THE UNIVERSITY OF CHICAGO

ALGORITHMIC AND STATISTICAL OPTIMALITY FOR HIGH-DIMENSIONAL DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

HAOYANG LIU

CHICAGO, ILLINOIS
AUGUST 2020

CONTENTS

ACKNOWLEDGMENTS

I would like to thank my advisors, Prof. Rina Foygel Barber and Prof. Chao Gao, for their constant support and encouragement throughout my graduate study, for showing me how to keep balance on the boundary of the current scope, and most importantly, for the privilege of intellectual freedom they give me as a PhD student. I would like to thank my committee member Prof. Mihai Anitescu for valuable discussions and career advices. I would like to thank Prof. Richard Samworth for his guidance and support throughout my graduate study. I would like to thank Prof. Emmanuel Candès for many stimulating discussions. I would like to thank all faculty and staff in the Department of Statistics where I have spent my wonderful five years of graduate school. I would like to thank my fellow graduate students in the department for the excellent academic environment. I would like to thank Minqi Liu for her love and support during my graduate study. I am thankful to Fan Yang, Fengshuo Zhang, Ran Dai, Changji Xu, Bumeng Zhuo, Wooseok Ha, Shihao Gu, Yao Tong for their care and support, and to many other friends at UChicago for their patience and help. Finally, I would like to thank my parents, Rongxia Hao and Hengliang Liu, for their love and support throughout my life.

ABSTRACT

For high-dimensional data, two of the most important questions are the question of algorithmic optimality, which asks for the optimal algorithm within a certain class of computationally feasible procedures, and the question of statistical optimality, which asks for the optimal statistical procedure under a generating model. In this thesis the question of algorithmic optimality is investigated for the class of iterative thresholding algorithms on sparse and low rank structures under the framework of restricted optimality. The question of statistical optimality is investigated for the high-dimensional sparse changepoint detection problem and the contaminated density estimation problem under the minimax framework.

CHAPTER 1

INTRODUCTION

Data are numerical information collected through different manners. In a statistical problem, we typically have repeated observations on a single object, and the goal is to distill different kind of knowledge about the underlying object through statistical analysis on the data, where the knowledge could for example be in the form of estimates of underlying parameters of a statistical model. High-dimensional data are observations made on complicated objects for which the number of underlying parameters is large compared to the amount of observations we have. And without any assumptions on the model or any prior knowledge on the nature of the underlying object, the question of distilling knowledge about a complicated objects from a limited amount of data is typically ill-posed. Therefore different kind of simplifying assumptions are made, within which sparsity is a dominating example. The sparsity assumption states that although the potential model class is big, the true model should be simple in its nature, involving only a small number of non-zero parameters. The sparsity assumption comes with its own challenges, both algorithmic and statistical, which we now expand on.

Since most statistical procedures work in searching for some parameter in a parameter space that best fits the data according to some standard, it is naturally formulated into an optimization problem. With the sparsity assumption, the parameter space becomes all vectors that has sparsity smaller than some value, and therefore the optimization problem becomes:

$$\min_{x\in\mathbb{R}^d, \|x\|_0\leq s} \mathsf{f}(x),$$

with some sparsity level s and some loss function f. The algorithmic difficulty then lies in the fact that the constraint $\|x\|_0 \leq s$ is non-convex. Classical results in optimization show that it is computationally hard to solve this problem globally (i.e. to find the global optimum) even within the class of quadratic loss functions (see [1]). However, such a negative statement is not enough for statistical applications for the following reasons. Firstly, the loss functions appearing

in statistical applications are typically well-conditioned when restricted to sparse parameters, and thus are rather well-behaved. Secondly, we are not aiming to solve the optimization problem globally, but are usually satisfied if the loss function value is low enough. Thirdly, instead of a qualitative statement, we care more about a quantitative characterization of the difficulty of the problem in statistical applications. We will provide partial answers to these challenges within the class of iterative thresholding algorithms in Chapter 2. Chapter 3 studies the relation between the constrained approach and the factorized approach for solving low-rank optimization problem.

The statistical challenge associated with sparsity lies in the task of determining the optimal (or near optimal) statistical procedures within all possible statistical procedures for a wide range of generating models where sparsity can play a role. Such a task of course depends on our standard for comparing different procedures, and we mainly work under the minimax framework in search for the statistical procedures with the optimal (or near optimal) worst-case performance. The optimal statistical performance is then used as a benchmark against which different practical statistical procedures are compared to. In the literature, the question of statistical optimality is investigated thoroughly for the sparse gaussian sequence model and the sparse linear model, where a wide range of phenomena are revealed that are believed to be universal for models with sparsity. Then, many other models are studied where sparsity either interacts with other structures of the underlying parameters, or interacts with different data generating mechanisms. For each of these special models, the interest is to discover new phenomena not covered in these universal ones. Chapter 4 contains such a study of statistical optimality under a model where the sparsity structure interacts with the changepoint structure. Chapter 5 studies statistical optimality under a contaminated density estimation model.

For notation, we use lowercase letters for vectors and uppercase letters for matrices. Notation specific to each chapter is introduced in the starting section of the chapter.

CHAPTER 2

BETWEEN HARD AND SOFT THRESHOLDING: OPTIMAL ITERATIVE THRESHOLDING ALGORITHMS

In this chapter, we consider the problem of analyzing the performance of iterative thresholding algorithms used on sparse constrained optimization problem, and provide a framework to compare the performance of different thresholding operators.[1]

2.1 Introduction

Consider the general problem of sparse optimization, where we seek to optimize a likelihood function or loss function subject to a sparsity constraint,

$$\min_{x\in\mathbb{R}^d, \|x\|_0\leq s} \mathsf{f}(x).$$

Here $\mathsf{f}:\mathbb{R}^d\to\mathbb{R}$ is the target function that we would like to minimize, while the constraint $\|x\|_0\leq s$ requires that the solution vector x has at most s many nonzero entries. Similarly, we may work with a matrix parameter $X\in\mathbb{R}^{n\times m}$ and search for a low-rank solution,

$$\min_{X\in\mathbb{R}^{n\times m}, \mathrm{rank}(X)\leq s} \mathsf{f}(X).$$

Optimization problems over a sparsity constraint or a rank constraint are ubiquitous in high-dimensional statistics and machine learning. Sparsity of a vector parameter x represents the idea that we can model the data using a small fraction of the available features, which, for instance, may correspond to covariates in a regression model or to basis expansion terms in a nonparametric function estimation problem. Similarly, a rank constraint on a matrix parameter X might correspond to an underlying factor model with a small number of factors. We will focus on problems

1. This work is joint with Rina Foygel Barber and published in [2].

where f is a differentiable function, as is often the case for many likelihood models and other loss functions.

In this chapter, we will study the iterative thresholding approach, where gradient steps that lower the value of the target function f are alternated with thresholding steps to enforce the sparsity constraint—for instance, *hard thresholding* sets all but the largest s entries to zero, while *soft thresholding* shrinks all values towards zero equally until the sparsity constraint is satisfied. (The same ideas apply to a rank constraint, by thresholding or shrinking singular values instead of vector entries. For simplicity, we will primarily discuss the sparse minimization problem, and will return to the low-rank problem later on.)

For sparse minimization of a differentiable target function $\mathsf{f}(x)$, many existing algorithms can be broadly described as iterating steps of the following form:

$$\begin{cases} \text{Gradient step: } x_t' = x_{t-1} - \eta_t \cdot \nabla \mathsf{f}(x_{t-1}) \text{ for some step size } \eta_t, \\ \text{Sparsity step: } x_t = \text{some sparse (or nearly sparse) approximation to } x_t'. \end{cases} \tag{2.1}$$

Our aim is to characterize the type of thresholding operators that are likely to be most successful at converging to a good solution, i.e. to a value of $\mathsf{f}(x)$ that is as low as possible. Is an iterative thresholding algorithm most likely to succeed if we use hard thresholding, soft thresholding, or yet another form of thresholding to enforce the sparsity constraint?

We find that the worst-case performance of a thresholding operator, relative to a broad class of target functions f that we may want to minimize, is fully characterized by a simple measure that we call the *relative concavity*. The relative concavity studies the behavior of the sparse thresholding map $x_t' \mapsto x_t$ in the iterative algorithm (2.1), viewed as an approximate projection onto the space of s-sparse vectors. Using relative concavity as a tool to evaluate and compare different thresholding operators, we find that commonly used thresholding operators, for example hard thresholding and soft thresholding, are indeed suboptimal. Instead, we characterize a general class of thresholding operators, lying between hard thresholding and soft thresholding, that we show to be optimal. This class includes ℓ_q norm thresholding, where $q \in (0,1)$ is chosen adaptively relative to the particular

problem; furthermore, choosing $q = 2/3$ is "universal" in the sense that it is nearly optimal across all sparse thresholding problems. We also develop the *reciprocal thresholding* operator, which enjoys the same optimality guarantees as ℓ_q thresholding, but with a closed-form equation for the iterative thresholding step. These simple and efficient iterative thresholding methods are then applied to the statistical setting of sparse linear regression problem:

$$y = X\theta_0 + z, \tag{2.2}$$

and are shown to match the Lasso in terms of the resulting guarantee on estimating the true mean vector $X\theta_0$.

2.2 Problem formulation

To rigorously formulate the problem, we need to specify the assumptions on the objective function, the class of algorithms under consideration, and the optimization guarantee of interest.

Restricted strong convexity and restricted smoothness In many problems in high-dimensional statistics, we aim to optimize loss functions that may be very poorly conditioned in general, but nonetheless exhibit convergence properties of a well-conditioned function when working only with sparse or approximately sparse vectors. This behavior is captured in the notions of restricted strong convexity and restricted smoothness (see e.g. Negahban et al. [3], Loh and Wainwright [4] for background). A differentiable function $\mathsf{f} : \mathbb{R}^d \to \mathbb{R}$ satisfies *restricted strong convexity* with parameter α at sparsity level s, abbreviated as (α, s)-RSC, if

$$\mathsf{f}(y) \geq \mathsf{f}(x) + \langle \nabla \mathsf{f}(x), y - x \rangle + \frac{\alpha}{2}\|x - y\|_2^2 \text{ for all } x, y \in \mathbb{R}^d \text{ with } \|x\|_0 \leq s, \|y\|_0 \leq s.$$

Similarly, f satisfies *restricted smoothness* with parameter β at sparsity level s, abbreviated as (β, s)-RSM, if

$$\mathsf{f}(y) \leq \mathsf{f}(x) + \langle \nabla \mathsf{f}(x), y - x \rangle + \frac{\beta}{2}\|x-y\|_2^2 \text{ for all } x, y \in \mathbb{R}^d \text{ with } \|x\|_0 \leq s, \|y\|_0 \leq s.$$

Our results will focus on $\kappa = \beta/\alpha$, the *condition number* of the function f (at the given sparsity level s).

Iterative thresholding algorithms After initializing at any point $x_0 \in \mathbb{R}^d$, an iterative thresholding algorithm proceeds by alternating between taking a gradient descent step, and applying a thresholding operator:

$$x_t = \Psi_s\big(x_{t-1} - \eta_t \nabla \mathsf{f}(x_{t-1})\big), \tag{2.3}$$

where $\Psi_s : \mathbb{R}^d \to \{x \in \mathbb{R}^d : \|x\|_0 \leq s\}$ is some thresholding operator that enforces s-sparsity at each step. For example, the "hard thresholding" operator Ψ_s^{HT}, is defined to be the operator that truncates any vector $z \in \mathbb{R}^d$ to its s largest entries, i.e.

$$\big(\Psi_s^{\mathrm{HT}}(z)\big)_i = \begin{cases} z_i, & i \in S, \\ 0, & i \notin S, \end{cases}$$

where $S \subset \{1, \ldots, d\}$ indexes the s largest-magnitude entries of z.[2]

Step size choice Throughout this chapter, we will primarily study this generalized iterative thresholding algorithm under the choice of a universal fixed step size $\eta = 1/\beta$, where β is the restricted smoothness parameter for the function f. When β is unknown, we will also consider the

2. To be fully precise, in the case of a tie between different entries of z, we may need to choose which entries to keep and which to set to zero. This choice will not matter from the point of view of our theoretical analysis, and from this point on, we will assume that we have fixed some map $z \mapsto S$, mapping each vector $z \in \mathbb{R}^d$ to a set $S \subset \{1, \ldots, d\}$ corresponding to the indices of the s largest entries, so that $|S| = s$ and $\min_{i \in S} |z_i| \geq \max_{j \notin S} |z_j|$, for every z. For instance, in the case of a tie between z_i and z_j for the position of the sth largest-magnitude entry, we might follow the rule that we choose to keep entry i if $i < j$ and to keep entry j otherwise. Since the exact choice of the rule for breaking ties is not relevant for our results here, we will implicitly assume it to be fixed for the remainder of this chapter.

following adaptive choice of step size based on exact line search:

$$\begin{cases} \text{Define } \widetilde{x}_t(\eta) = \Psi_s\big(x_{t-1} - \eta \nabla \mathsf{f}(x_{t-1})\big), \\ \text{Choose } \eta_t = \max\left\{\eta \geq 0 : \mathsf{f}(\widetilde{x}_t(\eta)) \leq \mathsf{f}(x_{t-1}) + \langle \widetilde{x}_t(\eta) - x_{t-1}, \nabla \mathsf{f}(x_{t-1})\rangle + \frac{1}{2\eta}\|\widetilde{x}_t(\eta) - x_{t-1}\|_2^2\right\}, \\ \text{Set } x_t = \widetilde{x}_t(\eta_t). \end{cases} \tag{2.4}$$

Note that, since x_{t-1} and $\widetilde{x}_t(\eta)$ are both s-sparse, the curvature condition

$$\mathsf{f}(\widetilde{x}_t(\eta)) \leq \mathsf{f}(x_{t-1}) + \langle \widetilde{x}_t(\eta) - x_{t-1}, \nabla \mathsf{f}(x_{t-1})\rangle + \frac{1}{2\eta}\|\widetilde{x}_t(\eta) - x_{t-1}\|_2^2 \tag{2.5}$$

is necessarily satisfied for any $\eta \leq \frac{1}{\beta}$ due to the restricted smoothness property. Therefore we will always have $\eta_t \geq \frac{1}{\beta}$. Intuitively, the rule not only helps us get rid of the need to know β, but also allows the algorithm to take larger step size for more progress when possible. In practice, we would consider using a backtracking line search, that is, starting from a large step size and iteratively shrinking it until condition (2.5) is satisfied. In this way, condition (2.5) is similar to the classical Armijo rule for backtracking line search. For simplicity of our theoretical result we do not treat inexact linesearch in the following.

Restricted optimality It is well known that, due to the nonconvexity of the sparsity constraint $\|x\|_0 \leq s$, computationally feasible algorithms cannot be guaranteed to find the global minimum, $\min_{\|x\|_0 \leq s} \mathsf{f}(x)$—at least, not without strong assumptions. In other words, it may be the case that $\lim_{t\to\infty} \mathsf{f}(x_t)$ is strictly larger than $\min_{\|x\|_0 \leq s} \mathsf{f}(x)$. However, Jain et al. [5]'s analysis of the iterative hard thresholding algorithm proves that IHT achieves a weaker optimization guarantee, converging to a loss value that is at least as small as the best value attained under a more restricted constraint $\|x\|_0 \leq s'$ where $s' < s$. In this work, we will refer to this type of result as a *restricted optimality* guarantee, where the output of an s-sparse optimization algorithm is guaranteed to perform well relative to a more restrictive s'-sparsity constraint, for some $s' < s$. In particular, we will be interested in the sparsity ratio s'/s—the ratio between the sparsity level s used in the algorithm, versus

the level s' appearing in the guarantee. We will assess a thresholding operator Ψ_s based on its ability to guarantee restricted optimality relative to a sparsity level s' that is as close to s as possible, i.e. a sparsity ratio $\rho = s'/s$ that is as close to 1 as possible.

Related works on iterative thresholding algorithms There exists a vast literature on the properties of iterative thresholding algorithms, especially iterative hard thresholding, regarding the optimization properties and statistical guarantees of these algorithms. Recent results in this area include the work of Blumensath and Davies [6], Jain et al. [5], Chen and Wainwright [7], Bhatia et al. [8], Jain et al. [9], Cai et al. [10], Kyrillidis and Cevher [11]. Accelerated forms of the iterative hard thresholding algorithm are studied in Kyrillidis and Cevher [12], Blumensath [13], Khanna and Kyrillidis [14]. In particular, Khanna and Kyrillidis [14] finds substantial theoretical and empirical improvement over the original non-accelerated version of the algorithm. Nguyen et al. [15] studies iterative hard thresholding in the context of stochastic gradient descent, where at each step t we only have access to a noisy vector that approximates the true current gradient, $\nabla \mathsf{f}(x_t)$. The works mentioned here also consider thresholding algorithms for the low-rank setting, truncating singular values instead of vector entries. More broadly, Nguyen et al. [15]'s work considers approximate thresholding procedures and more general definitions of sparsity.

Related works on penalized methods for sparsity The sparse optimization problem can alternately be approximated by a penalized minimization problem,

$$\min_{x \in \mathbb{R}^d} \left\{ \mathsf{f}(x) + \lambda \mathsf{R}(x) \right\},$$

where $\mathsf{R}(x)$ is a sparsity-promoting regularizer, and λ is tuning parameters controlling the penalization or constraint. Of course, choosing $\mathsf{R}(x) = \|x\|_0$ would reduce to the original target optimization problem, but these minimizations are generally only feasible to solve if $\mathsf{R}(x)$ is some relaxation of the sparsity constraint/penalty. For example, the Lasso [16] uses a convex regularizer, $\mathsf{R}(x) = \|x\|_1$, which enjoys many strong guarantees of accurate estimation of the true sparse signal

x and of its support. More recently, many nonconvex penalties have been proposed that reduce the shrinkage bias of the Lasso, at the cost of a more challenging optimization problem, such as the SCAD [17] and MCP [18] penalties. The ℓ_q norm, for $q \in (0,1)$, has also been extensively studied as a compromise between the convex but biased ℓ_1 norm (as in the Lasso), and the theoretically optimal but computationally infeasible ℓ_0 norm (i.e. the sparsity constraint, $\|x\|_0 \leq s$). Results for the ℓ_q norm include work by Chartrand [19], Foucart and Lai [20], Kabashima et al. [21], Lai and Wang [22]. Zheng et al. [23]'s recent work studies the ℓ_q norm using the framework of approximate message passing to characterize its superior performance relative to the convex ℓ_1 norm. While the resulting optimization problem is nonconvex for these alternatives to the ℓ_1 norm, Loh and Wainwright [4] show that restricted strong convexity in the objective function f is sufficient to outweigh bounded concavity in the penalty, to ensure successful optimization within a small error tolerance.

2.3 Convergence guarantee for iterative thresholding algorithms

To state our main convergence result, we have to first define the relative concavity of a thresholding operator.

Relative concavity of a thresholding operator Let $s \in \{1,\ldots,d\}$ be any fixed sparsity level and let $\rho \in [0,1]$. We define the *relative concavity* of an s-sparse thresholding operator Ψ_s relative to sparsity proportion ρ as

$$\gamma_{s,\rho}(\Psi_s) = \sup\left\{ \frac{\langle y-\Psi_s(z), z-\Psi_s(z)\rangle}{\|y-\Psi_s(z)\|_2^2} \;:\; y,z \in \mathbb{R}^d,\ \|y\|_0 \leq \rho s,\ y \neq \Psi_s(z) \right\}.$$

Note that $\frac{\langle y-\Psi_s(z), z-\Psi_s(z)\rangle}{\|y-\Psi_s(z)\|_2^2}$ is the coefficient of projection when projecting $z-\Psi_s(z)$ onto $y-\Psi_s(z)$, and measures how much these two vectors align. To understand the term "relative concavity" in the name, we note that if Ψ_s were a projection operator to some convex constraint set $\mathscr{C}$, then we would have $\langle y-\Psi_s(z), z-\Psi_s(z)\rangle \leq 0$ for any $y \in \mathscr{C}$, by the properties of convex

projections. For sparse estimation, the constraint $\|x\|_0 \leq s$ is not convex; any positive values of $\langle y - \Psi_s(z), z - \Psi_s(z)\rangle$ with $\|y\|_0 \leq s$ measure the extent to which the thresholding operator Ψ_s behaves *differently* from a convex projection. By taking a more restrictive constraint on y, namely $\|y\|_0 \leq \rho s$ rather than $\|y\|_0 \leq s$, we reduce this measure of concavity; the relative concavity of Ψ_s will be smaller for lower values of ρ.

Relative concavity and convergence We now examine how the relative concavity of Ψ_s relates to the convergence behavior of iterative thresholding with a fixed step size. The main message, casted informally, is this:

> Given sparsity levels s and $s' = \rho s$, and an s-sparse thresholding operator Ψ_s, the condition $\gamma_{s,\rho}(\Psi_s) \leq \frac{1}{2\kappa}$ is both necessary and sufficient for restricted optimality to hold relative to sparsity level s'.

Our first theorem accounts for the sufficiency of the condition.

Theorem 2.3.1. *Consider any objective function* $\mathsf{f} : \mathbb{R}^d \to \mathbb{R}$*, any sparsity levels* $s \geq s'$*, and any* s*-sparse thresholding operator* Ψ_s*. Assume the objective function* f *satisfies* (α, s)*-RSC and* (β, s)*-RSM. Let* $\rho = s'/s$ *and* $\kappa = \beta/\alpha$*, and assume that*

$$\gamma_{s,\rho}(\Psi_s) < \frac{1}{2\kappa}.$$

Then, for any s*-sparse* $x_0 \in \mathbb{R}^d$ *and any* s'*-sparse* $y \in \mathbb{R}^d$*, the iterated thresholding algorithm* (2.3) *initialized at* x_0 *and run with fixed step size* $\eta = 1/\beta$ *satisfies*

$$\min_{t=1,\ldots,T} \mathsf{f}(x_t) \leq \mathsf{f}(y) + \left(\frac{1 - 1/\kappa}{1 - 2\gamma_{s,\rho}(\Psi_s)}\right)^T \cdot \frac{\beta}{2}\|x_0 - y\|_2^2$$

for each $T \geq 1$*. The same result holds for the iterative thresholding algorithm with adaptive step size* (2.4)*.*

In other words, the condition $\gamma_{s,\rho}(\Psi_s) < \frac{1}{2\kappa}$ guarantees restricted optimality on the class of κ-

conditioned objective functions at sparsity proportion ρ. Next, we examine the necessity of the bound on $\gamma_{s,\rho}(\Psi_s)$. The following result proves that, if $\gamma_{s,\rho}(\Psi_s) > \frac{1}{2\kappa}$, then there exists an objective function $\mathsf{f}(x)$ on which the restricted optimality guarantee fails, when we run iterative thresholding with fixed step size $\eta = \frac{1}{\beta}$.

