COMPLEXITY, CONDITIONING, AND SADDLE AVOIDANCE IN NONSMOOTH OPTIMIZATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Mateo Díaz Díaz

August 2021

COMPLEXITY, CONDITIONING, AND SADDLE AVOIDANCE IN NONSMOOTH OPTIMIZATION

Mateo Díaz Díaz, Ph.D.

Cornell University 2021

Continuous optimization has become a prevalent tool across the sciences and engineering. Modern applications have displayed steady growth in problem sizes. Such sizes often prohibit the use of classical algorithmic solutions that heavily rely on costly operations, such as matrix inversion, and do not scale well. To counter this phenomenon practitioners have turned their focus to simpler first-order heuristics, such as gradient descent, that are often highly successful, yet are not well-understood. In this thesis, we study a few nonsmooth settings where simple algorithms are provably convergent.

We start with the problem of detecting infeasibility of large-scale linear programming problems using the primal-dual hybrid gradient method of Chambolle and Pock (2011). The literature on PDHG has focused chiefly on feasible problems. When the problem is not feasible, the iterates of the algorithm do not converge. In this scenario, we show that the iterates diverge at a controlled rate towards a well-defined ray. Leveraging this fact, we design a simple scheme to extract certificates of infeasibility from the iterates.

We then turn to unconstrained convex optimization and consider the classic proximal bundle methods, an algorithmic family dating back to the 70s. We prove convergence rates for bundle methods under a variety of assumptions. In particular, we show that these algorithms automatically adapt to problem regularity, exhibiting faster convergence rates. We complement these findings

with a new parallelizable variant of the bundle method that attains near-optimal rates without prior knowledge of function parameters. These results improve on the limited existing convergence rates and provide a unified approach across problem settings and algorithmic details.

After that, we study rapid local convergence guarantees for nonconvex formulations of low-rank matrix recovery problems, a problem family that includes phase retrieval, blind deconvolution, matrix completion, and robust PCA. Standard approaches for solving these problems use smooth penalty functions and often exhibit an undesirable phenomenon: the condition number, classically defined, scales poorly with the dimension. In contrast, we show that natural nonsmooth penalty formulations have two clear advantages: (1) they do not suffer from the same type of ill-conditioning, and (2) they are robust against noise and gross outliers. Consequently, we prove that off-the-shelf algorithms for nonsmooth optimization converge at a rapid dimension-independent rate when initialized close to the solution, even when a constant fraction of the measurements are adversarially corrupted.

To complement these local convergence guarantees, we turn to the question of escaping saddle points of nonsmooth functions. Recent work has shown that stochastically perturbed gradient methods can efficiently strict saddle points of smooth functions. We extend this body of work to nonsmooth optimization, by analyzing an inexact analogue of a stochastically perturbed gradient method applied to the Moreau envelope. The main conclusion is that a variety of algorithms can escape strict saddle points of the Moreau envelope at a controlled rate.

BIOGRAPHICAL SKETCH

Mateo Díaz was born in 1993. He grew up at the border between Brazil and Colombia in the neighboring towns of Tabatinga and Leticia. There, he started his education at Colegio Georges Charpak. He then moved to Bogotá in 2010, where he attended Universidad de Los Andes and received undergraduate degrees in Mathematics, and Systems and Computing Engineering, followed by a master's degree in Mathematics. He began his Ph.D. in the Center for Applied Mathematics at Cornell University in 2016. Upon completing his doctoral studies, he will start a postdoctoral appointment at Caltech.

A mis padres, a mis abuelos y a todos sus sueños imposibles.

ACKNOWLEDGEMENTS

I want to start by thanking my advisor, Damek Davis. A great deal of what I learned during my Ph.D. about the crafts of research, teaching, writing, and public speaking I learned from him. He spent countless hours with me in his office (and nowadays on zoom), helping me furnish and polish my ideas, papers, and talks. I'll miss our coffee interludes and the conversations about philosophy, meditation, and knowledge representation.

I would also like to thank the other members of my committee, Adrian Lewis, Jim Renegar, and Bobby Kleinberg. They have been a continuous source of inspiration and wisdom. Two of the best classes I took at Cornell, Algorithms and Convex Analysis, were taught by Bobby and Adrian. I hope one day I can deliver lectures with a pinch of their elegance and excitement.

I wish to thank Mauricio Velasco. While I was still an undergrad, Mauricio taught me about the beautiful interplay between mathematics and algorithms and made me fall madly in love with it. Even though he will never admit it, his influence shaped my path.

I have had the fortune of having many informal mentors throughout my Ph.D. I especially wish to thank Alex Townsend, Alex Vladmirsky, David Applegate, Dmitriy Drusvyatskiy, Javier Peña, Mauricio Junca, and Miles Lubin. Additionally, I'd like to thank all my collaborators, Adolfo Quiroz, Ben Grimmer, Brendan O'Donoghue, Carla Gomes, Felipe Rincón, Haihao Lu, Kaizheng Wang, Lijun Ding, Oliver Hinder, Ronan Le Bras, Vasilis Charisopoulos, Yudong Chen, and Yuling Yan.

I wish to thank the staff members that make CAM and ORIE work as well as they do. Primarily, I'd like to thank Erika Fowler-Decatur. CAM would not be the same without you, Erika. Thank you for all your help.

A wise person once told me "*More than a place, Ithaca is the people that you meet there*". My version of Ithaca wouldn't have been the same without Abby, Andrew, Avery, Ayah, Ben, Sergio, Camila, Jose, Heather, Marc, Matt, and Nick. Beyond Ithaca, I'd like to thank Gonche, Luis, Mariana, Ramiro, and Sergio for always being a phone call away.

I want to thank my closest friends, Artur, Maru, and Sebas, for keeping me sane throughout these winters and putting up with my random rants about mechanical keyboards, obscure Linux distributions, and DIY projects.

I am deeply thankful to Tatiana for the love, kindness, and companionship throughout all these years. And for believing in me even when I couldn't. Until you came along, I didn't know how ridiculously happy I could be with another human being — mil gracias, cosa.

Finally, I'd like to thank my family. If anything, this thesis is a tribute to their endless love, support, and sacrifice. A special thank you goes to my parents. Ma, Pa, en contra de todo pronóstico, lo lograron. Para ustedes todas las gracias hoy y siempre.

TABLE OF CONTENTS

1

INTRODUCTION

"Good problems and mushrooms of certain kinds have something in common; they grow in clusters."

— George Pólya, *How to Solve It: A New Aspect of Mathematical Method*

Optimization has become a widespread tool throughout the sciences and engineering. Modern applications dealing with learning and estimation tasks have led to a steady increase in problem sizes. Classical algorithms, such as simplex and interior-point methods, rely on costly operations and tend to struggle with such large-scale applications. Motivated by this drawback, practitioners have switched to simpler first-order heuristics, e.g., gradient descent, that are often highly successful, yet are not well-understood.

This thesis investigates several large-scale settings where simple first-order algorithms are provably convergent. In particular, in

- Chapter 3: We study huge-scale linear programming problems and propose a practical approach to detect infeasibility using the iterates of the primal-dual hybrid gradient method.
- Chapter 4: We analyze a classical and widely-used algorithm for convex optimization called the proximal bundle method. We design adaptive stepsize rules that lead to optimal convergence for a number of nonsmooth settings.
- Chapter 5: We consider nonconvex formulations for low-rank matrix recovery problems. We link notions of strong identifiability, e.g., restricted isometry properties, with favorable conditioning of nonsmooth optimization prob-

lems, leading to fast local convergence of off-the-shelf numerical algorithms, such as subgradient and prox-linear methods, for a suite of high-impact data science tasks.

- Chapter 6: We focus on the blind deconvolution problem and leverage the results in Chapter 5 to develop a two-stage method that can handle a constant fraction of gross outliers.
- Chapter 7: We investigate the question of escaping saddle points of non-smooth functions and design efficient algorithms that escape strict saddle points of the Moreau envelope of weakly convex functions.

In what follows, we describe the main contributions of these chapters in detail.

Infeasibility detection for large-scale linear programming

Our first subject of study is classical Linear Programming (LP). Formally, we consider the canonical primal-dual problems,

$$\begin{aligned} &\text{minimize} \quad c^\top x && \text{maximize} \quad b^\top y \\ &\text{subject to} \quad Ax = b && \text{subject to} \quad A^\top y \le c\,. \\ &\qquad\quad x \ge 0\,, \end{aligned} \tag{1.1}$$

where $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, and $c \in \mathbf{R}^n$ are given. The state-of-the-art algorithmic solutions — simplex [63] and interior point methods [188, 210] — have continuously improved over the past few decades and exhibit great practical performance for medium scale-problems, thus, leaving little room for first order methods to make inroads.

However, for large-scale problems, where the input data bearly fits in memory, these classical algorithms tend to struggle due to their dependence on ma-

trix inversion. To counter this drawback, a recent paper [14] proposed to apply the *Primal-Dual Hybrid Gradient method* (PDHG) [43] — a first-order method for general convex optimization problems — to LP. When specialized to (1.1) PDHG updates

$$\begin{aligned} x^{k+1} &= \text{proj}_{\mathbf{R}^n_+}(x^k - \eta A^\top y^k - \eta c) \\ y^{k+1} &= y^k + \tau A(2x^{k+1} - x^k) - \tau b \ . \end{aligned} \tag{1.2}$$

Said paper showed empirically that with the right enhancements, a PDHG-based solver presents moderate to significant gains compared to state-of-the-art solvers in the large scale regime.

Nonetheless, the existing theory for PDHG falls short of providing theoretical foundations for many of the features that a modern LP solver requires. In paticular, the literature has mostly focused on settings where the problem at hand is assumed to have at least one solution. Infeasibility detection and computation of certificates are an essential aspect of solving LP, not only to provide feedback on modeling errors but also for algorithms that directly exploit LP certificates such as Benders decomposition, and branch-and-cut [2]. Chapter 3 studies the problem of detecting infeasibility of LP problems using PDHG. When the problem is not feasible, the iterates of the algorithm do not converge. In this scenario, we show that the iterates diverge at a controlled rate towards a well-defined ray. The direction of this ray is known as the infimal displacement vector v.

The first contribution of Chapter 3 is to prove that v recovers certificates of primal and dual infeasibility whenever they exist. Based on this fact, we propose a simple way to extract approximate infeasibility certificates from the iterates of PDHG. We study three different sequences that converge to the infimal

displacement vector:

$$\begin{aligned} \textbf{(Difference of iterates)} \quad & (z^{k+1} - z^k), \\ \textbf{(Normalized iterates)} \quad & \frac{z^k}{k}, \\ \textbf{(Normalized average iterates)} \quad & \frac{2}{k+1}\bar{z}^k. \end{aligned}$$

where $z^k = (x^k, y^k)$ denotes the kth iterate and $\bar{z}^k = \frac{1}{k}\sum_{j=1}^{k} z^j$, the average of iterates. All of them are easy to compute, and thus the approach is suitable for large-scale applications.

Our second contribution is to establish tight convergence rates for these sequences. We demonstrate that the normalized iterates and the normalized average achieve a convergence rate of $O\left(\frac{1}{k}\right)$, improving over the known rate for the difference $O\left(\frac{1}{\sqrt{k}}\right)$ [163]. This rate is general and applies to any fixed-point iteration of a nonexpansive operator. Thus, the result covers a broad family of algorithms beyond PDHG, including, for example, the Alternating Direction Method of Multipliers (ADMM), and can be applied settings beyond linear programming, such as quadratic and semidefinite programming. Further, in the case of linear programming, we show that, under non-degeneracy assumptions, the iterates of PDHG identify the active set of an auxiliary feasible problem in finite time, ensuring that the difference of iterates exhibits eventual linear convergence to the infimal displacement vector.

Optimal convergence rates for the proximal bundle method

Next, we turn to the arguably more general problem of unconstrained convex optimization. Formally, we aim to minimize

$$\text{minimize}_{x \in \mathbf{R}^d} f(x)$$

where $f \colon \mathbf{R}^d \to \mathbf{R}$ is a convex function. We consider the classic *proximal bundle methods* [143, 242], an algorithmic family dating back to the 70s. They are conceptually similar to model-based methods [66, 193, 85]. That is, methods that update their iterates by applying a proximal step to an approximation of the function, known as the model f_k:

$$x_{k+1} \leftarrow \arg\min_x f_k(x) + \frac{\rho_k}{2}\|x - x_k\|^2. \tag{1.3}$$

However, bundle methods only update the next iterate x_{k+1} when the decrease in objective value is at least a fraction of the decrease that the model predicted. If the next iterate is not updated, they use the solution of (1.3) to update f_k. A common choice for the model is

$$f_k(x) = \max_j f(z_j) + \langle g_j, x - z_j \rangle$$

where $\{z_j\}$ are the solutions to (1.3) and $g_j \in \partial f(z_j)$ are subgradients. Unlike other local algorithms, bundle methods retain information about the geometry of the function around many iterates as opposed to the last one.

Though bundle methods are known to converge under different settings [128, 177, 12] and have been successfully used in applications [221, 220, 73], nonasymptotic guarantees have remained mostly elusive. In Chapter 4, we prove convergence rates for bundle methods under a variety of assumptions. We show that, without any modification, these algorithms adapt to converge

faster in the presence of smoothness or Hölder growth. Our analysis reveals that with a constant stepsize, the bundle methods are adaptive, yet they attain suboptimal convergence rates.

We overcome this shortcoming by proposing nonconstant stepsize schemes with optimal rates. These schemes use function information such as growth constants, which might be prohibitive in practice. We complete the chapter with a new parallelizable variant of the bundle method that attains near-optimal rates without prior knowledge of function parameters. These results improve on the limited existing convergence rates and provide a unified approach across problem settings and algorithmic details.

Composite optimization for low-rank matrix recovery

The task of recovering a structured signal from its noisy measurements plays a central role in data science. Relevant examples include compressed sensing, phase retrieval, matrix completion, and robust PCA. Optimization-based approaches naturally lead to nonconvex formulations, which are NP-hard in general. To bypass this issue, researchers developed convex relaxations based on linear and semidefinite programming [80, 36, 231, 46, 41, 38, 77]. These relaxations enjoy strong guarantees and can be solved in polynomial time. Yet, in practice, they do not scale well. This motivated the community to change its focus to nonconvex iterative methods better suited for large scale datasets [54, 39, 190, 228].

In Chapter 5, we investigate nonsmooth nonconvex formulations for a handful of concrete recovery problems. We show that these nonsmooth formulations

present two clear advantages over their smooth counterparts: first, they are robust against gross outliers, and second, their condition number does not degrade as the dimension grows. In turn, this implies that local algorithms, such as the *subgradient and prox-linear method*, are robust and exhibit fast dimension-independent convergence rates.

Let us elaborate. For the analysis we consider a broad class of functions that extends the convex and smooth classes. Given finite dimensional Euclidian spaces $\mathbf{E}$ and $\mathbf{Y}$, we study $f\colon \mathbf{E} \to \mathbf{R} \cup \{\infty\}$ with

$$f = h \circ F \tag{1.4}$$

where $F\colon \mathbf{E} \to \mathbf{Y}$ is a smooth map and $h\colon \mathbf{Y} \to \mathbf{R} \cup \{\infty\}$ is a convex function.

To illustrate our results, recall the *quadratic sensing problem*, a problem with applications to X-ray crystallography, astronomy, and microscopy among others [27, 173, 238]. The objective of the problem is to recover a matrix $\bar{X} \in \mathbf{R}^{d\times r}$ from a set of m quadratic measurements

$$b = \mathcal{A}(\bar{X}\bar{X}^\top) + \xi \in \mathbf{R}^m \quad \text{with} \quad \mathcal{A}_i(M) = p_i^\top M p_i,$$

where the p_i's are known random vectors and ξ represents noise. Notably, when the measurement vectors $(p_i)_{i=1}^n$ are sampled values of complex sinusoids and $r = 1$, these measurements correspond to X-ray diffraction images, an imagining modality that enabled the discovery of the double helix [241]. In this context, we propose to minimize

$$\arg\min_X f(X) \qquad \text{where} \qquad f(X) = \frac{1}{m}\|\mathcal{A}(XX^\top) - b\|_1. \tag{1.5}$$

Notice that f decomposes as (1.4) with $h(\cdot) = \|\cdot\|_1$ and $F(\cdot) = \mathcal{A}(\cdot) - b$.

To solve this problem we study the subgradient method, which iterates

$$X_{k+1} \leftarrow X_k - \alpha_k G_k \qquad \text{with} \qquad G_k \in \partial f(X_k)$$

where $\partial f(X)$ is the subdifferential set of f at X, a generalization of the gradient mapping, and $\alpha_k > 0$ are stepsizes. We also investigate the prox-linear method, which recursively updates

$$X_{k+1} \leftarrow \min_Y f_{X_k}(Y) + \frac{\beta}{2}\|Y - X_k\|^2 \quad \text{where} \qquad f_X(Y) = h(F(X) + \nabla F(X)^\top (Y - X)).$$

where $\beta > 0$ is a fixed parameter. In other words, at each iteration, we minimize the composition of h with a linear approximation of F at X_k plus a quadratic term. The quadratic ensures that the next iterate is not far from a region where the approximation is good. The inner problems are convex and can be solved efficiently with first-order methods, such as ADMM or PDHG.

Rates for these algorithms were understood for convex functions [109, 31], but have only recently been studied in the nonconvex setting [68, 82]. Inspired by this line of work, we show that the subgradient and prox-linear methods exhibit linear and quadratic convergence, respectively, as long as the function f satisfies the following local regularity properties. For any X, Y near the solution set S:

$$\begin{aligned} \textbf{(Sharp growth)} \quad & f(x) - \inf f \geq \mu \cdot \operatorname{dist}(x, S), \\ \textbf{(Lipschitz)} \quad & |f(X) - f(Y)| \leq L \cdot \|X - Y\|, \\ \textbf{(Quadratic approximation)} \quad & |f(x) - f_X(Y)| \geq \frac{\rho}{2} \cdot \|X - Y\|^2. \end{aligned}$$

The first two inequalities parallel the variational description of the minimum and maximum singular values of linear maps and, in fact, the condition number L/μ determines the efficiency rates of the methods. While the third condition is linked to the size of the basin of attraction where the algorithms are fast.

In turn, these regularity conditions are closely related to the restricted isometry property (RIP) of $\frac{1}{m}\mathcal{A}$; a seminal concept that ignited a decade of theoretical

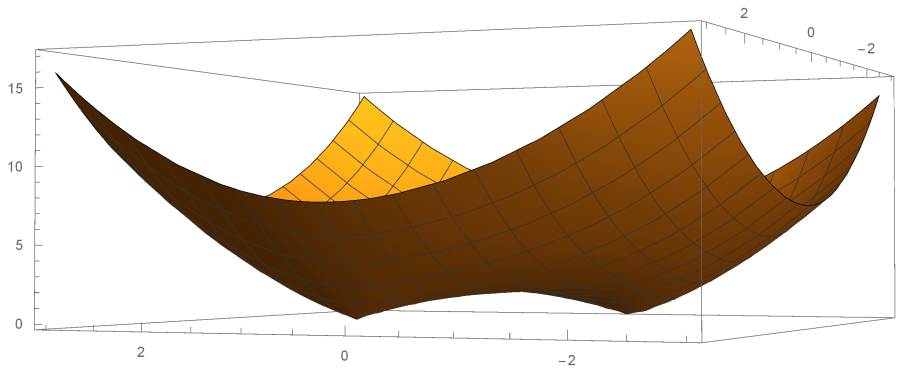


Figure 1.1: Expected quadratic sensing loss $\mathbb{E}f(x) = \mathbb{E}|(p^\top x)^2 - (p^\top \mathbf{1})^2|$

and computational advances in compressive sampling [80, 38, 231]. By leveraging this connection, we prove that the quadratic sensing loss (1.5) satisfies the desired regularity conditions, with high probability, as soon as the number of measurements m exceeds a constant multiple of $d \cdot r$, the information-theoretical limit required for recovery. Further, we show these properties hold even when a constant fraction of the measurements is corrupted by gross outliers.

We use this framework to analyze several statistical recovery problems. In particular, Chapter 5 sketches a similar picture for bilinear sensing, matrix completion, and robust PCA.

In Chapter 6, we specialize these local convergence rates to the so-called *blind deconvolution* problem — a rank-one bilinear recovery problem with applications to signal processing — and complement them with a robust spectral initialization method. We prove that using this initialization algorithm in tandem with any of the two local refinement methods provides a convergent algorithm that can stand a constant fraction of gross outliers.

Even though the initialization procedure is necessary to make the theory hold, numerical experiments show that a randomly initialized subgradient

method consistently solves the blind deconvolution problem. In a preliminary attempt to understand this phenomenon, Chapter 6 characterizes the critical points of the nonsmooth blind deconvolution problem and shows that the set of spurious critical points concentrate near a co-dimension two subspace. Thus, suggesting that there is a vast region of the space with benign geometry.

Escaping strict saddle points of weakly-convex functions

Though nonconvex optimization problems are NP-hard in general, simple nonconvex optimization techniques, e.g., gradient descent, are broadly used and often highly successful in high-dimensional statistical estimation and machine learning problems. For smooth formulations, a common explanation for this phenomenon is that nonconvex functions found in machine learning have benign geometry: all local minima are (nearly) global minima, and all saddle points are strict — meaning that they have a direction of negative curvature. This explanation is well-grounded: several important estimation and learning problems have amenable geometry [106, 226, 23, 105, 227, 240] and recent works [124, 123] have shown that when this property holds, stochastically perturbed gradient methods can efficiently converge to a global minimum.

While impressive in scope, these works fall short of establishing rates in the nonsmooth setting. In Chapter 7, we propose and analyze an algorithm that extends these ideas to the context of ρ-weakly convex functions; functions f for which $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex. This is a large family of nonsmooth nonconvex functions that contains, for example, the composite class (1.4).

Weakly convex functions admit a global C^1 smoothing: for all $\mu < \rho^{-1}$, define

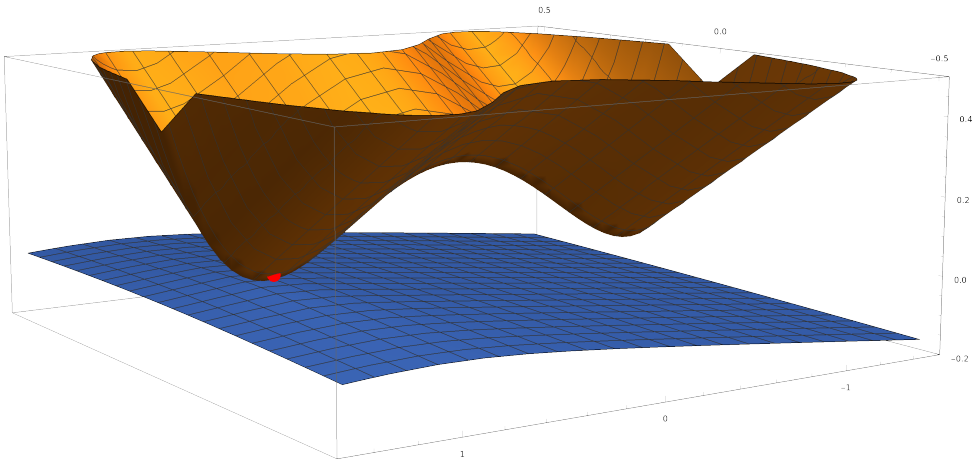


Figure 1.2: Function $f(x, y) = |x| + \frac{1}{4}(y^2 - 1)^2$, a point $(x, f(x))$ with x an approximate second-order critical point of f_μ and its corresponding quadratic $q(\cdot)$.

the *Moreau envelope* and the *proximal mapping* to be

$$f_\mu(x) = \min_{y \in \mathbf{R}^d} \; f(y) + \frac{1}{2\mu}\|y - x\|^2 \quad \text{and} \quad \mathbf{prox}_{\mu f}(x) = \arg\min_{y \in \mathbf{R}^d} \left\{ f(y) + \frac{1}{2\mu}\|y - x\|^2 \right\},$$

respectively. Moreover, the first order critical points of f and f_μ agree. Although the Moreau envelope f_μ is not C^2 in general, we prove that for generic semialgebraic functions it is C^2 around any x with $\|\nabla f_\mu(x)\| \approx 0$.

Inspired by these facts, we investigate the problem of finding an $(\varepsilon_1, \varepsilon_2)$-second-order critical point x of f_μ:

$$\|\nabla f_\mu(x)\| \le \varepsilon_1 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 f_\mu(x)) \ge -\varepsilon_2. \tag{1.6}$$

We argue that the geometry of f around any such x is favorable. Indeed, (1.6) implies the existence of an approximate quadratic minorant q of f with small slope and curvature at a nearby point, see Figure 1.2.

We propose an efficient outer-inner loop scheme for the task of finding x satistying (1.6). The outer loop executes a perturbed and inexact variation of the proximal point method, while the inner loop solves the proximal subproblem. For the inner loop, we consider model-based methods (1.3), and estab-

lish nonasymptotic complexity guarantees. The theory covers a comprehensive class of algorithms, including variants of the subgradient, prox-gradient, and prox-linear methods.

1.1 A comment about structure and related publications

This manuscript assumes a certain familiarity with nonsmooth analysis and high dimensional probability. However, for convenience, we have compiled most of the necessary notation and background in Chapter 2. The remaining chapters in this thesis follow a similar structure. They start with a few sections, introducing the problem of interest, an algorithmic solution, and theoretical results, followed by a section with numerical experiments. We defer long proofs to the last section of each chapter named "Analysis", which might be omitted in a first read.

This thesis wouldn't have been possible without my incredible collaborators:

- **Chapter 3** is based on joint work with David Applegate, Haihao Lu, and Miles Lubin [13]. This work was done during an internship at Google.
- **Chapter 4** is based on joint work with Ben Grimmer [76]. This project was a byproduct of a topics course taught by Adrian Lewis.
- **Chapter 5** is based on joint work with Vasilis Charisopoulos, Yudong Chen, Damek Davis, Lijun Ding, and Dmitriy Drusvyatskiy [48].
- **Chapter 6** is based on joint work with Vasilis Charisopoulos, Damek Davis, and Dmitriy Drusvyatskiy [49] and [75].
- **Chapter 7** is based on joint work with Damek Davis and Dmitriy Drusvyatskiy [69].

2

PRELIMINARIES

"Aunque siembren las raíces como les dé la gana,
los palos de guanábana no dan manzanas."

— Residente, *Hijos del cañaveral*

In this chapter, we summarize the notation and results we will use throughout this thesis. We note that the material presented here is not new but serves as the starting point for our studies.

2.1 Notation

Henceforth, the symbols $\mathbf{E}$ and $\mathbf{Y}$ will denote a Euclidean spaces with inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\|_2 = \sqrt{\langle x, x \rangle}$. The symbol $\mathbf{B}$ will denote the closed unit ball in $\mathbf{E}$, while a closed ball of radius $\epsilon > 0$ around a point x will be written as $B_\epsilon(x)$. For any point $x \in \mathbf{E}$ and a set $Q \subset \mathbf{E}$, the distance and the nearest-point projection in ℓ_2-norm are defined by

$$\operatorname{dist}(x; Q) = \inf_{y \in Q} \|x - y\|_2 \qquad \text{and} \qquad \operatorname{proj}_Q(x) = \arg\min_{y \in Q} \|x - y\|_2,$$

respectively. The symbol $\operatorname{cl}(Q)$ denotes the closure of Q. For any pair of functions f and g on $\mathbf{E}$, the notation $f \lesssim g$ will mean that there exists a numerical constant C such that $f(x) \leq Cg(x)$ for all $x \in \mathbf{E}$. Given a linear map between Euclidean spaces, $\mathcal{A} \colon \mathbf{E} \to \mathbf{Y}$, the adjoint map will be written as $\mathcal{A}^* \colon \mathbf{Y} \to \mathbf{E}$. Given a map $T : \mathbf{R}^d \to \mathbf{R}^d$, its range is defined as $\operatorname{range}(T) = \{T(z) \mid z \in \mathbf{R}^d\}$. Given two mappings $T_1, T_2 : \mathbf{R}^d \to \mathbf{R}^d$, we denote their composition as $T_1 \circ T_2$, that is

$T_1 \circ T_2(z) = T_1(T_2(z))$. We will use I_d for the d-dimensional identity matrix and $\mathbf{0}$ for the zero matrix with variable sizes. The symbol $[m]$ will be shorthand for the set $\{1, \ldots, m\}$.

We will always endow the Euclidean space of vectors $\mathbf{R}^d$ with the usual dot-product $\langle x, y \rangle = x^\top y$ and the induced ℓ_2-norm. More generally, the ℓ_p norm of a vector x will be denoted by $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$. We use $\mathrm{supp}(x) := \{i \in [d] \mid x_i \neq 0\}$ to denote the support of the vector x. We will equip the space of rectangular matrices $\mathbf{R}^{d_1 \times d_2}$ with the trace product $\langle X, Y \rangle = \mathrm{Tr}(X^\top Y)$ and the induced Frobenius norm $\|X\|_F = \sqrt{\mathrm{Tr}(X^\top X)}$. The operator norm of a matrix $X \in \mathbf{R}^{d_1 \times d_2}$ will be written as $\|X\|_{\mathrm{op}}$. The symbol $\sigma(X)$ will denote the vector of singular values of a matrix X in nonincreasing order. With this notation, we may equivalently write $\|X\|_{\mathrm{op}} = \sigma_1(X)$ and $\|X\|_F = \|\sigma(X)\|_2$. We also define the row-wise matrix norms $\|X\|_{b,a} = \|(\|X_{1\cdot}\|_b, \|X_{2\cdot}\|_b \ldots, \|X_{d_1\cdot}\|_b)\|_a$. We denote the pseudo inverse of X by $X^\dagger$. Given a matrix $M \in \mathbf{R}^{d \times d}$ we use $\lambda_1(M), \ldots, \lambda_d(M)$ to denote its eigenvalues. We define the spectral radius of a matrix M as $\rho(M) = \max_{j \in [d]} |\lambda_j(M)|$. We use the symbol $M \succ 0$ to denote that M is positive definite. Every positive definite matrix $M \succ 0$ defines an inner product and norm given by $\langle x, y \rangle_M = x^\top M y$ and $\|x\|_M^2 = \langle x, x \rangle_M$, respectively. The symbols $\mathcal{S}^d$, $\mathcal{S}^d_+$, $O(d)$, and $GL(d)$ will denote the sets of symmetric, positive semidefinite, orthogonal, and invertible matrices, respectively.

2.2 Nonsmooth analysis

Nonsmooth functions will play a central role in this work. Consequently, we will require some basic constructions of generalized differentiation, as described

for example in the monographs [214, 182, 24]. Consider a function $f : \mathbf{E} \to \mathbf{R} \cup \{+\infty\}$ and a point x, with $f(x)$ finite. The *subdifferential* of f at x, denoted by $\partial f(x)$, is the set of all vectors $\xi \in \mathbf{E}$ satisfying

$$f(y) \geq f(x) + \langle \xi, y - x \rangle + o(\|y - x\|_2) \quad \text{as } y \to x. \tag{2.1}$$

Here $o(r)$ denotes any function satisfying $o(r)/r \to 0$ as $r \to 0$. Thus, a vector ξ lies in the subdifferential $\partial f(x)$ precisely when the linear function $y \mapsto f(x) + \langle \xi, y - x \rangle$ lower-bounds f up to first-order around x. Standard results show that for a convex function f the subdifferential $\partial f(x)$ reduces to the subdifferential in the sense of convex analysis, in the sense that

$$f(y) \geq f(x) + \langle \xi, y - x \rangle \quad \text{for all } y \in \mathbf{E}.$$

While for a differentiable function it consists only of the gradient $\partial f(x) = \{\nabla f(x)\}$. We say that a point x is *stationary* for f whenever the inclusion $0 \in \partial f(x)$ holds. Equivalently, stationary points are precisely those that satisfy first-order necessary conditions for minimality: the directional derivative is nonnegative in every direction.

For any closed convex set $C \subseteq \mathbf{E}$, we define its *normal cone* at $\bar{x} \in C$ to be

$$N_C(\bar{x}) := \{g \in \mathbf{R}^d \mid \langle g, x - \bar{x} \rangle \leq 0 \quad \text{for all } x \in C\}.$$

Analogously, $N_C(\bar{x})$ can be seen as the set of all points x such that $\text{proj}_C(x) = \bar{x}$. For convex functions the subdifferential satisfies a nice of set of calculus rules. For any pair of closed convex functions $f_1 : \mathbf{E} \to \mathbf{R}$, $f_2 : \mathbf{E} \to \mathbf{R} \cup \{\infty\}$, and closed convex set C, we have

$$\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x) \qquad \text{and} \qquad \partial \iota_C(x) = N_C(x) \qquad \text{for all } x \in \mathbf{E}. \tag{2.2}$$

Additionally, for any closed convex functions $h\colon \mathbf{Y} \to \mathbf{R}$ and $g\colon \mathbf{E} \to \mathbf{R} \cup \{+\infty\}$ and C^1-smooth map $F\colon \mathbf{E} \to \mathbf{Y}$, the chain rule holds:

$$\partial(h \circ F + g)(x) = \nabla F(x)^* \partial h(F(x)) + \partial g(x).$$

We say that a function $f\colon \mathbf{E} \to \mathbf{R} \cup \{+\infty\}$ is *ρ-weakly convex*[1] if the function $x \mapsto f(x) + \frac{\rho}{2}\|x\|_2^2$ is convex. This encompasses a broad family of nonsmooth nonxonvex functions. In particular, composite functions $f = h \circ F$ satisfying the approximation guarantee

$$|f_x(y) - f(y)| \le \frac{\rho}{2}\|y - x\|_2^2 \qquad \forall x, y$$

are automatically ρ-weakly convex [83, Lemma 4.2]. Subgradients of weakly convex functions are very well-behaved. Indeed, notice that in general the little-o term in the expression (2.1) may depend on the basepoint x, and may therefore be nonuniform. The subgradients of weakly convex functions, on the other hand, automatically satisfy a uniform type of lower-approximation property. Indeed, a lower-semicontinuous function f is ρ-weakly convex if and only if it satisfies:

$$f(y) \ge f(x) + \langle \xi, y - x \rangle - \frac{\rho}{2}\|y - x\|_2^2 \qquad \forall x, y \in \mathbf{E}, \xi \in \partial f(x).$$

Although such functions are nonsmooth in general, they admit a smoothing: for all $\mu < \rho^{-1}$, we define the *Moreau envelope* and the *proximal mapping* to be

$$f_\mu(x) := \min_{y \in \mathbf{R}^d} f(y) + \frac{1}{2\mu}\|y - x\|^2 \quad \text{and} \quad \mathbf{prox}_{\mu f}(x) := \operatorname*{arg\,min}_{y \in \mathbf{R}^d} \left\{ f(y) + \frac{1}{2\mu}\|y - x\|^2 \right\}, \tag{2.3}$$

[1] Weakly convex functions also go by other names such as lower-C^2, uniformly prox-regularity, paraconvex, and semiconvex. We refer the reader to the seminal works on the topic [213, 204, 192, 215, 11].

respectively. These two constuctions are well-defined thanks to weak-convexity and moreover the Moreau envelope is a C^1 function [66]. The proximal map can be characterized via the subdifferential as

$$x \in (I + \partial f)(x^+) \iff x^+ = \mathbf{prox}_f(x). \tag{2.4}$$

2.3 High-dimensional probability

We will leverage the concentration of measure phenomena to control several random quantities. Here, we summarize the main inequalities that we use and refer the interested reader to the excellent monographs [235, 236, 25].

We start with the quintessential Hoeffding's inequality for symmetric Bernoulli random variables X, i.e., $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = 1/2$.

Theorem 2.3.1 (Theorem 2.2.2 in [235])**.** *Let $X_1, \dots, X_N$ be independent symmetric Bernoulli random variables. Then for any $t \geq 0$, we have*

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

Similarly, Bernstein's Inequality gives a bound that depends on the variance.

Theorem 2.3.2 (Theorem 2.8.4 in [235])**.** *Let $X_1, \dots, X_N$ be independent mean-zero random variables, such that for $|X_i| \leq K$ for all i. Then for any $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\left(\sigma^2 + Kt/3\right)}\right)$$

here $\sigma^2 = \sum \mathbb{E}[X_i^2]$ is the variance of the sum.

A common way to generalize concentration for unbounded random variables is to consider variables X with Gaussian-like tails: for some $\eta > 0$,

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{\eta^2 t^2}{2}\right) \quad \text{for all } t > 0.$$

In turn, this is equivalent (up to constants) to

$$\mathbb{E} \exp\left(\frac{X^2}{\eta^2}\right) \leq 2. \tag{2.5}$$

A random variable X is η-sub-gaussian whenever this inequality holds. One of the benefits of using this definition is that it defines a norm. The *sub-gaussian norm*[2] of a random variable X is given by

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E} \exp\left(\frac{X^2}{t^2}\right) \leq 2\right\}.$$

Theorem 2.3.3 (Theorem 2.6.3 in [235])**.** *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-gaussian random variables and $(a_1, \ldots, a_N) \in \mathbf{R}^N$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i X_i\right| \geq t\right) \leq 2 \exp\left(-\frac{ct^2}{K^2 \|a\|_2^2}\right)$$

where $K = \max_i \|X_i\|_{\psi_2}$.

Concentration also manifests for random variables with heavier tails, albeit at worst rates. We define *sub-exponential norm* of a random variable X as

$$\|X\|_{\psi_1} = \inf\left\{t > 0 : \mathbb{E} \exp\left(\frac{|X|}{t}\right) \leq 2\right\}$$

and say that X is η-sub-exponential if $\|X\|_{\psi_1} \leq \eta$.

Theorem 2.3.4 (Theorem 2.8.2 in [235])**.** *Let $Z_1, \ldots, Z_m$ be an independent, mean zero, sub-exponential random variables and let $a \in \mathbf{R}^m$ be a fixed vector. Then, for any $t \geq 0$ we have that*

$$\mathbb{P}\left(\sum_{i=1}^m a_i Z_i \leq -t\right) \leq \exp\left(-c \min\left\{\frac{t^2}{K^2 \|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right\}\right)$$

where $K := \max_i \|Z_i\|_{\psi_1}$ and $c > 0$ is a numerical constant.

[2] Also known as Orlicz-2 norm.

More generally, we say a that a random vector X in $\mathbf{R}^d$ is η-sub-gaussian or η-sub-exponential if its projection $\langle u, X \rangle$ onto any direction $u \in \mathbb{S}^{d-1}$ is sub-gaussian or sub-exponential, respectively.

3

INFEASIBILITY DETECTION WITH THE PRIMAL-DUAL HYBRID GRADIENT METHOD

"Sonhar mais um sonho impossível
Lutar quando é fácil ceder."

— Maria Bethânia, *Sonho Impossível*

3.1 Introduction

First-order methods (FOMs) have been extensively studied by the optimization community since the late 2000s, following a long period where interior-point methods dominated research in continuous optimization. FOMs are appealing for their simplicity and low computational overhead, in particular when solving large-scale optimization problems that arise in machine learning and data science. These methods have matured in many aspects [21] and are known to be useful for obtaining moderately accurate solutions to convex and non-convex optimization problems in a reasonable amount of time. Despite this progress, FOMs have made only modest inroads into linear programming (LP), a fundamental problem in mathematical optimization.

FOMs applied to LP provide relatively simple methods whose most expensive operations are matrix-vector multiplications with the (typically sparse) constraint matrix. Such matrix-vector products are amenable to scale efficiently given increasingly ubiquitous computing resources like multi-threaded CPUs,

Algorithm 1: Primal-dual hybrid gradient

Data: $x_0 \in \mathbf{R}^d$

Step k**:** $(k \geq 0)$

Update $x^{k+1} \leftarrow \mathbf{prox}_{\eta f}(x^k - \eta A^\top y^k)$,

Update $y^{k+1} \leftarrow \mathbf{prox}_{\tau h}\left(y^k + \tau A(2x^{k+1} - x^k)\right)$.

GPUs [232], or distributed clusters [93]. In contrast, interior-point and simplex-based methods that dominate current practice are limited in how they use available computing resources because they depend on matrix inversion. To mark this distinction, Nesterov [187] defines methods that use at most matrix-vector products as capable of handling *large-scale* problems and methods that use matrix inversion as handling *medium-scale* problems. In the context of LP, these definitions of scale perhaps belie the reliability and practical efficiency of interior-point and simplex methods, but nevertheless the contrast in the computing requirements of the algorithms is an important one. Even though such computational aspects are outside the scope of this work, it is this practical potential to efficiently solve large-scale LP that motivates the theoretical developments in this work.

While FOMs are typically studied in more general settings, the underlying assumptions and convergence rates in these settings do not necessarily hold or may not be tight for the special case of LP. Of particular relevance to this work, theory for FOMs is often developed under the assumption that an optimal solution exists, whereas LP solvers need to be able to detect infeasibility (i.e., when no optimal solution exists) and compute corresponding certificates. Infeasibility detection and computation of certificates are an essential aspect of solving LP, not only to provide feedback on modeling errors but also for algorithms that directly exploit LP certificates like Benders decomposition and branch-and-cut [2].

This chapter addresses the question of how to detect infeasibility in LP using the *Primal-Dual Hybrid Gradient method* (PDHG). PDHG is a popular first-order method introduced by Chambolle and Pock [43] to solve *convex-concave minimax problems*, that is, problems of the form

$$\min_{x \in \mathbf{R}^n} \max_{y \in \mathbf{R}^m} \langle Ax, y \rangle + g(x) - h(y) \tag{3.1}$$

where $g : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ and $h : \mathbf{R}^m \to \mathbf{R} \cup \{\infty\}$ are proper lower semicontinuous convex functions and $A \in \mathbf{R}^{m \times n}$. LP can be recast as a minimax problem through duality, and hence PDHG is applicable. The method consists of alternating updates between the primal and dual variables, see Algorithm 1, where **prox** is the proximal operator (see the definition in (2.3)). In particular, when instantiated for LP, these updates correspond to matrix-vector products and projections onto simple sets (such as the positive orthant). In contrast with other methods like the *Alternating Direction Method of Multipliers* (ADMM), PDHG does not require projections onto linear subspaces, which involve matrix inversions by direct or indirect methods.

The behavior of PDHG for *feasible* problems (i.e., problems that have an optimal solution) has been studied in depth under several regularity assumptions. In their seminal work, Chambolle and Pock [43] show that the algorithm converges at a rate of $O(1/k)$ given appropriate choices for the step sizes η and τ. However, the situation for infeasible problems remains largely unstudied.

While it is relatively straightforward to formulate always-feasible auxiliary problems that can be used to detect infeasibility, for example, by penalizing violations of primal and dual constraints, this approach is unappealing for two reasons: First is the aesthetic interest of having a single algorithm that robustly handles all possible input [136]. Second is the practical interest in effectively

using available computing resources, as solving such auxiliary problems would approximately double the necessary work. Instead, we aim to use *one* execution of PDHG and ask the following question:

Do the PDHG iterates encode information about infeasibility?

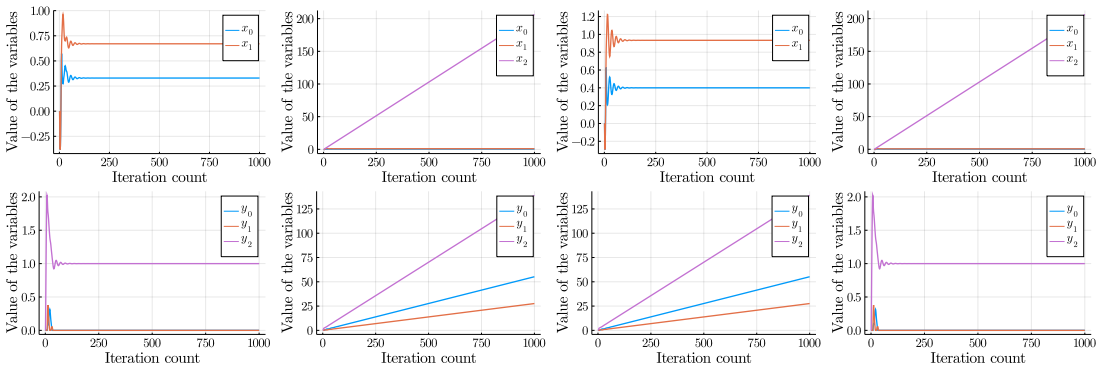
We answer this question in the affirmative. We show that if the primal (and/or dual) problem is infeasible, the iterates of PDHG recover primal (and/or dual) infeasibility certificates. Moreover, we completely characterize the behavior of the iterates under different infeasibility settings. Before diving into our main contributions, let us present an illustrative example. Recall that for a primal-dual LP pair, there exist three exhaustive and mutually exclusive possibilities: (1) both primal and dual are feasible, (2) both primal and dual are infeasible, and (3) one of the two problems is unbounded, and consequently, the other problem is infeasible. Small numerical experiments reveal that the behavior of PDHG is different depending on the setting.

Example 3.1.1. *Consider the LP problem with constants $\alpha, \beta \in \mathbf{R}$:*

$$\begin{aligned} \textit{minimize} \quad & x_0 + x_1 - \alpha x_2 \\ \textit{subject to} \quad & x_0 + 2x_1 \leq 2 \\ & 3x_0 + x_1 \leq 2 \\ & x_0 + x_1 \geq \beta \,. \end{aligned}$$

Figure 3.1 displays four choices of α and β

1. ***Both feasible**. Set $\alpha = 0$ and $\beta = 1$, then both primal and dual problems are feasible. In this case, both the primal and dual variables converge to a solution.*



(a) Both feasible (b) Both infeasible (c) Unbound. dual (d) Unbound. primal

Figure 3.1: Four different settings depicted in Example 3.1.1. Every subplot shows the component-wise value of the iterates against the iteration count. The first and the second rows correspond to the primal and dual iterates, respectively.

2. ***Both infeasible***. *Set $\alpha = 1$ and $\beta = 2$, then both primal and dual are infeasible. We observe that both primal and dual iterates diverge at a rate proportional to the number of iterations.*

3. ***Unbounded dual***. *Set $\alpha = 0$ and $\beta = 2$, then the dual problem is unbounded and, thus, the primal problem is infeasible and the dual is feasible. Then the dual iterates diverge, and, interestingly, the primal iterates converge.*

4. ***Unbounded primal***. *Set $\alpha = 1$ and $\beta = 1$, then the primal problem is unbounded and, thus, the dual problem is infeasible and the primal is feasible. Then the primal iterates diverge, and the dual iterates converge.*

From the experiments, we see that the iterates have a very stable asymptotic behavior. In particular, if the primal is feasible, then the dual variables converge, and analogously if the dual is feasible, then the primal iterates converge. Similarly, whenever the primal is infeasible, the dual iterates diverge at a controlled linear rate and vice-versa. Such behavior has not been previously observed or characterized in the literature.

Main contributions

For notational convenience, we use $z = (x, y)$ as the primal-dual pair, and $\bar{z}^k := \frac{1}{k}\sum_{j=1}^{k} z^j$ as the average of iterates. We propose to detect infeasibility using three sequences:

$$\textbf{(Difference of iterates)} \quad d_k = (z^{k+1} - z^k)\,, \tag{3.2a}$$

$$\textbf{(Normalized iterates)} \quad \frac{z^k}{k}\,, \tag{3.2b}$$

$$\textbf{(Normalized average iterates)} \quad \frac{2}{k+1}\bar{z}^k\,. \tag{3.2c}$$

Our proposal to detect infeasibility is as follows:

> Use these three sequences' primal and dual components as candidates for dual and primal infeasibility certificates. The algorithm should periodically check if any of these iterates satisfy the conditions that define an infeasibility certificate within numerical tolerances. If at any point this happens, it should conclude that the problem is (primal or dual) infeasible.

The overhead cost of extracting the certificates is negligible, making it suitable for large-scale problems. Most of the content of this work justifies this strategy theoretically.

Operator theory shows that all three of these sequences converge to a point v known as the *infimal displacement vector*. Section 3.2 will give a formal definition of this and other relevant concepts. We list our contributions assuming, for now, the existence of such a vector v.

Sublinear convergence rate to the infimal displacement vector (Section 3.3). It is natural to wonder how fast the three sequences (3.2) converge to the infimal displacement vector. We study this question through the lenses of general operators and fixed-point iteration. To the best of our knowledge, the only known result in this vein ensures a rate of $O\left(\frac{1}{\sqrt{k}}\right)$ for the difference of iterates (3.2a), which is known to be tight [163, 71]. In contrast, we show that two other sequences, the normalized iterates (3.2b) and the normalized average iterates (3.2c), converge at a faster rate of $O\left(\frac{1}{k}\right)$ in this same situation. Furthermore, this faster sublinear result generalizes to *any fixed-point iteration of a nonexpansive operator*, not only the firmly nonexpansive operators studied in [163]. Specifically, it also applies to many popular first-order methods, including but not limited to PDHG, ADMM [205], and Mirror-prox [185], and it extends to other settings beyond LP such as quadratic convex programming and semidefinite programming. Furthermore, we show that this result is tight for PDHG; i.e., there exist instances with a convergence rate lower bounded by $\Omega\left(\frac{1}{k}\right)$. This result suggests that current ADMM-based codes like OSQP [225] that use exclusively the difference of iterates to detect infeasibility should additionally consider the normalized iterates and normalized average iterates.

Characterization of the iterates for infeasible problems (Section 3.4). We characterize the behavior of PDHG for all the LP feasibility scenarios (see Table 3.1). In particular, we show that if the primal (or dual) iterates diverge, then the iterates diverge in the direction of a ray, where the direction of the ray recovers certificates of dual (or primal) infeasibility. Such direction turns out to be the infimal displacement vector (v_x, v_y). This justifies using the sequences (3.2) as infeasibility certificate candidates. Furthermore, we show that when the primal problem is feasible, then the dual iterates, without any normalization, converge

to some $y^\star$ that is closely related to v. An analogous result holds if the dual is feasible. This describes the dynamics of PDHG for unbounded problems. The next table summarizes our findings for the four possible cases.

Primal \ Dual	**Feasible**	**Infeasible**
Feasible	x^k, y^k both converge	x^k diverges, y^k converges
Infeasible	x^k converges, y^k diverges	x^k, y^k both diverge

Table 3.1: Behavior of PDHG for solving under different feasibility assumptions.

Eventual linear convergence for nondegenerate problems (Section 3.5). In the process of characterizing the dynamics of PDHG, we show that the iterates (x^k, y^k) always converge to a unique ray $\{(x^\star, y^\star) + \lambda v \mid \lambda \in \mathbf{R}_+\}$. We show that under a non-degeneracy condition (a direct extension of the strict complementary condition to the infeasible LPs), the iterates (x^k, y^k) fix their active set after finitely many iterations. In turn, this leads to the eventual linear convergence of the difference of iterates (3.2a). Formally, we show that there exists $K \geq 0$ such that for all sufficiently large $k \geq K$ we have

$$\|d^k - v\| \leq O\left(\gamma^{k-K}\right) \qquad \text{for some } \gamma \in (0, 1)\,.$$

We further show that even after the active set is fixed, the normalized iterates and normalized average do not exhibit faster convergence. Thus, it is strictly better to use the difference of iterates to detect infeasibility in this regime.

Computational experiments (Section 6.5). We verify our theoretical results by presenting numerical experiments displaying the efficacy of the different certificate candidates (3.2). Specifically, our experiments show that using all three sequences in (3.2) is beneficial. On the one hand, if the active set's finite time identification occurs relatively quickly, then the differences of iterates (3.2a) exhibit faster convergence. On the other hand, for some problems, identifying the

active set might not happen in a reasonable amount of time. In this case, both the normalized iterates (3.2b) recovers approximate infeasibility certificates much more efficiently than the differences.

Related work

Chambolle and Pock [44] review PDHG among other methods and describe its applications in computer vision. O'Connor and Vandenberghe [194] show that PDHG is in fact a particular application of Douglas-Rachford Splitting (DRS) [81, 108, 101, 162].

Lan et al. [138] and Renegar [211] develop FOMs for LP, considered as a special case of semidefinite programming, with $O\left(\frac{1}{k}\right)$ convergence rates. Gilpin et al. [107] obtain a restarted FOM for LP with a linear convergence rate. These analyses assume an optimal solution exists. Pock and Chambolle [203] apply PDHG with diagonal preconditioning to LP on a small number of test instances. They note that on small-scale problems, interior-point methods clearly dominate, while their method outperforms MATLAB's LP solver on one larger LP motivated by a computer vision application. Most recently, Basu et al. [17] apply accelerated gradient descent to a specialized LP instance, obtaining solutions to industrial problems with up to 10^{12} variables.

Classically, the primal simplex method for LP detects primal infeasibility while solving a "phase-one" auxiliary problem for an initial feasible basis and detects dual infeasibility based on conditions when computing a step size (i.e., the ratio test) [169]. Infeasibility certificates are extracted from the iterates of interior-point methods without substantial extra work [230]. Infeasibility detec-

tion is only the first step of diagnosing the cause of the infeasibility in an LP model [55].

Most research on infeasibility detection capabilities for FOMs for convex optimization has focused on ADMM or equivalently Douglas-Rachford Splitting. Eckstein and Bertsekas [92] show that when no solution exists, then the iterates diverge. Recent practical successes motivated further research in this direction, characterizing the asymptotic behavior of the iterates under additional assumptions. For example, the line of work [18, 19, 183] studies Douglas-Rachford applied to problems that look for a point at the intersection of two non-intersecting convex sets. On the other hand, Raghunathan and Di Cairano [206] investigate the asymptotic dynamics of ADMM for convex quadratic problems when the matrices involved are full rank.

Banjak et al. [16] show that the infimal displacement vector of ADMM recovers certificates of infeasibility for convex quadratic problems with conic constraints. Based on this, they proposed to use the difference of iterates to test infeasible. Complementing this work, [163] establishes a $O\left(\frac{1}{\sqrt{k}}\right)$ convergence rate for the difference of iterates of any algorithm that induces a firmly nonexpansive operator and introduced a scheme that utilizes multiple runs of ADMM to detect infeasibility. This type of scheme aims to handle pathological scenarios that do not occur in LP.

O'Donoghue et al. [196] propose to apply ADMM to a homogeneous self-dual embedding of a convex conic problem[1]. A nice byproduct of this approach is that it automatically produces infeasibility certificates. Subsequent work [195] extends this approach to Linear Complementarity Problems, which

[1]Linear objective with conic constraints.

cover quadratic convex losses with conic constraints.

To our knowledge, the only work analyzing the behavior of PDHG on potentially infeasible instances is by Malitsky [168], which considers linearly constrained problems. This analysis applies only to linear equality constraints, not to linear inequalites present in LP.

Finite time identifiability has a long history in the field of optimization. This phenomenon is first documented for the projected gradient descent method [91, 34, 32, 29]. Soon after it is studied for other methods, such as the Proximal Point Method [96] and Projected subgradient descent [97], among others [10]. Identifiability is also exploited as tool for algorithmic design for the so called "$\mathcal{UV}$-algorithms" [176]. Recent works [179, 158, 159] study finite time identification for popular FOMs. In particular, Liang et al. [159] show that the iterates of PDHG identify the active constraints in finite time, provided the limit point is nondegenerate. All of these works assume the underlying problem is feasible. The significant number of algorithms exhibiting this behavior motivated researchers to develop general theory (even beyond the realms of optimization) [243, 174, 175, 151, 86, 148]. We refer the interested reader to [148] for an elegant geometrical definition that generalizes most notions of nondegeneracy.

Outline of the chapter. Section 3.2 presents all the necessary background. In Section 3.3, we show a convergence rate of $O(1/k)$ for the normalized iterates and normalized average generated by the fixed-point iteration of a nonexpansive operator. Then, Section 3.4 shows a complete characterization of the behavior of PDHG under different infeasibility assumptions. In Section 3.5, we study a condition that ensures finite time identifiability of the active set. We show

that under this condition, the difference of iterates exhibits eventual linear convergence. We present numerical experiments that complement the theoretical results in Section 3.6. Section 3.7 contains all the omitted proofs of the first sections.

3.2 Preliminaries of the chapter

In this section, we introduce the notation we use throughout the chapter, summarize the LP formulations we solve, introduce PDHG algorithm for LP, and discuss some existing results.

Notation. We use the symbols $\mathbf{N}$ and $\mathbf{R}$ to denote the natural numbers and reals. Our results take place in a finite dimensional spaces $\mathbf{R}^d$. We denote the cone of nonnegative vectors as $\mathbf{R}^d_+$. Often, $\|\cdot\|$ will denote a norm with respect to which an operator is (firmly) nonexpansive; see a formal definition below. Since we study primal and dual problems, we use $z = (x, y) \in \mathbf{R}^{n+m}$ as a placeholder for primal and dual variables. We will sometimes refer to a vector $v \in \mathbf{R}^{n+m}$ and use v_x and v_y to denote its primal and dual components. In this chapter, we use superscripts to denote iteration counts, consequently z^k is the kth iterate.

Linear programming [62, 170]. LP problems can be parameterized using multiple equivalent forms. For our theoretical results we focus on the *standard form* of LP:

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & Ax = b \ \ (\in \mathbf{R}^m) \\ & x \geq 0 \quad (\in \mathbf{R}^n)\,, \end{array} \tag{P}$$

where $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, and $c \in \mathbf{R}^n$ are given. The dual of this problem is given by

$$\begin{aligned} &\text{maximize} && b^\top x \\ &\text{subject to} && A^\top y \leq c \ \ (\in \mathbf{R}^n) \ . \end{aligned} \tag{\textbf{D}}$$

Although the proofs in this chapter are tailored to this form, the techniques we use should extend easily to any other form.

Farkas' Lemma states that a feasible solution of (**P**) exists if, and only if, the following set is empty

$$\{y \in \mathbf{R}^m \mid b^\top y < 0 \text{ and } A^\top y \geq 0\} \ . \tag{3.3}$$

We call the elements of this set *certificates of primal infeasibility*, as their existence guarantees that the primal problem is infeasible. Analogously, the certificates of infeasibility for the dual problem (**D**) are

$$\{x \in \mathbf{R}^n \mid c^\top x < 0, Ax = 0 \text{ and } x \geq 0\} \ . \tag{3.4}$$

Primal-dual hybrid gradient. Chambolle and Pock [43] establish convergence to a saddle point at a rate of $O(1/k)$ provided that a *saddle exists* and $\eta\tau\|A\|_2^2 < 1$.[2] The primal-dual problems (**P**)-(**D**) can be recast as a *convex-concave saddle point problem*. In particular we choose $g(x) = c^\top x + \iota_{\{x \geq 0\}}(x)$ and $h(y) = b^\top y$. In this case the proximal updates can be computed in closed form. In fact, a PDHG update reduces to

$$\begin{aligned} x^+ &= \text{proj}_{\mathbf{R}^n_+}(x - \eta A^\top y - \eta c) \\ y^+ &= y + \tau A(2x^+ - x) - \tau b \ . \end{aligned} \tag{3.5}$$

[2] More precisely, Chambolle and Pock proved this rate for the primal-dual gap of the averaged iterates.

Observe that the most complex operations in the update formula are matrix-vector products, and all other operations are separable by component.

An update of PDHG (Algorithm 1) can be equivalently defined with a differential inclusion of the form

$$M \begin{bmatrix} x^k - x^{k+1} \\ y^k - y^{k+1} \end{bmatrix} \in \begin{bmatrix} \partial g(x^{k+1}) \\ \partial h(y^{k+1}) \end{bmatrix} + \begin{bmatrix} A^\top y^{k+1} \\ -Ax^{k+1} \end{bmatrix} \qquad \text{with} \qquad M := \begin{bmatrix} \frac{1}{\eta} I_n & -A^T \\ -A & \frac{1}{\tau} I_m \end{bmatrix}, \tag{3.6}$$

this follows from (2.4). We will later leverage this inclusion in our proofs.

Operators and the fixed-point iteration. We will find it useful to think of iterative algorithms from an operator viewpoint. Given an arbitrary map $T : \mathbf{R}^d \to \mathbf{R}^d$, the corresponding fixed-point iteration is defined as

$$z^{k+1} = T(z^k) \,. \tag{3.7}$$

Most first-order methods can be described in this form. The primal-dual hybrid gradient method can be encoded as $T(x, y) = (x^+, y^+)$ where the output pair is defined in (3.5). When looking at an algorithm from this perspective, we transform the problem of finding a solution of the optimization problem to that of finding a fixed-point of the operator, i.e., $z^\star = T(z^\star)$. This idea has proven fruitful for proving optimal converge rates for a variety of algorithms [71].

Here we make a minimal assumption that is sufficient to analyze PDHG in the infeasible case. An operator T is said to be *nonexpansive* if it is 1-Lipschitz continuous with respect to a matrix norm $\|\cdot\|$, meaning that for any $z_1, z_2 \in \mathbf{R}^d$ we have

$$\|T(z_1) - T(z_2)\| \le \|z_1 - z_2\| \,. \tag{3.8}$$

Nonexpansiveness does not ensure the convergence of iterates in the feasible case. Yet a slightly stronger condition does. An operator T is *firmly nonexpansive*

if it satisfies

$$\|T(z_1) - T(z_2)\|^2 \leq \|z_1 - z_2\|^2 - \|(T - I)(z_1) - (T - I)(z_2)\|^2 \qquad \text{for all } z_1, z_2 \in \mathbf{R}^d \ .$$

Note that the norm here is not necessarily the Euclidean norm. All the results concerning (firmly) nonexpansiveness in this section and the next one are with respect to the norm in which these properties hold. The following is a beautiful geometrical result proved by Pazy that defines a pivotal object in our studies.

Lemma 3.2.1 (Lemma 4 in [201])**.** *Let T be a nonexpansive operator, then the set* $\mathrm{cl}(\mathrm{range}(T - I))$ *is convex. Consequently, there exists a unique minimum norm vector in this set:*

$$v_T := \operatorname*{arg\,min}_{z \in \mathrm{cl}(\mathrm{range}(T-I))} \frac{1}{2}\|z\|^2 \ . \tag{3.9}$$

This vector is known as the *infimal displacement vector*. We drop the subscript T and make the corresponding operator clear from the context. Intuitively, v is the minimum size perturbation we should subtract from T to ensure it has a fixed point.

Theorem 3.2.2 ([201] and [15])**.** *Let T be a nonexpansive operator and (z^k) be a sequence generated by the fixed-point iteration* (3.7)*. Then, we have*

$$\lim_{k\to\infty} \frac{z^k}{k} = v \ . \tag{3.10}$$

If further T is firmly nonexpansive, then

$$\lim_{k\to\infty} z^{k+1} - z^k = v \ . \tag{3.11}$$

That is the normalized iterate converges to the infimal displacement vector when T is nonexpansive and if T is firmly nonexpansive the difference of iterates also converge. One might wonder whether the the stronger condition is necessary. This turns out to be the case.

The following proposition shows two things: first, (3.11) is provably stronger than (3.10) and second, the convergence of the iterates ensures converges of the normalized average. Recall from the previous section that we use $\bar{z}^k := \frac{1}{k}\sum_{j=1}^{k} z^j$ to denote the average.

Proposition 3.2.3. *Let $(z^k)_{k=1}^{\infty} \subseteq \mathbf{R}^d$ be an arbitrary sequence and let $v \in \mathbf{R}^d$ be a fixed vector. Then the following implications hold:*

1. ***Difference convergence implies normalized iterate convergence.***

$$\lim_{k\to\infty}(z^{k+1} - z^k) = v \implies \lim_{k\to\infty} \frac{z^k}{k} = v\,.$$

2. ***Normalized iterate convergence implies normalized average convergence.***

$$\lim_{k\to\infty} \frac{z_k}{k} \to v \implies \lim_{k\to\infty} \frac{2\bar{z}^k}{(k+1)} \to v\,.$$

Moreover, these implications cannot be reversed as there exist simple counterexamples in $\mathbf{R}$.

The proof of this proposition is technical, so it is deferred to Section 3.7.1.

Naturally when concerned with practical algorithms one would like to have convergence rates for (3.10) and (3.11). As far as we know, the state-of-the-art result in this vein is due to Liu, Ryu, and Yin [163].

