)\right) - x.\end{aligned}$$

*There is a constant $c_1 > 0$ that depends only on $H_1$ and $H_2$ such that*

$$\|\Delta x_1 - \Delta x_2\|_2 \leq \sqrt{\tfrac{1+c_1}{\mu_{l,1}}}\left\|(H_2 - H_1)\Delta x_1\right\|_2^{1/2}\|\Delta x_1\|_2^{1/2}.$$

*Further, $c_1$ is bounded as long as the eigenvalues of $H_1$ and $H_2$ are bounded.*

**Theorem 4.9.** *Under the conditions of Lemma 4.7, a proximal quasi-Newton method converges q-superlinearly to $x^*$.*

*Proof.* Since the assumptions of Lemma 4.7 are satisfied, the unit step satisfies the sufficient descent condition after sufficiently many iterations:

$$x_{t+1} = x_t + \Delta x_t.$$

Since the proximal Newton method converges $q$-quadratically (cf. Theorem 4.6),

$$\begin{aligned}\|x_{t+1} - x^*\|_2 &\leq \|x_t + \Delta x_{\text{nt},t} - x^*\|_2 + \|\Delta x_t - \Delta x_{\text{nt},t}\|_2\\ &\leq \frac{\mu_u'}{\mu_l}\|x_{\text{nt},t} - x^*\|_2^2 + \|\Delta x_t - \Delta x_{\text{nt},t}\|_2,\end{aligned} \tag{4.12}$$

where $\Delta x_{\text{nt},t}$ is the proximal-Newton search direction and $x^{\text{nt}} = x_t + \Delta x_{\text{nt},t}$. By Lemma 4.8, the second term is bounded by

$$\|\Delta x_t - \Delta x_{\text{nt},t}\|_2 \le \sqrt{\tfrac{1+c_t}{\mu_l}} \left\|(\nabla^2\phi_{\text{sm}}(x_t) - H_t)\Delta x_t\right\|_2^{1/2} \|\Delta x_t\|_2^{1/2}. \qquad (4.13)$$

Since the Hessian $\nabla^2\phi_{\text{sm}}$ is Lipschitz continuous and $\Delta x_t$ satisfies the Dennis-Moré criterion, we have

$$\begin{aligned}
&\left\|\left(\nabla^2\phi_{\text{sm}}(x_t) - H_t\right)\Delta x_t\right\|_2 \\
&\quad \le \left\|\left(\nabla^2\phi_{\text{sm}}(x_t) - \nabla^2\phi_{\text{sm}}(x^*)\right)\Delta x_t\right\|_2 + \left\|\left(\nabla^2\phi_{\text{sm}}(x^*) - H_t\right)\Delta x_t\right\|_2 \\
&\quad \le \mu_u' \|x_t - x^*\|_2 \|\Delta x_t\|_2 + o(\|\Delta x_t\|_2).
\end{aligned}$$

By Tseng and Yun (2009), Lemma 3, we know that $\|\Delta x_t\|_2$ is within a constant $\bar{\theta}_t$ of $\|\Delta x_{\text{nt},t}\|_2$. We also know that the proximal Newton method converges $q$-quadratically. Thus

$$\begin{aligned}
\|\Delta x_t\|_2 &\le c_t \|\Delta x_{\text{nt},t}\|_2 = c_t \|x_{\text{nt},t+1} - x_t\|_2 \\
&\le c_t \left(\|x_{\text{nt},t+1} - x^*\|_2 + \|x_t - x^*\|_2\right) \\
&\le O\left(\|x_t - x^*\|_2^2\right) + c_t \|x_t - x^*\|_2 .
\end{aligned}$$

Substituting in the bound on $\|\Delta x_t\|$ into (4.13), we obtain

$$\|\Delta x_t - \Delta x_{\text{nt},t}\|_2 = o(\|x_t - x^*\|_2).$$

We substitute this expression into (4.12) to deduce that $x_t$ converges to $x^*$ superlinearly:

$$\|x_{t+1} - x^*\| \le \frac{\mu_u'}{\mu_l} \|x_{\text{nt},t} - x^*\|_2^2 + o(\|x_t - x^*\|_2).$$

□

There has been a flurry of recent activity around the development of Newton-type methods for minimizing composite functions: Hsieh et al. (2011), Becker and Fadili (2012), Olsen et al. (2012). We have shown that proximal Newton-type methods converge rapidly near the optimal solution, and can produce a solution of high accuracy. The main drawback of proximal Newton-type methods is the cost of solving the subproblems. As we shall see, it is possible to reduce the cost by solving the subproblems inexactly and retain the fast convergence rate.

## 4.4 Computational results

We compare the performance of the proximal L-BFGS method with that of SpaRSA and FISTA on sparse logistic regression:

$$\underset{w \in \mathbf{R}^n}{\text{minimize}} \ \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y^{(i)} w^T x^{(i)})) + \lambda \|w\|_1 ,$$

where $\{(x^{(i)}, y^{(i)})\}_{i \in [m]}$ are feature-label pairs. The $\ell_1$ regularizer encourages sparse solutions, and the parameter $\lambda$ trades off goodness-of-fit and sparsity.

We train on two datasets: (i) `gisette`, a handwritten digits dataset from the NIPS 2003 feature selection challenge, and (ii) `rcv1`, an archive of categorized news stories from Reuters.[1] The parameter $\lambda$ was chosen to match the values reported by Yuan et al. (2012), where it was chosen by five-fold cross validation. Figures 1 and 2 show relative suboptimality versus number of function evaluations and wall time on the `gisette` and `rcv1` datasets.

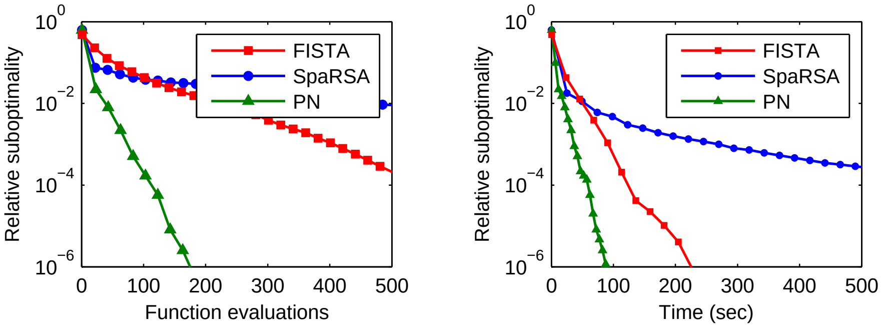


Figure 1: Sparse logistic regression on `gisette` dataset

On the dense `gisette` dataset, evaluating $\phi_{\text{sm}}$ dominates the training cost. The proximal L-BFGS method outperforms FISTA and SpaRSA because it evaluates $\phi_{\text{sm}}$ less often. On the sparse `rcv1` dataset (40 million nonzero entries in a $542000 \times 47000$ design matrix), evaluating $\phi_{\text{sm}}$ only makes up a small portion of the training cost, and the proximal L-BFGS method barely outperforms SpaRSA.

[1] These datasets are available at `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets`.

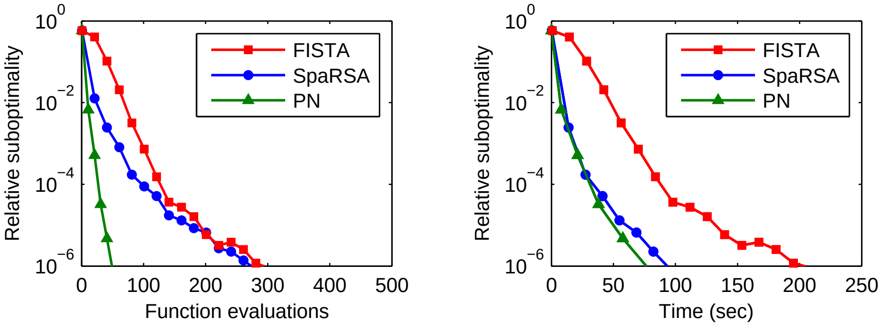


Figure 2: Sparse logistic regression on `rcv1` dataset

# 5 INEXACT PROXIMAL NEWTON-TYPE METHODS

Proximal Newton-type methods are like most second-order methods in terms of their computational cost: they take only a few iterations to converge, but each iteration is costly. The main cost per iteration is the cost of solving subproblem (4.5) for a search direction.

Inexact proximal Newton-type methods reduce the cost per iteration by solving the subproblem inexactly. These methods can be more efficient than their exact counterparts because they require less computation per iteration. Indeed, many practical implementations of proximal Newton-type methods such as `glmnet`, `newGLMNET`, and QUIC solve the subproblems inexactly.

In practice, how exactly the subproblem is solved is critical to the efficiency and reliability of the method. Most practical implementations use a variety of heuristics to decide how accurately to solve the subproblem. Although these methods perform admirably in practice, there are few results on how the inexact subproblem solutions affect their convergence behavior. In this chapter, we propose a criterion for deciding how exactly to solve the subproblem, and study the convergence rate of an inexact proximal Newton method that implements the proposed criterion.

## 5.1 AN ADAPTIVE STOPPING CONDITION

To begin, we observe that the subproblem (4.5) is itself a composite function minimization problem:

$$\arg\min_d \hat{\phi}_t(x_t + d) := \hat{\phi}_{\text{sm},t}(x_t + d) + \phi_{\text{ns}}(x_t + d).$$

Thus the size of the composite gradient step (on the subproblem)

$$\hat{g}_{t,\alpha}(x) := \frac{1}{\alpha}\left(x - \operatorname{prox}_{\phi_{\text{ns}}}(x - \alpha \nabla \hat{\phi}_{\text{sm},t}(x))\right)$$

is a measure of the exactness of the search direction. We propose a stopping condition that mimics the one used by *inexact Newton-type methods* for minimizing smooth functions. Let $\mu_u \geq \mu_l > 0$ be bounds on the eigenvalues of $H_t$. We stop the subproblem solver when the subproblem iterate $x_t + \Delta x_t$ satisfies

$$\|\hat{g}_{t,1/\mu_u}(x_t + \Delta x_t)\|_2 \le \eta_t \|g_{1/\mu_u}(x_t)\|_2, \tag{5.1}$$

where $\eta_t$ is a *forcing term* that requires the left-hand side to be small. We set $\eta_t$ based on how well $\hat{g}_{t-1}$ approximates $g$ near $x_t$: we require

$$\eta_t = \min\left\{\frac{\mu_l}{2}, \frac{\|\hat{g}_{t-1,1/\mu_u}(x_t) - g_{1/\mu_u}(x_t)\|_2}{\|g_{1/\mu_u}(x_{t-1})\|_2}\right\}. \tag{5.2}$$

This choice due to Eisenstat and Walker (1996) yields desirable convergence results and performs admirably in practice.

Intuitively, we should solve the subproblem exactly if

1. $x_t$ is close to the optimum,
2. $\hat{\phi}_t$ is a good model of $\phi$ near $x_t$.

If the former, we seek to preserve the fast local convergence behavior of proximal Newton-type methods; if the latter, then minimizing $\hat{\phi}_t$ is a good surrogate for minimizing $\phi$. In these cases, (5.1) and (5.2) are small, so the subproblem is solved accurately.

We can derive an expression like (4.6) for an inexact search direction in terms of an explicit gradient, an implicit subgradient, and a residual term $r_t$. The adaptive stopping condition (5.1) is equivalent to

$$\begin{aligned} 0 &\in \hat{g}_{t,1/\mu_u}(x_t + \Delta x_t) + r_t \\ &= \nabla \phi_{\mathrm{sm},t}(x_t + \Delta x_t) + \partial \phi_{\mathrm{ns}}(x_t + \Delta x_t + \hat{g}_{t,1/\mu_u}(x_t + \Delta x_t)) + r_t \\ &= \nabla \phi_{\mathrm{sm}}(x_t) + H_t \Delta x_t + \partial \phi_{\mathrm{ns}}(x_t + \Delta x_t + \hat{g}_{t,1/\mu_u}(x_t + \Delta x_t)) + r_t \end{aligned}$$

for some $r_t : \|r_t\|_2 \le \eta_t \|g(x_t)\|_2$. Thus

$$H_t \Delta x_t \in -\nabla \phi_{\mathrm{sm}}(x_t) - \partial \phi_{\mathrm{ns}}(x_t + \Delta x_t + \hat{g}_{t,1/\mu_u}(x_t + \Delta x_t)) + r_t. \tag{5.3}$$

## 5.2 CONVERGENCE OF AN INEXACT PROXIMAL NEWTON METHOD

As we shall see, the inexact proximal Newton method with unit step converges locally

- at a linear rate when the forcing terms $\eta_t$ are uniformly smaller than the inverse of the Lipschitz constant of $g$;

- at a superlinear rate when the forcing terms $\eta_t$ are chosen according to (5.2).

Before delving into the convergence analysis, we review some recent results by Byrd et al. (2013). They analyze the inexact proximal Newton method with a more stringent stopping condition:

$$\|\hat{g}_{t,1/\mu_u}(x_t+\Delta x_t)\|_2 \le \eta_t \|g_{1/\mu_u}(x_t)\|_2 \text{ and } \hat{\phi}_t(x_t+\Delta x_t) - \hat{\phi}_t(x_t) \le \frac{\delta_t}{2}. \quad (5.4)$$

The latter is a sufficient descent condition (on the subproblem). When the nonsmooth is the $\ell_1$ norm, they show that the inexact proximal Newton method with stopping condition (5.4)

1. converges globally,
2. eventually accepts the unit step size,
3. converges linearly or superlinearly depending on the choice of forcing terms.

Although their first two results generalize readily to composite functions with other nonsmooth part, their third result depends on the separability of the $\ell_1$ norm. We generalize their third result to composite functions with a other nonsmooth parts. In other words, when combined with their first two results, our result implies the inexact proximal Newton method with the more stringent stopping condition converges globally, and converges linearly or superlinearly (depending on the choice of forcing terms)

As before, we assume

1. the smooth part $\phi_{\text{sm}}$ is twice-continuously differentiable and strongly convex with constant $\mu_l$,
2. the gradient $\nabla\phi_{\text{sm}}$ and Hessian $\nabla^2\phi_{\text{sm}}$ are Lipschitz continuous with constants $\mu_u$ and $\mu_u'$.

We also assume $x_0$ is sufficiently close to $x^*$ so that the unit step always satisfies the sufficient descent condition. These are the same assumptions made by Dembo et al. (1982) and Eisenstat and Walker (1996) in their analysis of *inexact Newton methods* for minimizing smooth functions.

First, we show that (i) $\hat{g}_t$ is a good approximation of $g$, (ii) $\hat{g}_t$ inherits the Lipschitz continuity and strong monotonicity of $\nabla\phi_{\text{sm},t}$.

**Lemma 5.1.** *Under the aforementioned assumptions,* $\|g(x) - \hat{g}_t(x)\|_2 \le \frac{\mu_u'}{2}\|x - x_t\|_2^2$.

**Lemma 5.2.** *If $\nabla\phi_{\text{sm}}$ is Lipschitz continuous with constant $\mu_u$, $g$ is Lipschitz continuous with constant $\mu_u + 1$ :*

$$\|g(x) - g(x^*)\|_2 \leq (\mu_u + 1)\, \|x - x^\star\|_2 .$$

The next result generalizes Byrd et al. (2013), Lemma 4.1 to composite functions with a generic nonsmooth part. To our knowledge, it is a novel result concerning the composite gradient step.

**Lemma 5.3.** *If $\nabla\phi_{\text{sm}}$ is Lipschitz continuous with constant $\mu_u$ and strongly monotone with constant $\mu_l$, $g_\alpha$ is strongly monotone with constant $\frac{\mu_l}{2}$ for any $\alpha \leq \frac{1}{\mu_u}$ :*

$$(x-y)^T(g_\alpha(x) - g_\alpha(y)) \geq \frac{\mu_l}{2}\, \|x-y\|_2^2 .$$

We use these two results to show that the inexact proximal Newton method with unit step sizes converges locally at a linear or superlinear rate depending on the choice of forcing terms.

**Theorem 5.4.** *Under the aforementioned conditions,*

1. *when $\eta_t$ is smaller than $\bar{\eta} < \frac{\mu_l}{2}$, the inexact proximal Newton method with unit steps converges q-linearly to $x^*$;*
2. *when $\eta_t$ decays to zero, the inexact proximal Newton method with unit steps converges q-superlinearly to $x^*$.*

*Proof.* The local quadratic model $\hat{\phi}_t$ is strongly convex with constant $\mu_l$. By Lemma 5.3, $\hat{g}_{t,1/\mu_u}$ is strongly monotone with constant $\frac{\mu_l}{2}$:

$$(x-y)^T\left(\hat{g}_{t,1/\mu_u}(x) - \hat{g}_{t,1/\mu_u}(y)\right) \geq \frac{\mu_l}{2}\, \|x-y\|_2^2 .$$

By the Cauchy-Schwarz inequality, we have

$$\|\hat{g}_{t,1/\mu_u}(x) - \hat{g}_{t,1/\mu_u}(y)\|_2 \geq \frac{\mu_l}{2}\, \|x-y\|_2 .$$

We apply this result to $x_t + \Delta x_t$ and $x^*$ to obtain

$$\|x_{k+1} - x^*\|_2 = \|x_t + \Delta x_t - x^*\|_2 \leq \frac{2}{\mu_l} \|\hat{g}_{t,1/\mu_u}(x_t + \Delta x_t) - \hat{g}_{t,1/\mu_u}(x^*)\|_2. \tag{5.5}$$

Let $r_t$ be the residual $-g_{t,1/\mu_u}(x_t + \Delta x_t)$. The adaptive stopping condition (5.1) requires $\|r_t\|_2 \leq \eta_t \|g_{1/\mu_u}(x_t)\|_2$. We substitute this expression into (5.5) to obtain

$$\begin{aligned}
\|x_{k+1} - x^*\|_2 &\leq \frac{2}{\mu_l}\| - \hat{g}_{t,1/\mu_u}(x^*) - r_t\|_2 \\
&\leq \frac{2}{\mu_l}\Big(\|\hat{g}_{t,1/\mu_u}(x^*)\|_2 + \|r_t\|_2\Big) \\
&\leq \frac{2}{\mu_l}\Big(\|\hat{g}_{t,1/\mu_u}(x^*)\|_2 + \eta_t\|g_{1/\mu_u}(x_t)\|_2\Big). \qquad (5.6)
\end{aligned}$$

Applying Lemma 5.1 to $\frac{1}{\mu_u}\phi$ and $\frac{1}{\mu_u}\hat{\phi}$ gives

$$\|\hat{g}_{t,1/\mu_u}(x^*)\|_2 \leq \frac{\mu_u'}{2\mu_u}\|x_t - x^*\|_2^2 + \|g_{1/\mu_u}(x^*)\|_2 = \frac{\mu_u'}{2\mu_u}\|x_t - x^*\|_2^2.$$

We substitute this bound into (5.6) to obtain

$$\begin{aligned}
\|x_{k+1} - x^*\|_2 &\leq \frac{2}{\mu_l}\Big(\frac{\mu_u'}{2\mu_u}\|x_t - x^*\|_2^2 + \eta_t\|g_{1/\mu_u}(x_t)\|_2\Big) \\
&\leq \frac{\mu_u'}{\mu_u\mu_l}\|x_t - x^*\|_2^2 + \frac{2\eta_t}{\mu_l}\|x_t - x^*\|_2.
\end{aligned}$$

We deduce that (i) $x_t$ converges $q$-linearly to $x^*$ when $\eta_t \leq \bar{\eta}$ for some $\bar{\eta} < \frac{\mu_l}{2}$, and (ii) $x_t$ converges $q$-superlinearly to $x^*$ when $\eta_t$ decays to zero. □

Finally, we justify our choice of forcing terms. If we set $\eta_t$ according to (5.2), then the inexact proximal Newton method converges $q$-superlinearly. When minimizing smooth functions, we recover the result by Eisenstat and Walker (1996) on choosing forcing terms in an inexact Newton method.