Theorem 2.3.2. *Consider any sparsity levels $s \geq s'$, any s-sparse thresholding operator Ψ_s, and any constants $\beta \geq \alpha > 0$. Let $\rho = s'/s$ and $\kappa = \beta/\alpha$, and assume that*

$$\gamma_{s,\rho}(\Psi_s) > \frac{1}{2\kappa}.$$

Then there exists an objective function $\mathsf{f}(x)$ that satisfies (α, s)-RSC and (β, s)-RSM, and an s-sparse $x_0 \in \mathbb{R}^d$ and s'-sparse $y \in \mathbb{R}^d$, such that the iterated thresholding algorithm (2.3) *run with step size $\eta = 1/\beta$ and initialization point x_0 satisfies*

$$\lim_{t \to \infty} \mathsf{f}(x_t) > \mathsf{f}(y).$$

This result is proved by constructing an objective function f and an s-sparse point x_0, such that $\mathsf{f}(x_0) > \mathsf{f}(y)$, but x_0 is a stationary point of the iterated thresholding algorithm, i.e. by initializing at x_0, we obtain $x_t = x_0$ for all $t \geq 1$. This proves that the iterated thresholding algorithm does not satisfy restricted optimality (at the given sparsity levels), since it is trapped at an s-sparse point x_0 whose objective value is strictly worse than that of the s'-sparse point y.

2.4 Upper and lower bounds on relative concavity

We have now seen that the relative concavity $\gamma_{s,\rho}(\Psi_s)$ fully characterizes the worst-case performance of the thresholding operator Ψ_s in the gradient descent algorithm, with a convergence guarantee in Theorem 2.3.1 and a matching lower bound in Theorem 2.3.2 (assuming a fixed step size). In this next section, we turn to the question of investigating the relative concavity in greater detail, in order to determine which thresholding operators are most likely to lead to successful optimiza-

tion. Along the way, we will focus on the following questions:

- What is the relative concavity of commonly used thresholding operators, for example, hard thresholding and soft thresholding?
- What is the best (i.e. lowest) possible relative concavity $\gamma_{s,\rho}(\Psi_s)$ among all thresholding operators Ψ_s, and which thresholding operators are optimal?

Throughout this section, for providing upper and lower bounds on $\gamma_{s,\rho}(\Psi_s)$, we will assume without comment that $s, s' \in \{1, \dots, d\}$ are two sparsity levels satisfying $1 \leq s' \leq s \leq d$ and $s + s' \leq d$, and we will define $\rho = s'/s$ as usual.

2.4.1 Relative concavity of hard and soft thresholding

First, we consider hard thresholding, $\Psi_s = \Psi_s^{\mathrm{HT}}$. The following result computes the relative concavity for the hard thresholding operator:

Lemma 2.4.1. *The relative concavity of hard thresholding is given by*

$$\gamma_{s,\rho}(\Psi_s^{\mathrm{HT}}) = \frac{\sqrt{\rho}}{2}$$

for every sparsity proportion $\rho \in (0, 1]$.

In particular, with Lemma 2.4.1, the condition $\gamma_{s,\rho}(\Psi_s^{\mathrm{HT}}) < \frac{1}{2\kappa}$ becomes $\rho < \frac{1}{\kappa^2}$. In light of Theorems 2.3.1 and 2.3.2, we see that for iterative hard thresholding algorithm, $\rho < \frac{1}{\kappa^2}$ is necessary and sufficient to guarantee restricted optimality with sparsity level s and s', tightening the condition obtained in Jain et al. [5] where they prove restricted optimality with the sparsity proportion $\rho = \frac{1}{32\kappa^2}$.

We might wonder whether the highly discontinuous nature of the hard thresholding function might not be ideal—by smoothing out the discontinuity, could we attain better performance? However, we find that any continuous thresholding operator with respect to the Euclidean distance in $\mathbb{R}^d$ is necessarily worse than hard thresholding:

Lemma 2.4.2. *For any continuous map $\Psi_s : \mathbb{R}^d \to \{x \in \mathbb{R}^d : \|x\|_0 \le s\}$, its relative concavity satisfies*

$$\gamma_{s,\rho}(\Psi_s) \ge 1$$

for every sparsity proportion $\rho \in (0,1]$.

In particular, since $\kappa \ge 1$, the condition $\gamma_{s,\rho}(\Psi_s) < \frac{1}{2\kappa}$ never holds if Ψ_s is continuous. Comparing to Theorem 2.3.2, we see that no continuous operator can guarantee restricted optimality at any sparsity ratio ρ, even in the ideal setting where f is well-conditioned. This includes soft thresholding at a fixed sparsity level, i.e., the map Ψ_s^{ST} that shrinks all entries of z equally until the desired sparsity level is reached:

$$\left(\Psi_s^{\mathrm{ST}}(z)\right)_i = \begin{cases} z_i - \lambda, & z_i > \lambda, \\ 0, & |z_i| \le \lambda, \\ z_i + \lambda, & z_i < -\lambda, \end{cases} \qquad \text{taking } \lambda \ge 0 \text{ to be the smallest value s.t. } \|\Psi_s^{\mathrm{ST}}(z)\|_0 \le s.$$

In practice, it is much more common to implement soft thresholding at a fixed λ, rather than at a fixed s. We will discuss the fixed-λ formulation of soft thresholding later on, in Section 2.4.6.

2.4.2 Optimal value of relative concavity

In this section we turn to the question of optimality: what is the optimal value of relative concavity among all thresholding operators at a given sparsity proportion ρ? We will establish that

$$\inf_{\Psi_s : \mathbb{R}^d \to \{x \in \mathbb{R}^d : \|x\|_0 \le s\}} \gamma_{s,\rho}(\Psi_s) = \frac{\rho}{1+\rho}.$$

That is, the lowest relative concavity among all thresholding operators at a given sparsity proportion ρ is exactly $\frac{\rho}{1+\rho}$. Since this is much smaller than $\frac{\sqrt{\rho}}{2}$ when ρ is small, we see that hard thresholding is suboptimal.

We start with the following lower bound for all thresholding operators:

Lemma 2.4.3. *For any map $\Psi_s : \mathbb{R}^d \to \{x \in \mathbb{R}^d : \|x\|_0 \le s\}$ and any sparsity proportion $\rho \in (0,1]$, the relative concavity is lower-bounded as*

$$\gamma_{s,\rho}(\Psi_s) \ge \frac{\rho}{1+\rho}.$$

To show that this lower bound is indeed tight, we will consider ℓ_q thresholding and establish upper bound for its relative concavity that matches this lower bound with proper choices of q. ℓ_q thresholding encourage sparsity without exerting too much shrinkage by constraining the ℓ_q norm of the vector after thresholding for some $q \in (0,1)$. To be precise, let

$$P_{\ell_q}(z;t) = \arg\min \left\{ \|x-z\|_2 : \|x\|_q \le t \right\}$$

denote projection to the ℓ_q ball, where $\|x\|_q = (\sum_i |x_i|^q)^{1/q}$ is the ℓ_q "norm" (in fact a nonconvex function since $q < 1$). Then define

$$\Psi_s^{\ell_q}(z) = P_{\ell_q}(z;t(z)), \text{ where } t(z) = \sup \left\{ t : \|P_{\ell_q}(z;t)\|_0 \le s \right\}.$$

In words, $\Psi_s^{\ell_q}(z)$ projects z to an ℓ_q ball whose radius is chosen to be as large as possible while still ensuring s-sparsity.[3] The following result computes the relative concavity for ℓ_q thresholding:

Lemma 2.4.4. *The relative concavity of ℓ_q thresholding $\Psi_s^{\ell_q}$ is equal to*

$$\gamma_{s,\rho}(\Psi_s^{\ell_q}) = \frac{\frac{\rho}{\min\{1,(\frac{2-q}{q})^2(1-\rho)\}}}{\frac{4q(1-q)}{(2-q)^2}(1+\sqrt{1+(\frac{2-q}{q})^2\frac{\rho}{\min\{1,(\frac{2-q}{q})^2(1-\rho)\}}})}$$

3. Note that $P_{\ell_q}(z;t)$ may be non-unique. To be fully precise, we define $\Psi_s^{\ell_q}(z)$ by first fixing some map $z \mapsto S$, the possibly non-unique support of its largest s entries, and then defining $t(z)$ and choosing the possibly non-unique projection $P_{\ell_q}(z;t(z))$ in such a way that the nonzero entries in the projection are exactly on this support.

for every sparsity proportion $\rho \in (0,1)$. In particular, if we choose

$$q = \frac{2(1-\rho)}{3-\rho},$$

then the resulting thresholding operator attains the lowest possible relative concavity,

$$\gamma_{s,\rho}(\Psi_s^{\ell_q}) = \frac{\rho}{1+\rho}, \text{ for } q = \frac{2(1-\rho)}{3-\rho}.$$

In addition, the universal choice $q = 2/3$ yields relative concavity equal to,

$$\gamma_{s,\rho}(\Psi_s^{\ell_{2/3}}) = \frac{\frac{\rho}{\min\{1,4(1-\rho)\}}}{\frac{1}{2}+\frac{1}{2}\sqrt{1+\frac{4\rho}{\min\{1,4(1-\rho)\}}}} \leq \frac{\rho}{\min\{1,4(1-\rho)\}}, \text{ for all } \rho \in (0,1).$$

Now we provide some explanation for this result. If we are allowed to choose q depend on ρ, then the choice $q = \frac{2(1-\rho)}{3-\rho}$ would lead to a relative concavity of $\frac{\rho}{1+\rho}$, which exactly matches the lower bound in Lemma 2.4.3. Of course this specific choice of q is chosen for a specific sparsity proportion ρ and might not work well for other values of the sparsity proportion. To avoid this drawback or the need to tune the parameter q, one can have the universal choice $q = 2/3$. Due to the expression for $\gamma_{s,\rho}(\Psi_s^{\ell_{2/3}})$, we see that $\gamma_{s,\rho}(\Psi_s^{\ell_{2/3}}) \approx \rho$ when ρ is small, thus nearly matching the lower bound $\frac{\rho}{1+\rho}$.

In particular, with the optimal value of relative concavity $\gamma_{s,\rho} = \frac{\rho}{1+\rho}$, the condition $\gamma_{s,\rho} < \frac{1}{2\kappa}$ becomes $\rho < \frac{1}{2\kappa-1}$. In light of Theorem 2.3.1 and Theorem 2.3.2, we see that $\rho < \frac{1}{2\kappa-1}$ is both necessary and sufficient for restricted optimality to hold with sparsity proportion ρ. Compare this with the condition $\rho < \frac{1}{\kappa^2}$ required by hard thresholding, we see that the dependence on condition number is greatly improved!

2.4.3 A general class of thresholding operators

Now that we have seen that ℓ_q thresholding operators enjoy good properties in terms of relative concavity, we can ask whether there are other thresholding operators of such optimal and near-optimal properties. In this section we address this problem by showing ℓ_q thresholding can be characterized as a special case of a larger class of thresholding operators, which all enjoy the same optimality properties in the sense of their relative concavity. Consider any nonincreasing function

$$\sigma : [1, \infty) \to [0, 1],$$

which we call the "shrinkage function", which will determine the amount of shrinkage on each entry of a vector z at the thresholding step. Defining the support S and thresholding level $\tau = \max_{i \notin S} |z_i|$ as before, we then define the thresholding operator $\Psi_{s;\sigma}$ as

$$\big(\Psi_{s;\sigma}(z)\big)_i = \begin{cases} z_i - \tau\sigma\big(|z_i|/\tau\big), & i \in S, \\ 0, & i \notin S. \end{cases}$$

In other words, for entry $i \in S$, $\sigma\big(|z_i|/\tau\big)$ determines the *relative* amount of shrinkage on this entry. The intuitive meaning of σ is illustrated in Figure 2.1. (If $\tau = 0$, i.e. z is already s-sparse, then we would simply take $\Psi_{s;\sigma}(z) = z$; we will ignore this case from this point on.)

Note that since σ is nondecreasing, the maximum shrinkage occurs when $|z_i| = \tau$ exactly; the amount of shrinkage in this setting is governed by $\sigma(1)$.

We can now examine the relationship of the choice of σ to the relative concavity:

Lemma 2.4.5. *For any nonincreasing shrinkage function $\sigma : [1, \infty) \to [0, 1]$ such that $0 < \sigma(1) < 1$ and*

$$t \mapsto \sigma(t)(t - \sigma(t)) \text{ is nondecreasing over } t \geq 1, \tag{2.6}$$

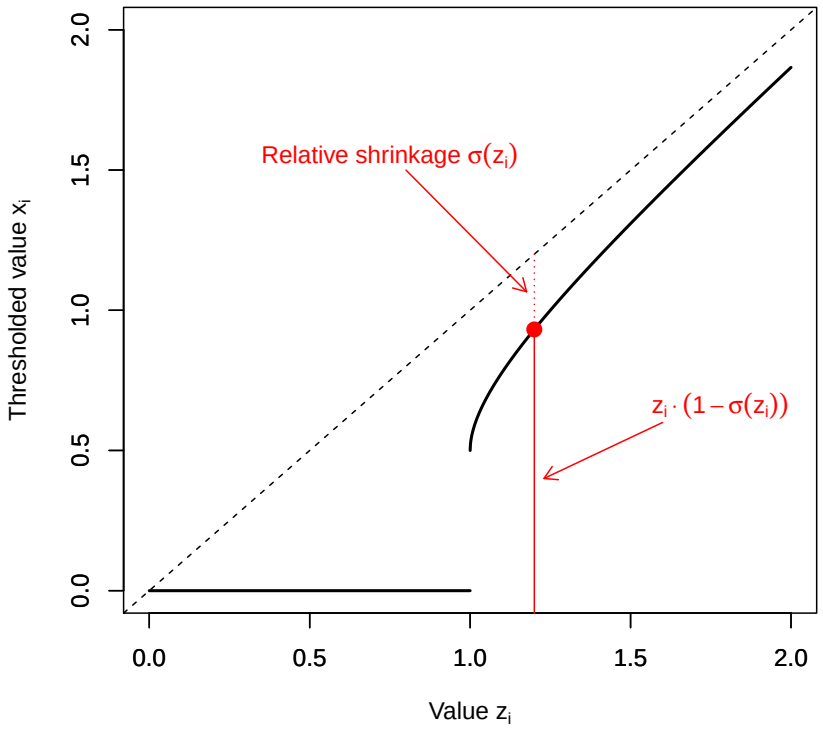


Figure 2.1: An illustration of the definition of the relative shrinkage function σ (for simplicity we use thresholding level $\tau = 1$ in this illustration). Here we use the relative shrinkage function $\sigma(t) = \frac{t-\sqrt{t^2-1}}{2}$, corresponding to the reciprocal thresholding operator defined in Section 2.4.4.

the thresholding operator $\Psi_{s;\sigma}$ has relative concavity

$$\gamma_{s,\rho}(\Psi_{s;\sigma}) = \frac{\frac{\rho}{\min\{1,(1-\rho)/\sigma(1)^2\}}}{2\sigma(1)\big(1-\sigma(1)\big)\left(1+\sqrt{1+\frac{\rho/\sigma(1)^2}{\min\{1,(1-\rho)/\sigma(1)^2\}}}\right)}.$$

In particular, the resulting operator attains the lowest possible relative concavity,

$$\gamma_{s,\rho}(\Psi_{s;\sigma}) = \frac{\rho}{1+\rho},$$

if and only if $\sigma(1) = \frac{1-\rho}{2}$. If instead we take a universal shrinkage level $\sigma(1) = \frac{1}{2}$, then the relative

concavity is given by

$$\gamma_{s,\rho}(\Psi_{s;\sigma}) = \frac{\frac{\rho}{\min\{1,4(1-\rho)\}}}{\frac{1}{2}+\frac{1}{2}\sqrt{1+\frac{4\rho}{\min\{1,4(1-\rho)\}}}} \leq \frac{\rho}{\min\{1,4(1-\rho)\}}.$$

Examining the definition of this general family of thresholding operators, we can see that ℓ_q thresholding corresponds to setting

$$\sigma(t) = t - \left(\text{the larger-magnitude root } x \text{ of the equation } t = x + \frac{q(2-2q)^{1-q}/(2-q)^{2-q}}{x^{1-q}}\right),$$

for which we have $\sigma(1) = \frac{q}{2-q}$ and which satisfies (2.6). We also have that $\sigma(1) = \frac{1}{2}$ corresponds to the "universal" choices $q = 2/3$, and $\sigma(1) = \frac{1-\rho}{2}$ (the optimal value) corresponds to the ρ-specific choices $q = \frac{2(1-\rho)}{3-\rho}$. As a consequence, the previous result for ℓ_q thresholding, Lemma 2.4.4, is simply special case of this more general lemma.

On the other hand, the hard thresholding operator Ψ_s^{HT} can be obtained by setting $\sigma(t) = 0$ for all $t \in [1,\infty)$, but this does not satisfy the assumption $\sigma(1) > 0$ required in the lemma. However, if we informally consider fixing $\rho > 0$ and taking a limit $\sigma(1) \to 0$ in the upper bound in the lemma, we see

$$\lim_{\sigma(1)\to 0} \frac{\frac{\rho}{\min\{1,\frac{1-\rho}{\sigma(1)^2}\}}}{2\sigma(1)(1-\sigma(1))\left(1+\sqrt{1+\frac{\rho}{\sigma(1)^2\min\{1,\frac{1-\rho}{\sigma(1)^2}\}}}\right)} = \frac{\sqrt{\rho}}{2},$$

obtaining the relative concavity of hard thresholding calculated earlier.

2.4.4 Reciprocal thresholding and minimal shrinkage

Practically, for two thresholding operators with the same restricted optimality guarantees, i.e. with the exact same value of relative concavity, we may favor the one that exerts smaller amount of shrinkage. Thus it makes sense to ask among the general class of thresholding operators defined in Section 2.4.3, which operators exert the minimal amount of shrinkage? Consider all operators of

the form $\Psi_{s;\sigma}$, with some fixed value of $\sigma(1) \in (0, 1/2]$. For any σ satisfying the assumption (2.6), for all $t \geq 1$ we have

$$\sigma(t)(t - \sigma(t)) \geq \sigma(1)(1 - \sigma(1)).$$

For convenience, we reparametrize this equation by setting $c = 1 - 2\sigma(1) \in [0, 1)$, and so we are considering all nonincreasing functions $\sigma : [1, \infty) \to [0, 1]$ that satisfy $\sigma(1) = \frac{1-c}{2}$ and

$$\sigma(t)(t - \sigma(t)) \geq \sigma(1)(1 - \sigma(1)) = \frac{1 - c^2}{4}.$$

Thus, we must have

$$\sigma(t) \geq \frac{t - \sqrt{t^2 - (1 - c^2)}}{2} \tag{2.7}$$

for all $t \geq 1$.

This motivates a new family of thresholding operators, *reciprocal thresholding with parameter c*, which is designed to make the inequality (2.7) an equality. To be specific, we define reciprocal thresholding with parameter c to be

$$\Psi_s^{\mathrm{RT},c} = \Psi_{s;\sigma} \text{ with shrinkage function } \sigma(t) = \frac{t - \sqrt{t^2 - (1 - c^2)}}{2}.$$

To apply this operator to some vector $z \in \mathbb{R}^d$, we first let $S \subset \{1, \ldots, d\}$ be the indices of the largest s entries of z (with our usual caveat about needing to establish some rule for breaking ties) and let $\tau = \max_{i \notin S} |z_i|$ be the magnitude of the $(s+1)$-st largest entry of z. Then $\Psi_s^{\mathrm{RT},c}(z)$ operates entry-wise as follows:

$$\left(\Psi_s^{\mathrm{RT},c}(z)\right)_i = \begin{cases} \operatorname{sign}(z_i) \cdot \left(\frac{1}{2}|z_i| + \frac{1}{2}\sqrt{|z_i|^2 - \tau^2(1 - c^2)}\right), & \text{if } i \in S, \\ 0, & \text{if } i \notin S. \end{cases} \tag{2.8}$$

Here the thresholded value $\left(\Psi_s^{\mathrm{RT},c}(z)\right)_i$ is equal to the larger-magnitude root t of the equation

$$z_i = t + \frac{\tau^2 \cdot \frac{1-c^2}{4}}{t}, \tag{2.9}$$

hence the name "reciprocal thresholding".

As before, to avoid the need for selecting c adaptively, we might want to consider some fixed choices. At one extreme, taking $c = 1$ yields $\Psi_s^{\mathrm{RT},1} = \Psi_s^{\mathrm{HT}}$, the hard thresholding operator. At the other extreme, taking $c = 0$ defines the "universal" *reciprocal thresholding* operator:

$$\Psi_s^{\mathrm{RT}} = \Psi_s^{\mathrm{RT},0}.$$

For any $z \in \mathbb{R}^d$, Ψ_s^{RT} operate entry-wise as:

$$\left(\Psi_s^{\mathrm{RT}}(z)\right)_i = \begin{cases} \operatorname{sign}(z_i) \cdot \left(\frac{1}{2}|z_i| + \frac{1}{2}\sqrt{|z_i|^2 - \tau^2}\right), & \text{if } i \in S, \\ 0, & \text{if } i \notin S. \end{cases} \tag{2.10}$$

The following lemma calculates the relative concavity of $\Psi_s^{\mathrm{RT},c}$ and Ψ_s^{RT} as a direct consequence of Lemma 2.4.5.