Theorem 3.2.4 ([163])**.** *Let T be a firmly nonexpansive operator and (z^k) be a sequence generated by* (3.7)*. Then, for any $\varepsilon > 0$, then there exists a point z_ε such that*

(Average iterate rate).

$$\left\| v - \frac{2}{k+1}(\bar{z}^k - z^0) \right\| \leq \sqrt{\frac{2}{k+1}}\|z^0 - z_\varepsilon\| + \varepsilon\,,$$

(Last iterate rate).

$$\left\| v - \frac{1}{k}(z^k - z^0) \right\| \leq \sqrt{\frac{1}{k}} \|z^0 - z_\varepsilon\| + \varepsilon \,,$$

(Difference rate).

$$\min_{j \leq k} \|v - z^{j+1} - z^j\| \leq \sqrt{\frac{1}{k}} \|z^0 - z_\varepsilon\| + \varepsilon \,.$$

Remark 1. *In the paper [163], this result is presented only for the difference of iterates, yet a simple modification of their argument proves the other two results.*

The theorem guarantees a rate of convergence that depends on a target accuracy ε. The rate could get worse as $\varepsilon \to 0$. Indeed, z_ε could diverge as ε goes to zero, see Example 3.7.3 in Section 3.7.3. We will see in the next section that for LP it is possible to get rates that are independent of the accuracy ε.

Since the algorithm of interest is PDHG, we might wonder whether or not its operator is firmly nonexpansive. It turns out that it is, but with respect to the norm induced by the matrix M.

Proposition 3.2.5. *If $\eta\tau\|A\|_2^2 < 1$, then the operator defined by a PDHG iteration is firmly nonexpansive with respect to the norm the $\|\cdot\|_M$ with M defined as in* (3.6).

Proof. This is a known result [116], yet we include a proof for the interested reader. A Schur complement argument proves that the condition $\eta\tau\|A\|_2^2 < 1$ ensures that $M > 0$ is positive definite. Then a direct application of Proposition 4.2 in [20] proves the result. □

3.3 Sublinear convergence of nonexpansive operators

This section presents the $O(1/k)$ convergence rate of the normalized iterates and the normalized average for nonexpansive operators. This rate applies to a broader class of operators than the previously known results (restated in Theorem 3.2.4) as it does not require the operator to be firmly nonexpansive. The resulting rate applies to many popular FOMs for convex optimization, including but not limited to PDHG [43], the Alternating Method of Multipliers (ADMM) or equivalently Douglas-Rachford Splitting (DRS) [205], and Mirror-Prox [185].

Theorem 3.3.1 presents our main result in this section.

Theorem 3.3.1. *Let T be a nonexpansive operator for some norm $\|\cdot\|$ and define v to be the minimum norm element in* $\mathrm{cl}(\mathrm{range}(T-I))$. *Then, for any $\varepsilon > 0$, there exists z_ε such that the following two inequalities hold*

(Average iterate rate).

$$\left\| v - \frac{2}{(k+1)}\left(\bar{z}^k - z^0\right)\right\| \leq \frac{4}{k+1}\|z^0 - z_\varepsilon\| + \varepsilon\,. \tag{3.12}$$

(Last iterate rate).

$$\left\| v - \frac{1}{k}(z^k - z^0)\right\| \leq \frac{2}{k}\|z^0 - z_\varepsilon\| + \varepsilon\,. \tag{3.13}$$

Furthermore, if $\mathrm{range}(T-I)$ *is closed, then there exists a finite $z^\star$ such that $T(z^\star) = z^\star + v$ and for any such $z^\star$ and all k:*

(Average iterate rate).

$$\left\| v - \frac{2}{(k+1)}\left(\bar{z}^k - z^0\right)\right\| \leq \frac{4}{k+1}\|z^0 - z^\star\|\,. \tag{3.14}$$

(Last iterate rate).

$$\left\| v - \frac{1}{k}(z^k - z^0) \right\| \leq \frac{2}{k}\|z^0 - z^\star\| \,. \tag{3.15}$$

Remark 2. *We comment that when* range$(T - I)$ *is not closed, Theorem 3.3.1 may not imply a* $O(1/k)$ *sublinear convergence rate. In fact, as* $\varepsilon \to 0$*, the vector* $\|z_\varepsilon\|$ *could grow to infinity, see Example 3.7.3 in Section 3.7.3 for a one dimensional example. When* range$(T - I)$ *is closed, we obtain a* $O(1/k)$ *sublinear rate.*

Remark 3. *When* range$(T - I)$ *is closed, the above result together with the lower bound proved later in Theorem 3.5.3 shows that the normalized iterates and normalized average of a nonexpansive operator exhibit a* $\Theta\left(\frac{1}{k}\right)$ *convergence rate. It is faster than the difference of iterates, by noticing that the difference of iterates converges at rate* $\Theta\left(\frac{1}{\sqrt{k}}\right)$ *(see Theorem 8 of [71]).*

The next Lemma is used in the proof of Theorem 3.3.1.

Lemma 3.3.2. *Suppose the assumptions of Theorem 3.3.1. Fix* $\varepsilon > 0$*, then there exists a point* z_ε*, such that the following two inequalities hold for all* $k \geq 0$*:*

1. $\|T^{k+1}(z_\varepsilon) - T^k(z_\varepsilon) - v\| \leq \varepsilon$.
2. $\|(T^k(z_\varepsilon) - z_\varepsilon) - kv\| \leq k\varepsilon$.

Furthermore, if range$(T - I)$ *is closed, then there exists a point* $z^\star$ *such that* $T(z^\star) = z^\star + v$*. For all such* $z^\star$*, for all* $k \geq 0$*:*

3. $T^{k+1}(z^\star) - T^k(z^\star) = v$.
4. $T^k(z^\star) - z^\star = kv$.

Proof. Without loss of generality we assume that $\varepsilon \leq 1$. Notice $v \in \mathrm{cl}(\mathrm{range}(T-I))$, thus there exists z_ε such that

$$\|T(z_\varepsilon) - z_\varepsilon - v\| \leq \frac{\varepsilon^2}{\max\{1, 2(\|v\| + 1)\}} \,. \tag{3.16}$$

We start by proving the first claim. Fix an arbitrary $k \geq 0$. We will make use of two facts in the proof. Since T is a nonnexpansive operator, an application of the triangle inequality yields

$$\|T^k(z_\varepsilon) - T^{k-1}(z_\varepsilon)\| - \|v\| \leq \|T(z_\varepsilon) - z_\varepsilon\| - \|v\| \leq \|T(z_\varepsilon) - z_\varepsilon - v\| \leq \frac{\varepsilon^2}{2(\|v\| + 1)} \,. \tag{3.17}$$

Noticing v is the nearest point to zero in $W = \mathrm{cl}(\mathrm{range}(T - I))$ with respect to the norm $\|\|$ and the set W is convex, it follows from the optimality conditions of this problem that

$$\langle w, v\rangle \geq \|v\|^2 \qquad \text{for all } w \in W \,. \tag{3.18}$$

Armed with these two facts, we derive for any arbitrary k:

$$\begin{aligned}
\|T^k(z_\varepsilon) - T^{k-1}(z_\varepsilon) - v\|^2 &= \|T^k(z_\varepsilon) - T^{k-1}(z_\varepsilon)\|^2 - 2\langle T^k(z_\varepsilon) - T^{k-1}(z_\varepsilon), v\rangle + \|v\|^2 \\
&\leq \|T^k(z_\varepsilon) - T^{k-1}(z_\varepsilon)\|^2 - 2\|v\|^2 + \|v\|^2 \\
&= \Big(\|T^k(z_\varepsilon) - T^{k-1}(z_\varepsilon)\| + \|v\|\Big)\Big(\|T^k(z_\varepsilon) - T^{k-1}(z_\varepsilon)\| - \|v\|\Big) \\
&\leq (\|T(z_\varepsilon) - z_\varepsilon - v\| + 2\|v\|)\,\frac{\varepsilon^2}{2(\|v\| + 1)} \\
&\leq \Big(\varepsilon^2 + 2\|v\|\Big)\,\frac{\varepsilon^2}{2(\|v\| + 1)} \leq \varepsilon^2 \,,
\end{aligned}$$

where the first inequality utilizes (3.18) by noticing $T^k(z_\varepsilon) - T^{k-1}(z_\varepsilon) \in W$, the second inequality uses (3.17) and the triangle inequality, the third inequality is from (3.16), and the last inequality uses $\varepsilon \leq 1$. This proves the first statement.

The second claim follows by induction. The base case $k = 0$ holds directly. For the inductive step, assume that the statement holds for $k - 1$. Then

$$\|(T^k(z_\varepsilon) - z_\varepsilon) - kv\| \leq \|T^k(z_\varepsilon) - T^{k-1}(z_\varepsilon) - v\| + \|T^{k-1}(z_\varepsilon) - z_\varepsilon - (k-1)v\| \leq k\varepsilon \,,$$

where we used the first claim and the inductive hypothesis.

Furthermore, if range$(T - I)$ is closed, the statements follow by taking $\varepsilon = 0$ in the previous proofs. □

Proof of Theorem 3.3.1. Let z_ε be the point given by Lemma 3.3.2. We proceed to prove the first two statements.

1. It follows from $z^j = T^j(z^0)$ that

$$\begin{aligned}\left\|\frac{2}{k(k+1)}\sum_{j=1}^{k}(z^j - z^0) - v\right\| &= \left\|\frac{2}{k(k+1)}\sum_{j=1}^{k}\left(T^j(z^0) - z^0 - jv\right)\right\| \\ &= \left\|\frac{2}{k(k+1)}\sum_{j=1}^{k}\left((T^j(z^0) - T^j(z_\varepsilon)) + (z_\varepsilon - z^0) + (T^j(z_\varepsilon) - z_\varepsilon - jv)\right)\right\| \\ &\leq \left\|\frac{2}{k(k+1)}\sum_{j=1}^{k}\left(T^j(z^0) - T^j(z_\varepsilon)\right)\right\| + \frac{2}{(k+1)}\left\|z^0 - z_\varepsilon\right\| + \varepsilon\,,\end{aligned}$$

where the inequality uses Lemma 3.3.2 and the triangle inequality. Applying the triangle inequality to the first term yields

$$\begin{aligned}\left\|\frac{2}{k(k+1)}\sum_{j=1}^{k}\left(T^j(z^0) - T^j(z_\varepsilon)\right)\right\| &\leq \frac{2}{k(k+1)}\sum_{j=1}^{k}\left\|T^j(z^0) - T^j(z_\varepsilon)\right\| \\ &\leq \frac{2}{k(k+1)}\sum_{j=1}^{k}\left\|z^0 - z_\varepsilon\right\| = \frac{2}{(k+1)}\left\|z^0 - z_\varepsilon\right\|\,,\end{aligned}$$

where the second inequality follows since T is nonexpansive.

2. Notice that

$$\begin{aligned}\left\|\frac{1}{k}(z^k - z^0) - v\right\| &= \left\|\frac{1}{k}\left((T^k(z^0) - z^0) - (T^k(z_\varepsilon) - z_\varepsilon) + (T^k(z_\varepsilon) - z_\varepsilon - kv)\right)\right\| \\ &\leq \left\|\frac{1}{k}\left(\left(T^k(z^0) - z^0\right) - (T^k(z_\varepsilon) - z_\varepsilon)\right)\right\| + \varepsilon \\ &\leq \frac{1}{k}\left(\|T^k(z^0) - T^k(z_\varepsilon)\| + \|z^0 - z_\varepsilon\|\right) + \varepsilon \\ &\leq \frac{2}{k}\|z^0 - z_\varepsilon\| + \varepsilon\,,\end{aligned}$$

where the first inequality uses triangle inequality and Lemma 3.3.2, and the last inequality is from non-expansiveness of T.

Finally, if range$(T - I)$ is closed, we obtain the results following the same argument as above with $\varepsilon = 0$, using the results for closed range$(T - I)$ from Lemma 3.3.2. □

A drawback of (3.12) and (3.13) in Theorem 3.3.1, as well as the results in Theorem 3.2.4, is that the constants accompanying the rates depend on ε. Nonetheless, we can bypass this issue, using (3.14) and (3.15), for problems where range$(T - I)$ is closed. The next proposition guarantees that range$(T - I)$ is indeed closed for a broad family of algorithms for solving LP.

Proposition 3.3.3. *Let $T : \mathbf{R}^d \to \mathbf{R}^d$ be an operator that can be decomposed as $T = T_k \circ \cdots \circ T_1$ where T_j is either an affine mapping or a projection onto a polyhedron. Then,* range(T) *is a finite union of polyhedra.*

Proof. The proof follows inductively. Assume that $C = \text{range}(T_j \circ \cdots \circ T_1)$ is a finite union of polyhedra. Without loss of generality, we can assume that C is equal to a single polyhedron. Now we consider two cases:

Case 1. Assume that T_{j+1} is an affine transformation. This is a well-known consequence of Fourier-Motzkin elimination [170].

Case 2. Assume that T_{j+1} is a projection onto a polyhedron Q. First, we start with an intuitive sketch of the proof and then formalize it. In this case, different pieces of the polyhedron C are going to be projected to different faces of the polyhedron Q. Each one of these pieces is a polyhedron and

since there are only finitely many faces of Q, the projection is a finite union of polyhedra.

More formally, any polytope Q defines a finite polyhedral partition of the space $\{P_F\}_{F\in\Delta}$ where Δ is the collection of faces of the polyhedron Q.[3] Each cell P_F corresponds to the region of the space that projects onto F, that is $\text{proj}_Q(P_F) = F$. Define a partition of the polyhedron C as $\{C_F\}_{F\in\Delta}$ given by $C_F = P_F \cap C$. Within each cell P_F the projection $\text{proj}_Q\ |_{P_F}$ is an affine transform. Thus, by Case 1 we have that $\text{proj}_Q(C_F)$ is a polyhedron and thus

$$T_{j+1}(C) = \text{proj}_Q(C) = \bigcup_{F\in F} \text{proj}_Q(C_F)$$

is a finite union of polyhedra.

□

As a result of Proposition 3.3.3 applied to the PDHG update for solving an LP problem, range$(T - I)$ is a polyhedron, thus closed:

Corollary 3.3.4. *Let T be the PDHG operator for an LP problem, then range$(T - I)$ is a finite union of closed polyhedra.*

Proof. Notice that the operator $T - I$ is composite of linear operators and projection operators, thus we obtain the results by using Proposition 3.3.3. □

Remark 4. *Proposition 3.3.3 shows that range$(T - I)$ is closed when T is an operator that corresponds to other first-order algorithms, such as ADMM and mirror-prox, to solve an LP problem. Further, one can extend the result to cover convex quadratic problems with polyhedral constraints.*

[3] I.e., each cell $P_F \subseteq \mathbf{R}^d$ is a polyhedron and $\cup_{F\in\Delta} P_F = \mathbf{R}^d$.

3.4 The complete behavior of PDHG for solving LP problems

In Figure 3.1, we saw low-dimensional examples of the dynamics of PDHG when solving LP problems in different feasibility settings. Indeed, such convergence/divergence dynamics generally hold when using PDHG to solve arbitrary LP problems. In this section, we present a complete description of the behavior of PDHG for feasible and infeasible LP problems and discuss how to recover the infeasibility certificate from the iterates of PDHG. The next theorem compiles the full characterization. The proof of this theorem will be deferred to Section 3.4.3.

Theorem 3.4.1. *Consider the primal* (**P**) *and dual* (**D**) *problems. Assume that* $\eta\tau\|A\|_2^2 < 1$, *let* T *be the operator induced by* (3.5), *and let* $\{z^k\}_k$ *be a sequence generated by the fixed-point iteration for an arbitrary starting point* z^0, *i.e.,* $z^k = T^k(z^0)$ *Then, one of the following holds:*

1. *If* ***both primal and dual are feasible****, then the iterates* (x^k, y^k) *converge to a primal-dual solution* $z^\star = (x^\star, y^\star)$ *and* $v = (T - I)(z^\star) = 0$.
2. *If* ***both primal and dual are infeasible****, then both primal and dual iterates diverge to infinity towards the direction of the infimal displacement vector* $v = (v_x, v_y)$. *Moreover, the primal and dual components of the infimal displacement vector* v_x *and* v_y *give certificates of dual and primal infeasibility, respectively.*
3. *If* ***the primal is infeasible and the dual is feasible****, then the dual iterates diverge to infinity in the direction of* v_y, *while the primal iterates converge to a vector* $x^\star$. *Furthermore, the dual-component* v_y *is a certificate of primal infeasibility, and there exists a vector* $y^\star$ *such that* $v = (T - I)(x^\star, y^\star)$.

4. *If **the primal is feasible and the dual is infeasible**, then the same conclusions as in the previous item hold by swapping primal with dual.*

To show this characterization, we establish two intermediate results: first, the infimal displacement vector v is nonzero if, and only if, either the primal or dual problems are infeasible; and second, the iterates (x^k, y^k) "converge" to a well-defined ray of the form $(x^\star, y^\star) + \lambda v$ for $\lambda \in \mathbf{R}_+$. The first result describes the asymptotic divergent behavior of the primal (resp. dual) iterates when the dual (resp. primal) problem is infeasible. The second one, ensures the asymptotic convergence of the primal (resp. dual) iterates without any normalization when the dual (resp. primal) problem is feasible. These two intermediate results are proved in Section 3.4.1 and Section 3.4.2, respectively.

3.4.1 The infimal displacement vector recovers certificates

In Section 3.3, we demonstrated that the differences of iterates, the normalized iterates, and the normalized average for a nonexpansive operator converge to the infimal displacement vector v. Here, we show that the infimal displacement vector v for PDHG applied to LP recovers infeasibility certificates whenever it is nonzero. First, some simple properties of v.

Lemma 3.4.2. *Consider the primal* (**P**) *and dual* (**D**) *problems. Assume that* $\eta\tau\|A\|_2^2 < 1$, *let* T *be the operator induced by* (3.5), *and let* $v = (v_x, v_y)$ *be the infimal displacement vector of* T. *Then* $v_x \geq 0$, $Av_x = 0$, *and* $A^\top v_y \geq 0$.

Proof. From Theorem 3.2.2,

$$\frac{1}{k}(x^k, y^k) \to v = (v_x, v_y) \quad \text{and} \quad \frac{1}{k}(z_{k+1} - z_k) \to 0\,. \tag{3.19}$$

Notice that PDHG for LP has the following iteration update in terms of a differential inclusion,

$$M \begin{bmatrix} x^k - x^{k+1} \\ y^k - y^{k+1} \end{bmatrix} \in \begin{bmatrix} N_{\mathbf{R}^n_+}(x^{k+1}) + A^\top y^{k+1} + c \\ -Ax^{k+1} + b \end{bmatrix}, \tag{3.20}$$

where this relation comes from (3.6) and (2.2). Dividing (3.20) by k and letting $k \to \infty$, we have from (3.19) that

$$0 \in \lim_{k\in\infty} N_{\mathbf{R}^n_+}(x^k) + A^\top \frac{1}{k} y_k \subseteq -\mathbf{R}^n_+ + A^\top v_y \implies A^\top v_y \geq 0\,, \tag{3.21}$$

where we utilize the fact that $N_{\mathbf{R}^n_+}(x) \subseteq -\mathbf{R}^n_+$ for any $x \in \mathbf{R}^n_+$ and $\lim_{k\to\infty} \frac{1}{k} y_k = v_y$; and

$$0 = \lim_{k\in\infty} -\frac{1}{k} A x^k = -A v_x \implies A v_x = 0\,. \tag{3.22}$$

Furthermore, note that $v_x \geq 0$ since $v_x = \lim_{k\to\infty} x^k/k$ and $x^k \geq 0$ for all k. □

Now we derive the main result of this section.

Proposition 3.4.3. *Consider the primal* (**P**) *and dual* (**D**) *problems. Assume that* $\eta\tau\|A\|_2^2 < 1$, *let* T *be the operator induced by* (3.5), *and let* $(z^k)_{k\in\mathbf{N}}$ *be a sequence generated by the fixed-point iteration for an arbitrary starting point* z^0. *Then, the primal problem* (**P**) *is infeasible if and only if* v_y *is a nonzero vector, and in this case,* v_y *is an infeasibility certificate for the primal problem. Analogously, the dual problem* (**D**) *is infeasible if and only if* v_x *is a nonzero vector, and in this case,* v_x *is an infeasibility certificate for the dual problem.*

Proof. To establish the first implication in this result we have to prove that if v_y is non-zero, then v_y is an infeasibility certificate for the primal problem, namely,

$$A^\top v_y \geq 0 \text{ and } b^\top v_y < 0\,,$$

thus the primal problem is infeasible. Similarly if v_x is non-zero, then v_x is an infeasibility certificate for the dual problem, namely

$$Av_x = 0, v_x \geq 0, \text{ and } c^\top v_x < 0 \,,$$

thus the dual problem is infeasible. We proved all the nonstrict inequalities in Lemma 3.4.2, so it suffices to show the strict ones.

First, consider the case when $v_x \neq 0$. Let $B = \{i \in [n] \mid (v_x)_i > 0\}$ and let $N = \{i \in [n] \mid (v_x)_i = 0\}$, then $B \cup N = [n]$ by noticing $v_x \geq 0$ (from Lemma 3.4.2). Given a vector x, let x_B be the vector of entries of x with indices in B; similarly given a matrix A, let A_B be the submatrix with columns of A with indices in B. Then for any $i \in B$, we have $(v_x)_i > 0$, thus there exists some K such that $(x^k/k)_i > 0$ for all $k \geq K$, and furthermore

$$(x^{k+1})_B = (x^k)_B - \eta A_B^\top y^k - \eta c_B \,.$$

Taking the limit $k \to \infty$ and noticing $\lim_{k\to\infty}(x^{k+1})_B - (x^k)_B = (v_x)_B$, we obtain

$$(v_x)_B = \lim_{k\to\infty} -\eta(A_B^\top y^k + c_B) \,.$$

Thus it holds that

$$c^\top v_x = c_B^\top (v_x)_B = -\frac{1}{\eta}\|v_x\|_2^2 - \lim_{k\to\infty}(y^k)^\top A_B (v_x)_B = -\frac{1}{\eta}\|v_x\|_2^2 < 0 \,, \tag{3.23}$$

where the last equality uses $A_B(v_x)_B = Av_x = 0$. Combining with $v_x \geq 0$ and $Av_x = 0$ proves that v_x is a certificate of infeasibility whenever it is nonzero.

Second, consider the case when $v_y \neq 0$. By taking $k \to \infty$, we have

$$v_y = \lim_{k\to\infty} y_{k+1} - y_k = \lim_{k\to\infty} \tau A(2x^{k+1} - x^k) - \tau b = \lim_{k\to\infty} \tau A x^{k+2} - \tau b \,, \tag{3.24}$$

where the second equality uses the update rule (3.5), and the third equality uses $\lim_{k\to\infty} 2x^{k+1} - x^k = \lim_{k\to\infty} x^{k+1} + v_x = \lim_{k\to\infty} x^{k+2}$.

Now we claim the following two facts:

Fact 3.4.4. *There exists some K such that if $(A^\top v_y)_i > 0$ then $x_i^k = 0$ for all $k \geq K$.*

Fact 3.4.5. *The support (nonzero components) of $A^\top v_y$ satisfies* $\operatorname{supp}(A^\top v_y) \subseteq N$.

The first fact is because if $(A^\top v_y)_i > 0$ then we have that $(A^\top y^k/k)_i \geq (A^\top v_y)_i/2 > 0$ for large enough k. Dividing (3.20) by k yields

$$-\frac{1}{k}(A^\top y^k)_i + \frac{1}{\eta k}\left(x^k - x^{k+1} - \eta c\right)_i \in N_{\mathbf{R}_+^n}(x_i^{k+1}) \,.$$

For large enough k, the second term on the left-hand side of the inclusion will be as small as $(A^\top y^k/k)_i/2$ and hence the sign of entire expression on the left-hand side will be negative. If $N_{\mathbf{R}_+}((x^{k+1})_i)$ contains a negative number, then $(x^{k+1})_i = 0$, which implies that $(x^{k+1})_i = 0$ for large enough k.

The second fact is because for any entry i in the support of $A^\top v_y$, namely $(A^\top v_y)_i > 0$, it follows from the first part that $(x^k)_i = 0$ for all k large enough, thus $(v_x)_i = \lim_{k\to\infty} \frac{1}{k} x_i^k = 0$, which proves the second fact by the definition of the set N.

Returning to the proof of Proposition 3.4.3, notice that

$$\lim_{k\to\infty} v_y^\top A x^k = \lim_{k\to\infty} \sum_{i\in N} (A^\top v_y)_i x_i^k = 0 \,, \tag{3.25}$$

where the first equality uses Fact 3.4.5, and the second equality uses Fact 3.4.4. Therefore, it holds that

$$v_y^\top b = -\frac{1}{\tau}\|v_y\|_2^2 + \lim_{k\to\infty} v_y^\top A x^{k+2} = -\frac{1}{\tau}\|v_y\|^2 < 0 \,,$$

where the first inequality uses (3.24) and the second equality is from (3.25). Together with (3.21), we know v_y is an infeasibility certificate for the primal problem.

Now we turn to the inverse direction. Recall that it follows from the closedness of the set range$(T - I)$ that there exists a pair $z^\star = (x^\star, y^\star)$ such that $T(z^\star) = z^\star + v$. If the dual problem is infeasible, we will show that $v_x \neq 0$ by contradiction. Assume $v_x = 0$; then it follows from the update rule (3.5) that

$$x^\star = \text{proj}_{\mathbf{R}^n_+}(x^\star - \eta(A^\top y^\star + \eta c)) \,,$$

thus $A^\top y^\star + \eta c \geq 0$ by noticing $x^\star \geq 0$, which contradicts the assumption that the dual problem is infeasible. If the primal problem is infeasible, we will show that $v_y \neq 0$ by contradiction. Suppose $v_y = 0$, then it follows from the update rule (3.5) that

$$y^\star = y^\star + \tau A(2(x^\star + v_x) - x^\star) - \tau b \,,$$

thus $Ax^\star = b$ by noticing $Av_x = 0$ from (3.22). Furthermore, we know $x^\star \geq 0$, thus the primal is a feasible problem, which contradicts with assumption.

This concludes the proof.

□

3.4.2 The iterates converge to a ray

Combining facts from the previous sections we know that if both primal and dual problems are feasible then the iterates (without normalization) will converge to a solution, and when both primal and dual problems are infeasible then the normalized iterates converge to a vector (v_x, v_y) with nonzeros on both primal and dual components. Yet the techniques used to prove these results do not explain what happens when one of the problems is feasible and the other one is infeasible. In this scenario the convergence of the primal and dual iterates hap-

pen at different scales, one with normalization by $\frac{1}{k}$ and the other without it. In this section, we fill in this gap by showing that the iterates of PDHG always converge to ray with direction v, emanating from a point $z^\star$. In turn, this allows us to connect the convergence results for the two scales.

Definition 3.4.6 (Ray). *Given a starting point $z^\star \in \mathbf{R}^{n+m}$ and a direction $v \in \mathbf{R}^{n+m}$, we define their* ray *as*

$$[z^\star, v] = \{z^\star + \lambda v \mid \lambda \in \mathbf{R}_+\} .$$

With this definition at hand we can now state the main result of this section.

Theorem 3.4.7. *Consider the primal* (**P**) *and dual* (**D**) *problems. Assume that $\eta\tau\|A\|_2^2 < 1$, let T be the operator induced by* (3.5)*, and let $(z^k)_{k\in\mathbf{N}}$ be a sequence generated by the fixed-point iteration for an arbitrary starting point z^0. Then, the iterates of PDHG converge to a ray $[z^\star, v]$, in particular*

$$\|z^k - z^\star - kv\| \to 0 \quad \textit{for some} \quad z^\star \in (T - I)^{-1}(v) .$$

To prove this result, we establish a connection between the iterates of PDHG applied to the original (possibly infeasible) problem and the iterates of PDHG applied to a feasible auxiliary LP problem. Let us start by defining this auxiliary problem. Define the index sets

$$\begin{aligned} B &= \{i \in [n] \mid (v_x)_i > 0\} , \\ N_1 &= \{i \in [n] \mid (v_x)_i = 0, (A^\top v_y)_i = 0\} , \\ N_2 &= \{i \in [n] \mid (v_x)_i = 0, (A^\top v_y)_i > 0\} . \end{aligned} \tag{3.26}$$

Define the operator $\tilde{T} : \mathbf{R}^{n+m} \to \mathbf{R}^{n+m}$ given by $\tilde{T}(z) := z^+$ with

$$\begin{aligned}
(x^+)_B &= x_B - \eta A_B^\top y - \eta c_B - (v_x)_B \\
(x^+)_{N_1} &= \mathrm{proj}_{\mathbf{R}_+^{|N_1|}}(x_{N_1} - \eta A_{N_1}^\top y - \eta c_{N_1}) - (v_x)_{N_1} \\
(x^+)_{N_2} &= -(v_x)_{N_2} \\
y^+ &= y + \tau A(2x^+ - x) - \tau b - v_y \, .
\end{aligned} \tag{3.27}$$

In turn, this is a PDHG operator for the auxiliary LP problem:

$$\begin{aligned}
\text{minimize} \quad & (c_B + (v_x)_B/\eta)^\top x_B + c_{N_1}^\top x_{N_1} + c_{N_2}^\top x_{N_2} \\
\text{subject to} \quad & A_B x_B + A_{N_1} x_{N_1} + A_{N_2} x_{N_2} = b + \frac{v_y}{\tau} \\
& x_{N_1} \geq 0, \; x_{N_2} = 0 \, .
\end{aligned} \tag{3.28}$$

Then we claim the following connection between T and $\tilde{T}$.

Proposition 3.4.8. *Given an arbitrary initial solution z^0, there exists a large enough $K \in \mathbf{N}$ such that*

$$\tilde{T}^k(z^K) = T^k(z^K) - kv \qquad \textit{for all } k \geq 0 \, . \tag{3.29}$$

Proof. For any initial solution, we know that there exists some K such that it holds for any $k \geq K$ that

$$(x^k)_B > 0 \text{ and } (x^k)_{N_2} = 0. \tag{3.30}$$

The former is because $(v_x)_B > 0$, and the latter follows from Fact 3.4.4. With some abuse of notation, we let $z^0 \leftarrow z^K$, so that we may study the iterates starting at z^0 (rather than starting at z^K), for notational convenience. From Lemma 3.4.2:

$$v_x \geq 0, \quad Av_x = 0, \quad \text{and} \quad A^\top v_y \geq 0 \, .$$

In addition, from the converse of Fact 3.4.4,

$$(A^\top v_y)_B = 0 \, .$$

We show the stated claim by induction. Denote $z^k = T^k(z^0)$ and $\tilde{z}^k = \tilde{T}^k(z^0)$. First, (3.29) holds with $k = 0$. Now suppose (3.29) holds for k, and consider $k+1$. Then by induction we have $\tilde{z}^{k+1} = \tilde{T}(z^k - kv)$, thus it holds by (3.27) that

$$\begin{aligned}(\tilde{x}^{k+1})_B &= (x^k)_B - k(v_x)_B - \eta A_B^\top (y^k - kv_y) - \eta c_B - (v_x)_B \\ &= (x^{k+1})_B - (k+1)(v_x)_B \,,\end{aligned}$$

where the second equality utilizes the update rule of PDHG by noticing $A_B^\top v_y = 0$ and $(x^{k+1})_B > 0$. For the components in N_1 we get

$$\begin{aligned}(\tilde{x}^{k+1})_{N_1} &= \mathrm{proj}_{\mathbf{R}_+^{|N_1|}}((x_k)_{N_1} - k(v_x)_{N_1} - \eta A_{N_1}^\top (y^k - kv_y) - \eta c_{N_1}) - (v_x)_{N_1} \\ &= \mathrm{proj}_{\mathbf{R}_+^{|N_1|}}((x_k)_{N_1} - \eta A_{N_1}^\top y^k - \eta c_{N_1}) \\ &= (x^{k+1})_{N_1} - (k+1)(v_x)_{N_1} \,,\end{aligned}$$

where the second equality follows from $A_{N_1}^\top v_y = 0$ and $(v_x)_{N_1} = 0$, the third one utilizes $(v_x)_{N_1} = 0$ and the update rule of PDHG. Similarly, for the N_2 block

$$(\tilde{x}^{k+1})_{N_2} = -(v_x)_{N_2} = 0 = (x^{k+1})_{N_2} - (k+1)(v_x)_{N_2} \,,$$

where the equations follow from $(x^{k+1})_{N_2} = 0$ and $(v_x)_{N_2} = 0$. Finally, for the dual iterates we have

$$\begin{aligned}\tilde{y}^{k+1} &= y^k - kv_y + \tau A(2\tilde{x}^{k+1} - \tilde{x}_k) - \tau b - v_y \\ &= y^k + \tau A(2(x^{k+1} - (k+1)v_x) - (x^k - kv_x)) - \tau b - (k+1)v_y \\ &= y^k + \tau A(2x^{k+1} - x^k) - \tau b - (k+1)v_y \\ &= y^{k+1} - (k+1)v_y \,,\end{aligned}$$

where the third equality utilizes $Av_x = 0$, and the last equality is from the update rule of PDHG. □

Equipped with this proposition we can now prove the theorem.

Proof of Theorem 3.4.7. Since $\tilde{T}$ is a PDHG operator, it is firmly nonexpansive with respect to $\|\cdot\|_M$. Thus, if $\tilde{T}$ has a fixed point, then the iteration $\tilde{T}^k(z^K)$ should converge to it. To see that $\tilde{T}$ has a fixed point, let $z^\star$ be a point such that $(T - I)(z^\star) = v$ and let K be the iteration after which $\tilde{T}^k(T^K(z^\star)) = T^{k+K}(z^\star) - kv$, which exists thanks to Proposition 3.4.8. We claim that $T^K(z^\star)$ is a fixed point of $\tilde{T}$. To see this, note that

$$\tilde{T}(T^K(z^\star)) = T^{K+1}(z^\star) - v = T^K(z^\star),$$

where the last equality follows from Lemma 3.3.2.

Now, let z^0 an arbitrary initial point and recall that K is defined in (3.29). Now that we know that the set of fixed points of $\tilde{T}$ is nonempty, we can define $z^\star = \lim_{k\to\infty} \tilde{T}^k(T^K(z_0))$. We will prove that $z^\star$ satisfies $(T - I)(z^\star) = v$. Due to Proposition 3.4.8, we know that

$$z^\star = \tilde{T}(z^\star) = T(z^\star) - v.$$

Finally, using decomposition (3.29) we get

$$\|z^{k+K} - z^\star - kv\| = \|\tilde{T}(z^k) - z^\star\| \to 0\,. \tag{3.31}$$

The statement of theorem claimed this convergence where the coefficient accompanying v is $(k + K)$. We can get around this by setting $z^\star \leftarrow z^\star - Kv$, a point that also satisfies $(T - I)(z^\star) = v$ thanks to Lemma 3.3.2. This establishes the result. □

3.4.3 Proof of Theorem 3.4.1

Proof. As a direct result of Proposition 3.4.3, we know that if both primal and dual are feasible, then $v = 0$ and Theorem 3.4.7 ensures that PDHG converges

to an optimal solution (or equivalently a fixed point of T). If primal (and/or dual) is infeasible, then the dual iterate of PDHG (and/or primal) diverges to infinity, and the diverging direction recovers primal (and/or dual) infeasibility certificate.

Thus, the only thing left to prove is the final conclusion of item 3 (and analogously item 4). Assume that the primal problem is infeasible and the dual is feasible. By Proposition 3.4.3 we know that $v_x = 0$. Then, Theorem 3.4.7 guarantees the existence of some $z^\star = (x^\star, y^\star)$ such that $x^k \to x^\star + kv_x = x^\star$ and $(T - I)(z^\star) = v$. The proof for the case where the primal is feasible and the dual is infeasible follows from an analogous argument. This completes the proof of the theorem. □

3.5 Finite time identifiability and eventual linear convergence

In this section, we introduce a nondegeneracy condition that ensures that after a finite amount the difference of iterates converges linearly to the infimal convergence vector. To show this, we demostrate under said condition the iterates "identify" the support of x^k, i.e., the support freezes after a finite number iterations. Finite-time identification has a long history in the analysis of iterative algorithms for feasible problems [91, 34, 32, 29, 97, 151, 159]. Roughly speaking, these algorithms' behaviors exhibit two phases: a first one that only takes finitely many steps but suffers from slow sublinear convergence, and then a second one after the active set is identified where the convergence is significantly faster and becomes linear.

For PDHG, finite-time identifiability is known to hold for *feasible* minimax

problems under suitable nondegeneracy conditions [159]. In contrast, here we study this phenomenon for infeasible LP problems. We demonstrate that *even when there is no primal (and/or dual) feasible solution, active set of the iterates with respect to an auxiliary feasible LP problem is fixed after finitely many iterations*.

Recall that the iterates of PDHG converge to a ray $[z^\star, v] = \{z^\star + kv \mid k \in \mathbf{N}\}$ where $z^\star$ is a solution to the feasible LP problem given by (3.28). Consider the constraint set defined by said auxiliary problem, that is

$$Ax = b + \frac{v_y}{\tau}, \qquad x_{N_1} \geq 0 \qquad \text{and} \qquad x_{N_2} = 0\,. \tag{3.32}$$

Here, the active set is the set of inequality constraints that attain their extreme values, namely, $\{i \in N_1 \mid x_i^\star = 0\}$. Note that when the problem is feasible, $N_1 = \{1, \ldots, n\}$ and thus the constrained set defined by the auxliary problem (3.32) matches that of the original problem.

Now, we introduce the nondegeneracy condition for possibly infeasible problems, which generalizes the classical identifibility theory of PDHG for feasible LP problems [151, 159]. Similar variations of the nondegeneracy condition have appeared in numerous works that deal with finite time identifiability. Further generalizations of this idea have led to conditions beyond the context of optimization, we refer the interested reader to [148] for a perspective from differential geometry.

Definition 3.5.1. *A ray* $[z^\star, v]$ *is* nondegenerate *if for any* $i \in N_1$ *(recall* N_1 *is defined in* (3.26)*), the pair* $(x_i^\star, (A^\top y^\star + c)_i)$ *satisfies strict complementarity with respect to the auxiliary problem* (3.28)*:* $x_i^\star > 0$ *if, and only if,* $(A^\top y^\star + c)_i = 0$*.*

Although here we chose to define nondegeneracy in terms of the extreme point $z^\star$, this definition is independent of the point we take in the ray $[z^\star, v]$. This

follows easily from the fact that $(v_x)_{N_1}$ and $(A^\top v_y)_{N_1}$ are zero vectors. Additionally, notice that when the original problem is feasible, this definition reduces to the classical strict complementarity of the original problem.

We now state the finite time identifibility of PDHG for infeasible LP:

Lemma 3.5.2. *Suppose* $[z^\star, v]$ *is a nondegenerate ray. Then, every PDHG iterate sequence* $z^k = (x^k, y^k)$*, converging to the ray* $[z^\star, v]$*, fixes the active set of* (3.32) *after finitely many steps. Furthermore, this ensures that the support of* x^k *is fixed for all large enough* k*.*

Proof. First let us prove that the active set of (3.32) is identified in finite time. Notice that this is equivalent to saying that the support of $x^k_{N_1}$ freezes after finitely many iterations. Let $i \in N_1$, due to strict complementary it is enough to consider two cases:

> **Case 1**. Assume that $(Ay^\star + c)_i > 0$, then complementary slackness implies $(x^k)_i \to 0$. By construction, we should have $\eta \cdot (A^\top y^k + c)_i > x^k_i$ for all k large enough. After this condition starts to hold, the PDHG update at i gives
>
> $$x^{k+1}_i = \left(x^k_i - \eta \cdot (A^\top y^k + c)_i\right)_+ = 0 \,.$$
>
> Hence, we have $x^k_i = 0$ for all large k.
>
> **Case 2**. Assume that $x^\star_i > 0$, then after finitely many iterations we have $x^k_i > 0$.

Thus, the support of $x^k_{N_1}$ is identified in finite time. Now, we argue that the same happens to the support of x^k. Assume that $i \in B$, then $(v_x)_i > 0$ and consequently for all large k we have $x^k_i/k > 0$, as we wanted. Lastly, if $i \in N_2$ then $A^\top v_y > 0$

and Fact 3.4.4 guarantees that $x_i^k = 0$ for large enough k. This concludes the proof. □

When nondegeneracy holds, PDHG eventually identifies the support of the primal iterate x^k. This simplifies the form of each iteration. Let S be the support of any x^k with k large enough. The projection to the positive orthant applied by PDHG (3.5) becomes a projection to the subspace $\{x \mid \text{supp}(x) = S\}$. As a consequence, one can recast each iteration (3.5) as an affine transformation:

$$\begin{bmatrix} x^{k+1} \\ y^{k+1} \end{bmatrix} = \underbrace{\begin{bmatrix} I & -\eta DA^\top \\ \tau AD & I - 2\tau\eta ADA^\top \end{bmatrix}}_{Q:=} \begin{bmatrix} x^k \\ y^k \end{bmatrix} - \underbrace{\begin{bmatrix} \eta Dc \\ 2\tau\eta ADc + \tau b \end{bmatrix}}_{p:=} \tag{3.33}$$

where D is a diagonal matrix with ones on the indices (i, i) such that $i \in S$ and zeros everywhere else, and matrix Q and vector p are defined in (3.33).

The next theorem presents upper and lower bounds for the convergence of the three sequences (3.2) under the nondegeneracy condition. In particular, we show that the difference of iterates (3.2a) exhibit eventual linear convergence, while normalized iterates (3.2b) and normalized average iterates (3.2c) exhibit eventual sublinear convergence.

Theorem 3.5.3 (**Eventual convergence rate under nondegeneracy**). *Consider the primal* (**P**) *and dual* (**D**) *problems. Assume that* $\eta\tau\|A\|_2^2 < 1$, *and let* $(z^k)_{k\in\mathbf{N}}$ *be a sequence generated by PDHG. Suppose that the iterates* $z^k = (x^k, y^k)$ *converge to a nondegenerate ray. Then, the kth power of* Q *converges* $\lim_{k\to\infty} Q^k = Q^\infty$ *to a projection matrix, and there exists a finite* K *such that for any* $k \geq 0$, *the active set of* x^{K+k} *is fixed. Furthermore, there exist positive constants* $\underline{\mu}, c_1, c_2, c_3, c_4 > 0$ *such that the following holds:*

1. ***Linear convergence of the differences***. *For any* $\mu \in (\sqrt{1 - \eta\tau\sigma_{\min}(A)^2}, 1)$, *the differences* $z^{K+k+1} - z^{K+k}$ *converge at a linear rate to the infimal displacement vector*

v, i.e., for all sufficiently large k

$$\underline{\mu}^k \|(Q-I)z^K + (Q^\infty - I)p\|_2 \leq \|z^{K+k+1} - z^{K+k} - v\|_2 \leq \mu^k \|(Q-I)z^K - p\|_2. \quad (3.34)$$

2. ***Sublinear convergence of the iterates***. *The normalized iterates converge to v at a* $\Theta\left(\frac{1}{k}\right)$ *rate, i.e., for all sufficiently large k*

$$\frac{c_1}{k}\mathcal{L}^k \leq \left\|\frac{1}{k}z^{K+k} - v\right\|_2 \leq \frac{c_2}{k}\mathcal{U}^k \quad (3.35)$$

where $\mathcal{L}^k = \|(I-Q^\infty)p\|_2$ *and* $\mathcal{U}^k = \left(\|(Q-I)^\dagger(I-Q^\infty)p - z^K\|_2 + \|z^K\|_2\right)$.

3. ***Sublinear convergence of the average***. *The normalized average converges to v at a* $\Theta(\frac{1}{k})$ *rate, i.e., for all sufficiently large k,*

$$\frac{c_3}{k+1}\mathcal{L}^k \leq \left\|\frac{2}{k(k+1)}\sum_{j=1}^{k} z^{K+j} - v\right\|_2 \leq \frac{c_4}{k+1}\mathcal{U}^k \quad (3.36)$$

where $\mathcal{L}^k = \|(I-Q)p\|_2$ *and* $\mathcal{U}^k = \left(\|(Q-I)^\dagger(I-Q^\infty)p - z^K\|_2 + \|z^K\|_2\right)$.

Some remarks are in order. Although equations (3.35) and (3.36) state a bound for the normalized iterates and normalized average of PDHG started from z^K, this result implies the same asymptotic bounds (with sightly worse constants) for the normalized iterates and normalized averaged started from z^0. Thus, the result concludes that under nondegeneracy, the difference of iterates converges much faster than the iterates and average. Furthermore, such conclusion is tight, as we provide both the upper and lower bounds for each sequence.

The proof of this result shows the same rates for Bilinear games. Recall that a Bilinear game is a minimax problem of the form

$$\min_{x\in\mathbf{R}^n}\max_{x\in\mathbf{R}^m} c^\top x + \langle Ax, y\rangle - b^\top y\,.$$

For these problems, the updates of PDHG (Algorithm 1) take the form of (3.33) with $D = I$. Thus, all the arguments in the proof of this result follow with $K = 0$.

In particular, this shows that the upper bound derived in Theorem 3.3.1 is tight for Bilinear minimax problems.

Every LP problem that has an optimal solution furthermore has at least one primal-dual solution that satisfies strict complementarity [170, Theorem 2.35]. Consequently, every infeasible problem has at least one nondegenerate ray. Thus there exists at least one initial point z^0, such that if PDHG is initialized at this point, then the iterates converge to the nondegenerate ray and thus enjoy linear convergence.

3.6 Numerical experiments

In this section, we test numerically the proposed approach to check infeasibility using PDHG. For the experiments, we implemented PDHG for LP problems with the following primal and dual form

$$\begin{array}{lll} \text{minimize} & c^\top x & \text{maximize} \quad b^\top y + l^\top r_+ - u^\top r_- \\ \text{subject to} & Ax \geq b & \text{subject to} \quad c - A^\top y = r \\ & l \leq x \leq u & \qquad\qquad y \geq 0 \end{array} \quad , \tag{3.37}$$

where $b \in \mathbf{R}^m, l \in (\mathbf{R} \cup \{-\infty\})^n, u \in (\mathbf{R} \cup \{\infty\})^n, A \in \mathbf{R}^{m \times n}$ are given and $r_+ = \text{proj}_{\mathbf{R}^n_+}(r)$ and $r_- = -\text{proj}_{\mathbf{R}^n_+}(-r)$ are the projections of r onto the positive and negative orthant, respectively. We chose this form over the standard form (**P**)-(**D**) since it is algorithmically easier to reduce arbitrary LP problems to it. Given that the PDHG algorithm for this formulation generates iterates (x^k, y^k), for our computations we generate r^k by finding the closest point to $c - A^\top y^k$ that makes the dual

objective finite; i.e., $r^k = \text{proj}_\Lambda(c - A^\top y^k)$ with

$$\Lambda = \left\{ x \in \mathbf{R}^n : \text{ for } i \in [n],\, x_i \begin{cases} = 0 & \text{if } l_i = -\infty, u_i = \infty, \\ \geq 0 & \text{if } l_i \in \mathbf{R}, u_i = \infty, \\ \leq 0 & \text{if } l_i = -\infty, u_i \in \mathbf{R}, \end{cases} \right\} .$$

All the results proved in this chapter also apply to this form under suitable modifications of the statements.

For our experiments we use the *Netlib dataset of infeasible LP instances* [102]. We use this dataset to illustrate the different dynamics that PDHG exhibits. For all our experiments we measure statistics that quantify how close are the candidate iterates (3.2) to being approximate certificates of inteasibility.

Before we describe these statistics, let us define what we mean by approximate infeasibility certificates. The set of (exact) primal infeasibility certificates for (3.37) is given by all the vectors $(y, r) \in \mathbf{R}^m_+ \times \mathbf{R}^n$ satisfying

$$b^\top y + l^\top r_+ - u^\top r_- > 0 \,, \quad \text{and} \quad r = -A^\top y \,, \tag{3.38}$$

while the set of (exact) dual infeasibility certificates is given by all the vectors $x \in \mathbf{R}^n$ satisfying

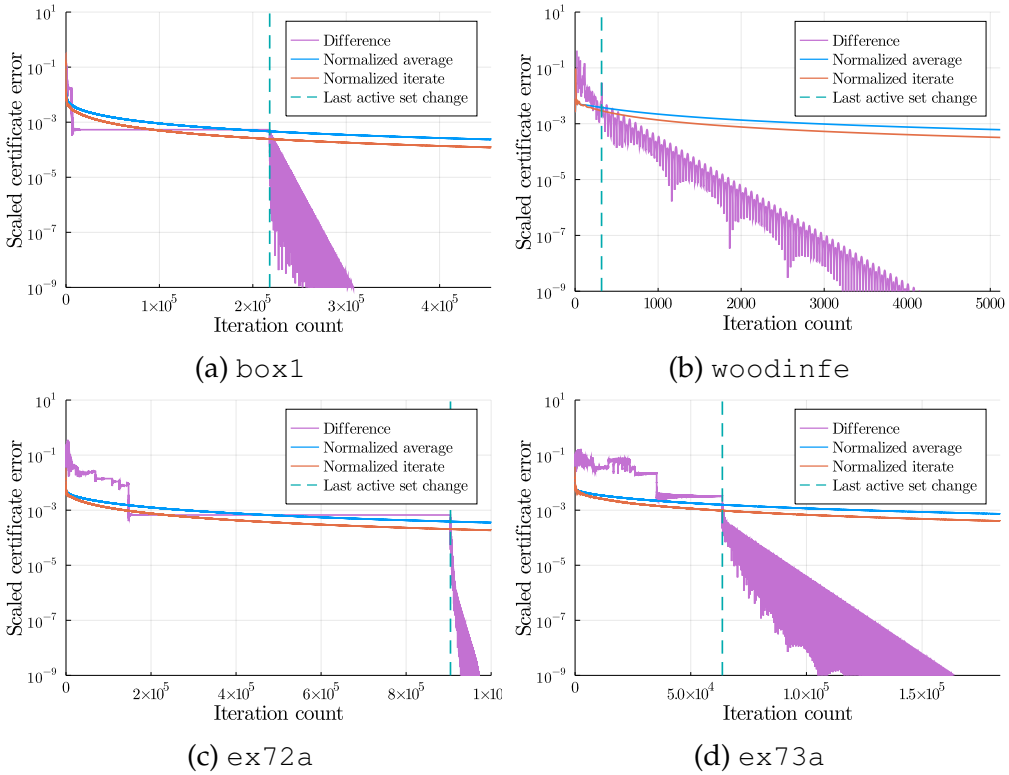
$$c^\top x < 0, \quad x \in C_v \,, \quad \text{and} \quad Ax \geq 0 \,, \tag{3.39}$$

where the set C_v is given by

$$C_v = \left\{ x \in \mathbf{R}^n : \text{ for } i \in [n],\, x_i \begin{cases} = 0 & \text{if } l_i, u_i \in \mathbf{R}, \\ \geq 0 & \text{if } l_i \in \mathbf{R}, u_i = \infty, \\ \leq 0 & \text{if } l_i = -\infty, u_i \in \mathbf{R}, \end{cases} \right\} .$$

We define an ε-approximate primal infeasibility certificate to be any point $(y, r) \in \mathbf{R}^m_+ \times \mathbf{R}^n$ satisfying

$$b^\top y + l^\top r_+ - u^\top r_- > 0 \,, \quad \text{and} \quad (b^\top y + l^\top r_+ - u^\top r_-)^{-1} \|r + A^\top y\|_\infty \leq \varepsilon \,. \tag{3.40}$$



(a) `box1` (b) `woodinfe`

(c) `ex72a` (d) `ex73a`

Figure 3.2: Scaled certificate error (3.42) for the three sequences defined in (3.2) for four instances of the Netlib infeasible dataset [102]. Vertical dotted lines denote the last observed active set change.

Similarly we say that a point $x \in \mathbf{R}^m$ is an ϵ-approximate dual infeasibility certificate if it satisfies

$$c^\top x < 0, \quad \frac{1}{-c^\top x} \cdot \|x - \mathrm{proj}_{C_v}(x)\|_\infty \leq \varepsilon\,, \qquad \text{and} \qquad \frac{1}{-c^\top x} \cdot \|Ax - \mathrm{proj}_{\mathbf{R}^m_+}(Ax)\|_\infty \leq \varepsilon\,. \tag{3.41}$$

These definitions parallel the criteria to detect infeasibility used by SCS [224], a popular open-source solver.

Since all the instances in the Netlib infeasible data set are primal infeasible, we will only plot information about the dual components of the candidate certificates (3.2). To illustrate how close is each candidate to being a certificate we

will plot

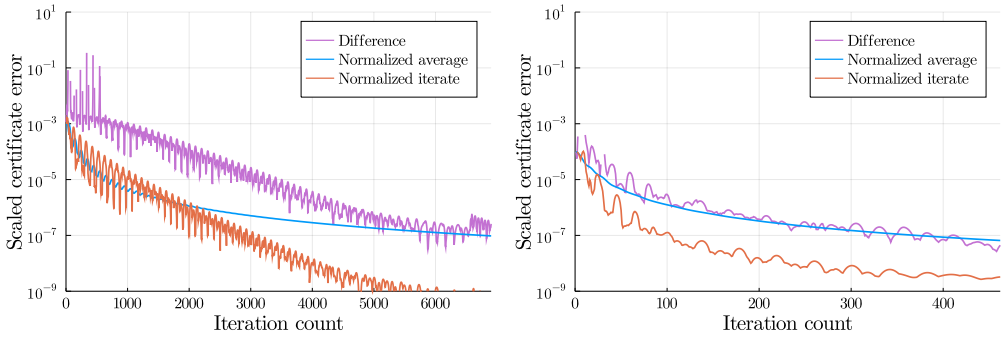
$$\frac{\|r^k + A^\top y^k\|_\infty}{b^\top y^k + l^\top r_+^k - u^\top r_-^k}\,. \tag{3.42}$$

We call this quantity the *scaled certificate error*. If the objective term, i.e., the denominator, is negative at any iteration then we do not plot it for that iteration. For almost all the problems we consider (3.42) remains positive for almost all iterations.

Nondegeneracy in practice. Our first batch of experiments showcases the faster convergence of the difference of iterates in practice. We found empirically that for a subset of instances in the dataset, the difference of iterates (3.2a) detects infeasibility faster than the other two sequences. Based on the theory, we expect that for these instances, the difference exhibits eventual faster convergence. To test this claim, we run an experiment on four of these instances: `box1`, `woodinfe`, `ex73a` and `ext72a`.

Figure 3.2 displays the scaled certificate error (3.42) against the number of iterations for the four instances. For all of them, we can see a clear phase transition between a first stage of slow convergence and a second stage that displays linear convergence. This transition is unequivocally marked by the last change of the active set of the solution (also depicted in the figure). Notice, however, that the iteration number at which the active set is fixed might be large; the point at which this happens ranges among multiple orders of magnitude in our experiments.

Normalized iterates can be faster. Even if eventual identifiability holds, this might take a significant number of iterations. In these cases it might be beneficial to check infeasibility using the normalized iterated (3.2b) and the normalized average (3.2c). In this batch of experiments we run PDHG on `bgdbg1` and



(a) `bgdbg1` (b) `chemcom`

Figure 3.3: Scaled certificate error (3.42) for the three sequences defined in (3.2) for two instances of the Netlib infeasible dataset [102].

`chemcom`, the results are displayed in Figure 3.3. Just as before we plot the scaled certificate error against the number of iterations.

The normalized average is consistently slower at converging than the normalized iterates. This is most likely due to the fact that it retains a tail of initial iterates, which are far away from the infimal displacement vector. For both these problems, the difference takes at least twice the number of iterations than the normalized iterates to obtain a highly accurate certificate, i.e., $\varepsilon = 10^{-8}$. This suggests that solvers may benefit from checking infeasibility with both the normalized iterates (3.2b) and difference of iterates (3.2a).

3.7 Analysis

3.7.1 Proof of Proposition 3.2.3

Proof. Assume that $(z^{k+1} - z^k) \to v$. Fix $\varepsilon > 0$. Our goal is to show that for all k large enough $\|z^k/k - v\| \leq \varepsilon$. Due to convergence, there exist $K_1 \in \mathbf{N}$ such that for

all $k \geq K_2$ we have $\|z^{k+1} - z^k - v\| \leq \varepsilon/3$. Define

$$B := \max_{k \leq K_1} \|z^{k+1} - z^k - v\|,$$

let $K_2 \in \mathbf{N}$ be such that for all $k \geq K_2$ we get $K_1 B/k \leq \varepsilon/3$, and let $K_3 \in \mathbf{N}$ be such that if $k \geq K_3$ then $\|z^0\|/k \leq \varepsilon/3$. Then, for any $k \geq \max\{K_1, K_2, K_3\}$ we have

$$\begin{aligned}
\left\| \frac{z^k}{k} - v \right\| &\leq \left\| \frac{1}{k}(z^k - z^0) - v \right\| + \frac{1}{k}\|z^0\| \\
&= \left\| \frac{1}{k} \sum_{j=1}^{k} (z^j - z^{j-1}) - v \right\| + \frac{1}{k}\|z^0\| \\
&\leq \frac{1}{k} \sum_{j=1}^{k} \left\| (z^j - z^{j-1}) - v \right\| + \frac{1}{k}\|z^0\| \\
&\leq \frac{K_1}{k} B + \frac{1}{k} \sum_{j=K_1}^{k} \left\| (z^j - z^{j-1}) - v \right\| + \frac{1}{k}\|z^0\| \\
&\leq \frac{2}{3}\varepsilon + \frac{1}{3k} \sum_{j=K_1}^{k} \varepsilon \leq \varepsilon \,.
\end{aligned}$$

This proves the first statement.

Now, assume that $\frac{z^k}{k} \to v$, and fix $\varepsilon > 0$. Just as before define $K_1 \in \mathbf{N}$ to be such that for all $\|z^k/k - v\| \leq \varepsilon/2$, define the constant $B = \max_{k \leq K_1} \|z^k/k - v\|$, and let K_2 be such that $B(K_1 + 1)K_1/((K_2 + 1)K_2) \leq \varepsilon/2$. Then, we have that for any $k \geq \max\{K_1, K_2\}$,

$$\begin{aligned}
\left\| \frac{2}{(k+1)k} \sum_{j=1}^{k} z^j - v \right\| &= \frac{2}{(k+1)k} \left\| \sum_{j=1}^{k} z^j - \frac{k(k+1)}{2} v \right\| \\
&= \frac{2}{(k+1)k} \left\| \sum_{j=1}^{k} (z^j - jv) \right\| \\
&\leq \frac{2}{(k+1)k} \sum_{j=1}^{k} j \left\| \frac{z^j}{j} - v \right\| \\
&\leq \frac{(K_1+1)K_1}{(k+1)k} B + \frac{2}{(k+1)k} \sum_{j=K_1} j \left\| \frac{z^j}{j} - v \right\| \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \left(\frac{2}{(k+1)k} \sum_{j=K_1}^{k} j \right) \leq \varepsilon \,.
\end{aligned}$$

The counterexamples can be found in Section 3.7.3. This concludes the proof.

□

3.7.2 Proof of Theorem 3.5.3

We start by making a few simplifying assumptions. First we assume $D = I$. If that is not the case, we can consider a submatrix of A where we trim out the columns indexed by $\{i \in n \mid D_{ii} = 0\}$. This has no effect in the end result since for all these indices $x_i^k = 0$, and thus it does not affect the nonzero entries of x^k nor the entries of y^k. Furthermore, without loss of generality we assume that A is diagonal. This is because, otherwise, we can decompose the matrix Q as

$$Q = \begin{bmatrix} V & 0 \\ 0 & U \end{bmatrix} \begin{bmatrix} I & \eta\Sigma^\top \\ \tau\Sigma & I - 2\eta\tau\Sigma\Sigma^\top \end{bmatrix} \begin{bmatrix} V^\top & 0 \\ 0 & U^\top \end{bmatrix} \tag{3.43}$$

where $A = U\Sigma V^\top$ is the SVD decomposition of A. Then we can change the primal and dual basis using V and U, which is equivalent to applying orthogonal maps to the primal and dual variables; thus it does not alter the metric nor the algorithm. Therefore, we can change the basis to enforce this assumption. Notice that the columns and rows of Q can be further permuted so that it becomes a block diagonal matrix of the form

$$Q = \begin{bmatrix} B_1 & & & \\ & \ddots & & \\ & & B_{\min\{n,m\}} & \\ & & & I \end{bmatrix} \quad \text{where} \quad B_i = \begin{bmatrix} 1 & -\eta\sigma_i \\ \tau\sigma_i & 1 - 2\tau\eta\sigma_i^2 \end{bmatrix}, \tag{3.44}$$

where σ_i is a the ith singular value of A. Note that when $n > m$ (or $n < m$), we might also have block corresponding to an identity of size $n - m$ (resp. $m - n$),

the arguments below can be easily extended to cover the identity block (as it follows from the rationale applied when $\sigma_i = 0$) and so we assume that $n = m$. Thus, from now on Q has the form (3.44) with $n = m$ and hence without the last identity block.

Now, we can compute closed-form formulas for the three certificate candidates (3.2a)-(3.2c). Let K be the smallest integer after which the PDHG iteration can be written as (3.33). By recursively expanding, we obtain

$$\begin{aligned} z^{k+1+K} &= Qz^{k+K} - p \\ &= Q^2 z^{k-1+K} - Qp - p \\ &\vdots \\ &= Q^{k+1} z^K - \sum_{i=0}^{k} Q^i p \,. \end{aligned} \tag{3.45}$$

If we take the difference between two consecutive iterates, this yields

$$z^{k+1+K} - z^{k+K} = Q^k(Q - I)z^K - Q^k p = Q^k((Q - I)z^K - p) \,. \tag{3.46}$$

On the other hand, summing the first k iterates (3.45) gives

$$\sum_{j=1}^{k} z^{j+K} = \sum_{j=1}^{k} Q^j z^K - \sum_{j=1}^{k} \sum_{l=0}^{j-1} Q^l p \,. \tag{3.47}$$

We will show that Q^k converges to a matrix Q^∞. We define the matrix Q^∞ as a block diagonal matrix

$$Q^\infty = \begin{bmatrix} B_1^\infty & & \\ & \ddots & \\ & & B_n^\infty \end{bmatrix} \qquad \text{where} \qquad B_i^\infty = \begin{cases} I_2 & \text{if } \sigma_i = 0 \\ 0 & \text{otherwise} \end{cases} , \tag{3.48}$$

where I_2 is the 2-by-2 identity matrix. Since each block is independent of each other, we can analyze Q^k by studying B_i^k. A simple calculation reveals that the

ith block has two eigenvalues of the form

$$\lambda_i^{\pm} = (1 - \eta\tau\sigma_i^2) \pm i\left(\eta\tau\sigma_i^2\left(1 - \eta\tau\sigma_i^2\right)\right)^{\frac{1}{2}} .$$

Taking the norm, we find $\rho(B_i) = |\lambda_i^{\pm}| = \sqrt{1 - \eta\tau\sigma_i^2}$. Then, we have that the iterated product of the ith block B_i^k converges to B_i^∞. To see this, consider two cases:

Case 1. Assume that $\sigma_i > 0$. By assumption $0 < 1 - \eta\tau\sigma_i^2 < 1$ hence $B_i^k \to 0 = B_i^\infty$. This follows since the spectral radius $\rho(B_i^k) = (1-\eta\tau\sigma_i^2)^{k/2} \to 0$ as k.

Case 2. Assume that $\sigma_i = 0$. Then $B_i^k = B_i = I = B_i^\infty$.

The matrix Q^∞ turns out to be the projection onto the kernel of $Q - I$ (that is, $Q^\infty(Q - I) = 0$). We use $\Delta \subseteq [n]$ to denote the set of indices such that $\sigma_i > 0$.