**Theorem 5.5.** *Under the conditions of Theorem 5.4, if we set $\eta_t$ according to (5.2), then the inexact proximal Newton method with unit steps converges $q$-superlinearly.*

*Proof.* To show superlinear convergence, we must show

$$\frac{\|\hat{g}_{t-1,1/\mu_u}(x_t) - g_{1/\mu_u}(x_t)\|_2}{\|g_{1/\mu_u}(x_{t-1})\|_2} \to 0. \qquad (5.7)$$

By Lemma 5.2, we have

$$\begin{aligned}\|\hat{g}_{t-1,1/\mu_u}(x_t) - g_{1/\mu_u}(x_t)\|_2 &\le \frac{\mu_u'}{2\mu_u}\left\|x_t - x_{t-1}\right\|_2^2 \\ &\le \frac{\mu_u'}{2\mu_u}\left(\|x_t - x^*\|_2 + \|x^* - x_{t-1}\|_2\right)^2.\end{aligned}$$

By Lemma 5.3, we also have

$$\|g_{1/\mu_u}(x_{t-1})\|_2 = \|g_{1/\mu_u}(x_{t-1}) - g_{1/\mu_u}(x^*)\|_2 \ge \frac{\mu_l}{2}\left\|x_{t-1} - x^*\right\|_2.$$

We substitute these expressions into (5.7) to obtain

$$\begin{aligned}&\frac{\|\hat{g}_{t-1,1/\mu_u}(x_t) - g_{1/\mu_u}(x_t)\|_2}{\|g_{1/\mu_u}(x_{t-1})\|_2} \\ &\quad\le \frac{\frac{\mu_u'}{2\mu_u}\left(\|x_t - x^*\|_2 + \|x^* - x_{t-1}\|_2\right)^2}{\frac{\mu_l}{2}\left\|x_{t-1} - x^*\right\|_2} \\ &\quad= \frac{\mu_u'}{\mu_u\mu_l}\frac{\|x_t - x^*\|_2 + \|x_{t-1} - x^*\|_2}{\|x_{t-1} - x^*\|_2}\left(\|x_t - x^*\|_2 + \|x_{t-1} - x^*\|_2\right) \\ &\quad= \frac{\mu_u'}{\mu_u\mu_l}\left(1 + \frac{\|x_t - x^*\|_2}{\|x_{t-1} - x^\star\|_2}\right)\left(\|x_t - x^*\|_2 + \|x_{t-1} - x^*\|_2\right).\end{aligned}$$

By Theorem 5.4, we have $\frac{\|x_t - x^*\|_2}{\|x_{t-1} - x^\star\|_2} < 1$ and

$$\frac{\|\hat{g}_{t-1,1/\mu_u}(x_t) - g_{1/\mu_u}(x_t)\|_2}{\|g_{1/\mu_u}(x_{t-1})\|_2} \le \frac{2\mu_u'}{\mu_u\mu_l}\left(\|x_t - x^*\|_2 + \|x_{t-1} - x^*\|_2\right).$$

Thus the forcing terms decay to zero. By Theorem 5.4, the inexact proximal Newton method with adaptive stopping condition (5.1) converges $q$-superlinearly. □

## 5.3 COMPUTATIONAL EXPERIMENTS

We study how inexact search directions affect the convergence of proximal Newton-type methods on a sparse inverse covariance estimation problem

$$\underset{\Theta \in \mathbf{R}^{n\times n}}{\text{minimize}}\ \operatorname{tr}\left(\hat{\Sigma}\Theta\right) - \log\det(\Theta) + \lambda\left\|\Theta\right\|_1, \tag{5.8}$$

where $\hat{\Sigma}$ is the sample covariance matrix. We regularize using the (entrywise) $\ell_1$ norm to encourage sparse solutions.

We fit a sparse inverse covariance matrix to two datasets: (i) Estrogen, a gene expression dataset consisting of 682 probe sets collected from 158 patients, and (ii) Leukemia, another gene expression dataset consisting of 1255 genes from 72 patients.[1] The regularization parameter $\lambda$ was chosen to match the value reported by Hsieh et al. (2011).

We solve the inverse covariance estimation problem (5.8) using a proximal BFGS method, i. e., $H_t$ is updated according to the BFGS update. (The proximal Newton method would be computationally very expensive on such large datasets.) To study the effects of inexact search directions, we compare the convergence of the proximal BFGS method with two rules to decide how accurately to solve the subproblem (4.7):

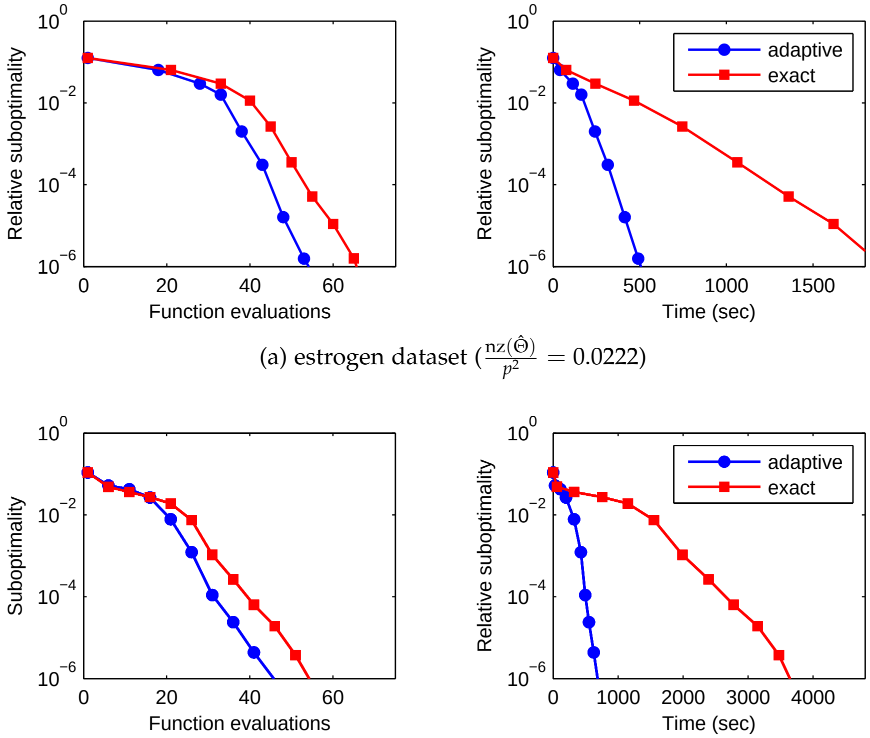
1. adaptive: stop when the subproblem iterate satisfies the adaptive stopping condition (5.1);

2. exact: stop when the norm of the composite gradient step (on the subproblem) is smaller than $10^{-6}$.

Figure 3 shows relative suboptimality versus number of function evaluations and wall time on the two datasets. We observe that the proximal BFGS method with both rules converges at roughly the same rate in terms of function evaluations. However, the exact rule spends more time per iteration solving the subproblem exactly. If we account for the time per iteration, the adaptive stopping rule converges significantly faster than the exact stopping rule.

Finally, we observe that although the conditions for superlinear convergence of proximal quasi-Newton methods are not met ($\log\det$ is strongly convex), we observe in Figure 3 that the proximal BFGS method with both stopping rules transitions from linear to superlinear convergence. The transition is characteristic of BFGS and other quasi-Newton methods on smooth problems.

[1] These datasets are available from `http://www.math.nus.edu.sg/~mattohkc/` with the SPINCOVSE package.



(a) estrogen dataset ($\frac{\mathrm{nz}(\hat{\Theta})}{p^2} = 0.0222$)

(b) leukemia dataset ($\frac{\mathrm{nz}(\hat{\Theta})}{p^2} = 0.0221$)

Figure 3: Convergence of the proximal BFGS method on the sparse inverse covariance estimation problem with three subproblem stopping conditions.

# 6

# COMMUNICATION EFFICIENT DISTRIBUTED SPARSE REGRESSION

Explosive growth in the size of modern datasets has fueled interest in distributed statistical learning. For examples, we refer to Boyd et al. (2011), Dekel et al. (2012), Duchi et al. (2012), Zhang et al. (2013) and the references therein. The problem arises, for example, when working with datasets that are too large to fit on a single machine and must be distributed across multiple machines. The main bottleneck in the distributed setting is usually communication between machines/processors, so the overarching goal of algorithm design is to minimize communication costs.

In distributed statistical learning, the simplest and most popular approach is *averaging*: each machine forms a local estimator $\hat{\theta}_k$ with the portion of the data stored locally, and a "master" averages the local estimators to produce an aggregate estimator: $\bar{\theta} = \frac{1}{m}\sum_{k=1}^{m}\hat{\theta}_k$. Averaging was first studied by Mcdonald et al. (2009) for multinomial regression. They derive non-asymptotic error bounds on the estimation error that show averaging reduces the variance of the local estimators, but has no effect on the bias (from the centralized solution). In follow-up work, Zinkevich et al. (2010) studied a variant of averaging where each machine computes a local estimator with stochastic gradient descent (SGD) on a random subset of the dataset. They show, among other things, that their estimator converges to the centralized estimator.

More recently, Zhang et al. (2013) studied averaged empirical risk minimization (ERM). They show that the mean squared error (MSE) of the averaged ERM decays like $O\big(N^{-\frac{1}{2}} + \frac{m}{N}\big)$, where $m$ is the number of machines and $N$ is the total number of samples. Thus, so long as $m \lesssim \sqrt{N}$, the averaged ERM matches the $N^{-\frac{1}{2}}$ convergence rate of the centralized ERM. Even more recently, Rosenblatt and Nadler (2014) studied the optimality of averaged ERM in two asymptotic settings: $N \to \infty$, $m, p$ fixed and $p, n \to \infty$, $\frac{p}{n} \to \mu_l \in (0,1)$, where $n = \frac{N}{m}$ is the number of samples per machine. They show that in the $n \to \infty$, $p$ fixed setting, the averaged ERM is first-order equivalent to the centralized ERM. However, when $p, n \to \infty$, the averaged ERM is suboptimal (versus the centralized ERM).

We develop an approach to distributed statistical learning in the high-dimensional setting. Since $p \gtrsim n$, regularization is essential. At a high level, the key idea is to average local *debiased* regularized M-estimators. We show that our averaged estimator converges at the same rate as the centralized regularized M-estimator. The material in this chapter appears in Lee et al. (2015a).

## 6.1 BACKGROUND ON THE LASSO AND THE DEBIASED LASSO

To keep things simple, we focus on sparse linear regression. Consider the sparse linear model

$$y = X\beta^* + \epsilon,$$

where the rows of $X \in \mathbf{R}^{n\times p}$ are predictors, and the components of $y \in \mathbf{R}^n$ are the responses. To keep things simple, we assume

(A1) the predictors $x \in \mathbf{R}^p$ are independent subgaussian random vectors whose covariance $\Sigma$ has smallest has smallest eigenvalue $\lambda_{\min}(\Sigma)$;

(A2) the regression coefficients $\beta^* \in \mathbf{R}^p$ are $s$-sparse, i. e. all but $s$ components of $\beta^*$ are zero;

(A3) the components of $\epsilon \in \mathbf{R}^n$ are independent, mean zero subgaussian random variables.

Given the predictors and responses, the lasso estimates $\beta^*$ by

$$\hat{\beta} = \arg\min_{\beta\in\mathbf{R}^p} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

There is a well-developed theory of the lasso that says, under suitable assumptions on $X$, the lasso estimator $\hat{\beta}$ is nearly as close to $\beta^*$ as the *oracle estimator*: $X^\dagger_{\mathrm{nz}(\beta^*)}y$ (e. g. see Hastie et al. (2015), Chapter 11 for an overview). More precisely, under some conditions on $\frac{1}{n}X^TX$, the MSE of the lasso estimator is roughly $\frac{s\log p}{n}$. Since the MSE of the oracle estimator is (roughly) $\frac{s}{n}$, the lasso estimator is almost as good as the oracle estimator.

However, the lasso estimator is also biased[1]. Since averaging only reduces variance, not bias, we gain (almost) nothing by averaging the biased lasso estimators. That is, it is possible to show if we naively averaged local lasso estimators, the MSE of the averaged estimator is of the same order as that of the local estimators. The key to overcoming the bias of the averaged lasso estimator is to "debias" the lasso estimators before averaging.

[1] We refer to Section 2.2 in Javanmard and Montanari (2013a) for a more formal discussion of the bias of the lasso estimator.

The *debiased lasso estimator* by Javanmard and Montanari (2013a) is

$$\hat{\beta}^d = \hat{\beta} + \frac{1}{n}\hat{\Theta}X^T(y - X\hat{\beta}), \tag{6.1}$$

where $\hat{\beta}$ is the lasso estimator and $\hat{\Theta} \in \mathbf{R}^{p\times p}$ is an approximate inverse to $\hat{\Sigma} = \frac{1}{n}X^TX$. Intuitively, the debiased lasso estimator trades bias for variance. The trade-off is obvious when $\hat{\Sigma}$ is non-singular: setting $\hat{\Theta} = \hat{\Sigma}^{-1}$ gives the ordinary least squares (OLS) estimator $(X^TX)^{-1}X^Ty$.

Another way to interpret the debiased lasso estimator is a corrected estimator that compensates for the bias incurred by shrinkage. By the optimality conditions of the lasso, the correction term $\frac{1}{n}X^T(y - X\hat{\beta})$ is a subgradient of $\lambda \left\|\cdot\right\|_1$ at $\hat{\beta}$. By adding a term proportional to the subgradient of the regularizer, the debiased lasso estimator compensates for the bias incurred by regularization. The debiased lasso estimator has previously been used to perform inference on the regression coefficients in high-dimensional regression models. We refer to the papers by Javanmard and Montanari (2013a), van de Geer et al. (2013), Zhang and Zhang (2014), Belloni et al. (2011) for details.

The choice of $\hat{\Theta}$ in the correction term is crucial to the performance of the debiased estimator. Javanmard and Montanari (2013a) suggest forming $\hat{\Theta}$ row by row: the $j$-th row of $\hat{\Theta}$ is the optimum of

$$\begin{aligned} \underset{\theta\in\mathbf{R}^p}{\text{minimize}} \quad & \theta^T\hat{\Sigma}\theta \\ \text{subject to} \quad & \|\hat{\Sigma}\theta - e_j\|_\infty \le \delta. \end{aligned} \tag{6.2}$$

The parameter $\delta$ should large enough to keep the problem feasible, but as small as possible to keep the bias (of the debiased lasso estimator) small. As we shall see, when the rows of $X$ are subgaussian, setting $\delta \sim \big(\frac{\log p}{n}\big)^{\frac{1}{2}}$ is usually large enough to keep (6.2) feasible.

**Definition 6.1** (Generalized coherence). *Given $X \in \mathbf{R}^{n\times p}$, let $\hat{\Sigma} = \frac{1}{n}X^TX$. The* generalized coherence *between $\hat{\Sigma}$ and $\Theta \in \mathbf{R}^{p\times p}$ is*

$$\mathrm{GC}(\hat{\Sigma},\Theta) = \max\nolimits_{j\in[p]} \|\hat{\Sigma}\Theta_j^T - e_j\|_\infty.$$

**Lemma 6.2** (Javanmard and Montanari (2013a)). *Under (A1), when $16\kappa\sigma_x^4 n >$ $\log p$, the event*

$$\mathcal{E}_{\mathrm{GC}}(\hat{\Sigma}) := \left\{\mathrm{GC}(\hat{\Sigma},\Sigma^{-1}) \le \frac{8}{\sqrt{c_1}}\sqrt{\kappa}\sigma_x^2\Big(\frac{\log p}{n}\Big)^{\frac{1}{2}}\right\}$$

*occurs with probability at least* $1 - 2p^{-2}$ *for some* $c_1 > 0$*, where* $\kappa := \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ *is the condition number of* $\Sigma$*.*

As we shall see, the bias of the debiased lasso estimate is of higher order than its variance under suitable conditions on $\hat{\Sigma}$. In particular, we require $\hat{\Sigma}$ to satisfy the *restricted eigenvalue (RE) condition*.

**Definition 6.3** (RE condition)**.** *For any* $\mathcal{S} \subset [p]$*, let*

$$\mathcal{C}(\mathcal{S}) := \{\Delta \in \mathbf{R}^p \mid \|\Delta_{\mathcal{S}^c}\|_1 \leq 3\,\|\Delta x_{\mathcal{S}}\|_1\}.$$

*We say* $\hat{\Sigma}$ *satisfies the RE condition on the cone* $\mathcal{C}(\mathcal{S})$ *when*

$$\Delta^T \hat{\Sigma} \Delta \geq \mu_l \|\Delta_{\mathcal{S}}\|_2^2$$

*for some* $\mu_l > 0$ *and any* $\Delta \in \mathcal{C}(\mathcal{S})$*.*

The RE condition requires $\hat{\Sigma}$ to be positive definite on $\mathcal{C}(\mathcal{S})$. When the rows of $X \in \mathbf{R}^{n \times p}$ are *i.i.d.* Gaussian random vectors, Raskutti et al. (2010) show there are constants $\mu_1, \mu_2 > 0$ such that

$$\frac{1}{n}\,\|X\Delta\|_2^2 \geq \mu_1\,\|\Delta\|_2^2 - \mu_2 \frac{\log p}{n}\,\|\Delta\|_1^2 \text{ for any } \Delta \in \mathbf{R}^p$$

with probability at least $1 - c_2 \exp(-c_2 n)$. Their result implies the RE condition holds on $\mathcal{C}(\mathcal{S})$ (for any $\mathcal{S} \subset [p]$) as long as $n \gtrsim |S| \log p$, even when there are dependencies among the predictors. Their result was extended to subgaussian designs by Rudelson and Zhou (2013), also allowing for dependencies among the covariates. We summarize their result in a lemma.

**Lemma 6.4.** *Under (A1), when* $n > 4000 \tilde{s} \sigma_x^2 \log(\frac{60\sqrt{2}ep}{\tilde{s}})$ *and* $p > \tilde{s}$*, where* $\tilde{s} := s + 25920\kappa s$*, the event*

$$\mathcal{E}_{\text{RE}}(X) = \left\{\Delta^T \hat{\Sigma} \Delta \geq \frac{1}{2} \lambda_{\min}(\Sigma) \|\Delta_S\|_2^2 \textit{ for any } \Delta \in \mathcal{C}(S)\right\}$$

*occurs with probability at least* $1 - 2e^{-\frac{n}{4000\sigma_x^4}}$.

*Proof.* The lemma is a consequence of Rudelson and Zhou (2013), Theorem 6. In their notation, we set $\delta = \frac{1}{\sqrt{2}}$, $k_0 = 3$ and bound $\max_{j \in [p]} \left\|A e_j\right\|_2^2$ and $K(s_0, k_0, \Sigma^{\frac{1}{2}})$ by $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)^{-\frac{1}{2}}$. □

When the RE condition holds, the lasso and debiased lasso estimators are consistent for a suitable choice of the regularization parameter $\lambda$. The parameter $\lambda$ should be large enough to dominate the "empirical process" part of the problem: $\frac{1}{n}\left\|X^T y\right\|_\infty$, but as small as possible to reduce the bias incurred by regularization. As we shall see, setting $\lambda \sim \sigma_y\left(\frac{\log p}{n}\right)^{\frac{1}{2}}$ is a good choice.