Lemma 2.4.6. *For any sparsity proportion $\rho \in (0,1]$, the thresholding operator $\Psi_s^{\mathrm{RT},c}$ with parameter $c = \rho$ has relative concavity equal to*

$$\gamma_{s,\rho}(\Psi_s^{\mathrm{RT},\rho}) = \frac{\rho}{1+\rho}.$$

The reciprocal thresholding operator Ψ_s^{RT} has relative concavity equal to

$$\gamma_{s,\rho}(\Psi_s^{\mathrm{RT}}) = \frac{\frac{\rho}{\min\{1,4(1-\rho)\}}}{\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4\rho}{\min\{1,4(1-\rho)\}}}} \leq \frac{\rho}{\min\{1,4(1-\rho)\}}$$

for every sparsity proportion $\rho \in (0,1)$.

Thus, $\Psi_s^{\mathrm{RT},c}$ with $c = \rho$ is exactly optimal among all thresholding operators relative to the sparsity proportion ρ (as is $\Psi_s^{\ell_q}$ with $q = \frac{2(1-\rho)}{3-\rho}$), while Ψ_s^{RT} is near optimal when ρ is small (as is $\Psi_s^{\ell_{2/3}}$).

2.4.5 An illustrative comparison

Through the development in this section, we see that there are three important benchmarks for relative concavity: the bound $\sqrt{\rho}/2$ attained by hard thresholding Ψ_s^{HT}, the near otpimal bound $\frac{\rho}{\frac{1}{2}+\frac{1}{2}\sqrt{1+\frac{4\rho}{\min\{1,4(1-\rho)\}}}}$ attained by reciprocal thresholding Ψ_s^{RT} and $\ell_{2/3}$ thresholding $\Psi_s^{\ell_{2/3}}$, and the optimal value $\frac{\rho}{1+\rho}$. In this section we provide a comparison between these three values.

The left-hand plot of Figure 2.2 displays the three values of relative concavity as functions of the sparsity proportion ρ. We see that at small values $\rho \approx 0$, the relative concavity of reciprocal thresholding and $\ell_{2/3}$ thresholding is nearly identical to the optimal bound $\frac{\rho}{1+\rho}$, and is substantially better than the relative concavity for hard thresholding, given by $\sqrt{\rho}/2$. At larger values of ρ, the relative concavity for hard thresholding is instead lower.

To view this comparison in another light, given any fixed thresholding operator Ψ_s with certain relative concavity, and given an objective function f with condition number κ, for what sparsity ratio $\rho = s'/s$ is the iterative thresholding algorithm guaranteed to achieve restricted optimality? Using the condition $\gamma_{s,\rho}(\Psi_s) \leq \frac{1}{2\kappa}$, for each relative concavity $\gamma_{s,\rho}$ we can solve for the largest possible κ for which restricted optimality is assured, as a function of ρ.

This is illustrated in the right-hand plot of Figure 2.2, where we see that the reciprocal thresholding operator Ψ_s^{RT} and the $\ell_{2/3}$ thresholding operator achieve a nearly-optimal sparsity ratio ρ when the condition number κ is large and ρ is correspondingly close to zero, while hard thresholding Ψ_s^{HT} is closer to optimal for κ and ρ close to 1. Thus, we can conclude that reciprocal thresholding and $\ell_{2/3}$ thresholding offer stronger theoretical guarantees when $\kappa > 2$, while hard thresholding may be better for very well-conditioned problems where $1 \leq \kappa < 2$. (Empirically, we have observed that it is often the case that the three perform nearly identically in "generic" problems, and only show substantial differences in problems constructed to mimic our lower bound result, Theorem 2.3.2, for example, in linear regression problems where a small subset of the fea-

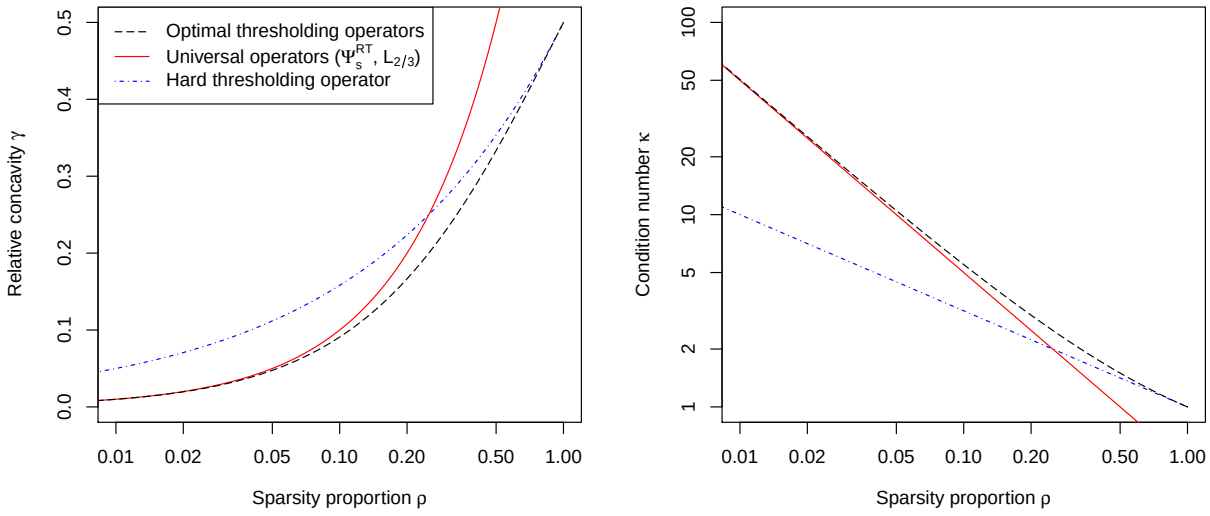


Figure 2.2: A comparison of three values of relative concavity: the optimal relative concavity (attained by, for instance, $\Psi_s^{\mathrm{RT},c}$ with $c=\rho$, or by $\Psi_s^{\ell_q}$ with $q=\frac{2(1-\rho)}{3-\rho}$); the relative concavity obtained by the "universal" operators, including Ψ_s^{RT} and by $\ell_{2/3}$ thresholding; and the relative concavity of hard thresholding. The left plot shows the relative concavity as a function of the sparsity proportion ρ. The right plot shows the largest possible condition number κ of the objective function f for which a restricted optimality guarantee can be attained (Theorems 2.3.1 and 2.3.2 show that $\gamma_{s,\rho}(\Psi_s)\leq\frac{1}{2\kappa}$ is necessary and sufficient for a restricted optimality guarantee).

tures are generated to have covariance structure similar to the construction in Theorem 2.3.2.)

2.4.6 A closer look at soft thresholding

In many applications, it is common to use a sparsity-inducing penalty rather than an explicit sparsity constraint. For example, we may solve

$$\widehat{x}=\arg\min_{x\in\mathbb{R}^d}\left\{\mathsf{f}(x)+\lambda\|x\|_1\right\},$$

which is known as the Lasso [16] in the context of regression problems. More generally, we can consider

$$\widehat{x}_\lambda=\arg\min_{x\in\mathbb{R}^d}\left\{\mathsf{f}(x)+\lambda\mathsf{R}(x)\right\}, \tag{2.11}$$

where $\mathsf{R} : \mathbb{R}^d \to \mathbb{R}$ is any proper convex function acting as a regularizer. This class of problems can be solved iteratively with a proximal gradient method,

$$x_t = \mathrm{Prox}_{\lambda\eta\mathsf{R}}\big(x_{t-1} - \eta\nabla\mathsf{f}(x_{t-1})\big), \tag{2.12}$$

where for any $t \geq 0$, the proximal map is defined as

$$\mathrm{Prox}_{t\mathsf{R}}(z) = \arg\min_{x\in\mathbb{R}^d}\left\{\frac{1}{2}\|x-z\|_2^2 + t\mathsf{R}(x)\right\}.$$

Note that convexity of $\mathsf{R}(x)$ ensures continuity of the proximal map. More properties of the proximal map and the proximal method can be found in [24]. Examining the iterations of proximal gradient descent (2.12), we see that it is very similar to the iterative thresholding method (2.3) for a fixed sparsity level s (using some particular thresholding operator Ψ_s); we simply replace the thresholding operator Ψ_s with the proximal map $\mathrm{Prox}_{\lambda\eta\mathsf{R}}\cdot$

In particular, if we consider $\mathsf{R}(x) = \|x\|_1$, the resulting proximal map is known as "soft thresholding", and can be computed with elementwise shrinkage:

$$\big(\mathrm{Prox}_{t\|\cdot\|_1}(z)\big)_i = \begin{cases} z_i - t, & z_i > t, \\ 0, & |z_i| \leq t, \\ z_i + t, & z_i < -t. \end{cases}$$

Now, recall that in Section 2.4.1, we considered a "soft thresholding" operator at a *fixed* sparsity level s, which we can now rewrite as

$$\Psi_s^{\mathrm{ST}}(z) = \mathrm{Prox}_{t\|\cdot\|_1}(z) \quad \text{where } t \geq 0 \text{ is the smallest value s.t. } \|\mathrm{Prox}_{t\|\cdot\|_1}(z)\|_0 \leq s.$$

We might ask whether the suboptimal worst-case performance of iterative thresholding with the operator Ψ_s^{ST}, as established by Lemma 2.4.2 and Theorem 2.3.2, is due to the unusual definition

of Ψ_s^{ST}, using a fixed sparsity level s, rather than the usual form of soft thresholding where we would iterate (2.12) at a fixed value of λ in order to minimize $\mathsf{f}(x) + \lambda\|x\|_1$.

In fact, we will now see that this is not the case—even if we use a fixed λ rather than a fixed sparsity level s, we can still find worst-case examples where restricted optimality is not achieved.

Theorem 2.4.1. *Let $d \geq 2$, let $\beta \geq \alpha > 0$, and let $\mathsf{R} : \mathbb{R}^d \to \mathbb{R}$ be a proper convex function that satisfies the following assumptions:*

$$\text{For any } z \in \mathbb{R}^d \text{ and any } t' > t \geq 0, \text{ if } \|\mathrm{Prox}_{t\mathsf{R}}(z)\|_0 < d \text{ then } \|\mathrm{Prox}_{t'\mathsf{R}}(z)\|_0 < d. \tag{2.13}$$

$$\text{There exist } v, w \in \mathbb{R}^d \text{ that are both dense, i.e., } \|v\|_0 = \|w\|_0 = d, \text{ with } w \in \partial\mathsf{R}(v). \tag{2.14}$$

Then there exists an objective function $\mathsf{f}(x)$ that satisfies (α, d)-RSC and (β, d)-RSM, and a 1-sparse vector $y \in \mathbb{R}^d$, such that defining $\widehat{x}_\lambda$ as in (2.11)*,*

$$\text{For all } \lambda \geq 0, \text{ either } \|\widehat{x}_\lambda\|_0 = d \text{ or } \mathsf{f}(\widehat{x}_\lambda) > \mathsf{f}(y).$$

In other words, this result means there is no value of λ that produces a solution that is both sparse (at any sparsity level $< d$) and has an objective function value at least as good as the best 1-sparse solution y.

We remark that our conditions (2.13) and (2.14) on the regularizer R are satisfied by many common regularizers—for example, the ℓ_1 norm (Lasso), any ℓ_p norm for $1 \leq p \leq \infty$, the elastic net (a combination of the ℓ_1 and ℓ_2 norms), the weighted ℓ_1 norm, and many others. To help interpret the first condition (2.13), this essentially requires that a sparse solution $\widehat{x}_\lambda$ will stay sparse if we increase the penalty parameter λ, as we would expect for any sparsity-promoting regularizer.

This theorem implies that, just as continuous thresholding operators Ψ_s at a fixed level s can fail to attain restricted optimality in a worst-case scenario, the same holds for regularization with convex penalties (such as soft thresholding with the ℓ_1 norm). An open question remains here, namely, is there a measure in the style of relative concavity, which can characterize the worst-case performance of penalty functions R (covering both convex and nonconvex penalty functions, just

as relative concavity treats both continuous and non-continuous thresholding operators)?

2.5 Iterative thresholding for low-rank matrices

We next extend our analysis of iterative thresholding methods to the setting of a low-rank constraint. In fact, our results carry over fully into this setting. Given a rank constraint, $\text{rank}(X) \leq s$, the hard thresholding operator is defined as

$$\widetilde{\Psi}_s^{\text{HT}} : X \mapsto U \cdot \text{diag}(\Psi_s^{\text{HT}}(d)) \cdot V^\top,$$

where $X = U \cdot \text{diag}(d) \cdot V^\top$ is the singular value decomposition of X.[4] That is, hard thresholding is performed on the singular values of the matrix X, rather than on its entries. Of course, we can extend this to any thresholding operator—given any $\Psi_s : \mathbb{R}^{\min\{n,m\}} \to \{x \in \mathbb{R}^{\min\{n,m\}} : \|x\|_0 \leq s\}$, we can "lift" this thresholding operator to the matrix setting by defining

$$\widetilde{\Psi}_s : X \mapsto U \cdot \text{diag}(\Psi_s(d)) \cdot V^\top. \tag{2.15}$$

Of course, its possible to construct a rank-s thresholding operator $\widetilde{\Psi}_s$ that is not of the form given in (2.15), for example, if $\widetilde{\Psi}_s$ does not preserve the left and right singular vectors of Z.

We next extend our convergence results, Theorems 2.3.1 and 2.3.2, to the low-rank setting. In order to do so, we need to define the matrix version of relative concavity—this definition is analogous to the vector case, with rank constraints in place of sparsity constraints:

$$\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) = \sup \left\{ \frac{\langle Y - \widetilde{\Psi}_s(Z), Z - \widetilde{\Psi}_s(Z) \rangle}{\|Y - \widetilde{\Psi}_s(Z)\|_{\text{F}}^2} \ : Y, Z \in \mathbb{R}^{n \times m}, \text{rank}(Y) \leq \rho s, Y \neq \widetilde{\Psi}_s(Z) \right\}.$$

As for the vector case, relative concavity is necessary and sufficient for guaranteeing restricted

4. In the case of repeated singular values, the singular value decomposition will not be unique, and we assume that we have some mechanism for specifying a specific singular value decomposition. This is analogous to the sparse vector problem, where if the sth largest entry in z is not unique, we need to assume some mechanism for breaking ties and choosing the support of the thresholded vector.

optimality—in fact, the proofs of these are completely identical to the vector case. For completeness, we state the results here, for the matrix version of the iterated thresholding algorithm:

$$X_t = \widetilde{\Psi}_s\big(X_{t-1} - \eta_t \nabla \mathsf{f}(X_{t-1})\big), \tag{2.16}$$

with either fixed step size $\eta_t = 1/\beta$ or adaptive step size defined as in (2.4).

Theorem 2.5.1. *Consider any objective function $\mathsf{f} : \mathbb{R}^{n\times m} \to \mathbb{R}$, any ranks $s \geq s'$, and any rank-s thresholding operator $\widetilde{\Psi}_s$. Assume the objective function f satisfies (α, s)-RSC and (β, s)-RSM relative to the rank constraint.*[5] *Let $\rho = s'/s$ and $\kappa = \beta/\alpha$, and assume that $\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) < \frac{1}{2\kappa}$. Then, for any $X_0, Y \in \mathbb{R}^{n\times m}$ with $\mathrm{rank}(X_0) \leq s$ and $\mathrm{rank}(Y) \leq s'$, the iterated thresholding algorithm* (2.16) *run with step size $\eta = 1/\beta$ and initialization point X_0 satisfies*

$$\min_{t=1,\ldots,T} \mathsf{f}(X_t) \leq \mathsf{f}(Y) + \left(\frac{1 - 1/\kappa}{1 - 2\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s)}\right)^T \cdot \frac{\beta}{2}\|X_0 - Y\|_{\mathrm{F}}^2$$

for each $T \geq 1$.

Theorem 2.5.2. *Consider any ranks $s \geq s'$, any rank-s thresholding operator $\widetilde{\Psi}_s$, and any constants $\beta \geq \alpha > 0$. Let $\rho = s'/s$ and $\kappa = \beta/\alpha$, and assume that $\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) > \frac{1}{2\kappa}$. Then there exists an objective function $\mathsf{f}(X)$ that satisfies (α, s)-RSC and (β, s)-RSM relative to the rank constraint, and matrices $X_0, Y \in \mathbb{R}^{n\times m}$ with $\mathrm{rank}(X_0) \leq s$ and $\mathrm{rank}(Y) \leq s'$, such that the iterated thresholding algorithm* (2.16) *run with step size $\eta = 1/\beta$ and initialization point X_0 satisfies*

$$\lim_{t\to\infty} \mathsf{f}(X_t) > \mathsf{f}(Y).$$

In other words, just as for the sparse optimization problem, the relationship between relative concavity and condition number gives a necessary and sufficient condition for guaranteed convergence. We note that these results apply to *any* rank-s thresholding operator $\widetilde{\Psi}_s$, whether or not it can be

5. In the low-rank setting, the RSC and RSM conditions are defined with rank in place of sparsity—specifically, we are assuming that $\frac{\alpha}{2}\|X - Y\|_{\mathrm{F}}^2 \leq \mathsf{f}(Y) - \mathsf{f}(X) - \langle \nabla \mathsf{f}(X), Y - X\rangle \leq \frac{\beta}{2}\|X - Y\|_{\mathrm{F}}^2$ whenever $\mathrm{rank}(X) \leq s, \mathrm{rank}(Y) \leq s$.

constructed by "lifting" a s-sparse thresholding operator as in (2.15).

Next, how can we calculate relative concavity of a thresholding operator in the matrix setting? For simplicity, from this point on we assume that we are working with ranks $s \geq s' \geq 1$ with $s+s' \leq \min\{n,m\}$. For this question, we will again see that results from the sparse setting transfer to the low-rank setting. First, we have the same lower bound uniformly over all operators:

Lemma 2.5.1. *For any map $\widetilde{\Psi}_s : \mathbb{R}^{n\times m} \to \{X \in \mathbb{R}^{n\times m} : \operatorname{rank}(X) \leq s\}$ and any sparsity proportion $\rho \in (0,1]$, the relative concavity is lower-bounded as*

$$\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) \geq \frac{\rho}{1+\rho}.$$

Furthermore, if we restrict our attention to "lifted" thresholding operators of the form (2.15), the relative concavity of Ψ_s is inherited by the lifted operator $\widetilde{\Psi}_s$—as long as we restrict ourselves to s-sparse thresholding operators Ψ_s that satisfy a natural sign condition:

$$\text{For any } z \in \mathbb{R}^d \text{ and any } a \in \{\pm 1\}^d,\ \Psi_s\big(\operatorname{diag}(a)\cdot z\big) = \operatorname{diag}(a)\cdot \Psi_s(z). \tag{2.17}$$

This effectively means that $\Psi_s(z)$ preserves the signs of z, but the signs of z do not affect the amount of shrinkage in the thresholded vector $\Psi_s(z)$. For example, this requires that $\Psi_s(-z) = -\Psi_s(z)$. Under this assumption, the relative concavity of Ψ_s carries over into the matrix setting.

Lemma 2.5.2. *Let Ψ_s be a s-sparse thresholding operator satisfying the sign condition* (2.17)*, and let $\widetilde{\Psi}_s$ be the lifted thresholding operator defined in* (2.15)*. Then for every sparsity proportion $\rho \in (0,1]$,*

$$\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) = \gamma_{s,\rho}(\Psi_s).$$

It is obvious that all the thresholding operators we have considered satisfy the sign condition (2.17). Thus, all the results of relative concavity that we have proved in the sparse setting, carry over directly to the low-rank setting. In particular, as for the sparse setting, the hard thresholding operator

has relative concavity

$$\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s^{\mathrm{HT}}) = \frac{\sqrt{\rho}}{2}, \tag{2.18}$$

while any thresholding operator $\widetilde{\Psi}_{s;\sigma}$ constructed with some shrinkage function σ satisfying $\sigma(1) = 1/2$ and the conditions of Lemma 2.4.5, such as the reciprocal thresholding operator, $\widetilde{\Psi}_s^{\mathrm{RT}}$, or ℓ_q thresholding with $q = 2/3$, $\widetilde{\Psi}_s^{\ell_{2/3}}$, satisfy

$$\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_{s;\sigma}) = \widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s^{\ell_{2/3}}) = \widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s^{\mathrm{RT}}) \leq \frac{\rho}{\min\{1, 4(1-\rho)\}}.$$

If the desired rank proportion $\rho = s'/s$ is fixed in advance, then as before, choosing reciprocal thresholding with parameter $c = \rho$, or ℓ_q thresholding with $q = \frac{2(1-\rho)}{3-\rho}$, we again obtain the optimal relative concavity of $\frac{\rho}{1+\rho}$. As before, we can conclude that reciprocal thresholding and $\ell_{2/3}$ each offer lower relative concavity than hard thresholding whenever ρ is small—and, correspondingly, are a safer choice for objective functions f whose condition number is not close to 1.

2.6 Sparse linear regression

Now that we have discussed the deterministic optimization setting in depth, it is natural to ask what is the implication of these guarantee for a statistically random setting. In this section, we apply our developed machinery to the concrete statistical setting of sparse linear regression. We work with the Gaussian linear model

$$y = X\theta_0 + z \tag{2.19}$$

where $X \in \mathbb{R}^{n \times p}$ is a fixed design matrix, $\theta_0 \in \mathbb{R}^p$ is the true coefficient vector assumed to be fixed and s_0-sparse, and $z \sim N(0, \sigma^2 \mathbf{I}_n)$ is the noise vector, with fixed unknown noise level $\sigma^2 > 0$. In this section we will mainly be interested in prediction error, i.e. how well we can estimate the true mean vector $X\theta_0$. One way of capturing the conditioning of the design matrix is by the following

definition: at some given sparsity level s, we define a set of design matrices $\mathscr{X}(\alpha,\beta,s)$ as

$$\mathscr{X}(\alpha,\beta,s) = \left\{ X \in \mathbb{R}^{n\times p} : \text{ the map } \theta \mapsto \theta^\top \left(\frac{X^\top X}{2n}\right)\theta \text{ satisfies } (\alpha,s)\text{-RSC and } (\beta,s)\text{-RSM} \right\}. \tag{2.20}$$

As usual, we will be interested in the condition number $\kappa = \beta/\alpha$. A similar definition is the *restricted eigenvalue* condition on the design matrix X, which constrains X to the following set

$$\mathscr{X}_{\mathrm{RE}}(\kappa,s_0) = \left\{ X \in \mathbb{R}^{n\times p} : \max_{j=1,\ldots,p} \frac{\|X_j\|_2}{\sqrt{n}} \le 1, \text{ and } \theta^\top \left(\frac{X^\top X}{2n}\right)\theta \ge \frac{1}{2\kappa}\|\theta\|_2^2 \right.$$
$$\left. \text{for all } \theta \in \mathbb{R}^d \text{ with } \|\theta\|_1 \le 4\max_{|S|=s_0}\|\theta_S\|_1 \right\}. \tag{2.21}$$

To gain some intuition for when these conditions may hold, for a design matrix X whose rows are i.i.d. draws from a normal distribution $N(0,\Sigma)$, Raskutti et al. [25, Theorem 1] show that the population-level eigenvalues of the covariance Σ are approximately preserved in the design matrix, at any sparsity level $s \ll \frac{n}{\log(p)}$.