Differences. We start by analyzing the differences (3.46). To prove the upper bound, we expand

$$\begin{aligned}
\|z^{K+k+1} - z^{K+k} - v\|_2 &= \|Q^k(((Q - I)z^K - p) - Q^\infty((Q - I)z^K - p)\|_2 \\
&\leq \|Q^k - Q^\infty\|_2\|(Q - I)z^K - p\|_2 \\
&= \max_{i\in\Delta} \|B_i^k\|_2\|(Q - I)z^K - p\|_2 \\
&\leq \mu^k\|(Q - I)z^K - p\|_2 ,
\end{aligned}$$

where the first equality comes from taking the limit $k \to \infty$ of (3.46), the first inequality used the fact that $\|Wz\|_2 \leq \|W\|_2\|z\|_2$ for any matrix W and vector z, and the last inequality follows since $\rho(B_i) = \lim \|B_i^k\|^{\frac{1}{k}}$ and hence for any $\mu \in (\rho(B_i), 1)$ we have that $\|B_i^k\| \leq \mu^k$ for all large enough k.

Now we turn our attention to the lower bound. Given a vector z, we define z_i to be the vector with the two components that multiply the block B_i in the matrix-vector product Qz. Using the same expansion as above, we get

$$\begin{aligned}
\|z^{K+k+1} - z^{K+k} - v\|_2^2 &= \|Q^k(((Q-I)z^K - p) - Q^\infty((Q-I)z^K - p)\|_2^2 \\
&= \sum_{i\in[n]} \|(B_i^k - B_i^\infty)((Q-I)z^K - p)_i\|_2^2 \\
&= \sum_{i\in\Delta} \|B_i^k((Q-I)z^K - p)_i\|_2^2 \\
&\geq \sum_{i\in\Delta} \sigma_{\min}(B_i^k)^2 \|((Q-I)z^K - p)_i\|_2^2 \\
&\geq \min_{i\in\Delta} \sigma_{\min}(B_i^k)^2 \sum_{i\in\Delta} \|((Q-I)z^K - p)_i\|_2^2 \\
&= \min_{i\in\Delta} \sigma_{\min}(B_i^k)^2 \|(I - Q^\infty)((Q-I)z^K - p)\|_2^2 \\
&= \left(\min_{i\in\Delta} \sigma_{\min}(B_i)\right)^{2k} \|(Q-I)z^K + (Q^\infty - I)p\|_2^2 ,
\end{aligned}$$

where the second equality and the penultimate one use the block-diagonal structure of the matrix Q to decompose the norm of the matrix-vector product into orthogonal components, and the first inequality follows since B_i^k is a rank two matrix for any $i \in \Delta$.

Normalized iterates. We now turn our attention to the normalized iterates. The upper bound follows almost immediately from Theorem (3.3.1) if we consider the PDHG algorithm started at z^K. To show the bound with the theorem, it suffices to note that 1) in this case $v = -Q^\infty p$ and so we might pick $z^\star = (Q-I)^\dagger(I - Q^\infty)p$, where $(I-Q)^\dagger$ is the pseudo inverse of $(I-Q)$; and 2) all the norms in finite dimensional spaces are equivalent so we can upper bound $\|\cdot\|_M \leq C\|\cdot\|_2$ for some constant $C > 0$. To see the first point note that

$$\begin{aligned}
Qz^\star - p - z^\star = -Q^\infty p \iff (Q-I)z^\star &= (I - Q^\infty)p \\
\impliedby z^\star &= (Q-I)^\dagger(I-Q^\infty)p .
\end{aligned}$$

Now we turn our attention to the lower bound. Just as before we analyze the dynamics of Q^k by studying the individual blocks B_i^k. We will use the following identity for blocks satisfying $\rho(B) < 1$:

$$\sum_{j=0}^{k} B^j = (I - B)^{-1}(I - B^{k+1}) \,. \tag{3.49}$$

Recall that $\Delta = \{i \mid \sigma_i > 0\}$, which corresponds with blocks satisfying $\rho(B_i) < 1$. Additionally, recall that p_i is the vector with the two components of p that multiply the block B_i in the matrix-vector product Qp. Expanding we get

$$\begin{aligned}
\left\| v - \frac{1}{k} z^{K+k+1} \right\|^2 &= \left\| Q^\infty p + \frac{1}{k} \sum_{j=0}^{k} Q^j p - Q^{k+1} z^K \right\|^2 \\
&= \sum_{i \in [n]} \left\| B_i^\infty p_i + \frac{1}{k} \sum_{j=0}^{k} B_i^j p_i - \frac{1}{k} B_i^{k+1} z_i^K \right\|^2 \\
&= \frac{1}{k^2} \sum_{i \in \Delta} \left\| (I - B_i)^{-1}(I - B_i^{k+1}) p_i - B_i^{k+1} z_i^K \right\|^2 + \frac{1}{k^2} \sum_{i \notin \Delta} \| B_i^{k+1} z_i^K \|^2 \\
&\geq \frac{1}{k^2} \sum_{i \in \Delta} (\sigma_{\max}(I - B_i))^{-2} \left\| (I - B_i^{k+1}) p_i - (I - B_i) B_i^{k+1} z_i^K \right\|^2 \,,
\end{aligned}$$

where for the last two equalities we used the fact that Q is block diagonal, and the last inequality follows since $I - B_i$ is invertible for $i \in \Delta$. Then, taking the minimum coefficient we get

$$\begin{aligned}
&\frac{1}{k^2} \sum_{i \in \Delta} \sigma_{\max}(I - B_i)^{-2} \left\| (I - B_i^{k+1}) p_i - (I - B_i) B_i^{k+1} z^K \right\|^2 \\
&\qquad \geq \frac{1}{k^2} \min_{j \in \Delta} \left\{ \sigma_{\max}(I - B_j)^{-2} \right\} \sum_{i \in \Delta} \left\| (I - B_i^{k+1}) p_i - (I - B_i) B_i^{k+1} z^K \right\|^2 \\
&\qquad = \frac{1}{k^2} \min_{j \in \Delta} \left\{ \sigma_{\max}(I - B_j)^{-2} \right\} \left\| (I - Q^{k+1}) p - (I - Q) Q^{k+1} z^K \right\|^2 \\
&\qquad \geq \frac{1}{2k^2} \min_{j \in \Delta} \left\{ \sigma_{\max}(I - B_j)^{-2} \right\} \| (I - Q^\infty) p \|^2 \,,
\end{aligned}$$

where the last equality uses the fact that the matrices we handle are block diagonal, and the last line follows since $\|(I - Q^{k+1})p - (I - Q)Q^{k+1} z^K\| \to \|(I - Q^\infty)p\|$, and so the inequality holds for sufficiently large $k \geq 0$.

Normalized average. The upper bound follows from the exact same argument as the normalized iterates by using Theorem 3.3.1.

The lower bound for this case is sightly more intricate. We expand and apply (3.49)

$$\begin{aligned}
&\left\| v - \frac{2}{k(k+1)} \sum_{j=1}^{k} z^{K+j} \right\|^2 \\
&\quad = \left\| Q^\infty p - \frac{2}{k(k+1)} \sum_{j=1}^{k} \sum_{l=0}^{j-1} Q^l p + \frac{2}{k(k+1)} \sum_{j=1}^{k} Q^j z^K \right\|^2 \\
&\quad = \sum_{i \in [n]} \left\| B_i^\infty p_i - \frac{2}{k(k+1)} \sum_{j=1}^{k} \sum_{l=0}^{j-1} B_i^l p_i + \frac{2}{k(k+1)} \sum_{j=1}^{k} B^j z_i^K \right\|^2 \\
&\quad = \sum_{i \in \Delta} \left\| \frac{2}{k(k+1)} \sum_{j=1}^{k} \sum_{l=0}^{j-1} B_i^l p_i + \frac{2}{k(k+1)} \sum_{j=1}^{k} B_i^j z_i^K \right\|^2 \sum_{i \notin \Delta} \left\| \frac{2}{k(k+1)} \sum_{j=1}^{k} B_i^j z_i^K \right\|^2 \\
&\quad = \frac{4}{k^2(k+1)^2} \sum_{i \in \Delta} \left\| (I - B_i)^{-1} \left(\sum_{j=1}^{k} (I - B_i^j) p_i + (I - B_i^{k+1}) z_i^K \right) - z_i^K \right\|^2 + \sum_{i \notin \Delta} \left\| \frac{2}{k(k+1)} \sum_{j=1}^{k} B_i^j z_i^K \right\|^2 .
\end{aligned}$$

The second equality follows from the fact that Q and Q^∞ are block diagonal. Then, dropping the second sum and using the fact that

$$\|(I - B_i)^{-1} z\|_2 \geq \sigma_{\max}(I - B_i) \|z\|_2$$

for all z and $i \in \Delta$, we can lower bound

$$\begin{aligned}
&\frac{4}{k^2(k+1)^2}\sum_{i\in\Delta}\left\|(I-B_i)^{-1}\left(\sum_{j=1}^{k}(I-B_i^j)p_i + B_i(I-B_i^k)z_i^K\right)\right\|^2 + \sum_{i\notin\Delta}\left\|\frac{2}{k(k+1)}\sum_{j=1}^{k}B_i^j z_i^K\right\|^2 \\
&\quad\geq \frac{4}{k^2(k+1)^2}\sum_{i\in\Delta}\sigma_{\max}(I-B_i)^{-2}\left\|\sum_{j=1}^{k}(I-B_i^j)p_i + B_i(I-B_i^k)z_i^K\right\|^2 \\
&\quad\geq \frac{4}{k^2(k+1)^2}\min_{i\in\Delta}\left\{\sigma_{\max}(I-B_i)^{-2}\right\}\sum_{i\in\Delta}\left\|kp_i - (I-B_i)^{-1}B_i(I-B_i^k)p_i + B_i(I-B_i^k)z_i^K\right\|^2 \\
&\quad\geq \frac{4}{k^2(k+1)^2}\min_{i\in\Delta}\left\{\sigma_{\max}(I-B_i)^{-4}\right\}\sum_{i\in\Delta}\left\|k(I-B_i)p_i - B_i(I-B_i^k)p_i + (I-B_i)B_i(I-B_i^k)z_i^K\right\|^2 \\
&\quad= \frac{4}{k^2(k+1)^2}\min_{i\in\Delta}\left\{\sigma_{\max}(I-B_i)^{-4}\right\}\left\|k(I-Q)p - Q(I-Q^k)p + (I-Q)Q(I-Q^k)z^K\right\|^2 \\
&\quad= \frac{4}{(k+1)^2}\min_{i\in\Delta}\left\{\sigma_{\max}(I-B_i)^{-4}\right\}\left\|(I-Q)p - \frac{1}{k}Q(I-Q^k)p + \frac{1}{k}(I-Q)Q(I-Q^k)z^K\right\|^2 \\
&\quad\geq \frac{2}{(k+1)^2}\min_{i\in\Delta}\left\{\sigma_{\max}(I-B_i)^{-4}\right\}\|(I-Q)p\|^2 \ ,
\end{aligned}$$

where the last inequality follows for large enough k since

$$\left\|-\frac{1}{k}Q(I-Q^k)p + \frac{1}{k}(I-Q)Q(I-Q^k)z^K\right\|^2 \to 0.$$

This completes the proof.

3.7.3 Counterexamples

Example 3.7.1 (Differences don't converge, but normalized iterates do). *Consider the sequence $(z^k) \subseteq \mathbf{R}$ that alternates $z^k = (-1)^k$. For this example, the differences of iterates $z^{k+1} - z^k$ also alternate between -2 and 2, and, consequently, do not converge. Nonetheless, since the iterates are bounded $\frac{1}{k}z^k \to 0$.*

Example 3.7.2 (Normalized iterates diverge, but normalized averages converge). *Consider the sequence $(z^k) \subseteq \mathbf{R}$ given by $z^k = (-1)^k k^{\frac{3}{2}}$ with $k \in \mathbf{N}$. Then, it is clear that $|z^k|/k > \sqrt{k}$, and so the normalized iterates diverge. On the other hand,*

notice that

$$\frac{2}{(k+1)}\bar{z}^k = \sum_{j=1}^{k}(-1)^j \frac{2k^{\frac{1}{2}}}{k+1},$$

and it is easy to show that this series converges using the Leibniz Test.

Example 3.7.3 (**Nonexpansive operator with divergent** z_ϵ)**.** *Let* $T : \mathbf{R} \to \mathbf{R}$ *given by*

$$T(z) = z + f(z) \qquad \textit{where} \qquad f(z) = \begin{cases} \exp(-z^2) + 1 & \textit{if } z > 0, \textit{ or} \\ 2 & \textit{otherwise.} \end{cases} \tag{3.50}$$

Since the derivative of T *is bounded by* 1*, we get that* T *is a nonexpansive operator. Furthermore,* $\text{range}(T - I) = \text{range}(f) = (1, 2]$*, and so* $v = 1$*. If we define* z_ϵ *to be a point such that* $|v - (T - I)(z_\epsilon)| \leq \epsilon$*, we see that* $z_\epsilon > \Omega\left(\log\left(\frac{1}{\epsilon}\right)^{\frac{1}{2}}\right)$*, and thus it diverges as* $\epsilon \to 0$*.*

4

OPTIMAL CONVERGENCE RATES FOR THE PROXIMAL BUNDLE METHOD

"If you only read the books that everyone else is reading,
you can only think what everyone else is thinking."

— Haruki Murakami, *Norwegian Wood*

4.1 Introduction

Convex optimization has played a fundamental role in recent developments in high-dimensional statistics, signal processing, and data science. Large-scale applications have motivated researchers to develop first-order methods with computationally simple iterations. Although impressive in scope, these methods often require delicate parameter tunning involving geometrical information about the objective function. Thus, imposing an obstacle for practitioners that rarely have access to such information.

In this work, we develop efficiency guarantees for *proximal bundle methods*, which date back to the 70s, that solve unconstrained convex problems

$$\underset{x\in\mathbf{R}^d}{\text{minimize}}\ f(x) \tag{4.1}$$

where $f\colon \mathbf{R}^d \to \mathbf{R}$ is a proper closed convex. Our core finding is that classic bundle methods, without any modification, are adaptive, which means that they speed up in the presence of smoothness or error bounds, with little to no tunning.

Proximal bundle methods were independently proposed in [143] and [242]. They are conceptually similar to model-based methods [66, 193, 85]. That is, methods that update their iterates by applying a proximal step to an approximation of the function, known as the model f_k:

$$x_{k+1} \leftarrow \arg\min_x f_k(x) + \frac{\rho_k}{2}\|x - x_k\|^2.$$

Unlike these schemes, bundle methods only update their iterates x_k when the decrease in objective value is at least a fraction of the decrease that the model predicted. Moreover, bundle methods incorporate information from past iterations into their models, allowing f_k to capture more than the just objective's geometry near x_k.

This seemingly subtle change has a rather surprising consequence: the iterates generated by a bundle method, with *any* constant parameter configuration, converge to a minimizer of f; see [129, Thm. 4.9], [118, Thm. XV.3.2.4], or [219, Thm. 7.16] for different variations of this result. This stands in harsh contrast to other first-order algorithms; for example, gradient descent and its accelerated variants rely on selecting a stepsize inversely proportional to the level of smoothness. Similarly, subgradient methods rely on carefully controlled decreasing stepsize sequences. These simpler algorithms may fail to converge when the stepsizes are not carefully managed. Thus, providing a compelling reason to consider bundle methods.

Although bundle methods are known to converge under a number of assumptions [128, 177, 12, 112, 166, 72, 181, 180] and have been successfully used in applications [221, 220, 73], nonasympotic guarantees have remained mostly evasive. The purpose of this chapter is to close this gap. We study convergence rates for finding an ε-minimizer, e.g., $f(x) - \inf f \leq \varepsilon$, under a variety of different

assumptions on f. We consider settings where the objective function is either M-Lipschitz continuous

$$|f(x) - f(y)| \leq M\|x - y\| \qquad \text{for all } x, y \in \mathbf{R}^d \tag{4.2}$$

or differentiable with an L-Lipschitz gradient, often referred to as L-smoothness,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \qquad \text{for all } x, y \in \mathbf{R}^d\,. \tag{4.3}$$

In either setting, we investigate the method's rate of convergence with and without the presence of Hölder growth

$$f(x) - \inf f \geq \mu \cdot \operatorname{dist}(x, X^*)^p \qquad \text{for all } x \in \mathbf{R}^d\,, \tag{4.4}$$

where $X^* = \{x \mid f(x) = \inf f\}$ is the set of minimizers.[1] Particularly important cases are when $p = 1$ and $p = 2$, which correspond to sharp growth (μ-SG) [30] and quadratic growth (μ-QG), generalizing strong convexity, respectively.

4.1.1 Main contributions

Our **first contribution** is to establish convergence rates under every realizable combination of continuity/smoothness (4.2) or (4.3) and growth assumptions (4.4), see Table 4.1. Full theorem statements are given in Section 4.2 and apply for any Hölder growth exponent (rather than just the cases of $p = 1$ and $p = 2$ shown in the table). Our analysis technique is fairly general as we apply it seamlessly to every combination of assumptions as well as different stepsize rules. We show rates for any constant stepsize $\rho_k = \rho$, which tend to be suboptimal. Yet, they improve under amenable geometry. Tuning the constant ρ

[1] Here $\operatorname{dist}(x, S) = \inf_{y \in S} \|x - y\|$.

to depend on a target accuracy ϵ yields faster convergence rates. Further, we propose nonconstant stepsize rules ρ_k with two clear advantages: they yield yet faster convergence and their convergence does not slow down after reaching the target accuracy.

Assumptions		Rate for generic ρ	Rate for tuned ρ	Rate for adaptive ρ_k
M-Lipschitz	No Growth	$O\left(\frac{M^2\|x_0 - x^*\|^4}{\rho\epsilon^3}\right)$	$O\left(\frac{M^2\|x_0 - x^*\|^2}{\epsilon^2}\right)$	$O\left(\frac{M^2\|x_0 - x^*\|^2}{\epsilon^2}\right)$
	μ-QG	$O\left(\frac{M^2}{\min\{\mu,\rho\}\epsilon}\right)$	$O\left(\frac{M^2}{\mu\epsilon}\right)$	$O\left(\frac{M^2}{\mu\epsilon}\right)$
	μ-SG	$O\left(\frac{M^2}{\rho\epsilon}\right)$	$O\left(\frac{M^2}{\mu^2}\sqrt{\frac{\Delta_f}{\epsilon}}\right)$	$O\left(\frac{M^2}{\mu^2}\log\left(\frac{\Delta_f}{\epsilon}\right)\right)$
L-Smooth	No Growth	$O\left(\frac{L^3\|x_0 - x^*\|^2}{\rho^2\epsilon}\right)$	$O\left(\frac{L\|x_0 - x^*\|^2}{\epsilon}\right)$	$O\left(\frac{L\|x_0 - x^*\|^2}{\epsilon}\right)$
	μ-QG	$O\left(\frac{L^3}{\rho^2\mu}\log\left(\frac{\Delta_f}{\epsilon}\right)\right)$	$O\left(\frac{L}{\mu}\log\left(\frac{\Delta_f}{\epsilon}\right)\right)$	$O\left(\frac{L}{\mu}\log\left(\frac{\Delta_f}{\epsilon}\right)\right)$

Table 4.1: Convergence rates. We denote $\Delta_f := f(x_0) - \inf f$. The first column applies for any choice of the parameter ρ, showing progressively faster convergence as more structure is introduced. The second column shows the rate after optimizing the choice of ρ. The third column further improves these by allowing nonconstant stepsizes ρ_k.

The existing convergence theory for the proximal bundle method applies to settings comparable to the first two rows of our table. Kiwiel [131] derived a $O(\epsilon^{-3})$ convergence rate for Lipschitz problems, which agrees with our theory. Du and Ruszczynski [88] and subsequently Liang and Monteiro [157] showed a $O(\log(1/\epsilon)/\epsilon)$ convergence rate for Lipschitz, strongly convex problems, which we improve on by removing the extra logarithmic term and thus achieve the optimal convergence rate for this setting of $O(1/\epsilon)$. To our knowledge, the rest of our convergence results apply to wholly new settings for the proximal bundle method. In all of the M-Lipschitz settings considered, we show that using a nonconstant stepsize the bundle method attains the optimal nonsmooth convergence rate. In the L-smooth settings considered, the bundle method converges

at the same rate as gradient descent. Although, unlike gradient descent, our convergence theory applies to any configuration of its algorithmic parameters.

Our **second contribution** is proposing a parallelizable variant of the bundle method that avoids the reliance on tuning a stepsize or sequence of stepsizes based on potentially unrealistic knowledge of underlying problem constants. This approach too seamlessly falls under the umbrella of our analysis. It attains the optimal nonsmooth convergence rates for Lipschitz problems with any level of Hölder growth, up to the cost of running a logarithmic number of instances of the bundle method in parallel.

4.1.2 Related work

In 2000, Kiwiel [131] gave the first convergence rate for the proximal bundle method, showing that an ϵ-minimizer x_k is found with $k \leq O\left(\frac{\|x_0 - x^*\|^4}{\epsilon^3}\right)$. More recently, Du and Ruszczyński [88] gave the first analysis of bundle methods when applied to problems satisfying a quadratic growth bound. In this case, an ϵ-minimizer is found within $O(\log(1/\epsilon)/\epsilon)$ iterations. Following this, Liang and Monteiro [157] showed a variant of the proximal bundle method with proper stepsize selection attains the optimal convergence rate for convex and strongly convex optimization, up to logarithmic terms.

Despite historically having weaker convergence rate guarantees than simple alternatives like the subgradient method, bundle methods have persisted as a method of choice for nonsmooth convex optimization. See [100, 144] for a survey of much of the bundle method literature. In practice, bundle methods have proven to be efficient methods for solving many nonsmooth problems,

see [221, 220, 73] for further discussion. Extensions that apply to nonconvex problems have been considered in in [128, 177, 12, 112, 166, 72, 181, 180] and as well as an extension to problems where only an inexact first-order oracle is available in [113, 74, 165].

Stronger convergence rates have been established for related level bundle methods [145], which share many core elements with proximal bundle methods. Variations of level bundle methods were studied in [130] and [137]. The results of Lan [137] are particularly impressive as their proposed method has optimal convergence rates for both smooth and nonsmooth problems while requiring little input.

Outline of the chapter . Section 4.2 introduces the Proximal Bundle Method and provides the formal convergence guarantees under different regularity assumptions. This section also introduces simple stepsize rules that guarantee optimal convergence rates for all nonsmooth settings. Practical implementations of these rules require access growth constants of the function. To bypass this issue, in Section 4.3 we propose an adaptive parallel bundle method that exhibits nearly the same convergence rates without knowledge of such constants. We complement our findings with numerical experiments in Section 4.4. Finally, Section 4.5 presents a broadly applicable proof technique to analyze bundle methods and uses it to establish the theoretical results.

Algorithm 2: Proximal Bundle Method

Data: $z_0 = x_0 \in \mathbf{R}^n$, $f_0(z) = f(x_0) + \langle g_0, z - x_0 \rangle$

Step k**:** $(k \geq 0)$

Compute candidate iterate $z_{k+1} \leftarrow \arg\min_{z \in \mathcal{X}} f_k(z) + \frac{\rho_k}{2}\|z - x_k\|^2$.

If $\beta(f(x_k) - f_k(z_{k+1})) \leq f(x_k) - f(z_{k+1})$ (Descent step)

set $x_{k+1} \leftarrow z_{k+1}$,

Else (Null step)

set $x_{k+1} \leftarrow x_k$.

Update f_{k+1} and ρ_{k+1} without violating Assumption 4.2.1.

4.2 Bundle methods

In this section, we formally define the family of proximal bundle methods that our theory applies to. We present the convergence rates for the classic method with constant stepsizes. Additionally, we introduce and analyze nonconstant stepsize rules that guarantee faster convergence rates.

Proximal bundle methods work by maintaining a model function $f_k \colon \mathbf{R}^n \to \mathbf{R}$ at each iteration k and a current iterate x_k. The method computes a candidate for the next iterate as

$$z_{k+1} = \arg\min_{z \in \mathcal{X}} f_k(z) + \frac{\rho_k}{2}\|z - x_k\|^2.$$

However, unlike other model-based algorithms, bundle methods do not necessarily move their next iterate to z_{k+1}. Instead, it first checks whether the candidate z_{k+1} has at least $\beta \in (0, 1)$ fraction of the decrease in objective value that our model $f_k(\cdot)$ predicts. If it does, it updates $x_{k+1} = z_{k+1}$ as the next iterate, this is called a *Descent Step*. Otherwise the method keeps the iterate the same $x_{k+1} = x_k$ and updates the model function f_{k+1}, called a *Null Step*.

The proximal bundle method is stated fully in Algorithm 2. Our analysis does not presume a particular parametrization or form of the models. We only

assume that the models satisfy mild assumptions, typical of bundle methods in the literature. To state the assumptions, note the first-order optimality conditions define a subgradient

$$s_{k+1} = \rho_k(z_{k+1} - x_k) \in \partial f_k(z_{k+1}) \quad \text{for each } k \geq 0$$

where $\partial f(x) = \{g \mid f(x') \geq f(x) + \langle g, x' - x\rangle \ \ \forall x' \in \mathbf{R}^d\}$ denotes the subdifferential of f at x.

Assumption 4.2.1. *Let $\left\{f_k \colon \mathbf{R}^d \to \mathbf{R}\right\}$ and $\{\rho_k\}$ be the sequence of models and stepsizes used throughout the execution of a bundle method. Assume that for any iteration $k \geq 0$, the next model f_{k+1} and stepsize ρ_{k+1} satisfy the following:*

1. ***Minorant.***

$$f_{k+1}(x) \leq f(x) \qquad \text{for all } x \in \mathbf{R}^d \,. \tag{4.5}$$

2. ***Subgradient lowerbound.*** *There is a subgradient $g_{k+1} \in \partial f(z_{k+1})$ such that*

$$f_{k+1}(x) \geq f(z_{k+1}) + \langle g_{k+1}, x - z_{k+1}\rangle \qquad \text{for all } x \in \mathbf{R}^d \,. \tag{4.6}$$

3. ***Model subgradient lowerbound.*** *After a null step k*

$$f_{k+1}(x) \geq f_k(z_{k+1}) + \langle s_{k+1}, x - z_{k+1}\rangle \qquad \text{for all } x \in \mathbf{R}^d \,. \tag{4.7}$$

4. ***Constant stepsize between null steps.*** *After a null step k*

$$\rho_{k+1} = \rho_k \,. \tag{4.8}$$

The first two conditions are natural as they ensure that a new model incorporates first-order information from the objective at z_{k+1}. The third condition is mild and, intuitively, requires the new model to retain some of the approximation accuracy of the previous model. The last assumption is trivial to enforce and guarantees the algorithm only changes its stepsize after it decides to move.

4.2.1 Model function choices

Several methods for constructing model functions f_k that satisfy (4.5)-(4.7) have been considered. In practice, the main consideration lies in weighing the potentially greater per iteration gains from having more complex models against the lower iteration costs from having simpler models.

Full-memory proximal bundle method. The earliest proposed bundle methods [143, 242] rely on using all of the past subgradient evaluations to construct the models as

$$f_{k+1}(x) = \max_{j=0\ldots k+1} \left\{ f(z_j) + \langle g_j, x - z_j \rangle \right\}. \tag{4.9}$$

In this case, solving the quadratically regularized subproblem at each iteration amounts to solving a quadratic programming problem.

Finite Memory Proximal Bundle Method. Using cut-aggregation[127, 129], the collection of $k+1$ lower bounds used by (4.9) can be simplified down to just two linear lower bounds. The only two necessary lower bounds are exactly those required by (4.6) and (4.7). Namely, one could construct the model functions as

$$f_{k+1}(x) = \max \left\{ f_k(z_{k+1}) + \langle \rho_k(z_{k+1} - x_k), x - z_{k+1} \rangle,\ f(z_{k+1}) + \langle g_{k+1}, x - z_{k+1} \rangle \right\}. \tag{4.10}$$

Then the subproblem that needs to be solved at each iteration can be done in closed form, see (4.17). Hence the iteration cost using this model is limited primarily by the cost of one subgradient evaluation.

Spectral bundle methods. Both of the above models rely on constructing piecewise linear models of the objective. For more structure problems, richer models can be constructed. For example, in eigenvalue optimization or more broadly semidefinite programming, better spectral lower bounds can be constructed instead of using simple polyhedral bounds [117, 198]. Primal-dual convergence rate guarantees for such spectral bundle methods were recently developed by Ding and Grimmer [79].

4.2.2 Convergence rates from constant stepsize choice

We now formalize our convergence theory for the proximal bundle method using any constant choice of the stepsize parameter $\rho_k = \rho$ and any $\beta \in (0, 1)$. These guarantees match those claimed in the first column of Table 4.1. After each theorem, we remark on the tuned choice of ρ that gives rise to the claimed rate in the second column of Table 4.1. We start by considering the setting where only Lipschitz continuity is assumed.

Theorem 4.2.2 (**Lipschitz**). *For any M-Lipschitz convex objective function f, consider applying the bundle method using a constant stepsize $\rho_k = \rho$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most*

$$\frac{2\rho D^2}{\beta\epsilon} + \left\lceil \frac{\log\left(\frac{f(x_0)-f(x^*)}{\rho D^2}\right)}{-\log(1-\beta/2)} \right\rceil_+$$

and the number of null steps is at most

$$\frac{12\rho M^2 D^4}{\beta(1-\beta)^2\epsilon^3} + \frac{8M^2}{\beta(1-\beta)^2\rho^2 D^2}$$

where $D^2 = \sup_k \|x_k - x^*\|^2 < \infty$.

Remark 5. *It follows from [219][(7.64)] that* $D^2 \leq \|x_0 - x^*\|^2 + \frac{2(1-\beta)(f(x_0)-f^*)}{\beta\rho}$. *Alternatively, if the level sets of* f *are bounded, the fact that* $f(x_k)$ *is non-increasing ensures* $D^2 \leq \sup\{\|x - x^*\|^2 \mid f(x) \leq f(x_0)\}$.

Remark 6. *Selecting* $\rho = \epsilon/D^2$ *gives an overall complexity bound of*

$$O\left(\frac{M^2D^2}{\epsilon^2}\right)$$

and matches the optimal rate for nonsmooth, Lipschitz convex optimization.

If instead of Lipschitz continuity of the objective, we assume the objective has Lipschitz gradient, the bundle method adapts to give the following faster rate.

Theorem 4.2.3 (Smooth). *For any L-smooth convex objective function* f, *consider applying the bundle method using a constant stepsize* $\rho_k = \rho$. *Then for any* $0 < \epsilon \leq f(x_0) - f(x^*)$, *the number of descent steps before an* ϵ*-minimizer is found is at most*

$$\frac{2\rho D^2}{\beta\epsilon} + \left\lceil \frac{\log\left(\frac{f(x_0)-f(x^*)}{\rho D^2}\right)}{-\log(1-\beta/2)} \right\rceil_+$$

and the number of null steps is at most

$$\frac{4(L+\rho)^3}{(1-\beta)^2\rho^3}\left(\frac{2\rho D^2}{\beta\epsilon} + \left\lceil \frac{\log\left(\frac{f(x_0)-f(x^*)}{\rho D^2}\right)}{\log(1-\beta/2)} \right\rceil_+ + 1\right)$$

where $D^2 = \sup_k \|x_k - x^*\|^2 < \infty$.

Remark 7. *Selecting* $\rho = L$ *gives an overall complexity bound of*

$$\frac{16LD^2}{\beta(1-\beta)^2\epsilon}.$$

This matches the standard convergence rate for gradient descent.

Next, we reconsider the settings of Lipschitz continuity and smoothness with additional structure in the form of a Hölder growth bound. We find that the convergence guarantees divide into three regions depending on the growth exponent p, whether it is large, equal to, or smaller than 2. Here two is the critical exponent value since the proximal subproblem is adding in quadratic regularization. Regardless, as p decreases, the bundle method converges faster.

Theorem 4.2.4 (Lipschitz with Hölder growth). *For any M-Lipschitz objective function f satisfying the Hölder growth condition* (4.4)*, consider applying the bundle method using a constant stepsize $\rho_k = \rho$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most*

$$\begin{cases} \dfrac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + \left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{-\log(1-\beta/2)} \right\rceil_+ & \text{if } p > 2 \\ \left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta\min\{\mu/2\rho, 1/2\})} \right\rceil & \text{if } p = 2 \\ \left\lceil \dfrac{\log\left(\frac{(\rho/\mu^{2/p})^{1/(1-2/p)}}{\epsilon}\right)}{-\log(1-\beta/2)} \right\rceil_+ + \dfrac{2\rho(f(x_0)-f^*)^{2/p-1}}{(1-2^{1-2/p})\beta\mu^{2/p}} & \text{if } 1 \leq p < 2 \end{cases}$$

and the number of null steps is at most

$$\begin{cases} \dfrac{12\rho M^2}{(1-2/p)\beta(1-\beta)^2\mu^{4/p}\epsilon^{3-4/p}} + \dfrac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} & \text{if } p > 2 \\ \dfrac{2M^2}{\beta(1-\beta)^2\min\{\mu/2\rho, 1/2\}\rho\epsilon} & \text{if } p = 2 \\ \dfrac{4M^2}{\beta(1-\beta)^2\rho\epsilon} + \dfrac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}}C & \text{if } 1 \leq p < 2 \end{cases}$$

with $C = \max\left\{\frac{(f(x_0)-f^*)^{4/p-3}}{(\rho/\mu^{2/p})^{(4/p-3)/(1-2/p)}}, 1\right\}\min\left\{\frac{1}{1-2^{-|4/p-3|}}, \left\lceil \log_2\left(\frac{f(x_0)-f^*}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)\right\rceil\right\}$.

Remark 8. *When $p = 2$, selecting $\rho = \mu$ gives an optimal overall complexity bound of $O(M^2/\mu\epsilon)$. Selecting $\rho = O(\epsilon^{1-2/p})$ matches the optimal rate for Lipschitz optimization with growth exponent $p > 2$. When $p = 1$, selecting $\rho = O(1/\sqrt{\epsilon})$ minimizes this bound, but the resulting sublinear $O(1/\sqrt{\epsilon})$ rate falls short of the best possible rate*

(linear convergence) for sharp, Lipschitz optimization. In the next section where we consider nonconstant stepsizes, this disconnect will be remedied and a linear convergence guarantee will follow.

Theorem 4.2.5 (Smooth with Hölder growth). *For any L-smooth objective function f satisfying the Hölder growth condition* (4.4)*, consider applying the bundle method using a constant stepsize $\rho_k = \rho$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most*

$$\begin{cases} \dfrac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + \left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{-\log(1-\beta/2)} \right\rceil_+ & \text{if } p > 2 \\ \left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta\min\{\mu/2\rho, 1/2\})} \right\rceil & \text{if } p = 2 \end{cases}$$

and the number of null steps is at most

$$\begin{cases} \dfrac{4(L+\rho)^3}{(1-\beta)^2\rho^3}\left(\dfrac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + \left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{-\log(1-\beta/2)} \right\rceil_+ + 1 \right) & \text{if } p > 2 \\ \dfrac{4(L+\rho)^3}{(1-\beta)^2\rho^3}\left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta\min\{\mu/2\rho, 1/2\})} \right\rceil & \text{if } p = 2\,. \end{cases}$$

Remark 9. *Selecting $\rho = L$ gives an overall complexity bound matching gradient descent.*

4.2.3 Convergence Rates from Improved Stepsize Choice

Picking ρ_k to vary throughout the execution of the bundle method allows for stronger convergence guarantees. These rates are formalized in the following pair of theorems that consider settings with and without Hölder growth. In the latter case, we find that our stepsize choice removes the need for piecewise guarantees around growth exponent $p = 2$, which notably simplifies the statement of our results.

Intuitively, the stepsize choices aim to mimic the following idealistic (and impractical) stepsize rule that naturally arises from our theory

$$\rho_k = \frac{f(x_k) - f(x^*)}{\|x_k - x^*\|^2} \,. \tag{4.11}$$

The proof techniques we develop could be extended to study other interesting nonconstant stepsizes. For instance, stepsizes that shrink/grow polynomial with the number of descent steps, mirroring those used for subgradient methods. The analysis of such schemes is beyond the scope of this work.

Theorem 4.2.6 (**Lipschitz**). *For any M-Lipschitz objective function f, consider applying the bundle method using the stepsize policy*

$$\rho_k = (f(x_k) - f(x^*))/D^2 \tag{4.12}$$

with any choice of $D^2 \geq \sup\{\|x - x^*\|^2 \mid f(x) \leq f(x_0)\}$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta/2)} \right\rceil$$

and the number of null steps is at most

$$\left(\frac{1}{1-(1-\beta/2)^2}\right)\frac{2M^2D^2}{(1-\beta)^2\epsilon^2} \,.$$

Theorem 4.2.7 (**Lipschitz with Hölder growth**). *For any M-Lipschitz objective function f satisfying the Hölder growth condition* (4.4)*, consider applying the bundle method using the stepsize policy*

$$\rho_k = \mu^{2/p}(f(x_k) - f(x^*))^{1-2/p}. \tag{4.13}$$

Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta/2)} \right\rceil$$

and the number of null steps is at most

$$\begin{cases} \left(\dfrac{1}{1-(1-\beta/2)^{2-2/p}}\right)\dfrac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} & \text{if } p > 1 \\ \dfrac{2M^2}{(1-\beta)^2\mu^2}\left\lceil\dfrac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta/2)}\right\rceil & \text{if } p = 1\ . \end{cases}$$

4.3 The parallel bundle method

We now give a practical scheme for applying the bundle method that attains the same complexity as our optimally tuned nonconstant stepsizes without any knowledge of the presence of smoothness or growth bounds. We do this by employing a logarithmic number of instances of the bundle method with different constant stepsizes in parallel that continually share their progress with each other. By doing so, we recover our optimal rates, up to the cost of running a logarithmic number of algorithms which can be mitigated through parallelization. This scheme is inspired by the ideas of [212].

The core observation behind our parallel method is that our nonconstant stepsize rules (4.12) and (4.13) before an ϵ-minimizer is found are always in the following interval

$$\rho_k \in \left[O(\epsilon), O(\epsilon^{-1})\right]\ .$$

As input, we only assume the following are given: a lower bound $\bar{\rho}$ and an upper bound $2^J\bar{\rho}$ on the range of stepsizes to consider. Provided our stepsize rules (4.12) and (4.13) lie in this interval,

$$\rho_k \in \left[\bar{\rho}, 2^J\bar{\rho}\right]\ ,$$

we are able to recover our optimal convergence rates. Notice that the interval $[\bar{\rho}, 2^J\bar{\rho}]$ can span the whole range of stepsizes needed for our Hölder growth

analysis by setting $\bar{\rho} = O(\epsilon)$ and $J = O(\log(1/\epsilon^2))$. Our resulting convergence guarantees only depends logarithmically on the size of this interval (a cost which can be mitigated through parallelization), so $\bar{\rho}$ and $2^J\bar{\rho}$ can bet set conservatively at little cost.

Description of the algorithm. We propose running J copies of the bundle method in parallel, which share their progress with each other as described below. Each bundle method $j \in \{0, \dots J-1\}$ uses a constant stepsize $\rho^{(j)} = 2^j\bar{\rho}$. Denote the iterates of bundle method j by $x_k^{(j)}$ and its model objectives by $f_k^{(j)}$. Each bundle method j proceeds as normal with the only modification being that after it takes a descent step, the algorithm checks if any other bundle method j' has an iterate with an even lower objective value $f(x_k^{(j')}) < f(x_{k+1}^{(j)})$. If such an improvement exists, the bundle method instead descends to the best such iterate, setting

$$\begin{cases} x_{k+1}^{(j)} & \leftarrow x_k^{(j')} \\ f_{k+1}^{(j)}(z) & \leftarrow f(x_k^{(j')}) + \langle g_k^{(j')}, z - x_k^{(j')} \rangle \end{cases}$$

and then proceeds.

For analysis sake, we will assume that each parallel instance of the bundle method operates synchronously, with every instance completing one iteration before any instance completes a second iteration. This process can be implemented sequentially by cycling through the bundle method instances computing one iteration for each before repeating. An asynchronous variant of this procedure could be analyzed as well, using similar techniques as those in [212]. However, this is beyond the focus of this work. Note the choice to use powers of two here is arbitrary. In the following numerical section, we use powers of 10 and 100 demonstrating the effectiveness of this scheme even when using a

sparse selection of sample stepsizes.

4.3.1 Convergence Rates for the Parallel Bundle Method

First, we remark that all of our previous convergence theory for constant stepsizes (Theorems 4.2.2, 4.2.3, 4.2.4, and 4.2.5) immediately apply to the Parallel Bundle Method fixing $\rho = 2^j\bar{\rho}$ for any $j \in \{0, \dots J-1\}$. This follows as our convergence theory on relies on a lemma ensuring sufficient decrease at each descent step (Lemma 4.5.1) and the new case of a bundle method restarting at another method's lower objective value iterate can only further improve on this decrease. Hence any individual instance of the bundle method with $\rho^{(j)} = 2^j\bar{\rho}$ in our parallel scheme will converge at least as fast as Theorems 4.2.2, 4.2.3, 4.2.4, and 4.2.5 guarantee it would converge on its own.

Further and more importantly, when our nonconstant stepsize rules (4.12) and (4.13) lie in the interval $[\bar{\rho}, 2^J\bar{\rho}]$, we find that their convergence theory (Theorems 4.2.6 and 4.2.7) also extends to our parallel algorithm. This is formalized as follows.

Theorem 4.3.1. *For any M-Lipschitz objective function f that satisfies the Hölder growth condition* (4.4)*, consider applying the Parallel Bundle Method with stepsizes $\rho = 2^j\bar{\rho}$ for $j \in \{0, \dots, J-1\}$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, if*

$$\bar{\rho} \leq \frac{1}{4}\mu^{2/p} \min\{\epsilon^{1-2/p}, (f(x_0) - f(x^*))^{1-2/p}\}$$

and

$$J \geq \log_2\left(\frac{\mu^{2/p}(\max\{\epsilon^{1-2/p}, (f(x_0) - f(x^*))^{1-2/p}\}}{4\bar{\rho}}\right),$$

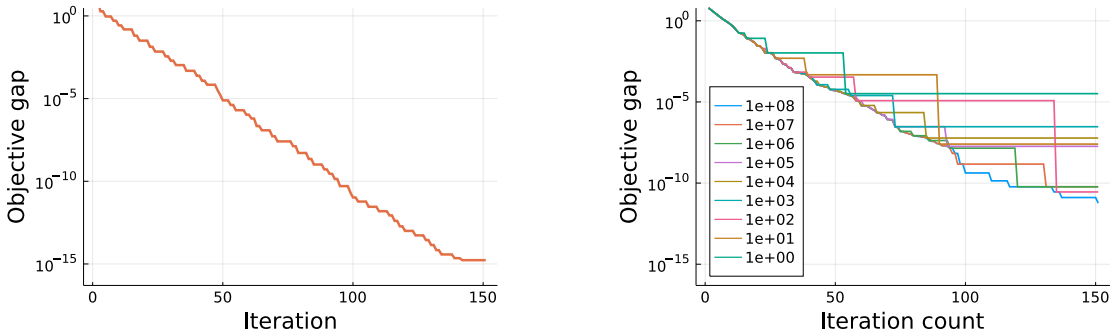


Figure 4.1: Objective gap against iteration count: using ideal stepsize (4.11) (left) and using the parallel bundle method, plotting each instance deployed with stepsizes from $10^0, \ldots, 10^8$ (right).

then one of our J bundle methods will find an ϵ-minimizer within its first

$$\begin{cases} \left(\dfrac{2}{1-(1-\beta/2)^{2-2/p}}\right)\dfrac{16M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} + 2\left\lceil \dfrac{\log(\frac{f(x_0)-f^*}{\epsilon})}{-\log(1-\beta/2)}\right\rceil & \text{if } p > 1 \\ 2\left(\dfrac{16M^2}{(1-\beta)^2\mu^2}+1\right)\left\lceil \dfrac{\log(\frac{f(x_0)-f^*}{\epsilon})}{-\log(1-\beta/2)}\right\rceil & \text{if } p = 1 \end{cases}$$

iterations.

4.4 Numerical experiments

In this section, we present two examples that illustrate numerically the theory for the bundle method. These experiments were implemented in `Julia`, see the github repository `https://github.com/mateodd25/proximal-bundle-method`.

4.4.1 Sharp linear regression

The first experiment aims to exemplify the fast convergence of the bundle method under sharp growth. We consider a simple linear regression problem of

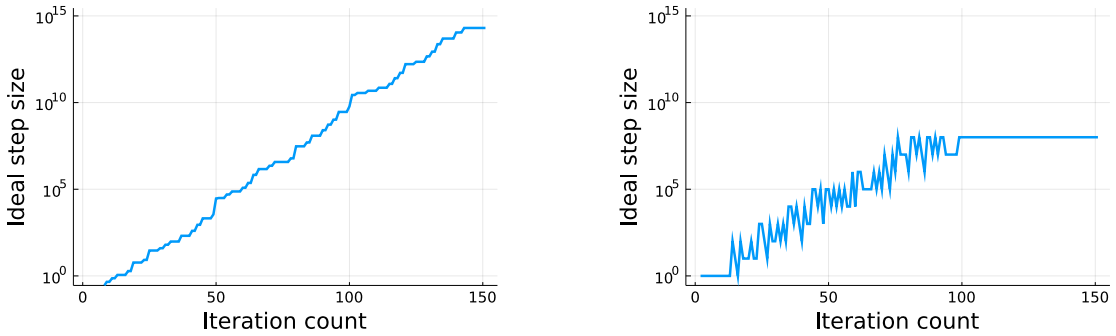


Figure 4.2: Stepsize against iteration count: using ideal stepsize (4.11) (left) and using the parallel bundle method (right).

the form

$$\min_{x \in \mathbf{R}^d} f(x) := \|Ax - b\|$$

where $A \in \mathbf{R}^{n \times d}$ is a matrix and $b = Ax^\star$ for a fixed $x^\star$. This problem is equivalent to the classic least-squares problem after taking squares. Yet, without the square it is well known that for Gaussian matrices, $(A)_{ij} \sim N(0, \frac{1}{\sqrt{n}})$, this function is sharp and Lipchitz continuous provided n is large enough.

We generate a random Gaussian matrix $A \in \mathbf{R}^{100 \times 50}$ and random solution $x^\star \sim N(0, I_d)$. We run two algorithms: the proximal bundle method with the "ideal" stepsize (4.11) and the parallel bundle method described in Section 4.3. The ideal stepsize is impractical since it requires knowing the optimal solution. However, the theoretical analysis shows that it gives optimal convergence rates. In fact, the stepsizes proposed in our results (4.12) and (4.13) try to mimic its behavior. Thus, the method with ideal stepsize serves as a point of comparison. The parallel bundle method uses 9 parallel instances with stepsizes in $\rho \in \{1, 10, \ldots, 10^8\}$. We let both methods run for 150 iterations.

Figure 4.1 displays the objective gap $f - \min f$ against the iteration count for both methods. On the other hand, Figure 4.2 shows the stepsize used at each iteration. For the parallel bundle method, we display the stepsize used by the

last instance to reduce the best objective value seen.

As the theory predicts the convergence of both methods is linear. The bundle method with ideal stepsize exhibits steady progress and reaches an objective gap of $1.70 \cdot 10^{-15}$, while the parallel version slows down around 100 iterations and only achieves $5.87 \cdot 10^{-12}$. This behavior is explained by the stepsize plots. Figure 4.2 plots how the parallel algorithm roughly emulates the ideal stepsize until it exceeds the largest instance's stepsize 10^8. After which, the instance with stepsize 10^8 consistently leads the method's progress, albeit sublinearly.

4.4.2 Support Vector Machine

To illustrate the adaptive features of the parallel bundle method we consider the standard Support Vector Machine (SVM) formulation: we are given datapoints $(x_1, y_1), \dots (x_n, y_n)$ with $x_i \in \mathbf{R}^d$ and $y_i \in \{\pm 1\}$ and our goal is to solve

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum \max\{0, 1 - y_i \langle w, x_i \rangle\} + \frac{\lambda}{2} \|w\|^2 \tag{4.14}$$

where $\lambda \in \mathbf{R}$ is a fixed constant. This problem is not smooth due to the first term. For this experiment we compare against a subgradient method based on Pegasos [222], a state-of-the art solver for SVM. Our vanilla implementation of the parallel bundle method is not tuned for efficiency and does not aim to be competitive with commercial solvers. Instead, we aim to show that an out-of-the-box implementation is immediately comparable to a specialized first-order method for this problem.

We generate SVM problems using three datasets from the LIBSVM Binary Classification Database [1]. In particular, we use `colon-cancer`, `duke`, and

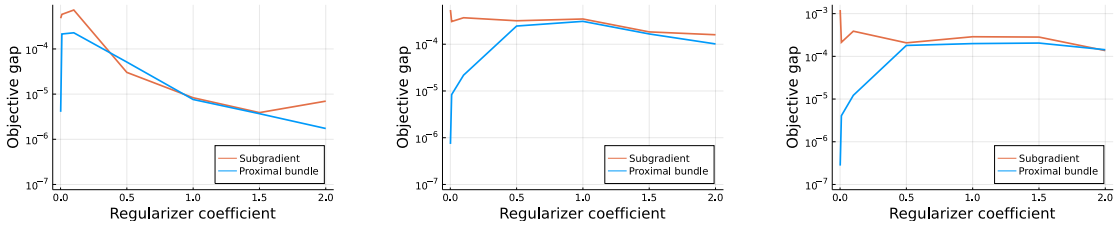


Figure 4.3: Objective gap against coefficient λ for the three problems, solved by a subgradient method and by the parallel bundle method: `colon-cancer` (left), `duke` (center), and `leu` (right).

`leu`.[2] We preprocess the data by deleting empty features, normalizing the features, and adding an extracomponent $x_k = (x_k, 1)$ to allow for affine functions.

The implementation of the subgradient algorithm updates

$$w_{k+1} \leftarrow (1 - \eta_k \lambda) w_k + \eta_k \sum_{i=1}^{n} \mathbf{1}\{1 \le y_i \langle w_k, x_i \rangle\} y_i x_i$$

where $\eta_k = \frac{1}{\lambda k}$ and $\mathbf{1}\{\cdot\}$ is one if $\cdot$ holds true and zero otherwise. This is analogous to Pegasos with the exception that it does full, instead of stochastic, subgradient evaluations.

For the parallel bundle method, we use stepsizes 11 instances with constant stepsizes

$$\rho \in \{10^{-15} \cdot 100^j \mid j = 0, \ldots, 10\}.$$

We run both methods for 2000 iterations and measure the objective gap $f - \min f$. To compute the minimum we use Gurobi with accuracy set to 10^{-10}. Figure 4.3 plots the gap against while varying the regularizer coefficient within $\lambda \in \{0.001, 0.01, 0.1, 0.5, 1.5, 2.0\}$.

In this simple setting, the parallel bundle method out of the box performs similarly to the tuned subgradient method while only requiring a constant amount of extra work (that can be parallelized). We see that the parallel method

[2] We refer the reader to LIBSVM for the origin of each of these datasets.

with the same parameter configuration can handle a wide range of parameters λ. While for small λ the performance of the subgradient method tends to deteriorate, the performance of the bundle method improves (outperforming the subgradient method by several orders of magnitude).

4.5 Analysis

In this section, we develop the proofs of the convergence rates. We start by introducing the general strategy that we use to establish all of our results and then specialize it to each scenario.

4.5.1 Analysis Overview and Proof Sketch

Each iteration of the bundle method can be viewed as an attempt to mimic the proximal point method, using the model f_k instead of the true objective function f. At each iteration k, we denote the objective gap of the proximal subproblem, called the *proximal gap*, by

$$\Delta_k := f(x_k) - \left(f(\bar{x}_{k+1}) + \frac{\rho_k}{2} \|\bar{x}_{k+1} - x_k\|^2 \right)$$

where $\bar{x}_{k+1} = \arg\min_{x \in \mathbf{R}^d} \left\{ f(x) + \frac{\rho_k}{2} \|x - x_k\|^2 \right\}$.

Regardless of which continuity, smoothness and growth assumptions are made, our analysis works by relating the proximal steps computed by the bundle method on the models f_k to proximal steps on f. The following pair of observations show that the behavior on both descent steps and null steps is controlled by the proximal gap Δ_k.

(i) **Descent steps attain decrease proportional to the proximal gap.**

Lemma 4.5.1. *A descent step, at iteration k, has*

$$f(x_{k+1}) \leq f(x_k) - \beta\Delta_k.$$

(ii) **The number of consecutive null steps is bounded by the proximal gap.**

Lemma 4.5.2. *A descent step, at iteration k, followed by T consecutive null steps has at most*

$$T \leq \frac{2G_{k+1}^2}{(1-\beta)^2\rho_{k+1}\Delta_{k+1}}$$

where $G_{k+1} = \sup\{\|g_{t+1}\| \mid k \leq t \leq k+T\}$. This simplifies to

$$T \leq \begin{cases} \dfrac{2M^2}{(1-\beta)^2\rho_{k+1}\Delta_{k+1}} & \text{if } f \text{ is } M\text{-Lipschitz, or} \\ \dfrac{4(L+\rho_k)^3}{(1-\beta)^2\rho_{k+1}^3} & \text{if } f \text{ is } L\text{-smooth .} \end{cases}$$

With these two observations in hand, convergence guarantees for the bundle method follow from specifying any choice of the parameter ρ_k. Given a choice of ρ_k, bounding the proximal gap is a classic, well-understand problem, independent from the details of the bundle method being used. Standard analysis [219] of the proximal gap shows the following bound for any minimizer x^*.

Lemma 4.5.3. *For any $x_k \in \mathbf{R}^n$, the proximal gap is lower bounded by*

$$\Delta_k \geq \begin{cases} \dfrac{1}{2\rho_k}\left(\dfrac{f(x_k) - f(x^*)}{\|x_k - x^*\|}\right)^2 & \text{if } f(x_k) - f(x^*) \leq \rho_k\|x_k - x^*\|^2 \\ f(x_k) - f(x^*) - \dfrac{\rho_k}{2}\|x_k - x^*\|^2 & \text{otherwise.} \end{cases} \tag{4.15}$$

Our ideal stepsize (4.11) is chosen to balance the two cases of this classic bound.

All of our analysis follows directly from applying these core lemmas. We bound the number of descent steps by combining Lemmas 4.5.1 and 4.5.3 to

give a recurrence relation describing the decrease in the objective gap. Then Lemmas 4.5.2 and 4.5.3 together allow us to bound the number of consecutive null steps between each of these descent steps, which can then be summed up to bound the total number of iterations required.

Proof of the Descent Step Lemma 4.5.1

Let $\bar{x}_{k+1} = \arg\min\{f(\cdot) + \frac{\rho_k}{2}\|\cdot - x_k\|^2\}$. From (4.5), we have

$$\begin{aligned} f_k(x_{k+1}) &\leq f_k(x_{k+1}) + \frac{\rho_k}{2}\|x_{k+1} - x_k\|^2 \\ &\leq f_k(\bar{x}_{k+1}) + \frac{\rho_k}{2}\|\bar{x}_{k+1} - x_k\|^2 \\ &\leq f(\bar{x}_{k+1}) + \frac{\rho_k}{2}\|\bar{x}_{k+1} - x_k\|^2 \,. \end{aligned}$$

Hence $f(x_k) - f_k(x_{k+1}) \geq \Delta_k$. Since we have assumed that iteration k was a descent step, this implies $(f(x_k) - f(x_{k+1}))/\beta \geq \Delta_k$. Concluding the proof.

Proof of the Null Step Lemma 4.5.2

Consider some descent step, at iteration k, followed by T consecutive null steps. Denote the proximal subproblem gap at iteration $k < t \leq k + T$ on the model f_t by

$$\widetilde{\Delta}_t := f(x_{k+1}) - \left(f_t(z_{t+1}) + \frac{\rho_{k+1}}{2}\|z_{t+1} - x_{k+1}\|^2\right).$$

Note every such null step t has the same stepsize $\rho_t = \rho_{k+1}$ and the same proximal center $x_t = x_{k+1}$. The core of this null step bound relies on the following recurrence showing $\widetilde{\Delta}_t$ decreases at each step

$$\widetilde{\Delta}_{t+1} \leq \widetilde{\Delta}_t - \frac{(1-\beta)^2\rho_{k+1}\widetilde{\Delta}_t^2}{2G_{k+1}^2} \,. \tag{4.16}$$

Before deriving this inequality, we show how it completes the proof of this lemma. After T consecutive null steps, the fact that $f_{k+T} \leq f$ ensures $\widetilde{\Delta}_{k+T} \geq \Delta_{k+T} = \Delta_{k+1}$. Thus, to bound T it suffices to bound the minimum iteration at which the reversed inequality hold. By solving the recurrence, see Lemma 4.5.4 with $\epsilon = \Delta_{k+1}$, we conclude the number of consecutive null steps is at most

$$T \leq \frac{2G_{k+1}^2}{(1-\beta)^2 \rho_{k+1} \Delta_{k+1}} .$$

Now all that remains is to derive the recurrence (4.16). Consider some null step $k < t \leq k+T$ in the sequence of consecutive null steps. We will use the following claim mulitiple times in the proof.

Claim 1. *The following inequalities hold true* $\|s_{t+1}\|^2 \leq 2\rho_{k+1}\tilde{\Delta}_t \leq G_{k+1}^2$.

Proof of the Claim. Due to the ρ_{k+1}-strongly convexity of the proximal subproblem $f_t(z) + \frac{\rho_{k+1}}{2}\|z - x_{k+1}\|^2$ and the fact that z_{t+1} is its unique minimizer, we derive

$$\begin{aligned}
\frac{\rho_{k+1}}{2}\|z_{t+1} - x_{k+1}\|^2 &\leq f_t(x_{k+1}) - \left(f_t(z_{t+1}) + \frac{\rho_{k+1}}{2}\|z_{t+1} - x_{k+1}\|^2\right) \\
&\leq \widetilde{\Delta}_t \\
&\leq \widetilde{\Delta}_{k+1} \leq \frac{1}{2\rho_k}\|g_{k+1}\|^2.
\end{aligned}$$

The last inequality follows by (4.6) since

$$\begin{aligned}
f_{k+1}(z_{k+2}) &\geq f(x_{k+1}) + \langle g_{k+1}, z_{k+2} - x_{k+1}\rangle \\
&\geq f(x_{k+1}) - \frac{1}{2}\left(\frac{\|g_{k+1}\|^2}{\rho_{k+1}} + \rho_{k+1}\|z_{k+2} - x_{k+1}\|^2\right).
\end{aligned}$$

□

Define the necessary lower bound on f_{t+1} given by (4.6) and (4.7) as

$$\tilde{f}_{t+1}(\cdot) := \max\left\{f_t(z_{t+1}) + \langle s_{t+1}, \cdot - z_{t+1}\rangle,\ f(z_{t+1}) + \langle g_{t+1}, \cdot - z_{t+1}\rangle\right\} \leq f_{t+1}(\cdot) .$$

Denote the result of a proximal step on $\tilde{f}_{t+1}$ by

$$y_{t+2} = \arg\min \left\{ \tilde{f}_{t+1}(\cdot) + \frac{\rho_{k+1}}{2} \| \cdot - x_{k+1} \|^2 \right\} .$$

A simple computation gives an explicit form for the minimizer of this problem

$$\begin{aligned} \theta_{t+1} &= \min \left\{ 1, \frac{\rho_{k+1} \left(f(z_{t+1}) - f_t(z_{t+1}) \right)}{\| g_{t+1} - s_{t+1} \|^2} \right\} \\ y_{t+2} &= x_{k+1} - \frac{1}{\rho_{k+1}} \left(\theta_{t+1} g_{t+1} + (1 - \theta_{t+1}) s_{t+1} \right) . \end{aligned} \tag{4.17}$$

Hence the objective of the proximal subproblem at iteration $t + 1$ is lower bounded by

$$\begin{aligned} & f_{t+1}(z_{t+2}) + \frac{\rho_{k+1}}{2} \| z_{t+2} - x_{k+1} \|^2 \\ & \geq \tilde{f}_{t+1}(y_{t+2}) + \frac{\rho_{k+1}}{2} \| y_{t+2} - x_{k+1} \|^2 \\ & \geq \theta_{t+1} \left(f(z_{t+1}) + \langle g_{t+1}, y_{t+2} - z_{t+1} \rangle \right) \\ & \qquad + (1 - \theta_{t+1}) \left(f_t(z_{t+1}) + \langle s_{t+1}, y_{t+2} - z_{t+1} \rangle \right) + \frac{\rho_{k+1}}{2} \| y_{t+2} - x_{k+1} \|^2 \\ & = f_t(z_{t+1}) + \theta_{t+1} \left(f(z_{t+1}) - f^t(z_{t+1}) \right) \\ & \qquad + \langle \theta_{t+1} g_{t+1} + (1 - \theta_{t+1}) s_{t+1}, y_{t+2} - z_{t+1} \rangle + \frac{\rho_{k+1}}{2} \| y_{t+2} - x_{k+1} \|^2 \\ & = f_t(z_{t+1}) + \theta_{t+1} \left(f(z_{t+1}) - f^t(z_{t+1}) \right) \\ & \qquad + \theta_{t+1}^2 \| g_{t+1} - s_{t+1} \|^2 / \rho_{k+1} + \frac{\rho_{k+1}}{2} \| z_{t+1} - x_{k+1} \|^2 , \end{aligned}$$

where the first inequality uses that $f_{t+1} \geq \tilde{f}_{t+1}$, the second inequality takes a convex combination of the two affine functions defining $\tilde{f}_{t+1}$, and the second equality uses the definition of y_{t+2}. Thus we have

$$\widetilde{\Delta}_{t+1} \leq \widetilde{\Delta}_t - \theta_{t+1} \left(f(z_{t+1}) - f_t(z_{t+1}) \right) + \theta_{t+1}^2 \| g_{t+1} - s_{t+1} \|^2 / \rho_{k+1} .$$

The amount of decrease guaranteed above can be lower bounded as follows

$$\begin{aligned}
&\theta_{t+1}\left(f(z_{t+1}) - f_t(z_{t+1})\right) + \theta_{t+1}^2 \|g_{t+1} - s_{t+1}\|^2/\rho_{k+1} \\
&\geq \min\left\{f(z_{t+1}) - f_t(z_{t+1}),\ \frac{2\rho_{k+1}(f(z_{t+1}) - f_t(z_{t+1}))^2}{\|g_{t+1} - s_{t+1}\|^2}\right\} \\
&\geq \min\left\{(1-\beta)\widetilde{\Delta}_t,\ \frac{2\rho_{k+1}(1-\beta)^2\widetilde{\Delta}_t^2}{\|g_{t+1} - s_{t+1}\|^2}\right\} \\
&\geq \min\left\{(1-\beta)\widetilde{\Delta}_t,\ \frac{\rho_{k+1}(1-\beta)^2\widetilde{\Delta}_t^2}{\|g_{t+1}\|^2 + \|s_{t+1}\|^2}\right\} \\
&\geq \min\left\{2\frac{\rho_{k+1}(1-\beta)\widetilde{\Delta}_t^2}{G_{k+1}^2},\ \frac{\rho_{k+1}(1-\beta)^2\widetilde{\Delta}_t^2}{2G_{k+1}^2}\right\} \\
&\geq \frac{\rho_{k+1}(1-\beta)^2\widetilde{\Delta}_t^2}{2G_{k+1}^2}
\end{aligned}$$

where the first inequality uses the definition of θ_{t+1} and drops a norm squared term, the second inequality uses the definition of a null step, and the fourth inequality uses Claim 1 and $\|g_{t+1}\|^2 \leq G_{k+1}^2$. This verifies (4.16) and completes the proof of our general bound.