**Lemma 6.5.** *Under (A3),*

$$\frac{1}{n}\|X^T\epsilon\|_\infty \le \max_{j\in[p]}(\hat{\Sigma}_{j,j})^{\frac{1}{2}}\sigma_y\left(\frac{3\log p}{c_2 n}\right)^{\frac{1}{2}}$$

*with probability at least* $1 - ep^{-2}$ *for any (non-random)* $X \in \mathbf{R}^{n\times p}$*.*

When $\hat{\Sigma}$ satisfies the RE condition and $\lambda$ is large enough, the lasso and debiased lasso estimators are consistent.

**Lemma 6.6** (Negahban et al. (2012))**.** *Under (A2) and (A3), suppose* $\hat{\Sigma}$ *satisfies the RE condition on* $\mathcal{C}^*$ *with constant* $\mu_l$ *and* $\frac{1}{n}\|X^T\epsilon\|_\infty \le \lambda$*,*

$$\|\hat{\beta} - \beta\|_1 \le \frac{3}{\mu_l}s\lambda \text{ and } \|\hat{\beta} - \beta\|_2 \le \frac{3}{\mu_l}\sqrt{s}\lambda.$$

When the lasso estimator is consistent, the debiased lasso estimator is also consistent. Further, it is possible to show that the bias of the debiased estimator is of higher order than its variance. Similar results by Javanmard and Montanari (2013a), van de Geer et al. (2013), Zhang and Zhang (2014), Belloni et al. (2011) are the key step in showing the asymptotic normality of the (components of) the debiased lasso estimator. The result we state is essentially Javanmard and Montanari (2013a), Theorem 2.3.

**Lemma 6.7.** *Under the conditions of Lemma 6.6, when* $(\hat{\Sigma}, \hat{\Theta})$ *has generalized incoherence* $\delta$*, the debiased lasso estimator has the form*

$$\hat{\beta}^d = \beta^* + \frac{1}{n}\hat{\Theta}X^T\epsilon + \hat{\Delta},$$

*where* $\|\hat{\Delta}\|_\infty \le \frac{3\delta}{\mu_l}s\lambda$*.*

Lemma 6.7, together with Lemmas 6.5 and 6.2, shows that the bias of the debiased lasso estimator is of higher order than its variance. In particular, setting $\lambda$ and $\delta$ according to Lemmas 6.5 and 6.2 gives a bias term $\|\hat{\Delta}\|_\infty$ that is $O\left(\frac{s\log p}{n}\right)$. By comparison, the variance term $\frac{1}{n}\|\hat{\Theta}X^T\epsilon\|_\infty$ is the maximum of $p$ subgaussian random variables with mean zero and variances of $O(1)$, which is $O\left(\left(\frac{\log p}{n}\right)^{\frac{1}{2}}\right)$. Thus the bias term is of higher order than the variance term as long as $n \gtrsim s^2 \log p$.

**Corollary 6.8.** *Under (A2), (A3), and the conditions of Lemma 6.6, when* $(\hat{\Sigma}, \hat{\Theta})$ *has generalized incoherence* $\delta'\big(\frac{\log p}{n}\big)^{\frac{1}{2}}$ *and we set* $\lambda = \max_{j\in[p]}(\hat{\Sigma}_{j,j})^{\frac{1}{2}}\sigma_y\big(\frac{3\log p}{c_2 n}\big)^{\frac{1}{2}}$,

$$\|\hat{\Delta}\|_\infty \le \frac{3\sqrt{3}}{\sqrt{c_2}}\frac{\delta' \max_{j\in[p]}(\hat{\Sigma}_{j,j})^{\frac{1}{2}}}{\mu_l}\sigma_y\frac{s\log p}{n}.$$

## 6.2 AVERAGING DEBIASED LASSOS

Recall the problem setup: we are given $N$ samples of the form $z_i = (x_i, y_i)$ distributed across $m$ machines:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

The $k$-th machine has local predictors $X_k \in \mathbf{R}^{n_k \times p}$ and responses $y_k \in \mathbf{R}^{n_k}$. To keep things simple, we assume the data is evenly distributed, i. e. $n_1 = \cdots = n_k = n = \frac{N}{m}$. The *averaged debiased lasso* estimator (for lack of a better name) is

$$\bar{\beta} = \frac{1}{m}\sum_{k=1}^m \hat{\beta}_k^d = \frac{1}{m}\sum_{k=1}^m \hat{\beta}_k + \hat{\Theta}_k X_k^T(y_k - X_k\hat{\beta}_k), \tag{6.3}$$

We study the error of the averaged debiased lasso in the $\ell_\infty$ norm.

**Lemma 6.9.** *Suppose the local sparse regression problem on each machine satisfies the conditions of Corollary 6.8, that is when* $m \le p$,

1. $\{\hat{\Sigma}_k\}_{k\in[m]}$ *satisfy the RE condition on* $\mathcal{C}^*$ *with constant* $\mu_l$,
2. $\{(\hat{\Sigma}_k, \hat{\Theta}_k)\}_{k\in[m]}$ *have generalized incoherence* $c_{\mathrm{GC}}\big(\frac{\log p}{n}\big)^{\frac{1}{2}}$,
3. *we set* $\lambda_1 = \cdots = \lambda_m = c_\Sigma \sigma_y\big(\frac{3\log p}{c_2 n}\big)^{\frac{1}{2}}$.

*Then*

$$\|\bar{\beta} - \beta^*\|_\infty \le c\sigma_y\Big(\Big(\frac{c_\Omega \log p}{N}\Big)^{\frac{1}{2}} + \frac{c_{\mathrm{GC}}c_\Sigma}{\mu_l}\sigma_y\frac{s\log p}{n}\Big)$$

*with probability at least* $1 - ep^{-1}$, *where* $c > 0$ *is a universal constant,* $c_\Omega := \max_{j\in[p],k\in[m]}((\hat{\Theta}_k\hat{\Sigma}_k\hat{\Theta}_k^T)_{j,j})$ *and* $c_\Sigma := \max_{j\in[p],k\in[m]}((\hat{\Sigma}_k)_{j,j})^{\frac{1}{2}}$.

Lemma 6.9 hints at the performance of the averaged debiased lasso. In particular, we note the first term is $O\big(\big(\frac{\log p}{N}\big)^{\frac{1}{2}}\big)$, which matches the convergence rate of the centralized estimator. When $n$ is large enough, $\frac{s\log p}{n}$ is negligible compared to $\big(\frac{\log p}{N}\big)^{\frac{1}{2}}$, and the error is $O\big(\big(\frac{\log p}{N}\big)^{\frac{1}{2}}\big)$.

Finally, we show the conditions of Lemma 6.9 occur with high probability when the rows of $X$ are independent subgaussian random vectors.

**Theorem 6.10.** *Under (A1), (A2), and (A3), when $m < p$, $p > \tilde{s}$,*

1. $n > \max\{4000\tilde{s}\sigma_x^2\log(\frac{60\sqrt{2}ep}{\tilde{s}}),\, 8000\sigma_x^4\log p,\, \frac{3}{c_1}\max\{\sigma_x^2, \sigma_x\}\log p\}$,
2. *we set* $\lambda_1 = \cdots = \lambda_m = \max_{j\in[p],k\in[m]}\big((\hat{\Sigma}_k)_{j,j}\big)^{\frac{1}{2}}\sigma_y\big(\frac{3\log p}{c_2 n}\big)^{\frac{1}{2}}$,
3. *we set* $\delta_1 = \cdots = \delta_m = \frac{8}{\sqrt{c_1}}\sqrt{\kappa}\sigma_x^2\big(\frac{\log p}{n}\big)^{\frac{1}{2}}$ *and form* $\{\hat{\Theta}_k\}_{k\in[m]}$ *by* (6.2),

$$\|\bar{\beta} - \beta^*\|_\infty \le c\Big(\sigma_y\Big(\frac{\max_{j\in[p]}\Sigma_{j,j}^{-1}\log p}{N}\Big)^{\frac{1}{2}} + \frac{\sqrt{\kappa}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s\log p}{n}\Big)$$

*with probability at least* $1 - (8+e)p^{-1}$ *for some universal constant* $c > 0$.

*Proof.* We start with the conclusion of Lemma 6.9:

$$\|\bar{\beta} - \beta^*\|_\infty \le \sigma_y\Big(\frac{2c_\Omega\log p}{c_2 N}\Big)^{\frac{1}{2}} + \frac{3\sqrt{3}}{\sqrt{c_2}}\frac{c_{\mathrm{GC}}c_\Sigma}{\mu_l}\sigma_y\frac{s\log p}{n}.$$

First, we show that the two constants $c_\Omega = \max_{j\in[p],k\in[m]}(\hat{\Theta}_k\hat{\Sigma}_k\hat{\Theta}_k^T)_{j,j}$ and $c_\Sigma := \max_{j\in[p],k\in[m]}((\hat{\Sigma}_k)_{j,j})^{\frac{1}{2}}$ are bounded with high probability.

**Lemma 6.11.** *Under (A1),*

$$\mathbf{Pr}\big(\max_{j\in[p]}\Sigma_j^{-1}\hat{\Sigma}\Sigma_j^{-1} > 2\max_{j\in[p]}\Sigma_{j,j}^{-1}\big) \le 2pe^{-c_1\min\{\frac{n}{\sigma_x^2},\frac{n}{\sigma_x}\}}$$

*for some universal constant* $c_1 > 0$.

Since we form $\{\hat{\Theta}_k\}_{k\in[m]}$ by (6.2),

$$(\hat{\Theta}_k\hat{\Sigma}_k\hat{\Theta}_k^T)_{j,j} \le \max_{j\in[p]}(\Sigma^{-1}\hat{\Sigma}_k\Sigma^{-1}))_{j,j}.$$

Lemma 6.11 implies

$$\max_{j\in[p]}(\Sigma^{-1}\hat{\Sigma}_k\Sigma^{-1}))_{j,j} \le 2\max_{j\in[p]}\Sigma_{j,j}^{-1} \text{ for each } k\in[m]$$

with probability at least $1 - 2pe^{-c_1\min\{\frac{n}{\sigma_x^2},\frac{n}{\sigma_x}\}}$.

**Lemma 6.12.** *Under (A1),*

$$\mathbf{Pr}(\max_{j\in[p]}(\hat{\Sigma}_{j,j})^{\frac{1}{2}} > \sqrt{2}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}) \le 2pe^{-c_1\min\{\frac{n}{16\sigma_x^2},\frac{n}{4\sigma_x}\}}$$

*for some universal constant* $c_1 > 0$.

We put the pieces together to obtain the stated result:

1. By Lemma 6.11 (and a union bound over $k \in [m]$),

   $$\mathbf{Pr}(c_\Omega \ge 2\max_j \Sigma_{j,j}^{-1}) \le 2mpe^{-c_1\min\{\frac{n}{\sigma_x^2},\frac{n}{\sigma_x}\}}.$$

   Since $m \le p$, when $n > \frac{3}{c_1}\max\{\sigma_x^2, \sigma_x\}\log p$,

   $$\mathbf{Pr}\big(c_\Omega < 2\max_j \Sigma_{j,j}^{-1}\big) \ge 1 - 2p^{-1}.$$

2. By Lemma 6.12 (and a union bound over $k \in [m]$),

   $$\mathbf{Pr}(c_\Sigma < \sqrt{2}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}) \ge 1 - 2mpe^{-c_1\min\{\frac{n}{16\sigma_x^2},\frac{n}{4\sigma_x}\}}.$$

   When $n > \frac{3}{c_1}\max\{\sigma_x^2, \sigma_x\}\log p$, the right side is again at most $2p^{-1}$.

3. By Lemma 6.4, as long as

   $$n > \max\{4000\tilde{s}\sigma_x^2\log(\tfrac{60\sqrt{2}ep}{\tilde{s}}), 8000\sigma_x^4\log p\},$$

   $\hat{\Sigma}_1, \dots, \hat{\Sigma}_m$ all satisfy the RE condition with probability at least

   $$1 - 2me^{-\frac{n}{4000\sigma_x^4}} \ge 1 - 2p^{-1}.$$

4. By Lemma 6.2,

   $$\mathbf{Pr}\big(\cap_{k\in[m]}\mathcal{E}_{\mathrm{GC}}(\hat{\Sigma}_k)\big) \ge 1 - 2p^{-2}.$$

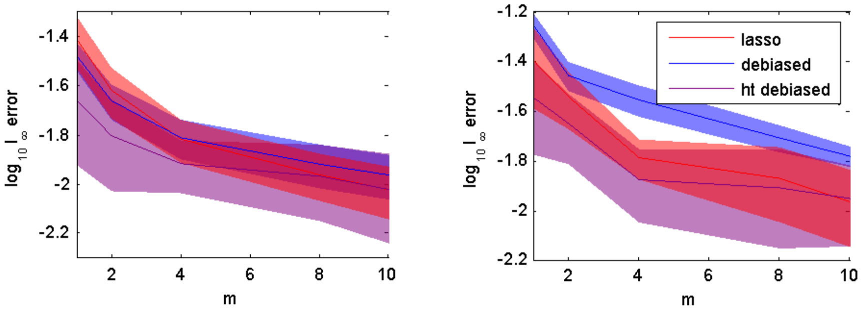
   Since $m < p$, the probability is at least $1 - 2p^{-1}$.

We apply the bounds $c_\Omega \le 2\max_{j\in[p]}\Sigma_{j,j}^{-1}$, $c_\Sigma \le \sqrt{2}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}$, $c_{\mathrm{GC}} = \frac{8}{\sqrt{c_1}}\sqrt{\kappa}\sigma_x^2$, and $\mu_l = \frac{1}{2}\lambda_{\min}(\Sigma)$ to obtain

$$\|\bar{\beta} - \beta^*\|_\infty \le \sigma_y\left(\frac{4\max_{j\in[p]}\Sigma_{j,j}^{-1}\log p}{c_2 N}\right)^{\frac{1}{2}} + \frac{48\sqrt{6}}{\sqrt{c_1c_2}}\frac{\sqrt{\kappa}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s\log p}{n}.$$

□

We validate our theoretical results with simulations. First, we study the estimation error of the averaged debiased lasso in $\ell_\infty$ norm. To focus on the effect of averaging, we grow the number of machines $m$ linearly with the (total) sample size $N$. In other words, we fix the sample size per machine $n$ and grow the total sample size by adding machines. Figure 4 compares the estimation error (in $\ell_\infty$ norm) of the averaged debiased lasso estimator with that of the centralized lasso. We see the estimation error of the averaged debiased lasso estimator is comparable to that of the centralized lasso.



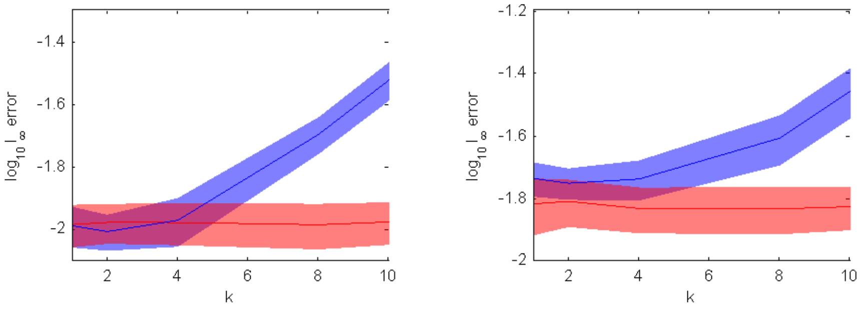
(a) $\Sigma = I$

(b) $(\Sigma)_{i,j} = 0.5^{|i-j|}$

Figure 4: The estimation error (in $\ell_\infty$ norm) of the averaged debiased lasso estimator versus that of the centralized lasso when the predictors are Gaussian. In both settings, the estimation error of the averaged debiased estimator is comparable to that of the centralized lasso.

We conduct a second set of simulations to study the effect of the number of machines on the estimation effor of the averaged estimator. To focus on the effect of the number of machines $k$, we fix the (total) sample size $N$ and vary the number of machines the samples are distributed across. Figure 5 shows how the estimation error (in $\ell_\infty$ norm) of the averaged estimator grows as the number of machines grows. When the number of machines is small, the estimation error of the averaged estimator is comparable to that of the centralized lasso. However, when the number of machines exceeds a certain threshold, the estimation error grows with the number of machines. This is consistent with the prediction of Theorem 6.10: when the number of machines exceeds a certain threshold, the bias term of order $\frac{s \log p}{n}$ becomes dominant.

The averaged debiased lasso has one serious drawback versus the lasso: $\bar{\beta}$ is usually dense. The density of $\bar{\beta}$ detracts from the intrepretability of the



(a) $\Sigma = I$

(b) $(\Sigma)_{i,j} = 0.5^{|i-j|}$

Figure 5: The estimation error (in $\ell_\infty$ norm) of the averaged estimator as the number of machines $m$ vary. When the number of machines is small, the error is comparable to that of the centralized lasso. However, when the number of machines exceeds a certain threshold, the bias term (which grows linearly in $m$) is dominant, and the performance of the averaged estimator degrades.

coefficients and makes the estimation error large in the $\ell_2$ and $\ell_1$ norms. To remedy both problems, we threshold the averaged debiased lasso:

$$\begin{aligned} \mathrm{HT}_t(\bar{\beta}) &\leftarrow \bar{\beta}_j \cdot \mathbf{1}_{\{|\bar{\beta}_j| \geq t\}}, \\ \mathrm{ST}_t(\bar{\beta}) &\leftarrow \mathrm{sign}(\bar{\beta}_j) \cdot \max\{|\bar{\beta}_j| - t, 0\}. \end{aligned}$$

As we shall see, both hard and soft-thresholding give sparse aggregates that are close to $\beta^*$ in $\ell_2$ norm.

**Lemma 6.13.** *As long as $t > \|\bar{\beta} - \beta^*\|_\infty$, $\bar{\beta}^{ht} = \mathrm{HT}_t(\bar{\beta})$ satisfies*

1. $\|\bar{\beta}^{ht} - \beta^*\|_\infty \leq 2t$,
2. $\|\bar{\beta}^{ht} - \beta^*\|_2 \leq 2\sqrt{2s}t$,
3. $\|\bar{\beta}^{ht} - \beta^*\|_1 \leq 2\sqrt{2}st$.

*The analogous result also holds for $\bar{\beta}^{st} = \mathrm{ST}_t(\bar{\beta})$.*

*Proof.* By the triangle inequality,

$$\begin{aligned} \|\bar{\beta}^{ht} - \beta^*\|_\infty &\leq \|\bar{\beta}^{ht} - \bar{\beta}\|_\infty + \|\bar{\beta} - \beta^*\|_\infty \\ &\leq t + \|\bar{\beta} - \beta^*\|_\infty \\ &\leq 2t. \end{aligned}$$

Since $t > \|\bar\beta - \beta^*\|_\infty$, $\bar\beta^{ht}_j = 0$ whenever $\beta^*_j = 0$. Thus $\bar\beta^{ht}$ is $s$-sparse and $\bar\beta^{ht} - \beta^*$ is $2s$-sparse. By the equivalence between the $\ell_\infty$ and $\ell_2$, $\ell_1$ norms,

$$\|\bar\beta^{ht} - \beta^*\|_2 \le 2\sqrt{2s}t,$$
$$\|\bar\beta^{ht} - \beta^*\|_1 \le 2\sqrt{2s}t.$$

The argument for $\bar\beta^{st}$ is similar. □

By combining Lemma 6.13 with Theorem 6.10, we show that $\bar\beta^{ht}$ converges at the same rates as the centralized lasso.