Computational lower bound In terms of prediction error, the optimal method, ℓ_0 constrained least squares method, is not computable. Thus from the lower bound side, it is of interest to ask what is the lowest prediction error achievable in the class of computational feasible estimator. Recently, Zhang et al. [26] provide a partial answer to this question, restricting to the class of s_0 sparse estimator. Their main result (see Theorem 1 in Zhang et al. [26]) states the following (informally):

> Under the assumption that $NP \nsubseteq P \setminus poly$, for any $\delta \in (0,1)$, under some assumption on n,d,s_0 and for any κ in a wide range, there exists a design matrix $X \in \mathscr{X}_{\mathrm{RE}}(\kappa,s_0)$ such that for any computational efficient methods, the maximum prediction error (over all s_0 sparse θ_0) is lower bounded by (up to some constant) $\kappa \cdot \frac{\sigma^2 s_0^{1-\delta}\log(d)}{n}$.

Thus if we restrict ourselves to all computationally feasible s_0 sparse estimator, then the best achievable squared prediction error is of order $\kappa \cdot \frac{\sigma^2 s_0 \log(d)}{n}$.

Upper bounds for iterative thresholding methods In this section we establish prediction error bounds for iterative thresholding algorithm. First we provide some intuition on how to connect restricted optimality guarantee with statistical performance. It is well known that the global optimum of ℓ_0-constrained least squared loss, i.e.

$$\widehat{\theta} \in \arg \min_{\|\theta\|_0 \leq s_0} \|y - X\theta\|_2^2,$$

achieves a squared prediction error scaling as $\frac{\sigma^2 s_0 \log(d)}{n}$. For iterative thresholding algorithms, since we only have restricted optimality rather than global optimality, we are forced to work over a constraint at a larger sparsity $s \geq s_0$ to guarantee $\|y - X\widehat{\theta}\|_2^2 \leq \min_{\|\theta\|_0 \leq s_0} \|y - X\theta\|_2^2$. The statistical price one has to pay for this computational strategy is the inflation in noise level corresponding to the inflation in sparsity—that is, we have error on s many nonzero coefficients, rather than s_0 many—so the final upper bound for prediction error would scale as $\frac{\sigma^2 s \log(d)}{n}$ instead of $\frac{\sigma^2 s_0 \log(d)}{n}$, where s is chosen to be the smallest sparsity level that guaratees restricted optimality relative to the lower sparsity level $s' = s_0$. Now recall from Section 2.4 that, while hard thresholding offers restricted optimality guarantees at sparsity levels $s \sim \kappa^2 s_0$, the optimal and near-optimal thresholding operators (for example reciprocal thresholding and $\ell_{2/3}$ thresholding) improves this scaling to $s \sim \kappa s_0$. This allows us to improve the upper bound for squared prediction error from scaling as κ^2 to κ, when we switch our method from iterative hard thresholding, to iterative thresholding with an operator Ψ_s that enjoys a lower relative concavity. Indeed in Jain et al. [5], it is shown that iterative hard thresholding achieves a prediction error upper bounded by $\kappa^2 \cdot \frac{\sigma^2 s_0 \log(d)}{n}$. In view of our lower bound result Theorem 2.3.2, which states that the restircted optimality guarantee is tight, we postulate that the corresponding prediction error bound is also tight for iterative hard thresholding method.

Now we formulate this rigorously. Consider the iterative thresholding algorithm with some thresholding operator Ψ_s applied to the objective function $\mathsf{f}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$, whose iteration

takes the form

$$\widehat{\theta}_t = \Psi_s\big(\widehat{\theta}_{t-1} + \eta_t \cdot \frac{1}{n} X^\top (y - X\widehat{\theta}_{t-1})\big). \tag{2.22}$$

As usual, for the step size we may choose $\eta_t = 1/\beta$ if β is known, or we may choose η_t adaptively as in (2.4). We will work with any thresholding operator Ψ_s satisfying

$$\gamma_{s,\rho}(\Psi_s) \le \rho \text{ for all } \rho \in (0, 1/2). \tag{2.23}$$

From Section 2.4, we see that on the one hand, this condition rules out hard thresholding and any continuous thresholding operator; on the other hand, it is satisfied by the reciprocal thresholding operator, Ψ_s^{RT}, by ℓ_q thresholding with $q = 2/3$, $\Psi_s^{\ell_{2/3}}$, and by any shrinkage operator $\Psi_{s;\sigma}$ where $\sigma(1) = 1/2$ and σ satisfies the conditions of Lemma 2.4.5. We now present our result for this setting:

Theorem 2.6.1. *Suppose that $y = X\theta_0 + N(0, \sigma^2 \mathbf{I}_n)$, where θ_0 is s_0-sparse, and where X belongs to the set $\mathscr{X}(\alpha, \beta, s)$, where $s = C\kappa s_0$ for some $C > 2$. Suppose that Ψ_s is any s-sparse thresholding operator satisfying* (2.23)*.*

Let $\widehat{\theta}_t$ be the estimate produced at step t of the iterative thresholding algorithm (2.22) *initialized at some s-sparse $\widehat{\theta}_0 \in \mathbb{R}^d$. Let $\tilde{\theta}_t \in \arg\min_{\theta \in \{\widehat{\theta}_1, \dots, \widehat{\theta}_t\}} \frac{1}{2n}\|y - X\theta\|_2^2$, that is, $\tilde{\theta}_t$ is the best estimate seen before time t, relative to the loss function $\mathsf{f}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.*

Then, for any $\delta > 0$ and any $t \ge 1$,

$$\frac{1}{n}\|X(\tilde{\theta}_t - \theta_0)\|_2^2 \le \kappa \cdot \frac{28C\sigma^2 s_0 \log(d)}{n} + \frac{12\sigma^2 \log(1/\delta)}{n} + \left(\frac{1 - 1/\kappa}{1 - 2/C\kappa}\right)^t \cdot 2\beta\|\widehat{\theta}_0 - \theta_0\|_2^2,$$

with probability at least $1 - \delta$.

Since t can be taken to be large (each iteration is very cheap), the dominant term is the first one, so we essentially have

$$\frac{1}{n}\|X(\tilde{\theta}_t - \theta_0)\|_2^2 \lesssim \kappa \cdot \frac{\sigma^2 s_0 \log(d)}{n}.$$

Comparing with the upper bound for iterative hard thresholding, we see that we now attains the

ideal κ, rather than κ^2, scaling.

Comparison with Lasso The Lasso estimate of θ_0, given by the convex optimization problem

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2}\|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\},$$

is proved in Bickel et al. [27] to achieve a squared prediction error bounded as

$$\frac{1}{n}\|X(\widehat{\theta} - \theta_0)\|_2^2 \lesssim \kappa \cdot \frac{\sigma^2 s_0 \log(d)}{n} \tag{2.24}$$

with a penalty parameter value $\lambda \sim \sigma\sqrt{\frac{\log(d)}{n}}$, under the assumption that $X \in \mathscr{X}_{\mathrm{RE}}(\kappa, s_0)$. Compared with Lasso, due to Theorem 2.6.1, iterative thresholding algorithms with proper thresholding operators, for example the simple and efficient reciprocal thresholding, achieve the same squared prediction error bound. Moreover, both Lasso and iterative reciprocal thresholding method are guaranteed to give an estimator that is $\mathscr{O}(\kappa s_0)$ sparse (this sparsity level for Lasso is proved in Bickel et al. [27, Eqn. (7.9)]), and thus nearly match the computational lower bound with a gap in sparsity. An open question for future work is whether the larger sparsity level, i.e. $\mathscr{O}(\kappa s_0)$ rather than s_0, is unavoidable to achieve the squared prediction error $\kappa \cdot \frac{\sigma^2 s_0 \log(d)}{n}$, or whether there may be an $\mathscr{O}(s_0)$-sparse and computationally efficient estimator that achieves this bound.

CHAPTER 3

AN EQUIVALENCE BETWEEN CRITICAL POINTS FOR RANK CONSTRAINTS VERSUS LOW-RANK FACTORIZATIONS

In this chapter we consider low-rank optimization problem, and establish a near equivalent relation between the critical points of the constrained approach and of the factorized approach.[1]

3.1 Introduction

Consider the following low rank optimization problem

$$\min_{X \in \mathbb{R}^{m\times n}} \big\{ \mathsf{f}(X) : \operatorname{rank}(X) \leq r \big\}, \tag{3.1}$$

for a differentiable function $\mathsf{f} : \mathbb{R}^{m\times n} \to \mathbb{R}$. Due to a wide range of applications, this type of optimization problem has been studied extensively in the past decade. Two common approaches in low-rank optimization problems are either working directly with a rank constraint on the matrix variable, or optimizing over a low-rank factorization so that the rank constraint is implicitly ensured. When working with the full variable, a standard approach is to treat the rank-constrained set as a subset of the Euclidean space $\mathbb{R}^{m\times n}$, and apply constrained optimization algorithms. As our central example of this work, we consider the projected gradient descent method (also known as iterative hard thresholding, see Jain et al. [5]):

$$X \leftarrow \mathscr{P}_r\big(X - \eta \nabla \mathsf{f}(X)\big), \tag{3.2}$$

where $\mathscr{P}_r(\cdot)$ denotes projection to the rank-r constraint (calculated by taking the top r components of a singular value decomposition). Note that $\mathscr{P}_r$ is just $\widetilde{\Psi}_r^{\mathrm{HT}}$ in the notation of Chapter 2. On the

1. The work presented in this chapter can be found in [28] and is under revision. It is joint with Wooseok Ha and Rina Foygel Barber.

other hand, if we work instead in the factorized setting, we would aim to solve

$$\min_{A\in\mathbb{R}^{m\times r},B\in\mathbb{R}^{n\times r}} \mathsf{f}(AB^\top). \tag{3.3}$$

For instance, we might apply any unconstrained optimization techniques to this minimization, which attempt to update each of the two factors A, B. In contrast to the full-dimensional approach, these methods implicitly explore the space of low rank matrix manifold embedded in $\mathbb{R}^{m\times n}$. Comparing these options naturally raises the following question: is there a connection between the output of full-dimensional approaches such as PGD (3.2) versus factorized approaches aiming to solve (3.3)? In this chapter, we give a postive answer to this question by establishing an equivalence between the critical points of the factorized approach and the constrained approach. Casted informally, our main result is as follows:

> Any second-order stationary point (SOSP) of the factorized objective function $\mathsf{g}(A,B) = \mathsf{f}(AB^\top)$, must also be a fixed point of projected gradient descent on the original objective function $\mathsf{f}(X)$.

We will see in the following that, as a consequence of this result, these two approaches (factorized and constrained), treated more or less separately in the literature, can in fact be considered to be equivalent for a wide class of low-rank optimization problems, and thus lead to the same guarantees in a range of settings.

Notation Throughout this chapter, $\mathsf{f} : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a twice-differentiable objective function. Its gradient $\nabla\mathsf{f}(X)$ is represented as a matrix in $\mathbb{R}^{m\times n}$ while its second derivative $\nabla^2\mathsf{f}(X) : \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R}$ will be written as a quadratic form, i.e., $\nabla^2\mathsf{f}(X)\big(X_1, X_2\big)$. We will work also with $\mathsf{g}(A,B) = \mathsf{f}(AB^\top)$, the function defining the factorized problem. Writing $\mathsf{g} : \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r} \to \mathbb{R}$, the first derivative $\nabla\mathsf{g}(A,B) = \big(\nabla_A \mathsf{g}(A,B), \nabla_B \mathsf{g}(A,B)\big)$ lies in $\mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$, while the second derivative $\nabla^2\mathsf{g}(A,B)$ is a quadratic form mapping from $\big(\mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}\big) \times \big(\mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}\big)$ to $\mathbb{R}$. For a matrix X, we write, respectively, $\|X\|_{\mathrm{F}}$ and $\|X\|$ to denote the Frobenius norm and the

spectral norm, while $\|X\|_{2,\infty}$ will be denoted as the largest ℓ_2 norm of any row. The ℓ_0 norm, $\|\cdot\|_0$, will denote the number of nonzero entries in a vector. If $\mathrm{rank}(X) \leq r$, we will write $X = U_X \cdot \mathrm{diag}\{\sigma_1, \ldots, \sigma_r\} \cdot V_X^\top$ to denote a (possibly non-unique) singular value decomposition of X, with $\sigma_1 \geq \cdots \geq \sigma_r$.

3.2 Main result

We now turn to our main result, relating critical points of factorized optimization of $\mathsf{g}(A,B) = \mathsf{f}(AB^\top)$ to the fixed points of PGD on the full-dimensional problem $\mathsf{f}(X)$. Before proceeding, we need one additional piece of notation that allow us to quantify the smoothness of f on the space of low-rank matrices:

$$\beta_{\mathsf{local}}(X) = \lim_{\varepsilon \to 0} \left\{ \sup_{\substack{0 < \|Y-X\|_{\mathrm{F}} \leq \varepsilon \\ \mathrm{rank}(Y) \leq r}} \frac{\mathsf{f}(Y) - \mathsf{f}(X) - \langle \nabla \mathsf{f}(X), Y - X \rangle}{\frac{1}{2}\|X - Y\|_{\mathrm{F}}^2} \right\}. \tag{3.4}$$

Note that, if f is twice differentiable, then $\beta_{\mathsf{local}}(X) \leq \|\nabla^2 \mathsf{f}(X)\|$. This local curvature measure will relate to the step size of PGD, since the step size for PGD is typically chosen with respect to the curvature of f—in particular, if the second derivative of f is globally bounded by some β, then a constant step size $\eta \leq 1/\beta$ ensures that each step of PGD will make progress towards minimizing f.

As a starting point in establishing relations between the critical points of the constrained and the factorized approach, the following inclusion is straightforward:

Lemma 3.2.1. *Let $(A,B) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$. If $X = AB^\top$ is a critical point of $\min_{\mathrm{rank}(X) \leq r} \mathsf{f}(X)$, then the pair (A,B) is a FOSP of the factorized objective function $\mathsf{g}(A,B)$.*

In words, any fixed point of the PGD is a first-order stationary point (FOSP) of the factorized objective function. Our main theoretical result to follow establish a partial converse to this inclusion, proving that any second-order stationary point (SOSP) of the factorized objective function $\mathsf{g}(A,B)$ must also be a fixed point of projected gradient descent on the original function $\mathsf{f}(X)$.

Theorem 3.2.1. *Assume that* f *is twice differentiable, and let* $(A,B) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{n\times r}$.

(a) If (A,B) *is a SOSP of the factorized objective function* $\mathsf{g}(A,B)$*, then* $X = AB^\top$ *is a fixed point of the projected gradient descent algorithm on* $\min_{\mathrm{rank}(X)\le r} \mathsf{f}(X)$ *with any step size* $\eta \le 1/\beta_{\mathsf{local}}(X)$.

(b) Conversely, if (A,B) *is not a SOSP of* g*, then* $X = AB^\top$ *is not a local minimum of* $\min_{\mathrm{rank}(X)\le r} \mathsf{f}(X)$.

This result, together with some other readily obtained relations, actually establishes the following chain of inclusion for a second differentiable f:

$$\left\{\begin{array}{c}\text{Local minima}\\ \text{of } \min\limits_{\mathrm{rank}(X)\le r} \mathsf{f}(X)\end{array}\right\} \subseteq \left\{\begin{array}{c}AB^\top \text{ for SOSPs}\\ (A,B) \text{ of } \mathsf{g}(A,B)\end{array}\right\} \subseteq \left\{\begin{array}{c}\text{Fixed pts. of PGD}\\ \text{on } \min\limits_{\mathrm{rank}(X)\le r} \mathsf{f}(X)\\ \text{with } \eta \le 1/\beta_{\mathsf{local}}\end{array}\right\} \subseteq \left\{\begin{array}{c}\text{Critical pts.}\\ \text{of } \min\limits_{\mathrm{rank}(X)\le r} \mathsf{f}(X)\end{array}\right\} \subseteq \left\{\begin{array}{c}AB^\top \text{ for FOSPs}\\ (A,B) \text{ of } \mathsf{g}(A,B)\end{array}\right\}.$$

3.3 Convergence guarantees

In this section, we investigate the implications of our main result Theorem 3.2.1 on the landscape of the factorized problem (3.3). Before stating our explicit result, some discussion on the different type of guarantee and different type of assumption is needed. From strongest to weakest, the three main styles of guarantees that appear in the literature are:

- Global optimality: the algorithm converges to a global minimizer.
- Local optimality, or basin of attraction: if initialized near a global minimizer, then the algorithm converges to that global minimizer.
- Restricted optimality: the algorithm converges to a matrix X that satisfies $\mathsf{f}(X) \le \mathsf{f}(X')$ for any rank-r' matrix X', where $r' < r$ is a strictly lower rank constraint.

To simplify our comparison of these three styles of guarantees, we will consider the setting where the original objective function f satisfies the α-RSC condition and the β-RSM condition defined

in Section 2.2 with respect to the rank constraint r. We recall the definition here for completeness. We say that f satisfies α-RSC with respect to the rank constraint r if for all $X, Y \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(X), \operatorname{rank}(Y) \le r$,

$$\mathsf{f}(Y) \ge \mathsf{f}(X) + \langle \nabla \mathsf{f}(X), Y - X\rangle + \frac{\alpha}{2}\|X - Y\|_{\mathrm{F}}^2. \tag{3.5}$$

We say f satisfies β-RSM with respect to the rank constraint r if for all X, Y with $\operatorname{rank}(X), \operatorname{rank}(Y) \le r$,

$$\mathsf{f}(Y) \le \mathsf{f}(X) + \langle \nabla \mathsf{f}(X), Y - X\rangle + \frac{\beta}{2}\|X - Y\|_{\mathrm{F}}^2. \tag{3.6}$$

Throughout this section, we will always write $\kappa = \beta/\alpha$ to denote the rank-restricted condition number of f. Note that $\kappa \ge 1$ always. We will consider two different regimes for the condition number κ:

$$\text{Near-isometry } (\kappa \approx 1) \quad \text{vs.} \quad \text{Arbitrary conditioning } (\kappa \gg 1).$$

We can expect to see $\kappa \approx 1$ in certain well-behaved problems, for instance the matrix sensing problem, where $\mathsf{f}(X)$ represents matching X with random linear measurements of the form $\langle A_i, X\rangle$, where e.g., the measurement matrices A_i have i.i.d. entries. In general, however, most problems do not have $\kappa \approx 1$.

In some cases, the restricted strong convexity and/or restricted smoothness conditions might not be satisfied globally (i.e., for all rank-r matrices), but is satisfied for a more restricted subset of matrices X, Y; in these settings we may write, for instance, that f satisfies α-RSC over a particular subset.

We also need to consider a second important distinction between different classes of problems. In many statistical settings, we may have an objective function $\mathsf{f}(X)$ that comes from a data likelihood, where $\mathbb{E}[\mathsf{f}(X)]$ is minimized at some true low-rank parameter matrix $X_\star$. When this is the case, it is common to see $\|\nabla \mathsf{f}(X_\star)\| \approx 0$. In other settings, though, there might not be any natural underlying low-rank structure, and the gradient $\nabla \mathsf{f}(X)$ is large at any low-rank X. We will therefore

distinguish between two scenarios:

$$\text{Vanishing gradient } (\min_{\text{rank}(X)\leq r} \|\nabla \mathsf{f}(X)\| \approx 0) \quad \text{vs.} \quad \text{Arbitrary gradient } (\min_{\text{rank}(X)\leq r} \|\nabla \mathsf{f}(X)\| \gg 0).$$

3.3.1 Existing results

We now summarize the existing results as well as our own findings, for the different types of assumptions and different styles of guarantees outlined above:

- Near-isometry + Vanishing gradient $\Rightarrow$ Global optimality.

  For the most well-behaved problems, where the objective function $\mathsf{f}(X)$ exhibits both near-isometry and a vanishing gradient, it is possible to prove convergence to an (approximate) globally optimal estimate $\widehat{X}$. For full-dimensional projected gradient descent algorithm, this has been established in the case of a least squares objective [29]; for factorized algorithms, an analogous result (no spurious local minima) has been established for certain least squares objectives [30, 31, 32, 33, 34] and more generally for functions f with a near-isometry property [35]. (We will show in the present work that under near-isometry + vanishing gradient, both full-dimensional and factorized approaches contain no spurious local minima.)

- Arbitrary conditioning + Vanishing gradient $\Rightarrow$ Local optimality.

  With a non-ideal condition number $\kappa > 1$, assuming a vanishing gradient condition is sufficient to prove a local optimality result, or the existence of basin of attraction, both for full-dimensional PGD [36] and for factorized approaches [7]; in the stronger setting of a near-isometry and a vanishing gradient, the local optimality result for factorized approaches has been also established by many works, including Candes et al. [37], Zheng and Lafferty [38, 39], Tu et al. [40], Bhojanapalli et al. [41], Jain et al. [42]. Note that all of the previous local optimality results for factorized problems are built upon identifying local region of attraction for globally optimal solution $\widehat{X}$ in the *factorized* space (A,B). (We will give in the present work the local region of attraction in the *full-dimensional* representations $X = AB^\top$.)

- Arbitrary conditioning + Arbitrary gradient $\Rightarrow$ Restricted optimality.

  In the most challenging setting, where we allow both arbitrary condition number κ and an arbitrarily large gradient, restricted optimality guarantees can still be obtained. This is established for the full-dimensional PGD algorithm [5, 2], as well as its variants, such as approximate low-rank projection [43, 44], and projection with debiasing step [45]; for sparse problems specifically, the analogous restricted optimality result has been established [46]. On the other hand, there is no known result for restricted optimality guarantees within the factorized approach. (We will show in the present work that it holds also for the factorized approach.)