For any M-Lipschitz objective, our specialized result follows from observing that $G_k \leq M$ as subgradients everywhere are uniformly bounded in norm by the Lipschitz constant. For any L-smooth objective, the following three inequalities hold for any null step t in the sequence of consecutive null steps following a descent step $k < t$:

$$\|g_{t+1}\| \leq \|g_{k+1}\| + L\|z_{t+1} - x_{k+1}\| \tag{4.18}$$

$$\|z_{t+1} - x_{k+1}\| \leq \|g_{k+1}\|/\rho_{k+1} \tag{4.19}$$

$$\|g_{k+1}\| \leq \sqrt{2(L + \rho_{k+1})\Delta_{k+1}}\,. \tag{4.20}$$

Before proving these three inequalities, we note that combined they give the

claimed bound as

$$\begin{aligned} G_{k+1} = \sup_t\{\|g_{t+1}\|\} &\le \sup_t\{\|g_{k+1}\| + L\|z_{t+1} - x_{k+1}\|\} \\ &\le (1 + L/\rho_{k+1})\|g_{k+1}\| \\ &\le (1 + L/\rho_{k+1})\sqrt{2(L + \rho_{k+1})\Delta_{k+1}} \end{aligned}$$

and thus $G_{k+1}^2 \le 2(L + \rho_{k+1})^3\Delta_{k+1}/\rho_{k+1}^2$. First (4.18) follows directly from the gradient being L-Lipschitz continuous. Second (4.19) follows from Claim 1. Third (4.20) follows from the L-smoothness of f and considering the full proximal subproblem $f(z) + \frac{\rho_{k+1}}{2}\|z - x_{k+1}\|^2$ since

$$\begin{aligned} \Delta_{k+1} &= f(x_{k+1}) - \min_z \left\{ f(z) + \frac{\rho_{k+1}}{2}\|z - x_{k+1}\|^2 \right\} \\ &\ge f(x_{k+1}) - \min_z \left\{ f(x_{k+1}) + \langle g_{k+1}, z - x_{k+1}\rangle + \frac{L + \rho_{k+1}}{2}\|z - x_{k+1}\|^2 \right\} \\ &= \frac{\|g_{k+1}\|^2}{2(L + \rho_{k+1})} . \end{aligned}$$

4.5.2 Proofs in Section 4.2

Proof of Theorem 4.2.2

For a constant stepsize $\rho_k = \rho$, we can simplify the lower bound (4.15) to only depend on x_k through a simple threshold on $f(x_k) - f^*$ as

$$\Delta_k \ge \begin{cases} \dfrac{1}{2\rho}\left(\dfrac{f(x_k) - f^*}{D}\right)^2 & \text{if } f(x_k) - f^* \le \rho D^2 \\ \dfrac{1}{2}\left(f(x_k) - f^*\right) & \text{otherwise.} \end{cases} \tag{4.21}$$

Combining this with Lemma 4.5.1 gives a recurrence relation describing the decrease in the objective gap $\delta_k = f(x_k) - f^*$ on any descent step k of

$$\delta_{k+1} \leq \begin{cases} \delta_k - \dfrac{\beta\delta_k^2}{2\rho D^2} & \text{if } \delta_k \leq \rho D^2 \\ (1-\beta/2)\delta_k & \text{if } \delta_k > \rho D^2 \end{cases}.$$

Our analysis of the bundle method then proceeds by considering these two cases separately. In each case, solving the given recurrence relation bounds the number of descent steps and applying Lemma 4.5.2 bounds the number of null steps.

Bounding steps with $\delta_k > \rho D^2$. First we show that the number of descent steps with $\delta_k > \rho D^2$ is bounded by

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f^*}{\rho D^2}\right)}{-\log(1-\beta/2)} \right\rceil_+ \tag{4.22}$$

and the number of null steps with $\delta_k > \rho D^2$ is at most

$$\frac{8M^2}{\beta(1-\beta)^2\rho^2 D^2}. \tag{4.23}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta/2)\delta_k$. This immediately bounds the number of descent steps by (4.22). Index the descent steps before a ρD^2-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than ρD^2. Define $k_0 = -1$. Then for each $i = 0 \ldots n-1$, $f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}\rho D^2$. It follows from (4.15) that $\Delta_{k_i+1} \geq (f(x_{k_i+1}) - f^*)/2 \geq (1-\beta/2)^{i-(n-1)}\rho D^2/2$. Plugging this into Lemma 4.5.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(n-1)-i} \frac{4M^2}{(1-\beta)^2\rho^2 D^2}.$$

Summing this over $i = 0 \dots n-1$ bounds the total number of null steps before a ρD^2-minimizer is found by (4.23) as

$$\sum_{i=0}^{n-1} (1-\beta/2)^{(n-1)-i} \frac{4M^2}{(1-\beta)^2\rho^2 D^2} \leq \frac{8M^2}{\beta(1-\beta)^2\rho^2 D^2} .$$

Bounding steps with $\rho D^2 \geq \delta_k > \epsilon$. Now we complete our proof of Theorem 4.2.2 by bounding the number of descent steps with $\rho D^2 \geq \delta_k > \epsilon$ by

$$\frac{2\rho D^2}{\beta\epsilon} \tag{4.24}$$

and the number of null steps with $\rho D^2 \geq \delta_k > \epsilon$ by

$$\frac{12\rho D^4 M^2}{(1-\beta)^2\epsilon^3} . \tag{4.25}$$

After the bundle method has passed objective value ρD^2, the recurrence relation becomes

$$\delta_{k+1} \leq \delta_k - \frac{\beta\delta_k^2}{2\rho D^2}.$$

Solving this recurrence with Lemma 4.5.4 implies $\delta_k > \epsilon$ holds for at most (4.24) descent stwhereeps. Then we can bound the number of null steps between these descent steps by noting (4.21) implies $\Delta_k \geq (f(x_k) - f^*)^2/2\rho D^2 \geq \epsilon^2/2\rho D^2$. Then Lemma 4.5.2 upper bounds the number of consecutive null steps by $4D^2M^2/(1-\beta)^2\epsilon^2$. Then multiplying this by our bound on the number of descent steps gives (4.25) as

$$\left(\frac{2\rho D^2}{\beta\epsilon} + 1\right) \frac{4D^2M^2}{(1-\beta)^2\epsilon^2} \leq \frac{12\rho D^4 M^2}{\beta(1-\beta)^2\epsilon^3} .$$

Proof of Theorem 4.2.3

Our bound on the number of descent steps comes directly from Theorem 4.2.2. Our claimed bound on the total number of null steps follows by multiply-

ing this by the constant bound on the number of consecutive null steps from Lemma 4.5.2.

Proof of Theorem 4.2.4

Assuming Hölder growth (4.4) holds and fixing $\rho_k = \rho$, the lower bound (4.15) simplifies to only depend on a simple threshold with $f(x_k) - f^*$ as

$$\Delta_k \geq \begin{cases} \dfrac{\mu^{2/p}(f(x_k) - f^*)^{2-2/p}}{2\rho} & \text{if } (f(x_k) - f^*)^{1-2/p} \leq \rho/\mu^{2/p} \\ \dfrac{1}{2}(f(x_k) - f^*) & \text{otherwise .} \end{cases} \tag{4.26}$$

From this, we arrive at a recurrence relation on the objective gap $\delta_k = f(x_k) - f^*$ decrease at each descent step k by plugging this lower bound into Lemma 4.5.1 of

$$\delta_{k+1} \leq \begin{cases} \delta_k - \dfrac{\beta\mu^{2/p}\delta_k^{2-2/p}}{2\rho} & \text{if } \delta_k^{1-2/p} \leq \rho/\mu^{2/p} \\ (1 - \beta/2)\delta_k & \text{if } \delta_k^{1-2/p} > \rho/\mu^{2/p} \text{ .} \end{cases}$$

Our analysis proceeds by considering the two cases of this recurrence and the three cases of $p > 2$, $p = 2$, and $1 \leq p < 2$ separately. In each case, solving the given recurrence relation bounds the number of descent steps and applying Lemma 4.5.2 bounds the number of null steps.

Given $p > 2$, bounding steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$. First we show that the number of descent steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is bounded by

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f^*}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{-\log(1-\beta/2)} \right\rceil_+ \tag{4.27}$$

and the number of null steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is at most

$$\frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \text{ .} \tag{4.28}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta/2)\delta_k$. This immediately bounds the number of descent steps by (4.27). Index the descent steps before a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than $(\rho/\mu^{2/p})^{1/(1-2/p)}$. Define $k_0 = -1$. Then for each $i = 0 \dots n-1$, $f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}(\rho/\mu^{2/p})^{1/(1-2/p)}$. It follows from (4.15) that

$$\Delta_{k_i+1} \geq (f(x_{k_i+1}) - f^*)/2 \geq (1-\beta/2)^{i-(n-1)} (\rho/\mu^{2/p})^{1/(1-2/p)}/2.$$

Plugging this into Lemma 4.5.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(n-1)-i} \frac{4M^2}{(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} .$$

Summing this over $i = 0 \dots n-1$ bounds the total number of null steps before a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer is found by (4.28) as

$$\sum_{i=0}^{n-1} (1-\beta/2)^{(n-1)-i} \frac{4M^2}{(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \leq \frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} .$$

Given $p > 2$, bounding steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$. Next we show that the total number of descent steps with$(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is bounded by

$$\frac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} \tag{4.29}$$

and the number of null steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is at most

$$\frac{12\rho M^2}{(1-2/p)\beta(1-\beta)^2\mu^{4/p}\epsilon^{3-4/p}} . \tag{4.30}$$

In this case, the recurrence relation on objective value decrease becomes

$$\delta_{k+1} \leq \delta_k - \frac{\beta\mu^{2/p}\delta_k^{2-2/p}}{2\rho}.$$

Applying Lemma 4.5.4 gives our bound on the number of descent steps with $\delta_k > \epsilon$ in (4.29). Plugging the lower bound $\Delta_k \geq \mu^{2/p}(f(x_k) - f^*)^{2-2/p}/2\rho \geq \mu^{2/p}\epsilon^{2-2/p}/2\rho$ into Lemma 4.5.2, the number of consecutive null steps after a descent step is at most

$$\frac{4M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} .$$

Then multiplying our limit on consecutive null steps by the number of descent steps between finding a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer and finding an ϵ-minimizer gives the bound (4.30) as

$$\left(\frac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + 1\right)\frac{4M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} \leq \frac{12\rho M^2}{(1-2/p)\beta(1-\beta)^2\mu^{4/p}\epsilon^{3-4/p}} .$$

Given $p = 2$, bounding steps with $\delta_k > \epsilon$. Here both cases of our recurrence relation have a similar form, and so we directly bound the total number of descent steps with $\delta_k > \epsilon$ by

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f^*}{\epsilon}\right)}{-\log(1-\beta\min\{\mu/2\rho, 1/2\})} \right\rceil \tag{4.31}$$

and the number of null steps with $\delta_k > \epsilon$ by

$$\frac{2M^2}{\beta(1-\beta)^2\min\{\mu/2\rho, 1/2\}\rho\epsilon} . \tag{4.32}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta\min\{\mu/2\rho, 1/2\})\delta_k$. This immediately bounds the number of descent steps by (4.31). Index the descent steps before an ϵ-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than ϵ. Define $k_0 = -1$. Then for each $i = 0 \ldots n-1$,

$$f(x_{k_i+1}) - f^* \geq (1-\beta\min\{\mu/2\rho, 1/2\})^{i-(n-1)}\epsilon .$$

It follows from (4.15) that $\Delta_{k_i+1} \geq (1-\beta\min\{\mu/2\rho, 1/2\})^{i-(n-1)}\,\epsilon/2$. Plugging this into Lemma 4.5.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1-\beta\min\{\mu/2\rho, 1/2\})^{(n-1)-i}\,\frac{2M^2}{(1-\beta)^2\rho\epsilon}\,.$$

Summing this over $i = 0 \ldots n-1$ bounds the total number of null steps before an ϵ-minimizer is found by

$$\sum_{i=0}^{n-1}(1-\beta\min\{\mu/2\rho, 1/2\})^{(n-1)-i}\,\frac{2M^2}{(1-\beta)^2\rho\epsilon} \leq \frac{2M^2}{\min\{\mu/2\rho, 1/2\}\beta(1-\beta)^2\rho\epsilon}\,.$$

Given $1 \leq p < 2$, bounding steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$. Then we show that the number of descent steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is bounded by

$$\frac{2\rho(f(x_0)-f^*)^{2/p-1}}{(1-2^{1-2/p})\beta\mu^{2/p}} \tag{4.33}$$

and the number of null steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is at most

$$\frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}}C \tag{4.34}$$

with $C = \max\left\{\frac{(f(x_0)-f^*)^{4/p-3}}{(\rho/\mu^{2/p})^{(4/p-3)/(1-2/p)}}, 1\right\}\min\left\{\frac{1}{1-2^{-|4/p-3|}}, \left\lceil\log_2\left(\frac{f(x_0)-f^*}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)\right\rceil\right\}$. Notice that since $p < 2$, the power $1-2/p$ of δ_k in the threshold condition of our recurrence is negative. In this case, the recurrence relation on objective value decrease becomes

$$\delta_{k+1} \leq \delta_k - \frac{\beta\mu^{2/p}\delta_k^{2-2/p}}{2\rho}\,.$$

As an intermediate step, for any $i \geq 0$, we first bound the number of descent and null steps with

$$2^{i+1}(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > 2^i(\rho/\mu^{2/p})^{1/(1-2/p)}\,.$$

Since descent steps decreases the objective gap by at least $\beta\mu^{2/p}\delta_k^{2-2/p}/2\rho$, there are at most

$$\frac{2\rho(2^i(\rho/\mu^{2/p})^{1/(1-2/p)})^{2/p-1}}{\beta\mu^{2/p}} = \frac{2^{(2/p-1)i+1}}{\beta}$$

descent steps in this interval. Further, noting that in this interval

$$\Delta_k \geq \frac{\mu^{2/p}(2^i(\rho/\mu^{2/p})^{1/(1-2/p)})^{2-2/p}}{2\rho} = 2^{(2-2/p)i-1}(\rho/\mu^{2/p})^{1/(1-2/p)},$$

we can bound the number of consecutive null steps following any of these descent steps via Lemma 4.5.2. Hence there are at most

$$\frac{2^{(4/p-3)i+3}M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}}$$

null steps in this interval.

The bundle method halves its objective value at most $N = \lceil \log_2((f(x_0) - f^*)/(\rho/\mu^{2/p})^{1/(1-2/p)}) \rceil$ times before an $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer is found. Then summing up these bounds on the descent and null steps in each interval limits the number of descent steps needed to find a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer by (4.33) as

$$\sum_{i=0}^{N-1} \frac{2^{(2/p-1)i+1}}{\beta} \leq \frac{2}{\beta}\sum_{i=0}^{N-1} 2^{(2/p-1)i} \leq \frac{2^{(2/p-1)(N-1)+1}}{(1-2^{1-2/p})\beta} \leq \frac{2\rho(f(x_0)-f^*)^{2/p-1}}{(1-2^{1-2/p})\beta\mu^{2/p}}$$

and similarly, the number of null steps needed by (4.34) as

$$\begin{aligned}
&\sum_{i=0}^{N-1} \frac{2^{(4/p-3)i+3}M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \\
&\leq \frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \sum_{i=0}^{N-1} 2^{(4/p-3)i} \\
&\leq \frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \\
&\qquad \max\left\{\frac{(f(x_0)-f^*)^{4/p-3}}{(\rho/\mu^{2/p})^{(4/p-3)/(1-2/p)}}, 1\right\} \min\left\{\frac{1}{1-2^{-|4/p-3|}}, N\right\}
\end{aligned}$$

where the last inequality bounds the geometric sum regardless of the sign of the exponent $4/p - 3$.

Given $1 \leq p < 2$, bounding steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$. Finally, we show that the number of descent steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is bounded

by

$$\left\lceil \frac{\log\left(\frac{(\rho/\mu^{2/p})^{1/(1-2/p)}}{\epsilon}\right)}{-\log(1-\beta/2)} \right\rceil \tag{4.35}$$

and the number of null steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is at most

$$\frac{4M^2}{\beta(1-\beta)^2\rho\epsilon}\,. \tag{4.36}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta/2)\delta_k$. This immediately bounds the number of descent steps by (4.35). Index the descent steps after a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer but before an ϵ-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than ϵ. Then for each $i = 0 \ldots n-1$, $f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}\epsilon$. It follows from (4.15) that $\Delta_{k_i+1} \geq (f(x_{k_i+1}) - f^*)/2 \geq (1-\beta/2)^{i-(n-1)}\,\epsilon/2$. Plugging this into Lemma 4.5.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(n-1)-i}\,\frac{2M^2}{(1-\beta)^2\rho\epsilon}\,.$$

Summing this over $i = 0 \ldots n-1$ bounds the additional number of null steps before an ϵ-minimizer is found by (4.36) as

$$\sum_{i=0}^{n-1} (1-\beta/2)^{(n-1)-i}\,\frac{2M^2}{(1-\beta)^2\rho\epsilon} \leq \frac{4M^2}{\beta(1-\beta)^2\rho\epsilon}\,.$$

Proof of Theorem 4.2.5

Our bound on the number of descent steps comes directly from Theorem 4.2.4. Our claimed bound on the total number of null steps follows by multiplying this by the constant bound on the number of consecutive null steps from Lemma 4.5.2.

Proof of Theorem 4.2.6

Combining the lower bound $\Delta_k \geq \frac{1}{2}(f(x_k) - f^*)$ with Lemma 4.5.1 shows linear decrease in the objective every descent step

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\beta}{2}\right)(f(x_k - f^*).$$

Our bound on the number of descent steps follows immediately from this. Combining the lower bound $\Delta_k \geq \frac{1}{2}(f(x_k) - f^*)$ with Lemma 4.5.2 shows that at most

$$\frac{2M^2D^2}{(1-\beta)^2(f(x_{k+1}) - f^*)^2}$$

null steps occur between each descent step. Denote the sequence of descent steps taken by the bundle method by $k_1, k_2, k_3 \ldots$ and as a base case define $k_0 = -1$. Let k_n be the first descent step finding an ϵ-minimizer, which must have $n \leq \lceil \log_{(1-\beta/2)}(\frac{\epsilon}{f(x_0)-f^*}) \rceil_+$. From our linear decrease condition, we know for any $i = 0, 1, 2, 3, \ldots n-1$

$$f(x_{k_i+1}) - f^* \geq (1 - \beta/2)^{i-(n-1)}\,\epsilon$$

and from our null step bound, we know for any $i = 0, 1, 2, \ldots n-1$

$$k_{i+1} - k_i - 1 \leq \frac{2M^2D^2}{(1-\beta)^2(f(x_{k_i+1}) - f^*)^2} \leq (1 - \beta/2)^{2(i-(n-1))}\frac{2M^2D^2}{(1-\beta)^2\epsilon^2}\,.$$

Then summing up our null step bounds ensures

$$k_n - n \leq \sum_{i=1}^{n}(1 - \beta/2)^{2(i-1-(n-1))}\frac{2M^2D^2}{(1-\beta)^2\epsilon^2}.$$

Bounding this geometric series shows us that the bundle method finds an ϵ-minimizer with the number of null steps bounded by

$$\left(\frac{1}{1-(1-\beta/2)^2}\right)\frac{2M^2D^2}{(1-\beta)^2\epsilon^2}\,.$$

Proof of Theorem 4.2.7

Our bound on the number of descent steps follows from Theorem 4.2.6. Our proof of the null step bound follows the same approach as Theorem 4.2.6 with only minor differences. Applying Lemma 4.5.2 with our stepsize choice (4.13) bounds the number of consecutive null steps after some descent step k by

$$\frac{2M^2}{(1-\beta)^2\mu^{2/p}(f(x_{k+1})-f^*)^{2-2/p}}.$$

Denote the descent steps $-1 = k_0 < k_1 < k_2 < \dots$ and suppose the x_{k_n+1} is the first ϵ-minimizer. Then

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(2-2/p)(i-(n-1))}\frac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}}$$

since $f(x_{k_i+1}) - f^* \geq \left(1-\frac{\beta}{2}\right)^{i-(n-1)}\epsilon$. Summing this up gives

$$k_n - n \leq \sum_{i=1}^{n}(1-\beta/2)^{(2-2/p)(i-1-(n-1))}\frac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}}.$$

When $p > 1$, this geometric series shows us that the bundle method finds an ϵ-minimizer with the number of null steps bounded by

$$\left(\frac{1}{1-(1-\beta/2)^{2-2/p}}\right)\frac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}}.$$

When $p = 1$, we have a constant upper bound on the number of null steps following a descent step. Hence the number of null steps is bounded by

$$\frac{2M^2}{(1-\beta)^2\mu^2}\left\lceil\frac{\log\left(\frac{f(x_0)-f^*}{\epsilon}\right)}{-\log(1-\beta/2)}\right\rceil.$$

4.5.3 Proofs in Section 4.3

Proof of Theorem 4.3.1

Let $\delta_k = \min_{j\in\{0,\dots,J-1\}}\{f(x_k^{(j)}) - f^*\}$ denote the lowest objective gap among all of our J instances of the bundle method after they have taken k synchronous steps. Then the core of our convergence proof is bounding the number of iterations where this lowest objective gap is in the interval

$$(1-\beta/2)^{-n}\epsilon \le \delta_k \le (1-\beta/2)^{-(n+1)}\epsilon\ .$$

for any integer $0 \le n < N := \left\lceil \frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)} \right\rceil$. Within this interval, we focus on the instance

$$j = \left\lceil \log_2\left(\frac{\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{1-2/p}}{4\bar{\rho}}\right)\right\rceil.$$

This instance of the bundle method's constant stepsize $\rho^{(j)} = 2^j\bar{\rho}$ approximates the stepsize (4.13) as

$$\frac{1}{4}\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{1-2/p} \le \rho^{(j)} \le \frac{1}{2}\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{1-2/p}\ .$$

Then (4.26) bounds this method's proximal gap before an $(1-\beta/2)^{-n}\epsilon$-minimizer is found by

$$\Delta_k^{(j)} \ge \frac{1}{2}(f(x_k^{(j)}) - f^*) \ge (1-\beta/2)^{-n}\epsilon/2\ .$$

Letting $\delta_k^{(j)} = f(x_k^{(j)}) - f^*$, each descent step k improves method j's objective gap according to the recurrence $\delta_{k+1}^{(j)} \le \min\{(1-\beta/2)\delta_k^{(j)}, \delta_k\}$ where the first term in the minimum comes from Lemma 4.5.1 and the second term comes from method j taking any further improvement from the other bundle methods. By assumption, we have $\delta_k \le (1-\beta/2)^{-(n+1)}\epsilon$, and so after one descent step $k' > k$ we must have $\delta_{k'+1}^{(j)} \le (1-\beta/2)^{-(n+1)}\epsilon$. Thus after a second descent step $k'' > k'$, our intermediate target accuracy is met as $\delta_{k''+1} \le \delta_{k''+1}^{(j)} \le (1-\beta/2)^{-n}\epsilon$.

Applying Lemma 4.5.2 bounds the number of null steps between descent steps by

$$\frac{2M^2}{(1-\beta)^2\rho^{(j)}\Delta_{k+1}^{(j)}} \leq \frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}}.$$

Hence the total number of steps before $\delta_k^{(j)} < 2^n\epsilon$ (and consequently $\delta_k < 2^n\epsilon$) is at most

$$2\left(\frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}} + 1\right).$$

Summing over this bound completes our proof. When $p > 1$, this gives

$$\begin{aligned}
&\sum_{n=0}^{N-1} 2\left(\frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}} + 1\right) \\
&= 2\sum_{n=0}^{N-1} \frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}} + 2\left\lceil \frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)} \right\rceil \\
&\leq \left(\frac{2}{1-(1-\beta/2)^{2-2/p}}\right) \frac{16M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} + 2\left\lceil \frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)} \right\rceil.
\end{aligned}$$

When $p = 1$, the number of steps in each of our intervals is constant. Consequently, the total number of iterations before an ϵ minimizer is found is at most

$$\sum_{n=0}^{N-1} 2\left(\frac{16M^2}{(1-\beta)^2\mu^2} + 1\right) = 2\left(\frac{16M^2}{(1-\beta)^2\mu^2} + 1\right)\left\lceil \frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)} \right\rceil.$$

4.5.4 Auxiliary lemmas

Throughout our analysis, we frequently encounter recurrence relations of the form $\delta_{k+1} \leq \delta_k - \alpha\delta_k^q$ for some $\alpha > 0$ and $q > 1$. The follow lemma bounds the number of steps of such a recurrence to reach a desired level of accuracy $\delta_k \leq \epsilon$.

Lemma 4.5.4. *For any $\epsilon > 0$, the recurrence $\delta_{k+1} \leq \delta_k - \alpha\delta_k^q$ has $\delta_k \leq \epsilon$ satisfied by some*

$$k \leq \left\lceil \frac{1}{(q-1)\alpha\epsilon^{q-1}} \right\rceil.$$

Proof. It suffices to show the following upper bound on δ_k as a function of k

$$\delta_k \leq \left(\frac{1}{(q-1)\alpha k}\right)^{1/(q-1)}.$$

First we show this bound holds with $k = 1$. This follows as

$$\delta_1 \leq \delta_0 - \alpha\delta_0^q \leq \max_{\delta \in \mathbf{R}} \{\delta - \alpha\delta^q\} \leq \left(\frac{1}{q\alpha}\right)^{1/(q-1)}.$$

Then we complete our proof by induction using the following *weighted arithmetic-geometric mean (AM-GM) inequality*, which ensures for any $a, \alpha, b, \beta > 0$ we have $a^\alpha b^\beta \leq \left(\frac{\alpha a + \beta b}{\alpha + \beta}\right)^{\alpha+\beta}$. This implies that for any $k \geq 1$, $(k - (q-1)^{-1})(k+1)^{1/(q-1)} \leq k^{q/(q-1)}$ by taking $a = k - (q-1)^{-1}$, $\alpha = 1$, $b = k+1$, $\beta = 1/(q-1)$. By expanding the recurrence at $k+1$ and applying this inequality we get

$$\begin{aligned}
\delta_{k+1} \leq \delta_k - \alpha\delta_k^q &\leq \left(\frac{1}{(q-1)\alpha k}\right)^{1/(q-1)} - \alpha\left(\frac{1}{(q-1)\alpha k}\right)^{q/(q-1)} \\
&= \left(\frac{1}{(q-1)\alpha}\right)^{1/(q-1)} \left(\frac{k}{k^{q/(q-1)}} - \frac{1}{(q-1)k^{q/(q-1)}}\right) \\
&= \left(\frac{1}{(q-1)\alpha}\right)^{1/(q-1)} \frac{k - (q-1)^{-1}}{k^{q/(q-1)}} \\
&\leq \left(\frac{1}{(q-1)\alpha(k+1)}\right)^{1/(q-1)}.
\end{aligned}$$

Proving the result. □

5

COMPOSITE OPTIMIZATION FOR LOW-RANK MATRIX RECOVERY

"I wish there was a way to know you are in the good old days,
before you've actually left them."

— "Andrew Bernard" in the finale of *The Office*

5.1 Introduction

Recovering a low-rank matrix from noisy linear measurements has become an increasingly central task in data science. Important and well-studied examples include phase retrieval [223, 39, 167], blind deconvolution [8, 153, 160, 233], matrix completion [35, 64, 228], covariance matrix estimation [52, 155], and robust principal component analysis [45, 38]. Optimization-based approaches for low-rank matrix recovery naturally lead to nonconvex formulations, which are NP hard in general. To overcome this issue, in the last two decades researchers have developed convex relaxations that succeed with high probability under appropriate statistical assumptions. Convex techniques, however, have a well-documented limitation: the parameter space describing the relaxations is usually much larger than that of the target problem. Consequently, standard algorithms applied on convex relaxations may not scale well to the large problems. Consequently, there has been a renewed interest in directly optimizing nonconvex formulations with iterative methods within the original parameter space of the problem. Aside from a few notable exceptions on specific problems [106, 23, 103], most algorithms of this type proceed in two-stages. The first

stage—*initialization*—yields a rough estimate of an optimal solution, often using spectral techniques. The second stage—*local refinement*—uses a local search algorithm that rapidly converges to an optimal solution, when initialized at the output of the initialization stage.

This chapter focuses on developing provable low-rank matrix recovery algorithms based on nonconvex problem formulations. We focus primarily on local refinement and describe a set of unifying sufficient conditions leading to rapid local convergence of iterative methods. In contrast to the current literature on the topic, which typically relies on smooth problem formulations and gradient-based methods, our primary focus is on *nonsmooth formulations* that exhibit sharp growth away from the solution set. Such formulations are well-known in the nonlinear programming community to be amenable to rapidly convergent local-search algorithms. Along the way, we will observe an apparent benefit of nonsmooth formulations over their smooth counterparts. All nonsmooth formulations analyzed in this chapter are "well-conditioned," resulting in fast "out-of-the-box" convergence guarantees. In contrast, standard smooth formulations for the same recovery tasks can be poorly conditioned, in the sense that classical convergence guarantees of nonlinear programming are overly pessimistic. Overcoming the poor conditioning typically requires nuanced problem and algorithmic specific analysis (e.g. [233, 51, 167, 184, 50]), which nonsmooth formulations manage to avoid for the problems considered here.

Setting the stage, consider a rank r matrix $M_\sharp \in \mathbf{R}^{d_1 \times d_2}$ and a linear map $\mathcal{A}\colon \mathbf{R}^{d_1 \times d_2} \to \mathbf{R}^m$ from the space of matrices to the space of measurements. The goal of low-rank matrix recovery is to recover $M_\sharp$ from the image vector $b = \mathcal{A}(M_\sharp)$, possibly corrupted by noise. Typical nonconvex approaches proceed by

choosing some penalty function $h(\cdot)$ with which to measure the residual $\mathcal{A}(M) - b$ for a trial solution M. Then, in the case that $M_\sharp$ is symmetric and positive semidefinite, one may focus on the formulation

$$\min_{X \in \mathbf{R}^{d \times r}} \; f(X) := h\left(\mathcal{A}(XX^\top) - b\right) \qquad \text{subject to } X \in \mathcal{D}, \tag{5.1}$$

or when $M_\sharp$ is rectangular, one may instead use the formulation

$$\min_{X \in \mathbf{R}^{d_1 \times r},\; Y \in \mathbf{R}^{r \times d_2}} \; f(X, Y) := h\left(\mathcal{A}(XY) - b\right) \qquad \text{subject to } (X, Y) \in \mathcal{D}. \tag{5.2}$$

Here, $\mathcal{D}$ is a convex set that incorporates prior knowledge about $M_\sharp$ and is often used to enforce favorable structure on the decision variables. The penalty h is chosen specifically to penalize measurement misfit and/or enforce structure on the residual errors.

Algorithms and conditioning for smooth formulations

Most widely-used penalties $h(\cdot)$ are smooth and convex. Indeed, the *squared* ℓ_2-norm $h(z) = \frac{1}{2}\|z\|_2^2$ is ubiquitous in this context. With such penalties, problems (5.1) and (5.2) are smooth and thus are amenable to gradient-based methods. The linear rate of convergence of gradient descent is governed by the "local condition number" of f. Indeed, if the estimate, $\mu I \preceq \nabla^2 f(X) \preceq LI$, holds for all X in a neighborhood of the solution set, then gradient descent converges to the solution set at the linear rate $1 - \mu/L$. It is known that for several widely-studied problems including phase retrieval, blind deconvolution, and matrix completion, the ratio μ/L scales inversely with the problem dimension. Consequently, generic nonlinear programming guarantees yield efficiency estimates that are far too pessimistic. Instead, near-dimension independent guarantees can be obtained by arguing that $\nabla^2 f$ is well conditioned along the "relevant" directions or

that $\nabla^2 f$ is well-conditioned within a restricted region of space that the iterates never escape (e.g. [233, 167, 184]). Techniques of this type have been elegantly and successfully used over the past few years to obtain algorithms with near-optimal sample complexity. One byproduct of such techniques, however, is that the underlying arguments are finely tailored to each particular problem and algorithm at hand. We refer the reader to the recent surveys [54] for details.

Algorithms and conditioning for nonsmooth formulations

The goal of our work is to justify the following principle:

> Statistical assumptions for common recovery problems guarantee that (5.1) and (5.2) *are well-conditioned* when h is an appropriate *nonsmooth convex penalty*.

To explain what we mean by "good conditioning," let us treat (5.1) and (5.2) within the broader *convex composite* problem class:

$$\min_{x \in \mathcal{X}} \; f(x) := h(F(x)), \tag{5.3}$$

where $F(\cdot)$ is a smooth map on the space of matrices and $\mathcal{X}$ is a closed convex set. Indeed, in the symmetric and positive semidefinite case, we identify x with matrices X and define $F(X) = \mathcal{A}(XX^\top) - b$, while in the asymmetric case, we identify x with pairs of matrices (X, Y) and define $F(X, Y) = \mathcal{A}(XY) - b$. Though compositional problems (5.3) have been well-studied in nonlinear programming [28, 31, 98], their computational promise in data science has only begun recently to emerge. For example, the papers [90, 66, 83] discuss stochastic and inexact algorithms on composite problems, while the papers [89, 70], [47],

and [154] investigate applications to phase retrieval, blind deconvolution, and matrix sensing, respectively.

A number of algorithms are available for problems of the form (5.3), and hence for (5.1) and (5.2). Two most notable ones are the projected subgradient method [68, 109]

$$x_{t+1} = \text{proj}_{\mathcal{X}}(x_t - \alpha_t v_t) \qquad \text{with} \qquad v_t \in \partial f(x_t),$$

and the prox-linear algorithm [28, 152, 82]

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathcal{X}} \; h\Big(F(x_t) + \nabla F(x_t)(x - x_t)\Big) + \frac{\beta}{2}\|x - x_t\|_2^2.$$

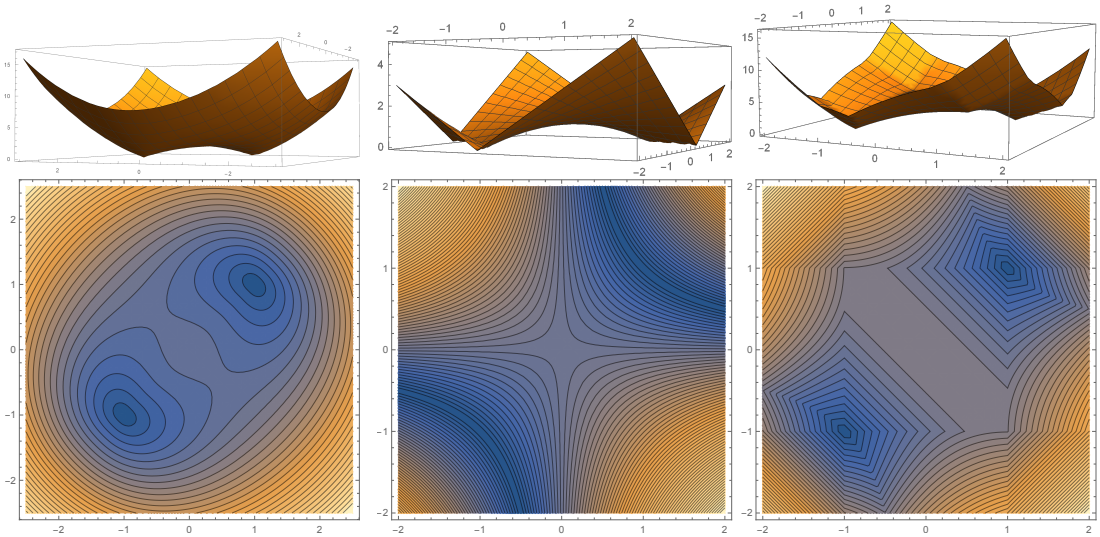
Notice that each iteration of the subgradient method is relatively cheap, requiring access only to the subgradients of f and the nearest-point projection onto $\mathcal{X}$. The prox-linear method in contrast requires solving a strongly convex problem in each iteration. That being said, the prox-linear method has much stronger convergence guarantees than the subgradient method, as we will review shortly.

The local convergence guarantees of both methods are straightforward to describe, and underlie what we mean by "good conditioning". Define $\mathcal{X}^* := \arg\min_{\mathcal{X}} f$, and for any $x \in \mathcal{X}$ define the convex model $f_x(y) = h(F(x) + \nabla F(x)(y - x))$. Suppose there exist constants $\rho, \mu > 0$ satisfying the two properties:

- **(approximation)** $|f(y) - f_x(y)| \le \frac{\rho}{2}\|y - x\|_2^2$ for all $x, y \in \mathcal{X}$,
- **(sharpness)** $f(x) - \inf f \ge \mu \cdot \text{dist}(x, \mathcal{X}^*)$ for all $x \in \mathcal{X}$.

The approximation and sharpness properties have intuitive meanings. The former says that the nonconvex function $f(y)$ is well approximated by the convex model $f_x(y)$, with quality that degrades quadratically as y deviates from x. In

particular, this property guarantees that the quadratically perturbed function $x \mapsto f(x) + \frac{\rho}{2}\|x\|_2^2$ is convex on $\mathcal{X}$. Yet another consequence of the approximation property is that the epigraph of f admits a supporting concave quadratic with amplitude ρ at each of its points. Sharpness, in turn, asserts that f must grow at least linearly as x moves away from the solution set. In other words, the function values should robustly distinguish between optimal and suboptimal solutions. In statistical contexts, one can interpret sharpness as strong identifiability of the statistical model. The three figures below illustrate the approximation and sharpness properties for idealized objectives in phase retrieval, blind deconvolution, and robust PCA problems.



(a) $f(x) = \mathbb{E}|(a^\top x)^2 - (a^\top \mathbf{1})^2|$ (phase retrieval)

(b) $f(x, y) = |xy - 1|$ (blind deconvolution)

(c) $f(x) = \|xx^\top - \mathbf{1}\mathbf{1}^\top\|_1$ (robust PCA)

Approximation and sharpness, taken together, guarantee rapid convergence of numerical methods when initialized within the tube:

$$\mathcal{T} = \left\{x \in \mathcal{X} : \operatorname{dist}(x, \mathcal{X}^*) \leq \frac{\mu}{\rho}\right\}.$$

For common low-rank recovery problems, $\mathcal{T}$ has an intuitive interpretation: it consists of those matrices that are within constant relative error of the solution.

We note that standard spectral initialization techniques, in turn, can generate such matrices with nearly optimal sample complexity. We will cover an example of these techniques in Chapter 6. We also refer the interested reader to the survey [54], and references therein, for details.

Guiding strategy. The following is the guiding algorithmic principle of this work:

> When initialized at $x_0 \in \mathcal{T}$, the prox-linear algorithm converges quadratically to the solution set $\mathcal{X}^*$; the subgradient method, in turn, converges linearly with a rate governed by ratio $\frac{\mu}{L} \in (0, 1)$, where L is the Lipschitz constant of f on $\mathcal{T}$.[1]

In light of this observation, our strategy can be succinctly summarized as follows. We will show that for a variety of low-rank recovery problems, the parameters $\mu, L, \rho > 0$ (or variants) are dimension independent under standard statistical assumptions. Consequently, the formulations (5.1) and (5.2) are "well-conditioned", and subgradient and prox-linear methods converge rapidly when initialized within constant relative error of the optimal solution.

Good conditioning via the Restricted Isometry Property

We begin verifying our thesis by showing that the composite problems, (5.1) and (5.2), are well-conditioned under the following Restricted Isometry Prop-

[1] Both the parameters α_t and β must be properly chosen for these guarantees to take hold.

erty (RIP): there exists a norm $|||\cdot|||$ and numerical constants $\kappa_1, \kappa_2 > 0$ so that

$$\kappa_1 \|W\|_F \leq |||\mathcal{A}(W)||| \leq \kappa_2 \|W\|_F, \tag{5.4}$$

for all matrices $W \in \mathbf{R}^{d_1 \times d_2}$ of rank at most $2r$. We argue that under RIP, the *nonsmooth* norm $h = |||\cdot|||$ is a natural penalty function to use. Indeed, as we will show, the composite loss $h(F(x))$ in the symmetric setting admits constants μ, ρ, L that depend only on the RIP parameters and the extremal singular values of $M_\sharp$:

$$\mu = 0.9\kappa_1 \sqrt{\sigma_r(M_\sharp)}, \qquad \rho = \kappa_2, \qquad L = 0.9\kappa_1 \sqrt{\sigma_r(M_\sharp)} + 2\kappa_2 \sqrt{\sigma_1(M_\sharp)}.$$

In particular, the initialization ratio scales as $\frac{\mu}{\rho} \asymp \frac{\kappa_1}{\kappa_2} \sqrt{\sigma_r(M_\sharp)}$ and the condition number scales as $\frac{L}{\mu} \asymp 1 + \frac{\kappa_2}{\kappa_1} \sqrt{\frac{\sigma_1(M_\sharp)}{\sigma_r(M_\sharp)}}$. Consequently, the rapid local convergence guarantees previously described immediately take-hold. The asymmetric setting is slightly more nuanced since the objective function is sharp only on bounded sets. Nonetheless, it can be analyzed in a similar way leading to analogous rapid convergence guarantees. Incidentally, we show that the prox-linear method converges rapidly without any modification; this is in contrast to smooth methods, which typically require incorporating an auxiliary regularization term into the objective (e.g. [233]). We note that similar results in the symmetric setting were independently obtained in the complimentary work [154], albeit with a looser estimate of L; the two treatments of the asymmetric setting are distinct, however.[2]

After establishing basic properties of the composite loss, we turn our attention to verifying RIP in several concrete scenarios. We note that the seminal works [209, 40] showed that if $\mathcal{A}(\cdot)$ arises from a Gaussian ensemble, then in

[2]The authors of [154] provide a bound on L that scales with the Frobenius norm $\sqrt{\|M_\sharp\|_F}$. We instead derive a sharper bound that scales as $\sqrt{\|M_\sharp\|_{\text{op}}}$. As a byproduct, the linear rate of convergence for the subgradient method scales only with the condition number $\sigma_1(M_\sharp)/\sigma_r(M_\sharp)$ instead of $\|M_\sharp\|_F/\sigma_r(M_\sharp)$.

the regime $m \gtrsim r(d_1 + d_2)$ RIP holds with high probability for the scaled ℓ_2 norm $|||z||| = m^{-1/2}||z||_2$. More generally when $\mathcal{A}$ is highly structured, RIP may be most naturally measured in a non-Euclidean norm. For example, RIP with respect to the scaled ℓ_1 norm $|||z||| = m^{-1}||z||_1$ holds for phase retrieval [94, 89], blind deconvolution [47], and quadratic sensing [52]; in contrast, RIP relative to the scaled ℓ_2 norm fails for all three problems. In particular, specializing our results to the aforementioned recovery tasks yields solution methodologies with best known sample and computational complexity guarantees. Notice that while one may "smooth-out" the ℓ_2 norm by squaring it, we argue that it may be more natural to optimize the ℓ_1 norm directly as a nonsmooth penalty. Moreover, we show that ℓ_1 penalization enables exact recovery even if a constant fraction of measurements is corrupted by outliers.

Beyond RIP: matrix completion and robust PCA

The RIP assumption provides a nice vantage point for analyzing the problem parameters $\mu, \rho, L > 0$. There are, however, a number of important problems, which do not satisfy RIP. Nonetheless, the general paradigm based on the interplay of sharpness and approximation is still powerful. We consider two such settings, matrix completion and robust principal component analysis (PCA), leveraging some intermediate results from [53].

The goal of the matrix completion problem [35] is to recover a low rank matrix $M_\sharp$ from its partially observed entries. We focus on the formulation

$$\arg\min_{X \in \mathcal{X}} \; f(X) = \|\Pi_\Omega(XX^\top) - \Pi_\Omega(M_\sharp)\|_2,$$

where Π_Ω is the projection onto the index set of observed entries Ω and

$$\mathcal{X} = \left\{ X \in \mathbb{R}^{d \times r} : \|X\|_{2,\infty} \leq \sqrt{\frac{\nu r \|M_\sharp\|_{\mathrm{op}}}{d}} \right\}$$

is the set of incoherent matrices.[3] To analyze the conditioning of this formulation, we assume that the indices in Ω are chosen as i.i.d. Bernoulli with parameter $p \in (0,1)$ and that all nonzero singular values of $M_\sharp$ are equal to one. Using results of [53], we quickly deduce sharpness with high probability. The error in approximation, however, takes the following nonstandard form. In the regime $p \geq \frac{c}{\epsilon^2}(\frac{\nu^2 r^2}{d} + \frac{\log d}{d})$ for some constants $c > 0$ and $\epsilon \in (0,1)$, the estimate holds with high probability:

$$|f(Y) - f_X(Y)| \leq \sqrt{1+\epsilon}\|Y - X\|_2^2 + \sqrt{\epsilon}\|X - Y\|_F \qquad \text{for all } X, Y \in \mathcal{X}.$$

The following modification of the prox-linear method therefore arises naturally:

$$X_{k+1} = \operatorname*{arg\,min}_{X \in \mathcal{X}} \; f_{X_k}(X) + \sqrt{1+\epsilon}\|X - X_k\|_F^2 + \sqrt{\epsilon}\|X - X_k\|_F.$$

We show that subgradient methods and the prox-linear method, thus modified, both converge at a dimension independent linear rate when initialized near the solution. Namely, as long as ϵ and $\mathrm{dist}(X_0, \mathcal{X}^*)$ are below some constant thresholds, both the subgradient and the modified prox-linear methods converge linearly with high probability:

$$\mathrm{dist}(X_k, \mathcal{X}^*) \lesssim \begin{cases} \left(1 - \frac{c}{\nu r}\right)^{k/2} & \text{subgradient} \\ 2^{-k} & \text{prox-linear} \end{cases}.$$

Here $c > 0$ is a numerical constant. Notice that the prox-linear method enjoys a much faster rate of convergence that is independent of any unknown constants

[3] Incoherence is necessary for recovery, e.g, $M_\sharp = e_1 e_1^\top$ cannot be recovered in nontrivial settings [35].

or problem parameters—an observation fully supported by our numerical experiments.

As the final example, we consider the problem of robust PCA [38, 45], which aims to decompose a given matrix W into a sum of a low-rank and a sparse matrix. We consider two different problem formulations:

$$\min_{(X,S)\in\mathcal{D}_1} F((X,S)) = \|XX^\top + S - W\|_F, \tag{5.5}$$

and

$$\min_{X\in\mathcal{D}_2} f(X) = \|XX^\top - W\|_1, \tag{5.6}$$

where $\mathcal{D}_1$ and $\mathcal{D}_2$ are appropriately defined convex regions. Under standard incoherence assumptions, we show that the formulation (5.5) is well-conditioned, and therefore subgradient and prox-linear methods are applicable. Still, formulation (5.5) has a major drawback in that one must know properties of the optimal sparse matrix $S_\sharp$ in order to define the constraint set $\mathcal{D}_1$, in order to ensure good conditioning. Consequently, we analyze formulation (5.6) as a more practical alternative.

The analysis of (5.6) is more challenging than that of (5.5). Indeed, it appears that we must replace the Frobenius norm $\|X\|_F$ in the approximation/sharpness conditions with the sum of the row norms $\|X\|_{2,1}$. With this set-up, we verify the convex approximation property in general:

$$|f(Y) - f_X(Y)| \le \|Y - X\|_{2,1}^2 \qquad \text{for all } X, Y$$

and sharpness only when $r = 1$. We conjecture, however, that an analogous sharpness bound holds for all r. It is easy to see that the quadratic convergence guarantees for the prox-linear method do not rely on the Euclidean nature of the norm, and the algorithm becomes applicable. To the best of our knowledge,

it is not yet known how to adapt linearly convergent subgradient methods to the non-Euclidean setting.

Robust recovery with sparse outliers and dense noise

The aforementioned guarantees lead to exact recovery of $M_\sharp$ under noiseless or sparsely corrupted measurements b. A more realistic noise model allows for further corruption by a dense noise vector e of small norm. Exact recovery is no longer possible with such errors. Instead, we should only expect to recover $M_\sharp$ up to a tolerance proportional to the size of e. Indeed, we show that appropriately modified subgradient and prox-linear algorithms converge linearly and quadratically, respectively, up to the tolerance $\delta = O(|||e|||/\mu)$ for an appropriate norm $|||\cdot|||$. Finally, we discuss in detail the case of recovering a low rank PSD matrix $M_\sharp$ from the corrupted measurements $\mathcal{A}(M_\sharp) + \Delta + e$, where Δ represents sparse outliers and e represents small dense noise. To the best of our knowledge, theoretical guarantees for this error model have not been previously established in the nonconvex low-rank recovery literature. Surprisingly, we show it is possible to recover the matrix $M_\sharp$ up to a tolerance *independent* of the norm or location of the outliers Δ.

Outline of the chapter. Section 5.2 informally discusses the sharpness and approximation properties, and their impact on convergence of the subgradient and prox-linear methods. Section 5.3 analyzes the parameters μ, ρ, L under RIP. Section 5.4 rigorously discusses convergence guarantees of numerical methods under regularity conditions. Section 5.5 reviews examples of problems satisfying RIP and deduces convergence guarantees for subgradient and prox-linear

algorithms. Sections 5.6 and 5.7 discuss the matrix completion and robust PCA problems, respectively. Section 5.8 discusses robust recovery up to a noise tolerance. Section 5.9 illustrates the developed theory and algorithms with numerical experiments on quadratic/bi-linear sensing, matrix completion, and robust PCA problems.

5.2 Regularity conditions and algorithms

As outlined in Section 5.1, we consider the low-rank matrix recovery problem within the framework of compositional optimization:

$$\min_{x \in \mathcal{X}} f(x) := h(F(x)), \tag{5.7}$$

where $\mathcal{X} \subset \mathbf{E}$ is a closed convex set, $h \colon \mathbf{Y} \to \mathbf{R}$ is a finite convex function and $F \colon \mathbf{E} \to \mathbf{Y}$ is a C^1-smooth map. We depart from previous work on low-rank matrix recovery by allowing h to be nonsmooth. We primary focus on those algorithms for (5.7) that converge rapidly (linearly or faster) when initialized sufficiently close to the solution set.

Such rapid convergence guarantees rely on some regularity of the optimization problem. In the compositional setting, regularity conditions take the following appealing form.

Assumption 5.2.1. *Suppose that the following properties hold for the composite optimization problem* (5.7) *for some real numbers* $\mu, \rho, L > 0$.

1. **(Approximation accuracy)** *The convex models* $f_x(y) := h(F(x) + \nabla F(x)(y - x))$ *satisfy the estimate*

$$|f(y) - f_x(y)| \leq \frac{\rho}{2}\|y - x\|_2^2 \qquad \forall x, y \in \mathcal{X}.$$

2. **(Sharpness)** *The set of minimizers* $\mathcal{X}^* := \arg\min_{x \in \mathcal{X}} f(x)$ *is nonempty and we have*

$$f(x) - \inf_{\mathcal{X}} f \geq \mu \cdot \operatorname{dist}(x, \mathcal{X}^*) \qquad \forall x \in \mathcal{X}.$$

3. **(Subgradient bound)** *The bound,* $\sup_{\zeta \in \partial f(x)} \|\zeta\|_2 \leq L$*, holds for any* x *in the tube*

$$\mathcal{T} := \left\{ x \in \mathcal{X} : \operatorname{dist}(x, \mathcal{X}) \leq \frac{\mu}{\rho} \right\}.$$

As pointed out in the introduction, these three properties are quite intuitive: The approximation accuracy guarantees that the objective function f is well approximated by the convex model f_x, up to a quadratic error relative to the basepoint x. Sharpness stipulates that the objective function should grow at least linearly as one moves away from the solution set. The subgradient bound, in turn, asserts that the subgradients of f are bounded in norm by L on the tube $\mathcal{T}$. In particular, this property is implied by Lipschitz continuity on $\mathcal{T}$.

Lemma 5.2.2 (Subgradient bound and Lipschitz continuity [214, Theorem 9.13])**.** *Suppose a function* $f : \mathbf{E} \to \mathbf{R}$ *is* L*-Lipschitz on an open set* $U \subset \mathbf{E}$*. Then the estimate* $\sup_{\zeta \in \partial f(x)} \|\zeta\|_2 \leq L$ *holds for all* $x \in U$*.*

The definition of the tube $\mathcal{T}$ might look unintuitive at first. Some thought, however, shows that it arises naturally since it provably contains no extraneous stationary points of the problem. In particular, $\mathcal{T}$ will serve as a basin of attraction of numerical methods; see the forthcoming Section 5.4 for details. The following general principle has recently emerged [68, 89, 70, 47]. Under Assumption 5.2.1, basic numerical methods converge rapidly when initialized within the tube $\mathcal{T}$. Let us consider three such procedures and briefly describe their convergence properties. Detailed convergence guarantees are deferred to Section 5.4.

Algorithm 2: Polyak Subgradient Method

Data: $x_0 \in \mathbf{R}^d$

Step k**:** $(k \geq 0)$

Choose $\zeta_k \in \partial f(x_k)$. **If** $\zeta_k = 0$, then exit algorithm.

Set $x_{k+1} = \mathrm{proj}_{\mathcal{X}}\left(x_k - \dfrac{f(x_k) - \min_{\mathcal{X}} f}{\|\zeta_k\|_2^2}\zeta_k\right)$.

Algorithm 3: Subgradient method with decreasing stepsize

Data: Real $\lambda > 0$ and $q \in (0, 1)$.

Step k**:** $(k \geq 0)$

Choose $\zeta_k \in \partial g(x_k)$. **If** $\zeta_k = 0$, then exit algorithm.

Set stepsize $\alpha_k = \lambda \cdot q^k$.

Update iterate $x_{k+1} = \mathrm{proj}_{\mathcal{X}}\left(x_k - \alpha_k \frac{\zeta_k}{\|\zeta_k\|_2}\right)$.

Algorithm 4: Prox-linear algorithm

Data: Initial point $x_0 \in \mathbf{R}^d$, proximal parameter $\beta > 0$.

Step k**:** $(k \geq 0)$

Set $x_{k+1} \leftarrow \underset{x \in \mathcal{X}}{\arg\min}\left\{h\left(F(x_k) + \nabla F(x_k)(x - x_k)\right) + \dfrac{\beta}{2}\|x - x_k\|_2^2\right\}$.

Algorithm 2 is the so-called Polyak subgradient method. In each iteration k, the method travels in the negative direction of a subgradient $\zeta_k \in \partial f(x_k)$, followed by a nearest-point projection onto $\mathcal{X}$. The step-length is governed by the current functional gap $f(x_k) - \min_{\mathcal{X}} f$. In particular, one must have the value $\min_{\mathcal{X}} f$ explicitly available to implement the procedure. This value is sometimes known; case in point, the minimal value of the penalty formulations (5.1) and (5.2) for low-rank recovery is zero when the linear measurements are exact. When the minimal value $\min_{\mathcal{X}} f$ is not known, one can instead use Algorithm 3, which replaces the step-length $(f(x_k) - \min_{\mathcal{X}} f)/\|\zeta_k\|_2$ with a preset geometrically decaying sequence. Notice that the per iteration cost of both subgra-

dient methods is dominated by a single subgradient evaluation and a projection onto $\mathcal{X}$. Under appropriate parameter settings, Assumption 5.2.1 guarantees that both methods converge at a linear rate governed by the ratio $\frac{\mu}{L}$, when initialized within $\mathcal{T}$. The prox-linear algorithm (Algorithm 3), in contrast, converges quadratically to the optimal solution, when initialized within $\mathcal{T}$. The caveat is that each iteration of the prox-linear method requires solving a strongly convex subproblem. Note that for low-rank recovery problems (5.1) and (5.2), the size of the subproblems is proportional to the size of the factors and not the size of the matrices.

In the subsequent sections, we show that Assumption 5.2.1 (or a close variant) holds with favorable parameters $\rho, \mu, L > 0$ for common low-rank matrix recovery problems.

5.3 Regularity under RIP

In this section, we consider the low-rank recovery problems (5.1) and (5.2), and show that restricted isometry properties of the map $\mathcal{A}(\cdot)$ naturally yield well-conditioned compositional formulations.[4] The arguments are short and elementary, and yet apply to such important problems as phase retrieval, blind deconvolution, and covariance matrix estimation.

Setting the stage, consider a linear map $\mathcal{A}\colon \mathbf{R}^{d_1 \times d_2} \to \mathbf{R}^m$, an arbitrary rank

[4]The guarantees we develop in the symmetric setting are similar to those in the recent preprint [154], albeit we obtain a sharper bound on L; the two sets of results were obtained independently. The guarantees for the asymmetric setting are different and are complementary to each other: we analyze the conditioning of the basic problem formulation (5.2), while [154] introduces a regularization term $\|X^\top X - YY^\top\|_F$ that improves the basin of attraction for the subgradient method by a factor of the condition number of $M_\sharp$.

r matrix $M_\sharp \in \mathbf{R}^{d_1 \times d_2}$, and a vector $b \in \mathbf{R}^m$ modeling a corrupted estimate of the measurements $\mathcal{A}(M_\sharp)$. Recall that the goal of low-rank matrix recovery is to determine $M_\sharp$ given $\mathcal{A}$ and b. By the term *symmetric setting*, we mean that $M_\sharp$ is symmetric and positive semidefinite, whereas by *asymmetric setting* we mean that $M_\sharp$ is an arbitrary rank r matrix. We will treat the two settings in parallel. In the symmetric setting, we use $X_\sharp$ to denote any fixed $d \times r$ matrix for which the factorization $M_\sharp = X_\sharp X_\sharp^\top$ holds. Similarly, in the asymmetric case, $X_\sharp$ and $Y_\sharp$ denote any fixed $d_1 \times r$ and $r \times d_2$ matrices, respectively, satisfying $M_\sharp = X_\sharp Y_\sharp$.

We are interested in the set of all possible factorization of $M_\sharp$. Consequently, we will often appeal to the following representations:

$$\{X \in \mathbf{R}^{d_1 \times r} : XX^\top = M_\sharp\} = \{X_\sharp R : R \in O(r)\}, \tag{5.8}$$

$$\{(X, Y) \in \mathbf{R}^{d_1 \times r} \times \mathbf{R}^{r \times d_2} : XY = M_\sharp\} = \{(X_\sharp A, A^{-1} Y_\sharp) : A \in GL(r)\}. \tag{5.9}$$

Throughout, we will let $\mathcal{D}^*(M_\sharp)$ refer to the set (5.8) in the symmetric case and to (5.9) in the asymmetric setting.

Henceforth, fix an arbitrary norm $|||\cdot|||$ on $\mathbf{R}^m$. The following property, widely used in the literature on low-rank recovery, will play a central role in this section.

Assumption 5.3.1 (Restricted Isometry Property (RIP))**.** *There exist constants $\kappa_1, \kappa_2 > 0$ such that for all matrices $W \in \mathbf{R}^{d_1 \times d_2}$ of rank at most $2r$ the following bound holds:*

$$\kappa_1 \|W\|_F \leq |||\mathcal{A}(W)||| \leq \kappa_2 \|W\|_F.$$

Assumption 5.3.1 is classical and is satisfied in various important problems with the rescaled ℓ_2-norm $|||\cdot||| = \frac{1}{\sqrt{m}}\|\cdot\|_2$ and ℓ_1-norm $|||\cdot||| = \frac{1}{m}\|\cdot\|_1$.[5] In Section 5.5

[5] In the latter case, RIP also goes by the name of Restricted Uniform Boundedness (RUB) [33].

we discuss a number of such examples including matrix sensing under (sub-)Gaussian design, phase retrieval, blind deconvolution, and quadratic/bilinear sensing. We summarize the RIP properties for these examples in Table 5.1 and refer the reader to Section 5.5 for the precise statements.

Problem	Measurement $\mathcal{A}(M)_i$	(κ_1, κ_2)	Regime
(sub-)Gaussian sensing	$\langle P_i, M \rangle$	(c, C)	$m \gtrsim \frac{rd}{(1-2p_{\text{fail}})^2} \ln(1 + \frac{1}{1-2p_{\text{fail}}})$
Quadratic sensing I	$p_i^\top M p_i$	$(c, C\sqrt{r})$	$m \gtrsim \frac{r^2 d}{(1-2p_{\text{fail}})^2} \ln(1 + \frac{\sqrt{r}}{1-2p_{\text{fail}}})$
Quadratic sensing II	$p_i^\top M p_i - \tilde{p}_i^\top M \tilde{p}_i$	(c, C)	$m \gtrsim \frac{rd}{(1-2p_{\text{fail}})^2} \ln\left(1 + \frac{1}{1-2p_{\text{fail}}}\right)$
Bilinear sensing	$p_i^\top M q_i$	(c, C)	$m \gtrsim \frac{rd}{(1-2p_{\text{fail}})^2} \ln\left(1 + \frac{1}{1-2p_{\text{fail}}}\right)$

Table 5.1: Common problems satisfying ℓ_1/ℓ_2 RIP in Assumption 5.3.1. The table summarizes the ℓ_1/ℓ_2 RIP for (sub-)Gaussian sensing, quadratic sensing (e.g., phase retrieval), and bilinear sensing (e.g., blind deconvolution) under standard (sub-)Gaussian assumptions on the data generating mechanism. In all cases, we set $|||\cdot||| = \frac{1}{m}\|\cdot\|_1$ and assume for simplicity $d_1 = d_2 = d$. The symbols c and C refer to numerical constants, p_{fail} refers to the proportion of corrupted measurements, κ_3 is a constant multiple of $(1 - 2p_{\text{fail}})$. See Section 5.5 for details.

In light of Assumption 5.3.1, it it natural to take the norm $|||\cdot|||$ as the penalty $h(\cdot)$ in (5.1) and (5.2) . Then the symmetric problem (5.1) becomes

$$\min_{X \in \mathbf{R}^{d \times r}} f(X) := |||\mathcal{A}(XX^\top) - b|||, \tag{5.10}$$

while the asymmetric formulation (5.2) becomes

$$\min_{X \in \mathbf{R}^{d_1 \times r},\ Y \in \mathbf{R}^{r \times d_2}} f(X, Y) := |||\mathcal{A}(XY) - b|||. \tag{5.11}$$

Our immediate goal is to show that under Assumption 5.3.1, the problems (5.10) and (5.11) are well-conditioned in the sense of Assumption 5.2.1. We note that the asymmetric setting is more nuanced that its symmetric counterpart because Assumption 5.2.1 can only be guaranteed to hold on bounded sets. Nonetheless, as we discuss in Section 5.4, a localized version of Assumption 5.2.1 suffices to guarantee rapid local convergence of subgradient and prox-

linear methods. In particular, our analysis of the local sharpness in the asymmetric setting is new and illuminating; it shows that the regularization technique suggested in [154] is not needed at all for the prox-linear method. This conclusion contrasts with known techniques in the smooth setting, where regularization is often used.

5.3.1 Approximation and Lipschitz continuity

We begin with the following elementary proposition, which estimates the subgradient bound L and the approximation modulus ρ in the symmetric setting. In what follows, we will use the expressions

$$f_X(Z) = \|\mathcal{A}(XX^\top + X(Z-X)^\top + (Z-X)X^\top) - b\|,$$
$$f_{(X,Y)}(\widehat{X},\widehat{Y}) = \|\mathcal{A}(XY + X(\widehat{Y}-Y) + (\widehat{X}-X)Y) - b\|.$$

Proposition 5.3.2 (Approximation accuracy and Lipschitz continuity (symmetric))**.**
Suppose Assumption 5.3.1 holds. Then for all $X, Z \in \mathbf{R}^{d\times r}$ the following estimates hold:

$$|f(Z) - f_X(Z)| \le \kappa_2 \|Z-X\|_F^2,$$
$$|f(X) - f(Z)| \le \kappa_2 \|X+Z\|_{op}\|X-Z\|_F.$$

Proof. To see the first estimate, observe that

$$\begin{aligned}
|f(Z) - f_X(Z)| &= |||\mathcal{A}(ZZ^\top) - b||| - |||\mathcal{A}(XX^\top + X(Z-X)^\top + (Z-X)X^\top) - b||| \\
&\leq |||\mathcal{A}(ZZ^\top - XX^\top - X(Z-X)^\top - (Z-X)X^\top||| \qquad (5.12) \\
&= |||\mathcal{A}((Z-X)(Z-X)^\top)||| \\
&\leq \kappa_2 \left\|(Z-X)(Z-X)^\top\right\|_F \qquad (5.13) \\
&\leq \kappa_2 \|Z-X\|_F^2,
\end{aligned}$$

where (5.12) follows from the reverse triangle inequality and (5.13) uses Assumption 5.3.1. Next, for any $X, Z \in \mathcal{X}$ we successively compute:

$$\begin{aligned}
|f(X) - f(Z)| &= \left| |||\mathcal{A}(XX^\top) - b||| - |||\mathcal{A}(ZZ^\top) - b||| \right| \\
&\leq \left|||\mathcal{A}(XX^\top - ZZ^\top)\right||| \qquad (5.14) \\
&\leq \kappa_2 \|XX^\top - ZZ^\top\|_F \qquad (5.15) \\
&= \frac{\kappa_2}{2} \|(X+Z)(X-Z)^\top + (X-Z)(X+Z)^\top\|_F \\
&\leq \kappa_2 \|(X+Z)(X-Z)\|_F \\
&\leq \kappa_2 \|X+Z\|_{op} \|X-Z\|_F,
\end{aligned}$$

where (5.14) follows from the reverse triangle inequality and (5.15) uses Assumption 5.3.1. The proof is complete. □

The estimates of L and ρ in the asymmetric setting are completely analogous; we record them in the following proposition.