**Theorem 6.14.** *Under the conditions of Theorem 6.10, hard-thresholding $\bar\beta$ at $\sigma_y\left(\frac{4\max_{j\in[p]}\Sigma^{-1}_{j,j}\log p}{c_2N}\right)^{\frac12} + \frac{48\sqrt6}{\sqrt{c_1c_2}}\frac{\sqrt\kappa\max_{j\in[p]}(\Sigma_{j,j})^{\frac12}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s\log p}{n}$ gives*

1. $\|\bar\beta^{ht} - \beta^*\|_\infty \lesssim_P \sigma_y\left(\frac{\max_{j\in[p]}\Sigma^{-1}_{j,j}\log p}{N}\right)^{\frac12} + \frac{\sqrt\kappa\max_{j\in[p]}(\Sigma_{j,j})^{\frac12}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s\log p}{n}$,
2. $\|\bar\beta^{ht} - \beta^*\|_2 \lesssim_P \sigma_y\left(\frac{\max_{j\in[p]}\Sigma^{-1}_{j,j}s\log p}{N}\right)^{\frac12} + \frac{\sqrt\kappa\max_{j\in[p]}(\Sigma_{j,j})^{\frac12}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s^{\frac32}\log p}{n}$,
3. $\|\bar\beta^{ht} - \beta^*\|_1 \lesssim_P \sigma_y\left(\frac{\max_{j\in[p]}\Sigma^{-1}_{j,j}s^2\log p}{N}\right)^{\frac12} + \frac{\sqrt\kappa\max_{j\in[p]}(\Sigma_{j,j})^{\frac12}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s^2\log p}{n}$.

**Remark 6.15.** *By Theorem 6.14, when $m \lesssim \frac{n}{s^2\log p}$, the variance term is dominant and the convergence rates given by the theorem simplify:*

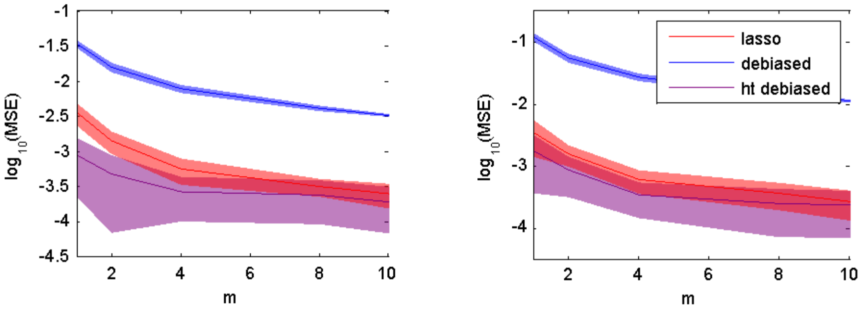
1. $\|\bar\beta^{ht} - \beta^*\|_\infty \lesssim_P \left(\frac{\log p}{N}\right)^{\frac12}$,
2. $\|\bar\beta^{ht} - \beta^*\|_2 \lesssim_P \left(\frac{s\log p}{N}\right)^{\frac12}$,
3. $\|\bar\beta^{ht} - \beta^*\|_1 \lesssim_P \left(\frac{s^2\log p}{N}\right)^{\frac12}$.

*The convergence rates for the centralized lasso estimator $\hat\beta$ are identical (modulo constants):*

1. $\|\hat\beta - \beta^*\|_\infty \lesssim_P \left(\frac{\log p}{N}\right)^{\frac12}$,
2. $\|\hat\beta - \beta^*\|_2 \lesssim_P \left(\frac{s\log p}{N}\right)^{\frac12}$,
3. $\|\hat\beta - \beta^*\|_1 \lesssim_P \left(\frac{s^2\log p}{N}\right)^{\frac12}$.

*The estimator $\bar\beta^{ht}$ matches the convergence rates of the centralized lasso in $\ell_1$, $\ell_2$, and $\ell_\infty$ norms. Furthermore, $\bar\beta^{ht}$ can be evaluated in a communication-efficient manner by a one-shot averaging approach.*

We conduct a third set of simulations to study the effect of thresholding on the estimation error in $\ell_2$ norm. Figure 6 compares the estimation error incurred by the averaged estimator with and without thresholding versus that of the centralized lasso. Since the averaged estimator is usually dense, its estimation error (in $\ell_2$ norm) is large compared to that of the centralized lasso. However, after thresholding, the averaged estimator performs comparably versus the centralized lasso.



(a) $\Sigma = I$

(b) $(\Sigma)_{i,j} = 0.5^{|i-j|}$

Figure 6: The estimation error (in $\ell_2$ norm) of the averaged estimator with and without thresholding versus that of the centralized lasso when the predictors are Gaussian. Although the estimation error of the averaged estimator is large compared to that of the centralized lasso, the thresholded averaged estimator performs comparably versus the centralized lasso.

## 6.3 A DISTRIBUTED APPROACH TO DEBIASING

The averaged estimator we studied has the form

$$\bar{\beta} = \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_k + \hat{\Theta}_k X_k^T (y - X_k \hat{\beta}_k).$$

The estimator requires each machine to form $\hat{\Theta}_k$ by the solution of (6.2). Since the dual of (6.2) is an $\ell_1$-regularized quadratic program:

$$\operatorname*{minimize}_{\gamma \in \mathbf{R}^p} \frac{1}{2}\gamma^T \hat{\Sigma}_k \gamma - \hat{\Sigma}_k \gamma + \delta \left\|\gamma\right\|_1, \tag{6.4}$$

forming $\hat{\Theta}_k$ is (roughly speaking) $p$ times as expensive as solving the local lasso problem, making it the most expensive step (in terms of FLOPS) of evaluating the averaged estimator. To trim the cost of the debiasing step, we consider an estimator that forms only a single $\hat{\Theta}$ :

$$\tilde{\beta} = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_k + \frac{1}{N}\hat{\Theta}\sum_{k=1}^{m}X_k^T(y - X_k\hat{\beta}_k). \tag{6.5}$$

To evaluate (6.5),

1. each machine sends $\hat{\beta}_k$ and $\frac{1}{n}X_k^T(y - X_k\hat{\beta}_k)$ to a central server,
2. the central server forms $\frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_k$ and $\frac{1}{N}\sum_{k=1}^{m}X_k^T(y - X_k\hat{\beta}_k)$ and sends the averages to all the machines,
3. each machine, given the averages, forms $\frac{p}{m}$ rows of $\hat{\Theta}$ and debiases $\frac{p}{m}$ coefficients:

   $$\tilde{\beta}_j = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_j + \hat{\Theta}_j\Big(\frac{1}{N}\sum_{k=1}^{m}X_k^T(y - X_k\hat{\beta}_k)\Big),$$

   where $\hat{\Theta}_j \in \mathbf{R}^p$ is a row vector.

As we shall see, each machine can perform debiasing with only the data stored locally. Thus, forming the estimator (6.5) requires two rounds of communication.

The question that remains is how to form $\hat{\Theta}_j$. We consider an estimator proposed by van de Geer et al. (2013): nodewise regression on the predictors. For some $j \in [p]$ that machine $k$ is debiasing, the machine solves

$$\hat{\gamma}_j := \arg\min_{\gamma\in\mathbf{R}^{p-1}} \frac{1}{2n}\|x_{k,j} - X_{k,-j}\gamma\|_2^2 + \lambda_j\|\gamma\|_1,\ j \in [p],$$

where $X_{k,-j} \in \mathbf{R}^{n\times(p-1)}$ is $X_k$ less its $j$-th column $x_{k,j}$, and forms

$$\hat{C} := \begin{bmatrix} 1 & -\hat{\gamma}_{1,2} & \dots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \dots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \dots & -\hat{\gamma}_{p,p} \end{bmatrix},$$

where the components of $\hat{\gamma}_j$ are indexed by $k \in \{1,\dots,j-1,j+1,\dots,p\}$. Finally, we scale the rows of $\hat{C}$ by $\mathbf{diag}\Big(\big[\hat{\tau}_1,\dots,\hat{\tau}_p\big]\Big)$, where

$$\hat{\tau}_j = (\frac{1}{n}\|x_j - X_{-j}\hat{\gamma}_j\|_2^2 + \lambda_j\|\hat{\gamma}_j\|_1)^{\frac{1}{2}},$$

to form $\hat{\Theta} = \hat{T}^{-2}\hat{C}$. Each row of $\hat{\Theta}$ is given by

$$\hat{\Theta}_j = -\frac{1}{\hat{\tau}_j^2}\left[\hat{\gamma}_{j,1} \quad \dots \quad \hat{\gamma}_{j,j-1} \quad 1 \quad \hat{\gamma}_{j,j+1} \quad \dots \quad \hat{\gamma}_{j,p}\right]. \tag{6.6}$$

Since $\hat{\gamma}_j$ and $\hat{\tau}_j$ only depend on $X_k$, they can be formed without any communication.

van de Geer et al. (2013) show that when the rows of $X$ are *i.i.d.* subgaussian random vectors and the precision matrix $\Sigma^{-1}$ is sparse, $\hat{\Theta}_j$ converges to $\Sigma_j^{-1}$ at the usual convergence rate of the lasso. For completeness, we restate their result.

We consider a sequence of regression problems indexed by the sample size $N$, dimension $p$, sparsity $s_0$ that satisfies (A1), (A2), and (A3). As $N$ grows to infinity, both $p = p(N)$ and $s = s(N)$ may also grow as a function of $N$. To keep notation manageable, we drop the index $N$. We further assume

(A4) the covariance of the predictors (rows of $X$) has smallest eigenvalue $\lambda_{\min}(\Sigma) \sim \Omega(1)$ and largest diagonal entry $\max_{j\in[p]} \Sigma_{j,j} \sim O(1)$,

(A5) the rows of $\Sigma^{-1}$ are sparse: $\max_{j\in[p]} \frac{s_j^2 \log p}{n} \sim o(1)$, where $s_j$ is the sparsity of $\Sigma_j^{-1}$.

**Lemma 6.16** (van de Geer et al. (2013), Theorem 2.4)**.** *Under (A1)–(A5),* (6.6) *with suitable parameters $\lambda_j \sim \left(\frac{\log p}{n}\right)^{\frac{1}{2}}$ satisfies*

$$\|\hat{\Theta}_j - \Sigma_j^{-1}\|_1 \lesssim_P \left(\frac{s_j^2 \log p}{n}\right)^{\frac{1}{2}} \text{ for any } j \in [p].$$

We show the estimator (6.5) matches the convergence rate of the centralized lasso. The argument is similar to the proof of Theorem 6.10.

**Theorem 6.17.** *Under (A1)–(A5),* (6.5)*, where $\hat{\Theta}$ is given by* (6.6)*, with suitable parameters $\lambda$, $\lambda_k \sim \left(\frac{\log p}{n}\right)^{\frac{1}{2}}$, $k \in [m]$ satisfies*

$$\|\bar{\beta} - \beta^*\|_\infty \lesssim_P \left(\frac{\log p}{N}\right)^{\frac{1}{2}} + \frac{\max_{j\in[p+1]} s_j \log p}{n}.$$

*Proof.* We start by substituting the linear model into (6.5):

$$\begin{aligned}\tilde{\beta} &= \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_k - \hat{\Theta}\hat{\Sigma}_k(\hat{\beta}_k - \beta^*) + \frac{1}{n}\hat{\Theta}X_k^T\epsilon_k \\ &= \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_k - \hat{\Theta}\hat{\Sigma}_k(\hat{\beta}_k - \beta^*) + \frac{1}{N}\hat{\Theta}X^T\epsilon.\end{aligned}$$

Subtracting $\beta^*$ and taking norms, we obtain

$$\|\tilde{\beta} - \beta^*\|_\infty \le \frac{1}{m}\sum_{k=1}^{m}\|(I - \hat{\Theta}\hat{\Sigma}_k)(\hat{\beta}_k - \beta^*)\|_\infty + \big\|\frac{1}{N}\hat{\Theta}X^T\epsilon\big\|_\infty. \tag{6.7}$$

By Vershynin (2010), Proposition 5.16, and Lemma (6.11), it is possible to show that

$$\big\|\frac{1}{N}\hat{\Theta}X^T\epsilon\big\|_\infty \lesssim_P \Big(\frac{\log p}{N}\Big)^{\frac{1}{2}}.$$

We turn our attention to the first term in (6.7). It's straightforward to see each term in the sum is bounded by

$$\begin{aligned}&\|(I - \hat{\Theta}\hat{\Sigma}_k)(\hat{\beta}_k - \beta^*)\|_\infty \\ &\quad\le \|(I - \Sigma^{-1}\hat{\Sigma}_k)(\hat{\beta}_k - \beta^*)\|_\infty + \|(\Sigma^{-1} - \hat{\Theta})\hat{\Sigma}_k(\hat{\beta}_k - \beta^*)\|_\infty \\ &\quad\le \max\nolimits_{j\in[p]} \|e_j^T - \Sigma_j^{-1}\hat{\Sigma}_k\|_\infty\|\hat{\beta}_k - \beta^*\|_1 + \|\Sigma_j^{-1} - \hat{\Theta}_j\|_1\|\hat{\Sigma}_k(\hat{\beta}_k - \beta^*)\|_\infty.\end{aligned}$$

We put the pieces together to deduce each term is $O\big(\frac{\max_{j\in[p]} s_j \log p}{n}\big)$ :

1. By Lemmas 6.4, 6.6, 6.12, $\|\hat{\beta}_k - \beta^*\|_1 \lesssim_P \sqrt{s_0}\lambda$.

2. By Lemma 6.16, $\|\Sigma_j^{-1} - \hat{\Theta}_j\|_1 \lesssim_P s_j\big(\frac{\log p}{n}\big)^{\frac{1}{2}}$.

3. By the triangle inequality,

$$\|\hat{\Sigma}_k(\hat{\beta}_k - \beta^*)\|_\infty \le \Big\|\frac{1}{n}X_k^T(y_k - X_k\hat{\beta}_k)\Big\|_\infty + \Big\|\frac{1}{n}X_k^T\epsilon_k\Big\|_\infty.$$

By the optimality conditions of the (local) lasso estimator, the first term is $\lambda$, and it is possible to show, by Lemma 6.11 and Vershynin (2010), Proposition 5.16, that the second term is $O_P\big(\big(\frac{\log p}{n}\big)^{\frac{1}{2}}\big)$.

Since $\lambda \sim \big(\frac{\log p}{n}\big)^{\frac{1}{2}}$, by a union bound over $k \in [m]$, we obtain

$$\|\bar{\beta} - \beta^*\|_\infty \sim O_P\Big(\Big(\frac{\log p}{N}\Big)^{\frac{1}{2}} + \frac{\max_{j\in[p]} s_j \log p}{n}\Big).$$

□

By combining the Lemma 6.13 with Theorem 6.17, we can show that $\tilde{\beta}^{ht} := \mathrm{HT}(\tilde{\beta}, t)$ for an appropriate threshold $t$ converges to $\beta^*$ at the same rates as the centralized lasso.

**Theorem 6.18.** *Under the conditions of Theorem 6.17, hard-thresholding $\tilde{\beta}$ at $t \sim \big(\frac{\log p}{N}\big)^{\frac{1}{2}} + \frac{\max_{j\in[p]} s_j \log p}{n}$ gives*

1. $\|\tilde{\beta}^{ht} - \beta^*\|_\infty \lesssim_P \big(\frac{\log p}{N}\big)^{\frac{1}{2}} + \frac{\max_{j\in[p]} s_j \log p}{n}$,
2. $\|\tilde{\beta}^{ht} - \beta^*\|_2 \lesssim_P \big(\frac{s_0 \log p}{N}\big)^{\frac{1}{2}} + \frac{\sqrt{s_0} \max_{j\in[p]} s_j \log p}{n}$,
3. $\|\tilde{\beta}^{ht} - \beta^*\|_1 \lesssim_P \big(\frac{s_0^2 \log p}{N}\big)^{\frac{1}{2}} + \frac{s_0 \max_{j\in[p]} s_j \log p}{n}$.

Assuming $s \sim s_j$, for any $j \in [p]$, Theorem 6.18 shows that for $m \lesssim \frac{n}{s_0^2 \log p}$, the variance term is dominant, so the convergence rates simplify:

1. $\|\tilde{\beta}^{ht} - \beta^*\|_\infty \lesssim_P \big(\frac{\log p}{N}\big)^{\frac{1}{2}}$,
2. $\|\tilde{\beta}^{ht} - \beta^*\|_2 \lesssim_P \big(\frac{s_0 \log p}{N}\big)^{\frac{1}{2}}$,
3. $\|\tilde{\beta}^{ht} - \beta^*\|_1 \lesssim_P \big(\frac{s_0^2 \log p}{N}\big)^{\frac{1}{2}}$.

Thus, estimator $\tilde{\beta}^{ht}$ shares the advantages of $\bar{\beta}^{ht}$ over the centralized lasso (cf. Remark 6.15). It also achieves computational gains over $\bar{\beta}^{ht}$ by amortizing the cost of debiasing across $m$ machines.

Finally, we mention that it is possible to obtain a sharper result by forgoing the $\ell_\infty$ norm convergence rate. We state the result, but defer its proof to Appendix C.7.

**Theorem 6.19.** *Under the conditions of Theorem 6.17, hard-thresholding $\tilde{\beta}$ at $t = |\tilde{\beta}|_{(\hat{s}_0)}$ for some $\hat{s}_0 \sim s_0$, i. e. setting all but the largest $s_0'$ debiased coefficients to zero, gives*

1. $\|\tilde{\beta}^{ht} - \beta^*\|_2 \lesssim_P \big(\frac{s_0 \log p}{N}\big)^{\frac{1}{2}} + \frac{s_0 \log p}{n}$,
2. $\|\bar{\beta}^{ht} - \beta^*\|_1 \lesssim_P \big(\frac{s_0^2 \log p}{N}\big)^{\frac{1}{2}} + \frac{s_0^{3/2} \log p}{n}$.

As long as $n \gtrsim s_0 \log p$, it is possible to obtain a good estimate of $s_0$ by the *empirical sparsity* of any of the local lasso estimators. Let $\hat{\mathcal{E}} \subset [p]$ be the *equicorrlation set* of the lasso estimator.

$$\{j \in [p] : |x_j^T(y - X\hat{\beta})| = \lambda\}.$$

The empirical sparsity $\hat{s}_0$ is the size of $\hat{\mathcal{E}}$.

**Lemma 6.20.** *Under (A1)–(A3), when*

$$n > \max\{4000\tilde{s}_0\sigma_x^2 \log(\tfrac{60\sqrt{2}ep}{\tilde{s}_0}), 4000\sigma_x^4 \log p, s_0 \log p\},$$

*where* $\tilde{s}_0 := s_0 + 25920\kappa s_0$, *we have*

$$\hat{s}_0 \leq \Big(\frac{192\sigma_x^2 + 384\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} + \frac{384^2\sigma_x^4}{c_1\lambda_{\min}(\Sigma)^2}\Big)^2 s$$

*with probability at least* $1 - 2p^{-(s_0+1)}$.