This extensive literature has enabled us to understand the landscape of the nonconvex low-rank optimization problem, but the various results have been proved somewhat disjointly, using very different techniques for analyzing full-dimensional PGD type algorithms versus factorized algorithms. It is natural to ask whether this collection of results can be unified into a single framework. Our main result, Theorem 3.2.1, allows us to connect established results between PGD algorithms and factorized algorithms, allowing us to establish simpler proofs of some existing results, and provide new results in certain settings. Overall, it is the goal of this section to provide a broader view of the landscape of results known for low-rank optimization problems through the lens of the equivalence between PGD and factorized algorithms established in Theorem 3.2.1.

3.3.2 Results for global and local optimality

In the special case of least squares objective, i.e., $\mathsf{f}(X) = \frac{1}{2}\|\mathscr{A}(X) - b\|_{\mathrm{F}}^2$ for a linear operator $\mathscr{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, Oymak et al. [29] show that, in the near-isometry setting ($\kappa \approx 1$), projected gradient descent offers a global convergence guarantee starting from any initialization point. Here we extend some of their technical tools to general functions $\mathsf{f}(X)$. We will write $\mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$ to denote the set of $m \times n$ matrices with rank $\leq r$.

Lemma 3.3.1. *Suppose that* $\mathsf{f} : \mathbb{R}^{m\times n} \to \mathbb{R}$ *satisfies* α*-RSC* (3.5) *over a subset* $\mathscr{X} \subseteq \mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$,

where $\alpha > 0$. If $X_0, X_1 \in \mathscr{X}$ are both fixed points of PGD run with step size $\eta_0 > 0$ or $\eta_1 > 0$, respectively, then one of the following must hold:

- $X_0 = X_1$, *or*
- $\operatorname{rank}(X_0) = r$ *and* $\operatorname{rank}(X_1) < r$ *and* $\frac{\|\nabla \mathsf{f}(X_0)\|}{\sigma_r(X_0)} \geq 2\alpha$, *or*
- $\operatorname{rank}(X_1) = r$ *and* $\operatorname{rank}(X_0) < r$ *and* $\frac{\|\nabla \mathsf{f}(X_1)\|}{\sigma_r(X_1)} \geq 2\alpha$, *or*
- $\operatorname{rank}(X_0) = \operatorname{rank}(X_1) = r$ *and*

$$\frac{\|\nabla \mathsf{f}(X_0)\|}{\sigma_r(X_0)} + \frac{\|\nabla \mathsf{f}(X_1)\|}{\sigma_r(X_1)} \geq 2\alpha.$$

We also verify a simple result:

Lemma 3.3.2. *Suppose that $\mathsf{f} : \mathbb{R}^{m\times n} \to \mathbb{R}$ satisfies β-RSM (3.6) over an open subset $\mathscr{X} \subseteq \mathbb{R}^{m\times n}_{\operatorname{rank}(r)}$. If $\widehat{X}$ is a global minimizer (i.e., $\mathsf{f}(\widehat{X}) = \min_{\operatorname{rank}(X)\leq r} \mathsf{f}(X)$) and $\widehat{X} \in \mathscr{X}$, then $\widehat{X}$ is a fixed point of projected gradient descent run with rank constraint r and any step size $\eta \leq 1/\beta$.*

These lemmas will allow us to easily prove global optimality and local optimality results under the appropriate assumptions. We now turn to the question of obtaining global and local optimality results for PGD and factorized algorithms. While results of this flavor are already known in the literature (see Section 3.3.1 for some references), our goal here is to give extremely short and clean proofs that illuminate the connection between the full-dimensional and factorized representations of the optimization problem, and thereby also highlight the utility of our main result, Theorem 3.2.1. In some cases, our work also establishes guarantees in a broader setting than previous results.

Global optimality

In the setting where $\mathsf{f}(X)$ satisfies the near-isometry property, with condition number $\kappa < 2$, we can obtain global optimality guarantees for both PGD and factorized methods whenever $\|\nabla \mathsf{f}(X)\|$

is sufficiently small, i.e., the vanishing gradient condition. (See Section 3.3.1 for related existing results in the literature.)

Theorem 3.3.1. *Assume that* $\mathsf{f}(X)$ *satisfies* α*-RSC* (3.5) *and* β*-RSM* (3.6) *over an open subset* $\mathscr{X} \subseteq \mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$*, and that* $\beta < 2\alpha$*. Suppose* $\widehat{X}$ *is a global minimizer, i.e.,* $\mathsf{f}(\widehat{X}) = \min_{\mathrm{rank}(X)\le r} \mathsf{f}(X)$*. If* $\widehat{X} \in \mathscr{X}$ *and* $\widehat{X}$ *satisfies*

$$\text{Either } \mathrm{rank}(\widehat{X}) < r, \text{ or } \mathrm{rank}(\widehat{X}) = r \text{ and } \|\nabla \mathsf{f}(\widehat{X})\| < (2\alpha-\beta)\cdot\sigma_r(\widehat{X})\ ,$$

then

- $\widehat{X}$ *is the unique fixed point of PGD in* $\mathscr{X}$ *for any step size* $1/(2\alpha) < \eta \le 1/\beta$ *in the case that* $\mathrm{rank}(\widehat{X}) < r$*, or in the case* $\mathrm{rank}(\widehat{X}) = r$*, for any step size satisfying*

$$\frac{1}{2\alpha - \frac{\|\nabla \mathsf{f}(\widehat{X})\|}{\sigma_r(\widehat{X})}} < \eta \le \frac{1}{\beta}. \tag{3.7}$$

- *If* $X = AB^\top \in \mathscr{X}$ *where* (A,B) *is a SOSP of* $\mathsf{g}(A,B)$*, then* $X = \widehat{X}$*.*

Note that, in the case that $\mathrm{rank}(\widehat{X}) = r$, due to the condition $\|\nabla \mathsf{f}(\widehat{X})\| < (2\alpha-\beta)\cdot\sigma_r(\widehat{X})$ the interval (3.7) given for step size η is always non-empty. Theorem 3.3.1 proves that global optimality guarantees can be achieved as long as $\kappa < 2$, i.e., the map f is a near-isometry. This type of assumption on κ is crucial to achieving global optimality guarantees. For instance, Zhang et al. [47, Example 3] construct an example of objective function $\mathsf{f}(X)$ with $\beta = 3\alpha$, i.e., $\kappa = 3$, where there exists a fixed point X that is *not* globally optimal. This proves that $\kappa < 3$ is necessary for achieving a global optimality guarantee, while our work shows $\kappa < 2$ is sufficient. While it is not the goal of the present work, an interesting open question is to close the gap between these necessary and sufficient conditions to identify an exact correspondence between condition number and the global optimality guarantee; see also Zhang et al. [48] for the sufficient and necessary conditions when rank $r = 1$.

Local optimality

Next we turn to the local optimality guarantees, i.e., the existence of local region of attraction, that can be obtained when f exhibits a vanishing gradient, but may have an arbitrarily large condition number κ. (See Section 3.3.1 for related existing results in the literature.)

Theorem 3.3.2. *Assume that* $\mathsf{f}(X)$ *satisfies* α*-RSC* (3.5) *over a subset* $\mathscr{X} \subseteq \mathbb{R}^{m\times n}_{\mathrm{rank}(r)}$. *Assume that* $\widehat{X}$ *is a global minimizer, i.e.,* $\mathsf{f}(\widehat{X}) = \min_{\mathrm{rank}(X)\leq r} \mathsf{f}(X)$, *that* $\widehat{X} \in \mathscr{X}$, *and that* $\widehat{X}$ *satisfies*

$$\text{Either } \mathrm{rank}(\widehat{X}) < r, \text{ or } \mathrm{rank}(\widehat{X}) = r \text{ and } \|\nabla \mathsf{f}(\widehat{X})\| < \alpha \cdot \sigma_r(\widehat{X}) \ .$$

Let

$$\mathscr{N} = \left\{ X \in \mathscr{X} : \mathrm{rank}(X) < r \text{ or } \frac{\|\nabla \mathsf{f}(X)\|}{\sigma_r(X)} < 2\alpha \right\}$$

in the case that $\mathrm{rank}(\widehat{X}) < r$, *or*

$$\mathscr{N} = \left\{ X \in \mathscr{X} : \mathrm{rank}(X) < r \text{ or } \frac{\|\nabla \mathsf{f}(\widehat{X})\|}{\sigma_r(\widehat{X})} + \frac{\|\nabla \mathsf{f}(X)\|}{\sigma_r(X)} < 2\alpha \right\}$$

in the case that $\mathrm{rank}(\widehat{X}) = r$. *Then:*

- *For any fixed point* X *of PGD with any step size* $\eta > 0$, *if* $X \in \mathscr{N}$, *then* $X = \widehat{X}$.
- *If* $X = AB^\top \in \mathscr{N}$ *where* (A,B) *is a SOSP of* $\mathsf{g}(A,B)$, *then* $X = \widehat{X}$.

We note that $\widehat{X} \in \mathscr{N}$ by the assumptions of the theorem. In this setting where κ may be arbitrarily large, global optimality does not hold in general (as shown by Zhang et al. [47]'s counterexample, discussed in Section 3.3.2 above). Nonetheless, the results in Theorem 3.3.2 still ensure the existence of regions of attraction $\mathscr{N}$ within which the global minimum $\widehat{X}$ will be discovered, for both the full-dimensional and factorized methods.

To compare with the existing results, the first part of Theorem 3.3.2 (for fixed points of PGD) is an immediate result given the work in Barber and Ha [36]. Next, turning to the second part of the result, on the SOSPs of the factorized approach, some related results in the existing literature

have shown that certain rank-constrained problems exhibit local region of attraction near the global minimum $\widehat{X}$ [37, 38, 39, 40, 41, 42]. While these problems satisfy the near-isometry property with $\kappa \approx 1$, our result in Theorem 3.3.2 extends to a broader setting with an arbitrarily large condition number κ. Chen and Wainwright [7] have also established local convergence guarantees under conditions similar to restricted strong convexity and smoothness, but the difference is that they work with RSC and RSM type conditions defined directly on the factorized variable pair (A, B). In addition, many of these works address the positive semidefinite setting, $X = AA^\top$, rather than the generic setting $X = AB^\top$ considered here.

3.3.3 A restricted optimality guarantee

In this last setting, we will make no assumptions on either the gradient or the condition number, i.e., it may be possible that $\|\nabla \mathsf{f}(\widehat{X})\|$ is large and the condition κ is large as well. (See Section 3.3.1 for related existing results in the literature.)

Under such assumptions, to the best of our knowledge, there is no guaranteed result to solve the low-rank minimization problem either locally or globally—identifying a region of attraction in a deterministic way is a nontrivial task. Therefore, we may wish to instead establish a weaker *restricted optimality* guarantee, which entails proving that the algorithm converges to some matrix X satisfying

$$\mathsf{f}(X) \leq \min_{\operatorname{rank}(Y) \leq r'} \mathsf{f}(Y),$$

where the rank $r' < r$ proves a more restrictive constraint. In a statistical setting where we are aiming to recover some true low-rank parameter, we might think of r' as the true underlying rank, while $r \geq r'$ is a relaxed rank constraint that we place on our optimization scheme. More generally, we are simply aiming to show that optimizing over rank r, while not ensuring the best rank-r solution, is competitive with the best lower-rank solution.

Under these conditions, Theorem 2.5.1 and equation (2.18) together prove that any fixed point X of PGD with step size $\eta = 1/\beta$ satisfies restricted optimality with respect to any rank $r' < r/\kappa^2$.

Based on our main result, Theorem 3.2.1, the same guarantee also holds for any SOSP of the factorized problem. For completeness, we restate their result along with the new extension to the factorized problem:

Theorem 3.3.3. *Assume that* $\mathsf{f}(X)$ *satisfies* α*-RSC* (3.5) *and* β*-RSM* (3.6) *over an open subset* $\mathscr{X} \subseteq \mathbb{R}^{m\times n}_{\text{rank}(r)}$. *Let* $\kappa = \beta/\alpha$. *Then:*

- *[2] For any fixed point* $X \in \mathscr{X}$ *of PGD with step size* $\eta = 1/\beta$,

$$\mathsf{f}(X) \leq \min_{\text{rank}(Y)<r/\kappa^2, Y\in\mathscr{X}} \mathsf{f}(Y), \tag{3.8}$$

  i.e., X satisfies restricted optimality with respect to any rank $r' < r/\kappa^2$ *within* $\mathscr{X}$.

- *For any* $X = AB^\top \in \mathscr{X}$ *where* (A,B) *is a SOSP of the factorized problem* $\mathsf{g}(A,B)$,

$$\mathsf{f}(AB^\top) \leq \min_{\text{rank}(Y)<r/\kappa^2, Y\in\mathscr{X}} \mathsf{f}(Y),$$

  i.e., $X = AB^\top$ *satisfies restricted optimality with respect to any rank* $r' < r/\kappa^2$ *within* $\mathscr{X}$.

Conversely, Theorem 2.5.2 and equation (2.18) also establish that this factor of κ^2 is sharp in general (on all low-rank matrices), i.e., restricted optimality cannot be guaranteed relative to rank $r' > r/\kappa^2$. Here we establish the analogous result for the factorized problem. For completeness, we state the two results together.

Theorem 3.3.4. *For any parameters* $\beta \geq \alpha > 0$ *and any rank* $r' > r/\kappa^2$, *there exists a function* $\mathsf{f} : \mathbb{R}^{m\times n} \to \mathbb{R}$ *satisfying* α*-RSC* (3.5) *and* β*-RSM* (3.6) *over* $\mathbb{R}^{m\times n}_{\text{rank}(r)}$, *such that:*

- *[2] There exists a fixed point X of PGD with step size* $\eta = 1/\beta$, *such that*

$$\mathsf{f}(X) > \min_{\text{rank}(Y)\leq r'} \mathsf{f}(Y).$$

- *There exists a second-order stationary point* (A,B) *of the factorized problem, such that*

$$\mathsf{f}(AB^\top) > \min_{\mathrm{rank}(Y)\leq r'} \mathsf{f}(Y).$$

Unlike the restricted optimality guarantee above (Theorem 3.3.3), this converse result does not follow directly from Theorem 2.5.2, and instead requires a new construction.

CHAPTER 4

MINIMAX RATES IN SPARSE, HIGH-DIMENSIONAL CHANGEPOINT DETECTION

In this chapter we change gear to a statistical problem where the sparsity interacts with the changepoint structure. In particular, we study the detection of a sparse change in a high-dimensional mean vector and establish the worst-case difficulty of the problem.[1]

4.1 Introduction

The problem of changepoint detection has a long history [e.g. 50], but has undergone a remarkable renaissance over the last 5–10 years. This has been driven in part because these days sensors and other devices collect and store data on unprecedented scales, often at high frequency, which has placed a greater emphasis on the running time of changepoint detection algorithms [51, 52]. But it is also because nowadays these data streams are often monitored simultaneously as a multidimensional process, with a changepoint in a subset of the coordinates representing an event of interest. Examples include distributed denial of service attacks as detected by changes in traffic at certain internet routers [53] and changes in a subset of blood oxygen level dependent contrast in a subset of voxels in fMRI studies [54]. Away from time series contexts, the problem is also of interest, for instance in the detection of chromosomal copy number abnormality [55, 56]. Key to the success of changepoint detection methods in such settings is the ability to borrow strength across the different coordinates, in order to be able to detect much smaller changes than would be possible through observation of any single coordinate in isolation.

We consider the model where, for some $n \geq 2$, we observe a $p \times n$ matrix X that can be written as

$$X = \theta + E, \tag{4.1}$$

1. The work presented in this chapter can be found in [49] and is to appear in the Annals of Statistics. It is joint with Chao Gao and Richard Samworth.

where $\theta \in \mathbb{R}^{p\times n}$ is deterministic and the entries of E are independent $N(0,1)$ random variables. We wish to test the null hypothesis that the columns of θ are constant against the alternative that there exists a time $t_0 \in \{1,\dots,n-1\}$ at which these mean vectors change, in at most s out of the p coordinates. The difficulty of this problem is governed by a signal strength parameter ρ^2 that measures the squared Euclidean norm of the difference between the mean vectors, rescaled by $\frac{t_0(n-t_0)}{n}$; this latter quantity can be interpreted as an effective sample size. The goal is to identify the minimax testing rate in ρ^2 as a function of the problem parameters p, n and s, and we denote this by $\rho^*(p,n,s)^2$; this is the signal strength at which we can find a test making the sum of the Type I and Type II error probabilities arbitrarily small by choosing ρ^2 to be an appropriately large multiple of $\rho^*(p,n,s)^2$ (where the multiple is not allowed to depend on p, n and s), and at which any test has error probability sum arbitrarily close to 1 for a suitably small multiple of $\rho^*(p,n,s)^2$.

We are going to see that, despite the seemingly simplicity of the model, it already captures many subtleties of the interaction between the sparsity and the changepoint structure. In speicifc, we reveal a particularly subtle form of the exact minimax testing rate in the above problem, namely

$$\rho^*(p,n,s)^2 \asymp \begin{cases} \sqrt{p\log\log(8n)} & \text{if } s \geq \sqrt{p\log\log(8n)}, \\ s\log\left(\frac{ep\log\log(8n)}{s^2}\right) \vee \log\log(8n) & \text{if } s < \sqrt{p\log\log(8n)}. \end{cases}$$

This result provides a significant generalization of two known special cases in the literature, namely $\rho^*(1,n,1)^2$ and $\rho^*(p,2,s)^2$; see Section 4.2 for further discussion. Although our initial optimal testing procedure depends on the sparsity level s, which would often be unknown in practice, we show in Theorem 4.3.2 that it is possible to construct an adaptive test that achieves exactly the same rate (but is a little more complicated to describe). Exact asymptotic constant are also obtained for the dense regime.

Related work Most prior work on multivariate changepoint detection has proceeded without a sparsity condition and in an asymptotic regime with n growing to infinity with the dimension fixed, including Basseville and Nikiforov [57], Csörgő and Horváth [58], Ombao et al. [59], Aue

et al. [60], Kirch et al. [61], Zhang et al. [55] and Horváth and Hušková [62]. Bai [63] studied the least squares estimator of a change in mean for high-dimensional panel data. Jirak [64], Cho and Fryzlewicz [65], Cho [66] and Wang and Samworth [56] have all proposed CUSUM-based methods for the estimation of the location of a sparse, high-dimensional changepoint. Aston and Kirch [67] introduce a notion of efficiency that quantifies the detection power of different statistics in high-dimensional settings. Enikeeva and Harchaoui [68] study the sparse changepoint detection problem in an asymptotic regime in which $p \to \infty$, and at the same time $s \to \infty$ with $s/p \to \infty$ and the sample size not too large; we compare their results with ours in Section 4.3.3. Further related work on high-dimensional changepoint problems include the detection of changes in covariance [e.g. 60, 69, 70] and in sparse dynamic networks [71]. We emphasize that in this work we focus entirely on the offline version of the changepoint testing problem, where the entire data stream is observed prior to the statistician attempting to determine whether or not a change in mean has occurred. For recent work on the corresponding online problem, where the data are observed sequentially and one wishes to declare a change as soon as possible after it has occured, see, e.g., Xie and Siegmund [72] and Chen et al. [73].

Notation For $d \in \mathbb{N}$, we write $[d] := \{1, \ldots, d\}$. Given $a, b \in \mathbb{R}$, we write $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$. We also write $a \lesssim b$ to mean that there exists a universal constant $C > 0$ such that $a \leq Cb$; moreover, $a \asymp b$ means $a \lesssim b$ and $b \lesssim a$. For a set S, we use $\mathbb{1}_S$ and $|S|$ to denote its indicator function and cardinality respectively. For a vector $v = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, we define the norms $\|v\|_1 := \sum_{\ell=1}^d |v_\ell|$, $\|v\|^2 := \sum_{\ell=1}^d v_\ell^2$ and $\|v\|_\infty := \max_{\ell \in [d]} |v_\ell|$, and also define $\|v\|_0 := \sum_{\ell=1}^d \mathbb{1}_{\{v_\ell \neq 0\}}$. Given two vectors $u, v \in \mathbb{R}^d$ and a positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, we define $\langle u, v \rangle_\Sigma^{-1} := u^T \Sigma^{-1} v$ and $\|v\|_{\Sigma^{-1}} := (v^T \Sigma^{-1} v)^{1/2}$ and omit the subscripts when $\Sigma = I_d$. More generally, the trace inner product of two matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$ is defined as $\langle A, B \rangle := \sum_{\ell=1}^{d_1} \sum_{\ell'=1}^{d_2} A_{\ell\ell'} B_{\ell\ell'}$, while the Frobenius and operator norms of A are given by $\|A\|_{\mathrm{F}} := \sqrt{\langle A, A \rangle}$ and $\|A\|_{\mathrm{op}} := s_{\max}(A)$ respectively, where $s_{\max}(\cdot)$ denotes the largest singular value. The total variation distance between two probability measures P and Q on a measurable space $(\mathscr{X}, \mathscr{A})$

is defined as $\mathsf{TV}(P,Q) := \sup_{A\in\mathscr{A}} |P(A) - Q(A)|$. Moreover, if P is absolutely continuous with respect to Q, then the Kullback–Leibler divergence is defined as $D(P\|Q) := \int_{\mathscr{X}} \log \frac{dP}{dQ}\, dP$, and the chi-squared divergence is defined as $\chi^2(P\|Q) := \int_{\mathscr{X}} \left(\frac{dP}{dQ} - 1\right)^2 dQ$. The notation $\mathbb{P}$ and $\mathbb{E}$ are generic probability and expectation operators whose distribution is determined from the context.