Proposition 5.3.3 (Approximation accuracy and Lipschitz continuity (asymmetric))**.**

Suppose Assumption 5.3.1 holds. Then for all $X, \widehat{X} \in \mathbf{R}^{d_1 \times r}$ and $Y, \widehat{Y} \in \mathbf{R}^{r \times d_2}$ the follow-

ing estimates hold:

$$|f(\widehat{X},\widehat{Y}) - f_{(X,Y)}(\widehat{X},\widehat{Y})| \le \frac{\kappa_2}{2} \cdot \|(X,Y) - (\widehat{X},\widehat{Y})\|_F^2,$$
$$|f(X,Y) - f(\widehat{X},\widehat{Y})| \le \frac{\kappa_2 \max\{\|X+\widehat{X}\|_{\mathrm{op}}, \|Y+\widehat{Y}\|_{\mathrm{op}}\}}{\sqrt{2}} \cdot \|(X,Y) - (\widehat{X},\widehat{Y})\|_F.$$

Proof. To see the first estimate, observe that

$$\begin{aligned}
|f(\widehat{X},\widehat{Y}) - f_{(X,Y)}(\widehat{X},\widehat{Y})| &= \left| \|\mathcal{A}(\widehat{X}\widehat{Y}) - b\| - \|\mathcal{A}(XY + X(\widehat{Y}-Y) + (\widehat{X}-X)Y) - b\| \right| \\
&\le \|\mathcal{A}(\widehat{X}\widehat{Y} - XY - X(\widehat{Y}-Y) - (\widehat{X}-X)Y)\| \\
&= \|\mathcal{A}((X-\widehat{X})(Y-\widehat{Y}))\| \\
&\le \kappa_2 \left\| (X-\widehat{X})(Y-\widehat{Y}) \right\|_F \\
&\le \frac{\kappa_2}{2} \left(\|X-\widehat{X}\|_F^2 + \|Y-\widehat{Y}\|_F^2 \right),
\end{aligned}$$

where the last estimate follows from Young's inequality $2ab \le a^2 + b^2$. Next, we successively compute:

$$\begin{aligned}
|f(X,Y) - f(\widehat{X},\widehat{Y})| &\le \|\mathcal{A}(XY - \widehat{X}\widehat{Y})\| \le \kappa_2 \|XY - \widehat{X}\widehat{Y}\|_F \\
&= \frac{\kappa_2}{2} \|(X+\widehat{X})(Y-\widehat{Y})^\top + (X-\widehat{X})(Y+\widehat{Y})^\top\|_F \\
&\le \frac{\kappa_2 \max\{\|X+\widehat{X}\|_{\mathrm{op}}, \|Y+\widehat{Y}\|_{\mathrm{op}}\}}{2} (\|Y-\widehat{Y}\|_F + \|X-\widehat{X}\|_F).
\end{aligned}$$

The result follows by noting that $a + b \le \sqrt{2(a^2+b^2)}$ for all $a, b \in \mathbf{R}$.

□

5.3.2 Sharpness

We next move on to estimates of the sharpness constant μ. We first deal with the noiseless setting $b = \mathcal{A}(M_\sharp)$ in Section 5.3.2, and then move on to the general case when the measurements are corrupted by outliers in Section 5.3.2.

Sharpness in the noiseless regime

We begin with with the symmetric setting in the noiseless case $b = \mathcal{A}(M_\sharp)$. By Assumption 5.3.1, we have the estimate

$$f(X) = \|\mathcal{A}(XX^\top) - b\| = \|\mathcal{A}(XX^\top - X_\sharp X_\sharp^\top)\| \geq \kappa_1 \|XX^\top - X_\sharp X_\sharp^\top\|_F. \tag{5.16}$$

It follows that the set of minimizers $\arg\min_{X \in \mathbf{R}^{d\times r}} f(X)$ coincides with the set of minimizers of the function $X \mapsto \|XX^\top - X_\sharp X_\sharp^\top\|_F$, namely

$$\mathcal{D}^*(M_\sharp) := \{X_\sharp R : R \in O(r)\}.$$

Thus to argue sharpness of f it suffices to estimate the sharpness constant of the function $X \mapsto \|XX^\top - X_\sharp X_\sharp^\top\|_F$. Fortunately, this calculation was already done in [233, Lemma 5.4].

Proposition 5.3.4 ([233, Lemma 5.4])**.** *For any matrices $X, Z \in \mathbf{R}^{d\times r}$, we have the bound*

$$\|XX^\top - ZZ^\top\|_F \geq \sqrt{2(\sqrt{2}-1)}\sigma_r(Z) \cdot \min_{R\in O(r)} \|X - ZR\|_F.$$

Consequently if Assumption 5.3.1 holds in the noiseless setting $b = \mathcal{A}(M_\sharp)$, then the bound holds:

$$f(X) \geq \kappa_1 \sqrt{2(\sqrt{2}-1)\sigma_r(M_\sharp)} \cdot \operatorname{dist}(X, \mathcal{D}^*(M_\sharp)) \qquad \textit{for all } X \in \mathbf{R}^{d\times r}.$$

We next consider the asymmetric case. By exactly the same reasoning as before, the set of minimizers of $f(X, Y)$ coincides with the set of minimizers of the function $(X, Y) \mapsto \|XY - X_\sharp Y_\sharp\|_F$, namely

$$\mathcal{D}^*(M_\sharp) := \{(X_\sharp A, A^{-1} Y_\sharp) : A \in GL(r)\}.$$

Thus to argue sharpness of f it suffices to estimate the sharpness constant of the function $(X, Y) \mapsto \|XY - X_\sharp Y_\sharp\|_F$.

Notice that in contrast to the symmetric setting, the sharpness estimate is only valid on bounded sets. Indeed, this is unavoidable even in the setting $d_1 = d_2 = 2$. To see this, define $M_\sharp = e_2 e_2^\top$ and for any $\alpha > 0$ set $x = \alpha e_1$ and $w = \frac{1}{\alpha} e_1$. It is routine to compute

$$\frac{\|xw^\top - M_\sharp\|_F}{\operatorname{dist}((x, w), \mathcal{D}^*(M_\sharp))} = \sqrt{\frac{2}{2 + \alpha^2 + \frac{1}{\alpha^2}}}.$$

Therefore letting α tend to zero (or infinity) the quotient tends to zero.

The following theorem is a nonsymmetric variant of Proposition 6.2.1.

Theorem 5.3.5 (Sharpness (asymmetric and noiseless))**.** *Fix a constant $\nu > 0$ and define $X_\sharp := U\sqrt{\Lambda}$ and $Y_\sharp = \sqrt{\Lambda} V^\top$, where $M_\sharp = U\Lambda V^\top$ is any compact singular value decomposition of $M_\sharp$. Then for all $X \in \mathbf{R}^{d_1 \times r}$ and $Y \in \mathbf{R}^{r \times d_2}$ satisfying*

$$\begin{aligned} \max\{\|X - X_\sharp\|_F, \|Y - Y_\sharp\|_F\} &\leq \nu \sqrt{\sigma_r(M_\sharp)} \\ \operatorname{dist}((X, Y), \mathcal{D}^*(M_\sharp)) &\leq \frac{\sqrt{\sigma_r(M_\sharp)}}{1 + 2(1 + \sqrt{2})\nu} \end{aligned}, \tag{5.17}$$

the estimate holds:

$$\|XY - M_\sharp\|_F \geq \frac{\sqrt{\sigma_r(M_\sharp)}}{2 + 4(1 + \sqrt{2})\nu} \cdot \operatorname{dist}((X, Y), \mathcal{D}^*(M_\sharp)).$$

Proof. Define $\delta := \frac{1}{1+2(1+\sqrt{2})\nu}$ and consider a pair of matrices X and Y satisfying (5.17). Let $A \in GL(r)$ be an invertible matrix satisfying

$$A \in \operatorname*{arg\,min}_{A \in GL(r)} \left\{ \|X - X_\sharp A\|_F^2 + \|Y - A^{-1} Y_\sharp\|_F^2 \right\}. \tag{5.18}$$

As a first step, we successively compute

$$
\begin{aligned}
&\|XY - X_\sharp Y_\sharp\|_F \\
&= \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp) + (X - X_\sharp A)(Y - A^{-1}Y_\sharp)\|_F \\
&\geq \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F - \|(X - X_\sharp A)(Y - A^{-1}Y_\sharp)\|_F \\
&\geq \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F - \|X - X_\sharp A\|_F \cdot \|Y - A^{-1}Y_\sharp\|_F \\
&\geq \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F - \frac{1}{2}(\|X - X_\sharp A\|_F^2 + \|Y - A^{-1}Y_\sharp\|_F^2) \\
&= \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F - \frac{1}{2}\mathrm{dist}^2((X,Y), \mathcal{D}^*(M_\sharp)) \\
&\geq \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F - \frac{\delta\sqrt{\sigma_r(M_\sharp)}}{2} \cdot \mathrm{dist}((X,Y), \mathcal{D}^*(M_\sharp)).
\end{aligned}
\tag{5.19}
$$

We next aim to lower bound the first term on the right. To this end, observe

$$
\begin{aligned}
&\|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F^2 \\
&\qquad = \|(X - X_\sharp A)(A^{-1}Y_\sharp)\|_F^2 + \|X_\sharp A(Y - A^{-1}Y_\sharp)\|_F^2 \\
&\qquad\quad + 2\mathrm{Tr}((X - X_\sharp A)(A^{-1}Y_\sharp)(Y - A^{-1}Y_\sharp)^\top (X_\sharp A)^\top).
\end{aligned}
\tag{5.20}
$$

We claim that the cross-term is non-negative. To see this, observe that first order optimality conditions in (5.18) directly imply that A satisfies the equality

$$
A^\top X_\sharp^\top (X - X_\sharp A) = (Y - A^{-1}Y_\sharp)Y_\sharp^\top A^{-\top}.
$$

Thus we obtain

$$
\begin{aligned}
\mathrm{Tr}((X - X_\sharp A)(A^{-1}Y_\sharp)(Y - A^{-1}Y_\sharp)^\top (X_\sharp A)^\top) &= \mathrm{Tr}(A^\top X_\sharp^\top (X - X_\sharp A)(A^{-1}Y_\sharp)(Y - A^{-1}Y_\sharp)^\top) \\
&= \mathrm{Tr}((Y - A^{-1}Y_\sharp)Y_\sharp^\top A^{-T}(A^{-1}Y_\sharp)(Y - A^{-1}Y_\sharp)^\top) \\
&= \|(A^{-1}Y_\sharp)(Y - A^{-1}Y_\sharp)\|_F^2.
\end{aligned}
$$

Therefore, returning to (5.20) we conclude that

$$
\begin{aligned}
&\|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F \\
&\geq \sqrt{\|(X - X_\sharp A)(A^{-1}Y_\sharp)\|_F^2 + \|X_\sharp A(Y - A^{-1}Y_\sharp)\|_F^2} \\
&\geq \sqrt{\sigma_r(M_\sharp)} \cdot \min\{\sigma_r(A^{-1}), \sigma_r(A)\} \cdot \mathrm{dist}((X,Y), \mathcal{D}^*(M_\sharp)).
\end{aligned}
\tag{5.21}
$$

Combining (5.19) and (5.21), we obtain

$$\|XY - M_\sharp\|_F \geq \sqrt{\sigma_r(M_\sharp)} \cdot \left(\min\{\sigma_r(A^{-1}), \sigma_r(A)\} - \frac{\delta}{2}\right) \cdot \operatorname{dist}((X,Y), \mathcal{D}^*(M_\sharp)) \tag{5.22}$$

Finally, we estimate $\min\{\sigma_r(A^{-1}), \sigma_r(A)\}$. To this end, first note that

$$\begin{aligned}\|X_\sharp - X_\sharp A\|_F + \|Y_\sharp - A^{-1}Y_\sharp\|_F &\leq \|X_\sharp - X\|_F + \|Y_\sharp - Y\|_F + \sqrt{2} \cdot \operatorname{dist}((X,Y), \mathcal{D}^*(M_\sharp)) \\ &\leq 2\nu\sqrt{\sigma_r(M_\sharp)} \cdot (1 + \sqrt{2}).\end{aligned} \tag{5.23}$$

We now aim to lower bound the left-hand-side in terms of $\min\{\sigma_r(A^{-1}), \sigma_r(A)\}$. Observe

$$\|X_\sharp - X_\sharp A\|_F \geq \|X_\sharp - X_\sharp A\|_{\mathrm{op}} \geq \sqrt{\sigma_r(M_\sharp)} \cdot \|I - A\|_{\mathrm{op}} \geq \sqrt{\sigma_r(M_\sharp)} \cdot (\sigma_1(A) - 1).$$

Similarly, we have

$$\|Y_\sharp - A^{-1}Y_\sharp\|_F \geq \|Y_\sharp - A^{-1}Y_\sharp\|_{\mathrm{op}} \geq \sqrt{\sigma_r(M_\sharp)} \cdot \|I - A^{-1}\|_{\mathrm{op}} \geq \sqrt{\sigma_r(M_\sharp)} \cdot (\sigma_1(A^{-1}) - 1).$$

Hence using (5.23), we obtain the estimate

$$\min\{\sigma_r(A^{-1}), \sigma_r(A)\} \geq \left(1 + 2\nu \cdot (1 + \sqrt{2})\right)^{-1} = \delta.$$

Using this estimate in (5.22) completes the proof. □

Sharpness in presence of outliers

The most important example of the norm $|||\cdot|||$ for us is the scaled ℓ_1-norm $|||\cdot||| = \frac{1}{m}\|\cdot\|_1$. Indeed, all the examples in the forthcoming Section 5.5 will satisfy RIP relative to this norm. In this section, we will show that the ℓ_1-norm has an added advantage. Under reasonable RIP-type conditions, sharpness will hold even if up to a half of the measurements are grossly corrupted.

Henceforth, for any set $\mathcal{I}$, define the restricted map $\mathcal{A}_{\mathcal{I}} := (\mathcal{A}(X))_{i\in\mathcal{I}}$. We interpret the set $\mathcal{I}$ as corresponding to (arbitrarily) outlying measurements, while its complement corresponds to exact measurements. Motivated by the work [89] on robust phase retrieval, we make the following assumption.

Assumption 5.3.6 ($\mathcal{I}$-outlier bounds)**.** *There exists a set $\mathcal{I} \subset \{1,\dots,m\}$ and a constant $\kappa_3 > 0$ such that the following hold.*

(C1) *Equality holds $b_i = \mathcal{A}(M_\sharp)_i$ for all $i \notin \mathcal{I}$.*

(C2) *For all matrices W of rank at most $2r$, we have*

$$\kappa_3\|W\|_F \le \frac{1}{m}\|\mathcal{A}_{\mathcal{I}^c}(W)\|_1 - \frac{1}{m}\|\mathcal{A}_{\mathcal{I}}(W)\|_1. \tag{5.24}$$

The assumption is simple to interpret. To elucidate the bound (5.24), let us suppose that the restricted maps $\mathcal{A}_{\mathcal{I}}$ and $\mathcal{A}_{\mathcal{I}^c}$ satisfy Assumption 5.3.1 (RIP) with constants $\hat{\kappa_1}$, $\hat{\kappa_2}$ and κ_1, κ_2, respectively. Then for any rank $2r$ matrix X we immediately deduce the estimate

$$\frac{1}{m}\|\mathcal{A}_{\mathcal{I}^c}(W)\|_1 - \frac{1}{m}\|\mathcal{A}_{\mathcal{I}}(W)\|_1 \ge ((1-p_{\text{fail}})\kappa_1 - p_{\text{fail}}\hat{\kappa_2})\,\|W\|_F,$$

where $p_{\text{fail}} = \frac{|\mathcal{I}|}{m}$ denotes the corruption frequency. In particular, the right-hand side is positive as long as the corruption frequency is below the threshold $p_{\text{fail}} < \frac{\kappa_1}{\kappa_1+\hat{\kappa_2}}$.

Combining Assumption 5.3.6 with Proposition 6.2.1 quickly yields sharpness of the objective even in the noisy setting.

Proposition 5.3.7 (Sharpness with outliers (symmetric))**.** *Suppose that Assumption 5.3.6 holds. Then*

$$f(X) - f(X_\sharp) \ge \kappa_3\left(\sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp)\right)\operatorname{dist}(X, \mathcal{D}^*(M_\sharp)) \qquad \text{for all } X \in \mathbf{R}^{d\times r}.$$

Proof. Defining $\Delta := \mathcal{A}(X_\sharp X_\sharp^\top) - b$, we have the following bound:

$$\begin{aligned} m \cdot (f(X) - f(X_\sharp)) &= \|\mathcal{A}\left(XX^\top - X_\sharp X_\sharp^\top\right) + \Delta\|_1 - \|\Delta\|_1 \\ &= \|\mathcal{A}_{\mathcal{I}^c}(XX^\top - X_\sharp X_\sharp^\top)\|_1 + \sum_{i \in \mathcal{I}} \left(\left| \left(\mathcal{A}(XX^\top - X_\sharp X_\sharp^\top) \right)_i + \Delta_i \right| - |\Delta_i| \right) \\ &\geq \|\mathcal{A}_{\mathcal{I}^c}(XX^\top - X_\sharp X_\sharp^\top)\|_1 - \|\mathcal{A}_{\mathcal{I}}(XX^\top - X_\sharp X_\sharp^\top)\|_1 \\ &\geq \kappa_3 m \|XX^\top - X_\sharp X_\sharp^\top\|_F \geq \kappa_3 m \left(\sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp) \right) \operatorname{dist}(X, \mathcal{D}^*(M_\sharp)), \end{aligned}$$

where the first inequality follows by the reverse triangle inequality, the second inequality follows by Assumption (C2), and the final inequality follows from Proposition 6.2.1. The proof is complete. □

The argument in the asymmetric setting is completely analogous.

Proposition 5.3.8 (Sharpness with outliers (asymmetric))**.** *Suppose that Assumption 5.3.6 holds. Fix a constant $\nu > 0$ and define $X_\sharp := U\sqrt{\Lambda}$ and $Y_\sharp = \sqrt{\Lambda}V^\top$, where $M_\sharp = U\Lambda V^\top$ is any compact singular value decomposition of $M_\sharp$. Then for all $X \in \mathbf{R}^{d_1 \times r}$ and $Y \in \mathbf{R}^{r \times d_2}$ satisfying*

$$\begin{aligned} \max\{\|X - X_\sharp\|_F, \|Y - Y_\sharp\|_F\} &\leq \nu \sqrt{\sigma_r(M_\sharp)} \\ \operatorname{dist}((X, Y), \mathcal{D}^*(M_\sharp)) &\leq \frac{\sqrt{\sigma_r(M_\sharp)}}{1 + 2(1 + \sqrt{2})\nu} \end{aligned}$$

The estimate holds:

$$f(X, Y) - f(X_\sharp, Y_\sharp) \geq \frac{\kappa_3 \sqrt{\sigma_r(M_\sharp)}}{2 + 4(1 + \sqrt{2})\nu} \cdot \operatorname{dist}((X, Y), \mathcal{D}^*(M_\sharp)).$$

5.4 Guarantees for subgradient & prox-linear methods

In this section, we formally develop convergence guarantees for Algorithms 2, 3, and 4 under Assumption 5.2.1, and deduce performance guarantees in the

RIP setting. To this end, it will be useful to first consider a broader class than the compositional problems (5.7). Recall that a function $f\colon \mathbf{E} \to \mathbf{R} \cup \{+\infty\}$ is ρ-weakly convex if the perturbed function $x \mapsto f(x) + \frac{\rho}{2}\|x\|_2^2$ is convex. In particular, a composite function $f = h \circ F$ satisfying the approximation guarantee

$$|f_x(y) - f(y)| \leq \frac{\rho}{2}\|y - x\|_2^2 \qquad \forall x, y$$

is automatically ρ-weakly convex [83, Lemma 4.2].

Setting the stage, we introduce the following assumption.

Assumption 5.4.1. *Consider the optimization problem,*

$$\min_{x \in \mathcal{X}} \; f(x). \tag{5.25}$$

Suppose that the following properties hold for some real numbers $\mu, \rho > 0$.

1. **(Weak convexity)** *The set $\mathcal{X}$ is closed and convex, while the function $f\colon \mathbf{E} \to \mathbf{R}$ is ρ-weakly convex.*
2. **(Sharpness)** *The set of minimizers $\mathcal{X}^* := \operatorname*{arg\,min}_{x \in \mathcal{X}} f(x)$ is nonempty and the following inequality holds:*

$$f(x) - \inf_{\mathcal{X}} f \geq \mu \cdot \operatorname{dist}(x, \mathcal{X}^*) \qquad \forall x \in \mathcal{X}.$$

In particular, notice that Assumption 5.2.1 implies Assumption 5.4.1. Taken together, weak convexity and sharpness provide an appealing framework for deriving local rapid convergence guarantees for numerical methods. In this section, we specifically focus on two such procedures: the subgradient and prox-linear algorithms. We aim to estimate both the radius of rapid converge around the solution set and the rate of convergence. Note that both of the algorithms,

when initialized at a stationary point could stay there for all subsequent iterations. Since we are interested in finding global minima, we therefore estimate the neighborhood of the solution set that has no extraneous stationary points. This is the content of the following simple lemma.

Lemma 5.4.2 ([68, Lemma 3.1])**.** *Suppose that Assumption 5.4.1 holds. Then the problem* (5.25) *has no stationary points x satisfying*

$$0 < \operatorname{dist}(x; \mathcal{X}^*) < \frac{2\mu}{\rho}.$$

It is worthwhile to note that the estimate $\frac{2\mu}{\rho}$ of the radius in Lemma 5.4.2 is tight [47, Section 3]. Hence, let us define for any $\gamma > 0$ the tube

$$\mathcal{T}_\gamma := \left\{ z \in \mathcal{X} \colon \operatorname{dist}(z, \mathcal{X}^*) \leq \gamma \cdot \frac{\mu}{\rho} \right\}. \tag{5.26}$$

Thus we would like to search for algorithms whose basin of attraction is a tube $\mathcal{T}_\gamma$ for some numerical constant $\gamma > 0$. Such a basin of attraction is in essence optimal.

The rate of convergence of the subgradient methods (Algorithms 2 and 3) relies on the subgradient bound and the condition measure:

$$L := \sup\{\|\zeta\|_2 : \zeta \in \partial f(x), x \in \mathcal{T}_1\} \qquad \text{and} \qquad \tau := \frac{\mu}{L}.$$

A straightforward argument [68, Lemma 3.2] shows $\tau \in [0, 1]$. The following theorem appears as [68, Theorem 4.1], while its application to phase retrieval was investigated in [70].

Theorem 5.4.3 (Polyak subgradient method)**.** *Suppose that Assumption 5.4.1 holds and fix a real number $\gamma \in (0, 1)$. Then Algorithm 2 initialized at any point $x_0 \in \mathcal{T}_\gamma$ produces iterates that converge Q-linearly to $\mathcal{X}^*$, that is*

$$\operatorname{dist}^2(x_{k+1}, \mathcal{X}^*) \leq \left(1 - (1 - \gamma)\tau^2\right) \operatorname{dist}^2(x_k, \mathcal{X}^*) \qquad \forall k \geq 0.$$

The following theorem appears as [68, Theorem 6.1]. The convex version of the result dates back to Goffin [109].

Theorem 5.4.4 (Geometrically decaying subgradient method)**.** *Suppose that Assumption 5.4.1 holds, fix a real number $\gamma \in (0,1)$, and suppose $\tau \leq \sqrt{\frac{1}{2-\gamma}}$. Set $\lambda := \frac{\gamma\mu^2}{\rho L}$ and $q := \sqrt{1-(1-\gamma)\tau^2}$ in Algorithm 3. Then the iterates x_k generated by Algorithm 3, initialized at any point $x_0 \in \mathcal{T}_\gamma$, satisfy:*

$$\operatorname{dist}^2(x_k; \mathcal{X}^*) \leq \frac{\gamma^2\mu^2}{\rho^2}\left(1-(1-\gamma)\tau^2\right)^k \qquad \forall k \geq 0.$$

Let us now specialize to the composite setting under Assumption 5.2.1. Since Assumption 5.2.1 implies Assumption 5.4.1, both subgradient Algorithms 2 and 3 will enjoy a linear rate of convergence when initialized sufficiently close the solution set. The following theorem, on the other hand, shows that the prox-linear method will enjoy a quadratic rate of convergence (at the price of a higher per-iteration cost). Guarantees of this type have appeared, for example, in [89, 31, 82].

Theorem 5.4.5 (Prox-linear algorithm)**.** *Suppose Assumption 5.2.1 holds. Choose any $\beta \geq \rho$ in Algorithm 4 and set $\gamma := \rho/\beta$. Then Algorithm 4 initialized at any point $x_0 \in \mathcal{T}_\gamma$ converges quadratically:*

$$\operatorname{dist}(x_{k+1}, \mathcal{X}^*) \leq \tfrac{\beta}{\mu} \cdot \operatorname{dist}^2(x_k, \mathcal{X}^*) \qquad \forall k \geq 0.$$

We now apply the results above to the low-rank matrix factorization problem under RIP, whose regularity properties were verified in Section 5.3. In particular, we have the following efficiency guarantees of the subgradient and prox-linear methods applied to this problem.

Corollary 5.4.6 (Convergence guarantees under RIP (symmetric))**.** *Suppose Assumptions 5.3.1 and 5.3.6 are valid with $|||\cdot||| = \frac{1}{m}\|\cdot\|_1$ and consider the optimization*

problem

$$\min_{X \in \mathbf{R}^{d \times r}} \; f(X) = \frac{1}{m} \|\mathcal{A}(XX^\top) - b\|_1.$$

Choose any matrix X_0 satisfying

$$\frac{\operatorname{dist}(X_0, \mathcal{D}^*(M_\sharp))}{\sqrt{\sigma_r(M_\sharp)}} \leq 0.2 \cdot \frac{\kappa_3}{\kappa_2}.$$

Define the condition number $\chi := \sigma_1(M_\sharp)/\sigma_r(M_\sharp)$. Then the following are true.

1. **(Polyak subgradient)** *Algorithm 2 initialized at X_0 produces iterates that converge linearly to $\mathcal{D}^*(M_\sharp)$, that is*

$$\frac{\operatorname{dist}^2(X_k, \mathcal{D}^*(M_\sharp))}{\sigma_r(M_\sharp)} \leq \left(1 - \frac{0.2}{1 + \frac{4\kappa_2^2\chi}{\kappa_3^2}}\right)^k \cdot \frac{\kappa_3^2}{100\kappa_2^2} \qquad \forall k \geq 0.$$

2. **(geometric subgradient)** *Algorithm 3 with $\lambda = \frac{0.81\kappa_3^2\sqrt{\sigma_r(M_\sharp)}}{2\kappa_2(\kappa_3 + 2\kappa_2\sqrt{\chi})}$, $q = \sqrt{1 - \frac{0.2}{1+4\kappa_2^2\chi/\kappa_3^2}}$ and initialized at X_0 converges linearly:*

$$\frac{\operatorname{dist}^2(X_k, \mathcal{D}^*(M_\sharp))}{\sigma_r(M_\sharp)} \leq \left(1 - \frac{0.2}{1 + \frac{4\kappa_2^2\chi}{\kappa_3^2}}\right)^k \cdot \frac{\kappa_3^2}{100\kappa_2^2} \qquad \forall k \geq 0.$$

3. **(prox-linear)** *Algorithm 4 with $\beta = \rho$ and initialized at X_0 converges quadratically:*

$$\frac{\operatorname{dist}(X_k, \mathcal{D}^*(M_\sharp)))}{\sqrt{\sigma_r(M_\sharp)}} \leq 2^{-2^k} \cdot \frac{0.45\kappa_3}{\kappa_2} \qquad \forall k \geq 0.$$

5.4.1 Guarantees under local regularity

As explained in Section 5.3, Assumptions 5.2.1 and 5.4.1 are reasonable in the symmetric setting under RIP. The asymmetric setting is more nuanced. Indeed, the solution set is unbounded, while uniform bounds on the sharpness and subgradient norms are only valid on bounded sets. One remedy, discussed in [154],

is to modify the optimization formulation by introducing a form of regularization:

$$\min_{X,Y} \; \|\mathcal{A}(XY) - y\| + \lambda \|X^\top X - YY^\top\|_F.$$

In this section, we take a different approach that requires no modification to the optimization problem nor the algorithms. The key idea is to show that if the problem is well-conditioned only on a neighborhood of a particular solution, then the iterates will remain in the neighborhood provided the initial point is sufficiently close to the solution. In fact, we will see that the iterates themselves must converge. The proofs of the results in this section (Theorems 5.4.8, 5.4.9, and 5.4.11) are deferred to Section 5.10.1.

We begin with the following localized version of Assumption 5.4.1.

Assumption 5.4.7. *Consider the optimization problem,*

$$\min_{x \in \mathcal{X}} \; f(x). \tag{5.27}$$

Fix an arbitrary point $\bar{x} \in \mathcal{X}^*$ and suppose that the following properties hold for some real numbers $\epsilon, \mu, \rho > 0$.

1. **(Local weak convexity)** *The set $\mathcal{X}$ is closed and convex, and the bound holds:*

$$f(y) \geq f(x) + \langle \zeta, y - x \rangle - \frac{\rho}{2}\|y - x\|_2^2 \qquad \forall x, y \in \mathcal{X} \cap B_\epsilon(\bar{x}), \zeta \in \partial f(x).$$

2. **(Local sharpness)** *The inequality holds:*

$$f(x) - \inf_{\mathcal{X}} f \geq \mu \cdot \operatorname{dist}(x, \mathcal{X}^*) \qquad \forall x \in \mathcal{X} \cap B_\epsilon(\bar{x}).$$

The following two theorems establish convergence guarantees of the two subgradient methods under Assumption 5.4.7. Abusing notation slightly, we define the local quantities:

$$L := \sup_{\zeta \in \partial f(x)} \{\|\zeta\|_2 : x \in \mathcal{X} \cap B_\epsilon(\bar{x})\} \quad \text{and} \quad \tau := \frac{\mu}{L}.$$

Theorem 5.4.8 (Polyak subgradient method (local regularity))**.** *Suppose Assumption 5.4.7 holds and fix an arbitrary point $x_0 \in B_{\epsilon/4}(\bar{x})$ satisfying*

$$\operatorname{dist}(x_0, \mathcal{X}^*) \leq \min\left\{\frac{3\epsilon\mu^2}{64L^2}, \frac{\mu}{2\rho}\right\}.$$

Then Algorithm 2 initialized at x_0 produces iterates x_k that always lie in $B_\epsilon(\bar{x})$ and satisfy

$$\operatorname{dist}^2(x_{k+1}, \mathcal{X}^*) \leq \left(1 - \tfrac{1}{2}\tau^2\right)\operatorname{dist}^2(x_k, \mathcal{X}^*), \qquad \text{for all } k \geq 0. \tag{5.28}$$

Moreover the iterates converge to some point $x_\infty \in \mathcal{X}^*$ at the R-linear rate

$$\|x_k - x_\infty\|_2 \leq \frac{16L^3 \cdot \operatorname{dist}(x_0, \mathcal{X}^*)}{3\mu^3} \cdot \left(1 - \tfrac{1}{2}\tau^2\right)^{\frac{k}{2}} \qquad \text{for all } k \geq 0.$$

Theorem 5.4.9 (Geometrically decaying subgradient method (local regularity))**.** *Suppose that Assumption 5.4.7 holds and that $\tau \leq \frac{1}{\sqrt{2}}$. Define $\gamma = \frac{\epsilon\rho}{4L+\epsilon\rho}$, $\lambda = \frac{\gamma\mu^2}{\rho L}$, and $q = \sqrt{1-(1-\gamma)\tau^2}$. Then Algorithm 3 initialized at any point $x_0 \in B_{\epsilon/4}(\bar{x}) \cap \mathcal{T}_\gamma$ generates iterates x_k that always lie in $B_\epsilon(\bar{x})$ and satisfy*

$$\operatorname{dist}^2(x_k; \mathcal{X}^*) \leq \frac{\gamma^2\mu^2}{\rho^2}\left(1 - (1-\gamma)\tau^2\right)^k \qquad \text{for all } k \geq 0. \tag{5.29}$$

Moreover, the iterates converge to some point $x_\infty \in \mathcal{X}^*$ at the R-linear rate

$$\|x_k - x_\infty\|_2 \leq \tfrac{\lambda}{1-q} \cdot q^k \qquad \text{for all } k \geq 0.$$

We end the section by specializing to the composite setting and analyzing the prox-linear method. The following is the localized version of Assumption 5.2.1.

Assumption 5.4.10. Consider the optimization problem,

$$\min_{x \in \mathcal{X}} f(x) := h(F(x)),$$

where the function $h(\cdot)$ and the set $\mathcal{X}$ are convex and $F(\cdot)$ is differentiable. Fix a point $\bar{x} \in \mathcal{X}^*$ and suppose that the following properties holds for some real numbers $\epsilon, \mu, \rho > 0$.

1. **(Approximation accuracy)** The convex models $f_x(y) := h(F(x)+\nabla F(x)(y-x))$ satisfy the estimate:

$$|f(y) - f_x(y)| \leq \frac{\rho}{2}\|y - x\|_2^2 \qquad \forall x \in \mathcal{X} \cap B_\epsilon(\bar{x}), y \in \mathcal{X}.$$

2. **(Sharpness)** The inequality holds:

$$f(x) - \inf_{\mathcal{X}} f \geq \mu \cdot \operatorname{dist}(x, \mathcal{X}^*) \qquad \forall x \in \mathcal{X} \cap B_\epsilon(\bar{x}).$$

The following theorem provides convergence guarantees for the prox-linear method under Assumption 5.4.10.

Theorem 5.4.11 (Prox-linear (local))**.** *Suppose Assumption 5.4.10 holds, choose any $\beta \geq \rho$, and fix an arbitrary point $x_0 \in B_{\epsilon/2}(\bar{x})$ satisfying*

$$f(x_0) - \min_{\mathcal{X}} f \leq \min\left\{\frac{\beta\epsilon^2}{25}, \frac{\mu^2}{2\beta}\right\}.$$

Then Algorithm 4 initialized at x_0 generates iterates x_k that always lie in $B_\epsilon(\bar{x})$ and satisfy

$$\operatorname{dist}(x_{k+1}, \mathcal{X}^*) \leq \frac{\beta}{\mu} \cdot \operatorname{dist}^2(x_k, \mathcal{X}^*),$$
$$f(x_{k+1}) - \min_{\mathcal{X}} f \leq \frac{\beta}{\mu^2}\left(f(x_k) - \min_{\mathcal{X}} f\right)^2.$$

Moreover the iterates converge to some point $x_\infty \in \mathcal{X}^*$ at the quadratic rate

$$\|x_k - x_\infty\|_2 \leq \frac{2\sqrt{2}\mu}{\beta} \cdot \left(\frac{1}{2}\right)^{2^{k-1}} \qquad \text{for all } k \geq 0.$$

With the above generic results in hand, we can now derive the convergence guarantees for the subgradient and prox-linear methods for asymmetric low-rank matrix recovery problems. To summarize, the prox-linear method converges quadratically, as long as it is initialized within constant relative error of

the solution. The guarantees for the subgradient methods are less satisfactory: the size of the region of the linear convergence scales with the condition number of $M_\sharp$. The reason is that the proof estimates the region of convergence using the length of the iterate path, which scales with the condition number. The dependence on the condition number in general can be eliminated by introducing regularization $\|X^\top X - YY^\top\|_F$, as suggested in the work [154]. Still the results we present here are notable even for the subgradient method. For example, we see that for rank $r = 1$ instances satisfying RIP (e.g. blind deconvolution), the condition number of $M_\sharp$ is always one and therefore regularization is not required at all for subgradient and prox-linear methods.

Corollary 5.4.12 (Convergence guarantees under RIP (asymmetric))**.** *Suppose Assumptions 5.3.1 and 5.3.6 are valid*[6] *and consider the optimization problem*

$$\min_{X\in\mathbf{R}^{d_1\times r},\ Y\in\mathbf{R}^{r\times d_2}} f(X) = \frac{1}{m}\|\mathcal{A}(XY) - b\|_1.$$

Define $X_\sharp := U\sqrt{\Lambda}$ *and* $Y_\sharp = \sqrt{\Lambda}V^\top$*, where* $M_\sharp = U\Lambda V^\top$ *is any compact singular value decomposition of* $M_\sharp$*. Define also the condition number* $\chi := \sigma_1(M_\sharp)/\sigma_r(M_\sharp)$*. Then there exists* $\eta > 0$ *depending only on* κ_2*,* κ_3*, and* $\sigma(M_\sharp)$ *such that the following are true.*

1. **(Polyak subgradient)** *Algorithm 2 initialized at* (X_0, Y_0) *satisfying* $\frac{\|(X_0,Y_0)-(X_\sharp,Y_\sharp)\|_F}{\sqrt{\sigma_r(M_\sharp)}} \lesssim \min\{1, \frac{\kappa_3^2}{\kappa_2^2\chi}, \frac{\kappa_3}{\kappa_2}\}$*, will generate an iterate sequence that converges at the linear rate:*

$$\frac{\operatorname{dist}((X_k, Y_k), \mathcal{D}^*(M_\sharp))}{\sqrt{\sigma_r(M_\sharp)}} \le \delta \qquad \text{after} \qquad k \gtrsim \frac{\kappa_2^2\chi^2}{\kappa_3^2}\cdot\ln\left(\frac{\eta}{\delta}\right) \qquad \text{iterations.}$$

2. **(geometric subgradient)** *Algorithm 3 initialized at* (X_0, Y_0) *satisfying* $\frac{\|(X_0,Y_0)-(X_\sharp,Y_\sharp)\|_F}{\sqrt{\sigma_r(M_\sharp)}} \lesssim \min\{1, \frac{\kappa_3}{\kappa_2\chi}\}$*, will generate an iterate sequence that converges at the linear rate:*

$$\frac{\operatorname{dist}((X_k, Y_k), \mathcal{D}^*(M_\sharp))}{\sqrt{\sigma_r(M_\sharp)}} \le \delta \qquad \text{after} \qquad k \gtrsim \frac{\kappa_2^2\chi^2}{\kappa_3^2}\cdot\ln\left(\frac{\eta}{\delta}\right) \qquad \text{iterations.}$$

[6] with $|||\cdot||| = \frac{1}{m}\|\cdot\|_1$

3. **(prox-linear)** *Algorithm 4 initialized at* (X_0, Y_0) *satisfying* $\frac{f(x_0)-\min_{\mathcal{X}} f}{\sigma_r(M_\sharp)} \lesssim \min\{\kappa_2, \kappa_3^2/\kappa_2\}$ *and* $\frac{\|(X_0,Y_0)-(X_\sharp,Y_\sharp)\|_F}{\sqrt{\sigma_r(M_\sharp)}} \lesssim 1$*, will generate an iterate sequence that converges at the quadratic rate:*

$$\frac{\operatorname{dist}((X_k, Y_k), \mathcal{D}^*(M_\sharp))}{\sqrt{\sigma_r(M_\sharp)}} \lesssim \frac{\kappa_3}{\kappa_2} \cdot 2^{-2^k} \qquad \text{for all } k \geq 0.$$

5.5 Examples of ℓ_1/ℓ_2 RIP

In this section, we survey three matrix recovery problems from different fields, including physics, signal processing, control theory, wireless communications, and machine learning, among others. In all cases, the problems satisfy ℓ_1/ℓ_2 RIP and the $\mathcal{I}$-outlier bounds and consequently, the convergence results in Corollaries 5.4.6 and 5.4.12 immediately apply. Most of the RIP results in this section were previously known (albeit under more restrictive assumptions); we provide self-contained arguments in Section 5.10.2 for the sake of completeness. On the other hand, using nonsmooth optimization in these problems and the corresponding convergence guarantees based on RIP are, for the most part, new.

For the rest of this section we will assume the following data-generating mechanism.

Definition 5.5.1 (Data-generating mechanism). *A random linear map* $\mathcal{A}\colon \mathbf{R}^{d_1 \times d_2} \to \mathbf{R}^m$ *and a random index set* $\mathcal{I} \subset [m]$ *are drawn independently of each other. We assume moreover that the outlier frequency* $p_{\text{fail}} := |\mathcal{I}|/m$ *satisfies* $p_{\text{fail}} \in [0, 1/2)$ *almost surely. We then observe the corrupted measurements*

$$b_i = \begin{cases} \mathcal{A}(M_\sharp) & \text{if } i \notin \mathcal{I}, \text{ and} \\ \eta_i & \text{if } i \in \mathcal{I}, \end{cases} \tag{5.30}$$

where η is an arbitrary vector. In particular, η could be correlated with $\mathcal{A}$.

Throughout this section, we consider four distinct linear operators $\mathcal{A}$.

Matrix Sensing. In this scenario, measurements are generated as follows:

$$\mathcal{A}(M_\sharp)_i := \langle P_i, M_\sharp \rangle \qquad \text{for } i = 1, \ldots, m \tag{5.31}$$

where $P_i \in \mathbf{R}^{d_1 \times d_2}$ are fixed matrices.

Quadratic Sensing I . In this scenario, $M_\sharp \in \mathbf{R}^{d \times d}$ is assumed to be a PSD rank r matrix with factorization $M_\sharp = X_\sharp X_\sharp^\top$ and measurements are generated as follows:

$$\mathcal{A}(M_\sharp)_i = p_i^\top M_\sharp p_i = \|X_\sharp^\top p_i\|_2^2 \qquad \text{for } i = 1, \ldots, m, \tag{5.32}$$

where $p_i \in \mathbf{R}^d$ are fixed vectors.

Quadratic Sensing II . In this scenario, $M_\sharp \in \mathbf{R}^{d \times d}$ is assumed to be a PSD rank r matrix with factorization $M_\sharp = X_\sharp X_\sharp^\top$ and measurements are generated as follows:

$$\mathcal{A}(M_\sharp)_i = p_i^\top M_\sharp p_i - \tilde{p}_i^\top M_\sharp \tilde{p}_i = \|X_\sharp^\top p_i\|_2^2 - \|X_\sharp^\top \tilde{p}_i\|_2^2 \qquad \text{for } i = 1, \ldots, m, \tag{5.33}$$

where $p_i, \tilde{p}_i \in \mathbf{R}^d$ are fixed vectors.

Bilinear Sensing. In this scenario, $M_\sharp \in \mathbf{R}^{d_1 \times d_2}$ is assumed to be a r matrix with factorization $M_\sharp = XY$ and measurements are generated as follows:

$$\mathcal{A}(M_\sharp)_i = p_i^\top M_\sharp q_i \qquad \text{for } i = 1, \ldots, m, \tag{5.34}$$

where $p_i \in \mathbf{R}^{d_1}$ and $q_i \in \mathbf{R}^{d_2}$ are fixed vectors.

The matrix, quadratic, and bilinear sensing problems have been considered in a number of papers and in a variety of applications. The first theoretical properties for matrix sensing were discussed in [95, 209, 40]. Quadratic sensing in its full generality appeared in [52] and is a higher-rank generalization of the much older (real) phase retrieval problem [37, 41, 110]. Besides phase retrieval, quadratic sensing has applications to covariance sketching, shallow neural networks, and quantum state tomography; see for example [155] for a discussion. Bilinear sensing is a natural modification of quadratic sensing and is a higher-rank generalization of the blind deconvolution problem [8]; it was first proposed and studied in [33]. We will comeback to detailed study of the blind deconvolution problem in Chapter 6.

The reader is reminded that once ℓ_1/ℓ_2 RIP guarantees, in particular Assumptions 5.3.1 and 5.3.6, are established for the above four operators, the guarantees of Corollaries 5.4.6 and Corollary 5.4.12 immediately take hold for the problems

$$\min_{X \in \mathbf{R}^{d \times r}} \; f(X) = \frac{1}{m} \|\mathcal{A}(XX^\top) - b\|_1$$

and

$$\min_{X \in \mathbf{R}^{d_1 \times r},\; Y \in \mathbf{R}^{r \times d_2}} \; f(X) = \frac{1}{m} \|\mathcal{A}(XY) - b\|_1,$$

respectively. Thus, we turn our attention to establishing such guarantees.

5.5.1 Warm-up: ℓ_2/ℓ_2-RIP for matrix sensing

In this section, we are primarily interested in the ℓ_1/ℓ_2-RIP for the above four linear operators. However, as a warm-up, we first consider the ℓ_2/ℓ_2-RIP property for matrix sensing with Gaussian P_i. The following result appears in [209, 40].

Theorem 5.5.2 (ℓ_2/ℓ_2**-RIP for matrix sensing**). *For any $\delta \in (0,1)$ there exist constants $c, C > 0$ depending only on δ such that if the entries of P_i are i.i.d. standard Gaussian and $m \geq cr(d_1+d_2)\log(d_1 d_2)$, then with probability at least $1-\exp(-Cm)$, the estimate*

$$(1-\delta)\|M\|_F \leq \frac{1}{\sqrt{m}}\|\mathcal{A}(M)\|_2 \leq (1+\delta)\|M\|_F,$$

holds simultaneously for all $M \in \mathbf{R}^{d_1 \times d_2}$ of rank at most $2r$. Consequently, Assumption 5.3.1 is satisfied.

Following the general recipe of this Chapter, we see that the nonsmooth formulation

$$\min_{X \in \mathbf{R}^{d_1 \times r},\ Y \in \mathbf{R}^{r \times d_2}} \frac{1}{\sqrt{m}}\|\mathcal{A}(XY) - b\|_2 = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{Tr}(YP_i^\top X) - b_i\right)^2} \tag{5.35}$$

is immediately amenable to subgradient and prox-linear algorithms in the noiseless setting $\mathcal{I} = \emptyset$. In particular, a direct analogue of Corollary 5.4.12, which was stated for the penalty function $h = \frac{1}{m}\|\cdot\|_1$, holds; we omit the straightforward details.

5.5.2 The ℓ_1/ℓ_2-RIP and $\mathcal{I}$-outlier bounds

We now turn our attention to the ℓ_1/ℓ_2 RIP for more general classes of linear maps than the i.i.d. Gaussian matrices considered in Theorem 5.5.2. To establish

such guarantees, one must ensure that the linear maps $\mathcal{A}$ have light tails and are robustly injective on certain spaces of matrices. The first property leads to tight concentration results, while the second yields the existence of a lower RIP constant κ_1.

Assumption 5.5.3 (Matrix Sensing)**.** *The matrices $\{P_i\}$ are i.i.d. realizations of an η-sub-Gaussian random matrix*[7] *$P \in \mathbf{R}^{d_1 \times d_2}$. Furthermore, there exists a numerical constant $\alpha > 0$ such that*

$$\inf_{\substack{M:\ \mathrm{rank} M \leq 2r \\ \|M\|_F = 1}} \mathbb{E}|\langle P, M \rangle| \geq \alpha. \tag{5.36}$$

Assumption 5.5.4 (Quadratic Sensing I)**.** *The vectors $\{p_i\}$ are i.i.d. realizations of a η-sub-Gaussian random variable $p \in \mathbf{R}^d$. Furthermore, there exists a numerical constant $\alpha > 0$ such that*

$$\inf_{\substack{M \in \mathcal{S}^d:\ \mathrm{rank} M \leq 2r \\ \|M\|_F = 1}} \mathbb{E}|p^\top M p| \geq \alpha. \tag{5.37}$$

Assumption 5.5.5 (Quadratic Sensing II)**.** *The vectors $\{p_i\}, \{\tilde{p}_i\}$ are i.i.d. realizations of a η-sub-Gaussian random variable $p \in \mathbf{R}^d$. Furthermore, there exists a numerical constant $\alpha > 0$ such that*

$$\inf_{\substack{M \in \mathcal{S}^d:\ \mathrm{rank} M \leq 2r \\ \|M\|_F = 1}} \mathbb{E}|p^\top M p - \tilde{p}^\top M \tilde{p}| \geq \alpha. \tag{5.38}$$

Assumption 5.5.6 (Bilinear Sensing)**.** *The vectors $\{p_i\}$ and $\{q_i\}$ are i.i.d. realizations of η-sub-Gaussian random vectors $p \in \mathbf{R}^{d_1}$ and $q \in \mathbf{R}^{d_2}$, respectively. Furthermore, there exists a numerical constant $\alpha > 0$ such that*

$$\inf_{\substack{M:\ \mathrm{rank} M \leq 2r \\ \|M\|_F = 1}} \mathbb{E}|p^\top M q| \geq \alpha. \tag{5.39}$$

The Assumptions 5.5.3-5.5.6 are all valid for i.i.d. Gaussian realizations with

[7] By this we mean that the vectorized matrix $\mathbf{vec}(P)$ is a η-sub-gaussian random vector.

independent identity covariance, as the following lemma shows. We defer its proof to Section 5.10.2.

Lemma 5.5.7. *Assumption 5.5.3 holds for matrices P with i.i.d. standard Gaussian entries. Assumptions 5.5.4 and 5.5.5 hold for vectors $p, \tilde{p}$ with i.i.d. standard Gaussian entries. Assumption 5.5.6 holds for vectors p and q with i.i.d. standard Gaussian entries.*

We can now state the main RIP guarantees under the above assumptions. Throughout all the results, we fix the data generating mechanism as in Definition 5.5.1. Then, we wish to establish the inequalities

$$\kappa_1 \|M\|_F \leq \frac{1}{m} \|\mathcal{A}(M)\|_1 \leq \kappa_2 \|M\|_F \tag{5.40}$$

and

$$\kappa_3 \|M\|_F \leq \frac{1}{m} (\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1), \tag{5.41}$$

and, hence, Assumptions 5.3.1 and 5.3.6, respectively, for certain constants κ_1, κ_2, and κ_3. We defer the proof of this theorem to Section 5.10.2.

Theorem 5.5.8 (ℓ_1/ℓ_2 RIP and $\mathcal{I}$-outlier bounds)**.** *There exist numerical constants $c_1, \ldots, c_6 > 0$ depending only on α, η such that the following hold for the corresponding measurement operators described in Equations* (5.31)*,* (5.32)*,* (5.33)*, and* (5.34)*, respectively*

1. **(Matrix sensing)** *Suppose Assumption 5.5.3 holds. Then provided $m \geq \frac{c_1}{(1-2p_{\text{fail}})^2} r(d_1 + d_2 + 1) \ln\left(c_2 + \frac{c_2}{1-2p_{\text{fail}}}\right)$, we have with probability at least $1 - 4\exp\left(-c_3(1-2p_{\text{fail}})^2 m\right)$ that every matrix $M \in \mathbf{R}^{d_1 \times d_2}$ of rank at most $2r$ satisfies* (5.40) *and* (5.41) *with constants $\kappa_1 = c_4, \kappa_2 = c_5$ and $\kappa_3 = c_6(1 - 2p_{\text{fail}})$.*

2. **(Quadratic sensing I)** *Suppose Assumption 5.5.4 holds. Then provided* $m \geq \frac{c_1}{(1-2p_{\text{fail}})^2} r^2(2d+1)\ln\left(c_2 + \frac{c_2}{1-2p_{\text{fail}}}\sqrt{r}\right)$*, we have with probability at least* $1 - 4\exp\left(-c_3(1-2p_{\text{fail}})^2 m/r\right)$ *that every matrix* $M \in \mathbf{R}^{d\times d}$ *of rank at most* $2r$ *satisfies* (5.40) *and* (5.41) *with constants* $\kappa_1 = c_4, \kappa_2 = c_5 \cdot \sqrt{r}$ *and* $\kappa_3 = c_6(1-2p_{\text{fail}})$.

3. **(Quadratic sensing II)** *Suppose Assumption 5.5.5 holds. Then provided* $m \geq \frac{c_1}{(1-2p_{\text{fail}})^2} r(2d+1)\ln\left(c_2 + \frac{c_2}{1-2p_{\text{fail}}}\right)$*, we have with probability at least* $1 - 4\exp\left(-c_3(1-2p_{\text{fail}})^2 m\right)$ *that every matrix* $M \in \mathbf{R}^{d\times d}$ *of rank at most* $2r$ *satisfies* (5.40) *and* (5.41) *with constants* $\kappa_1 = c_4, \kappa_2 = c_5$ *and* $\kappa_3 = c_6(1-2p_{\text{fail}})$.

4. **(Bilinear sensing)** *Suppose Assumption 5.5.6 holds. Then provided* $m \geq \frac{c_1}{(1-2p_{\text{fail}})^2} r(d_1+d_2+1)\ln\left(c_2 + \frac{c_2}{1-2p_{\text{fail}}}\right)$*, we have with probability at least* $1 - 4\exp\left(-c_3(1-2p_{\text{fail}})^2 m\right)$ *that every matrix* $M \in \mathbf{R}^{d_1\times d_2}$ *of rank at most* $2r$ *satisfies* (5.40) *and* (5.41) *with constants* $\kappa_1 = c_4, \kappa_2 = c_5$ *and* $\kappa_3 = c_6(1-2p_{\text{fail}})$.

The guarantees of Theorem 5.5.8 were previously known under stronger assumptions. In particular, item (1) generalizes the results in [154] for the pure Gaussian setting. The case $r = 1$ of item (2) can be found, in a sightly different form, in [94, 89]. Item (3) sharpens slightly the analogous guarantee in [52] by weakening the assumptions on the moments of the measuring vectors to the uniform lower bound (5.38). Special cases of item (4) were established in [47], for the case $r = 1$, and [33], for Gaussian measurement vectors.

We note that all linear mappings require the same number of measurements in order to satisfy RIP and $\mathcal{I}$ outlier bounds, except for quadratic sensing I operator, which incurs an extra r-factor. This reveals the utility of the quadratic sensing II operator, which achieves optimal sample complexity. For larger scale problems, a shortcoming of matrix sensing operator (5.31) is that $m \cdot d_1 d_2$ scalars are required to represent the map $\mathcal{A}$. In contrast, all other measurement opera-

tors may be represented with only $m \cdot (d_1 + d_2)$ scalars.

5.6 Matrix Completion

In the previous sections, we saw that low-rank recovery problems satisfying RIP lead to well-conditioned nonsmooth formulations. We claim, however, that the general framework of sharpness and approximation is applicable even for problems without RIP. We consider two such problems, namely matrix completion in this section and robust PCA in Section 5.7, to follow. Both problems will be considered in the symmetric setting.

The goal of matrix completion problem is to recover a PSD rank r matrix $M_\sharp \in \mathcal{S}^d$ given access only to a subset of its entries. Henceforth, let $X_\sharp \in \mathbf{R}^{d\times r}$ be a matrix satisfying $M_\sharp = X_\sharp X_\sharp^\top$. Throughout, we assume incoherence condition, $\|X_\sharp\|_{2,\infty} \leq \sqrt{\frac{\nu r}{d}}$, for some $\nu > 0$. We also make the fairly strong assumption that the singular values of $X_\sharp$ are all equal $\sigma_1(X_\sharp) = \sigma_2(X_\sharp) = \ldots = \sigma_r(X_\sharp) = 1$. This assumption is needed for our theoretical results. We let $\Omega \subseteq [d]\times[d]$ be an index set generated by the Bernoulli model, that is, $\mathbb{P}((i,j),(j,i) \in \Omega) = p$ independently for all $1 \leq i \leq j \leq d$. Let $\Pi_\Omega \colon \mathcal{S}^d \to \mathbf{R}^{|\Omega|}$ be the projection onto the entries indexed by Ω. We consider the following optimization formulation of the problem

$$\min_{X\in\mathcal{X}} \; f(X) = \|\Pi_\Omega(XX^\top) - \Pi_\Omega(M_\sharp)\|_2 \quad \text{where } \mathcal{X} = \left\{X \in \mathbf{R}^{d\times r} \colon \; \|X\|_{2,\infty} \leq \sqrt{\frac{\nu r}{d}}\right\}.$$

We will show that both the Polyak subgradient method and an appropriately modified prox-linear algorithm converge linearly to the solution set under reasonable initialization. Moreover, we will see that the linear rate of convergence for the prox-linear method is much better than that for the subgradient method.

To simplify notation, we set

$$\mathcal{D}^* := \mathcal{D}^*(M_\sharp) = \{X \in \mathbf{R}^{d_1 \times r} : XX^\top = M_\sharp\}.$$

We begin by estimating the sharpness constant μ of the objective function. Fortunately, this estimate follows directly from inequalities (58) and (59a) in [53].

Lemma 5.6.1 (Sharpness [53])**.** *There are numerical constant $c_1, c_2 > 0$ such that the following holds. If $p \geq c_2(\frac{\nu^2 r^2}{d} + \frac{\log d}{d})$, then with probability $1 - c_1 d^{-2}$, the estimate*

$$\frac{1}{p}\|\Pi_\Omega(XX^\top - X_\sharp X_\sharp^\top)\|_F^2 \geq c_1 \|XX^\top - X_\sharp X_\sharp^\top\|_F^2$$

holds uniformly for all $X \in \mathcal{X}$ with $\operatorname{dist}(X, \mathcal{D}^*) \leq c_1$.

Let us next estimate the approximation accuracy $|f(Z) - f_X(Z)|$, where

$$f_X(Z) = \|\Pi_\Omega(XX - M_\sharp + X(Z - X)^\top + (Z - X)X^\top)\|_F.$$

To this end, we will require the following result.

Lemma 5.6.2 (Lemma 5 in [53])**.** *There is a numerical constant $c > 0$ such that the following holds. If $p \geq \frac{c}{\epsilon^2}(\frac{\nu^2 r^2}{d} + \frac{\log d}{d})$ for some $\epsilon \in (0, 1)$, then with probability at least $1 - 2d^{-4}$, the estimates*

1. $\frac{1}{\sqrt{p}}\|\Pi_\Omega(HH^\top)\|_F \leq \sqrt{(1+\epsilon)}\|H\|_F^2 + \sqrt{\epsilon}\|H\|_F$*; and*
2. $\frac{1}{\sqrt{p}}\|\Pi_\Omega(GH^\top)\|_F \leq \sqrt{\nu r}\|G\|_F$

hold uniformly for all matrices H with $\|H\|_{2,\infty} \leq 6\sqrt{\frac{\nu r}{d}}$ and $G \in \mathbf{R}^{d \times r}$.

An estimate of the approximation error $|f(Z) - f_X(Z)|$ is now immediate.

Lemma 5.6.3 (Approximation accuracy and Lipschitz continuity)**.** *There is a numerical constant $c > 0$ such that the following holds. If $p \geq \frac{c}{\epsilon^2}(\frac{\nu^2 r^2}{d} + \frac{\log d}{d})$ for some $\epsilon \in (0, 1)$, then with probability at least $1 - 2d^{-4}$, the estimates*

$$\frac{1}{\sqrt{p}}|f(X) - f_Y(X)| \leq \sqrt{(1+\epsilon)}\|X - Y\|_F^2 + \sqrt{\epsilon}\|X - Y\|_F,$$
$$|f(X) - f(Y)| \leq \sqrt{p\nu r}\|X - Y\|_F,$$

holds uniformly for all $X, Y \in \mathcal{X}$.

Proof. The first inequality follows immediately by observing the estimate

$$|f(X) - f_Y(X)| \leq \|\Pi_\Omega((X - Y)(X - Y)^\top)\|_F,$$

and using Lemma 5.6.2. To see the second inequality, observe

$$\begin{aligned}
|f(X) - f(Y)| &\leq \|\Pi_\Omega(XX^\top - YY^\top)\|_F \\
&= \frac{1}{2}\|\Pi_\Omega((X - Y)(X + Y)^\top - (X + Y)(X - Y)^\top)\|_F \\
&\leq \|\Pi_\Omega((X - Y)(X + Y)^\top)\|_F \\
&\leq \sqrt{p\nu r}\|X - Y\|_F,
\end{aligned}$$

where the last inequality follows by Part 2 of Lemma 5.6.2. □

Note that the approximation bound in Lemma 5.6.2 is not in terms of the square Euclidean norm. Therefore the results in Section 5.4 do not apply directly. Nonetheless, it is straightforward to modify the prox-linear method to take into account the new approximation bound. The proof of the following lemma appears in Section 5.10.3.

Lemma 5.6.4. *Suppose that Assumption 5.2.1 holds with the approximation property replaced by*

$$|f(y) - f_x(y)| \leq a\|y - x\|_2^2 + b\|y - x\|_2 \qquad \forall x, y \in \mathcal{X},$$

for some real $a, b \geq 0$. Consider the iterates generated by the process:

$$x_{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ f_{x_k}(x) + a\|x - x_k\|_2^2 + b\|x - x_k\|_2 \right\}.$$

Then as long as x_0 satisfies $\operatorname{dist}(x_0, \mathcal{X}^*) \leq \frac{\mu - 2b}{2a}$, the iterates converge linearly:

$$\operatorname{dist}(x_{k+1}, \mathcal{X}^*) \leq \frac{2(b + a\operatorname{dist}(x, \mathcal{X}^*))}{\mu} \cdot \operatorname{dist}(x_k, \mathcal{X}^*) \qquad \forall k \geq 0.$$

Combining Lemma 5.6.4 with our estimates of the sharpness and approximation accuracy, we deduce the following convergence guarantee for matrix completion.

Corollary 5.6.5 (Prox-linear method for matrix completion)**.** *There are numerical constants $c_0, c, C > 0$ such that the following holds. If $p \geq \frac{c}{\epsilon^2}(\frac{\nu^2 r^2}{d} + \frac{\log d}{d})$ for some $\epsilon \in (0, 1)$, then with probability at least $1 - c_0 d^{-2}$, the iterates generated by the modified prox-linear algorithm*

$$X_{k+1} = \arg\min_{X \in \mathcal{X}} \left\{ f_{X_k}(X) + \sqrt{p(1+\epsilon)} \cdot \|X - X_k\|_2^2 + \sqrt{p\epsilon}\|X - X_k\|_2 \right\} \tag{5.42}$$

satisfy

$$\operatorname{dist}(X_{k+1}, \mathcal{D}^*) \leq \frac{\sqrt{\epsilon} + \sqrt{1+\epsilon} \cdot \operatorname{dist}(X_k, \mathcal{D}^*)}{C} \cdot \operatorname{dist}(X_k, \mathcal{D}^*) \qquad \forall k \geq 0.$$

In particular, the iterates converge linearly as long as $\operatorname{dist}(X_0, \mathcal{D}^*) < \frac{C - 2\sqrt{\epsilon}}{2\sqrt{(1+\epsilon)}}$.