By Theorem 6.19, when $m \lesssim \frac{N}{s_0 \log p}$, the variance term is dominant and the convergence rates given by the theorem simplify to the convergence rates of the (centralized) lasso estimator:

1. $\|\bar{\beta}^{ht} - \beta^*\|_2 \lesssim_P \big(\frac{s_0 \log p}{N}\big)^{\frac{1}{2}}$,

2. $\|\bar{\beta}^{ht} - \beta^*\|_1 \lesssim_P \big(\frac{s_0^2 \log p}{N}\big)^{\frac{1}{2}}$.

Thus, by forgoing consistency in the $\ell_\infty$ norm, it is possible to reduce the sample complexity of the averaged estimator to $m \lesssim \frac{s \log p}{N}$. When $m = 1$, we recover the sample complexity of the lasso estimator.

# 7 SUMMARY AND DISCUSSION

This thesis has two parts: estimation and computing. Although estimation and inference have been core topics in statistics since its inception as a scientific discipline, computation is a recent addition. However, as the size and complexity of datasets continue to grow, the computational aspects of statistical practice become ever more important.

In this thesis, we study the statistical and computational properties of regularized M-estimators of the form (1.3). Although regularization is an old idea, the emergence of high-dimensional datasets in modern science and engineering has led to a resurgence of interest in the idea. Statistically, regularization is essential: it prevents overfitting and allows us to design estimators that discover latent low-dimensional structure in the data. Computationally, it improves the stability of the estimator and often leads to computational gains.

## 7.1 ESTIMATION AND INFERENCE

The first part focuses on the statistical properties of regularized M-estimators. The estimators are used in diverse areas of science and engineering to fit high-dimensional models with some low-dimensional structure. In Chapter 2, we develop a framework for establishing consistency and model selection consistency of regularized M-estimators on high-dimensional problems.[1] Our analysis identifies two key properties of regularized M-estimators that ensure consistency and model selection consistency: geometric decomposability and irrepresentability. We also showed that for an estimator to be consistent and model selection consistent, irrepresentability is necessary.

We only studied the "first-order correctness" of regularized M-estimators in the high-dimensional setting. As our understanding of first-order properties becomes more complete, attention has shifted to "second-order" properties, including testing (statistical) hypotheses and forming confidence intervals. In a separate series of papers, beginning with Lee et al. (2013), we cast the high-dimensional inference problem as a selective inference prob-

[1] Recall our notion of model selection consistency means the estimator falls in the model subspace with high probability. In the context of sparse regression, it means the fitted coefficients of the truly irrelevant predictors are zero.

lem and propose an approach based on a new framework for selective inference in linear models. At the core of the framework is a result that characterizes the distribution of a post-selection estimator conditioned on the selection event. We specialize the approach to model selection by the lasso to obtain *exact* (non-asymptotically valid) confidence intervals for the regression coefficients. Related work by Taylor et al. (2014), Lee and Taylor (2014), Sun and Taylor (2014), Reid and Tibshirani (2014), Fithian et al. (2014), Choi et al. (2014), Tian et al. (2015) generalizes the approach to other statistical models and selection strategies. Further investigation of the effects of regularization on the second-order correctness of estimators remains an exciting area of future research.

## 7.2 COMPUTING

The second part focuses on algorithms to evaluate regularized M-estimators. The estimators are usually expressed as the solution to composite function minimization problems. Recently there has been a flurry of activity around the development of Newton-type methods for minimizing composite functions. Most of the proposed methods fall under the umbrella of proximal Newton-type methods. In Chapter 4 we analyze the methods and show that they inherit the fast local convergence properties of Newton-type methods for minimizing smooth functions, even when the search directions are computed inexactly.

Follow-up work by Tran-Dinh et al. (2015) studies the convergence of the proximal Newton method on composite functions where the smooth part is *self-concordant*. By assuming self-concordance, they show global rates of convergence and characterize the region of quadratic convergence. It remains an open problem to study the convergence rate of inexact proximal Newton-type methods on self-concordant functions.

In Chapter 6 we describe a communication-efficient approach to sparse regression when the samples $\mathcal{Z}^n = \{(x_1, y_1), \dots, (x_N, y_N)\}$ are stored in a distributed fashion. On modern distributed computing platforms, communication is often the dominant cost of computing, and avoiding communication is usually the primary tenet of algorithm design. Our approach, based on the idea of averaging debiased lassos, requires only a single round of communication. We show that as long as the data is not split across too many machines, the averaged estimator achieves the convergence rate of the centralized lasso estimator.

Traditionally, the benchmark of an estimator is its statistical efficiency. However, as the size and complexity of datasets continue to grow, the computational efficiency of an estimator is becoming increasingly important. In recent years, the literature on the statistical benefits of regularization has grown vastly. For many high-dimensional problems, there are essentially minimax optimal estimators (modulo constants) that depend on regularization to achieve optimality. However, the computational benefits of regularization are not as well understood. A line of work beginning with Agarwal et al. (2012) show that the proximal gradient method on (1.3) converges linearly, even when the objective function is not strongly convex (and strong convexity is impossible when $n > p$). However, similar results on other methods are scarce. A key outstanding challenge is to design computationally and statistically optimal estimators for emerging applications.

Part III

# THE TECHNICAL DETAILS

# A

# PROOFS OF LEMMAS IN CHAPTERS 2 AND 3

## A.1 PROOF OF LEMMA 2.7

Since $\hat{\theta}$ is the solution to the restricted problem,

$$\ell_n(\hat{\theta}) + \lambda h_{\mathcal{A}}(\hat{\theta}) \le \ell_n(\theta^*) + \lambda h_{\mathcal{A}}(\theta^*).$$

Since $\hat{\theta} \in \mathcal{C} \cap \mathcal{M}$ and $\ell_n$ is strongly convex on $\mathcal{C} \cap \mathcal{M}$, $\hat{\theta}$ is the unique solution to (2.11). Again, by the restricted strong convexity of $\ell_n$,

$$\nabla \ell_n(\theta^*)^T P_{\mathcal{M}}(\hat{\theta} - \theta^*) + \frac{\mu_l}{2}\|\hat{\theta} - \theta^*\|_2^2 + \lambda(\rho(\hat{\theta}) - \rho(\theta^*)) \le 0.$$

We take norms to obtain

$$\begin{aligned} 0 &\ge -\bar{\rho}^*(P_{\mathcal{M}}\nabla \ell_n(\theta^*))\bar{\rho}(\hat{\theta} - \theta^*) + \frac{\mu_l}{2}\left\|\hat{\theta} - \theta^*\right\|_2^2 - \lambda\rho(\hat{\theta} - \theta^*) \\ &\ge -\kappa_{\bar{\rho}}\bar{\rho}^*(P_{\mathcal{M}}\nabla \ell_n(\theta^*))\left\|\hat{\theta} - \theta^*\right\|_2 + \frac{\mu_l}{2}\left\|\hat{\theta} - \theta^*\right\|_2^2 - \lambda\rho(\hat{\theta} - \theta^*). \end{aligned}$$

Further, since $\hat{\theta} - \theta^* \in \mathcal{M}$,

$$0 \ge -\kappa_{\bar{\rho}}\bar{\rho}^*(P_{\mathcal{M}}\nabla \ell_n(\theta^*))\left\|\hat{\theta} - \theta^*\right\|_2 + \frac{\mu_l}{2}\left\|\hat{\theta} - \theta^*\right\|_2^2 - \kappa_\rho\lambda\left\|\hat{\theta} - \theta^*\right\|_2.$$

We rearrange to obtain

$$\left\|\hat{\theta} - \theta^*\right\|_2 \le \frac{2}{\mu_l}\left(\kappa_{\bar{\rho}}\bar{\rho}^*(P_{\mathcal{M}}\nabla \ell_n(\theta^*)) + \kappa_\rho\lambda\right).$$

We substitute in $\lambda > \frac{4\kappa_{\text{ir}}}{\delta}\bar{\rho}^*(P_{\mathcal{M}}\nabla \ell_n(\theta^*))$ to obtain the stated conclusion.

## A.2 PROOF OF LEMMA 2.8

Suppose the original problem has two optimal solutions: i. e. there are two pairs $(\hat{\theta}_1, \hat{z}_{\mathcal{A},1}, \hat{z}_{\mathcal{I},1})$ and $(\hat{\theta}_2, \hat{z}_{\mathcal{A},2}, \hat{z}_{\mathcal{I},2})$ that satisfy

$$\begin{aligned} \nabla \ell_n(\hat{\theta}_1) + \lambda(\hat{z}_{\mathcal{A},1} + \hat{z}_{\mathcal{I},1}) &= 0 \\ \nabla \ell_n(\hat{\theta}_2) + \lambda(\hat{z}_{\mathcal{A},2} + \hat{z}_{\mathcal{I},2}) &= 0. \end{aligned} \tag{A.1}$$

Since the original problem is convex, the optimal value is unique:

$$\begin{aligned}\ell_n(\hat\theta_1) + P(\hat\theta_1) &= \ell_n(\hat\theta_1) + \lambda(\hat z_{\mathcal{A},1} + \hat z_{\mathcal{I},1})^T \hat\theta_1 \\ &= \ell_n(\hat\theta_2) + P(\hat\theta_2) = \ell_n(\hat\theta_2) + \lambda(\hat z_{\mathcal{A},2} + \hat z_{\mathcal{I},2})^T \hat\theta_2.\end{aligned}$$

We subtract $\lambda(\hat z_{\mathcal{A},1} + \hat z_{\mathcal{I},1})^T \hat\theta_2$ from both sides to obtain

$$\begin{aligned}&\ell_n(\hat\theta_1) + \lambda(\hat z_{\mathcal{A},1} + \hat z_{\mathcal{I},1})^T (\hat\theta_1 - \hat\theta_2) \\ &\quad = \ell_n(\hat\theta_2) + \lambda(\hat z_{\mathcal{A},2} + \hat z_{\mathcal{I},2} - \hat z_{\mathcal{A},1} - \hat z_{\mathcal{I},1})^T \hat\theta_2.\end{aligned}$$

Rearranging,

$$\begin{aligned}&\ell_n(\hat\theta_1) - \ell_n(\hat\theta_2) + \lambda(\hat z_{\mathcal{A},1} + \hat z_{\mathcal{I},1})^T (\hat\theta_1 - \hat\theta_2) \\ &\quad = \lambda(\hat z_{\mathcal{A},2} + \hat z_{\mathcal{I},2} - \hat z_{\mathcal{A},1} - \hat z_{\mathcal{I},1})^T \hat\theta_2.\end{aligned}$$

We substitute in (A.1) to obtain

$$\begin{aligned}&\ell_n(\hat\theta_1) - \ell_n(\hat\theta_2) - \nabla \ell_n(\hat\theta_1)^T (\hat\theta_1 - \hat\theta_2) \\ &\quad = \lambda(\hat z_{\mathcal{A},2} + \hat z_{\mathcal{I},2} - \hat z_{\mathcal{A},1} - \hat z_{\mathcal{I},1})^T \hat\theta_2.\end{aligned}$$

Since $\ell_n$ is convex, the left side is non-positive, which implies

$$(\hat z_{\mathcal{A},2} + \hat z_{\mathcal{I},2})^T \hat\theta_2 \le (\hat z_{\mathcal{A},1} + \hat z_{\mathcal{I},1})^T \hat\theta_2.$$

But we also know

$$(\hat z_{\mathcal{A},1} + \hat z_{\mathcal{I},1})^T \hat\theta_2 \le h_{\mathcal{A}}(\hat\theta_2) + \hat z_{\mathcal{I},1}^T \hat\theta_2 = \hat z_{\mathcal{A},2}^T \hat\theta_2 + \hat z_{\mathcal{I},1}^T \hat\theta_2.$$

We combine the two inequalities to obtain

$$(\hat z_{\mathcal{A},2} + \hat z_{\mathcal{I},2})^T \hat\theta_2 \le (\hat z_{\mathcal{A},1} + \hat z_{\mathcal{I},1})^T \hat\theta_2 \le \hat z_{\mathcal{A},2}^T \hat\theta_2 + \hat z_{\mathcal{I},1}^T \hat\theta_2,$$

which implies $\hat z_{\mathcal{I},2}^T \hat\theta_2 \le \hat z_{\mathcal{I},1}^T \hat\theta_2$. If $\hat z_{\mathcal{I},1} \in \operatorname{relint}(\mathcal{I})$ but $\hat\theta_2$ has a component in $\operatorname{span}(\mathcal{I})$, we arrive at a contradiction:

$$\hat z_{\mathcal{I},1}^T \hat\theta_2 < h_{\mathcal{I}}(\hat\theta_2) = \hat z_{\mathcal{I},2}^T \hat\theta_2.$$

Thus $\hat\theta_2$ has no component in $\operatorname{span}(\mathcal{I})$.

## A.3 PROOF OF LEMMA 2.9

The Taylor remainder term is

$$R = \nabla \ell_n(\hat{\theta}) - \nabla \ell_n(\theta^*) - Q(\hat{\theta} - \theta^*).$$

By the mean value theorem (along $\hat{\theta} - \theta^*$), we have

$$R = \int_0^1 \big(\nabla^2 \ell_n(\theta^* + \alpha(\hat{\theta} - \theta^*)) - Q_n\big)(\hat{\theta} - \theta^*)\, d\alpha.$$

Since $\ell_n$ is strongly smooth on $\mathcal{C} \cap \mathcal{M}$ with constant $\mu_l$,

$$\begin{aligned}
\|R\|_2 &= \left\| \int_0^1 \big(\nabla^2 \ell_n(\theta^* + \alpha(\hat{\theta} - \theta^*)) - Q_n\big)(\hat{\theta} - \theta^*)\, d\alpha \right\|_2 \\
&\leq \int_0^1 \left\|\nabla^2 \ell_n(\theta^* + \alpha(\hat{\theta} - \theta^*)) - Q_n\right\| \left\|\hat{\theta} - \theta^*\right\|_2 d\alpha \\
&\leq \int_0^1 \mu_u \left\|\hat{\theta} - \theta^*\right\|_2^2 \alpha\, d\alpha \\
&\leq \frac{\mu_u}{2} \left\|\hat{\theta} - \theta^*\right\|_2^2.
\end{aligned}$$

By Lemma 2.7,

$$\|R\|_2 \leq \frac{2\mu_u}{\mu_l^2}\left(\kappa_\rho + \frac{\delta \kappa_{\bar{\rho}}}{4\kappa_{\mathrm{ir}}}\right)^2 \lambda^2.$$

To ensure $\frac{\kappa_{\mathrm{ir}}}{\lambda}\bar{\rho}^*(R) \leq \frac{\delta}{4}$, it sufficies to ensure $\frac{\kappa_{\mathrm{ir}}}{\lambda}\|R\|_2 \leq \frac{\delta}{4\kappa_{\bar{\rho}^*}}$. We recall $\lambda$ is in the interval (2.10) to obtain the stated conclusion.

## A.4 PROOF OF LEMMA 3.4

For any $\Delta \in \mathrm{span}(I)^\perp$, we have

$$\begin{aligned}
&\|\Theta^* + \Delta\|_* - \|\Theta^*\|_* - \mathrm{tr}\big(V_r U_r^T \Delta\big) \\
&\quad = \mathrm{tr}\big(\tilde{V}_r \tilde{U}_r^T (\Theta^* + \Delta)\big) - \mathrm{tr}\big(V_r U_r^T \Theta^*\big) - \mathrm{tr}\big(V_r U_r^T \Delta\big),
\end{aligned}$$

where $\tilde{U} \in \mathbf{R}^{p_1 \times r}$ and $\tilde{V} \in \mathbf{R}^{p_2 \times r}$ are the left and right singular factors of $\Theta^* + \Delta$. Since $\mathrm{tr}\big(\tilde{V}_r \tilde{U}_r^T \Theta^*\big) \leq \mathrm{tr}\big(V_r U_r^T \Theta^*\big)$,

$$\begin{aligned}
&\|\Theta^* + \Delta\|_* - \|\Theta^*\|_* - \mathrm{tr}\big(V_r U_r^T \Delta\big) \leq \mathrm{tr}\big(\big(\tilde{U}_r \tilde{V}_r^T - U_r V_r^T\big)^T \Delta\big) \\
&\quad \leq \left\|\tilde{U}_r \tilde{V}_r^T - U_r V_r^T\right\|_{\mathrm{F}} \|\Delta\|_{\mathrm{F}}.
\end{aligned}$$

By Li and Sun (2002), Theorem 2.4,

$$\left\|\tilde{U}_r\tilde{V}_r^T - U_rV_r^T\right\|_{\mathrm{F}} \leq \frac{4}{3\sigma_r^*}\|\Delta\|_F$$

for any $\Delta$ such that $\|\Delta\|_2 \leq \frac{1}{2}\sigma_r^*$. We put the pieces together to obtain the stated bound.

# B

# PROOFS OF LEMMAS IN CHAPTERS 4 AND 5

## B.1 PROOF OF LEMMA 4.2

For any $\alpha \in (0,1]$,

$$\begin{aligned}
&f(x_{t+1}) - f(x_t) \\
&\quad = \phi_{\text{sm}}(x_{t+1}) - \phi_{\text{sm}}(x_t) + \phi_{\text{ns}}(x_{t+1}) - \phi_{\text{ns}}(x_t) \\
&\quad \leq \phi_{\text{sm}}(x_{t+1}) - \phi_{\text{sm}}(x_t) + \alpha\phi_{\text{ns}}(x_t + \Delta x_t) + (1-\alpha)\phi_{\text{ns}}(x_t) - \phi_{\text{ns}}(x_t) \\
&\quad = \phi_{\text{sm}}(x_{t+1}) - \phi_{\text{sm}}(x_t) + \alpha(\phi_{\text{ns}}(x_t + \Delta x_t) - \phi_{\text{ns}}(x_t)) \\
&\quad = \nabla\phi_{\text{sm}}(x_t)^T(\alpha\Delta x_t) + \alpha(\phi_{\text{ns}}(x_t + \Delta x_t) - \phi_{\text{ns}}(x_t)) + O(\alpha^2),
\end{aligned}$$

which shows (4.7).

By the optimality of $\Delta x_t$, $\alpha\Delta x_t$ satisfies

$$\begin{aligned}
&\nabla\phi_{\text{sm}}(x_t)^T\Delta x_t + \frac{1}{2}\Delta x_t^T H\Delta x_t + \phi_{\text{ns}}(x_t + \Delta x_t) \\
&\quad \leq \nabla\phi_{\text{sm}}(x_t)^T(\alpha\Delta x_t) + \frac{\alpha^2}{2}\Delta x_t^T H\Delta x_t + \phi_{\text{ns}}(x_{t+1}) \\
&\quad \leq \alpha\nabla\phi_{\text{sm}}(x_t)^T\Delta x_t + \frac{\alpha^2}{2}\Delta x_t^T H\Delta x_t + \alpha\phi_{\text{ns}}(x_t + \Delta x_t) + (1-\alpha)\phi_{\text{ns}}(x_t).
\end{aligned}$$

We rearrange and then simplify:

$$\begin{aligned}
(1-\alpha)\nabla\phi_{\text{sm}}(x_t)^T\Delta x_t + \frac{1-\alpha^2}{2}\Delta x_t^T H\Delta x_t + (1-\alpha)(\phi_{\text{ns}}(x_t + \Delta x_t) - \phi_{\text{ns}}(x_t)) &\leq 0 \\
\nabla\phi_{\text{sm}}(x_t)^T\Delta x_t + \frac{1+\alpha}{2}\Delta x_t^T H\Delta x_t + \phi_{\text{ns}}(x_t + \Delta x_t) - \phi_{\text{ns}}(x_t) &\leq 0 \\
\nabla\phi_{\text{sm}}(x_t)^T\Delta x_t + \phi_{\text{ns}}(x_t + \Delta x_t) - \phi_{\text{ns}}(x_t) &\leq -\frac{1+\alpha}{2}\Delta x_t^T H\Delta x_t.
\end{aligned}$$

Finally, we let $\alpha$ tend to 1 and rearrange to obtain (4.8).