4.2 Problem formulation

Recall that we consider the observation of a $p \times n$ matrix $X = \theta + E$, where $n \geq 2$, where θ is deterministic and where each entry of the error matrix E is an independent $N(0,1)$ random variable. In other words, writing X_t and θ_t for the tth columns of X and θ respectively, we have $X_t \sim N_p(\theta_t, I_p)$ independently across t. The goal is to test whether or not the sequence $\{\theta_t\}_{t\in[n]}$ has a changepoint. We define the parameter space of signals without a changepoint by

$$\Theta_0(p,n) := \left\{\theta \in \mathbb{R}^{p\times n} : \theta_t = \mu \text{ for some } \mu \in \mathbb{R}^p \text{ and all } t \in [n]\right\}.$$

For $s \in [p]$ and $\rho > 0$, the space consisting of signals with a sparse structural change at time $t_0 \in [n-1]$ is defined by

$$\begin{aligned}\Theta^{(t_0)}(p,n,s,\rho) := \Big\{\theta = (\theta_1,\ldots,\theta_n) \in \mathbb{R}^{p\times n} :& \\ \theta_t = \mu_1 \text{ for some } \mu_1 \in \mathbb{R}^p \text{ for all } 1 \leq t \leq t_0,& \\ \theta_t = \mu_2 \text{ for some } \mu_2 \in \mathbb{R}^p \text{ for all } t_0+1 \leq t \leq n,& \\ \|\mu_1 - \mu_2\|_0 \leq s, \frac{t_0(n-t_0)}{n}\|\mu_1-\mu_2\|^2 \geq \rho^2\Big\}.&\end{aligned}$$

In the definition of $\Theta^{(t_0)}(p,n,s,\rho)$, the parameters p and n determine the size of the problem, while t_0 is the location of the changepoint. The quantities s and ρ parametrize the sparsity level and the magnitude of the structural change respectively. It is worth noting that $\|\mu_1 - \mu_2\|^2$ is normalized by the factor $\frac{t_0(n-t_0)}{n}$, which plays the role of the effective sample size of the problem. To understand this, consider the problem of testing the changepoint at location t_0 when $p = 1$. Then the natural

test statistic is

$$\frac{1}{t_0}\sum_{t=1}^{t_0} X_t - \frac{1}{n-t_0}\sum_{t=t_0+1}^{n} X_t,$$

whose variance is $\frac{n}{t_0(n-t_0)}$. Hence the difficulty of changepoint detection problem depends on the location of the changepoint. Through the normalization factor $\frac{t_0(n-t_0)}{n}$, we can define a common signal strength parameter ρ across different possible changepoint locations. Taking a union over all such changepoint locations, the alternative hypothesis parameter space is given by

$$\Theta(p,n,s,\rho) := \bigcup_{t_0=1}^{n-1} \Theta^{(t_0)}(p,n,s,\rho).$$

We will address the problem of testing the two hypotheses

$$H_0 : \theta \in \Theta_0(p,n), \qquad H_1 : \theta \in \Theta(p,n,s,\rho). \tag{4.2}$$

To this end, we let Ψ denote the class of possible test statistics, i.e. measurable functions $\psi : \mathbb{R}^{p\times n} \to [0,1]$. We also define the minimax testing error by

$$\mathscr{R}(\rho) := \inf_{\psi\in\Psi}\left\{ \sup_{\theta\in\Theta_0(p,n)} \mathbb{E}_\theta \psi(X) + \sup_{\theta\in\Theta(p,n,s,\rho)} \mathbb{E}_\theta\big(1-\psi(X)\big)\right\},$$

where we use $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$ to denote probabilities and expectations under the data generating process (4.1). Our goal is to determine the order of the minimax rate of testing in this problem, as defined below.

Definition 4.2.1. We say $\rho^* = \rho^*(p,n,s)$ is the minimax rate of testing if the following two conditions are satisfied:

1. For any $\varepsilon \in (0,1)$, there exists $C_\varepsilon > 0$, depending only on ε, such that $\mathscr{R}(C\rho^*) \le \varepsilon$ for any $C > C_\varepsilon$.

2. For any $\varepsilon \in (0,1)$, there exists $c_\varepsilon > 0$, depending only on ε, such that $\mathscr{R}(c\rho^*) \ge 1-\varepsilon$ for any $c \in (0, c_\varepsilon)$.

Special cases Some special cases of $\rho^*(p,n,s)$ are well understood in the literature. For instance, when $p = s = 1$, we recover the one-dimensional changepoint detection problem. [74] showed that

$$\rho^*(1,n,1)^2 \asymp \log\log(8n). \tag{4.3}$$

The rate (4.3) involves an iterated logarithmic factor, in constrast to a typical logarithmic factor in the minimax rate of sparse signal detection [e.g., 75, 76, 77]. Another solved special case is when $n = 2$. In this setting, we observe $X_1 \sim N_p(\mu_1, I_p)$ and $X_2 \sim N_p(\mu_2, I_p)$, and the problem is to test whether or not $\mu_1 = \mu_2$. Since $X_1 - X_2$ is a sufficient statistic for $\mu_1 - \mu_2$, the problem can be further reduced to a sparse signal detection problem in a Gaussian sequence model. For this problem, [78] established the minimax detection boundary

$$\rho^*(p,2,s)^2 \asymp \begin{cases} \sqrt{p} & \text{if } s \geq \sqrt{p} \\ s\log\left(\frac{ep}{s^2}\right) & \text{if } s < \sqrt{p}. \end{cases} \tag{4.4}$$

It is interesting to notice the elbow effect in the rate (4.4). Above the sparsity level of $\sqrt{p}$, one obtains the parametric rate that can be achieved using the test that rejects H_0 if $\|X_1 - X_2\|_2^2 > 2p + c\sqrt{p}$ for an appropriate $c > 0$. It is straightforward to extend both rates (4.3) and (4.4) to cases where either p or n is of a constant order. However, the general form of $\rho^*(p,n,s)$ is unknown in the statistical literature.

4.3 Minimax detection boundary

The main result of this chapter is given by the following theorem.

Theorem 4.3.1. *The minimax rate of the detection boundary of the problem (4.2) is given by*

$$\rho^*(p,n,s)^2 \asymp \begin{cases} \sqrt{p\log\log(8n)} & \text{if } s \geq \sqrt{p\log\log(8n)} \\ s\log\left(\frac{ep\log\log(8n)}{s^2}\right) \vee \log\log(8n) & \text{if } s < \sqrt{p\log\log(8n)}. \end{cases} \tag{4.5}$$

It is important to note that the minimax rate (4.5) is not a simple sum or multiplication of the rates (4.3) and (4.4) for constant p or n. The high-dimensional changepoint detection problem differs fundamentally from both its low-dimensional version and the sparse signal detection problem.

We observe that the minimax rate exhibits the two regimes in (4.5) only when $p \geq \log\log(8n)$, since if $p < \log\log(8n)$, then the condition $s \geq \sqrt{p\log\log(8n)}$ is empty, and (4.5) has just one regime. Compared with the rate (4.4), the phase transition boundary for the sparsity s becomes $\sqrt{p\log\log(8n)}$. In fact, the minimax rate (4.5) can be obtained by first replacing the p in (4.4) with $p\log\log(8n)$, and then adding the extra term (4.3).

The dependence of (4.5) on n is very delicate. Consider the range of sparsity where

$$\frac{\log\log(8n)}{\log(e\log\log(8n))} \vee \frac{\sqrt{p}}{(\log\log(8n))^C} \lesssim s \lesssim \sqrt{p\log\log(8n)},$$

for some universal constant $C > 0$. The rate (4.5) then becomes

$$\rho^*(p,n,s)^2 \asymp s\log(e\log\log(8n)).$$

That is, it grows with n at a $\log\log\log(\cdot)$ rate. To the best of our knowledge, such a triple iterated logarithmic rate has not been found in any other problem before in the statistical literature.

Last but not least, we remark that when p or n is a constant, the rate (4.5) recovers (4.3) and (4.4) as special cases.

4.3.1 Upper bound

To derive the upper bound, we need to construct a testing procedure. We emphasize that the goal of hypothesis testing is to detect the existence of a changepoint; this is in contrast to the problem of changepoint estimation [65, 56, 79], where the goal is to find the changepoint's location.

If we knew that the changepoint were between t and $n-t+1$, it would be natural to define the

Cumulative Sum (CUSUM)-type statistic

$$Y_t := \frac{(X_1 + \ldots + X_t) - (X_{n-t+1} + \ldots + X_n)}{\sqrt{2t}}. \tag{4.6}$$

Note that the definition of Y_t does not use the observations between $t+1$ and $n-t$. This allows Y_t to detect any changepoint in this range, regardless of its location. The existence of a changepoint implies that $\mathbb{E}_\theta(Y_t) \neq 0$. Since the structural change only occurs in a sparse set of coordinates, we threshold the magnitude of each coordinate $Y_t(j)$ at level $a \geq 0$ to obtain

$$A_{t,a} := \sum_{j=1}^{p} \big\{ Y_t(j)^2 - \nu_a \big\} \mathbb{1}_{\{|Y_t(j)| \geq a\}},$$

where $\nu_a := \mathbb{E}\big(Z^2 \;\big|\; |Z| \geq a\big)$ is the conditional second moment of $Z \sim N(0,1)$, given that its magnitude is at least a. See [78] for a similar strategy for the sparse signal detection problem. Note that $A_{t,0} = \sum_{j=1}^{p} \big\{ Y_t(j)^2 - 1 \big\}$ has a centered χ^2_p distribution under H_0.

Since the range of the potential changepoint locations is unknown, a natural first thought is to take a maximum of $A_{t,a}$ over $t \in [n/2]$. It turns out, however, that in high-dimensional settings it is very difficult to control the dependence between these different test statistics at the level of precision required to establish the minimax testing rate. A methodological contribution of this work, then, is the recognition that it suffices to compute a maximum of $A_{t,a}$ over a candidate set $\mathscr{T}$ of locations, because if there exists a changepoint at time t_0 and $t_0/2 < \widetilde{t} \leq t_0$ for some $\widetilde{t} \in \mathscr{T}$, then $\|\mathbb{E}_\theta(Y_{\widetilde{t}})\|$ and $\|\mathbb{E}_\theta(Y_{t_0})\|$ are of the same order of magnitude. This observation reflects a key difference between the changepoint testing and estimation problems. To this end, we define

$$\mathscr{T} := \Big\{ 1, 2, 4, \ldots, 2^{\lfloor \log_2(n/2) \rfloor} \Big\},$$

so that $|\mathscr{T}| = 1 + \lfloor \log_2(n/2) \rfloor$. Then, for a given $r \geq 0$, the testing procedure we consider is given by

$$\psi \equiv \psi_{a,r}(X) := \mathbb{1}_{\{\max_{t \in \mathscr{T}} A_{t,a} > r\}}. \tag{4.7}$$

The theoretical performance of the test (4.7) is given by the following theorem. We use the notation $r^*(p,n,s)$ for the rate function on the right-hand side of (4.5).

Proposition 4.3.1. *For any $\varepsilon \in (0,1)$, there exists $C > 0$, depending only on ε, such that the testing procedure (4.7) with $a^2 = 4\log\left(\frac{ep\log\log(8n)}{s^2}\right)\mathbb{1}_{\{s<\sqrt{p\log\log(8n)}\}}$ and $r = Cr^*(p,n,s)$ satisfies*

$$\sup_{\theta\in\Theta_0(p,n)} \mathbb{E}_\theta \psi + \sup_{\theta\in\Theta(p,n,s,\rho)} \mathbb{E}_\theta(1-\psi) \le \varepsilon,$$

as long as $\rho^2 \ge 32Cr^*(p,n,s)$.

Just as the minimax rate (4.5) has two regimes, the testing procedure (4.7) also uses two different strategies. In the dense regime $s \ge \sqrt{p\log\log(8n)}$, we have $a^2 = 0$ and thus (4.7) becomes simply $\psi = \mathbb{1}_{\{\max_{t\in\mathscr{T}} \|Y_t\|^2 - p > r\}}$. In the sparse regime $s < \sqrt{p\log\log(8n)}$, a thresholding rule is applied at level a, where $a^2 = 4\log\left(\frac{ep\log\log(8n)}{s^2}\right)$. We discuss adaptivity to the sparsity level s in Section 4.3.3.

4.3.2 Lower bound

We show that the testing procedure (4.7) is minimax optimal by stating a matching lower bound.

Proposition 4.3.2. *For any $\varepsilon \in (0,1)$, there exists $c > 0$, depending only on ε, such that $\mathscr{R}(\rho) \ge 1-\varepsilon$ whenever $\rho^2 \le cr^*(p,n,s)$.*

The most important factor in the minimax rate (4.5) is $p\log\log(8n)$, which appears in both the sparse and the dense regimes as well as the phase transition boundary. We illustrate why this term is necessary by giving the lower bound construction for $s = p$. The construction considers a $p \times n$ random matrix $\theta = (\theta_{jt})$ with distribution denoted by ν, generated according to the following sampling process:

1. Let $k \sim \mathrm{Unif}(\{0,1,2,\ldots,\lfloor \log_2(n/2)\rfloor\})$;
2. Let $u_1,\ldots,u_p \overset{\text{iid}}{\sim} \mathrm{Unif}(\{-1,1\})$, independent of k, and let $u := (u_1,\ldots,u_p)^T$;

3. Given (k,u), define $\theta_{jt} := 2^{-k/2}\beta u_j$ for $j \in [p], t \in [2^k]$, where $\beta > 0$ will be specified later, and set $\theta_{jt} := 0$ otherwise.

The lower bound is then obtained through calculating the second moment of the likelihood ratio between the null and the alternative mixture.

4.3.3 Adaptation to sparsity

The testing procedure (4.7) that achieves the minimax detection rate depends on knowledge of the sparsity s. In this section, we present an alternative procedure that is adaptive to s. The idea is to take supremum over a grid of sparsity levels. Recall the definition of the testing procedure $\psi_{a,r}$ in (4.7), and let us makes the dependence on s explicit by writing

$$\psi^{(s)} := \psi_{a(s),r(s)},$$

where $a^2(s) := 4\log\left(\frac{ep\log\log(8n)}{s^2}\right)\mathbb{1}_{\{s<\sqrt{p\log\log(8n)}\}}$ and $r(s) := Cr^*(p,n,s)$ as in Proposition 4.3.1. Then our adaptive test is defined by

$$\psi_{\text{adaptive}} := \max_{s\in\mathscr{S}} \psi^{(s)},$$

where

$$\mathscr{S} := \left\{1,2,4,\ldots,2^{\left\lceil \log_2\left(\sqrt{p\log\log(8n)}\right)\right\rceil -1}\right\} \cup \{p\}.$$

The choice of this particular grid for $\mathscr{S}$ is not essential (we could also take $\mathscr{S} := [p]$), but it reduces computation.

Theorem 4.3.2. *For any $\varepsilon \in (0,1)$, there exists $C > 0$, depending only on ε, such that the testing procedure ψ_{adaptive} satisfies*

$$\sup_{\theta\in\Theta_0(p,n)} \mathbb{E}_\theta \psi_{\text{adaptive}} + \sup_{\theta\in\Theta(p,n,s,\rho)} \mathbb{E}_\theta(1-\psi_{\text{adaptive}}) \le \varepsilon,$$

as long as $\rho^2 \geq 64Cr^*(p,n,s)$.

Theorem 4.3.2 shows that the minimax detection boundary (4.3.1) can be achieved adaptively without the knowledge of the sparsity level s. In the literature, changepoint detection with unknown sparsity was also investigated by [68]. Their procedure has a vanishing testing error as long as

$$\rho^2 \gtrsim \min\left(\sqrt{p\log p} + \sqrt{p\log\log n}, s\log\frac{p}{s}\right), \tag{4.8}$$

under the additional assumptions that $p, s \to \infty$, $s/p \to 0$, and $\frac{\log n}{s\log(p/s)} \to 0$. Comparing (4.8) with the optimal rate (4.3.1), we see that [68] successfully identified the $\sqrt{p\log\log n}$ term in the dense regime and the $s\log p$ term in the sparse regime. However, we can also observe that the $\sqrt{p\log p}$ term is not necessary, and the rate (4.8) is in general not sharp without the assumption $\frac{\log n}{s\log(p/s)} \to 0$, especially when the sparsity level s is around $\sqrt{p\log\log n}$.

4.3.4 Asymptotic constants

A notable feature of our minimax detection boundary derived in Theorem 4.3.1 is that the rate is non-asymptotic, meaning that the result holds for arbitrary $n \geq 2$, $p \in \mathbb{N}$ and $s \in [p]$. On the other hand, if we are allowed to make a few asymptotic assumptions, we can give explicit constants for the lower and upper bounds. In this subsection, therefore, we let both the dimension p and the sparsity s be functions of n, and we consider asymptotics as $n \to \infty$.

Theorem 4.3.3 (Dense regime)**.** *Assume that* $s^2/(p\log\log n) \to \infty$ *as* $n \to \infty$*. Then, with*

$$\rho = \xi\,(p\log\log n)^{1/4},$$

we have $\mathscr{R}(\rho) \to 0$ *when* $\xi > \sqrt{2}$ *and* $\mathscr{R}(\rho) \to 1$ *when* $\xi < \sqrt{2}$*.*

Theorem 4.3.4 (Sparse regime)**.** *Assume that* $s^2/p \to 0$ *and* $s/\log\log n \to \infty$ *as* $n \to \infty$*. Then, with*

$$\rho = \xi\sqrt{s\log\left(\frac{p\log\log n}{s^2}\right)},$$

we have $\mathscr{R}(\rho) \to 0$ *when* $\xi > \sqrt{2}$ *and* $\mathscr{R}(\rho) \to 1$ *when* $\xi < 1$.

Theorem 4.3.3 characterizes the exact asymptotic minimax rate in the dense regime. On the other hand, under the current formulation, there is no exact constant in the sparse regime, and a different scaling of sparsity is needed in order to pin down the exact constant, and we therefore leave it as an open problem for future research.

CHAPTER 5

DENSITY ESTIMATION WITH CONTAMINATION: MINIMAX RATES AND ADAPTATION

This chapter considers the problem of density estimation with contamination under the point-wise loss. The worst-case difficulty of the problem and possibility for adaption are studied.[1]

5.1 Introduction

Nonparametric density estimation is a well-studied classical topic [81, 82, 83]. In this chapter, we consider this classical statistical task with a twist. Instead of assuming i.i.d. observations from a true density f, we assume

$$X_1, ..., X_n \sim (1-\varepsilon)f + \varepsilon g, \tag{5.1}$$

where g is a density not related to f, and the goal is to estimate $f(x_0)$ at some $x_0 \in \mathbb{R}$. In other words, for each observation, there is an ε probability that the observation is sampled from a distribution not related to the density of interest.

This problem naturally appears in both the robust statistics and the multiple testing literature. In robust statistics literature, g is recognized as the contamination distribution, and the task is interpreted as robustly estimating a density f with contaminated data points [84]. In multiple testing literature, f and g are respectively recognized as null density and alternative density, and the task is interpreted as estimating null density at a point [85]. In this chapter, we adopt the name "contamination" to refer to both g and the observations generated from it.

The nature of the problem heavily depends on the assumptions put on f and g. When there is no constraint on the contamination distribution g, the data generating process (5.1) is also recognized as Huber's ε-contamination model [86, 87]. Recent work on nonparametric estimation in such a setting includes [84, 88], and the influence of contamination on minimax rates is investigated

1. Work in this chapter is joint with Chao Gao and published in [80].

by [89, 84]. On the other hand, in the literature of multiple testing, it is more common to put parametric structural assumptions on the alternative g, and optimal rates of estimating the null density f are investigated by [90, 91].

In this chapter, we explore this problem with connections to nonparametric density estimation literature in mind. Specifically, the density function f is assumed to have a Hölder smoothness β_0 (it is also possible to put other kinds of assumptions, for example shape constraints, see [92, 93, 94]). Both cases of structured and arbitrary contamination are considered and fundamental limit of this problem is studied by establishing the minimax rate. The theory of adaptation in both settings are also investigated.

Notation For $a, b \in \mathbb{R}$, let $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. For an integer m, $[m]$ denotes the set $\{1,2,...,m\}$. For a positive real number x, $\lceil x \rceil$ is the smallest integer no smaller than x and $\lfloor x \rfloor$ is the largest integer no larger than x. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if $a_n \leq Cb_n$ for all n with some consntant $C > 0$ independent of n. The notation $a_n \asymp b_n$ means we have both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Given a set S, $|S|$ denotes its cardinality, and $\mathbb{1}_S$ is the associated indicator function. We use $\mathbb{P}$ and $\mathbb{E}$ to denote generic probability and expectation whose distribution is determined from the context. The notation $\mathbb{E}(X : S)$ stands for $\mathbb{E}(X\mathbb{1}_S)$. The class of infinitely differentiable functions on $\mathbb{R}$ is denoted by $\mathscr{C}^\infty(\mathbb{R})$. For two probability measures $\mathbb{P}$ and $\mathbb{Q}$, the chi-squared divergence is defined as $\chi^2(\mathbb{P},\mathbb{Q}) = \int \frac{d\mathbb{P}^2}{d\mathbb{Q}} - 1$, and the total variation distance is defined as $\mathsf{TV}(\mathbb{P},\mathbb{Q}) = \sup_B |\mathbb{P}(B) - \mathbb{Q}(B)|$. Throughout the chapter, C, c and their variants denote generic constants that do not depend on n. Their values may change from place to place. For an integer k, we use $f^{(k)}$ to denote the k'th derivative of a k'th differentiable function f, with the convention $f^{(0)} = f$.

5.2 Results with structured contamination

Consider i.i.d. observations $X_1,...,X_n \sim (1-\varepsilon)f + \varepsilon g$. The goal is to estimate f at a given point. Without loss of generality, we aim to estimate $f(0)$. In other words, for every $i \in [n]$, we have

$X_i \sim f$ with probability $1-\varepsilon$ and $X_i \sim g$ with probability ε. Thus, there are approximately $n\varepsilon$ observations that are not related to the density function f, which are referred to as contamination.