Proof. By invoking Proposition 6.2.1 and Lemmas 5.6.1 and 5.6.3 we may appeal to Lemma 5.6.4 with $a = \sqrt{p(1+\epsilon)}$, $b = \sqrt{p\epsilon}$, and $\mu = \sqrt{2c_1 p(\sqrt{2} - 1)}$. The result follows immediately. □

To summarize, there exist numerical constants $c_0, c_1, c_2, c_3 > 0$ such that the following is true with probability at least $1 - c_0 d^{-2}$. In the regime

$$p \geq \frac{c_2}{\epsilon^2}\left(\frac{\nu^2 r^2}{d} + \frac{\log d}{d}\right) \qquad \text{for some } \epsilon \in (0, c_1),$$

the prox-linear method will converge at the rapid linear rate,

$$\operatorname{dist}(X_k, \mathcal{D}^*) \leq \frac{c_2}{2^k},$$

when initialized at $X_0 \in \mathcal{X}$ satisfying $\operatorname{dist}(X_0, \mathcal{D}^*) < c_2$.

As for the prox-linear method, the results of Section 5.4 do not immediately yield convergence guarantees for the Polyak subgradient method. Nonetheless, it straightforward to show that the standard Polyak subgradient method still enjoys local linear convergence guarantees. The proof is a straightforward modification of the argument in [68, Theorem 3.1], and appears in Section 5.10.3.

Theorem 5.6.6. *Suppose that Assumption 5.2.1 holds with the approximation property replaced by*

$$|f(y) - f_x(y)| \leq a\|y - x\|_2^2 + b\|y - x\|_2 \qquad \forall x, y \in \mathcal{X},$$

for some real $a, b \geq 0$. Consider the iterates $\{x_k\}$ generated by the Polyak subgradient method in Algorithm 2. Then as long as the sharpness constant satisfies $\mu > 2b$ and x_0 satisfies $\operatorname{dist}(x_0, \mathcal{X}^*) \leq \gamma \cdot \frac{\mu - 2b}{2a}$ for some $\gamma < 1$, the iterates converge linearly

$$\operatorname{dist}^2(x_{k+1}, \mathcal{X}^*) \leq \left(1 - \frac{(1-\gamma)\mu(\mu - 2b)}{L^2}\right) \cdot \operatorname{dist}^2(x_k, \mathcal{X}^*) \qquad \forall k \geq 0.$$

Finally, combining Theorem 5.6.6 with our estimates of the sharpness and approximation accuracy, we deduce the following convergence guarantee for matrix completion.

Corollary 5.6.7 (Subgradient method for matrix completion)**.** *There are numerical constants $c_0, c, C > 0$ such that the following holds. If $p \geq \frac{c}{\epsilon^2}(\frac{\nu^2 r^2}{d} + \frac{\log d}{d})$ for some $\epsilon \in (0, 1)$, then with probability at least $1 - c_0 d^{-2}$, the iterates generated by the iterates $\{X_k\}$ generated by the Polyak Subgradient method in Algorithm 2 satisfy*

$$\operatorname{dist}(X_{k+1}, \mathcal{D}^*)^2 \leq \left(1 - \frac{C(C - 2\sqrt{\varepsilon})}{2\nu r}\right) \cdot \operatorname{dist}^2(X_k, \mathcal{D}^*) \qquad \forall k \geq 0.$$

In particular, the iterates converge linearly as long as $\operatorname{dist}(X_0, \mathcal{D}^*) < \frac{C-2\sqrt{\epsilon}}{4\sqrt{(1+\epsilon)}}$.

Proof. First, observe that we have the bound $L \leq \sqrt{pvr}$ by Lemma 5.6.3. By invoking Proposition 6.2.1 and Lemmas 5.6.1 and 5.6.3 we may appeal to Theorem 5.6.6 with $\gamma = 1/2$, $a = \sqrt{p(1+\epsilon)}$, $b = \sqrt{p\epsilon}$, and $\mu = \sqrt{2c_1 p(\sqrt{2}-1)}$. The result follows immediately. □

To summarize, there exist numerical constants $c_0, c_1, c_2, c_3 > 0$ such that the following is true with probability at least $1 - c_0 d^{-2}$. In the regime

$$p \geq \frac{c_2}{\epsilon^2}\left(\frac{\nu^2 r^2}{d} + \frac{\log d}{d}\right) \qquad \text{for some } \epsilon \in (0, c_1),$$

the Polyak subgradient method will converge at the linear rate,

$$\operatorname{dist}(X_k, \mathcal{D}^*) \leq \left(1 - \frac{c_3}{\nu r}\right)^{\frac{k}{2}} c_2,$$

when initialized at $X_0 \in \mathcal{X}$ satisfying $\operatorname{dist}(X_0, \mathcal{D}^*) < c_2$. Notice that the prox-linear method enjoys a much faster linear rate of convergence than the subgradient method—an observation fully supported by numerical experiments in Section 5.9. The caveat is that the per iteration cost of the prox-linear method is significantly higher than that of the subgradient method.

5.7 Robust PCA

The goal of robust PCA is to decompose a given matrix W into a sum of a low-rank matrix $M_\sharp$ and a sparse matrix $S_\sharp$, where $M_\sharp$ represents the principal components, $S_\sharp$ the corruption, and W the observed data [45, 38, 245]. In this section,

we explore methods of nonsmooth optimization for recovering such a decomposition, focusing on two different problem formulations. We only consider the symmetric version of the problem.

5.7.1 The Euclidean formulation

Setting the stage, we assume that the matrix $W \in \mathbf{R}^{d\times d}$ admits a decomposition $W = M_\sharp + S_\sharp$, where the matrices $M_\sharp$ and $S_\sharp$ satisfy the following for some parameters $\nu > 0$ and $k \in \mathbb{N}$:

1. The matrix $M_\sharp \in \mathbf{R}^{d\times d}$ has rank r and can be factored as $M_\sharp = X_\sharp X_\sharp^\top$ for some matrix $X_\sharp \in \mathbf{R}^{d\times r}$ satisfying $\|X_\sharp\|_{\mathrm{op}} \le 1$ and $\|X_\sharp\|_{2,\infty} \le \sqrt{\frac{\nu r}{d}}$.[8]
2. The matrix $S_\sharp$ is sparse in the sense that it has at most k nonzero entries per column/row.

The goal is to recover $M_\sharp$ and $S_\sharp$ given W. The first formulation we consider is the following:

$$\min_{X\in\mathcal{X}, S\in\mathcal{S}} F((X,S)) = \|XX^\top + S - W\|_F, \tag{5.43}$$

where the constraint sets are defined by

$$\mathcal{S} := \left\{S \in \mathbb{R}^{d\times d} : \|Se_i\|_1 \le \|S_\sharp e_i\|_1 \ \forall i\right\}, \qquad \mathcal{X} = \left\{X \in \mathbb{R}^{d\times r} : \|X\|_{2,\infty} \le \sqrt{\frac{\nu r}{d}}\right\}.$$

Note that the problem formulation requires knowing the ℓ_1 norms of the rows of $S_\sharp$. The same assumption was also made in [53, 103]. While admittedly unrealistic, this formulation provides a nice illustration of the paradigm we advocate here. The following technical lemma will be useful in proving the regularity conditions needed for rapid convergence. The proof is given in Section 5.10.4.

[8] Recall that $\|X\|_{2,\infty} = \max_{i\in[d]} \|X_{i\cdot}\|_2$ is the maximum row norm.

Lemma 5.7.1. *For all $X \in \mathcal{X}$ and $S \in \mathcal{S}$, the estimate holds:*

$$|\langle S - S_\sharp, XX^\top - X_\sharp X_\sharp^\top \rangle| \leq 10\sqrt{\frac{\nu r k}{d}} \cdot \|S - S_\sharp\|_F \cdot \|X - X_\sharp\|_F.$$

Equipped with the above lemma, we can estimate the sharpness and approximation parameters μ, ρ for the formulation (5.43).

Lemma 5.7.2 (Regularity constants)**.** *For all $X \in \mathcal{X}$ and $S \in \mathcal{S}$, the estimates hold:*

$$F((X,S))^2 \geq \left(\frac{1}{2}\sigma_r^2(X_\sharp) - 10\sqrt{\frac{\nu r k}{d}}\right) \cdot \left(\operatorname{dist}(X, \mathcal{D}^*(M_\sharp))^2 + \|S - S_\sharp\|_F^2\right) \tag{5.44}$$

and

$$|F((X,S)) - F_Y((X,S))| \leq \|X - Y\|_F^2. \tag{5.45}$$

Moreover, for any $X_1, X_2 \in \mathcal{X}$ and $S_1, S_2 \in \mathcal{S}$, the Lipschitz bounds holds:

$$|F((X_1,S_1)) - F((X_2,S_2))| \leq 2\sqrt{\nu r}\|X_1 - X_2\|_F + \|S_1 - S_2\|_F.$$

Proof. Let $X_\sharp \in \operatorname{proj}_{\mathcal{D}^*(M_\sharp)}(X)$. To establish the bound (5.44), we observe that

$$\begin{aligned}\|XX^\top + S - W\|_F^2 &= \|XX^\top - M_\sharp\|_F^2 + 2\langle S - S_\sharp, XX^\top - M_\sharp\rangle + \|S - S_\sharp\|_F^2 \\ &\geq \frac{1}{2}\sigma_r^2(X_\sharp)\|X - X_\sharp\|_F^2 - 20\sqrt{\frac{\nu r k}{d}}\|S - S_\sharp\|_F\|X - X_\sharp\|_F + \|S - S_\sharp\|_F^2,\end{aligned}$$

where the first inequality follows from Proposition 6.2.1 and Lemma 5.7.1. Now set

$$a := 10\sqrt{\frac{\nu r k}{d}}, \qquad b := \|X - X_\sharp\|_F, \qquad c := \|S - S_\sharp\|_F,$$

and $s := \frac{1}{2}\sigma_r^2(X_\sharp)$. With this notation, we apply the Fenchel-Young inequality to show that for any $\varepsilon > 0$, we have

$$2abc \leq a\varepsilon b^2 + (a/\varepsilon)c^2.$$

Thus, for any $\varepsilon > 0$, we have

$$\|XX^\top + S - W\|_F^2 \geq sb^2 - 2abc + c^2 \geq (s - a\varepsilon)b^2 + (1 - a/\varepsilon)c^2.$$

Now, let us choose $\varepsilon > 0$ so that $s - a\varepsilon = 1 - a/\varepsilon$. Namely set $\varepsilon = \frac{-(1-s)+\sqrt{(1-s)^2+4a^2}}{2a}$. With this choice of ε and the bound $s - a\varepsilon \geq \frac{1}{2}\sigma_r^2(X_\sharp) - 10\sqrt{\nu rk/d}$, the claimed bound (5.44) follows immediately. The bound (5.45) follows from the reverse triangle inequality:

$$\begin{aligned}
|F((X,S)) - F_Y((X,S))| &\leq \|XX^\top - YY^\top - (X-Y)Y^\top - Y^\top(X-Y)\|_F \\
&= \|XX^\top - XY^\top - YX^\top + YY^\top\|_F \\
&= \|(X-Y)(X-Y)^\top\|_F \\
&\leq \|X-Y\|_F^2.
\end{aligned}$$

Finally observe

$$\begin{aligned}
|F((X_1,S_1)) - F((X_2,S_2))| &\leq \|X_1X_1^\top - X_2X_2^\top\|_F + \|S_1 - S_2\|_F \\
&\leq \|X_1 + X_2\|_{\mathrm{op}}\|X_1 - X_2\|_F + \|S_1 - S_2\|_F \\
&\leq 2\sqrt{\nu r}\|X_1 - X_2\|_F + \|S_1 - S_2\|_F,
\end{aligned}$$

where we use the bound $\|X_i\|_{\mathrm{op}} \leq \sqrt{d}\|X_i\|_{2,\infty} \leq \sqrt{\nu r}$ in the final inequality. The proof is complete. □

To summarize, there exist numerical constants $c_0, c_1, c_2 > 0$ such that the following is true. In the regime

$$\sqrt{\frac{\nu rk}{d}} \leq c_0\sigma_r^2(X_\sharp),$$

the Polyak subgradient method will converge at the linear rate,

$$\mathrm{dist}(X_k, \mathcal{D}^*(M_\sharp)) \leq \left(1 - \frac{c_1\sigma_r^2(X_\sharp)}{\nu r}\right)^{\frac{k}{2}} \cdot c_2\mu,$$

and the prox-linear method will converge quadratically when initialized at $X_0 \in \mathcal{X}$ satisfying $\operatorname{dist}(X_0, \mathcal{D}^*(M_\sharp)) < c_2\sigma_r(X_\sharp)$.

5.7.2 The non-Euclidean formulation

We next turn to a different formulation for robust PCA that does not require knowledge of ℓ_1 row norms of $S_\sharp$. In particular, we consider the formulation

$$\min_{X \in \mathcal{X}} f(X) = \|XX^\top - W\|_1 \quad \text{where } \mathcal{X} = \{X \in \mathbf{R}^{d\times r} \mid \|X\|_{2,\infty} \le C\|X_\sharp\|_{2,\infty}\}, \tag{5.46}$$

for a constant $C > 1$. Unlike Section 5.7.1, here we consider a randomized model for the sparse matrix $S_\sharp$. We assume that there are real $\nu, \tau > 0$ such that

1. $M_\sharp \in \mathbf{R}^{d\times d}$ can be factored as $M_\sharp = X_\sharp X_\sharp^\top$ for some matrix $X_\sharp \in \mathbf{R}^{d\times r}$ satisfying $\|X_\sharp\|_{2,\infty} \le \sqrt{\frac{\nu r}{d}}\|X_\sharp\|_{\mathrm{op}}$.
2. We assume the random corruption model

   $$(S_\sharp)_{ij} = \delta_{ij}\hat{S}_{ij} \qquad \forall i, j$$

   where δ_{ij} are i.i.d. Bernoulli random variables with $\tau = \mathbb{P}(\delta_{ij} = 1)$ and $\hat{S}$ is an arbitrary and fixed $d \times d$ symmetric matrix.

In this setting, the approximation function at X is given by

$$f_X(Z) = \|XX - W + X(Z - X)^\top + (Z - X)X^\top\|_1.$$

We begin by computing an estimate of the approximation accuracy $|f(Z) - f_X(Z)|$.

Lemma 5.7.3 (Approximation accuracy)**.** *The estimate holds:*

$$|f(Z) - f_X(Z)| \le \|Z - X\|_{2,1}^2 \qquad \textit{for all } X, Z \in \mathbf{R}^{d\times r}.$$

Proof. As in the proof of Proposition 5.3.2, we compute

$$\begin{aligned}|f(Z) - f_X(Z)| &= \left| \|ZZ^\top - W\|_1 - \|XX - W + X(Z-X)^\top + (Z-X)X^\top\|_1 \right| \\ &\leq \|(Z-X)(Z-X)^\top\|_1 = \sum_{i,j} |e_i^\top (Z-X)(e_j^\top (Z-X))^\top| \\ &\leq \sum_{i,j} \|e_i^\top (Z-X)\|_2 \cdot \|e_j^\top (Z-X)\|_2 = \|Z-X\|_{2,1}^2,\end{aligned}$$

thereby completing the argument. □

Notice that the error $|f(Z) - f_X(Z)|$ is bounded in terms of the non-Euclidean norm $\|Z - X\|_{2,1}$. Thus, although in principle one may apply subgradient methods to the formulation (5.46), their convergence guarantees, which fundamentally relied on the Euclidean norm, would yield potentially overly pessimistic performance predictions. On the other hand, the convergence guarantees for the prox-linear method do not require the norm to be Euclidean. Indeed, the following is true, with a proof that is nearly identical as that of Theorem 5.4.11.

Theorem 5.7.4. *Suppose that Assumption 5.2.1 holds where $\|\cdot\|$ is replaced by an arbitrary norm $|||\cdot|||$. Choose any $\beta \geq \rho$ and set $\gamma := \rho/\beta$ in Algorithm 4. Then Algorithm 4 initialized at any point x_0 satisfying $\operatorname{dist}_{|||\cdot|||}(x_0, \mathcal{X}^*) < \frac{\mu}{\rho}$ converges quadratically:*

$$\operatorname{dist}_{|||\cdot|||}(x_{k+1}, \mathcal{X}^*) \leq \tfrac{\rho}{\mu} \cdot \operatorname{dist}^2_{|||\cdot|||}(x_k, \mathcal{X}^*) \qquad \forall k \geq 0.$$

To apply the above generic convergence guarantees for the prox-linear method, it remains to show that the objective function f in (5.46) is sharp relative to the norm $\|\cdot\|_{1,2}$. A key step in showing such a result is to prove that

$$\|XX^\top - X_\sharp X_\sharp^\top\|_1 \geq c \cdot \inf_{R^\top R = I} \|X - X_\sharp R\|_{2,1}$$

for a quantity c depending only on $X_\sharp$. One may prove this inequality using Proposition 6.2.1 together with the equivalence of the norms $\|\cdot\|_F$ and $\|\cdot\|_{1,2}$.

Doing so however leads to a dimension-dependent c, resulting in a poor rate of convergence and region of attraction. We instead seek to directly establish sharpness relative to the norm $\|\cdot\|_{2,1}$. In the rank one setting, this can be done using the following theorem.

Theorem 5.7.5 (Sharpness (rank one))**.** *Consider two vectors $x, \bar{x} \in \mathbf{R}^d$ satisfying*

$$\mathrm{dist}_{\|\cdot\|_1}(x, \{\pm\bar{x}\}) \leq (\sqrt{2} - 1)\|\bar{x}\|_1.$$

Then the estimate holds:

$$\|xx^\top - \bar{x}\bar{x}^\top\|_1 \geq (\sqrt{2} - 1) \cdot \|\bar{x}\|_1 \cdot \mathrm{dist}_{\|\cdot\|_1}(x, \{\pm\bar{x}\}).$$

The proof of this result appears in Section 5.10.4. We leave as an intriguing open question to determine if an analogous result holds in the higher rank setting.

Conjecture 5.7.6 (Sharpness (general rank))**.** *Fix a rank r matrix $X_\sharp \in \mathbf{R}^{d\times r}$ and set $\mathcal{D}^* := \{X \in \mathcal{X} : XX^\top = X_\sharp X_\sharp^\top\}$. Then there exist constants $c, \gamma > 0$ depending only on $X_\sharp$ such that the estimate holds:*

$$\|XX^\top - M\|_1 \geq c \cdot \mathrm{dist}_{\|\cdot\|_{2,1}}(X, \mathcal{D}^*),$$

for all $X \in \mathcal{X}$ satisfying $\mathrm{dist}_{\|\cdot\|_{2,1}}(X, \mathcal{D}^*) \leq \gamma$.

Assuming this conjecture, we can then show that the loss function f is sharp under the randomized corruption model. We first state the following technical lemma, whose proof is deferred to Section 5.10.4. In what what follows, given a matrix $X \in \mathbf{R}^{d\times r}$, the notation X_i always refers to the ith row of X.

Lemma 5.7.7. *Assume Conjecture 5.7.6. Then there exist constants $c_1, c_2, c_3 > 0$ so that for all d satisfying $d \geq \frac{c_1 \log d}{\tau}$, we have that with probability $1 - d^{-c_2}$, the following*

bound holds:

$$\sum_{i,j=1}^{d} \delta_{ij} |\langle X_i, X_j \rangle - \langle (X_\sharp)_i, (X_\sharp)_j \rangle| \leq \left(\tau + \frac{c_3 C \sqrt{\tau \nu r \log d}}{c} \|X_\sharp\|_{\text{op}} \right) \|XX^\top - X_\sharp X_\sharp^\top\|_1$$

for all $X \in \mathcal{X}$ *satisfying* $\text{dist}_{\|\cdot\|_{2,1}}(X, \mathcal{D}^*) \leq \gamma$.

We remark that we expect c to scale with $\|X_\sharp\|_{\text{op}}$ in the above bound, yielding a ratio $\|X_\sharp\|_{\text{op}}/c$ dependent on the conditioning of $X_\sharp$. Given the above lemma, sharpness of f quickly follows.

Lemma 5.7.8 (Sharpness of Non-Euclidean Robust PCA)**.** *Assume Conjecture 5.7.6. Then there exists a constants* $c_1, c_2, c_3 > 0$ *so that for all* d *satisfying* $d \geq \frac{c_1 \log d}{\tau}$*, we have that with probability* $1 - d^{-c_2}$*, the following bound holds:*

$$f(X) - f(X_\sharp) \geq c \cdot \left(1 - 2\tau - \frac{2c_3 C \sqrt{\tau \nu r \log d}}{c} \|X_\sharp\|_{op} \right) \cdot \text{dist}_{\|\cdot\|_{2,1}}(X, \mathcal{D}^*(M_\sharp))$$

for all $X \in \mathcal{X}$ *satisfying and* $\text{dist}_{\|\cdot\|_{2,1}}(X, \mathcal{D}^*(M_\sharp)) \leq \gamma$.

Proof. The reverse triangle inequality implies that

$$\begin{aligned}
&f(X) - f(X_\sharp) \\
&= \|XX^\top - W\|_1 - f(X_\sharp) \\
&= \|XX^\top - X_\sharp X_\sharp^\top\|_1 \\
&\quad + \sum_{i,j=1}^{d} \delta_{ij} \left(|\langle X_i, X_j \rangle - \langle (X_\sharp)_i, (X_\sharp)_j \rangle - (S_\sharp)_{ij}| - |\langle X_i, X_j \rangle - \langle (X_\sharp)_i, (X_\sharp)_j \rangle| \right) - f(X_\sharp) \\
&= \|XX^\top - X_\sharp X_\sharp^\top\|_1 \\
&\quad + \sum_{i,j=1}^{d} \delta_{ij} \left(|\langle X_i, X_j \rangle - \langle (X_\sharp)_i, (X_\sharp)_j \rangle - (S_\sharp)_{ij}| - |\langle X_i, X_j \rangle - \langle (X_\sharp)_i, (X_\sharp)_j \rangle| - |(S_\sharp)_{ij}| \right) \\
&\geq \|XX^\top - X_\sharp X_\sharp^\top\|_1 - 2 \sum_{i,j=1}^{d} \delta_{ij} |\langle X_i, X_j \rangle - \langle (X_\sharp)_i, (X_\sharp)_j \rangle|.
\end{aligned}$$

The result them follows from the the sharpness of the function $\|XX^\top - X_\sharp X_\sharp^\top\|_1$ together with Lemma 5.7.7. □

Combining Lemma 5.7.8 and Theorem 5.7.4, we deduce the following convergence guarantee.

Theorem 5.7.9 (Convergence for non-Euclidean Robust PCA)**.** *Assume Conjecture 5.7.6. Then there exist constants $c_1, c_2, c_3 > 0$ so that for all τ satisfying $1 - 2\tau - 2c_3 C\sqrt{\tau v r \log d}\|X_\sharp\|_{op}/c > 0$ and d satisfying $d \geq \frac{c_1 \log d}{\tau}$, we have that with probability $1 - d^{-c_2}$, the iterates generated by the prox-linear algorithm*

$$X_{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ f_{X_k}(X) + \frac{1}{2\gamma}\|X - X_k\|_{2,1}^2 \right\} \tag{5.47}$$

satisfy

$$\mathrm{dist}_{\|\cdot\|_{2,1}}(X_{k+1}, \mathcal{D}^*(M_\sharp)) \leq \frac{2}{c \cdot \left(1 - 2\tau - \frac{2c_3 C\sqrt{\tau v r \log d}}{c}\|X_\sharp\|_{op}\right)} \cdot \mathrm{dist}^2_{\|\cdot\|_{2,1}}(X_k, \mathcal{D}^*(M_\sharp)),\ \forall k \geq 0.$$

In particular, the iterates converge quadratically as long as the initial iterate $X_0 \in \mathcal{X}$ satisfies

$$\mathrm{dist}_{\|\cdot\|_{2,1}}(X_0, \mathcal{D}^*(M_\sharp)) < \min\left\{ (1/2)c \cdot \left(1 - 2\tau - \frac{2c_3 C\sqrt{\tau v r \log d}}{c}\|X_\sharp\|_{op}\right), \gamma \right\}.$$

5.8 Recovery up to a tolerance

Thus far, we have developed exact recovery guarantees under noiseless or sparsely corrupted measurements. We showed that sharpness together with weak convexity imply rapid local convergence of numerical methods under these settings. In practical scenarios, however, it might be unlikely that any, let alone a constant fraction of measurements, are perfectly observed. Instead, a more realistic model incorporates additive errors that are the sum of a sparse, but otherwise arbitrary vector and a dense vector with relatively small norm.

Exact recovery is in general not possible under this noise model. Instead, we should only expect to recover the signal up to an error.

To develop algorithms for this scenario, we need only observe that the previously developed sharpness results all yield a corresponding "sharpness up to a tolerance" result. Indeed, all problems considered thus far, are convex composite and sharp:

$$\min_{x \in \mathcal{X}} f(x) := h(F(x)) \qquad \text{and} \qquad f(x) - \inf_{\mathcal{X}} f \geq \mu \cdot \operatorname{dist}(x, \mathcal{X}^*),$$

where h is convex and η-Lipschitz with respect to some norm $\|\cdot\|$, F is a smooth map, and $\mu > 0$. Now consider a fixed additive error vector e, and the perturbed problem

$$\min_{x \in \mathcal{X}} \tilde{f}(x) := h(F(x) + e). \tag{5.48}$$

The triangle inequality immediately implies that the perturbed problem is sharp up to tolerance $2\eta\|e\|$:

$$\tilde{f}(x) - \inf_{\mathcal{X}} \tilde{f} \geq \mu \cdot \operatorname{dist}(x, \mathcal{X}^*) - 2\eta\|e\| \qquad \forall x \in \mathcal{X}.$$

In particular, any minimizer x^* of $\tilde{f}$ satisfies

$$\operatorname{dist}(x^*, \mathcal{X}^*) \leq (2\eta/\mu)\|e\|, \tag{5.49}$$

where as before we set $\mathcal{X}^* = \arg\min_{\mathcal{X}} f$. In this section, we show that subgradient and prox-linear algorithms applied to the perturbed problem (5.48) converge rapidly up to a tolerance on the order of $\eta\|e\|/\mu$. To see the generality of the above approach, we note that even the robust recovery problems considered in Section 5.3.2, in which a constant fraction of measurements are already corrupted, may be further corrupted through additive error vector e. We will study this problem in detail in Section 5.8.1.

Throughout the rest of the section, let us define the noise level:

$$\varepsilon := \eta |\!|\!| e |\!|\!|.$$

Mirroring the discussion in Section 5.4, define the annulus:

$$\widetilde{\mathcal{T}}_\gamma := \left\{ z \in \mathcal{X} : \frac{14\varepsilon}{\mu} < \operatorname{dist}(z, \mathcal{X}^*) < \frac{\gamma\mu}{4\rho} \right\}, \tag{5.50}$$

for some $\gamma > 0$. Note that for the annulus $\widetilde{\mathcal{T}}_\gamma$ to be nonempty, we must ensure $\epsilon < \frac{\mu^2\gamma}{56\rho}$. We will see that $\widetilde{\mathcal{T}}_\gamma$ serves as a region of rapid convergence for some numerical constant γ. As before, we also define subgradient bound and the condition measure:

$$\tilde{L} := \sup\{\|\zeta\|_2 : \zeta \in \partial \tilde{f}(x), x \in \widetilde{\mathcal{T}}_1\} \qquad \text{and} \qquad \tilde{\tau} := \mu/\tilde{L}.$$

In all examples considered in this Chapter, it is possible to show directly that $\tilde{L} \leq L$ as defined in Assumption 5.4.1. A similar result is true in the general case, as well. Indeed, the following Lemma provides a bound for $\tilde{L}$ in terms of the subgradients of f on a slight expansion of the tube $\mathcal{T}_1$ from (5.26); the proof appears in Section 5.10.5.

Lemma 5.8.1. *Suppose $\epsilon < \frac{\mu^2}{56\rho}$ so that $\widetilde{\mathcal{T}}_1$ is nonempty. Then the following bound holds:*

$$\tilde{L} \leq \sup\left\{ \|\zeta\|_2 : \zeta \in \partial f(x), \operatorname{dist}(x, \mathcal{X}^*) \leq \frac{\mu}{\rho}, \operatorname{dist}(x, \mathcal{X}) \leq 2\sqrt{\frac{\varepsilon}{\rho}} \right\} + 2\sqrt{8\rho\varepsilon}.$$

We will now design algorithms whose basin of attraction is the annulus $\widetilde{\mathcal{T}}_\gamma$ for some γ. To that end, the following modified sharpness bound will be useful for us. The reader should be careful to note the appearance of $\inf_{\mathcal{X}} f$, not $\inf_{\mathcal{X}} \tilde{f}$ in the following bound.

Lemma 5.8.2 (Approximate sharpness)**.** *We have the following bound:*

$$\tilde{f}(x) - \inf_{\mathcal{X}} f \geq \mu \cdot \operatorname{dist}(x, \mathcal{X}^*) - \varepsilon \qquad \forall x \in \mathcal{X}. \tag{5.51}$$

Proof. For any $x \in \mathcal{X}$, observe $\tilde{f}(x) - \inf f \geq f(x) - \inf f - \varepsilon \geq \mu \cdot \operatorname{dist}(x, \mathcal{X}^*) - \varepsilon$, as claimed. □

Next, we show that $\tilde{f}$ satisfies the following approximate subgradient inequality.

Lemma 5.8.3 (Approximate subgradient inequality)**.** *The following bound holds:*

$$f(y) \geq \tilde{f}(x) + \langle \zeta, y - x \rangle - \frac{\rho}{2}\|x - y\|^2 - 3\varepsilon \qquad \forall x, y \text{ and } \zeta \in \partial \tilde{f}(x).$$

Proof. First notice that $|f_x(y) - \tilde{f}_x(y)| \leq \varepsilon$ for all x, y. Furthermore, we have $\partial \tilde{f}(x) = \nabla F(x)^* \partial h(F(x) + e) = \partial \tilde{f}_x(x)$. Therefore, it follows that for any $\zeta \in \partial \tilde{f}_x(x)$ we have

$$\begin{aligned}
\langle \zeta, y - x \rangle &\leq \tilde{f}_x(y) - \tilde{f}_x(x) \\
&\leq f_x(y) - f_x(x) + 2\eta |\!|\!| e |\!|\!| \\
&\leq f(y) - f(x) + \frac{\rho}{2}\|x - y\|^2 + 2\varepsilon \\
&\leq f(y) - \tilde{f}(x) + \frac{\rho}{2}\|x - y\|^2 + 3\varepsilon,
\end{aligned}$$

as desired. □

Now consider the following modified Polyak method. It is important to note that the stepsize assumes knowledge of $\min_{\mathcal{X}} f$ rather than $\min_{\mathcal{X}} \tilde{f}$. This distinction is important because it often happens that $\min_{\mathcal{X}} f = 0$, whereas $\min_{\mathcal{X}} \tilde{f}$ is in general unknown; for example, consider any noiseless problem analyzed thus far. We note that the standard Polyak subgradient method may also be applied to $\tilde{f}$ without any changes and has similar theoretical guarantees. The proof appears in Section 5.10.5.

Algorithm 5: Modified Polyak Subgradient Method

t **Data:** $x_0 \in \mathbf{R}^d$

Step k**:** $(k \geq 0)$

Choose $\zeta_k \in \partial \tilde{f}(x_k)$. **If** $\zeta_k = 0$, then exit algorithm.

Set $x_{k+1} = \mathrm{proj}_{\mathcal{X}}\left(x_k - \dfrac{\tilde{f}(x_k) - \min_{\mathcal{X}} f}{\|\zeta_k\|^2}\zeta_k\right)$.

Theorem 5.8.4 (Polyak subgradient method)**.** *Suppose that Assumption 5.4.1 holds and suppose that* $\varepsilon \leq \mu^2/14\rho$*. Then Algorithm 5 initialized at any point* $x_0 \in \widetilde{\mathcal{T}}_1$ *produces iterates that converge Q-linearly to* $\mathcal{X}^*$ *up to tolerance* $14\varepsilon/\mu$*, that is*

$$\mathrm{dist}^2(x_{k+1}, \mathcal{X}^*) \leq \left(1 - \frac{13\tilde{\tau}^2}{56}\right)\mathrm{dist}^2(x_k, \mathcal{X}^*) \qquad \forall k \geq 0 \text{ with } \mathrm{dist}(x_k, \mathcal{X}^*) \geq 14\varepsilon/\mu.$$

Next we provide theoretical guarantees for Algorithm 5.4.4, where one does not know the optimal value $\min_{\mathcal{X}} f$. The proof of this result is a straightforward modification of [68, Theorem 6.1] based on the Lemmas 5.8.2 and 5.8.3, and therefore we omit it.

Theorem 5.8.5 (Geometrically decaying subgradient method)**.** *Suppose that Assumption 5.4.1 holds, fix a real number* $\gamma \in (0, 1)$*, and suppose* $\tilde{\tau} \leq \frac{14}{11}\sqrt{\frac{1}{2-\gamma}}$*. Suppose also* $\epsilon < \frac{\mu^2\gamma}{56\rho}$ *so that* $\widetilde{\mathcal{T}}_\gamma$ *is nonempty. Set* $\lambda := \frac{\gamma\mu^2}{4\rho\tilde{L}}$ *and* $q := \sqrt{1 - (1-\gamma)\tilde{\tau}^2}$ *in Algorithm 3. Then the iterates* x_k *generated by Algorithm 3 on the perturbed problem* (5.48)*, initialized at a point* $x_0 \in \tilde{\mathcal{T}}_\gamma$*, satisfy:*

$$\mathrm{dist}^2(x_k; \mathcal{X}^*) \leq \frac{\gamma^2\mu^2}{16\rho^2}\left(1 - (1-\gamma)\tilde{\tau}^2\right)^k \qquad \forall k \geq 0 \text{ with } \mathrm{dist}(x_k, \mathcal{X}^*) \geq 14\varepsilon/\mu.$$

Finally, we analyze the prox-linear algorithm applied to the problem $\min_{\mathcal{X}} \tilde{f}$. In contrast to the Polyak method, one does not need to know the optimal value $\min_{\mathcal{X}} f$. The proof appears in Section 5.10.5.

Theorem 5.8.6 (Prox-linear algorithm)**.** *Suppose Assumptions 5.2.1 holds. Choose any* $\beta \geq \rho$ *in Algorithm 4 applied to the perturbed problem* (5.48) *and set* $\gamma := \rho/\beta$*.*

Suppose moreover $\epsilon < \frac{\mu^2\gamma}{56\rho}$ *so that* $\tilde{\mathcal{T}}_\gamma$ *is nonempty. Then Algorithm 4 initialized at any point* $x_0 \in \widetilde{\mathcal{T}}_\gamma$ *converges quadratically up to tolerance* $14\varepsilon/\mu$*:*

$$\operatorname{dist}(x_{k+1}, \mathcal{X}^*) \leq \tfrac{7\beta}{6\mu} \cdot \operatorname{dist}^2(x_k, \mathcal{X}^*) \qquad \forall k \geq 0 \text{ with } \operatorname{dist}(x_{k+1}, \mathcal{X}^*) \geq 14\varepsilon/\mu.$$

5.8.1 Example: sparse outliers and dense noise under ℓ_1/ℓ_2 RIP

To further illustrate the ideas of this section, we now generalize the results of Section 5.3.2, in particular Assumption 5.3.6, to the following observation model.

Assumption 5.8.7 ($\mathcal{I}$-outlier bounds)**.** *There exists vectors* $e, \Delta \in \mathbf{R}^m$*, a set* $\mathcal{I} \subset \{1, \ldots, m\}$*, and a constant* $\kappa_3 > 0$ *such that the following hold.*

(C1) $b = \mathcal{A}(M_\sharp) + \Delta + e$.

(C2) *Equality holds* $\Delta_i = 0$ *for all* $i \notin \mathcal{I}$.

(C3) *For all matrices* W *of rank at most* $2r$*, we have*

$$\kappa_3 \|W\|_F \leq \frac{1}{m}\|\mathcal{A}_{\mathcal{I}^c}(W)\|_1 - \frac{1}{m}\|\mathcal{A}_{\mathcal{I}}(W)\|_1.$$

Given these assumptions we follow the notation of the previous section and let

$$f(X) := \frac{1}{m}\|\mathcal{A}(XX^\top - M_\sharp) - \Delta\|_1 \qquad \text{and} \qquad \tilde{f}(X) = \frac{1}{m}\|\mathcal{A}(XX^\top - M_\sharp) - \Delta - e\|_1.$$

Then we have the following proposition:

Proposition 5.8.8. *Suppose Assumption 5.3.1 and 5.8.7 are valid. Then the following hold:*

1. **(Sharpness)** *We have*

$$f(X) - f(X_\sharp) \geq \mu \cdot \operatorname{dist}(X, \mathcal{D}^*(M_\sharp)) \qquad \text{for all } X \in \mathbf{R}^{d\times r} \text{ and } \mu := \kappa_3 \sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp),$$

2. **(Weak Convexity)** *The function f is $\rho := 2\kappa_2$-weakly convex.*

3. **(Minimizers)** *All minimizers of $\tilde{f}$ satisfy*

$$\operatorname{dist}(X^*, \mathcal{X}^*) \leq \frac{2\frac{1}{m}\|e\|_1}{\kappa_3 \sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp)} \qquad \forall X^* \in \arg\min_{\mathcal{X}} \tilde{f}.$$

4. **(Lipschitz Bound)** *We have the bound*

$$\tilde{L} \leq 2\kappa_2 \cdot \left(\frac{\kappa_3 \sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp)}{8\kappa_2} + \sigma_1(X_\sharp) \right).$$

Proof. Sharpness follows from Proposition 6.2.2, while weak convexity follows from Proposition 5.3.2. The minimizer bound follows from (5.49). Finally, due to Lemma 5.2.2, the argument given in Proposition (5.3.2), but applied instead to $\tilde{f}$, guarantees that

$$\tilde{L} \leq 2\kappa_2 \cdot \sup \left\{ \|X\|_{op} \colon \operatorname{dist}(X, \mathcal{D}^*(M_\sharp)) \leq \frac{\kappa_3 \sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp)}{8\kappa_2} \right\}.$$

In turn the supremum may be bounded as follows: Let $X_\star = X_\sharp R$ denote the closest point to X in $\mathcal{D}^*(M)$. Then

$$\|X\|_{\text{op}} \leq \|X - X_\sharp R\|_{\text{op}} + \|X_\sharp R\|_{\text{op}} \leq \frac{\kappa_3 \sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp)}{8\kappa_2} + \sigma_1(X_\sharp),$$

as desired. □

In particular, combining Proposition 5.8.8 with the previous results in this section, we deduce the following. As long as the noise satisfies

$$\frac{1}{m}\|e\|_1 \leq \frac{c_0\kappa_3^2\sigma_r(M_\sharp)}{\kappa_2}$$

for a sufficiently small constant $c_0 > 0$, the subgradient and prox-linear methods converge rapidly to within tolerance

$$\delta \approx \frac{\frac{1}{m}\|e\|_1}{\kappa_3 \sigma_r(X_\sharp)},$$

when initialized at a matrix X_0 satisfying

$$\frac{\operatorname{dist}(X_0, \mathcal{D}^*(M_\sharp))}{\sqrt{\sigma_r(M_\sharp)}} \leq c_1 \cdot \frac{\kappa_3}{\kappa_2},$$

for some small constant c_1. The formal statement is summarized in the following corollary.

Corollary 5.8.9 (Convergence guarantees under RIP with sparse outliers and dense noise (symmetric))**.** *Suppose Assumptions 5.3.1 is and 5.8.7 are valid with* $|||\cdot||| = \frac{1}{m}\|\cdot\|_1$ *and define the condition number* $\chi = \sigma_1(M_\sharp)/\sigma_r(M_\sharp)$*. Then there exists numerical constants* $c_0, c_1, c_2, c_3, c_4, c_5, c_6 > 0$ *such that the following hold. Suppose the noise level satisfies*

$$\frac{1}{m}\|e\|_1 \leq \frac{2(\sqrt{2}-1)c_0\kappa_3^2\sigma_r(M_\sharp)}{28\kappa_2}$$

and define the tolerance

$$\delta = \frac{\frac{14}{m}\|e\|_1}{\kappa_3\sqrt{2(\sqrt{2}-1)\sigma_r(M_\sharp)}}.$$

Then as long as the matrix X_0 *satisfies*

$$\frac{\operatorname{dist}(X_0, \mathcal{D}^*(M_\sharp))}{\sqrt{\sigma_r(M_\sharp)}} \leq c_1 \cdot \frac{\kappa_3}{\kappa_2},$$

the following are true.

1. **(Polyak subgradient)** *Algorithm 2 initialized at* X_0 *produces iterates that converge linearly to* $\mathcal{D}^*(M_\sharp)$*, that is*

$$\frac{\operatorname{dist}^2(X_k, \mathcal{D}^*(M_\sharp))}{\sigma_r(M_\sharp)} \leq \left(1 - \frac{c_2}{1 + \frac{c_3\kappa_2^2\chi}{\kappa_3^2}}\right)^k \cdot \frac{c_4\kappa_3^2}{\kappa_2^2} \qquad \forall k \geq 0 \text{ with } \operatorname{dist}(X_k, \mathcal{X}^*) \geq \delta.$$

2. **(geometric subgradient)** *Algorithm 3 with* $\lambda = \frac{c_5\kappa_3^2\sqrt{\sigma_r(M_\sharp)}}{\kappa_2(\kappa_3+2\kappa_2\sqrt{\chi})}$, $q = \sqrt{1-\frac{c_2}{1+c_3\kappa_2^2\chi/\kappa_3^2}}$ *and initialized at* X_0 *converges linearly:*

$$\frac{\mathrm{dist}^2(X_k, \mathcal{D}^*(M_\sharp))}{\sigma_r(M_\sharp)} \leq \left(1 - \frac{c_2}{1+\frac{c_3\kappa_2^2\chi}{\kappa_3^2}}\right)^k \cdot \frac{c_4\kappa_3^2}{\kappa_2^2} \qquad \forall k \geq 0 \text{ with } \mathrm{dist}(X_k, \mathcal{X}^*) \geq \delta.$$

3. **(prox-linear)** *Algorithm 4 with* $\beta = \rho$ *and initialized at* X_0 *converges quadratically:*

$$\frac{\mathrm{dist}(X_k, \mathcal{D}^*(M_\sharp)))}{\sqrt{\sigma_r(M_\sharp)}} \leq 2^{-2^k} \cdot \frac{c_6\kappa_3}{\kappa_2} \qquad \forall k \geq 0 \text{ with } \mathrm{dist}(X_k, \mathcal{X}^*) \geq \delta.$$

5.9 Numerical experiments

In this section, we demonstrate the theory and algorithms developed in the previous sections on a number of low-rank matrix recovery problems, namely quadratic and bilinear sensing, low rank matrix completion, and robust PCA. All experiments were performed using the programming language `Julia` [22]. The code used for these experiments can be found in the github repository: `https://github.com/COR-OPT/CompOpt-LowRankMatrixRecovery`.

5.9.1 Robustness to outliers

In our first set of experiments, we empirically test the robustness of our optimization methods to outlying measurements. We generate *phase transition plots*, where each pixel corresponds to the empirical probability of successful recovery over 50 test runs using randomly generated problem instances. Brighter pixels represent higher recovery rates. All generated instances obey the following:

1. The initial estimate is specified reasonably close to the ground truth. In particular, given a target symmetric positive semidefinite matrix $X_\sharp$, we set

$$X_0 := X_\sharp + \delta \cdot \left\|X_\sharp\right\|_F \cdot \Delta, \quad \text{where } \Delta = \frac{G}{\|G\|_F},\ G_{ij} \sim_{\text{i.i.d.}} N(0, I).$$

   Here, δ is a scalar that controls the quality of initialization and Δ is a random unit "direction". The asymmetric setting is completely analogous.

2. When using the subgradient method with geometrically decreasing stepsize, we set $\lambda = 1.0,\ q = 0.98$.

3. For the quadratic sensing, bilinear sensing, and matrix completion problems, we mark a test run as a success when the normalized distance $\|M - M_\sharp\|_F / \|M_\sharp\|_F$ is less than 10^{-5}. Here we set $M = XX^\top$ in the symmetric setting and $M = XY$ in the asymmetric setting. For the robust PCA problem, we stop when $\|M - M_\sharp\|_1 / \|M_\sharp\|_1 < 10^{-5}$.

Moreover, we set the seed of the random number generator at the beginning of each batch of experiments to enable reproducibility.

Quadratic and Bilinear sensing. Figures 5.2 and 5.3 depict the phase transition plots for bilinear (5.34) and symmetrized quadratic (5.33) sensing formulations using Gaussian measurement vectors. In the experiments, we corrupt a fraction of measurements with additive Gaussian noise of unit entrywise variance. Empirically, we observe that increasing the variance of the additive noise does not affect recovery rates. Both problems exhibit a sharp phase transition at very similar scales. Moreover, increasing the rank of the generating signal does not seem to dramatically affect the recovery rate for either problem. Under additive noise, we can recover the true signal (up to natural ambiguity) even if we

corrupt as much as half of the measurements.

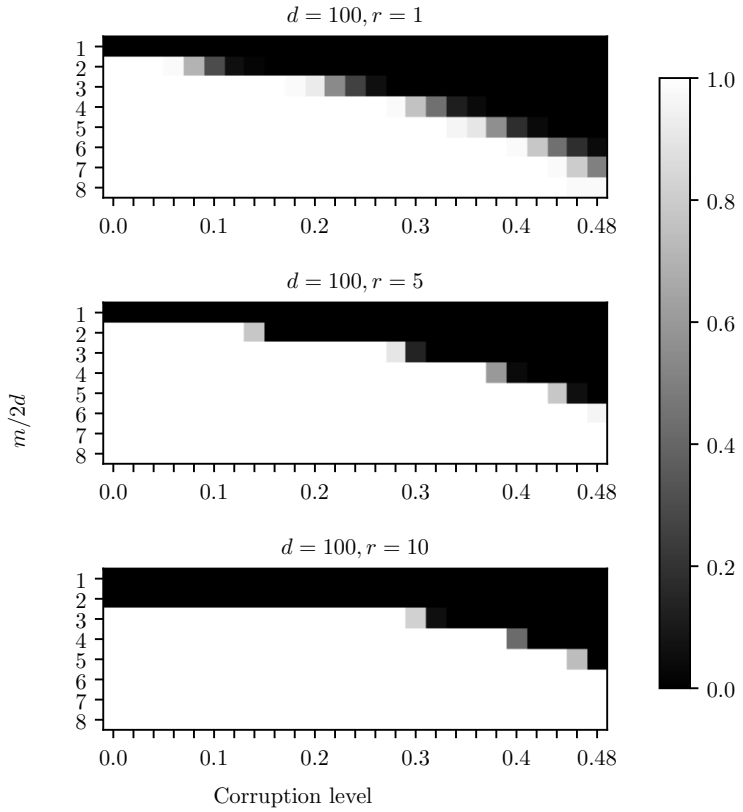


Figure 5.2: Bilinear sensing with $d_1 = d_2 = d = 100$ using Algorithm 3.

Robust PCA. We generate robust PCA instances for $d = 80$ and $r \in \{1, 2, 4, 8, 16\}$. The corruption matrix $S_\sharp$ follows the assumptions in Section 5.7.2, where for simplicity we set $\hat{S}_{ij} \sim \mathsf{N}(0, \sigma^2)$. We observed that increasing or decreasing the variance σ^2 did not affect the probability of successful recovery, so our experiments use $\sigma = 1$. We use the subgradient method, Algorithm 4, and the prox-linear algorithm (5.47). Notice that we have not presented any guarantees for the subgradient method on this problem, in contrast to the prox-linear method. The subproblems for the prox-linear method are solved by ADMM with graph splitting as in [200]. We set tolerance $\epsilon_k = \frac{10^{-4}}{2k}$ for the proximal sub-

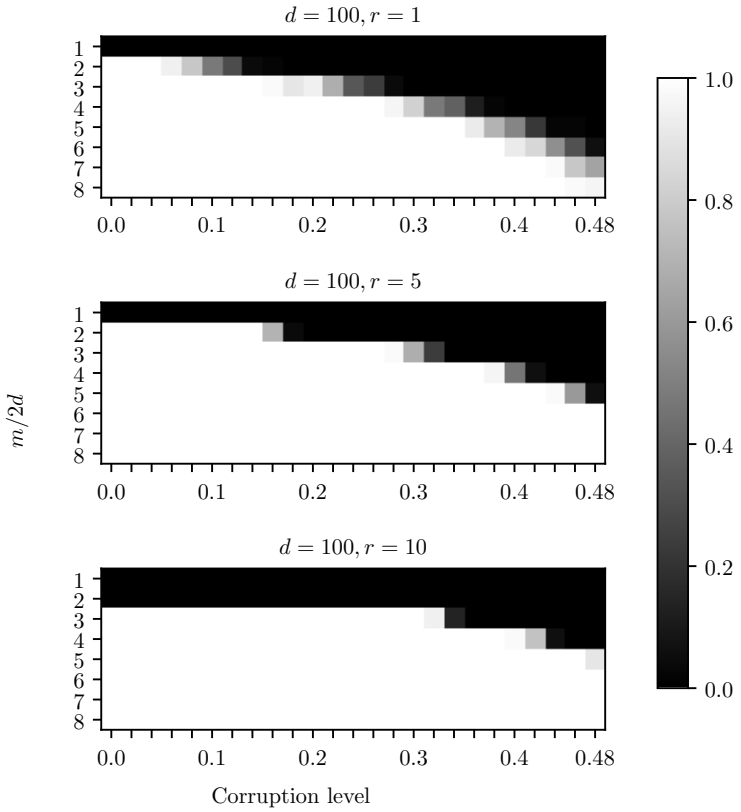


Figure 5.3: Quadratic sensing with symmetrized measurements using Algorithm 3.

problems, which we continue solve for at most 500 iterations. We choose $\gamma = 10$ in all subproblems. The phase transition plots are shown in Figure 5.4. It appears that the prox-linear method is more robust to additive sparse corruption, since the empirical recovery rate for the subgradient method decays faster as the rank increases.

Matrix completion. We next perform experiments on the low-rank matrix completion problem that test successful recovery against the sampling frequency. We generate random instances of the problem, where we let the probability of observing an entry, $\mathbb{P}(\delta_{ij} = 1)$, range in $[0.02, 0.6]$ with increments of 0.02. Figure 5.5 depicts the empirical recovery rate using the Polyak sub-

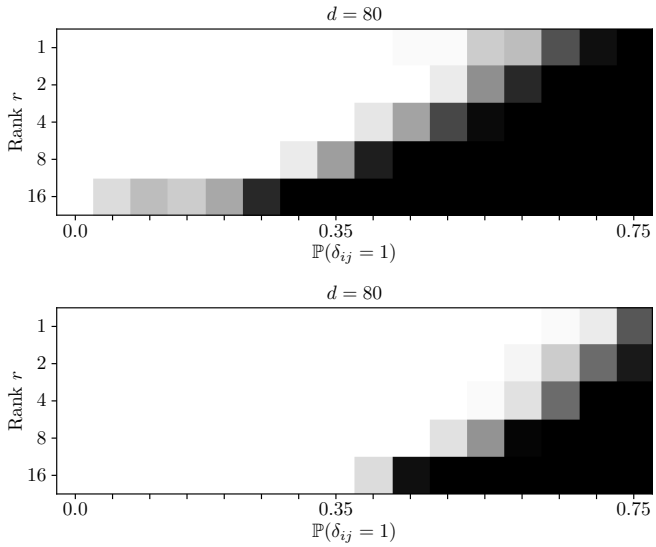


Figure 5.4: Robust PCA using the subgradient method, Algorithm 3, (top) and the modified prox-linear method (5.47) (bottom).

gradient method and the modified prox-linear algorithm (5.42). Similarly to the quadratic/bilinear sensing problems, low-rank matrix completion exhibits a sharp phase transition. As predicted in Section 5.6, the ratio $\frac{r^2}{d}$ appears to be driving the required observation probability for successful recovery. Finally, we empirically observe that the prox-linear method can "tolerate" slightly smaller sampling frequencies.

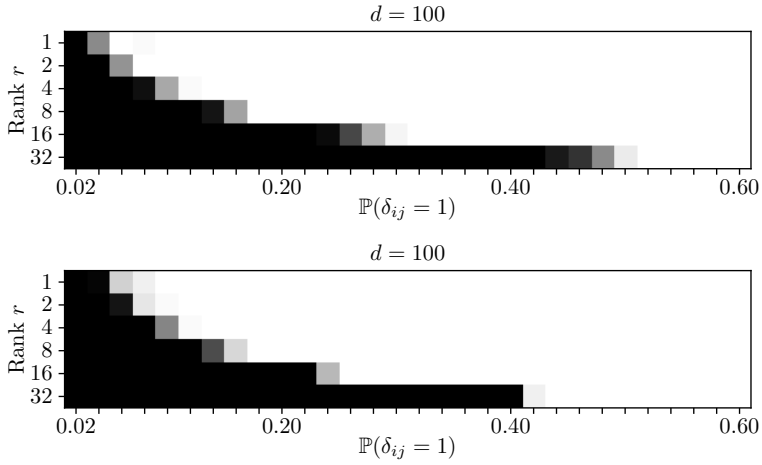


Figure 5.5: Low-rank matrix completion using the subgradient method, Algorithm 2 (top), and the modified prox-linear method (5.42) (bottom).

5.9.2 Convergence behavior

We empirically validate the rapid convergence guarantees of the subgradient and prox-linear methods, given a proper initialization. Moreover, we compare the subgradient method with gradient descent, i.e. gradient descent applied to a smooth formulation of each problem, using the same initial estimate in the noiseless setting. In all the cases below, the step sizes for the gradient method were tuned for best performance. Moreover, we noticed that the gradient descent method, equipped with the Polyak step size $\eta := \tau \frac{\nabla f}{\|\nabla f\|^2}$ performed at least as well as gradient descent with constant step size. That being said, we were unable to locate any theoretical guarantees in the literature for gradient descent with the Polyak step-size for the problems we consider here.

Quadratic and Bilinear sensing. For the quadratic and bilinear sensing problems, we apply gradient descent on the smooth formulations

$$\frac{1}{m}\left\|\mathcal{A}(XX^\top) - b\right\|_2^2 \quad \text{and} \quad \frac{1}{m}\left\|\mathcal{A}(XY) - b\right\|_2^2.$$

In Figure 5.6, we plot the performance of Algorithm 3 for matrix sensing problems with different rank / corruption level; remarkably, the level of noise does not significantly affect the rate of convergence. Additionally, the convergence behavior is almost identical for the two problems for similar rank/noise configurations. Figure 5.7 depicts the behavior of Algorithm 2 versus gradient descent with empirically tuned step sizes. The subgradient method significantly outperforms gradient descent. For completeness, we also depict the convergence rate of Algorithm 4 for both problems in Figure 5.8, where we solve the proximal subproblems approximately.

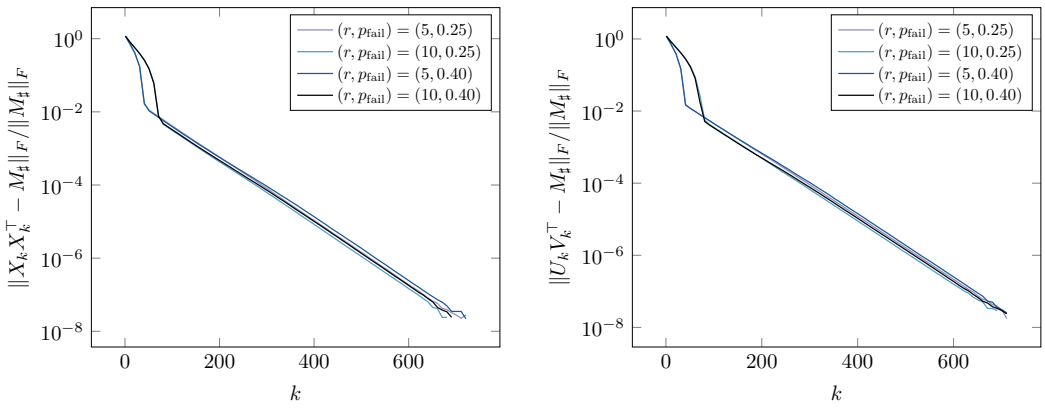


Figure 5.6: Quadratic (left) and bilinear (right) matrix sensing with $d = 200, m = 8 \cdot rd$, using the subgradient method, Algorithm 3.

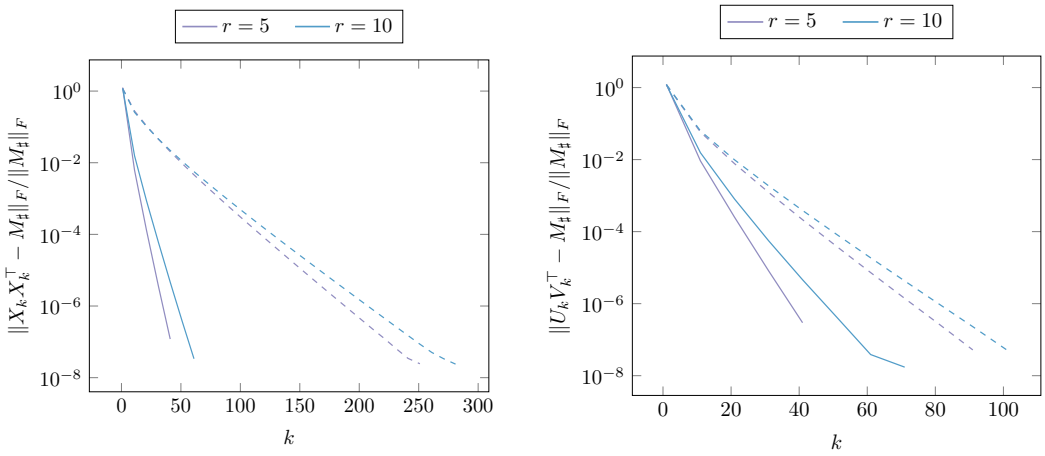


Figure 5.7: Algorithm 2 (solid lines) against gradient descent (dashed lines) with step size η. Left: quadratic sensing, $\eta = 10^{-4}$. Right: bilinear sensing, $\eta = 10^{-3}$.

Matrix completion. In our comparison with smooth methods, we apply gradient descent on the following minimization problem:

$$\min_{X \in \mathbf{R}^{d \times r} : \|X\|_{2,\infty} \le C} \left\| \Pi_\Omega(XX^\top) - \Pi_\Omega(M) \right\|_F^2 . \tag{5.52}$$

Figure 5.9 depicts the convergence behavior of Algorithm 2 (solid lines) versus gradient descent applied to Problem (5.52) with a tuned step size $\eta = 0.004$ (dashed lines), initialized under the same conditions for low-rank matrix completion instances. As the theory suggests, higher sampling frequency implies better convergence rates. The subgradient method outperforms gradient de-

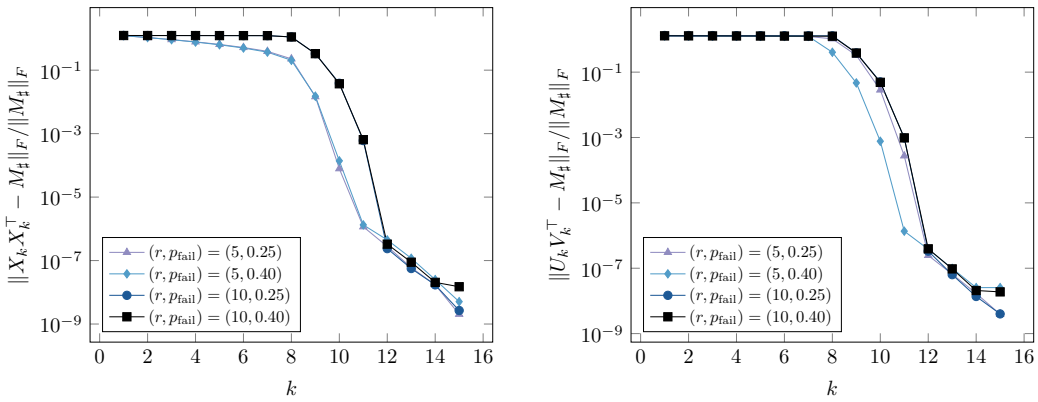


Figure 5.8: Quadratic (left) and bilinear (right) matrix sensing with $d = 100, m = 8 \cdot rd$, using the prox-linear method, Algorithm 4.

scent in all regimes.

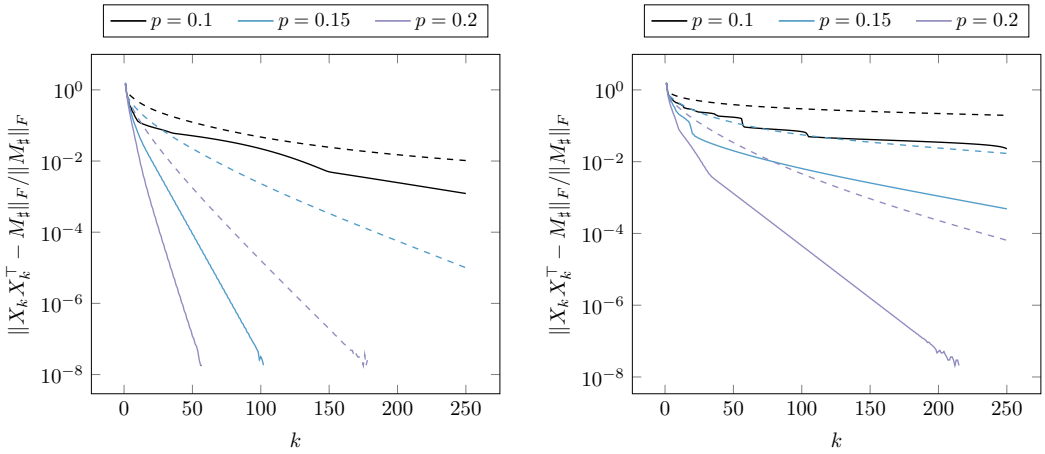


Figure 5.9: Low rank matrix completion with $d = 100$. Left: $r = 4$, right: $r = 8$. Solid lines use Algorithm 2, dashed lines use gradient descent with step $\eta = 0.004$.

Figure 5.10 depicts the performance of the modified prox-linear method (5.42) in the same setting as Figure 5.9. In most cases, the prox-linear algorithm converges within just 15 iterations, at what appears to be a rapid linear rate of convergence. Each convex subproblem is solved using a variant of the graph-splitting ADMM algorithm [200].

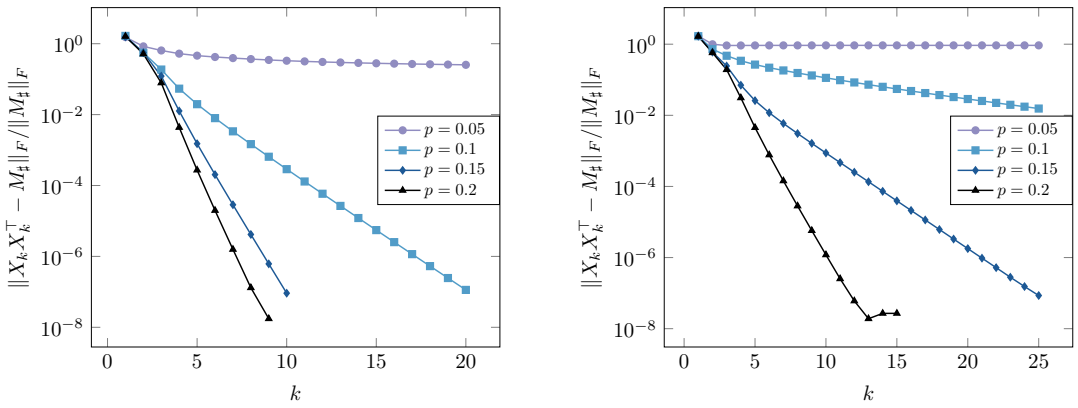


Figure 5.10: Low rank matrix completion with $d = 100$ using the modified prox-linear Algorithm (5.42). Left: $r = 4$, right: $r = 8$.

Robust PCA. For the robust PCA problem, we consider different rank/corruption level configurations to better understand how they affect convergence for the subgradient and prox-linear methods, using the non-Euclidean formulation of Section 5.7.2. We depict all configurations in the same plot for a fixed optimization algorithm to better demonstrate the effect of each parameter, as shown in Figure 5.11. The parameters of the prox-linear method are chosen in the same way reported in Section 5.9.1. In particular, our numerical experiments appear to support our sharpness Conjecture 5.7.6 for the robust PCA problem.