## B.2 PROOF OF LEMMA 4.3

We bound the decrease at each iteration by

$$\begin{aligned}\phi(x_{t+1}) - \phi(x_t) &= \phi_{\text{sm}}(x_{t+1}) - \phi_{\text{sm}}(x_t) + \phi_{\text{ns}}(x_{t+1}) - \phi_{\text{ns}}(x_t)\\ &\le \int_0^1 \nabla\phi_{\text{sm}}(x_t + t(\alpha\Delta x_t))^T(\alpha\Delta x_t)dt + \alpha\phi_{\text{ns}}(x_t+\Delta x_t)\\ &\quad + (1-\alpha)\phi_{\text{ns}}(x_t) - \phi_{\text{ns}}(x_t)\\ &= \nabla\phi_{\text{sm}}(x_t)^T(\alpha\Delta x_t) + \alpha(\phi_{\text{ns}}(x_t+\Delta x_t) - \phi_{\text{ns}}(x_t))\\ &\quad + \int_0^1 (\nabla\phi_{\text{sm}}(x_t + t(\alpha\Delta x_t)) - \nabla\phi_{\text{sm}}(x_t))^T(\alpha\Delta x_t)dt\\ &\le \alpha\big(\nabla\phi_{\text{sm}}(x_t)^T\Delta x_t + \phi_{\text{ns}}(x_t+\Delta x_t) - \phi_{\text{ns}}(x_t)\\ &\quad + \int_0^1 \left\|\nabla\phi_{\text{sm}}(x_t + t(\Delta x_t)) - \nabla\phi_{\text{sm}}(x_t)\right\|_2 \left\|\Delta x_t\right\|_2 dt\big).\end{aligned}$$

Since $\nabla\phi_{\text{sm}}$ is Lipschitz continuous with constant $\mu_u$,

$$\begin{aligned}\phi(x_{t+1}) - \phi(x_t) &\le \alpha\big(\nabla\phi_{\text{sm}}(x_t)^T\Delta x_t + \phi_{\text{ns}}(x_t+\Delta x_t) - \phi_{\text{ns}}(x_t) + \frac{\mu_u t}{2}\left\|\Delta x_t\right\|_2^2\big)\\ &= \alpha\big(\delta_t + \frac{\mu_u t}{2}\left\|\Delta x_t\right\|_2^2\big). \qquad \text{(B.1)}\end{aligned}$$

If we choose $\alpha \le \frac{\mu_l}{\mu_u}$, then

$$\frac{\alpha\mu_u}{2}\left\|\Delta x_t\right\|_2^2 \le \frac{\mu_l}{2}\left\|\Delta x_t\right\|_2^2 \le \frac{1}{2}\Delta x_t^T H \Delta x_t.$$

By (4.8), we have $\frac{\alpha\mu_u}{2}\left\|\Delta x_t\right\|_2^2 \le -\frac{1}{2}\delta_t$. We combine this bound with (B.1) to conclude

$$\phi(x_{t+1}) - \phi(x_t) \le \alpha\big(\delta_t - \frac{1}{2}\delta_t\big) = \frac{\alpha}{2}\delta_t.$$

## B.3 PROOF OF LEMMAS 4.5 AND 4.7

Since Lemma 4.5 is a special case of Lemma 4.7, we focus on proving Lemma 4.7. Since $\nabla^2\phi_{\text{sm}}$ is Lipschitz continuous,

$$\phi_{\text{sm}}(x_t+\Delta x_t) \le \phi_{\text{sm}}(x_t) + \nabla\phi_{\text{sm}}(x)^T\Delta x_t + \frac{1}{2}\Delta x_t^T\nabla^2\phi_{\text{sm}}(x_t)\Delta x_t + \frac{\mu_u'}{6}\left\|\Delta x_t\right\|_2^3.$$

We add $\phi_{\text{ns}}(x+\Delta x_t)$ to both sides to obtain

$$\begin{aligned}\phi(x_t+\Delta x_t) \le \phi_{\text{sm}}(x_t) &+ \nabla\phi_{\text{sm}}(x)^T\Delta x_t + \frac{1}{2}\Delta x_t^T\nabla^2\phi_{\text{sm}}(x_t)\Delta x_t \\ &+ \frac{\mu_u'}{6}\|\Delta x_t\|_2^3 + \phi_{\text{ns}}(x+\Delta x_t).\end{aligned}$$

We then add and subtract $\phi_{\text{ns}}(x)$ from the right-hand side to obtain

$$\begin{aligned}\phi(x_t+\Delta x_t) &\le \phi_{\text{sm}}(x_t) + \phi_{\text{ns}}(x) + \nabla\phi_{\text{sm}}(x)^T\Delta x_t + \phi_{\text{ns}}(x+\Delta x_t) - \phi_{\text{ns}}(x) \\ &\quad + \frac{1}{2}\Delta x_t^T\nabla^2\phi_{\text{sm}}(x_t)\Delta x_t + \frac{\mu_u'}{6}\|\Delta x_t\|_2^3 \\ &\le \phi(x_t) + \delta_t + \frac{1}{2}\Delta x_t^T\nabla^2\phi_{\text{sm}}(x_t)\Delta x_t + \frac{\mu_u'}{6}\|\Delta x_t\|_2^3 \\ &\le \phi(x_t) + \delta_t + \frac{1}{2}\Delta x_t^T\nabla^2\phi_{\text{sm}}(x_t)\Delta x_t + \frac{\mu_u'}{6\mu_l}\|\Delta x_t\|_2\,\delta_t,\end{aligned}$$

where we use (4.8). We add and subtract $\frac{1}{2}\Delta x_t^T H\Delta x_t$ to obtain

$$\begin{aligned}\phi(x_t+\Delta x_t) &\le \phi(x_t) + \delta_t + \frac{1}{2}\Delta x_t^T\left(\nabla^2\phi_{\text{sm}}(x_t) - H\right)\Delta x_t + \frac{1}{2}\Delta x_t^T H\Delta x_t + \frac{\mu_u'}{6\mu_l}\|\Delta x_t\|_2\,\delta_t \\ &\le \phi(x_t) + \delta_t + \frac{1}{2}\Delta x_t^T\left(\nabla^2\phi_{\text{sm}}(x_t) - H\right)\Delta x_t - \frac{1}{2}\delta_t + \frac{\mu_u'}{6\mu_l}\|\Delta x_t\|_2\,\delta_t,\end{aligned} \tag{B.2}$$

where we again use (4.8). Since $\nabla^2\phi_{\text{sm}}$ is Lipschitz continuous and the search direction $\Delta x_t$ satisfies the Dennis-Moré criterion,

$$\begin{aligned}&\frac{1}{2}\Delta x_t^T\left(\nabla^2\phi_{\text{sm}}(x_t) - H\right)\Delta x_t \\ &\quad = \frac{1}{2}\Delta x_t^T\left(\nabla^2\phi_{\text{sm}}(x_t) - \nabla^2\phi_{\text{sm}}(x_t^\star)\right)\Delta x_t + \frac{1}{2}\Delta x_t^T\left(\nabla^2\phi_{\text{sm}}(x_t^\star) - H\right)\Delta x_t \\ &\quad \le \frac{1}{2}\left\|\nabla^2\phi_{\text{sm}}(x_t) - \nabla^2\phi_{\text{sm}}(x_t^\star)\right\|_2\|\Delta x_t\|_2^2 + \frac{1}{2}\left\|\left(\nabla^2\phi_{\text{sm}}(x_t^\star) - H\right)\Delta x_t\right\|_2\|\Delta x_t\|_2 \\ &\quad \le \frac{\mu_u'}{2}\|x - x^*\|_2\|\Delta x_t\|_2^2 + o\left(\|\Delta x_t\|_2^2\right).\end{aligned}$$

We substitute this expression into (B.2) and rearrange to obtain

$$\phi(x_t+\Delta x_t) \le \phi(x_t) + \frac{1}{2}\delta_t + o\left(\|\Delta x_t\|_2^2\right) + \frac{\mu_u'}{6\mu_l}\|\Delta x_t\|_2\,\delta_t.$$

It is possible to show that $\Delta x_t$ decays to zero by the same argument used to prove Theorem 4.4. Thus $\phi(x_t + \Delta x_t) - \phi(x_t) < \frac{1}{2}\delta_t$ after sufficiently many iterations.

## B.4 PROOF OF LEMMA 4.8

By (4.7) and Fermat's rule, $\Delta x_1$ and $\Delta x_2$ are also the solutions to

$$\begin{aligned} \Delta x_1 &= \arg\min_d \; \nabla\phi_{\text{sm}}(x)^T d + \Delta x_1^T H_1 d + \phi_{\text{ns}}(x+d), \\ \Delta x_2 &= \arg\min_d \; \nabla\phi_{\text{sm}}(x)^T d + \Delta x_2^T H_2 d + \phi_{\text{ns}}(x+d). \end{aligned}$$

Thus $\Delta x_1$ and $\Delta x_2$ satisfy

$$\begin{aligned} &\nabla\phi_{\text{sm}}(x)^T \Delta x_1 + \Delta x_1^T H_1 \Delta x_1 + \phi_{\text{ns}}(x + \Delta x_1) \\ &\quad \le \nabla\phi_{\text{sm}}(x)^T \Delta x_2 + \Delta x_2^T H_1 \Delta x_2 + \phi_{\text{ns}}(x + \Delta x_2) \end{aligned}$$

and

$$\begin{aligned} &\nabla\phi_{\text{sm}}(x)^T \Delta x_2 + \Delta x_2^T H_2 \Delta x_2 + \phi_{\text{ns}}(x + \Delta x_2) \\ &\quad \le \nabla\phi_{\text{sm}}(x)^T \Delta x_1 + \Delta x_1^T H_2 \Delta x_1 + \phi_{\text{ns}}(x + \Delta x_1). \end{aligned}$$

We sum these two inequalities and rearrange to obtain

$$\Delta x_1^T H_1 \Delta x_1 - \Delta x_1^T (H_1 + H_2) \Delta x_2 + \Delta x_2^T H_2 \Delta x_2 \le 0.$$

We then complete the square on the left side and rearrange to obtain

$$\begin{aligned} &\Delta x_1^T H_1 \Delta x_1 - 2\Delta x_1^T H_1 \Delta x_2 + \Delta x_2^T H_1 \Delta x_2 \\ &\quad \le \Delta x_1^T (H_2 - H_1) \Delta x_2 + \Delta x_2^T (H_1 - H_2) \Delta x_2. \end{aligned}$$

The left side is $\|\Delta x_1 - \Delta x_2\|_{H_1}^2$ and the eigenvalues of $H_1$ are bounded. Thus

$$\begin{aligned} \|\Delta x_1 - \Delta x_2\|_2 &\le \frac{1}{\sqrt{\mu_{l,1}}} \left( \Delta x_1^T (H_2 - H_1) \Delta x_2 + \Delta x_2^T (H_1 - H_2) \Delta x_2 \right)^{1/2} \\ &\le \frac{1}{\sqrt{\mu_{l,1}}} \|(H_2 - H_1)\Delta x_2\|_2^{1/2} (\|\Delta x_1\|_2 + \|\Delta x_2\|_2)^{1/2}. \qquad \text{(B.3)} \end{aligned}$$

We apply Tseng and Yun (2009), Lemma 3 to bound $\|\Delta x_1\|_2 + \|\Delta x_2\|_2$. Let $\tilde{H}_1 = H_2^{-1/2} H_1 H_2^{-1/2}$. Then $\|\Delta x_1\|_2$ and $\|\Delta x_2\|_2$ satisfy

$$\|\Delta x_1\|_2 \le \frac{1 + \tilde{\mu}_{u,1} + (1 - 2\tilde{\mu}_{l,1} + \tilde{\mu}_{u,1}^2)^{1/2}}{2} \frac{\mu_{u,1}}{\mu_{l,2}} \|\Delta x_2\|_2 .$$

We denote the constant in front of $\|\Delta x_2\|_2$ by $c_1$ and conclude that

$$\|\Delta x_1\|_2 + \|\Delta x_2\|_2 \le (1 + c_1) \|\Delta x_2\|_2 . \tag{B.4}$$

We substitute (B.4) into (B.3) to obtain

$$\|\Delta x_1 - \Delta x_2\|_2^2 \le \sqrt{\tfrac{1+c_1}{\mu_{l,1}}} \left\|(H_2 - H_1)\Delta x_2\right\|_2^{1/2} \|\Delta x_2\|_2^{1/2} .$$

## B.5 PROOF OF LEMMA 5.1

*Proof.* By the non-expansiveness of the proximal mapping,

$$\begin{aligned} \|\phi_{\text{sm}}(x) - \hat{g}_t(x)\|_2 &= \left\|\text{prox}_{\phi_{\text{ns}}}(x - \nabla\phi_{\text{sm}}(x)) - \text{prox}_{\phi_{\text{ns}}}(x - \nabla\hat{\phi}_{\text{sm},t}(x))\right\|_2 \\ &\le \left\|\nabla\phi_{\text{sm}}(x) - \nabla\hat{\phi}_{\text{sm},t}(x)\right\|_2 . \end{aligned}$$

Since $\nabla\phi_{\text{sm}}$ and $\nabla^2\phi_{\text{sm}}$ are Lipschitz continuous,

$$\begin{aligned} \left\|\nabla\phi_{\text{sm}}(x) - \nabla\hat{\phi}_{\text{sm},t}(x)\right\|_2 &\le \left\|\nabla\phi_{\text{sm}}(x) - \nabla\phi_{\text{sm}}(x_t) - \nabla^2\phi_{\text{sm}}(x_t)(x - x_t)\right\|_2 \\ &\le \frac{\mu_u'}{2} \|x - x_t\|_2^2 . \end{aligned}$$

Combining the two inequalities gives the desired result. □

## B.6 PROOF OF LEMMA 5.2

The composite gradient steps at $x$ and the optimal solution $x^*$ satisfy

$$\begin{aligned} g(x) &\in \nabla\phi_{\text{sm}}(x) + \partial\phi_{\text{ns}}(x - g(x)), \\ g(x^*) &\in \nabla\phi_{\text{sm}}(x^*) + \partial\phi_{\text{ns}}(x^*). \end{aligned}$$

We subtract these two expressions and rearrange to obtain

$$\partial\phi_{\text{ns}}(x - g(x)) - \partial\phi_{\text{ns}}(x^*) \ni g(x) - (\nabla\phi_{\text{sm}}(x) - \nabla\phi_{\text{sm}}(x^*)).$$

Since $\phi_{\text{ns}}$ is convex, $\partial\phi_{\text{ns}}$ is monotone and

$$\begin{aligned} 0 &\le (x - g(x) - x^*)^T \partial\phi_{\text{ns}}(x - g(x)) \\ &= -g(x)^T g(x) + (x - x^*)^T g(x) + g(x)^T(\nabla\phi_{\text{sm}}(x) - \nabla\phi_{\text{sm}}(x^*)) \\ &\quad + (x - x^*)^T(\nabla\phi_{\text{sm}}(x) - \nabla\phi_{\text{sm}}(x^*)). \end{aligned}$$

We drop the last term because it is nonnegative ($\nabla\phi_{\text{sm}}$ is monotone) to obtain

$$\begin{aligned} 0 &\le -\|g(x)\|_2^2 + (x - x^*)^T g(x) + g(x)^T(\nabla\phi_{\text{sm}}(x) - \nabla\phi_{\text{sm}}(x^*)) \\ &\le -\|g(x)\|_2^2 + \|g(x)\|_2 \left(\|x - x^*\|_2 + \|\nabla\phi_{\text{sm}}(x) - \nabla\phi_{\text{sm}}(x^*)\|_2\right), \end{aligned}$$

so that

$$\|g(x)\|_2 \le \|x - x^*\|_2 + \|\nabla\phi_{\text{sm}}(x) - \nabla\phi_{\text{sm}}(x^*)\|_2 . \tag{B.5}$$

Since $\nabla\phi_{\text{sm}}$ is Lipschitz continuous, we conclude

$$\|g(x)\|_2 \le (\mu_u + 1)\,\|x - x^*\|_2 .$$

## B.7 PROOF OF LEMMA 5.3

The composite gradient step on $\phi$ has the form

$$g_\alpha(x) = \frac{1}{\alpha}(x - \operatorname{prox}_{\alpha\phi_{\text{ns}}}(x - \alpha\nabla\phi_{\text{sm}}(x))).$$

By Moreau's decomposition,

$$g_\alpha(x) = \nabla\phi_{\text{sm}}(x) + \frac{1}{\alpha}\operatorname{prox}_{[\alpha\cdot\phi_{\text{ns}}]^*}(x - \alpha\nabla\phi_{\text{sm}}(x)).$$

Thus $g_\alpha(x) - g_\alpha(y)$ has the form

$$\begin{aligned} g_\alpha(x) - g_\alpha(y) = \nabla\phi_{\text{sm}}(x) - \nabla\phi_{\text{sm}}(y) &+ \frac{1}{\alpha}\operatorname{prox}_{[\alpha\cdot\phi_{\text{ns}}]^*}(x - \alpha\nabla\phi_{\text{sm}}(x)) \\ &- \frac{1}{\alpha}\operatorname{prox}_{[\alpha\cdot\phi_{\text{ns}}]^*}(y - \alpha\nabla\phi_{\text{sm}}(y)). \end{aligned}$$

Let $w = \operatorname{prox}_{[\alpha\cdot\phi_{\text{ns}}]^*}(x - \alpha\nabla\phi_{\text{sm}}(x)) - \operatorname{prox}_{[\alpha\cdot\phi_{\text{ns}}]^*}(y - \alpha\nabla\phi_{\text{sm}}(y))$ and

$$\begin{aligned} d &= x - \alpha\nabla\phi_{\text{sm}}(x) - (y - \alpha\nabla\phi_{\text{sm}}(y)) \\ &= (x - y) - \alpha(\nabla\phi_{\text{sm}}(x) - \nabla\phi_{\text{sm}}(y)). \end{aligned}$$

We express (B.6) in terms of $W = \frac{ww^T}{w^T d}$ to obtain

$$\begin{aligned} g_\alpha(x) - g_\alpha(y) &= \nabla\phi_{\mathrm{sm}}(x) - \nabla\phi_{\mathrm{sm}}(y) + \frac{w}{\alpha} \\ &= \nabla\phi_{\mathrm{sm}}(x) - \nabla\phi_{\mathrm{sm}}(y) + \frac{1}{\alpha}Wd. \end{aligned}$$

We multiply by $x - y$ to obtain

$$\begin{aligned} &(x-y)^T(g_\alpha(x) - g_\alpha(y)) \\ &\quad = (x-y)^T(\nabla\phi_{\mathrm{sm}}(x) - \nabla\phi_{\mathrm{sm}}(y)) + \frac{1}{\alpha}(x-y)^T Wd \\ &\quad = (x-y)^T(\nabla\phi_{\mathrm{sm}}(x) - \nabla\phi_{\mathrm{sm}}(y)) + \frac{1}{\alpha}(x-y)^T W(x-y) \\ &\qquad - (\nabla\phi_{\mathrm{sm}}(x) - \nabla\phi_{\mathrm{sm}}(y)). \end{aligned} \tag{B.6}$$