To study the fundamental limit of estimating f with contaminated data, we need to specify appropriate regularity conditions on both f and g. We first define the Hölder class by

$$\Sigma(\beta, L) = \left\{ f : \mathbb{R} \to \mathbb{R} \middle| \left| f^{(\lfloor \beta \rfloor)}(x_1) - f^{(\lfloor \beta \rfloor)}(x_2) \right| \leq L|x_1 - x_2|^{\beta - \lfloor \beta \rfloor} \text{ for any } x_1, x_2 \in \mathbb{R} \right\}.$$

Here, β stands for the smoothness parameter, and L stands for the radius of the function space. The Hölder class of density functions is defined as

$$\mathscr{P}(\beta, L) = \left\{ f : \mathbb{R} \to [0, \infty) \middle| f \in \Sigma(\beta, L), \int f = 1 \right\}.$$

Finally, we define the class of mixtures in the form of $(1-\varepsilon)f + \varepsilon g$ by

$$\mathscr{M}(\varepsilon, \beta_0, \beta_1, L_0, L_1, m) = \left\{ (1-\varepsilon)f + \varepsilon g \middle| f \in \mathscr{P}(\beta_0, L_0), g \in \mathscr{P}(\beta_1, L_1), g(0) \leq m \right\}.$$

This class is indexed by several parameters. Throughout this chapter, we refer to ε as contamination proportion and m as contamination level at 0. The pair (β_0, L_0) controls the smoothness of the density function f that we want to estimate, and the pair (β_1, L_1) controls the smoothness of the contamination density g. Among the six numbers, ε and m are allowed to depend on the sample size n, but the numbers $\beta_0, \beta_1, L_0, L_1$ are all assumed to be constants that do not depend on n throughout this chapter. It is also assumed that $\varepsilon \leq 1/2$.

5.2.1 minimax rates

The minimax risk of estimation is defined as (notice that we suppress the dependence on n for $\mathscr{R}$)

$$\mathscr{R}(\varepsilon, \beta_0, \beta_1, L_0, L_1, m) = \inf_{\widehat{f}(0)} \sup_{p(\varepsilon, f, g) \in \mathscr{M}(\varepsilon, \beta_0, \beta_1, L_0, L_1, m)} \mathbb{E}_{X_1, \ldots, X_n \sim p} \left(\widehat{f}(0) - f(0) \right)^2,$$

where the notation $p(\varepsilon, f, g)$ is used to denote the density $(1-\varepsilon)f + \varepsilon g$. Later in this chapter, we will shorthand $\mathbb{E}_{X_1,\ldots,X_n \overset{i.i.d}{\sim} p}$ by $\mathbb{E}_{p^n}$. Obviously, the minimax risk becomes smaller if ε gets smaller or n gets larger. Besides the role of ε and n, the other model indices are also expected to affect the difficulty of the problem, as listed in the following.

- The smoothness of f: From classical density estimation theory, we know the smoother f is, the easier it is to estimate $f(0)$.
- The level of $g(0)$: Intuitively, the smaller $g(0)$ is, the smaller its influence is on $f(0)$, and thus the easier the problem is.
- The smoothness of g: Intuitively, the smoother g is, the less the contamination effect can spread, and thus the easier it is to account for the effect of g in the contamination model.

Now we present the following theorem of minimax rate, that justifies our intuition above.

Theorem 5.2.1. *Under the setting above, we have*

$$\mathscr{R}(\varepsilon, \beta_0, \beta_1, L_0, L_1, m) \asymp [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\varepsilon^2 (1 \wedge m)^2] \vee [n^{-\frac{2\beta_1}{2\beta_1+1}} \varepsilon^{\frac{2}{2\beta_1+1}}]. \tag{5.2}$$

In other words, $\mathscr{R}(\varepsilon, \beta_0, \beta_1, L_0, L_1, m)$ can be upper bounded by a constant multiple of the right hand side of (5.2) and lower bounded by another constant multiple of the right hand side of (5.2), where the constants only depend on $\beta_0, \beta_1, L_0, L_1$.

Theorem 5.2.1 completely characterizes the difficulty of estimating $f(0)$ with contaminated data. The three terms in the rate (5.2) have different but very clear meanings. The first term $n^{-\frac{2\beta_0}{2\beta_0+1}}$ is the classical minimax rate of estimating a smooth function at a given point without contamination. The second term $\varepsilon^2(1 \wedge m)^2$ is proportional to the squared of the product of contamination level and contamination proportion. The last term $n^{-\frac{2\beta_1}{2\beta_1+1}} \varepsilon^{\frac{2}{2\beta_1+1}}$ is perhaps the most interesting. Here the effect of ε is powered by an exponent depending on β_1, and it stands for the interaction between the contamination proportion and the contamination smoothness. The fact

that it does not depend on m implies that we have to pay this price with contaminated data even if $g(0) = 0$.

Upper bound Define the following class of kernel functions

$$\begin{aligned}\mathscr{K}_l(L) &= \left\{K : \mathbb{R} \to \mathbb{R} \Big| \int K = 1, \int x^j K(x)dx = 0 \text{ for all } j \in [l], \right. \\ &\qquad \left. \|K\|_\infty \vee \int K^2 \vee \int |x|^l |K(x)|dx \le L \right\}.\end{aligned}$$

which collects all bounded and squared integrable kernel functions of order l. The minimax rate (5.2) can be achieved by a simple kernel density estimator that takes the form

$$\widehat{f}_h(0) = \frac{1}{n(1-\varepsilon)} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i}{h}\right). \tag{5.3}$$

with some kernel function $K \in \mathscr{K}_{\lfloor \beta_0 \vee \beta_1 \rfloor}(L)$ and the bandwidth choice

$$h = n^{-\frac{1}{2\beta_0+1}} \wedge n^{-\frac{1}{2\beta_1+1}} \varepsilon^{-\frac{2}{2\beta_1+1}}.$$

This estimator is slightly different from the classical kernel density estimator because it is normalized by $\frac{1}{n(1-\varepsilon)}$ instead of $\frac{1}{n}$. The knowledge of the contamination proportion ε is very critical to achieve the minimax rate (5.2). Later, we will show in Section 5.2.2 that the minimax rate (5.2) cannot be achieved if ε is not known. Compared with the optimal bandwidth of order $n^{-\frac{1}{2\beta_0+1}}$ in classical nonparametric function estimation, the h in the structured contamination setting is always smaller. The choice of bandwidth is a consequences of the specific bias-variance tradeoff under the structured contamination model. As an interesting contrast, in the case of arbitrary contamination, the optimal choice of bandwidth is always larger than the usual one, see Section 5.3.

Lower Bounds The lower bound of the first two terms in the minimax rate can be readily obtained. On the other hand, the derivation of the lower bound corresponding to the third term

$\mathscr{R}(\varepsilon,\beta_0,\beta_1,L_0,L_1,m) \gtrsim n^{-\frac{2\beta_1}{2\beta_1+1}}\varepsilon^{\frac{2}{2\beta_1+1}}$ is rather intricate. The construction considers the following four functions, where the terms in blue and the term in red correspond to the blue functions and the red function plotted in figure 5.1.

$$
\begin{aligned}
f(x) &= f_0(x),\\
\widetilde{f}(x) &= f_0(x) + \frac{\varepsilon}{1-\varepsilon} c_2 \left[h^{\beta_0} l\left(\frac{x}{h}\right) - h^{\beta_0} l\left(\frac{2(x-c_4)}{h}\right) - h^{\beta_0} l\left(\frac{2(x+c_4)}{h}\right)\right],\\
g(x) &= c_1 a(c_1 x) + c_2 \left[h^{\beta_0} l\left(\frac{x}{h}\right) - h^{\beta_0} l\left(\frac{2(x-c_4)}{h}\right) - h^{\beta_0} l\left(\frac{2(x+c_4)}{h}\right)\right] - c_3 \widetilde{h}^{\beta_1} b\left(\frac{x}{\widetilde{h}}\right),\\
\widetilde{g}(x) &= c_1 a(c_1 x),
\end{aligned}
$$

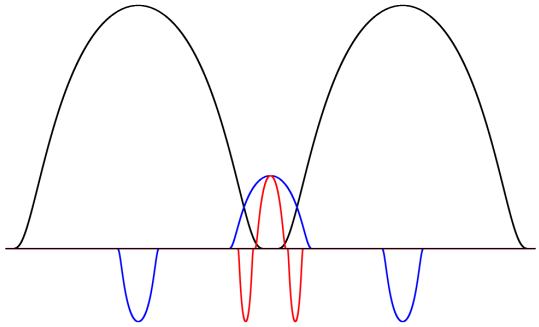
where the constants c_1, c_2, c_3, c_4 are chosen properly so that the constructed functions are well-defined densities in the desired function classes.

A dominant feature of this constructions is that g is a perturbation of $\widetilde{g}$ with two levels of perturbation, respectively with bandwidth h and $\widetilde{h}$, while usual lower bound proof in nonparametric estimation involves perturbing a function at a single bandwidth level. The first level of perturbation $h^{\beta_0} l\left(\frac{x}{h}\right)$ serves to cancel the effect of the corresponding perturbation on f, while the second perturbation $-\widetilde{h}^{\beta_1} b\left(\frac{x}{\widetilde{h}}\right)$ serves to ensure the constraint of contamination level. Indeed, if we relate h and $\widetilde{h}$ through the equation $h^{\beta_0} \asymp \widetilde{h}^{\beta_1}$, then it is direct that $\widetilde{g}(0) = g(0) = 0$. In other words, the constructed contamination density functions g and $\widetilde{g}$ both have contamination level 0. An illustration of this construction with a two-level perturbation is given by Figure 5.1. The colors of the plot correspond to those in the formulas.

5.2.2 Adaptation theory

To achieve the minimax rate in Theorem 5.2.1, the kernel density estimator (5.3) requires the knowledge of contamination proportion ε and smoothness (β_0,β_1). In this section, we discuss adaptive procedures to estimate $f(0)$ without the knowledge of these parameters. However, adaptation to ε or to (β_0,β_1) is not free, and one can only achieve slower rates than the minimax rate

Figure 5.1: An illustration of the construction of g.



(5.2). The adaptation cost varies for each different scenario. A summary of our results is listed below.

- When the contamination proportion is unknown, the best possible rate is

$$n^{-\frac{2\beta_0}{2\beta_0+1}} \vee \varepsilon^2.$$

- When the smoothness parameters are unknown, the best possible rate is

$$\left[\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}}\right] \vee \left[\varepsilon^2(1\wedge m)^2\right] \vee \left[\left(\frac{n}{\log n}\right)^{-\frac{2\beta_1}{2\beta_1+1}} \varepsilon^{\frac{2}{2\beta_1+1}}\right].$$

- When both the contamination proportion and the smoothness are unknown, the best possible rate becomes

$$\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}} \vee \varepsilon^2.$$

Compared with the minimax rate (5.2), the ignorance of the contamination proportion implies that m is replaced by 1 in the rate, while the ignorance of the smoothness implies that n is replaced by $n/\log n$ in the rate.

Unknown contamination proportion

The kernel density estimator (5.3) depends on ε in two ways: the normalization through $\frac{1}{n(1-\varepsilon)}$ and the optimal choice of bandwidth h. Without the knowledge of ε, we consider the following estimator

$$\widehat{f}_h(0) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{X_i}{h}\right). \tag{5.4}$$

The first difference between (5.4) and (5.3) is the normalization. When ε is not given, we can only use $\frac{1}{n}$ in (5.4). Moreover, the choice of h in (5.4) cannot depend on ε.

Theorem 5.2.2. *For the estimator $\widehat{f}(0) = \widehat{f}_h(0)$ with some $K \in \mathscr{K}_{\lfloor \beta_0 \vee \beta_1 \rfloor}(L)$ and $h = n^{-\frac{1}{2\beta_0+1}}$, we have*

$$\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\varepsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \lesssim n^{-\frac{2\beta_0}{2\beta_0+1}} \vee \varepsilon^2.$$

With the choice $h = n^{-\frac{1}{2\beta_0+1}}$, $\widehat{f}_h$ becomes the classical nonparametric density estimator. The contamination results in an extra ε^2 in the rate compared with the classical nonparametric minimax rate, regardless of the values of m and β_1. In view of the form of the minimax rate (5.2), the rate given by Theorem 5.2.2 can be obtained by replacing the $\varepsilon^2(1\wedge m)^2$ in (5.2) with ε^2. A matching lower bound for adaptivity to ε is given by the following theorem.

Theorem 5.2.3. *Consider two models $\mathscr{M}(\varepsilon,\beta_0,\beta_1,L_0,L_1,m)$ and $\mathscr{M}(\widetilde{\varepsilon},\beta_0,\beta_1,L_0,L_1,m)$ with different contamination proportions. For any estimator $\widehat{f}(0)$ that satisfies*

$$\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\widetilde{\varepsilon},\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \le C\widetilde{\varepsilon}^2,$$

for some constant $C > 0$, there must exist another constant $C' > 0$, such that for $\varepsilon \ge C'\widetilde{\varepsilon}$, we have

$$\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\varepsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \gtrsim \varepsilon^2.$$

Theorem 5.2.3 shows that it is impossible to achieve a rate that is faster than ε^2 even over only two different contamination proportions.

Unknown smoothness

In this section, we consider the case that the smoothness numbers are unknown, but the contamination proportion is given. In view of the kernel density estimator (5.3) that achieves the minimax rate, we can still use the normalization by $\frac{1}{n(1-\varepsilon)}$ because of the knowledge of ε, but the bandwidth h needs to be picked in a data-driven way. For a given h, define

$$\widehat{f}_h(0) = \frac{1}{n(1-\varepsilon)} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i}{h}\right).$$

With a discrete set $\mathscr{H}$ and some constant $c_1 > 0$, Lepski's method [95, 96, 97] selects a data-driven bandwidth through the following procedure,

$$\widehat{h} = \max\left\{ h \in \mathscr{H} : |\widehat{f}_h(0) - \widehat{f}_l(0)| \le c_1 \sqrt{\frac{\log n}{nl}}, \forall l \le h, l \in \mathscr{H} \right\}. \tag{5.5}$$

In words, we choose the largest bandwidth below which the variance dominates. If the set that is maximized over is empty, we will use the convention $\widehat{h} = \frac{1}{n}$. The estimator $\widehat{f}_{\widehat{h}}(0)$ that uses a data-driven bandwidth enjoys the following guarantee.

Theorem 5.2.4. *Consider the adaptive kernel density estimator $\widehat{f}(0) = \widehat{f}_{\widehat{h}}(0)$ with the bandwidth defined by (5.5). In (5.5), we set $\mathscr{H} = \left\{1, \frac{1}{2}, \cdots, \frac{1}{2^m}\right\}$ such that $\frac{1}{2^m} \le \frac{1}{n} < \frac{1}{2^{m-1}}$ and c_1 to be a sufficiently large constant. The kernel K is selected from $\mathscr{K}_l(L)$ with a large constant $l \ge \lfloor \beta_0 \vee \beta_1 \rfloor$. Then, we have*

$$\begin{aligned} &\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\varepsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \\ \lesssim\ & \left[\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}}\right] \vee \left[\varepsilon^2 (1 \wedge m)^2\right] \vee \left[\left(\frac{n}{\log n}\right)^{-\frac{2\beta_1}{2\beta_1+1}} \varepsilon^{\frac{2}{2\beta_1+1}}\right]. \end{aligned}$$

Lepski's method is known to be adaptive over various nonparametric classes, and it can achieve minimax rates up to a logarithmic factor without knowing the smoothness parameter [98]. Theorem 5.2.4 shows that this is also the case with contaminated observations. With an adaptive kernel

density estimator normalized by $\frac{1}{n(1-\varepsilon)}$, the minimax rate (5.2) is achieved up to a logarithmic factor in Theorem 5.2.4.

A comparison between the adaptive rate given by Theorem 5.2.4 and the minimax rate (5.2) reveals two differences. The first adaptation cost is given by $\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}}$, compared with $n^{-\frac{2\beta_0}{2\beta_0+1}}$ in (5.2). Previous work in adaptive nonparametric estimation [99, 98, 100] implies that this cost is unavoidable for adaptation to smoothness. The second adaptation cost is given by $\left(\frac{n}{\log n}\right)^{-\frac{2\beta_1}{2\beta_1+1}} \varepsilon^{\frac{2}{2\beta_1+1}}$, compared with $n^{-\frac{2\beta_1}{2\beta_1+1}} \varepsilon^{\frac{2}{2\beta_1+1}}$ in (5.2). In the next theorem, we show that this adaptations cost is also unavoidable without the knowledge of the smoothness parameters.

Theorem 5.2.5. *Consider two models $\mathscr{M}(\varepsilon, \beta_0, \beta_1, L_0, L_1, m)$ and $\mathscr{M}(\varepsilon, \widetilde{\beta}_0, \widetilde{\beta}_1, \widetilde{L}_0, \widetilde{L}_1, m)$ with different smoothness parameters. Assume that $\beta_0 \leq \tilde{\beta}_0$, $\beta_1 < \tilde{\beta}_1$, $\beta_0 \geq \beta_1$ and $n\varepsilon^2 \geq (\log n)^2$. For any estimator $\widehat{f}(0)$ that satisfies*

$$\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\varepsilon,\widetilde{\beta}_0,\widetilde{\beta}_1,\widetilde{L}_0,\widetilde{L}_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \leq C\left(\frac{n}{\log n}\right)^{-\frac{2\widetilde{\beta}_1}{2\widetilde{\beta}_1+1}} \varepsilon^{\frac{2}{2\widetilde{\beta}_1+1}},$$

for some constant $C > 0$, we must have

$$\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\varepsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \gtrsim \left(\frac{n}{\log n}\right)^{-\frac{2\beta_1}{2\beta_1+1}} \varepsilon^{\frac{2}{2\beta_1+1}}.$$

Unknown contamination proportion and unknown smoothness

When both the contamination proportion and the smoothness are unknown, we consider Lepski's method with a kernel density estimator normalized by $\frac{1}{n}$. Define

$$\widehat{f}_h(0) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i}{h}\right).$$

Then, a data-driven bandwidth $\widehat{h}$ is selected according to (5.5). Again, if the set that is maximized over is empty in (5.5), we will use the convention $\widehat{h} = \frac{1}{n}$. Note that this is a fully data-driven

estimator that is adaptive to both the contamination proportion and the smoothness. It enjoys the following guarantee.

Theorem 5.2.6. *Consider the adaptive kernel density estimator $\widehat{f}(0) = \widehat{f}_{\widehat{h}}(0)$ with the bandwidth defined by (5.5). In (5.5), we set $\mathscr{H} = \left\{1, \frac{1}{2}, \cdots, \frac{1}{2^m}\right\}$ such that $\frac{1}{2^m} \leq \frac{1}{n} < \frac{1}{2^{m-1}}$ and c_1 to be a sufficiently large constant. The kernel K is selected from $\mathscr{K}_l(L)$ with a large constant $l \geq \lfloor \beta_0 \vee \beta_1 \rfloor$. Then, we have*

$$\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\varepsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \lesssim \left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}} \vee \varepsilon^2.$$

Compared with the minimax rate in Theorem 5.2.1, the rate in Theorem 5.2.6 can be understood as replacing n and $\varepsilon^2(1\wedge m)^2$ respectively by $n/\log n$ and ε^2 in (5.2). In view of the results in both Section 5.2.2 and Section 5.2.2, this rate $\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}} \vee \varepsilon^2$ in Theorem 5.2.6 cannot be improved by any procedure that is adaptive to both contamination proportion and smoothness.

5.3 Results for arbitrary contamination

In this section, we study the contamination model without any structural assumption on the contamination distribution:

$$X_1, \ldots, X_n \sim (1-\varepsilon)P_f + \varepsilon G$$

where P_f is a distribution on $\mathbb{R}$ that has a density function f, and G is an arbitrary contamination distribution. This leads to the following model space

$$\mathscr{M}(\varepsilon, \beta_0, L_0) = \left\{(1-\varepsilon)P_f + \varepsilon G \middle| f \in \mathscr{P}(\beta_0, L_0) \text{ and } G \text{ is an arbitrary distribution}\right\}.$$

This is often referred to as Huber's ε-contamination model [86, 87]. Nonparametric function estimation under Huber's ε-contamination model has recently been studied by [84, 88] for global

loss functions. Here our focus is on the local estimation of $f(0)$.

5.3.1 Minimax rates

The corresponding minimax risk is defined by

$$\mathscr{R}(\varepsilon,\beta_0,L_0)=\inf_{\widehat{f}(0)}\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\varepsilon,\beta_0,L_0)}\mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2.$$

In contrast to the minimax rate studied in Section 5.2.1, we only have one parameter ε that indexes the influence of the contamination for $\mathscr{R}(\varepsilon,\beta_0,L_0)$.

Theorem 5.3.1. *Under the setting above, we have*

$$\mathscr{R}(\varepsilon,\beta_0,L_0)\asymp[n^{-\frac{2\beta_0}{2\beta_0+1}}]\vee[\varepsilon^{\frac{2\beta_0}{\beta_0+1}}]. \tag{5.6}$$

The minimax rate given by Theorem 5.3.1 only involves two terms. The first term $n^{-\frac{2\beta_0}{2\beta_0+1}}$ is the classical minimax rate for nonparametric estimation. The second term $\varepsilon^{\frac{2\beta_0}{\beta_0+1}}$ characterizes the influence of contamination. It is worth noticing that the smoothness index of f appears both in $n^{-\frac{2\beta_0}{2\beta_0+1}}$ and $\varepsilon^{\frac{2\beta_0}{\beta_0+1}}$. A larger value of β_0 implies a less influence of the contamination. This is in contrast to the rate of $\mathscr{R}(\varepsilon,\beta_0,\beta_1,L_0,L_1,m)$ in Theorem 5.2.1.

The phase transition boundary of $\mathscr{R}(\varepsilon,\beta_0,L_0)$ occurs at $\varepsilon=n^{-\frac{\beta_0+1}{2\beta_0+1}}$. Below this level, we have $\mathscr{R}(\varepsilon,\beta_0,L_0)\asymp n^{-\frac{2\beta_0}{2\beta_0+1}}$, and the contamination has no influence on the classical minimax rate. When ε is above $n^{-\frac{\beta_0+1}{2\beta_0+1}}$, the rate becomes $\varepsilon^{\frac{2\beta_0}{\beta_0+1}}$, dominated by the contamination of data. Since we have about $n\varepsilon$ contaminated observations in expectation, an optimal procedure can achieve the classical minimax rate $n^{-\frac{2\beta_0}{2\beta_0+1}}$ with at most $n\varepsilon\le n^{\frac{\beta_0}{2\beta_0+1}}$ contaminated data points. Note that the number $n^{\frac{\beta_0}{2\beta_0+1}}$ is an increasing function of β_0.