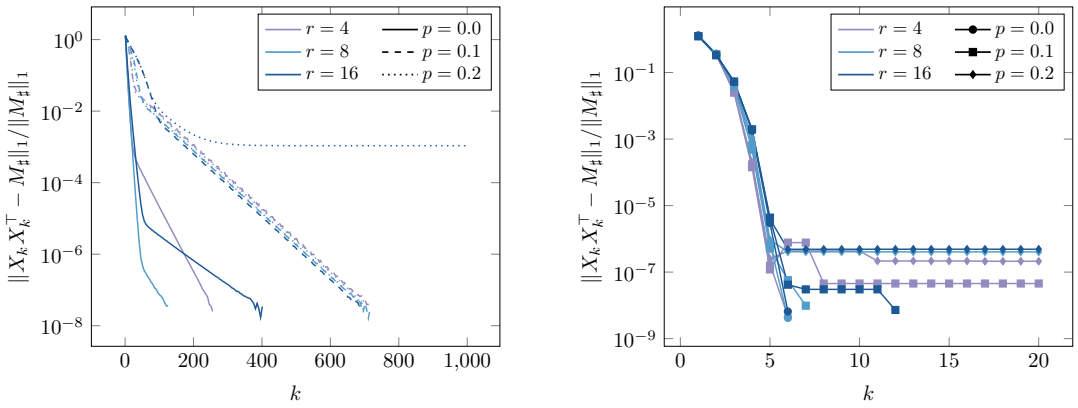


Figure 5.11: ℓ_1-robust PCA with $d = 100$ and $p := \mathbb{P}(\delta_{ij} = 1)$. Left: Algorithm 3, right: Algorithm (5.42).

Recovery up to tolerance

In this last section, we test the performance of the prox-linear method and the modified Polyak subgradient method (Algorithm 5) for the quadratic sensing and matrix completion problems, under a dense noise model of Section 5.8. In the former setting, we set $p_{\text{fail}} = 0.25$, so 1/4th of our measurements is corrupted with large magnitude noise. For matrix completion, we observe $p = 25\%$ of the entries. In both settings, we add Gaussian noise e which is rescaled to satisfy $\|e\|_F = \delta\sigma_r(X_\sharp)$, and test $\delta := 10^{-k}\sigma_r(X_\sharp)$, $k \in \{1, \ldots, 4\}$. The relevant plots can be found in Figures 5.12 and 5.13. The numerical experiments fully support the developed theory, with the iterates converging rapidly up to the tolerance that is proportional to the noise level. Incidentally, we observe that the modified prox-linear method (5.42) is more robust to additive noise for the matrix completion problem, with Algorithm 5 exhibiting heavy fluctuations and failing to converge for the highest level of dense noise.

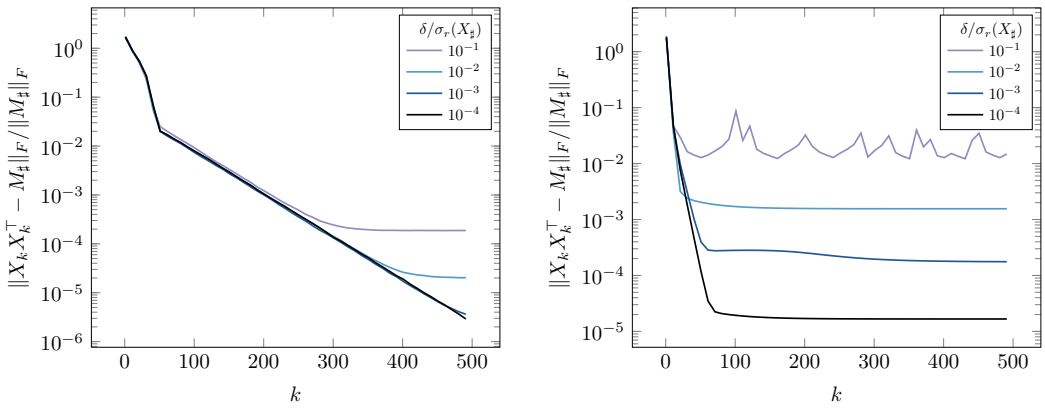


Figure 5.12: Quadratic sensing with $r = 5$ (left) and matrix completion with $r = 8$ (right), $d = 100$, using Algorithm 5.

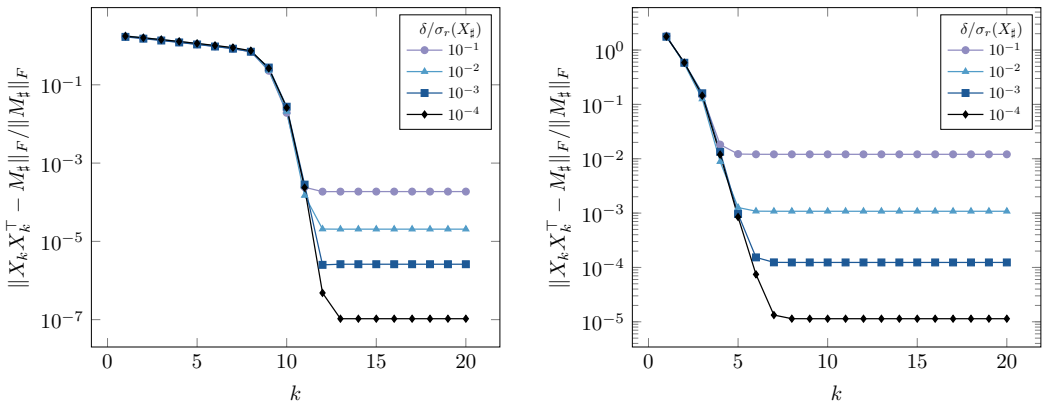


Figure 5.13: Quadratic sensing with $r = 5$ (left) and matrix completion with $r = 8$ (right), $d = 100$, using Algorithm (5.42).

5.10 Analysis

5.10.1 Proofs in Section 5.4

In this section, we prove rapid local convergence guarantees for the subgradient and prox-linear algorithms under regularity conditions that hold only locally around a particular solution. We will use the Euclidean norm throughout this section; therefore to simplify the notation, we will drop the subscript two. Thus $\|\cdot\|$ denotes the ℓ_2 on a Euclidean space $\mathbf{E}$ throughout.

We will need the following quantitative version of Lemma 5.4.2.

Lemma 5.10.1. *Suppose Assumption 5.4.7 holds and let $\gamma \in (0, 2)$ be arbitrary. Then for any point $x \in B_{\epsilon/2}(\bar{x}) \cap \mathcal{T}_\gamma \backslash \mathcal{X}^*$, the estimate holds:*

$$\operatorname{dist}(0, \partial f(x)) \geq \left(1 - \tfrac{\gamma}{2}\right)\mu.$$

Proof. Consider any point $x \in B_{\epsilon/2}(\bar{x})$ satisfying $\operatorname{dist}(x, \mathcal{X}^*) \leq \gamma\frac{\mu}{\rho}$. Let $x^* \in \operatorname{proj}_{\mathcal{X}^*}(x)$

be arbitrary and note $x^* \in B_\epsilon(\bar{x})$. Thus for any $\zeta \in \partial f(x)$ we deduce

$$\mu \cdot \operatorname{dist}(x, \mathcal{X}^*) \leq f(x) - f(x^*) \leq \langle \zeta, x - x^* \rangle + \frac{\rho}{2}\|x - x^*\|^2 \leq \|\zeta\| \operatorname{dist}(x, \mathcal{X}^*) + \frac{\rho}{2}\operatorname{dist}^2(x, \mathcal{X}^*).$$

Therefore we deduce the lower bound on the subgradients $\|\zeta\| \geq \mu - \frac{\rho}{2} \cdot \operatorname{dist}(x, \mathcal{X}^*) \geq \left(1 - \frac{\gamma}{2}\right)\mu$, as claimed. □

Proof of Theorem 5.4.8

Let k be the first index (possibly infinite) such that $x_k \notin B_{\epsilon/2}(\bar{x})$. We claim that (5.28) holds for all $i < k$. We show this by induction. To this end, suppose (5.28) holds for all indices up to $i-1$. In particular, we deduce $\operatorname{dist}(x_i, \mathcal{X}^*) \leq \operatorname{dist}(x_0, \mathcal{X}^*) \leq \frac{\mu}{2\rho}$. Let $x^* \in \operatorname{proj}_{\mathcal{X}^*}(x_i)$ and note $x^* \in B_\epsilon(\bar{x})$, since

$$\|x^* - \bar{x}\| \leq \|x^* - x_i\| + \|x_i - \bar{x}\| \leq 2\|x_i - \bar{x}\| \leq \epsilon.$$

Thus we deduce

$$\begin{aligned}
\|x_{i+1} - x^*\|^2 &= \left\|\operatorname{proj}_{\mathcal{X}}\left(x_i - \tfrac{f(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right) - \operatorname{proj}_{\mathcal{X}}(x^*)\right\|^2 \\
&\leq \left\|(x_i - x^*) - \tfrac{f(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right\|^2 \qquad (5.53) \\
&= \|x_i - x^*\|^2 + \frac{2(f(x_i) - \min_{\mathcal{X}} f)}{\|\zeta_i\|^2} \cdot \langle \zeta_i, x^* - x_i \rangle + \frac{(f(x_i) - f(x^*))^2}{\|\zeta_i\|^2} \\
&\leq \|x_i - x^*\|^2 + \frac{2(f(x_i) - \min f)}{\|\zeta_i\|^2}\left(f(x^*) - f(x_i) + \frac{\rho}{2}\|x_i - x^*\|^2\right) \\
&\qquad + \frac{(f(x_i) - f(x^*))^2}{\|\zeta_i\|^2} \qquad (5.54)
\end{aligned}$$

Here, the estimate (5.53) follows from the fact that the projection $\operatorname{proj}_Q(\cdot)$ is non-expansive, (5.54) uses local weak convexity.

Then, rearranging we get

$$
\begin{aligned}
&= \|x_i - x^*\|^2 + \frac{f(x_i) - \min f}{\|\zeta_i\|^2} \left(\rho\|x_i - x^*\|^2 - (f(x_i) - f(x^*))\right) \\
&\leq \|x_i - x^*\|^2 + \frac{f(x_i) - \min f}{\|\zeta_i\|^2} \left(\rho\|x_i - x^*\|^2 - \mu\|x_i - x^*\|\right) && (5.55) \\
&= \|x_i - x^*\|^2 + \frac{\rho(f(x_i) - \min f)}{\|\zeta_i\|^2} \left(\|x_i - x^*\| - \frac{\mu}{\rho}\right)\|x_i - x^*\| \\
&\leq \|x_i - x^*\|^2 - \frac{\mu(f(x_i) - \min f)}{2\|\zeta_i\|^2} \cdot \|x_i - x^*\| && (5.56) \\
&\leq \left(1 - \frac{\mu^2}{2\|\zeta_i\|^2}\right)\|x_i - x^*\|^2. && (5.57)
\end{aligned}
$$

where (5.56) follow from the estimate $\operatorname{dist}(x_i, \mathcal{X}^*) \leq \frac{\mu}{2\rho}$, while (5.55) and (5.57) use local sharpness. We therefore deduce

$$
\operatorname{dist}^2(x_{i+1}; \mathcal{X}^*) \leq \|x_{i+1} - x^*\|^2 \leq \left(1 - \frac{\mu^2}{2L^2}\right)\operatorname{dist}^2(x_i, \mathcal{X}^*). \tag{5.58}
$$

Thus (5.28) holds for all indices up to $k - 1$. We next show that k is infinite. To this end, observe

$$
\begin{aligned}
\|x_k - x_0\| \leq \sum_{i=0}^{k-1} \|x_{i+1} - x_i\| &= \sum_{i=0}^{k-1} \left\|\operatorname{proj}_{\mathcal{X}}\left(x_i - \tfrac{f(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right) - \operatorname{proj}_{\mathcal{X}}(x_i)\right\| \\
&\leq \sum_{i=0}^{k-1} \frac{f(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|} \\
&\leq \sum_{i=0}^{k-1} \left\langle \tfrac{\zeta_i}{\|\zeta_i\|}, x_i - \operatorname{proj}_{\mathcal{X}^*}(x_i)\right\rangle + \frac{\rho}{2\|\zeta_i\|}\|x_i - \operatorname{proj}_{\mathcal{X}^*}(x_i)\|^2 \\
&\leq \sum_{i=0}^{k-1} \operatorname{dist}(x_i, \mathcal{X}^*) + \frac{2\rho}{3\mu}\operatorname{dist}^2(x_i, \mathcal{X}^*) && (5.59) \\
&\leq \frac{4}{3} \cdot \sum_{i=0}^{k-1} \operatorname{dist}(x_i, \mathcal{X}^*) && (5.60) \\
&\leq \frac{4}{3} \cdot \operatorname{dist}(x_0, \mathcal{X}^*) \cdot \sum_{i=0}^{k-1} \left(1 - \frac{\mu^2}{2L^2}\right)^{\frac{i}{2}} && (5.61) \\
&\leq \frac{16L^2}{3\mu^2} \cdot \operatorname{dist}(x_0, \mathcal{X}^*) \leq \frac{\epsilon}{4},
\end{aligned}
$$

where (5.59) follows by Lemma 5.10.1 with $\gamma = 1/2$, the bound in (5.60) follows by (5.58) and the assumption on $\operatorname{dist}(x_0, \mathcal{X}^*)$, finally (5.61) holds thanks to (5.58).

Thus applying the triangle inequality we get the contradiction $\|x_k - \bar{x}\| \leq \epsilon/2$. Consequently all the iterates x_k for $k = 0, 1, \ldots, \infty$ lie in $B_{\epsilon/2}(\bar{x})$ and satisfy (5.28).

Finally, let x_∞ be any limit point of the sequence $\{x_i\}$. We then successively compute

$$\begin{aligned}
\|x_k - x_\infty\| \leq \sum_{i=k}^{\infty} \|x_{i+1} - x_i\| &\leq \sum_{i=k}^{\infty} \frac{f(x_i) - \min f}{\|\zeta_i\|} \\
&\leq \frac{4L}{3\mu} \cdot \sum_{i=k}^{\infty} \operatorname{dist}(x_i, \mathcal{X}^*) \\
&\leq \frac{4L}{3\mu} \cdot \operatorname{dist}(x_0, \mathcal{X}^*) \cdot \sum_{i=k}^{\infty} \left(1 - \frac{\mu^2}{2L^2}\right)^{\frac{i}{2}} \\
&\leq \frac{16L^3}{3\mu^3} \cdot \operatorname{dist}(x_0, \mathcal{X}^*) \cdot \left(1 - \frac{\mu^2}{2L^2}\right)^{\frac{k}{2}}.
\end{aligned}$$

This completes the proof.

Proof of Theorem 5.4.9

Fix an arbitrary index k and observe

$$\|x_{k+1} - x_k\| = \left\| \operatorname{proj}_Q(x_k) - \operatorname{proj}_Q\left(x_k - \alpha_k \frac{\xi_k}{\|\xi_k\|}\right) \right\| \leq \alpha_k = \lambda \cdot q^k.$$

Hence, we conclude the uniform bound on the iterates:

$$\|x_k - x_0\| \leq \sum_{i=0}^{\infty} \|x_{i+1} - x_i\| \leq \tfrac{\lambda}{1-q}$$

and the R-linear rate of convergence

$$\|x_k - x_\infty\| \leq \sum_{i=k}^{\infty} \|x_{i+1} - x_i\| \leq \tfrac{\lambda}{1-q} q^k,$$

where x_∞ is any limit point of the iterate sequence.

Let us now show that the iterates do not escape $B_{\epsilon/2}(\bar{x})$. To this end, observe

$$\|x_k - \bar{x}\| \leq \|x_k - x_0\| + \|x_0 - \bar{x}\| \leq \tfrac{\lambda}{1-q} + \tfrac{\epsilon}{4}.$$

We must therefore verify the estimate $\frac{\lambda}{1-q} \leq \frac{\epsilon}{4}$, or equivalently $\gamma \leq \frac{\epsilon\rho L(1-\gamma)\tau^2}{4\mu^2(1+\sqrt{1-(1-\gamma)\tau^2})}$. Clearly, it suffices to verify $\gamma \leq \frac{\epsilon\rho(1-\gamma)}{4L}$, which holds by the definition of γ. Thus all the iterates x_k lie in $B_{\epsilon/2}(\bar{x})$. Moreover $\tau \leq \sqrt{\frac{1}{2}} \leq \sqrt{\frac{1}{2-\gamma}}$, the rest of the proof is identical to that in [68, Theorem 5.1].

Proof of Theorem 5.4.11

Fix any index i such that $x_i \in B_\epsilon(\bar{x})$ and let $x \in \mathcal{X}$ be arbitrary. Since the function $z \mapsto f_{x_i}(z) + \frac{\beta}{2}\|z - x_i\|^2$ is β-strongly convex and x_{i+1} is its minimizer, we deduce

$$\left(f_{x_i}(x_{i+1}) + \frac{\beta}{2}\|x_{i+1} - x_i\|^2\right) + \frac{\beta}{2}\|x_{i+1} - x\|^2 \leq f_{x_i}(x) + \frac{\beta}{2}\|x - x_i\|^2. \tag{5.62}$$

Setting $x = x_i$ and appealing to approximation accuracy, we obtain the descent guarantee

$$\|x_{i+1} - x_i\|^2 \leq \frac{2}{\beta}(f(x_i) - f(x_{i+1})). \tag{5.63}$$

In particular, the function values are decreasing along the iterate sequence. Next choosing any $x^* \in \mathrm{proj}_{\mathcal{X}^*}(x_i)$ and setting $x = x^*$ in (5.62) yields

$$\left(f_{x_i}(x_{i+1}) + \frac{\beta}{2}\|x_{i+1} - x_i\|^2\right) + \frac{\beta}{2}\|x_{i+1} - x^*\|^2 \leq f_{x_i}(x^*) + \frac{\beta}{2}\|x^* - x_i\|^2.$$

Appealing to approximation accuracy and lower-bounding $\frac{\beta}{2}\|x_{i+1} - x^*\|^2$ by zero, we conclude

$$f(x_{i+1}) \leq f(x^*) + \beta\|x^* - x_i\|^2. \tag{5.64}$$

Using sharpness we deduce the contraction guarantee

$$\begin{aligned} f(x_{i+1}) - f(x^*) &\leq \beta \cdot \mathrm{dist}^2(x_i, \mathcal{X}^*) \\ &\leq \frac{\beta}{\mu^2}(f(x_i) - f(x^*))^2 \\ &\leq \frac{\beta(f(x_i) - f(x^*))}{\mu^2} \cdot (f(x_i) - f(x^*)) \leq \frac{1}{2} \cdot (f(x_i) - f(x^*)), \end{aligned} \tag{5.65}$$

where the last inequality uses the assumption $f(x_0) - \min_{\mathcal{X}} f \leq \frac{\mu^2}{2\beta}$. Let $k > 0$ be the first index satisfying $x_k \notin B_\epsilon(\bar{x})$. We then deduce

$$\begin{aligned}
\|x_k - x_0\| \leq \sum_{i=0}^{k-1} \|x_{i+1} - x_i\| &\leq \sqrt{\frac{2}{\beta}} \cdot \sum_{i=0}^{k-1} \sqrt{f(x_i) - f(x_{i+1})} && (5.66)\\
&\leq \sqrt{\frac{2}{\beta}} \cdot \sum_{i=0}^{k-1} \sqrt{f(x_i) - f(x^*)} \\
&\leq \sqrt{\frac{2}{\beta}} \cdot \sqrt{f(x_0) - f(x^*)} \cdot \sum_{i=0}^{k-1} \left(\frac{1}{2}\right)^{\frac{i}{2}} && (5.67)\\
&\leq \frac{1}{\sqrt{2}-1} \sqrt{\frac{f(x_0) - f(x^*)}{\beta}} \leq \epsilon/2,
\end{aligned}$$

where (5.66) follows from (5.63) and (5.67) follows from (5.65). Thus we conclude $\|x_k - \bar{x}\| \leq \epsilon$, which is a contradiction. Therefore all the iterates x_k, for $k = 0, 1, \ldots, \infty$, lie in $B_\epsilon(\bar{x})$. Combing this with (5.64) and sharpness yields the claimed quadratic converge guarantee

$$\mu \cdot \operatorname{dist}(x_{k+1}, \mathcal{X}^*) \leq f(x_{k+1}) - f(\bar{x}) \leq \beta \cdot \operatorname{dist}^2(x_k, \mathcal{X}).$$

Finally, let x_∞ be any limit point of the sequence $\{x_i\}$. We then deduce

$$\begin{aligned}
\|x_k - x_\infty\| \leq \sum_{i=k}^{\infty} \|x_{i+1} - x_i\| &\leq \sqrt{\frac{2}{\beta}} \cdot \sum_{i=k}^{\infty} \sqrt{f(x_i) - f(x_{i+1})} \\
&\leq \sqrt{\frac{2}{\beta}} \cdot \sum_{i=k}^{\infty} \sqrt{f(x_i) - \min_{\mathcal{X}} f} \\
&\leq \frac{\mu\sqrt{2}}{\beta} \cdot \sum_{i=k}^{\infty} \left(\frac{\beta}{\mu^2}(f(x_0) - \min f)\right)^{2^{i-1}} && (5.68)\\
&\leq \frac{\mu\sqrt{2}}{\beta} \cdot \sum_{i=k}^{\infty} \left(\frac{1}{2}\right)^{2^{i-1}} \\
&\leq \frac{\mu\sqrt{2}}{\beta} \sum_{j=0}^{\infty} \left(\frac{1}{2}\right)^{2^{k-1}+j} \leq \frac{2\sqrt{2}\mu}{\beta} \cdot \left(\frac{1}{2}\right)^{2^{k-1}},
\end{aligned}$$

where (5.68) follows from (5.65). The theorem is proved.

5.10.2 Proofs in Section 5.5

Proof of Lemma 5.5.7

In order to prove that the assumption in each case, we will prove a stronger "small-ball condition" [172, 171], which immediately implies the claimed lower bounds on the expectation by Markov's inequality. More precisely, we will show that there exist numerical constants $\mu_0, p_0 > 0$ such that

1. **(Matrix Sensing)**

$$\inf_{\substack{M:\ \mathrm{rank} M \le 2r \\ \|M\|_F = 1}} \mathbb{P}(|\langle P, M\rangle| \ge \mu_0) \ge p_0,$$

2. **(Quadratic Sensing I)**

$$\inf_{\substack{M \in \mathcal{S}^d:\ \mathrm{rank} M \le 2r \\ \|M\|_F = 1}} \mathbb{P}(|p^\top M p| \ge \mu_0) \ge p_0,$$

3. **(Quadratic Sensing II)**

$$\inf_{\substack{M \in \mathcal{S}^d:\ \mathrm{rank} M \le 2r \\ \|M\|_F = 1}} \mathbb{P}(|p^\top M p - \tilde{p}^\top M \tilde{p}| \ge \mu_0) \ge p_0,$$

4. **(Bilinear Sensing)**

$$\inf_{\substack{M:\ \mathrm{rank} M \le 2r \\ \|M\|_F = 1}} \mathbb{P}(|p^\top M q| \ge \mu_0) \ge p_0.$$

These conditions immediately imply Assumptions 5.5.3-5.5.6. Indeed, by Markov's inequality, in the case of matrix sensing we deduce

$$\mathbb{E}|\langle P, M\rangle| \ge \mu_0 \mathbb{P}\left(|\langle P, M\rangle| > \mu_0\right) \ge \mu_0 p_0.$$

The same reasoning applies to all the other problems.

Matrix sensing. Consider any matrix M with $\|M\|_F = 1$. Then, since $g := \langle P, M\rangle$ follows a standard normal distribution, we may set μ_0 to be the median of $|g|$ and $p_0 = 1/2$ to obtain

$$\inf_{\substack{M:\ \mathrm{rank} M \le 2r \\ \|M\|_F = 1}} \mathbb{P}(|\langle P, M\rangle| \ge \mu_0) = \mathbb{P}(|g| \ge \mu_0) \ge p_0.$$

Quadratic Sensing I. Fix a matrix M with $\mathrm{rank} M \le 2r$ and $\|M\|_F = 1$. Let $M = UDU^\top$ be an eigenvalue decomposition of M. Using the rotational invariance of the Gaussian distribution, we deduce

$$p^\top M p \overset{d}{=} p^\top D p = \sum_{k=1}^{2r} \lambda_k p_k^2,$$

where $\overset{d}{=}$ denotes equality in distribution. Next, let z be a standard normal variable. We will now invoke Proposition 5.10.8. Let $C > 0$ be the numerical constant appearing in the proposition. Notice that the function $\phi\colon \mathbf{R}_+ \to \mathbf{R}$ given by

$$\phi(t) = \sup_{u \in \mathbf{R}} \mathbb{P}(|z^2 - u| \le t)$$

is continuous and strictly increasing, and it satisfies $\phi(0) = 0$ and $\lim_{t\to\infty} \phi(t) = 1$. Hence we may set $\mu_0 = \phi^{-1}(\min\{1/2C, 1/2\})$. Proposition 5.10.8 then yields

$$\mathbb{P}(|p^\top M p| \le \mu_0) = \mathbb{P}\left(\left|\sum_{k=1}^{2r} \lambda_k p_k^2\right| \le \mu_0\right) \le \sup_{u\in\mathbf{R}} \mathbb{P}\left(\left|\sum_{k=1}^{2r} \lambda_k p_k^2 - u\right| \le \mu_0\right) \le C\phi(\mu_0) \le \frac{1}{2}.$$

By taking the supremum of both sides of the inequality we conclude that Assumption 5.5.4 holds with μ_0 and $p_0 = 1/2$.

Quadratic sensing II. Let $M = UDU^\top$ be an eigenvalue decomposition of M. Using the rotational invariance of the Gaussian distribution, we deduce

$$p^\top M p - \tilde{p}^\top M \tilde{p} \overset{d}{=} p^\top D p - \tilde{p}^\top D \tilde{p} = \sum_{k=1}^{2r} \lambda_k \left(p_k^2 - \tilde{p}_k^2\right) \overset{d}{=} 2\sum_{k=1}^{2r} \lambda_k p_k \tilde{p}_k,$$

where the last relation follows since $(p_k - \tilde{p}_k), (p_k + \tilde{p}_k)$ are independent standard normal random variables with mean zero and variance two. We will now invoke Proposition 5.10.8. Let $C > 0$ be the numerical constant appearing in the proposition. Let z and $\tilde{z}$ be independent standard normal variables. Notice that the function $\phi : \mathbf{R}_+ \to \mathbf{R}$ given by

$$\phi(t) = \sup_{u \in \mathbf{R}} \mathbb{P}(|2z\tilde{z} - u| \leq t)$$

is continuous, strictly increasing, satisfies $\phi(0) = 0$ and approaches one at infinity. Defining $\mu_0 = \phi^{-1}(\min\{1/2C, 1/2\})$ and applying Proposition 5.10.8, we get

$$\mathbb{P}\left(\left|2\sum_{k=1}^{2r} \sigma_k p_k \tilde{p}_k\right| \leq \mu_0\right) \leq \sup_{u \in \mathbf{R}} \mathbb{P}\left(\left|2\sum_{k=1}^{2r} \sigma_k p_k \tilde{p}_k - u\right| \leq \mu_0\right) \leq C\phi(\mu_0) \leq \frac{1}{2}.$$

By taking the supremum of both sides of the inequality we conclude that Assumption 5.5.5 holds with μ_0 and $p_0 = 1/2$.

We omit the details for the bilinear case, which follow by similar arguments.

Proof of Theorem 5.5.8

The proofs in this section rely on the following proposition, which shows that that pointwise concentration imply uniform concentration. We defer the proof to Section 5.10.2.

Proposition 5.10.2. *Let $\mathcal{A} : \mathbf{R}^{d_1 \times d_2} \to \mathbf{R}^m$ be a random linear mapping with property that for any fixed matrix $M \in \mathbf{R}^{d_1 \times d_2}$ of rank at most $2r$ with norm $\|M\|_F = 1$ and any fixed subset of indices $\mathcal{I} \subseteq \{1, \ldots, m\}$ satisfying $|\mathcal{I}| < m/2$, the following hold:*

(1) *The measurements $\mathcal{A}(M)_1, \ldots, \mathcal{A}(M)_m$ are i.i.d.*

(2) *RIP holds in expected value:*

$$\alpha \leq \mathbb{E}|\mathcal{A}(M)_i| \leq \beta(r) \qquad \text{for all } i \in \{1, \ldots, m\} \tag{5.69}$$

where $\alpha > 0$ is a universal constant and β is a positive-valued function that could potentially depend on the rank of M.

(3) *There exist a universal constant $K > 0$ and a positive-valued function $c(m, r)$ such that for any $t \in [0, K]$ the deviation bound*

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1]\right| \leq t \tag{5.70}$$

holds with probability at least $1 - 2\exp(-t^2 c(m, r))$.

Then, there exist universal constants $c_1, \ldots, c_6 > 0$ depending only on α and K such that if $\mathcal{I} \subseteq \{1, \ldots, m\}$ is a fixed subset of indices satisfying $|\mathcal{I}| < m/2$ and

$$c(m, r) \geq \frac{c_1}{(1 - 2|\mathcal{I}|/m)^2} r(d_1 + d_2 + 1) \ln\left(c_2 + \frac{c_2 \beta(r)}{1 - 2|\mathcal{I}|/m}\right)$$

then with probability at least $1 - 4\exp\left(-c_3(1 - 2|\mathcal{I}|/m)^2 c(m, r)\right)$ every matrix $M \in \mathbf{R}^{d_1 \times d_2}$ of rank at most $2r$ satisfies

$$c_4 \|M\|_F \leq \frac{1}{m}\|\mathcal{A}(M)\|_1 \leq c_5 \beta(r)\|M\|_F, \tag{5.71}$$

and

$$c_6\left(1 - \frac{2|\mathcal{I}|}{m}\right)\|M\|_F \leq \frac{1}{m}\left(\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}} M\|_1\right). \tag{5.72}$$

Due to scale invariance of the above result, we need only verify its assumptions in the case that $\|M\|_F = 1$. We implicitly use this observation below.

Part 1 of Theorem 5.5.8 (Matrix sensing)

Lemma 5.10.3. *The random variable $|\langle P, M\rangle|$ is sub-gaussian with parameter $C\eta$. Consequently,*

$$\alpha \leq \mathbb{E}|\langle P, M\rangle| \lesssim \eta. \tag{5.73}$$

Moreover, there exists a universal constant $c > 0$ such that for any $t \in [0, \infty)$ the deviation bound

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1]\right| \leq t \tag{5.74}$$

holds with probability at least $1 - 2\exp\left(-\frac{ct^2}{\eta^2}m\right)$.

Proof. Assumption 5.5.3 immediately implies the lower bound in (5.73). To prove the upper bound, first note that by assumption we have

$$\|\langle P, M\rangle\|_{\psi_2} \lesssim \eta.$$

This bound has two consequences, first $\langle P, M\rangle$ is a sub-gaussian random variable with parameter η and second $\mathbb{E}|\langle P, M\rangle| \lesssim \eta$ [235, Proposition 2.5.2]. Thus, we have proved (5.73).

To prove the deviation bound (5.74) we introduce the random variables

$$Y_i = \begin{cases} |\langle P_i, M\rangle| - \mathbb{E}|\langle P_i, M\rangle| & \text{if } i \notin \mathcal{I}, \text{ and} \\ -\left(|\langle P_i, M\rangle| - \mathbb{E}|\langle P_i, M\rangle|\right) & \text{otherwise.} \end{cases}$$

Since $|\langle P_i, M\rangle|$ is sub-gaussian, we have $\|Y_i\|_{\psi_2} \lesssim \eta$ for all i, see [235, Lemma 2.6.8]. Hence, Hoeffding's inequality for sub-gaussian random variables [235, Theorem 2.6.2] gives the desired upper bound on $\mathbb{P}\left(\frac{1}{m}\left|\sum_{i=1}^m Y_i\right| \geq t\right)$. □

Applying Proposition 5.10.2 with $\beta(r) \asymp \eta$ and $c(m, r) \asymp m/\eta^2$ now yields the result. □

Part 2 of Theorem 5.5.8 (Quadratic sensing I)

Lemma 5.10.4. *The random variable $|p^\top M p|$ is sub-exponential with parameter $\sqrt{2r}\eta^2$. Consequently,*

$$\alpha \le \mathbb{E}|p^\top M p| \lesssim \sqrt{2r}\eta^2. \tag{5.75}$$

Moreover, there exists a universal constant $c > 0$ such that for any $t \in [0, \sqrt{2r}\eta]$ the deviation bound

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1]\right| \le t \tag{5.76}$$

holds with probability at least $1 - 2\exp\left(-\frac{ct^2}{\eta^4}m/r\right)$.

Proof. Assumption 5.5.4 gives the lower bound in (5.75). To prove the upper bound, first note that $M = \sum_{k=1}^{2r}\sigma_k u_k u_k^\top$ where σ_k and u_k are the kth singular values and vectors of M, respectively. Hence

$$\begin{aligned}\|p^\top M p\|_{\psi_1} &= \left\|p^\top\left(\sum_{k=1}^{2r}\sigma_k u_k u_k^\top\right)p\right\|_{\psi_1} = \left\|\sum_{k=1}^{2r}\sigma_k\langle p, u_k\rangle^2\right\|_{\psi_1} \\ &\le \sum_{k=1}^{2r}\sigma_k\left\|\langle p, u_k\rangle^2\right\|_{\psi_1} \le \sum_{k=1}^{2r}\sigma_k\|\langle p, u_k\rangle\|_{\psi_2}^2 = \eta^2\sum_{k=1}^{2r}\sigma_k \le \sqrt{2r}\eta^2,\end{aligned}$$

where the first inequality follows since $\|\cdot\|_{\psi_1}$ is a norm, the second one follows since $\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$ [235, Lemma 2.7.7], and the third inequality holds since $\|\sigma\|_1 \le \sqrt{2r}\|\sigma\|_2$. This bound has two consequences, first $p^\top M p$ is a sub-exponential random variable with parameter $\sqrt{r}\eta^2$ and second $\mathbb{E}p^\top M p \le \sqrt{2r}\eta^2$ [235, Exercise 2.7.2]. Thus, we have proved (5.75).

To prove the deviation bound (5.76) we introduce the random variables

$$Y_i = \begin{cases} p_i^\top M p_i - \mathbb{E}p_i^\top M p_i & \text{if } i \notin \mathcal{I}, \text{ and} \\ -\left(p_i^\top M p_i - \mathbb{E}p_i^\top M p_i\right) & \text{otherwise.}\end{cases}$$

Since $p^\top Mp$ is sub-exponential, we have $\|Y_i\|_{\psi_1} \lesssim \sqrt{r}\eta^2$ for all i, see [235, Exercise 2.7.10]. Hence, Bernstein inequality for sub-exponential random variables [235, Theorem 2.8.2] gives the desired upper bound on $\mathbb{P}\left(\frac{1}{m}\left|\sum_{i=1}^m Y_i\right| \geq t\right)$. □

Applying Proposition 5.10.2 with with $\beta(r) \asymp \sqrt{r}\eta^2$ and $c(m,r) \asymp m/\eta^4 r$ now yields the result. □

Part 3 of Theorem 5.5.8 (Quadratic sensing II)

Lemma 5.10.5. *The random variable $|p^\top Mp - \tilde{p}^\top M\tilde{p}|$ is sub-exponential with parameter $C\eta^2$. Consequently,*

$$\alpha \leq \mathbb{E}|p^\top Mp - \tilde{p}^\top M\tilde{p}| \lesssim \eta^2. \tag{5.77}$$

Moreover, there exists a universal constant $c > 0$ such that for any $t \in [0, \eta^2]$ the deviation bound

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1]\right| \leq t \tag{5.78}$$

holds with probability at least $1 - 2\exp\left(-\frac{ct^2}{\eta^4}m\right)$.

Proof. Assumption 5.5.5 implies the lower bound in (5.77). To prove the upper bound, we will show that $\||p^\top Mp - \tilde{p}^\top M\tilde{p}^\top|\|_{\psi_1} \leq \eta^2$. By definition of the Orlicz norm $\|\|X\|\|_{\psi_1} = \|X\|_{\psi_1}$ for any random variable X, hence without loss of generality we may remove the absolute value. Recall that $M = \sum_{k=1}^{2r} \sigma_k u_k u_k^\top$ where σ_k and u_k are the kth singular values and vectors of M, respectively. Hence, the random variable of interest can be rewritten as

$$p^\top Mp - \tilde{p}^\top M\tilde{p}^\top \overset{d}{=} \sum_{k=1}^{2r} \sigma_k\left(\langle u_k, p\rangle^2 - \langle u_k, \tilde{p}\rangle^2\right). \tag{5.79}$$

By assumption the random variables $\langle u_k, p\rangle$ are η-sub-gaussian, this implies that $\langle u_k, p\rangle^2$ are η^2-sub-exponential, since $\|\langle u_k, p\rangle^2\|_{\psi_1} \leq \|\langle u_k, p\rangle\|_{\psi_2}^2$.

Recall the following characterization of the Orlicz norm for mean-zero random variables

$$\|X\|_{\psi_1} \leq Q \iff \mathbb{E}\exp(\lambda X) \leq \exp(\tilde{Q}^2\lambda^2) \text{ for all } \lambda \text{ satisfying } |\lambda| \leq 1/\tilde{Q}^2 \tag{5.80}$$

where the $Q \asymp \tilde{Q}$, see [235, Proposition 2.7.1]. To prove that the random variable (5.79) is sub-exponential we will exploit this characterization. Since each inner product squared $\langle u_k, p\rangle^2$ is sub-exponential, the equivalence implies the existence of a constant $c > 0$ for which the uniform bound

$$\mathbb{E}\exp(\lambda\langle u_k, p\rangle^2) \leq \exp\left(c\eta^4\lambda^2\right) \qquad \text{for all } k \in [2r] \text{ and } |\lambda| \leq 1/c\eta^4 \tag{5.81}$$

holds. Let λ be an arbitrary scalar with $|\lambda| \leq 1/c\eta^4$, then by expanding the moment generating function of (5.79) we get

$$\begin{aligned}
\mathbb{E}\exp\left(\lambda\sum_{k=1}^{2r}\sigma_k\left(\langle u_k, p\rangle^2 - \langle u_k, \tilde{p}\rangle^2\right)\right) &= \mathbb{E}\prod_{k=1}^{2r}\exp\left(\lambda\sigma_k\langle u_k, p\rangle^2\right)\exp\left(-\lambda\sigma_k\langle u_k, \tilde{p}\rangle^2\right) \\
&= \prod_{k=1}^{2r}\mathbb{E}\exp\left(\lambda\sigma_k\langle u_k, p\rangle^2\right)\mathbb{E}\exp\left(-\lambda\sigma_k\langle u_k, \tilde{p}\rangle^2\right) \\
&\leq \prod_{k=1}^{2r}\exp\left((c\eta)^2\lambda^2\sigma_k^2\right)\exp\left(c\eta^4\lambda^2\sigma_k^2\right) \\
&= \exp\left(2c\eta^4\lambda^2\sum_{k=1}^{2r}\sigma_k^2\right) = \exp\left(2c\eta^4\lambda^2\right).
\end{aligned}$$

where the inequality follows by (5.81) and the last relation follows since σ is unit norm. Combining this with (5.80) gives

$$\|\|p^\top Mp - \tilde{p}^\top M\tilde{p}^\top\|\|_{\psi_1} \lesssim \eta^2.$$

This bound has two consequences, first $|p^\top Mp - \tilde{p}^\top M\tilde{p}^\top|$ is a sub-exponential random variable with parameter $C\eta^2$ and second $\mathbb{E}|p^\top Mp - \tilde{p}^\top M\tilde{p}^\top| \leq C\eta^2$ [235, Exercise 2.7.2]. Thus, we have proved (5.77).

To prove the deviation bound (5.78) we introduce the random variables

$$Y_i = \begin{cases} \mathcal{A}(M)_i - \mathbb{E}\mathcal{A}(M)_i & \text{if } i \notin \mathcal{I}, \text{ and} \\ -\left(\mathcal{A}(M)_i - \mathbb{E}\mathcal{A}(M)_i\right) & \text{otherwise.} \end{cases}$$

The sub-exponentiality of $\mathcal{A}(M)_i$ implies $\|Y_i\|_{\psi_1} \lesssim \eta^2$ for all i, see [235, Exercise 2.7.10]. Hence, Bernstein inequality for sub-exponential random variables [235, Theorem 2.8.2] gives the desired upper bound on $\mathbb{P}\left(\frac{1}{m}\left|\sum_{i=1}^m Y_i\right| \geq t\right)$. □

Applying Proposition 5.10.2 with $\beta(r) \asymp \eta^2$ and $c(m,r) \asymp m/\eta^4$ now yields the result. □

Part 4 of Theorem 5.5.8 (Bilinear sensing)

Lemma 5.10.6. *The random variable $|p^\top Mq|$ is sub-exponential with parameter $C\eta^2$. Consequently,*

$$\alpha \leq \mathbb{E}|p^\top Mq| \lesssim \eta^2. \tag{5.82}$$

Moreover, there exists a universal constant $c > 0$ such that for any $t \in [0, \eta^2]$ the deviation bound

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1]\right| \leq t \tag{5.83}$$

holds with probability at least $1 - 2\exp\left(-\frac{ct^2}{\eta^4}m\right)$.

Proof. As before the lower bound in (5.82) is implied by Assumption 5.5.6. To prove the upper bound, we will show that $\|\|p^\top Mq\|\|_{\psi_1} \leq \eta^2$. By definition of the Orlicz norm $\|\|X\|\|_{\psi_1} = \|X\|_{\psi_1}$ for any random variable X, hence we may remove the absolute value. Recall that $M = \sum_{k=1}^{2r} \sigma_k u_k v_k^\top$ where σ_k and (u_k, v_k) are the kth

singular values and vectors of M, respectively. Hence, the random variable of interest can be rewritten as

$$p^\top M q \overset{d}{=} \sum_{k=1}^{2r} \sigma_k \langle p, u_k \rangle \langle v_k, q \rangle. \tag{5.84}$$

By assumption the random variables $\langle p, u_k \rangle$ and $\langle v_k, q \rangle$ are η-sub-gaussian, this implies that $\langle p, u_k \rangle \langle v_k, q \rangle$ are η^2-sub-exponential.

To prove that the random variable (5.84) is sub-exponential we will again use (5.80). Since each random variable $\langle p, u_k \rangle \langle v_k, q \rangle$ is sub-exponential, the equivalence implies the existence of a constant $c > 0$ for which the uniform bound

$$\mathbb{E} \exp(\lambda \langle p, u_k \rangle \langle v_k, q \rangle) \leq \exp\left(c\eta^4 \lambda^2\right) \qquad \text{for all } k \in [2r] \text{ and } |\lambda| \leq 1/c\eta^4 \tag{5.85}$$

holds. Let λ be an arbitrary scalar with $|\lambda| \leq 1/c\eta^4$, then by expanding the moment generating function of (5.84) we get

$$\begin{aligned} \mathbb{E} \exp\left(\lambda \sum_{k=1}^{2r} \sigma_k \langle p, u_k \rangle \langle v_k, q \rangle\right) &= \prod_{k=1}^{2r} \mathbb{E} \exp\left(\lambda \sigma_k \langle p, u_k \rangle \langle v_k, q \rangle\right) \\ &\leq \exp\left(2c\eta^4 \lambda^2 \sum_{k=1}^{r} \sigma_k^2\right) = \exp\left(2c\eta^4 \lambda^2\right). \end{aligned}$$

where the inequality follows by (5.85) and the last relation follows since σ is unitary. Combining this with (5.80) gives

$$|||p^\top M q|||_{\psi_1} \lesssim \eta^2.$$

Thus, we have proved (5.82).

Once again, to show the deviation bound (5.83) we introduce the random variables

$$Y_i = \begin{cases} |p_i^\top M q_i| - \mathbb{E}|p_i^\top M q_i| & \text{if } i \notin \mathcal{I}, \text{ and} \\ -\left(|p_i^\top M q_i| - \mathbb{E}|p_i^\top M q_i|\right) & \text{otherwise.} \end{cases}$$

and apply Bernstein's inequality for sub-exponential random variables [235, Theorem 2.8.2] to get the stated upper bound on $\mathbb{P}\left(\frac{1}{m}\left|\sum_{i=1}^{m} Y_i\right| \geq t\right)$. □

Applying Proposition 5.10.2 with $\beta(r) \asymp \eta^2$ and $c(m, r) \asymp m/\eta^4$ now yields the result. □

Proof of Proposition 5.10.2

Choose $\epsilon \in (0, \sqrt{2})$ and let $\mathcal{N}$ be the $(\epsilon/\sqrt{2})$-net guaranteed by Lemma 5.10.7. Pick some $t \in (0, K]$ so that (5.70) can hold, we will fix the value of this parameter later in the proof. Let $\mathcal{E}$ denote the event that the following two estimates hold for all matrices in $M \in \mathcal{N}$:

$$\frac{1}{m}\left|\|\mathcal{A}_{I^c}(M)\|_1 - \|\mathcal{A}_I(M)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{I^c}(M)\|_1 - \|\mathcal{A}_I(M)\|_1\right]\right| \leq t, \tag{5.86}$$

$$\frac{1}{m}\left|\|\mathcal{A}(M)\|_1 - \mathbb{E}\left[\|\mathcal{A}(M)\|_1\right]\right| \leq t. \tag{5.87}$$

Throughout the proof, we will assume that the event $\mathcal{E}$ holds. We will estimate the probability of $\mathcal{E}$ at the end of the proof. Meanwhile, seeking to establish RIP, define the quantity

$$c_2 := \sup_{M \in S_{2r}} \frac{1}{m}\|\mathcal{A}(M)\|_1.$$

We aim first to provide a high probability bound on c_2.

Let $M \in S_{2r}$ be arbitrary and let $M_\star$ be the closest point to M in $\mathcal{N}$. Then we have

$$\begin{aligned}
\frac{1}{m}\|\mathcal{A}(M)\|_1 &\leq \frac{1}{m}\|\mathcal{A}(M_\star)\|_1 + \frac{1}{m}\|\mathcal{A}(M - M_\star)\|_1 \\
&\leq \frac{1}{m}\mathbb{E}\|\mathcal{A}(M_\star)\|_1 + t + \frac{1}{m}\|\mathcal{A}(M - M_\star)\|_1 \qquad (5.88) \\
&\leq \frac{1}{m}\mathbb{E}\|\mathcal{A}(M)\|_1 + t + \frac{1}{m}\left(\mathbb{E}\|\mathcal{A}(M - M_\star)\|_1 + \|\mathcal{A}(M - M_\star)\|_1\right), \qquad (5.89)
\end{aligned}$$

where (5.88) follows from (5.87) and (5.89) follows from the triangle inequality. To simplify the third term in (5.89), using SVD, we deduce that there exist two orthogonal matrices M_1, M_2 of rank at most $2r$ satisfying $M - M_\star = M_1 + M_2$. With this decomposition in hand, we compute

$$\begin{aligned}\frac{1}{m}\|\mathcal{A}(M - M_\star)\|_1 &\leq \frac{1}{m}\|\mathcal{A}(M_1)\|_1 + \frac{1}{m}\|\mathcal{A}(M_2)\|_1 \\ &\leq c_2(\|M_1\|_F + \|M_2\|_F) \leq \sqrt{2}c_2\|M - M_\star\|_F \leq c_2\epsilon, \end{aligned} \tag{5.90}$$

where the second inequality follows from the definition of c_2 and the estimate $\|M_1\|_F + \|M_2\|_F \leq \sqrt{2}\|(M_1, M_2)\|_F = \sqrt{2}\|M_1 + M_2\|_F$. Thus, we arrive at the bound

$$\frac{1}{m}\|\mathcal{A}(M)\|_1 \leq \frac{1}{m}\mathbb{E}\|\mathcal{A}(M)\|_1 + t + 2c_2\epsilon. \tag{5.91}$$

As M was arbitrary, we may take the supremum of both sides of the inequality, yielding $c_2 \leq \frac{1}{m}\sup_{M\in S_{2r}} \mathbb{E}\|\mathcal{A}(M)\|_1 + t + 2c_2\epsilon$. Rearranging yields the bound

$$c_2 \leq \frac{\frac{1}{m}\sup_{M\in S_{2r}} \mathbb{E}\|\mathcal{A}(M)\|_1 + t}{1 - 2\epsilon}.$$

Assuming that $\epsilon \leq 1/4$, we further deduce that

$$c_2 \leq \bar{\sigma} := \frac{2}{m}\sup_{M\in S_{2r}} \mathbb{E}\|\mathcal{A}(M)\|_1 + 2t \leq 2\beta(r) + 2t, \tag{5.92}$$

establishing that the random variable c_2 is bounded by $\bar{\sigma}$ in the event $\mathcal{E}$.

Now let $\hat{\mathcal{I}}$ denote either $\hat{\mathcal{I}} = \emptyset$ or $\hat{\mathcal{I}} = \mathcal{I}$. We now provide a uniform lower

bound on $\frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1$. Indeed,

$$\begin{aligned}
&\frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1 \\
&= \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M_\star) + \mathcal{A}_{\hat{\mathcal{I}}^c}(M - M_\star)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(M_\star) + \mathcal{A}_{\hat{\mathcal{I}}}(M - M_\star)\|_1 \\
&\geq \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M_\star)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(M_\star)\|_1 - \frac{1}{m}\|\mathcal{A}(M - M_\star)\|_1 && (5.93) \\
&\geq \frac{1}{m}\mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M_\star)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M_\star)\|_1\right] - t - \frac{1}{m}\|\mathcal{A}(M - M_\star)\|_1 && (5.94) \\
&\geq \frac{1}{m}\mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1\right] - t - \frac{1}{m}\left(\mathbb{E}\|\mathcal{A}(M - M_\star)\|_1 + \|\mathcal{A}(M - M_\star)\|_1\right) && (5.95) \\
&\geq \frac{1}{m}\mathbb{E}\left[\|\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1\right] - t - 2\bar{\sigma}\epsilon, && (5.96)
\end{aligned}$$

where (5.93) uses the forward and reverse triangle inequalities, (5.94) follows from (5.86), the estimate (5.95) follows from the forward and reverse triangle inequalities, and (5.96) follows from (5.90) and (5.92). Switching the roles of $\mathcal{I}$ and $\mathcal{I}^c$ in the above sequence of inequalities, and choosing $\epsilon = t/4\bar{\sigma}$, we deduce

$$\frac{1}{m}\sup_{M\in S_{2r}}\left|\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1\right]\right| \leq \frac{3t}{2}.$$

In particular, setting $\hat{\mathcal{I}} = \emptyset$, we deduce

$$\frac{1}{m}\sup_{M\in S_{2r}}\left|\|\mathcal{A}(M)\|_1 - \mathbb{E}\left[\|\mathcal{A}(M)\|_1\right]\right| \leq \frac{3t}{2} \tag{5.97}$$

and therefore using (5.69), we conclude the RIP property

$$\alpha - \frac{3t}{2} \leq \frac{1}{m}\|\mathcal{A}(M)\|_1 \lesssim \beta(r) + \frac{3t}{2}, \qquad \forall X \in S_{2r}. \tag{5.98}$$

Next, let $\hat{\mathcal{I}} = \mathcal{I}$ and note that

$$\frac{1}{m}\mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1\right] = \frac{|\mathcal{I}^c| - |\mathcal{I}|}{m}\cdot\mathbb{E}|\mathcal{A}(M)_i| \geq \left(1 - \frac{2|\mathcal{I}|}{m}\right)\alpha,$$

where the equality follows by assumption (1). Therefore every $M \in S_{2r}$ satisfies

$$\frac{1}{m}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1\right] \geq \left(1 - \frac{2|\mathcal{I}|}{m}\right)\alpha - \frac{3t}{2}. \tag{5.99}$$

Setting $t = \frac{2}{3}\min\{\alpha, \alpha(1-2|\mathcal{I}|/m)/2\} = \frac{1}{3}\alpha(1-2|\mathcal{I}|/m)$ in (5.98) and (5.99), we deduce the claimed estimates (5.71) and (5.72). Finally, let us estimate the probability of $\mathcal{E}$. Using the union bound and Lemma 5.10.7 yields

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^c) &\le \sum_{M\in\mathcal{N}} \mathbb{P}\{\text{(5.86) or (5.87) fails at } M\} \\
&\le 4|\mathcal{N}| \exp\left(-t^2 c(m,r)\right) \\
&\le 4\left(\frac{9}{\epsilon}\right)^{2(d_1+d_2+1)r} \exp\left(-t^2 c(m,r)\right) \\
&= 4\exp\left(2(d_1+d_2+1)r\ln(9/\epsilon) - t^2 c(m,r)\right)
\end{aligned}$$

where $c(m,r)$ is the function guaranteed by assumption (3).

By (5.69) we get $1/\epsilon = 4\bar{\sigma}/t \lesssim 2 + \beta(r)/(1-2|\mathcal{I}|/m)$. Then we deduce

$$\mathbb{P}(\mathcal{E}^c) \le 4\exp\left(c_1(d_2+d_2+1)r\ln\left(c_2 + \frac{c_2\beta(r)}{1-2|\mathcal{I}|/m}\right) - \frac{\alpha^2}{9}(1-\frac{2|\mathcal{I}|}{m})^2 c(m,r)\right).$$

Hence as long as $c(m,r) \ge \frac{9c_1(d_1+d_2+1)r^2\ln\left(c_2+\frac{c_2\beta(r)}{1-2|\mathcal{I}|/m}\right)}{\alpha^2\left(1-\frac{2|\mathcal{I}|}{m}\right)^2}$, we can be sure

$$\mathbb{P}(\mathcal{E}^c) \le 4\exp\left(-\frac{\alpha^2}{18}\left(1-\frac{2|\mathcal{I}|}{m}\right)^2 c(m,r)\right).$$

Proving the desired result. □

5.10.3 Proof in Section 5.6

Proof of Lemma 5.6.4

Define $P(x,y) = a\|y-x\|_2^2 + b\|y-x\|_2$. Fix an iteration k and choose $x^* \in \mathrm{proj}_{\mathcal{X}^*}(x_k)$. Then the estimate holds:

$$f(x_{k+1}) \le f_{x_k}(x_{k+1}) + P(x_{k+1}, x_k) \le f_{x_k}(x^*) + P(x^*, x_k) \le f(x^*) + 2P(x^*, x_k).$$

Rearranging and using the sharpness and approximation accuracy assumptions, we deduce

$$\mu \cdot \text{dist}(x_{k+1}, \mathcal{X}^*) \leq 2(a \cdot \text{dist}^2(x, \mathcal{X}^*) + b \cdot \text{dist}(x, \mathcal{X}^*)) = 2(b + a\text{dist}(x, \mathcal{X}^*))\text{dist}(x, \mathcal{X}^*).$$

The result follows.

Proof of Theorem 5.6.6

First notice that for any y, we have $\partial f(y) = \partial f_y(y)$. Therefore, since f_y is a convex function, we have that for all $x, y \in \mathcal{X}$ and $v \in \partial f(y)$, the bound

$$f(y) + \langle v, x - y \rangle = f_y(y) + \langle v, x - y \rangle \leq f_y(x) \leq f(x) + a\|x - y\|_F^2 + b\|x - y\|_F. \quad (5.100)$$

Consequently, given that $\text{dist}(x_i, \mathcal{X}^*) \leq \gamma \cdot \frac{\mu - 2b}{2a}$, we have

$$\begin{aligned}
\|x_{i+1} - x^*\|^2 &= \left\| \text{proj}_{\mathcal{X}} \left(x_i - \tfrac{f(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2} \zeta_i \right) - \text{proj}_{\mathcal{X}}(x^*) \right\|^2 \\
&\leq \left\| (x_i - x^*) - \tfrac{f(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2} \zeta_i \right\|^2 \quad (5.101) \\
&= \|x_i - x^*\|^2 + \frac{2(f(x_i) - \min_{\mathcal{X}} f)}{\|\zeta_i\|^2} \cdot \langle \zeta_i, x^* - x_i \rangle + \frac{(f(x_i) - f(x^*))^2}{\|\zeta_i\|^2} \\
&\leq \|x_i - x^*\|^2 + \frac{2(f(x_i) - \min f)}{\|\zeta_i\|^2} \left(f(x^*) - f(x_i) + a\|x_i - x^*\|^2 + b\|x_i - x^*\| \right) \\
&\qquad + \frac{(f(x_i) - f(x^*))^2}{\|\zeta_i\|^2} \quad (5.102) \\
&= \|x_i - x^*\|^2 + \frac{f(x_i) - \min f}{\|\zeta_i\|^2} \left(2a\|x_i - x^*\|^2 + 2b\|x_i - x^*\| - (f(x_i) - f(x^*)) \right) \\
&\leq \|x_i - x^*\|^2 + \frac{f(x_i) - \min f}{\|\zeta_i\|^2} \left(a\|x_i - x^*\|^2 - (\mu - 2b)\|x_i - x^*\| \right) \quad (5.103) \\
&= \|x_i - x^*\|^2 + \frac{2a(f(x_i) - \min f)}{\|\zeta_i\|^2} \left(\|x_i - x^*\| - \frac{\mu - 2b}{2a} \right) \|x_i - x^*\| \\
&\leq \|x_i - x^*\|^2 - \frac{(1 - \gamma)(\mu - 2b)(f(x_i) - \min f)}{\|\zeta_i\|^2} \cdot \|x_i - x^*\| \quad (5.104) \\
&\leq \left(1 - \frac{(1 - \gamma)\mu(\mu - 2b)}{\|\zeta_i\|^2} \right) \|x_i - x^*\|^2. \quad (5.105)
\end{aligned}$$

Here, the estimate (5.101) follows from the fact that the projection $\mathrm{proj}_{\mathcal{X}}(\cdot)$ is nonexpansive, (5.102) uses the bound in (5.100), (5.104) follow from the estimate $\mathrm{dist}(x_i, \mathcal{X}^*) \leq \gamma \cdot \frac{\mu - 2b}{2a}$, while (5.103) and (5.105) use local sharpness. The result then follows by the upper bound $\|\zeta_i\| \leq L$.

5.10.4 Proofs in Section 5.7

Proof of Lemma 5.7.1

The inequality can be established using an argument similar to that for bounding the T_3 term in [53, Section 6.6]. We provide the proof below for completeness. Define the shorthand $\Delta_S := S - S_\sharp$ and $\Delta_X = X - X_\sharp$, and let $e_j \in \mathbb{R}^d$ denote the j-th standard basis vector of $\mathbb{R}^d$. Simple algebra gives

$$\begin{aligned} |\langle S - S_\sharp, XX^\top - X_\sharp X_\sharp^\top \rangle| &= |2\langle \Delta_S, \Delta_X X_\sharp^\top \rangle + \langle \Delta_S, \Delta_X \Delta_X^\top \rangle| \\ &\leq \left(2\|X_\sharp^\top \Delta_S\|_F + \|\Delta_X^\top \Delta_S\|_F \right) \cdot \|\Delta_X\|_F. \end{aligned}$$

We claim that $\|\Delta_S e_j\|_1 \leq 2\sqrt{k}\|\Delta_S e_j\|_2$ for each $j \in [d]$. To see this, fix any $j \in [d]$ and let $v := S e_j$, $v^* := S_\sharp e_j$, and $T := \mathrm{support}(v^*)$. We have

$$\begin{aligned} \|v_T^*\|_1 = \|v^*\|_1 &\geq \|v\|_1 && S \in \mathcal{S} \\ &= \|v_T\|_1 + \|v_{T^c}\|_1 && \text{decomposability of } \ell_1 \text{ norm} \\ &= \|v_T^* + (v - v^*)_T\|_1 + \|(v - v^*)_{T^c}\|_1 \\ &\geq \|v_T^*\|_1 - \|(v - v^*)_T\|_1 + \|(v - v^*)_{T^c}\|_1. && \text{reverse triangle inequality} \end{aligned}$$

Rearranging terms gives $\|(v - v^*)_{T^c}\|_1 \leq \|(v - v^*)_T\|_1$, whence

$$\begin{aligned} \|v - v^*\|_1 = \|(v - v^*)_T\|_1 + \|(v - v^*)_{T^c}\|_1 &\leq 2\|(v - v^*)_T\|_1 \\ &\leq 2\sqrt{k}\|(v - v^*)_T\|_2 \leq 2\sqrt{k}\|v - v^*\|_2, \end{aligned}$$

where step the second inequality holds because $|T| \leq k$ by assumption. The claim follows from noting that $v - v^* = \Delta_S e_j$.

Using the claim, we get that

$$\|X_\sharp^\top \Delta_S\|_F = \sqrt{\sum_{j\in[d]} \|X_\sharp^\top \Delta_S e_j\|_2^2} \leq \sqrt{\sum_{j\in[d]} \|X_\sharp\|_{2,\infty}^2 \|\Delta_S e_j\|_1^2}$$
$$\leq \|X_\sharp\|_{2,\infty} \sqrt{\sum_{j\in[d]} 4k\|\Delta_S e_j\|_2^2} \leq 2\sqrt{\frac{\nu r k}{d}}\|\Delta_S\|_F.$$

Using a similar argument and the fact that $\|\Delta_X\|_{2,\infty} \leq \|X\|_{2,\infty} + \|X_\sharp\|_{2,\infty} \leq 3\sqrt{\frac{\nu r}{d}}$, we obtain

$$\|\Delta_X^\top \Delta_S\|_F \leq 6\sqrt{\frac{\nu r k}{d}}\|\Delta_S\|_F.$$

Putting everything together, we have

$$|\langle S - S^*, XX^\top - X_\sharp X_\sharp^\top\rangle| \leq \left(2 \cdot 2\sqrt{\frac{\nu r k}{d}}\|\Delta_S\|_F + 6\sqrt{\frac{\nu r k}{d}}\|\Delta_S\|_F\right) \cdot \|\Delta_X\|_F.$$

The claim follows.

Proof of Theorem 5.7.5

Without loss of generality, suppose that x is closer to $\bar{x}$ than to $-\bar{x}$. Consider the following expression:

$$\begin{aligned}
\|\bar{x}(x-\bar{x})^\top + (x-\bar{x})\bar{x}^\top\|_1 &= \sup_{\|V\|_\infty=1, V^\top=V} \mathrm{Tr}((\bar{x}(x-\bar{x})^\top + (x-\bar{x})\bar{x}^\top)V) \\
&= \sup_{\|V\|_\infty=1, V^\top=V} \mathrm{Tr}(\bar{x}x^\top V + x\bar{x}^\top V - 2\bar{x}\bar{x}^\top V) \\
&= \sup_{\|V\|_\infty=1, V^\top=V} \mathrm{Tr}(x^\top V\bar{x} + \bar{x}^\top Vx - 2\bar{x}^\top V\bar{x}) \\
&= 2 \sup_{\|V\|_\infty=1, V^\top=V} \mathrm{Tr}(x^\top V\bar{x} - \bar{x}^\top V\bar{x}) \\
&= 2 \sup_{\|V\|_\infty=1, V^\top=V} \mathrm{Tr}((x-\bar{x})^\top V\bar{x}) \\
&= 2 \sup_{\|V\|_\infty=1, V^\top=V} \mathrm{Tr}(\bar{x}(x-\bar{x})^\top V).
\end{aligned}$$

We now produce a few different lower bounds by testing against different V. In what follows, we set $a = \sqrt{2} - 1$, i.e., the positive solution of the equation $1 - a^2 = 2a$.