Let $H(t) = \nabla^2 g(x + t(x-y))$. By the mean value theorem, we have

$$\begin{aligned} &(x-y)^T(g_\alpha(x) - g_\alpha(y)) \\ &\quad = \int_0^1 (x-y)^T\Big(H(t) - WH(t) + \frac{1}{\alpha}W\Big)(x-y)\; dt \\ &\quad = \int_0^1 (x-y)^T\Big(H(t) - \frac{1}{2}(WH(t) + H(t)W) + \frac{1}{\alpha}W\Big)(x-y)\; dt. \end{aligned} \tag{B.7}$$

To show the strong monotonicity of $g_\alpha$, it suffices to show that $H(t) + \frac{1}{\alpha}W - \frac{1}{2}(WH(t) + H(t)W)$ is positive definite for $\alpha \leq \frac{1}{\mu_u}$. We rearrange

$$(\sqrt{\alpha}H(t) - \frac{1}{\sqrt{\alpha}}W)(\sqrt{\alpha}H(t) - \frac{1}{\sqrt{\alpha}}W) \succeq 0$$

to obtain

$$tH(t)^2 + \frac{1}{\alpha}W^2 \succeq WH(t) + H(t)W.$$

Combining this expression with (B.7), we obtain

$$\begin{aligned} &(x-y)^T(g_\alpha(x) - g_\alpha(y)) \\ &\quad \geq \int_0^1 (x-y)^T\Big(H(t) - \frac{\alpha}{2}H(t)^2 + \frac{1}{\alpha}(W - \frac{1}{2}W^2)\Big)(x-y)\; dt. \end{aligned}$$

Since $\mathrm{prox}_{[\alpha\cdot\phi_{\mathrm{ns}}]^*}$ is firmly non-expansive, we have $\|w\|^2 \leq d^T w$ and

$$W = \frac{ww^T}{w^T d} = \frac{\|w\|_2^2}{w^T d}\frac{ww^T}{\|w\|_2^2} \preceq I.$$

Since $0 \preceq W \preceq I$, $W - W^2$ is also positive semidefinite and

$$(x-y)^T(g_\alpha(x) - g_\alpha(y)) \geq \int_0^1 (x-y)^T \Big( H(t) - \frac{\alpha}{2} H(t)^2 \Big) (x-y) \, dt.$$

If we set $\alpha \leq \frac{1}{\mu_u}$, the eigenvalues of $H(t) - \frac{\alpha}{2} H(t)^2$ are

$$\lambda_i(t) - \frac{\alpha}{2}\lambda_i(t)^2 \geq \lambda_i(t) - \frac{\lambda_i(t)^2}{2\mu_u} \geq \frac{\lambda_i(t)}{2} > \frac{\mu_l}{2},$$

where $\{\lambda_i(t)\}_{i\in[n]}$ are the eigenvalues of $H(t)$. We conclude

$$(x-y)^T(g_\alpha(x) - g_\alpha(y)) \geq \frac{\mu_l}{2} \|x-y\|_2^2 .$$

# C PROOFS OF LEMMAS IN CHAPTER 6

## C.1 PROOF OF LEMMA 6.2

Let $z_i = \Sigma^{-\frac{1}{2}} x_i$. The generalized coherence between $X$ and $\Sigma^{-1}$ is given by

$$|||\Sigma^{-1}\hat{\Sigma} - I|||_\infty = |||\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-\frac{1}{2}} z_i)(\Sigma^{\frac{1}{2}} z_i)^T - I|||_\infty.$$

Each entry of $\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-\frac{1}{2}} z_i)(\Sigma^{\frac{1}{2}} z_i)^T - I$ is a sum of independent subexponential random variables. Their subexponential norms are bounded by

$$\|(\Sigma^{-\frac{1}{2}} z_i)_j(\Sigma^{\frac{1}{2}} z_i)_k - \delta_{j,k}\|_{\psi_1} \leq 2\|(\Sigma^{-\frac{1}{2}} z_i)_j(\Sigma^{\frac{1}{2}} z_i)_k\|_{\psi_1}.$$

Recall for any two subgaussian random variables $X, Y$, we have

$$\|XY\|_{\psi_1} \leq 2\,\|X\|_{\psi_2}\,\|Y\|_{\psi_2}.$$

Thus

$$\|(\Sigma^{-\frac{1}{2}} z_i)_j(\Sigma^{\frac{1}{2}} z_i)_k - \delta_{j,k}\|_{\psi_1} \leq 4\|(\Sigma^{-\frac{1}{2}} z_i)_j\|_{\psi_2}\|(\Sigma^{\frac{1}{2}} z_i)_k\|_{\psi_2} \leq 4\sqrt{\kappa}\sigma_x^2,$$

where $\sigma_x = \|z_i\|_{\psi_2}$. By a Bernstein-type inequality,

$$\mathbf{Pr}\Big(\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-\frac{1}{2}} z_i)_j(\Sigma^{\frac{1}{2}} z_i)_k - \delta_{j,k} \geq t\Big) \leq 2e^{-c_1\min\{\frac{nt^2}{\tilde{\sigma}_x^4},\frac{nt}{\tilde{\sigma}_x^2}\}},$$

where $c_1 > 0$ is a universal constant and $\tilde{\sigma}_x^2 := 4\sqrt{\kappa}\sigma_x^2$. Since $\tilde{\sigma}_x^4 n > \log p$, we set $t = \frac{2\tilde{\sigma}_x^2}{\sqrt{c_1}}\big(\frac{\log p}{n}\big)^{\frac{1}{2}}$ to obtain

$$\mathbf{Pr}\Big(\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-\frac{1}{2}} z_i)_j(\Sigma^{\frac{1}{2}} z_i)_k - \delta_{j,k} \geq \frac{2\tilde{\sigma}_x^2}{\sqrt{c_1}}\Big(\frac{\log p}{n}\Big)^{\frac{1}{2}}\Big) \leq 2p^{-4}.$$

We obtain the stated result by taking a union bound over the $p^2$ entries of $\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-\frac{1}{2}} z_i)(\Sigma^{\frac{1}{2}} z_i)^T - I$.

## C.2 PROOF OF LEMMA 6.5

By Vershynin (2010), Proposition 5.10,

$$\mathbf{Pr}\Big(\frac{1}{n}|x_j^T\epsilon| > t\Big) \le e\exp\Big(-\frac{c_2n^2t^2}{\sigma_y^2\|x_j^T\|_2^2}\Big) \le e\exp\Big(-\frac{c_2n^2t^2}{\sigma_y^2\max_{j\in[p]}\hat{\Sigma}_{j,j}}\Big).$$

We take a union bound over the $p$ components of $\frac{1}{n}X^T\epsilon$ to obtain

$$\mathbf{Pr}\Big(\frac{1}{n}\|X^T\epsilon\|_\infty > t\Big) \le e\exp\Big(-\frac{c_2n^2t^2}{\sigma_y^2\max_{j\in[p]}\hat{\Sigma}_{j,j}} + \log p\Big).$$

We set $\lambda = \max_{j\in[p]}\hat{\Sigma}_{j,j}^{\frac{1}{2}}\sigma_y\big(\frac{3\log p}{c_2n}\big)^{\frac{1}{2}}$ to obtain the desired conclusion.

## C.3 PROOF OF LEMMA 6.7

We start by substituting in the linear model into (6.1):

$$\hat{\beta}^d = \hat{\beta} + \frac{1}{n}\hat{\Theta}X^T(y - X\hat{\beta}) = \beta^* + M\hat{\Sigma}(\beta^* - \hat{\beta}) + \frac{1}{n}MX^T\epsilon.$$

By adding and subtracting $\hat{\Delta} = \beta^* - \hat{\beta}$, we obtain

$$\hat{\beta}^d = \beta^* + \frac{1}{n}\hat{\Theta}X^T(y - X\hat{\beta}) = \beta^* + (M\hat{\Sigma} - I)(\beta^* - \hat{\beta}) + \frac{1}{n}MX^T\epsilon.$$

We obtain the expression of $\hat{\beta}^d$ by setting $\hat{\Delta} = (M\hat{\Sigma} - I)(\beta^* - \hat{\beta})$.

To show $\|\hat{\Delta}\|_\infty \le \frac{3\delta}{\mu}s\lambda$, we apply Hölder's inequality to each component of $\hat{\Delta}$ to obtain

$$|(M\hat{\Sigma} - I)(\beta^* - \hat{\beta})| \le \max_j \|\hat{\Sigma}m_j^T - e_j\|_\infty\|\hat{\beta} - \beta^*\|_1 \le \delta\|\hat{\beta} - \beta^*\|_1, \quad \text{(C.1)}$$

where $\delta$ is the generalized incoherence between $X$ and $M$. By Lemma 6.6, $\|\hat{\beta} - \beta^*\|_1 \le \frac{3}{\mu}s\lambda$. We combine the bound on $\|\hat{\beta} - \beta^*\|_1$ with (C.1) to obtain the stated bound on $\|\hat{\Delta}\|_\infty$.

## C.4 PROOF OF LEMMA 6.9

By Lemma 6.7,

$$\bar{\beta} - \beta^\star = \frac{1}{N}\sum_{k=1}^m \hat{\Theta}_k X_k^T\epsilon_k + \frac{1}{m}\sum_{k=1}^m \hat{\Delta}_k.$$

We take norms to obtain

$$\|\bar{\beta} - \beta^*\|_\infty \leq \Big\| \frac{1}{N} \sum_{k=1}^{m} \hat{\Theta}_k X_k^T \epsilon_k \Big\|_\infty + \frac{1}{m} \sum_{k=1}^{m} \|\hat{\Delta}_k\|_\infty.$$

We focus on bounding the first term. Let $a_j^T := e_j^T \begin{bmatrix} \hat{\Theta}_1 X_1^T & \dots & \hat{\Theta}_m X_m^T \end{bmatrix}$. By Vershynin (2010), Proposition 5.10,

$$\mathbf{Pr}\Big( \big| \frac{1}{N} a_j^T \epsilon \big| > t \Big) \leq e \exp\Big( -\frac{c_2 N^2 t^2}{\|a_j\|_2^2 \sigma_y^2} \Big)$$

for some universal constant $c_2 > 0$. Further,

$$\|a_j\|_2^2 = \sum_{k=1}^{m} \|X_k \hat{\Theta}_k^T e_j\|_2^2 = n \sum_{k=1}^{m} \big( \hat{\Theta}_k \hat{\Sigma}_k \hat{\Theta}_k^T \big)_{j,j} \leq c_\Omega N,$$

where $c_\Omega := \max_{j \in [p], k \in [m]} \big( \hat{\Theta}_k \hat{\Sigma}_k \hat{\Theta}_k^T \big)_{j,j}$. By a union bound over $j \in [p]$,

$$\mathbf{Pr}\Big( \max_{j \in [p]} \big| \frac{1}{N} a_j^T \epsilon \big| > t \Big) \leq e \exp\Big( -\frac{c_2 N t^2}{c_\Omega \sigma_y^2} + \log p \Big).$$

We set $t = \sigma_y \big( \frac{2 c_\Omega \log p}{c_2 N} \big)^{\frac{1}{2}}$ to deduce

$$\mathbf{Pr}\Big( \max_{j \in [p]} \big| \tfrac{1}{N} a_j^T \epsilon \big| \geq \sigma_y \big( \tfrac{2 c_\Omega \log p}{c_2 N} \big)^{\frac{1}{2}} \Big) \leq e p^{-1}.$$

We turn our attention to bounding the second term. By Lemma 6.5 and a union bound over $j \in [p]$, when we set

$$\lambda_1 = \dots = \lambda_m = \lambda := \max_{j \in [p], k \in [m]} \big( (\hat{\Sigma}_k)_{j,j} \big)^{\frac{1}{2}} \sigma_y \Big( \frac{3 \log p}{c_2 n} \Big)^{\frac{1}{2}},$$

we have $\frac{1}{n} \|X_k^T \epsilon\|_\infty \leq \lambda$ for any $k \in [m]$ with probability at least $1 - \frac{em}{p^2} \geq 1 - ep^{-1}$. By Lemma 6.7, when

1. $\{\hat{\Sigma}_k\}_{k \in [m]}$ satisfy the RE condition on $\mathcal{C}^*$ with constant $\mu_l$,
2. $\{(\hat{\Sigma}_k, \hat{\Theta}_k)\}_{k \in [m]}$ have generalized incoherence $c_{\mathrm{GC}} \big( \frac{\log p}{n} \big)^{\frac{1}{2}}$,

the second term is at most $\frac{3\sqrt{3}}{\sqrt{c_2}} \frac{c_{\mathrm{GC}} c_\Sigma}{\mu_l} \sigma_y \frac{s \log p}{n}$. We put the pieces together to obtain

$$\|\bar{\beta} - \beta^*\|_\infty \leq \sigma_y \Big( \frac{2 c_\Omega \log p}{c_2 N} \Big)^{\frac{1}{2}} + \frac{3\sqrt{3}}{\sqrt{c_2}} \frac{c_{\mathrm{GC}} c_\Sigma}{\mu_l} \sigma_y \frac{s \log p}{n},$$

## C.5 PROOF OF LEMMA 6.11

We express

$$\Sigma_{j,\cdot}^{-1}\hat{\Sigma}\Sigma_{j,\cdot}^{-1} = \Sigma_{j,\cdot}^{-1}\hat{\Sigma}\Sigma_{j,\cdot}^{-1} - \Sigma_{j,j}^{-1} + \Sigma_{j,j}^{-1} = \frac{1}{n}\sum_{i=1}^{n}(x_i^T\Sigma_{\cdot,j})^2 - \Sigma_{j,j}^{-1} + \Sigma_{j,j}^{-1}.$$

Since the subgaussian norm of $z_i = \Sigma^{-\frac{1}{2}}x_i$ is $\sigma_x$, $x_i^T\Sigma_{\cdot,j}$ is also subgaussian with subgaussian norm bounded by

$$\|x_i^T\Sigma_{\cdot,j}\|_{\psi_2} \le \|\Sigma^{\frac{1}{2}}z_i\|_{\psi_2}\|\Sigma_{\cdot,j}\|_2 \le \sigma_x(\Sigma_{j,j})^{\frac{1}{2}}.$$

We recognize $\frac{1}{n}\sum_{i=1}^{n}(x_i^T\Sigma_{\cdot,j})^2 - \Sigma_{j,j}^{-1}$ as a sum of *i.i.d.* subexponential random variables with subexponential norm bounded by

$$\|(x_i^T\Sigma_{\cdot,j})^2 - \Sigma_{j,j}^{-1}\|_{\psi_1} \le 2\|(x_i^T\Sigma_{\cdot,j})^2\|_{\psi_1} \le 4\|x_i^T\Sigma_{\cdot,j}\|_{\psi_2}^2 \le 4\sigma_x^2\Sigma_{j,j}^{-1}.$$

By Vershynin (2010), Proposition 5.16, we have

$$\mathbf{Pr}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i^T\Sigma_{\cdot,j})^2 - \Sigma_{j,j}^{-1} > t\Big) \le 2e^{-c_1\min\{\frac{nt^2}{16\sigma_x^2(\Sigma_{j,j}^{-1})^2},\frac{nt}{4\sigma_x\Sigma_{j,j}^{-1}}\}}$$

for some absolute constant $c_1 > 0$. For $t = \Sigma_{j,j}^{-1}$, the bound simplifies to

$$\mathbf{Pr}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i^T\Sigma_{\cdot,j})^2 - \Sigma_{j,j}^{-1} > \Sigma_{j,j}^{-1}\Big) \le 2e^{-c_1\min\{\frac{n}{16\sigma_x^2},\frac{n}{4\sigma_x}\}}.$$

We take a union bound over $j \in [p]$ to obtain the stated result.

## C.6 PROOF OF LEMMA 6.12

We follow a similar argument as the proof of Lemma 6.11:

$$\hat{\Sigma}_{k;j,j} = \hat{\Sigma}_{j,j} = \hat{\Sigma}_{j,j} - \Sigma_{j,j} + \Sigma_{j,j} = \frac{1}{n}\sum_{i=1}^{n}x_{i,j}^2 - \Sigma_{j,j} + \Sigma_{j,j}.$$

Since the $z_i = \Sigma^{-\frac{1}{2}}x_i$ is subgaussian with subgaussian norm $\sigma_x$, $x_{i,j}$ is also subgaussian with subgaussian norm bounded by

$$\|x_{i,j}\|_{\psi_2} \le \|\Sigma_{j,\cdot}^{\frac{1}{2}}z_i\|_{\psi_2} \le \sigma_x(\Sigma_{j,j})^{\frac{1}{2}}.$$

We recognize $\hat{\Sigma}_{j,j} - \Sigma_{j,j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2 - \Sigma_{j,j}$ as a sum of *i.i.d.* subexponential random variables with subexponential norm bounded by

$$\|\hat{\Sigma}_{j,j} - \Sigma_{j,j}\|_{\psi_1} \le 2\|x_{i,j}^2\|_{\psi_1} \le 4\|x_{i,j}\|_{\psi_2}^2 \le 4\sigma_x^2\Sigma_{j,j}.$$

By Vershynin (2010), Proposition 5.16, we have

$$\mathbf{Pr}(\hat{\Sigma}_{j,j} - \Sigma_{j,j} > t) \le 2e^{-c_1 \min\{\frac{nt^2}{16\sigma_x^2\Sigma_{j,j}^2}, \frac{nt}{\sigma_x\Sigma_{j,j}}\}}$$

for some absolute constant $c_1 > 0$. For $t = \Sigma_{j,j}$, the bound simplifies to

$$\mathbf{Pr}(\hat{\Sigma}_{j,j} - \Sigma_{j,j} > \Sigma_{j,j}) \le 2e^{-c_1 \min\{\frac{n}{16\sigma_x^2}, \frac{n}{4\sigma_x}\}}.$$

We take a union bound over $j \in [p]$ to obtain the stated result.

## C.7 PROOF OF THEOREM 6.19

The sharper consistency result depends on a result by Javanmard and Montanari (2013b), which we combine with Lemma 6.16 and restate for completeness. Before stating the results, we define the $(\infty, l)$ norm of a point $x \in \mathbf{R}^p$ as

$$\|x\|_{(\infty,l)} := \max_{\mathcal{A} \subset [p], |\mathcal{A}| \ge l} \frac{\|x_{\mathcal{A}}\|_2}{\sqrt{l}}.$$

When $l = 1$, the $(\infty, l)$ norm of $x$ is its $\ell_\infty$ norm. When $l = p$, the $(\infty, l)$ norm is the $\ell_2$ norm (rescaled by $\frac{1}{\sqrt{p}}$). Thus the $(\infty, l)$ norm interpolates between the $\ell_2$ and $\ell_\infty$ norms. Javanmard and Montanari (2013b), Theorem 2.3 shows that the bias of the debiased lasso is of order $\frac{\sqrt{s_0}\log p}{n}$.