For the upper bound of the minimax rate, we consider the kernel density estimator $\widehat{f}_h(0)=\frac{1}{n}\sum_{i=1}^n\frac{1}{h}K\left(\frac{X_i}{h}\right)$ with the choice of bandwidth $h=n^{-\frac{1}{2\beta_0+1}}\vee\varepsilon^{\frac{1}{\beta_0+1}}$. It is interesting to note that this choice of bandwidth is always larger than or equal to $n^{-\frac{1}{2\beta_0+1}}$. Recall that when the contamination

is smooth, the optimal bandwidth is smaller than $n^{-\frac{1}{2\beta_0+1}}$. The lower bound part of Theorem 5.3.1 can be viewed as an application of Theorem 5.1 in [89].

5.3.2 Adaptation to either contamination proportion or smoothness

The key to adaptation to either contamination proportion or smoothness is the following risk decomposition

$$\mathbb{E}\left(\widehat{f}_h(0)-f(0)\right)^2 \lesssim \frac{1}{nh}\vee h^{2\beta_0}\vee\frac{\varepsilon^2}{h^2} \asymp \left(\frac{\varepsilon^2}{h^2}+\frac{1}{nh}\right)+h^{2\beta_0}, \tag{5.7}$$

of the kernel density estimator $\widehat{f}_h(0)=\frac{1}{n}\sum_{i=1}^n\frac{1}{h}K\left(\frac{X_i}{h}\right)$. The first term $\frac{\varepsilon^2}{h^2}+\frac{1}{nh}$ is a decreasing function of h with a possibly unknown ε, while the second term $h^{2\beta_0}$ is an increasing function of h with a possibly unknown β_0. If we know ε but do not know β_0, then we can use Lespki's method with $\frac{\varepsilon^2}{h^2}+\frac{1}{nh}$ as a reference curve. On the other hand, if we know β_0 but do not know ε, we can then use a reverse version of Lepski's method with $h^{2\beta_0}$ as a reference curve. Specifically, when ε is known but β_0 is unknown, we use

$$\widehat{h}=\max\left\{h\in\mathscr{H}:|\widehat{f}_h(0)-\widehat{f}_l(0)|\le c_1\left(\sqrt{\frac{\log n}{nl}}+\frac{\varepsilon}{l}\right),\forall l\le h,l\in\mathscr{H}\right\}. \tag{5.8}$$

If the set that is maximized over is empty, we take $\widehat{h}=\frac{1}{n}$. When β_0 is known but ε is unknown, we use

$$\widehat{h}=\min\left\{h\in\mathscr{H}:|\widehat{f}_h(0)-\widehat{f}_l(0)|\le c_1 l^{\beta_0},\forall l\ge h,l\in\mathscr{H}\right\}. \tag{5.9}$$

If the set that is minimized over is empty, we take $\widehat{h}=1$.

Theorem 5.3.2. *Consider the adaptive kernel density estimator $\widehat{f}(0)=\widehat{f}_{\widehat{h}}(0)$ with the bandwidth $\widehat{h}$ given by (5.8) or (5.9). In either case, we set $\mathscr{H}=\left\{1,\frac{1}{2},\cdots,\frac{1}{2^m}\right\}$ such that $\frac{1}{2^m}\le\frac{1}{n}<\frac{1}{2^{m-1}}$ and c_1 to be a sufficiently large constant. The kernel K is selected from $\mathscr{K}_l(L)$ with a large constant*

$l \geq \lfloor \beta_0 \rfloor$. *Then, we have*

$$\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\varepsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \lesssim \left(\frac{\log n}{n}\right)^{\frac{2\beta_0}{2\beta_0+1}} \vee \varepsilon^{\frac{2\beta_0}{\beta_0+1}}.$$

With one of ε and β_0 given, Theorem 5.3.2 guarantees adaptive estimation with the rate $\left(\frac{\log n}{n}\right)^{\frac{2\beta_0}{2\beta_0+1}} \vee \varepsilon^{\frac{2\beta_0}{\beta_0+1}}$. Compared with the minimax rate in Theorem 5.3.1, we have an extra logarithmic factor due to the ignorance of either ε or β_0. This logarithmic factor cannot be removed by any adaptive procedure in view of the results of [99, 98, 100].

5.3.3 Adaptation to both contamination proportion and smoothness?

When both contamination proportion and smoothness are unknown, the adaptation theory with arbitrary contamination is completely different from the case with structured contamination. Since there is no constraint on the contamination distribution, a model with (ε, β_0) can also be written as a different model with $(\widetilde{\varepsilon}, \widetilde{\beta}_0)$. As a consequence, we can prove the following lower bound.

Lemma 5.3.1. *For any constants $c_1, c_2 > 0$, there exists a constant c_0, such that for any $\beta_0, \widetilde{\beta}_0 \leq c_1$, and any $L_0, \widetilde{L_0} \geq c_2$, and any estimator $\widehat{f}(0)$, one of the following lower bounds must be true,*

$$\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\varepsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \geq c_0\varepsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}},$$

$$\sup_{p(0,f,g)\in\mathscr{M}(0,\widetilde{\beta}_0,\widetilde{L}_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \geq c_0\varepsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}}.$$

As we will show in the following, this specific form of lower bound in this lemma has a profound implication, in that an adaptive estimation rate that is a function of an individual class is impossible! As a first step, the following definition formulates what adaptivity means in our specific setting.

Definition 5.3.1. An estimator $\widehat{f}(0)$ is called $(c_1, c_2, c_3, r_1(\cdot), r_2(\cdot))$ rate adaptive if the following

holds: for any $n \geq 1$, any $\varepsilon \leq 1/2$, any $\beta_0 \leq c_1$ and any $L_0 \leq c_2$, we have

$$\sup_{p(\varepsilon,f,g)\in\mathscr{M}(\varepsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \leq c_3 n^{-r_1(\beta_0)} \vee \varepsilon^{r_2(\beta_0)}. \tag{5.10}$$

As concrete examples, when the contamination distribution is restricted to those with density functions that are Hölder smooth, it is shown in Theorem 5.2.6 that adaptive estimation is possible with some $r_1(\beta_0) < \frac{2\beta_0}{2\beta_0+1}$ and $r_2(\beta_0) = 2$. When the contamination distribution is arbitrary, Theorem 5.3.2 shows that adaptive estimation is possible over (ε, β_0) if either ε or β_0 is fixed (known) with some $r_1(\beta_0) < \frac{2\beta_0}{2\beta_0+1}$ and $r_2(\beta_0) = \frac{2\beta_0}{\beta_0+1}$. In contrast, the following theorem shows that such a goal is impossible for any $r_1(\cdot)$ and $r_2(\cdot)$ when both ε and β_0 are unknown.

Theorem 5.3.3. *For any constants $c_1, c_2, c_3 > 0$ and any positive functions $r_1(\cdot)$ and $r_2(\cdot)$, there is no estimator $\widehat{f}(0)$ that is $(c_1, c_2, c_3, r_1(\cdot), r_2(\cdot))$ rate adaptive.*

In conclusion, when the contamination is arbitrary, the theory of adaptation to both contamination proportion and smoothness is qualitatively different from adaptation to only one of them. In comparison, when the contamination is structured, that difference is just quantitative according to the results in Section 5.2.2. Therefore, in order to achieve sensible error rates adaptively in a robust density estimation context, we need to either assume a given contamination proportion, a given smoothness index, or a structured contamination distribution.

CHAPTER 6

DISCUSSION

Here we discuss several loose ends left in the above chapters. For Chapter 2, the most outstanding problem lies in extending the lower bound to the class of all computationally feasible method, not only the class of iterative thresholding algorithms. In particular, could we show that it is impossible to achieve a restricted optimality guarantee better then what is achieved by the iterative reciprocal thresholding algorithm in the class of all computationally feasible method? In Chapter 3, we show that the factorized approach have the same restricted optimality guarantee as the projected gradient descent method, which is sub-optimal. So the first question is whether it is possible to construct an algorithm that only uses matrix multiplication and can still achieve the optimal restricted optimality guarantee. The second question lies in whether it is possible to generalize the equivalence relation to other settings, including problems with other types of reparametrization or problems with additional constraints other than the low-rank constraint. For Chapter 4, some natural extensions of the sparse changepoint detection problem that still require thorough study are data with temporal and spatial dependence and data arriving in an online fashion. Moreover, there is the question of pinning down the optimal constant in the sparse regime. Finally, let us conclude with the following:

> Perhaps statistical research is the process of coming up with statements that, if true, change the way we look at data.

BIBLIOGRAPHY

[1] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM journal on computing*, vol. 24, pp. 227–234, 1995.

[2] H. Liu and R. Foygel Barber, "Between hard and soft thresholding: optimal iterative thresholding algorithms," *Information and Inference: A Journal of the IMA*, 2018.

[3] S. Negahban, B. Yu, M. J. Wainwright, and P. K. Ravikumar, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," in *Advances in Neural Information Processing Systems*, 2009, pp. 1348–1356.

[4] P.-L. Loh and M. J. Wainwright, "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima," in *Advances in Neural Information Processing Systems*, 2013, pp. 476–484.

[5] P. Jain, A. Tewari, and P. Kar, "On iterative hard thresholding methods for high-dimensional M-estimation," in *Advances in Neural Information Processing Systems*, 2014, pp. 685–693.

[6] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and computational harmonic analysis*, vol. 27, pp. 265–274, 2009.

[7] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," *arXiv preprint arXiv:1509.03025*, 2015.

[8] K. Bhatia, P. Jain, and P. Kar, "Robust regression via hard thresholding," in *Advances in Neural Information Processing Systems*, 2015, pp. 721–729.

[9] P. Jain, N. Rao, and I. S. Dhillon, "Structured sparse regression via greedy hard thresholding," in *Advances in Neural Information Processing Systems*, 2016, pp. 1516–1524.

[10] T. T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow," *The Annals of Statistics*, vol. 44, pp. 2221–2251, 2016.

[11] A. Kyrillidis and V. Cevher, "Matrix recipes for hard thresholding methods," *Journal of mathematical imaging and vision*, vol. 48, pp. 235–265, 2014.

[12] ——, "Recipes on hard thresholding methods," in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on*. IEEE, 2011, pp. 353–356.

[13] T. Blumensath, "Accelerated iterative hard thresholding," *Signal Processing*, vol. 92, pp. 752–756, 2012.

[14] R. Khanna and A. Kyrillidis, "IHT dies hard: provable accelerated iterative hard thresholding," *arXiv preprint arXiv:1712.09379*, 2017.

[15] N. Nguyen, D. Needell, and T. Woolf, "Linear convergence of stochastic iterative greedy algorithms with sparse constraints," *IEEE Transactions on Information Theory*, vol. 63, pp. 6869–6895, 2017.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, pp. 267–288, 1996.

[17] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, pp. 1348–1360, 2001.

[18] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of statistics*, vol. 38, pp. 894–942, 2010.

[19] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, pp. 707–710, 2007.

[20] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$," *Applied and Computational Harmonic Analysis*, vol. 26, pp. 395–407, 2009.

[21] Y. Kabashima, T. Wadayama, and T. Tanaka, "A typical reconstruction limit for compressed sensing based on ℓ_p-norm minimization," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, art. no. L09003, 2009.

[22] M.-J. Lai and J. Wang, "An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems," *SIAM Journal on Optimization*, vol. 21, pp. 82–101, 2011.

[23] L. Zheng, A. Maleki, H. Weng, X. Wang, and T. Long, "Does ℓ_p-minimization outperform ℓ_1-minimization?" *CoRR*, 2015.

[24] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, pp. 127–239, 2014.

[25] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

[26] Y. Zhang, M. J. Wainwright, and M. I. Jordan, "Lower bounds on the performance of polynomial-time algorithms for sparse linear regression," in *Conference on Learning Theory*, 2014, pp. 921–948.

[27] P. J. Bickel, Y. Ritov, A. B. Tsybakov *et al.*, "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, vol. 37, pp. 1705–1732, 2009.

[28] W. Ha, H. Liu, and R. F. Barber, "An equivalence between stationary points for rank constraints versus low-rank factorizations," *arXiv preprint arXiv:1812.00404*, 2018.

[29] S. Oymak, B. Recht, and M. Soltanolkotabi, "Sharp time–data tradeoffs for linear inverse problems," *IEEE Transactions on Information Theory*, vol. 64, pp. 4129–4158, 2018.

[30] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[31] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[32] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," *arXiv preprint arXiv:1704.00708*, 2017.

[33] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, "Non-square matrix sensing without spurious local minima via the burer-monteiro approach," *arXiv preprint arXiv:1609.03240*, 2016.

[34] M. Yu, V. Gupta, M. Kolar *et al.*, "Recovery of simultaneous low rank and two-way sparse coefficient matrices, a nonconvex approach," *Electronic Journal of Statistics*, vol. 14, pp. 413–457, 2020.

[35] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "Global optimality in low-rank matrix optimization," *IEEE Transactions on Signal Processing*, vol. 66, pp. 3614–3628, 2018.

[36] R. F. Barber and W. Ha, "Gradient descent with nonconvex constraints: local concavity determines convergence," *arXiv preprint arXiv:1703.07755*, 2017.

[37] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, pp. 1985–2007, 2015.

[38] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," in *Advances in Neural Information Processing Systems*, 2015, pp. 109–117.

[39] ——, "Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent," *arXiv preprint arXiv:1605.07051*, 2016.

[40] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," *arXiv preprint arXiv:1507.03566*, 2015.

[41] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, "Dropping convexity for faster semi-definite optimization," in *Conference on Learning Theory*, 2016, pp. 530–582.

[42] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 665–674.

[43] S. Becker, V. Cevher, and A. Kyrillidis, "Randomized low-memory singular value projection," *arXiv preprint arXiv:1303.0167*, 2013.

[44] M. Soltani and C. Hegde, "Fast low-rank matrix estimation without the condition number," *arXiv preprint arXiv:1712.03281*, 2017.

[45] X.-T. Yuan, P. Li, and T. Zhang, "Gradient hard thresholding pursuit," *Journal of Machine Learning Research*, vol. 18, pp. 1–43, 2018.

[46] J. Shen and P. Li, "A tight bound of hard thresholding," *The Journal of Machine Learning Research*, vol. 18, pp. 7650–7691, 2017.

[47] R. Zhang, C. Josz, S. Sojoudi, and J. Lavaei, "How much restricted isometry is needed in nonconvex matrix recovery?" in *Advances in neural information processing systems*, 2018, pp. 5586–5597.

[48] R. Y. Zhang, S. Sojoudi, and J. Lavaei, "Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery," *arXiv preprint arXiv:1901.01631*, 2019.

[49] H. Liu, C. Gao, and R. J. Samworth, "Minimax rates in sparse, high-dimensional change-point detection," *arXiv preprint arXiv:1907.10012*, 2019.

[50] E. Page, "A test for a change in a parameter occurring at an unknown point," *Biometrika*, vol. 42, pp. 523–527, 1955.

[51] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, pp. 1590–1598, 2012.

[52] K. Frick, A. Munk, and H. Sieling, "Multiscale change point inference," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, pp. 495–580, 2014.

[53] T. Peng, C. Leckie, and K. Ramamohanarao, "Proactively detecting distributed denial of service attacks using source IP address monitoring," in *Networking 2004*, N. Mitrou, K. Kontovasilis, G. N. Rouskas, I. I., and L. Merakos, Eds. Berlin: Springer, 2004, pp. 771–782.

[54] J. A. Aston and C. Kirch, "Evaluating stationarity via change-point alternatives with applications to fMRI data," *The Annals of Applied Statistics*, vol. 6, pp. 1906–1948, 2012.

[55] N. R. Zhang, D. O. Siegmund, J. Hanlee, and J. Z. Li, "Detecting simultaneous changepoints in multiple sequences," *Biometrika*, vol. 97, pp. 631–645, 2010.

[56] T. Wang and R. J. Samworth, "High dimensional change point estimation via sparse projection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, pp. 57–83, 2018.

[57] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Prentice Hall Englewood Cliffs, 1993, vol. 104.

[58] M. Csörgő and L. Horváth, *Limit Theorems in Change-Point Analysis*. Wiley, Chichester, 1997.

[59] H. Ombao, R. Von Sachs, and W. Guo, "Slex analysis of multivariate nonstationary time series," *Journal of the American Statistical Association*, vol. 100, pp. 519–531, 2005.

[60] A. Aue, S. Hörmann, L. Horváth, and M. Reimherr, "Break detection in the covariance structure of multivariate time series models," *The Annals of Statistics*, vol. 37, pp. 4046–4087, 2009.

[61] C. Kirch, B. Muhsal, and H. Ombao, "Detection of changes in multivariate time series with application to eeg data," *Journal of the American Statistical Association*, vol. 110, pp. 1197–1216, 2015.

[62] L. Horváth and M. Hušková, "Change-point detection in panel data," *Journal of Time Series Analysis*, vol. 33, pp. 631–648, 2012.

[63] J. Bai, "Common breaks in means and variances for panel data," *Journal of Econometrics*, vol. 157, pp. 78–92, 2010.

[64] M. Jirak, "Uniform change point tests in high dimension," *The Annals of Statistics*, vol. 43, pp. 2451–2483, 2015.

[65] H. Cho and P. Fryzlewicz, "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 77, pp. 475–507, 2015.

[66] H. Cho, "Change-point detection in panel data via double CUSUM statistic," *Electronic Journal of Statistics*, vol. 10, pp. 2000–2038, 2016.

[67] J. A. Aston and C. Kirch, "High dimensional efficiency with applications to change point tests," *Electronic Journal of Statistics*, vol. 12, pp. 1901–1947, 2018.

[68] F. Enikeeva and Z. Harchaoui, "High-dimensional change-point detection under sparse alternatives," *The Annals of Statistics*, vol. 47, pp. 2051–2079, 2019.

[69] I. Cribben and Y. Yu, "Estimating whole-brain dynamics by using spectral clustering," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 66, pp. 607–627, 2017.

[70] D. Wang, Y. Yu, and A. Rinaldo, "Optimal covariance change point localization in high dimension," *arXiv preprint arXiv:1712.09912*, 2017.

[71] ——, "Optimal change point detection and localization in sparse dynamic networks," *arXiv preprint arXiv:1809.09602*, 2018.

[72] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection," *The Annals of Statistics*, vol. 41, pp. 670–692, 2013.

[73] Y. Chen, T. Wang, and R. J. Samworth, "High-dimensional, multiscale online changepoint detection," *arXiv preprint arXiv:2003.03668*, 2020.

[74] E. Arias-Castro, E. J. Candès, and A. Durand, "Detection of an anomalous cluster in a network," *The Annals of Statistics*, vol. 39, pp. 278–304, 2011.

[75] D. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *The Annals of Statistics*, vol. 32, pp. 962–994, 2004.

[76] E. Arias-Castro, D. L. Donoho, and X. Huo, "Near-optimal detection of geometric objects by fast multiscale methods," *IEEE Transactions on Information Theory*, vol. 51, pp. 2402–2425, 2005.

[77] Q. Berthet and P. Rigollet, "Optimal detection of sparse principal components in high dimension," *The Annals of Statistics*, vol. 41, pp. 1780–1815, 2013.

[78] O. Collier, L. Comminges, and A. B. Tsybakov, "Minimax estimation of linear and quadratic functionals on sparsity classes," *The Annals of Statistics*, vol. 45, pp. 923–958, 2017.

[79] D. Wang, Y. Yu, and A. Rinaldo, "Univariate mean change point detection: Penalization, CUSUM and optimality," *arXiv preprint arXiv:1810.09498*, 2018.

[80] H. Liu, C. Gao *et al.*, "Density estimation with contamination: minimax rates and theory of adaptation," *Electronic Journal of Statistics*, vol. 13, pp. 3613–3653, 2019.

[81] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.

[82] L. Devroye and G. Lugosi, "Combinatorial methods in density estimation," 2001.

[83] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer, 2009, vol. 11.

[84] M. Chen, C. Gao, and Z. Ren, "A general decision theory for huber's ϵ-contamination model," *Electronic Journal of Statistics*, vol. 10, pp. 3752–3774, 2016.

[85] B. Efron, "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis," *Journal of the American Statistical Association*, vol. 99, pp. 96–104, 2004.

[86] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, pp. 73–101, 1964.

[87] ——, "A robust version of the probability ratio test," *The Annals of Mathematical Statistics*, vol. 36, pp. 1753–1758, 1965.

[88] C. Gao, "Robust regression via mutivariate regression depth," *arXiv preprint arXiv:1702.04656*, 2017.

[89] M. Chen, C. Gao, and Z. Ren, "Robust covariance matrix estimation under huber's contamination model," *The Annals of Statistics (to appear)*, 2017.

[90] J. Jin and T. T. Cai, "Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons," *Journal of the American Statistical Association*, vol. 102, pp. 495–506, 2007.

[91] T. T. Cai and J. Jin, "Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing," *The Annals of Statistics*, vol. 38, pp. 100–145, 2010.

[92] R. J. Samworth *et al.*, "Recent progress in log-concave density estimation," *Statistical Science*, vol. 33, pp. 493–509, 2018.

[93] F. Yang, R. F. Barber *et al.*, "Contraction and uniform convergence of isotonic regression," *Electronic Journal of Statistics*, vol. 13, pp. 646–677, 2019.

[94] R. Dai, H. Song, R. F. Barber, and G. Raskutti, "The bias of isotonic regression," *Electronic journal of statistics*, vol. 14, art. no. 801, 2020.

[95] O. Lepskii, "On a problem of adaptive estimation in gaussian white noise," *Theory of Probability & Its Applications*, vol. 35, pp. 454–466, 1991.

[96] ——, "Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates," *Theory of Probability & Its Applications*, vol. 36, pp. 682–697, 1992.

[97] ——, "Asymptotically minimax adaptive estimation. ii. schemes without optimal adaptation: Adaptive estimators," *Theory of Probability & Its Applications*, vol. 37, pp. 433–448, 1993.

[98] O. Lepski and V. Spokoiny, "Optimal pointwise adaptive methods in nonparametric estimation," *The Annals of Statistics*, vol. 25, pp. 2512–2546, 1997.

[99] L. D. Brown and M. G. Low, "A constrained risk inequality with applications to nonparametric functional estimation," *The Annals of Statistics*, vol. 24, pp. 2524–2535, 1996.

[100] T. T. Cai, "Rates of convergence and adaptation over besov spaces under pointwise risk," *Statistica Sinica*, vol. 13, pp. 881–902, 2003.