**Lemma C.1.** *Under the conditions of Theorem 6.17,*

$$\|\hat{\Delta}_k\|_{(\infty, c's_0)} \lesssim_P \frac{c\sqrt{s_0}\log p}{n} \text{ for any } k \in [m] \text{ for any } c' > 0,$$

*where $c$ is a constant that depends only on $c'$ and $\Sigma$.*

*Proof.* The result is essentially Javanmard and Montanari (2013b), Theorem 2.3 with $\hat{\Omega} = \hat{\Theta}$ given by (6.6). Lemma 6.16 shows that

$$\max_{j \in [p]} \|\hat{\Theta}_j - \Sigma_j^{-1}\|_1 \lesssim_P s_j\left(\frac{\log p}{n}\right)^{\frac{1}{2}},$$

Since $\frac{\max_{j\in[p]} s_j^2 \log p}{n} \sim o(1)$, $\hat{\Theta}$ satisfies the conditions of Javanmard and Montanari (2013b), Theorem 2.3:

$$\|\hat{\Delta}_k\|_{(\infty,c's_0)} \lesssim_P \frac{c\sqrt{s_0}\log p}{n} \text{ for any } k \in [m],$$

The bound is uniform in $k \in [m]$ by a union bound for suitable parameters $\lambda_k \sim \left(\frac{\log p}{n}\right)^{\frac{1}{2}}$. □

By Lemma C.1, the estimator (6.5) is consistent in the $(\infty, s_0)$ norm. The argument is similar to the proof of Theorem 6.17.

**Theorem C.2.** *Under the conditions of Theorem 6.17,*

$$\|\bar{\beta} - \beta^*\|_{(\infty,c's_0)} \sim O_P\Big(\Big(\frac{\log p}{N}\Big)^{\frac{1}{2}} + \frac{\sqrt{s_0}\log p}{n}\Big).$$

*Proof.* We start by substituting the linear model into (6.5):

$$\tilde{\beta} = \frac{1}{m}\sum_{k=1}^{m}\hat{\Delta}_k + \frac{1}{N}\hat{\Theta}X^T\epsilon.$$

Subtracting $\beta^*$ and taking norms, we obtain

$$\|\tilde{\beta} - \beta^*\|_{(\infty,c's_0)} \le \frac{1}{m}\sum_{k=1}^{m}\|\hat{\Delta}_k\|_{(\infty,c's_0)} + \Big\|\frac{1}{N}\hat{\Theta}X^T\epsilon\Big\|_{(\infty,c's_0)}. \tag{C.2}$$

By Lemma C.1, the first (bias) term is of order $\frac{c\sqrt{s_0}\log p}{n}$. We focus on showing the second (variance) term is of order $\left(\frac{\log p}{N}\right)^{\frac{1}{2}}$. Since the $(\infty, l)$ norm is non-increasing in $l$,

$$\Big\|\frac{1}{N}\hat{\Theta}X^T\epsilon\Big\|_{(\infty,c's_0)} \le \Big\|\frac{1}{N}\hat{\Theta}X^T\epsilon\Big\|_{\infty}.$$

By Vershynin (2010), Proposition 5.16 and Lemma 6.11, it is possible to show that

$$\Big\|\frac{1}{N}\hat{\Theta}X^T\epsilon\Big\|_{\infty} \sim O_P\Big(\Big(\frac{\log p}{N}\Big)^{\frac{1}{2}}\Big).$$

Thus the second term in (C.2) is of order $\left(\frac{\log p}{N}\right)^{\frac{1}{2}}$. We put all the pieces together to obtain the stated conclusion. □

Finally, we prove Theorem 6.19. Since $\tilde{\beta}^{ht} - \beta^*$ is $2s_0$-sparse,

$$\|\tilde{\beta}^{ht} - \beta^*\|_2^2 \lesssim s_0\|\tilde{\beta}^{ht} - \beta^*\|_{(\infty,c's_0)}^2$$

or, equivalently,

$$\|\tilde{\beta}^{ht} - \beta^*\|_2 \lesssim \sqrt{s_0}\|\tilde{\beta}^{ht} - \beta^*\|_{(\infty,c's_0)}.$$

By the triangle inequality,

$$\begin{aligned}\|\tilde{\beta}^{ht} - \beta^*\|_{(\infty,c's_0)} &\leq \|\tilde{\beta}^{ht} - \tilde{\beta}\|_{(\infty,c's_0)} + \|\tilde{\beta} - \beta^*\|_{(\infty,c's_0)} \\ &\leq 2\|\tilde{\beta} - \beta^*\|_{(\infty,c's_0)},\end{aligned}$$

where the second inequality is by the fact that thresholding at $t = |\tilde{\beta}|_{(c's_0)}$ minimizes $\|\beta - \beta^*\|_{(\infty,c's_0)}$ over $c's_0$-sparse points $\beta$. Thus

$$\|\tilde{\beta}^{ht} - \beta^*\|_2 \sim O_P\Big(\Big(\frac{s_0 \log p}{N}\Big)^{\frac{1}{2}} + \frac{s_0 \log p}{n}\Big).$$

The consistency of $\tilde{\beta}^{ht}$ in the $\ell_1$ norm follows by the fact that $\tilde{\beta}^{ht} - \beta^*$ is $2s_0$-sparse.

## C.8 PROOF OF LEMMA 6.20

First, we derive an inequality in terms of $s_0$ and $\hat{s}_0$.

**Lemma C.3.** *Suppose $\hat{\Sigma}$ satisfies the RE condition on $\mathcal{C}^*$ with constant $\mu_l$ and $\lambda \geq \frac{2}{n}\|X^T\epsilon\|_\infty$. Then, the empirical sparsity satisfies*

$$\hat{s}_0 \leq \frac{192\mu_u(\hat{s}_0)}{\mu_l} s_0. \tag{C.3}$$

*Proof.* Since $\|\hat{z}_{\hat{\mathcal{E}}}\|_2^2 = \lambda^2 \hat{s}_0$, the optimality conditions of the lasso give

$$\hat{s}_0 = \frac{1}{(\lambda n)^2}\|X_{\hat{\mathcal{E}}}^T(y - X\hat{\beta})\|_2^2. \tag{C.4}$$

We express the right side as

$$\|\frac{1}{n}X_{\hat{\mathcal{E}}}^T(y - X\hat{\beta})\|_2^2 \leq \|\frac{1}{n}X_{\hat{\mathcal{E}}}^T X(\hat{\beta} - \beta^*)\|_2^2 + \|\frac{1}{n}X_{\hat{\mathcal{E}}}^T\epsilon\|_2^2.$$

By Hastie et al. (2015), Chapter 11, Theorem 2, the first term is bounded by

$$\|\frac{1}{n}X_{\hat{\mathcal{E}}}^T X(\hat{\beta} - \beta^*)\|_2^2 \leq \frac{1}{n}\|X_{\hat{\mathcal{E}}}^T\|_2^2 \frac{1}{n}\|X(\hat{\beta} - \beta^*)\|_2^2 \leq \frac{12^2\mu_u(\hat{s}_0)}{\mu_l} s_0\lambda^2,$$

where

$$\mu_u(s_0) := \sup_{\eta\in\mathcal{K}(s_0)} \eta^T\hat{\Sigma}\eta : \mathcal{K}(s_0) = \{x \in \mathbf{S}^{p-1} : \|x\|_0 \le s_0\}$$

is the $s$-sparse (upper) eigenvalue. Since $\lambda$ is at least $\frac{2}{n}\|X^T\epsilon\|_\infty$,

$$\|\frac{1}{n}X_{\hat{\mathcal{E}}}^T\epsilon\|_2^2 \le \hat{s}_0\|\frac{1}{n}X_{\hat{\mathcal{E}}}^T\epsilon\|_\infty^2 \le \frac{\hat{s}_0\lambda}{4}.$$

We substitute the bounds into (C.4) to obtain $\hat{s}_0 \le \frac{12^2\mu_u(\hat{s}_0)}{\mu_l}s_0 + \frac{\hat{s}_0}{4}$. Simplifying gives the stated bound. □

Lemma C.3 gives a fixed-point inequality. To obtain a bound on $\hat{s}_0$, we plug in an upper bound for $\mu_u(\hat{s}_0)$ and solve the inequality for $\hat{s}_0$.

**Lemma C.4** (Loh and Wainwright (2012), Lemma 15). *Under (A1),*

$$\mathbf{Pr}\big(\sup_{\eta\in\mathcal{K}(s_0)} |\eta^T\hat{\Sigma}\eta - \eta^T\Sigma\eta| > t\big) \le 2e^{-c_1 n\min\{\frac{t^2}{\sigma_x^4},\frac{t}{\sigma_x^2}\}+s_0\log p}$$

*for some constant* $c_1 > 0$.

Let $\gamma \in \mathbf{R}^p$ be the top (largest) $s$-sparse eigenvector of $\Sigma : \lambda_{u,s_0}(\Sigma) = \gamma^T\Sigma\gamma$. We have

$$\begin{aligned}\mu_u(s) - \lambda_{u,s_0}(\Sigma) &= \sup_{\eta\in\mathcal{K}(s)} \eta^T\hat{\Sigma}\eta - \gamma^T\Sigma\gamma \\ &\le \gamma^T(\hat{\Sigma}-\Sigma)\gamma.\end{aligned}$$

Thus $\mathbf{Pr}\big(\mu_u(s) - \lambda_{u,s_0}(\Sigma) > t\big) \le \mathbf{Pr}\big(\sup_{\eta\in\mathcal{K}(s_0)} |\gamma^T(\hat{\Sigma}-\Sigma)\gamma| > t\big)$. As long as $n > s_0\log p$, setting $t = \frac{2}{\sqrt{c_1}}\sigma_x^2\big(\frac{s_0\log p}{n}\big)^{\frac{1}{2}}$ shows that

$$\mu_u(s) \le \lambda_{u,s_0}(\Sigma) + \frac{2}{\sqrt{c_1}}\sigma_x^2\Big(\frac{s_0\log p}{n}\Big)^{\frac{1}{2}} \tag{C.5}$$

with probability at least $1 - 2p^{-s_0}$.

To complete the proof of Lemma 6.20, we substitute (C.5) into (C.3) to obtain

$$\hat{s}_0 \le \frac{192}{\mu_l}\Big(\lambda_{u,s_0}(\Sigma) + \frac{2}{\sqrt{c_1}}\sigma_x^2\Big(\frac{s_0\log p}{n}\Big)^{\frac{1}{2}}\Big)s_0. \tag{C.6}$$

(C.6) is a quadratic inequality in $\sqrt{\hat{s}_0}$. Rearranging,

$$\hat{s}_0 - \frac{384\sigma_x^2}{\sqrt{c_1}\mu_l}\sqrt{s_0}\Big(\frac{s_0\log p}{n}\Big)^{\frac{1}{2}}\sqrt{\hat{s}_0} \le \frac{192\lambda_{u,s_0}(\Sigma)}{\mu_l}s_0.$$

When $n > s_0 \log p$, the left side is larger than $\hat{s}_0 - \frac{384\sigma_x^2}{\sqrt{c_1}\mu_l}\sigma_x^2\sqrt{s_0}\sqrt{\hat{s}_0}$. Thus

$$\hat{s}_0 - \frac{384\sigma_x^2}{\sqrt{c_1}\mu_l}\sqrt{s_0}\sqrt{\hat{s}_0} \leq \frac{192\lambda_{u,s_0}(\Sigma)}{\mu_l}s_0.$$

Completing the square,

$$\Big(\sqrt{\hat{s}_0} - \frac{384\sigma_x^2}{\sqrt{c_1}\mu_l}\sqrt{s_0}\Big)^2 \leq \Big(\frac{192\lambda_{u,s_0}(\Sigma)}{\mu_l} + \frac{192^2\sigma_x^4}{c_1\mu_l^2}\Big)s_0.$$

We take the square root and rearrange to obtain

$$\sqrt{\hat{s}_0} \leq \Big(\frac{96\sigma_x^2 + 192\lambda_{u,s_0}(\Sigma)}{\mu_l} + \frac{192^2\sigma_x^4}{c_1\mu_l^2}\Big)\sqrt{s_0}.$$

Squaring both sides gives

$$\hat{s}_0 \leq \Big(\frac{96\sigma_x^2 + 192\lambda_{u,s_0}(\Sigma)}{\mu_l} + \frac{192^2\sigma_x^4}{c_1\mu_l^2}\Big)^2 s_0.$$

We have $\lambda_{u,s_0}(\Sigma) \leq \lambda_{\max}(\Sigma)$, and by Lemma 6.4, as long as

$$n > \max\{4000\tilde{s}_0\sigma_x^2\log(\tfrac{60\sqrt{2}ep}{\tilde{s}}), 4000\sigma_x^4\log p\},$$

$\mu_l$ is at least $\frac{1}{2}\lambda_{\min}(\Sigma)$ with probability at least $1-2p^{-1}$. Thus

$$\hat{s}_0 \leq \Big(\frac{96\sigma_x^2 + 192\lambda_{\max}(\Sigma)}{\mu_l} + \frac{192^2\sigma_x^4}{c_1\mu_l^2}\Big)^2 s_0$$

with probability at least $1-2p^{-(s_0+1)}$.

# BIBLIOGRAPHY

A. Agarwal, S. Negahban, M. J. Wainwright, et al. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.

F. R. Bach. Consistency of trace norm minimization. *The Journal of Machine Learning Research*, 9:1019–1048, 2008.

R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *Information Theory, IEEE Transactions on*, 56(4):1982–2001, 2010.

S. Becker and M. Fadili. A quasi-Newton proximal splitting method. In *Adv. Neural Inf. Process. Syst. (NIPS)*, 2012.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*, 2011.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statis.*, 37(4):1705–1732, 2009.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. 1:169–194, 2007.

R. H. Byrd, J. Nocedal, and F. Oztoprak. An inexact successive quadratic approximation method for convex l-1 regularized optimization. *arXiv preprint arXiv:1309.3529*, 2013.

E. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21(2): 572–596, 2011.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

Y. Choi, J. Taylor, and R. Tibshirani. Selecting the number of principal components: estimation of the true rank of a noisy matrix. *arXiv preprint arXiv:1410.8260*, 2014.

O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *The Journal of Machine Learning Research*, 13(1):165–202, 2012.

R. Dembo, S. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM J. Num. Anal.*, 19(2):400–408, 1982.

D. Donoho and J. Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53, 2009.

D. L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *Automatic Control, IEEE Transactions on*, 57(3):592–606, 2012.

S. Eisenstat and H. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM J. Sci. Comput.*, 17(1):16–32, 1996.

M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE, 2001.

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *Internat. J. Systems Sci.*, 12 (8):989–1000, 1981.

E. Greenshtein, Y. Ritov, et al. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6): 971–988, 2004.

D. Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and its generalizations*. CRC Press, 2015.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

C. Hsieh, M. Sustik, I. Dhillon, and P. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. In *Adv. Neural Inf. Process. Syst. (NIPS)*, 2011.

D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *Information Theory, IEEE Transactions on*, 57(11):7221–7234, 2011.

L. Jacob, G. Obozinski, and J. Vert. Group lasso with overlap and graph lasso. In *Int. Conf. Mach. Learn. (ICML)*, pages 433–440. ACM, 2009.

A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 964–972, 2010.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*, 2013a.

A. Javanmard and A. Montanari. Nearly optimal sample size in hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1311.0274*, 2013b.

J. Jia and K. Rohe. Preconditioning to comply with the irrepresentable condition. *arXiv preprint arXiv:1208.5584*, 2012.

Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16(2):375, 2006.

V. Koltchinskii, K. Lounici, A. B. Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

J. D. Lee and J. E. Taylor. Exact post model selection inference for marginal screening. In *Advances in Neural Information Processing Systems*, pages 136–144, 2014.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference with application to the lasso. *arXiv preprint arXiv:1311.6238*, 2013.

J. D. Lee, Y. Sun, and M. A. Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3): 1420–1443, 2014. doi: 10.1137/130921428.

J. D. Lee, Y. Sun, Q. Liu, and J. E. Taylor. Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*, 2015a.

J. D. Lee, Y. Sun, and J. E. Taylor. On model selection consistency of regularized m-estimators. *Electron. J. Statist.*, 9:608–642, 2015b. doi: 10.1214/15-EJS1013. URL http://dx.doi.org/10.1214/15-EJS1013.

W. Li and W. Sun. Perturbation bounds of unitary and subunitary polar factors. *SIAM J. Mat. Anal. Appl.*, 23(4):1183–1193, 2002.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664, 06 2012. doi: 10.1214/12-AOS1018. URL http://dx.doi.org/10.1214/12-AOS1018.

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

R. Mcdonald, M. Mohri, N. Silberman, D. Walker, and G. S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, pages 1231–1239, 2009.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statis.*, 34(3):1436–1462, 2006.

S. Negahban and M. Wainwright. Simultaneous support recovery in high dimensions: benefits and perils of block $\ell_1/\ell_\infty$-regularization. 57(6): 3841–3863, 2011.

S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

S. Negahban, M. J. Wainwright, et al. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39 (2):1069–1097, 2011.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer, 2003.

Y. Nesterov. Gradient methods for minimizing composite functions. *CORE Discussion Paper*, 2007.

P. Olsen, F. Oztoprak, J. Nocedal, and S. Rennie. Newton-like methods for sparse inverse covariance estimation. In *Adv. Neural Inf. Process. Syst. (NIPS)*, 2012.

M. Patriksson. *Nonlinear Programming and Variational Inequality Problems: A Unified Approach.* Kluwer Academic Publishers, 1999.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, 2010.

B. Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

S. Reid and R. Tibshirani. Post selection point and interval estimation of signal sizes in gaussian samples. *arXiv preprint arXiv:1405.3340*, 2014.

A. Rohde, A. B. Tsybakov, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

J. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *arXiv preprint arXiv:1407.2724*, 2014.

M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on*, 59(6):3434–3447, 2013.

M. Schmidt. *Graphical Model Structure Learning with $\ell_1$-regularization*. PhD thesis, University of British Columbia, 2010.

M. Schmidt, E. Van Den Berg, M. Friedlander, and K. Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm. In *Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2009.

M. Schmidt, D. Kim, and S. Sra. Projected Newton-type methods in machine learning. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Adv. Neural Inf. Process. Syst. (NIPS)*, pages 1329–1336, 2004.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, 1956.

Y. Sun and J. E. Taylor. Exact inference for censored regression problems. *arXiv preprint arXiv:1403.3457*, 2014.

J. Taylor, R. Lockhart, R. J. Tibshirani, and R. Tibshirani. Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 2014.

X. Tian, J. R. Loftus, and J. E. Taylor. Selective inference with unknown variance via the square-root lasso. *arXiv preprint arXiv:1504.08031*, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, pages 267–288, 1996.

A. N. Tikhonov. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.

Q. Tran-Dinh, A. Kyrillidis, and V. Cevher. Composite self-concordant minimization. *Journal of Machine Learning Research*, 16:371–416, 2015.

J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *Information Theory, IEEE Transactions on*, 52(3): 1030–1051, 2006.

P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Math. Prog. Ser. B*, 117(1):387–423, 2009.

B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

S. van de Geer. Weakly decomposable regularization penalties and structured sparsity. *arXiv preprint arXiv:1204.4813*, 2012.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*, 2013.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (lasso). 55(5): 2183–2202, 2009.

S. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.*, 57(7):2479–2493, 2009.

G. Yuan, C. Ho, and C. Lin. An improved glmnet for $\ell_1$-regularized logistic regression. *J. Mach. Learn. Res.*, 13:1999–2030, 2012.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statis.*, 36(4):1567–1594, 2008.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.

M. Zinkevich, M. Weimer, L. Li, and A. J. Smola. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2010.

# COLOPHON

This thesis was typeset using `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX at:

http://code.google.com/p/classicthesis/

*Final Version* as of June 3, 2015 (`YuekaiThesis`).