Case 1: Suppose that $|(x-\bar{x})^\top \mathrm{sign}(\bar{x})| \geq a\|x-\bar{x}\|_1$. Then set $\bar{V} = \mathrm{sign}((x-\bar{x})^\top \mathrm{sign}(\bar{x})) \cdot \mathrm{sign}(\bar{x})\mathrm{sign}(\bar{x})^\top$, to get

$$\begin{aligned}
\|\bar{x}(x-\bar{x})^\top + (x-\bar{x})\bar{x}^\top\|_1 &\geq 2\mathrm{Tr}(\bar{x}(x-\bar{x})^\top \bar{V}) \\
&= 2\mathrm{sign}((x-\bar{x})^\top \mathrm{sign}(\bar{x})) \cdot \mathrm{Tr}((x-\bar{x})^\top \mathrm{sign}(\bar{x})\mathrm{sign}(\bar{x})^\top \bar{x}) \\
&= 2\|\bar{x}\|_1 \mathrm{sign}((x-\bar{x})^\top \mathrm{sign}(\bar{x})) \cdot (x-\bar{x})^\top \mathrm{sign}(\bar{x}) \\
&\geq 2a\|\bar{x}\|_1 \|x-\bar{x}\|_1
\end{aligned}$$

Case 2: Suppose that $|\text{sign}(x-\bar{x})^\top \bar{x}| \geq a\|\bar{x}\|_1$. Then set $\bar{V} = \text{sign}(\text{sign}(x-\bar{x})^\top \bar{x}) \cdot \text{sign}(x-\bar{x})\text{sign}(x-\bar{x})^\top$, to get

$$\begin{aligned}
\|\bar{x}(x-\bar{x})^\top + (x-\bar{x})\bar{x}^\top\|_1 &\geq 2\text{Tr}(\bar{x}(x-\bar{x})^\top \bar{V}) \\
&= 2\text{sign}(\text{sign}(x-\bar{x})^\top \bar{x}) \cdot \text{Tr}((x-\bar{x})^\top \text{sign}(x-\bar{x})\text{sign}(x-\bar{x})^\top \bar{x}) \\
&= 2\|x-\bar{x}\|_1 \text{sign}(\text{sign}(x-\bar{x})^\top \bar{x}) \cdot \text{sign}(x-\bar{x})^\top \bar{x} \\
&\geq 2a\|\bar{x}\|_1 \|x-\bar{x}\|_1
\end{aligned}$$

Case 3: Suppose that

$$|(x-\bar{x})^\top \text{sign}(\bar{x})| \leq a\|x-\bar{x}\|_1 \qquad \text{and} \qquad |\text{sign}(x-\bar{x})^\top \bar{x}| \leq a\|\bar{x}\|_1.$$

Define $\bar{V} = \frac{1}{2}(\text{sign}(\bar{x}(x-\bar{x})^\top) + \text{sign}((x-\bar{x})\bar{x}^\top))$. Observe that

$$\text{Tr}(\bar{x}(x-\bar{x})^\top \text{sign}(\bar{x}(x-\bar{x})^\top)) = (x-\bar{x})^\top \text{sign}(\bar{x})\text{sign}(x-\bar{x})^\top \bar{x} \geq -a^2\|\bar{x}\|_1\|x-\bar{x}\|_1$$

and

$$\text{Tr}(\bar{x}(x-\bar{x})^\top \text{sign}((x-\bar{x})\bar{x}^\top)) = \text{Tr}(\bar{x}(x-\bar{x})^\top \text{sign}(x-\bar{x})\text{sign}(\bar{x}^\top)) = \|\bar{x}\|_1\|x-\bar{x}\|_1.$$

Putting these two bounds together, we find that

$$\|\bar{x}(x-\bar{x})^\top + (x-\bar{x})\bar{x}^\top\|_1 \geq 2\text{Tr}(\bar{x}(x-\bar{x})^\top \bar{V}) = (1-a^2)\|\bar{x}\|_1\|x-\bar{x}\|_1.$$

Altogether, we find that

$$\begin{aligned}
F(x) &= \|xx^\top - \bar{x}\bar{x}^\top\|_1 \\
&= \|\bar{x}(x-\bar{x})^\top + (x-\bar{x})\bar{x}^\top + (x-\bar{x})(x-\bar{x})^\top\|_1 \\
&\geq \|\bar{x}(x-\bar{x})^\top + (x-\bar{x})\bar{x}^\top\|_1 - \|(x-\bar{x})(x-\bar{x})^\top\|_1 \\
&\geq 2a\|\bar{x}\|_1\|x-\bar{x}\|_1 - \|(x-\bar{x})\|_1^2 \\
&= 2a\|\bar{x}\|_1 \left(1 - \frac{\|x-\bar{x}\|_1}{2a\|\bar{x}\|_1}\right)\|x-\bar{x}\|_1,
\end{aligned}$$

as desired.

Proof of Lemma 5.7.7

We start by stating a claim we will use to prove the lemma. Let us introduce some notation. Consider the set

$$S = \left\{ (\Delta_+, \Delta_-) \in \mathbf{R}^{d\times r} \times \mathbf{R}^{d\times r} \mid \|\Delta_+\|_{2,\infty} \le (1+C)\sqrt{\frac{\nu r}{d}}\|X_\sharp\|_{op}, \|\Delta_-\|_{2,1} \neq 0 \right\}.$$

Define the random variable

$$Z = \sup_{(\Delta_+,\Delta_-)\in S} \Big| \frac{1}{\|\Delta_-\|_{2,1}} \sum_{i,j=1}^{d} \delta_{ij} |\langle \Delta_{-,i}, \Delta_{+,j}\rangle + \langle \Delta_{+,i}, \Delta_{-,j}\rangle| \\ - \mathbb{E}\frac{1}{\|\Delta_-\|_{2,1}} \sum_{i,j=1}^{d} \delta_{ij} |\langle \Delta_{-,i}, \Delta_{+,j}\rangle + \langle \Delta_{+,i}, \Delta_{-,j}\rangle| \Big|.$$

Claim 2. *There exist constants $c_2, c_3 > 0$ such that with probability at least $1 - \exp(-c_2 \log d)$*

$$Z \le c_3 C \sqrt{\tau \nu r \log d}\, \left\|X_\sharp\right\|_{op}.$$

Before proving this claim, let us show how it implies the theorem. Let

$$R \in \operatorname*{arg\,min}_{\hat{R}^\top \hat{R} = I} \|X - X_\sharp \hat{R}\|_{2,1}.$$

Set $\Delta_- = X - X_\sharp R$ and $\Delta_+ = X + X_\sharp R$. Notice that

$$\|\Delta_+\|_{2,\infty} \le \|X\|_{2,\infty} + \|X_\sharp\|_{2,\infty} \le (1+C)\|X_\sharp\|_{2,\infty} \le \sqrt{\frac{\nu r}{d}}(1+C)\|X_\sharp\|_{op}.$$

Therefore, because $(\Delta_+, \Delta_-) \in S$ and

$$\frac{1}{\|\Delta_-\|_{2,1}} \sum_{i,j=1}^{d} \delta_{ij} |\langle X_i, X_j\rangle - \langle (X_\sharp)_i, (X_\sharp)_j\rangle| = \frac{1}{\|\Delta_-\|_{2,1}} \sum_{i,j=1}^{d} \delta_{ij} |\langle \Delta_{-,i}, \Delta_{+,j}\rangle + \langle \Delta_{+,i}, \Delta_{-,j}\rangle|,$$

we have that

$$\sum_{i,j=1}^{d} \delta_{ij} |\langle X_i, X_j\rangle - \langle (X_\sharp)_i, (X_\sharp)_j\rangle| \le \tau \|XX^\top - X_\sharp X_\sharp^\top\|_1 + c_3 C \sqrt{\tau\nu r \log d}\|X_\sharp\|_{op}\|X - X_\sharp R\|_{2,1} \\ \le \left(\tau + \frac{c_3 C \sqrt{\tau \nu r \log d}}{c}\|X_\sharp\|_{op}\right)\|XX^\top - X_\sharp X_\sharp^\top\|_1,$$

where the last line follows by Conjecture 5.7.6. This proves the desired result.

Proof of the Claim. Our goal is to show that the random variable Z is highly concentrated around its mean. We may apply the standard symmetrization inequality [25, Lemma 11.4] to bound the expectation $\mathbb{E}Z$ as follows:

$$\begin{aligned}\mathbb{E}Z &\leq 2\mathbb{E}\sup_{(\Delta_+,\Delta_-)\in S}\left|\frac{1}{||\Delta_-||_{2,1}}\sum_{i,j=1}^{d}\varepsilon_{ij}\delta_{ij}|\langle\Delta_{-,i},\Delta_{+,j}\rangle+\langle\Delta_{+,i},\Delta_{-,j}\rangle|\right| \\ &\leq \underbrace{2\mathbb{E}\sup_{(\Delta_+,\Delta_-)\in S}\left|\frac{1}{||\Delta_-||_{2,1}}\sum_{i,j=1}^{d}\varepsilon_{ij}\delta_{ij}|\langle\Delta_{-,i},\Delta_{+,j}\rangle|\right|}_{T_1}+\underbrace{2\mathbb{E}\sup_{(\Delta_+,\Delta_-)\in S}\left|\frac{1}{||\Delta_-||_{2,1}}\sum_{i,j=1}^{d}\varepsilon_{ij}\delta_{ij}|\langle\Delta_{+,i},\Delta_{-,j}\rangle|\right|}_{T_2}.\end{aligned}$$

Observing that T_1 and T_2 can both be bounded by

$$\begin{aligned}\max\{T_1,T_2\} &\leq 2\sup_{(\Delta_+,\Delta_-)\in S}\frac{1}{||\Delta_-||_{2,1}}||\Delta_+\Delta_-^\top||_{2,\infty}\mathbb{E}\max_j\left|\sum_{i=1}^{d}\varepsilon_{ij}\delta_{ij}\right| \\ &\leq 2\sup_{(\Delta_+,\Delta_-)\in S}||\Delta_+||_{2,\infty}\mathbb{E}\max_j\left|\sum_{i=1}^{d}\varepsilon_{ij}\delta_{ij}\right| \\ &\leq 2(1+C)\sqrt{\frac{\nu r}{d}}||X_\sharp||_{op}\mathbb{E}\max_j\left|\sum_{i=1}^{d}\varepsilon_{ij}\delta_{ij}\right| \\ &\lesssim C\sqrt{\frac{\nu r}{d}}||X_\sharp||_{op}(\sqrt{\tau d\log d}+\log d),\end{aligned}$$

where the final inequality follows from Bernstein's inequality and a union bound, we find that

$$\mathbb{E}Z \lesssim C\sqrt{\frac{\nu r}{d}}||X_\sharp||_{op}(\sqrt{\tau d\log d}+\log d).$$

To prove that Z is well concentrated around $\mathbb{E}Z$, we apply Theorem 5.10.9. To apply this theorem, we set $\mathcal{S}=S$ and define the collection $(Z_{ij,s})_{ij,s\in\mathcal{S}}$, where $s=(\Delta_+,\Delta_-)$ by

$$\begin{aligned}Z_{ij,s} &= \frac{1}{||\Delta_-||_{2,1}}\delta_{ij}|\langle\Delta_{-,i},\Delta_{+,j}\rangle+\langle\Delta_{+,i},\Delta_{-,j}\rangle|-\mathbb{E}\frac{1}{||\Delta_-||_{2,1}}\delta_{ij}|\langle\Delta_{-,i},\Delta_{+,j}\rangle+\langle\Delta_{+,i},\Delta_{-,j}\rangle| \\ &= \frac{(\delta_{ij}-\tau)}{||\Delta_-||_{2,1}}|\langle\Delta_{-,i},\Delta_{+,j}\rangle+\langle\Delta_{+,i},\Delta_{-,j}\rangle|.\end{aligned}$$

We also bound

$$b = \sup_{ij,s\in\mathcal{S}} |Z_{ij,s}| \leq \sup_{ij,(\Delta_+,\Delta_-)\in S} \left| \frac{(\delta_{ij}-\tau)}{||\Delta_-||_{2,1}} (||\Delta_{-,i}||_F ||\Delta_{+,j}||_F + ||\Delta_{+,i}||_F ||\Delta_{-,j}||_F) \right|$$
$$\leq (1+C)\sqrt{\frac{\nu r}{d}} ||X_\sharp||_{op} \sup_{ij,(\Delta_+,\Delta_-)\in S} \left| \frac{1}{||\Delta_-||_{2,1}} (||\Delta_{-,i}||_F + ||\Delta_{-,j}||_F) \right| \leq 2C\sqrt{\frac{\nu r}{d}} ||X_\sharp||_{op}$$

and

$$\sigma^2 = \sup_{(\Delta_+,\Delta_-)\in S} \mathbb{E} \frac{1}{||\Delta_-||_{2,1}^2} \sum_{ij=1}^{d} (\delta_{ij}-\tau)^2 |\langle \Delta_{-,i}, \Delta_{+,j}\rangle + \langle \Delta_{+,i}, \Delta_{-,j}\rangle|^2$$
$$\leq \tau \sup_{(\Delta_+,\Delta_-)\in S} \frac{1}{||\Delta_-||_{2,1}^2} \sum_{ij=1}^{d} (||\Delta_{-,i}||_F ||\Delta_{+,j}||_F + ||\Delta_{+,i}||_F ||\Delta_{-,j}||_F)^2$$
$$\leq \tau \sup_{(\Delta_+,\Delta_-)\in S} \frac{4}{||\Delta_-||_{2,1}^2} \sum_{ij=1}^{d} ||\Delta_{-,i}||_F^2 ||\Delta_{+,j}||_F^2$$
$$\leq \tau \frac{4(1+C)^2\nu r}{d} ||X_\sharp||_{op}^2 \sup_{(\Delta_+,\Delta_-)\in S} \frac{2}{||\Delta_-||_{2,1}^2} \sum_{ij=1}^{d} ||\Delta_{-,i}||_F^2$$
$$\leq \tau \frac{4(1+C)^2\nu r}{d} ||X_\sharp||_{op}^2 \sup_{(\Delta_+,\Delta_-)\in S} \frac{2d||\Delta_-||_F^2}{||\Delta_-||_{2,1}^2}$$
$$\leq 16\tau C^2 \nu r ||X_\sharp||_{op}^2.$$

Therefore, due to Theorem 5.10.9 there exists a constant $c_1, c_2, c_3 > 0$ so that with

$t = c_2 \log d$, we have that with probability $1 - e^{-c_2 \log d}$, the bound

$$
\begin{aligned}
Z &\leq \mathbb{E}Z + \sqrt{8\left(2b\mathbb{E}Z + \sigma^2\right)t} + 8bt \\
&\leq c_1 C \sqrt{\frac{\nu r}{d}} \|X_\sharp\|_{op} (\sqrt{\tau d \log d} + \log d) \\
&\quad + \sqrt{8c_2 \left(\frac{c_1^2 C^2 \nu r}{d} \|X_\sharp\|_{op}^2 (\sqrt{\tau d \log d} + \log d) + 16\tau C^2 \nu r \|X_\sharp\|_{op}^2 \right) \log d} \\
&\quad + 16 c_2 C \sqrt{\frac{\nu r}{d}} \|X_\sharp\|_{op} \log(d) \\
&\leq C \sqrt{\nu r \log d} \|X_\sharp\|_{op} \left(c_1 \sqrt{\tau} + c_1 \sqrt{\frac{\log d}{d}} + \sqrt{8c_2} \sqrt{c_1^2 \sqrt{\frac{\tau \log d}{d}} + c_1^2 \frac{\log d}{d} + 16\tau} \right. \\
&\qquad \left. + 16 c_2 \sqrt{\frac{\log d}{d}} \right) \\
&\leq c_3 C \sqrt{\tau \nu r \log d} \|X_\sharp\|_{op}.
\end{aligned}
$$

where the last line follows since by assumption $\log d / d \lesssim \tau$. □

5.10.5 Proofs in Section 5.8

Proof of Lemma 5.8.1

The proof follows the same strategy as [70, Theorem 6.1]. Fix $x \in \widetilde{\mathcal{T}}_1$ and let $\zeta \in \partial \tilde{f}(x)$. Then for all y, we have, from Lemma 5.8.3, that

$$f(y) \geq \tilde{f}(x) + \langle \zeta, y - x \rangle - \frac{\rho}{2} \|x - y\|_2^2 - 3\varepsilon.$$

Therefore, the function $g(y) := f(y) - \langle \zeta, y - x \rangle + \frac{\rho}{2} \|x - y\|_2^2 + 3\varepsilon$ satisfies

$$g(x) - \inf g \leq f(x) - \tilde{f}(x) + 3\varepsilon \leq 4\varepsilon.$$

Now, for $\gamma > 0$ to be determined momentarily, define

$$\hat{x} = \arg\min \left\{ g(x) + \frac{\varepsilon}{\gamma^2} \|x - y\|_2^2 \right\}.$$

First order optimality conditions and the sum rule immediately imply that

$$\frac{2\varepsilon}{\gamma^2}(x-\hat{x}) \in \partial g(\hat{x}) = \partial f(\hat{x}) - \zeta + \rho(\hat{x}-x).$$

Thus, $\operatorname{dist}(\zeta, \partial f(\hat{x})) \leq \left(\frac{2\varepsilon}{\gamma^2} + \rho\right)\|x-\hat{x}\|_2$. Now we estimate $\|x-\hat{x}\|_2$. Indeed, from the definition of $\hat{x}$ we have

$$\frac{\varepsilon}{\gamma^2}\|\hat{x}-x\|^2 \leq g(x) - g(\hat{x}) \leq g(x) - \inf g \leq 4\varepsilon.$$

Consequently, we have $\|x-\hat{x}\| \leq 2\gamma$. Thus, setting $\gamma = \sqrt{2\varepsilon/\rho}$ and recalling that $\varepsilon \leq \mu^2/56\rho$ we find that

$$\operatorname{dist}(\hat{x}, \mathcal{X}^*) \leq \|x-\hat{x}\| + \operatorname{dist}(x, \mathcal{X}^*) \leq 2\sqrt{\frac{2\varepsilon}{\rho}} + \frac{\mu}{4\rho} \leq \frac{\mu}{\rho}.$$

Likewise, we have

$$\operatorname{dist}(\hat{x}, \mathcal{X}) \leq \|x-\hat{x}\| \leq 2\sqrt{\frac{2\varepsilon}{\rho}}.$$

Therefore, setting $L = \sup\left\{\|\zeta\|_2 : \zeta \in \partial f(x), \operatorname{dist}(x, \mathcal{X}^*) \leq \frac{\mu}{\rho}, \operatorname{dist}(x, \mathcal{X}) \leq 2\sqrt{\frac{\varepsilon}{\rho}}\right\}$, we find that

$$\|\zeta\| \leq L + \operatorname{dist}(\zeta, \partial f(\hat{x})) \leq L + \frac{4\varepsilon}{\gamma} + 2\rho\gamma = L + 2\sqrt{8\rho\varepsilon},$$

as desired.

Proof of Theorem 5.8.4

Let $i \geq 0$, suppose $x_i \in \widetilde{\mathcal{T}}_1$, and let $x^* \in \mathrm{proj}_{\mathcal{X}^*}(x_i)$. Notice that Lemma 5.8.2 implies $\tilde{f}(x_i) - \min_{\mathcal{X}} f > 0$. We successively compute

$$
\begin{aligned}
\|x_{i+1} - x^*\|^2 &= \left\|\mathrm{proj}_{\mathcal{X}}\left(x_i - \tfrac{\tilde{f}(x_i)-\min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right) - \mathrm{proj}_{\mathcal{X}}(x^*)\right\|^2 \\
&\leq \left\|(x_i - x^*) - \tfrac{\tilde{f}(x_i)-\min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right\|^2 \qquad (5.106)\\
&= \|x_i - x^*\|^2 + \frac{2(\tilde{f}(x_i) - \min_{\mathcal{X}} f)}{\|\zeta_i\|^2} \cdot \langle \zeta_i, x^* - x_i\rangle + \frac{(\tilde{f}(x_i) - \min_{\mathcal{X}} f)^2}{\|\zeta_i\|^2} \\
&\leq \|x_i - x^*\|^2 + \frac{2(\tilde{f}(x_i) - \min_{\mathcal{X}} f)}{\|\zeta_i\|^2}\left(\min_{\mathcal{X}} f - \tilde{f}(x_i) + \frac{\rho}{2}\|x_i - x^*\|^2 + 3\varepsilon\right) \\
&\qquad + \frac{(\tilde{f}(x_i) - \min_{\mathcal{X}} f)^2}{\|\zeta_i\|^2} \qquad (5.107)\\
&= \|x_i - x^*\|^2 + \frac{\tilde{f}(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2}\left(\rho\|x_i - x^*\|^2 - (\tilde{f}(x_i) - \min_{\mathcal{X}} f) + 6\varepsilon\right) \\
&\leq \|x_i - x^*\|^2 + \frac{\tilde{f}(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2}\left(\rho\|x_i - x^*\|^2 - \mu\|x_i - x^*\| + 7\varepsilon\right) \qquad (5.108)\\
&\leq \|x_i - x^*\|^2 + \frac{\rho(\tilde{f}(x_i) - \min_{\mathcal{X}} f)}{\|\zeta_i\|^2}\left(\|x_i - x^*\| - \frac{\mu}{2\rho}\right)\|x_i - x^*\| \qquad (5.109)\\
&\leq \|x_i - x^*\|^2 - \frac{\mu(\tilde{f}(x_i) - \min_{\mathcal{X}} f)}{4\|\zeta_i\|^2} \cdot \|x_i - x^*\| \qquad (5.110)\\
&\leq \|x_i - x^*\|^2 - \frac{\mu(\mu\|x_i - x^*\| - \varepsilon)}{4\|\zeta_i\|^2} \cdot \|x_i - x^*\| \qquad (5.111)\\
&\leq \left(1 - \frac{13\mu^2}{56\|\zeta_i\|^2}\right)\|x_i - x^*\|^2.
\end{aligned}
$$

Here, the estimate (5.106) follows from the fact that the projection $\mathrm{proj}_Q(\cdot)$ is nonexpansive, (5.107) uses Lemma 5.8.3, the estimate (5.109) follows from the assumption $\epsilon < \frac{\mu}{14}\|x_k - x^*\|$, the estimate (5.110) follows from the estimate $\|x_i - x^*\| \leq \frac{\mu}{4\rho}$, while (5.108) and (5.111) use Lemma 5.8.2. We therefore deduce

$$
\mathrm{dist}^2(x_{i+1}; \mathcal{X}^*) \leq \|x_{i+1} - x^*\|^2 \leq \left(1 - \frac{13\mu^2}{56L^2}\right)\mathrm{dist}^2(x_i, \mathcal{X}^*).
$$

Consequently either we have $\mathrm{dist}(x_{i+1}, \mathcal{X}^*) < \frac{14\varepsilon}{\mu}$ or $x_{i+1} \in \widetilde{\mathcal{T}}_1$. Therefore, by induction, the proof is complete.

Proof of Theorem 5.8.6

Let $i \geq 0$, suppose $x_i \in \mathcal{T}_\gamma$, and let $x^* \in \text{proj}_{\mathcal{X}^*}(x_i)$. Then

$$\begin{aligned}
\mu \text{dist}(x_{i+1}, \mathcal{X}^*) \leq f(x_{i+1}) - \inf_{\mathcal{X}} f &\leq f_x(x_{i+1}) - \inf_{\mathcal{X}} f + \frac{\rho}{2}\|x_{i+1} - x_i\|^2 \\
&\leq \tilde{f}_x(x_{i+1}) - \inf_{\mathcal{X}} f + \frac{\rho}{2}\|x_{i+1} - x_i\|^2 + \varepsilon \\
&\leq \tilde{f}_x(x^*) - \inf_{\mathcal{X}} f + \frac{\beta}{2}\|x_i - x^*\|^2 + \varepsilon \\
&\leq f_x(x^*) - \inf_{\mathcal{X}} f + \frac{\beta}{2}\|x_i - x^*\|^2 + 2\varepsilon \\
&\leq f(x^*) - \inf_{\mathcal{X}} f + \beta\|x_i - x^*\|^2 + 2\varepsilon \\
&= \beta \text{dist}^2(x_i, \mathcal{X}^*) + 2\varepsilon.
\end{aligned}$$

Rearranging yields the result.

5.10.6 Auxiliary lemmas

The following are auxiliary lemmas that we used throughout this section. We record them here for convenience.

Lemma 5.10.7 (Lemma 3.1 in [40]). *Let* $S_r := \left\{X \in \mathbf{R}^{d_1 \times d_2} \mid \text{rank}(X) \leq r, \|X\|_F = 1\right\}$. *There exists an* ϵ*-net* $\mathcal{N}$ *(with respect to* $\|\cdot\|_F$*) of* S_r *obeying*

$$|\mathcal{N}| \leq \left(\frac{9}{\epsilon}\right)^{(d_1+d_2+1)r}.$$

Proposition 5.10.8 (Corollary 1.4 in [218]). *Consider* $X_1, \ldots, X_d$ *real-valued random variables and let* $\sigma \in \mathbb{S}^{d-1}$ *be a unit vector. Let* $t, p > 0$ *such that*

$$\sup_{u \in \mathbf{R}} \mathbb{P}\left(|X_i - u| \leq t\right) \leq p \qquad \text{for all } i = 1, \ldots, d.$$

Then the following holds

$$\sup_{u \in \mathbf{R}} \mathbb{P}\left(\left|\sum_k \sigma_k X_k - u\right| \leq t\right) \leq Cp,$$

where $C > 0$ is a universal constant.

Theorem 5.10.9 (Talagrand's Functional Bernstein for non-identically distributed variables [132, Theorem 1.1(c)])**.** *Let $\mathcal{S}$ be a countable index set. Let $Z_1, \ldots, Z_n$ be independent vector-valued random variables of the form $Z_i = (Z_{i,s})_{s \in \mathcal{S}}$. Let $Z := \sup_{s \in \mathcal{S}} \sum_{i=1}^{n} Z_{i,s}$. Assume that for all $i \in [n]$ and $s \in \mathcal{S}$, $\mathbb{E} Z_{i,s} = 0$ and $\left|Z_{i,s}\right| \leq b$. Let*

$$\sigma^2 = \sup_{s \in \mathcal{S}} \sum_{i=1}^{n} \mathbb{E} Z_{i,s}^2.$$

Then for each $t > 0$, we have the tail bound

$$P\left(Z - \mathbb{E}Z \geq \sqrt{8\left(2b\mathbb{E}Z + \sigma^2\right)t} + 8bt\right) \leq e^{-t}.$$

6

BLIND DECONVOLUTION: A CASE STUDY

"Las escaleras se suben de frente, pues hacia atrás o de costado resultan particularmente incómodas."

— Julio Cortázar, *Instrucciones para subir una escalera*

6.1 Introduction

In this chapter, we revisit the nonsmooth penalty technique, introduced in Chapter 5, for the *blind deconvolution* problem. We complement the local convergence guarantees, derived in the previous chapter, in two ways: first, we develop a robust spectral initialization method, which leads to a globally converging method; and second, we show that the spurious critical points of the nonsmooth problem lie close to a codimension two subspace, thus suggesting there might be a large region with friendly geometry. Unlike the convergence rates, the developments in this chapter are highly tailored for the blind deconvolution problem.

Formally, we consider the task of robustly recovering a pair $(\bar{w}, \bar{x}) \in \mathbf{R}^{d_1} \times \mathbf{R}^{d_2}$ from m bilinear measurements:

$$y_i = \langle \ell_i, \bar{w} \rangle \langle r_i, \bar{x} \rangle + \xi_i, \tag{6.1}$$

where ξ is an arbitrary noise corruption with a fraction of nonzeros $p_{\text{fail}} := |\text{supp } \xi|/m$ that is at most one half, and $\ell_i \in \mathbf{R}^{d_1}$ and $r_i \in \mathbf{R}^{d_2}$ are known mea-

surement vectors.[1] Such bilinear systems and their complex analogues arise often in biological systems, control theory, coding theory, and image deblurring, among others. Most notably such problems appear when recovering a pair $(u, v) \in \mathbf{C}^m \times \mathbf{C}^m$ from the convolution measurements $y = (Lu) * (Rv) \in \mathbf{C}^m$. When passing to the Fourier domain this problem is equivalent to that of solving a complex bilinear system of equations; see the pioneering work [8] on blind deconvolution. All the arguments we present can be extended to the complex case. We focus on the real case for simplicity.

Here we analyze the following nonsmooth formulation of the problem:

$$\min_{\|w\|_2,\, \|x\|_2 \le \nu\sqrt{\Phi}} f(w, x) := \frac{1}{m} \sum_{i=1}^{m} |\langle \ell_i, w\rangle\langle r_i, x\rangle - y_i|, \tag{6.2}$$

where $\nu \ge 1$ is a user-specified constant and $\Phi = \|\bar{w}\bar{x}^\top\|_F$. In contrast to the previous chapter, where we tackled asymmetric problems by developing guarantees under local regularity (Section 5.4.1), here we take the complementary approach of considering a bounded constraint set.

Our contributions are three-fold:

1. (**Local refinement)** Suppose that the vectors ℓ_i and r_i are both i.i.d. Sub-Gaussian and satisfy a mild growth condition (automatic for Gaussian random vectors). We show, levaraging results from the previous chapter, that as long as the number of measurements satisfies $m \gtrsim \frac{d_1+d_2}{(1-2p_{\text{fail}})^2} \ln(C + \frac{1}{1-2p_{\text{fail}}})$, where C is a small dimension-independent constant, the formulation (6.2) admits dimension independent constants ρ, L_f, and μ with high probability. Consequently, subgradient and prox-linear methods rapidly converge

[1] To avoid name clashes and be consistent with the literature, we relabel some objects from the previous section: $p_i \to \ell_i$, $q_i \to r_i$, $L \to \mathsf{L}_f$, and $b \to y$.

to the optimal solution at a dimension-independent rate when initialized at a point x_0 with constant relative error $\frac{\|w_0 x_0^\top - \bar{w}\bar{x}^\top\|_F}{\|\bar{w}\bar{x}^T\|F} \lesssim 1$.

2. (**Initialization)** Suppose now that ℓ_i and r_i are both i.i.d. Gaussian and are independent from the noise η. We develop an initialization procedure that in the regime $m \gtrsim d_1 + d_2$ and $p_{\text{fail}} \in [0, 1/10]$, will find a point x_0 satisfying $\frac{\|w_0 x_0^\top - \bar{w}\bar{x}^\top\|_F}{\|\bar{w}\bar{x}^T\|F} \lesssim 1$, with high probability. The proposed procedure is motivated by the analogous initialization algorithm for robust phase retrieval [89, 239]. To the best of our knowledge, this is the only available initialization procedure for rank-1 bilinear sensing with provable guarantees in presence of gross outliers. We also develop complementary guarantees under the weaker assumption that the vectors (ℓ_i, r_i) corresponding to exact measurements are independent from the noise η_i in the outlying measurements. This noise model allows one to plant outlying measurements from a completely different pair of signals, and is therefore computationally more challenging.

3. (**Landscape)** Suppose now that ℓ_i and r_i are both i.i.d. Gaussian and there is no noise $\eta = 0$. We show that when $m \gtrsim d_1 + d_2$ the set of spurious critical points of f lies close to the subspace $V = (\bar{w}, 0)^\perp \times (0, \bar{x})^\perp$. In particular, for any spurious critical point (w, x) we prove that $\text{dist}((w, x), V) \leq \tilde{O}\left(\left(\frac{d_1+d_2}{m}\right)^{\frac{1}{8}} \|(w, x)\|\right)$.

The literature studying bilinear systems is rich. It is well-known [156, 56, 126] that the optimal sample complexity in the noiseless regime is $m \gtrsim d_1 + d_2$ if no further assumptions (e.g. sparsity) are imposed on the signals. Therefore, from a sample complexity viewpoint, our guarantees are optimal. Incidentally, to our best knowledge, all alternative approaches are either suboptimal by a

polylogarithmic factor in d_1, d_2 or require knowing the sign pattern of one of the underlying signals [9, 8].

Recent algorithmic advances for blind deconvolution can be classified into two main approaches: works based on convex relaxations and those employing gradient descent on a smooth nonconvex function. The influential convex techniques of [9, 8] "lift" the objective to a higher dimension, thereby necessitating the resolution of a high-dimensional semidefinite program. The more recent work of [5, 6] instead relaxes the feasible region in the natural parameter space, under the assumption that the coordinate signs of either $\bar{w}$ or $\bar{x}$ are known a priori. Finally, with the exception of [8], the aforementioned works do not provide guarantees in the noisy regime.

Nonconvex approaches for blind deconvolution typically apply gradient descent to a smooth formulation of the problem [153, 167, 120]. Since the condition number of the problem scales with dimension, as we mentioned previously, these works introduce a nuanced analysis that is specific to the gradient method. The authors of [153] propose applying gradient descent on a regularized objective function, and identify a "basin of attraction" around the solution. The paper [167] instead analyzes gradient descent on the unregularized objective. They use the leave-one-out technique and prove that the iterates remain within a region where the objective function satisfies restricted strong convexity and smoothness conditions. The sample complexities of the methods in [153, 167, 120, 167] are optimal up to polylog factors.

The popular formulation for the blind deconvolution problem [8] necessitates one of the sets of measurement vectors r_i or ℓ_i to be deterministic. Indeed, they are built from the columns of a discrete Fourier transform matrix. Conse-

quently, our assumptions that both r_i and ℓ_i are random is an oversimplification. Similar assumptions are made in [5, 6, 7] for example. Nonetheless, extensive experiments in Section 6.5.4 show that even in this semi-stochastic setting the proposed algorithms work remarkably well. In particular, for difficult instances with a large incoherence parameter, we observe that the proposed algorithms perform on par and often better than gradient descent on the smooth formulations of the problem.

The nonconvex strategies mentioned above all use spectral methods for initialization. These methods are not robust to outliers, since they rely on the leading singular vectors/values of a potentially noisy measurement operator. Adapting the spectral initialization of [89] to bilinear inverse problems enables us to deal with gross outliers of arbitrary magnitude. Indeed, high variance noise makes it easier for our initialization to "reject" outlying measurements.

A related line of work [247, 134, 139] considers the original blind deconvolution problem of recovering a pair (u, v) from their convolution $u * v$ when u is low-dimensional and v is a sparse vector. These works are based on a very different approach to modeling the problem than the one we consider here. It would be interesting to see if similar ideas can be extended to this setting.

Outline of the chapter. Section 5.3 establishes estimates of weak convexity, sharpness, and Lipschitz moduli for the rank-1 bilinear sensing problem under statistical assumptions on the data. Section 6.3 introduces the initialization procedure and proves its correctness even if a constant fraction of measurements is corrupted by gross outliers. Section 6.4 studies the nonsmooth landscape of the blind deconvolution problem. The final Section 6.5 presents numerical experiments illustrating the theoretical results in this chapter.

6.2 Data generating model and local convergence guarantees

In this section, we formally specify the problem setting and state the corresponding local convergence guarantees proved in Chapter 5. The guarantees that we present here differ from the one established in Corollary 5.4.12 in that they only apply to the rank-1 case and they have slightly tighter estimate of the sharpness constant. The gains here are not significant, however we present the results since: they were derived, and uploaded to arXiv, before the guarantees in the previous chapter; and the sharpness estimate might be of independent interest.

The setting in this chapter is anologous to the previous one. Let us quickly remind the reader about it. We fix two disjoint sets $\mathcal{I}_{\text{in}} \subseteq [m]$ and $\mathcal{I}_{\text{out}} \subseteq [m]$, called the *inlier* and *outlier* sets. Intuitively, the index set $\mathcal{I}_{\text{in}}$ encodes exact measurements while $\mathcal{I}_{\text{out}}$ encodes measurements that have been replaced by gross outliers. Define the corruption frequency $p_{\text{fail}} := \frac{|\mathcal{I}_{\text{out}}|}{m}$; henceforth, we will suppose $p_{\text{fail}} \in [0, 1/2)$. Then for an arbitrary, potentially random sequence $\{\xi_i\}_{i=1}^m$, we consider the measurement model:

$$b_i := \begin{cases} \langle \ell_i, \bar{w} \rangle \langle r_i, \bar{x} \rangle & \text{if } i \in \mathcal{I}_{\text{in}}, \\ \xi_i & \text{if } i \in \mathcal{I}_{\text{out}}. \end{cases} \tag{6.3}$$

We define the linear map $\mathcal{A}\colon \mathbf{R}^{d_1 \times d_2} \to \mathbf{R}^m$ by $\mathcal{A}(X) = (\ell_i^\top X r_i)_{i=1}^m$. To simplify notation, we let $P \in \mathbf{R}^{m \times d_1}$ denote the matrix whose rows, in column form, are ℓ_i and we let $Q \in \mathbf{R}^{m \times d_2}$ denote the matrix whose rows are r_i. Note that we make no assumptions about the nature of ξ_i. In particular, ξ_i can even encode exact measurements for a different signal.

In acordance to Section 5.5.2, we focus on the following fully stochastic ma-

trix model. We should note that in more realistic circumstancees, such as the problem of blind decovonvolution, it is more appropriate for one of the matrices P or Q to be deterministic. Though the theoretical guarantees we present only hold in the fully stochastic setting, numerical experiments in Section 6.5.4 indicate that the proposed methods are effective even when one of the matrices is deterministic.

Random matrix model.

M The vectors ℓ_i and r_i are i.i.d. realizations of η-sub-gaussian random vectors $\ell \in \mathbf{R}^{d_1}$ and $r \in \mathbf{R}^{d_2}$, respectively. Suppose moreover that ℓ and r are independent and satisfy the nondegeneracy condition,

$$\inf_{\substack{X:\ \operatorname{rank} X \le 2 \\ \|X\|_F = 1}} \mathbb{E}|\ell^\top X r| \ge \alpha, \tag{6.4}$$

for some real $\alpha > 0$.

Thus the model **M** asserts that ℓ_i and r_i are generated by independent subgaussian random vectors. The nondegeneracy condition (6.4) essentially asserts that with positive probability, the products $\ell^\top X r$ are non-negligible, uniformly over all unit norm rank two matrices X. In particular, Gaussian matrices with i.i.d. entries are admissible under Model **M**, see Lemma 5.5.7.

6.2.1 Guarantees

Just as before, we will establish guarantees for the subgradient and the prox-linear method introduced in Section 5.2. To this end, consider two vectors $\bar{w} \in$

$\mathbf{R}_1^d$ and $\bar{x} \in \mathbf{R}^{d_2}$, and set $\Phi := \|\bar{x}\bar{w}^\top\|_F = \|\bar{x}\|_2 \cdot \|\bar{w}^\top\|_2$. Without loss of generality, henceforth, we suppose $\|\bar{w}\|_2 = \|\bar{x}\|_2$. Define the two sets:

$$\mathcal{S}_\nu := \nu\sqrt{\Phi} \cdot (\mathbb{B}^{d_1} \times \mathbb{B}^{d_2}), \qquad \mathcal{S}_\nu^* := \{(\alpha\bar{w}, (1/\alpha)\bar{x}) \colon 1/\nu \le |\alpha| \le \nu\}.$$

The set $\mathcal{S}_\nu$ simply encodes a bounded region, while $\mathcal{S}_\nu^*$ encodes all rank-1 factorizations of the matrix $\bar{w}\bar{x}^\top$ with bounded factors.

The objective function can be decomposed as $f = h \circ F$ with $h = \|\cdot\|_1$ and $F = \mathcal{A}$. Thus, all the machinery for rapid convergence, developed in the previous chapter, applies provided we show that f satisfies the Approximation accuracy, Sharpness and Subgradient bound conditions on $\mathcal{S}_\nu$; we refer the reader to Assumption 5.2 for definitions.

Theorem 5.5.8 establishes that there exists constants κ_1, κ_2, and κ_3 such that with high probability, the operator $\mathcal{A}$ satisfies RIP and the $\mathcal{I}$-outlier bound, i.e.,

$$\kappa_1\|M\|_F \le \frac{1}{m}\|\mathcal{A}(M)\|_1 \le \kappa_2\|M\|_F \qquad \text{for all rank-2 } M \in \mathbf{R}^{d_1 \times d_2}$$

and

$$\kappa_3\|M\|_F \le \frac{1}{m}\left(\|\mathcal{A}_{\mathcal{I}_{\text{in}}}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}_{\text{out}}}(M)\|_1\right) \qquad \text{for all rank-2 } M \in \mathbf{R}^{d_1 \times d_2},$$

respectively. In turn, RIP ensures that f satisfies both the approximation accuracy and subgradient bound conditions, see Proposition 5.3.3. We also proved that the last inequality implies sharpness, albeit only on a small neighborhood around the solution, see Proposition 5.3.8. We now show that, at least for the rank-one case, sharpness hold uniformly over $\mathcal{S}_\nu$.

To establish sharpness of f, we first show that the function $(x, w) \mapsto \|wx^\top - \bar{w}\bar{x}^\top\|_F$ is sharp on the set $\mathcal{S}_\nu$. This is a specialization of Theorem 5.3.5 with a slightly better constant. The proof is quite tedious, and therefore we have placed it in Section 6.6.1.

Theorem 6.2.1 (Sharpness rank-one (asymmetric))**.** *For any $\nu \geq 1$, we have the following bound*

$$\|wx^\top - \bar{w}\bar{x}^\top\|_F \geq \frac{\sqrt{\Phi}}{2\sqrt{2}(\nu+1)}\text{dist}((w,x),\mathcal{S}_\nu^*) \qquad \textit{for all } (w,x) \in \mathcal{S}_\nu.$$

Armed with this Theorem we now establish the sharpness guarantee for f.

Proposition 6.2.2 (Sharpness in the noisy regime)**.** *Suppose that Assumption 5.3.6 holds. Then*

$$f(w,x) - f(\bar{w},\bar{x}) \geq \frac{c_3\sqrt{\Phi}}{2\sqrt{2}(\nu+1)}\text{dist}((w,x),\mathcal{S}_\nu^*) \qquad \textit{for all } (w,x) \in \mathcal{S}_\nu.$$

Proof. Defining $\xi = \mathcal{A}(\bar{w}\bar{x}^T) - b$, we have the following bound:

$$\begin{aligned}
&f(w,x) - f(\bar{w},\bar{x}) \\
&\quad= \frac{1}{m}\left(\|\mathcal{A}\left(wx^\top - \bar{w}\bar{x}^\top\right) + \xi\|_1 - \|\xi\|_1\right) \\
&\quad= \frac{1}{m}\left(\|\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\|_1 + \sum_{i\in\mathcal{I}}\left(\left|\left(\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\right)_i + \xi_i\right| - \left|\left(\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\right)_i\right| - |\xi_i|\right)\right) \\
&\quad\geq \frac{1}{m}\left(\|\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\|_1 - 2\sum_{i\in\mathcal{I}}\left|\left(\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\right)_i\right|\right) \\
&\quad= \frac{1}{m}\sum_{i\in\mathcal{I}^c}\left|\left(\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\right)_i\right| - \frac{1}{m}\sum_{i\in\mathcal{I}}\left|\left(\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\right)_i\right| \\
&\quad\geq c_3\|wx^\top - \bar{w}\bar{x}^\top\|_F \geq \frac{c_3\sqrt{\Phi}}{2\sqrt{2}(\nu+1)}\text{dist}((w,x),\mathcal{S}_\nu^*),
\end{aligned}$$

where the first inequality follows by the reverse triangle inequality, the second inequality follows by Assumption (C2), and the final inequality follows from Theorem 6.2.1. The proof is complete. □

Combining this result with Theorem 5.5.8 and the convergence rates in Section 5.4 we obtain the following guarantee.

Corollary 6.2.3 (Convergence guarantees)**.** *Consider the measurement model* (6.3) *and suppose that model* $\boldsymbol{M}$ *is valid. Consider the optimization problem*

$$\min_{(x,w)\in\mathcal{S}_\nu} f(w,x) = \frac{1}{m}\sum_{i=1}^{m} |\langle \ell_i, w\rangle\langle r_i, x\rangle - b_i|.$$

Then there exist constants $c_1, c_2, c_3, c_4, c_5, c_6 > 0$ *depending only on* α, η *such that as long as* $m \geq \frac{c_1(d_1+d_2+1)}{(1-2p_{\text{fail}})^2}\ln\left(c_2 + \frac{c_2}{1-2p_{\text{fail}}}\right)$ *and you choose any pair* (w_0, x_0) *with relative error*

$$\frac{\operatorname{dist}((w_0,x_0),\mathcal{S}_\nu^*)}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \leq \frac{c_6\,(1-2p_{\text{fail}})}{4\sqrt{2}c_5(\nu+1)}, \tag{6.5}$$

then with probability at least $1 - 4\exp\left(-c_3(1-2p_{\text{fail}})^2 m\right)$ *the following are true.*

1. **(Polyak subgradient)** *Algorithm 2 initialized* (w_0, x_0) *produces iterates that converge linearly to* $\mathcal{S}_\nu^*$*, that is*

$$\frac{\operatorname{dist}^2((w_k,x_k),\mathcal{S}_\nu^*)}{\|\bar{w}\bar{x}^\top\|_F} \leq \left(1 - \frac{c_6^2\,(1-2p_{\text{fail}})^2}{32c_5^2(\nu+1)^4}\right)^k \cdot \frac{c_6^2\,(1-2p_{\text{fail}})^2}{32c_5^2(\nu+1)^2} \qquad \forall k \geq 0.$$

2. **(geometric subgradient)** *Set* $\lambda := \frac{c_6^2(1-2p_{\text{fail}})^2\sqrt{\|\bar{w}\bar{x}^\top\|_F}}{16\sqrt{2}c_5^2\nu(\nu+1)^2}$ *and* $q := \sqrt{1 - \frac{c_6^2(1-2p_{\text{fail}})^2}{32c_5^2(\nu+1)^4}}$*. Then the iterates* x_k *generated by Algorithm 3, initialized at* (w_0, x_0) *converge linearly:*

$$\frac{\operatorname{dist}^2((w_k,x_k),\mathcal{S}_\nu^*)}{\|\bar{w}\bar{x}^\top\|_F} \leq \left(1 - \frac{c_6^2\,(1-2p_{\text{fail}})^2}{32c_5^2(\nu+1)^4}\right)^k \cdot \frac{c_6^2\,(1-2p_{\text{fail}})^2}{32c_5^2(\nu+1)^2} \qquad \forall k \geq 0.$$

3. **(prox-linear)** *Algorithm 4 with* $\beta = \rho$ *and initialized at* (w_0, x_0) *converges quadratically:*

$$\frac{\operatorname{dist}((w_k,x_k),\mathcal{X}^*)}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \leq 2^{-2^k} \cdot \frac{c_6\,(1-2p_{\text{fail}})}{2\sqrt{2}c_5(\nu+1)} \qquad \forall k \geq 0.$$

Thus with high probability, if one initializes the subgradient and prox-linear methods at a pair (w_0, x_0) satisfying $\frac{\operatorname{dist}((w_0,x_0),\mathcal{S}_\nu^*)}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \leq \frac{c_6(1-2p_{\text{fail}})}{4\sqrt{2}c_5(\nu+1)}$, then the methods will converge to the optimal solution set at a dimension independent rate.

6.3 Initialization

Previous sections have focused on local convergence guarantees under various statistical assumptions. In particular, under Assumptions 5.3.1 and 5.3.6, one must initialize the local search procedures at a point (w, x), whose relative distance to the solution set $\frac{\operatorname{dist}((x,w),\mathcal{S}_\nu^*)}{\sqrt{\|\bar{x}\bar{w}^\top\|_F}}$ is upper bounded by a constant. In this section, we present a new spectral initialization routine (Algorithm 6) that is able to efficiently find such point (w, x). The algorithm is inspired by [89, Section 4] and [239].

Before describing the intuition behind the procedure, let us formally introduce our assumptions. Throughout this section, we make the following assumption on the data generating mechanism, which is stronger than Model **M**:

$\overline{\mathbf{M}}$ The entries of matrices L and R are i.i.d. Gaussian.

Our arguments rely heavily on properties of the Gaussian distribution. We note, however, that our experimental results suggest that Algorithm 6 provides high-quality initializations under weaker distributional assumptions.

Recall that in the previous sections, the noise ξ was arbitrary. In this section, however, we must assume more about the nature of the noise. We will consider two different settings.

N1 The measurement vectors $\{(\ell_i, r_i)\}_{i=1}^m$ and the noise sequence $\{\xi_i\}_{i=1}^m$ are independent.

N2 The inlying measurement vectors $\{(\ell_i, r_i)\}_{i\in\mathcal{I}_{\text{in}}}$ and the corrupted observations $\{\xi_i\}_{i\in\mathcal{I}_{\text{out}}}$ are independent.

The noise models **N1** and **N2** differ in how an adversary may choose to corrupt the measurements. Model **N1** allows an adversary to corrupt the signal, but does not allow observation of the measurement vectors $\{(\ell_i, r_i)\}_{i=1}^m$. On the other hand, Model **N2** allows an adversary to observe the outlying measurement vectors $\{(\ell_i, r_i)\}_{i\in\mathcal{I}_{\text{out}}}$ and arbitrarily corrupt those measurements. For example, the adversary may replace the outlying measurements with those taken from a completely different signal: $y_i = (\mathcal{A}(\widetilde{w}\widetilde{x}^\top))_i$ for $i \in \mathcal{I}_{\text{out}}$.

Algorithm 6: Initialization.

Data: $y \in \mathbf{R}^m, L \in \mathbf{R}^{m\times d_1}, R \in \mathbf{R}^{m\times d_2}$

$\mathcal{I}^{\text{sel}} \leftarrow \{i \mid |b_i| \le \texttt{med}(|y|)\}$

Form directional estimates:

$$L^{\text{init}} \leftarrow \tfrac{1}{m}\textstyle\sum_{i\in\mathcal{I}^{\text{sel}}} \ell_i\ell_i^\top, \quad R^{\text{init}} \leftarrow \tfrac{1}{m}\textstyle\sum_{i\in\mathcal{I}^{\text{sel}}} r_i r_i^\top$$

$$\widehat{w} \leftarrow \arg\min_{p\in\mathbb{S}^{d_1-1}} p^\top L^{\text{init}} p, \quad \text{and} \quad \widehat{x} \leftarrow \arg\min_{q\in\mathbb{S}^{d_2-1}} q^\top R^{\text{init}} q.$$

Estimate the norm of the signal:

$$\widehat{\Phi} \leftarrow \arg\min_{\beta\in\mathbf{R}} G(\beta) := \frac{1}{m}\sum_{i=1}^m \left|b_i - \beta\langle\ell_i, \widehat{w}\rangle\langle r_i, \widehat{x}\rangle\right|,$$

$$w_0 \leftarrow \operatorname{sign}(\widehat{\Phi})\left|\widehat{\Phi}\right|^{1/2}\widehat{w}, \quad \text{and} \quad x_0 \leftarrow \left|\widehat{\Phi}\right|^{1/2}\widehat{x}.$$

return (w_0, x_0)

We can now describe the intuition underlying Algorithm 6. Throughout we denote unit vectors parallel to $\bar{w}$ and $\bar{x}$ by $\bar{w}_\star$ and $\bar{x}_\star$, respectively. Algorithm 6 exploits the expected near orthogonality of the random vectors ℓ_i and r_i to the directions $\bar{w}_\star$ and $\bar{x}_\star$, respectively, in order to select a "good" set of measurement vectors. Namely, since $\mathbb{E}[\langle\ell_i, \bar{w}_\star\rangle] = \mathbb{E}[\langle r_i, \bar{x}_\star\rangle] = 0$, we expect minimal eigenvectors of L^{init} and R^{init} to be near $\bar{w}_\star$ and $\bar{x}_\star$, respectively. Since our measurements are bilinear, we cannot necessarily select vectors for which $|\langle\ell_i, \bar{w}_\star\rangle|$ and $|\langle r_i, \bar{x}_\star\rangle|$ are both small, rather, we may only select vectors for which the

product $|\langle \ell_i, \bar{w}_\star \rangle \langle r_i, \bar{x}_\star \rangle|$ is small, leading to subtle ambiguities not present in [89, Section 4] and [239]; see Figure 6.1. Corruptions add further ambiguities since the noise model **N2** allows a constant fraction of measurements to be adversarially modified.

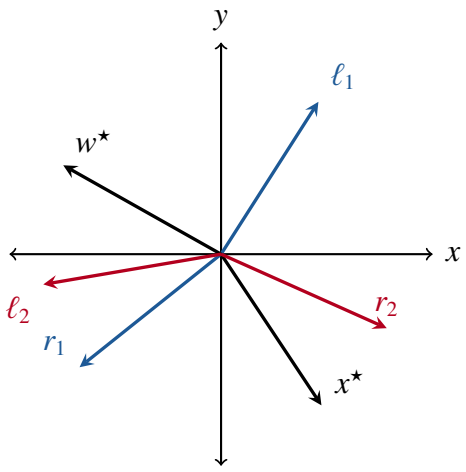


Figure 6.1: Intuition behind spectral initialization. The pair ℓ_1, r_1 will be included since both vectors are almost orthogonal to the true directions. ℓ_2, r_2 is unlikely to be included since r_2 is almost aligned with $x^\star$.

Formally, Algorithm 6 estimates an initial signal (w_0, x_0) in two stages: first it constructs a pair of directions $(\hat{w}, \hat{x})$ which estimate the true directions

$$\bar{w}_\star := \frac{1}{\|\bar{w}\|_2} \bar{w} \qquad \text{and} \qquad \bar{x}_\star := \frac{1}{\|\bar{x}\|_2} \bar{x}$$

(up to sign); then it constructs an estimate $\widehat{\Phi}$ of the signed signal norm $\pm\Phi$, which corrects for sign errors in the first stage. We now discuss both stages in more detail, starting with the direction estimate. Most proofs will be deferred to Section 6.6.2. The general proof strategy we follow is analogous to [89, Section 4] for phase retrieval, with some subtle modifications due to asymmetry.

Step 1: Direction Estimate. In the first stage of the algorithm, we estimate the directions $\bar{w}_\star$ and $\bar{x}_\star$, up to sign. Key to our argument is the following decom-

position for model **N1** (which will be proved in Section 6.6.2):

$$L^{\text{init}} = \tfrac{|\mathcal{I}^{\text{sel}}|}{m} \cdot I_{d_1} - \gamma_1 \bar{w}_\star \bar{w}_\star^\top + \Delta_L, \quad R^{\text{init}} = \tfrac{|\mathcal{I}^{\text{sel}}|}{m} \cdot I_{d_2} - \gamma_2 \bar{x}_\star \bar{x}_\star^\top + \Delta_R,$$

where $\gamma_1, \gamma_2 \gtrsim 1$ and the matrices Δ_L, Δ_R have small operator norm (decreasing with $(d_1 + d_2)/m$), with high probability. Using the Davis-Kahan $\sin\theta$ theorem [65], we can then show that the minimal eigenvectors of L^{init} and R^{init} are sufficiently close to $\{\pm\bar{w}_\star\}$ and $\{\pm\bar{x}_\star\}$, respectively.

Proposition 6.3.1 (Directional estimates)**.** *There exist numerical constants $c_1, c_2, C > 0$, so that for any $p_{\text{fail}} \in [0, 1/10]$ and $t \in [0, 1]$, with probability at least $1 - c_1 \exp(-c_2 mt)$, the following hold:*

$$\min_{s\in\{\pm 1\}} \left\| \widehat{w}\widehat{x}^\top - s w^\star x^{\star\top} \right\|_F \leq \begin{cases} C \cdot \left(\sqrt{\frac{\max\{d_1,d_2\}}{m} + t} \right) & \text{under Model } \textbf{N1}, \text{ and} \\ C \cdot \left(p_{\text{fail}} + \sqrt{\frac{\max\{d_1,d_2\}}{m} + t} \right) & \text{under Model } \textbf{N2}. \end{cases}$$

Step 2: Norm estimate. In the second stage of the algorithm, we estimate Φ as well as correct the sign of the direction estimates from the previous stage. In particular, for any $(\widehat{w}, \widehat{x}) \in \mathbb{S}^{d_1-1} \times \mathbb{S}^{d_2-1}$ define the quantity

$$\delta := \left(1 + \frac{c_5}{c_6(1 - 2p_{\text{fail}})} \right) \min_{s\in\{\pm 1\}} \left\| \widehat{w}\widehat{x}^\top - s\bar{w}_\star \bar{x}_\star^\top \right\|_F, \tag{6.6}$$

where c_5 and c_6 are as in Theorem 5.5.8. Then we prove the following estimate (see Section 6.6.2).

Proposition 6.3.2 (Norm Estimate)**.** *Under either noise model, **N1** and **N2**, there exist numerical constants $c_1, \ldots, c_6 > 0$ so that if $m \geq \frac{c_1(d_1+d_2+1)}{(1-2p_{\text{fail}})^2} \ln\left(c_2 + \frac{c_2}{1-2p_{\text{fail}}} \right)$, then with probability at least $1 - 4\exp\left(-c_3(1 - 2p_{\text{fail}})^2 m\right)$, we have that any minimizer $\widehat{\Phi}$ of the function*

$$G(\beta) := \frac{1}{m} \sum_{i=1}^{m} |y_i - \beta \langle \ell_i, \widehat{w} \rangle \langle \widehat{x}, r_i \rangle|$$

satisfies $|\widehat{\Phi}| - \Phi| \leq \delta\Phi$. *Moreover, if in this event* $\delta < 1$, *then we have* $\operatorname{sign}(\widehat{\Phi}) = \arg\min_{s\in\{\pm 1\}} \left\|\widehat{w}\widehat{x}^\top - s\bar{w}_\star \bar{x}_\star^\top\right\|_F$.

Thus, the preceding proposition shows that tighter estimates on the norm Φ result from better directional estimates in the first stage of Algorithm 6. In light of Proposition 6.3.2, we next estimate the probability of the event $\delta \leq 1/2$, which in particular implies with high probability $\operatorname{sign}(\widehat{\Phi}) = \arg\min_{s\in\{\pm 1\}} \left\|\widehat{w}\widehat{x}^\top - s\bar{w}_\star \bar{x}_\star^\top\right\|_F$.

Proposition 6.3.3 (Sign estimate)**.** *Under either Model* ***N1*** *and* ***N2****, there exist numerical constants* $c_0, c_1, c_2, c_3 > 0$ *such that if* $p_{\text{fail}} < c_0$ *and* $m \geq c_3(d_1 + d_2)$*, then the estimate holds:*[2]

$$\mathbb{P}\left(\delta > 1/2\right) \leq c_1 \exp\left(-c_2 m\right).$$

Proof. Using Theorem 5.5.8 and Propositions 6.3.1, we deduce that for any $t \in [0, 1]$, with probability $1 - c_1 \exp(-c_2 mt)$ we have

$$\delta \leq \begin{cases} C \cdot \left(\sqrt{\frac{\max\{d_1,d_2\}}{m}} + t\right) & \text{under Model } \mathbf{N1}, \text{ and} \\ C \cdot \left(p_{\text{fail}} + \sqrt{\frac{\max\{d_1,d_2\}}{m}} + t\right) & \text{under Model } \mathbf{N2}. \end{cases}$$

Thus under model **N1** it suffices to set $t = (2C)^{-2} - \frac{\max\{d_1,d_2\}}{m}$. Then the probability of the event $\delta \leq 1/2$ is at least $1 - c_1 \exp\left(-c_2((2C)^{-2}m - \max\{d_1, d_2\})\right)$. On the other hand, under model **N2**, it suffices to assume $2Cp_{\text{fail}} < 1$ and then we can set $t = (((2C)^{-1} - p_{\text{fail}})^2 - \frac{\max\{d_1,d_2\}}{m})$. The probability of the event $\delta \leq 1/2$ is then at least $1 - c_1(\exp\left(-c_2(m((2C)^{-1} - p_{\text{fail}})^2 - \max\{d_1, d_2\})\right)$. Finally using the bound $\max\{d_1, d_2\} \leq d_1 + d_2 \leq \frac{m}{c_3}$ yields the result. □

Step 3: Final estimate. Putting the directional and norm estimates together, we arrive at the following theorem.

[2] In the case of model **N1**, one can set $c_0 = 1/10$.

Theorem 6.3.4. *There exist numerical constants $c_0, c_1, c_2, c_3, C > 0$ such that if $p_{\text{fail}} \leq c_0$ and $m \geq c_4(d_1 + d_2)$, then for all $t \in [0, 1]$, with probability at least $1 - c_1 \exp(-c_3 mt)$, we have*

$$\frac{\left\|w_0 x_0^\top - \bar{w}\bar{x}^\top\right\|_F}{\|\bar{w}\bar{x}^\top\|_F} \leq \begin{cases} C \cdot \left(\sqrt{\frac{\max\{d_1,d_2\}}{m}} + t\right) & \text{under Model } \textbf{N1}, \text{ and} \\ C \cdot \left(p_{\text{fail}} + \sqrt{\frac{\max\{d_1,d_2\}}{m}} + t\right) & \text{under Model } \textbf{N2}. \end{cases}$$

Proof. Suppose that we are in the events guaranteed by Propositions 6.3.1,6.3.2, and 6.3.3. Then noting that

$$w_0 = \operatorname{sign}(\widehat{\Phi})|\widehat{\Phi}|^{1/2}\widehat{w}, \;\; x_0 = |\widehat{\Phi}|^{1/2}\widehat{x},$$

we find that

$$\begin{aligned} \left\|w_0 x_0^\top - \bar{w}\bar{x}^\top\right\|_F &= \left\|\operatorname{sign}(\widehat{\Phi})|\widehat{\Phi}|\widehat{w}\widehat{x}^\top - \Phi \bar{w}_\star \bar{x}_\star^\top\right\|_F \\ &= \Phi \left\|\widehat{w}\widehat{x}^\top - \operatorname{sign}(\widehat{\Phi})\bar{w}_\star \bar{x}_\star^\top + \frac{|\widehat{\Phi}| - \Phi}{\Phi}\widehat{w}\widehat{x}^\top\right\|_F \\ &\leq \Phi \left\|\widehat{w}\widehat{x}^\top - \operatorname{sign}(\widehat{\Phi})w^\star x^{\star\top}\right\|_F + \Phi\delta \\ &= \Phi \cdot \left(2 + \frac{c_5}{c_6(1 - 2p_{\text{fail}})}\right) \min_{s\in\{\pm 1\}} \left\|\widehat{w}\widehat{x}^\top - s\bar{w}_\star \bar{x}_\star^\top\right\|_F, \end{aligned}$$

where c_5 and c_6 are defined in Theorem 5.5.8. Appealing to Proposition 6.3.1, the result follows. □

Combining Corollary 6.2.3 and Theorem 6.3.4, we arrive at the following guarantee for the stage procedure.

Corollary 6.3.5 (Efficiency estimates)**.** *Suppose either of the models **N1** and **N2**. Let (w_0, x_0) be the output of the initialization Algorithm 6. Set $\widehat{\Phi} = \|w_0 x_0^\top\|_F$ and consider the optimization problem*

$$\min_{\|x\|_2, \|w\|_2 \leq \sqrt{2\widehat{\Phi}}} g(w, x) = \frac{1}{m}\|\mathcal{A}(wx^\top) - y\|_1. \tag{6.7}$$

Set $\nu := \sqrt{\frac{2\widehat{\Phi}}{\Phi}}$ *and notice that the feasible region of* (6.7) *coincides with* $\mathcal{S}_\nu$*. Then there exist constants* $c_0, c_1, c_2, c_3, c_5 > 0$ *and* $c_4 \in (0,1)$ *such that as long as* $m \geq c_3(d_1 + d_2)$ *and* $p_{\text{fail}} \leq c_0$*, the following properties hold with probability* $1 - c_1 \exp(-c_2 m)$*.*[3]

1. **(subgradient)** *Both Algorithms 2 and 3 (with appropriate* λ, q*) initialized* (w_0, x_0) *produce iterates that converge linearly to* $\mathcal{S}_\nu^*$*, that is*

$$\frac{\text{dist}^2((w_k, x_k), \mathcal{S}_\nu^*)}{\|\bar{w}\bar{x}^\top\|_F} \leq c_4 (1 - c_4)^k \qquad \forall k \geq 0.$$

2. **(prox-linear)** *Algorithm 4 initialized at* (w_0, x_0) *(with appropriate* $\beta > 0$*) converges quadratically:*

$$\frac{\text{dist}((w_k, x_k), \mathcal{S}_\nu^*))}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \leq c_5 \cdot 2^{-2^k} \qquad \forall k \geq 0.$$

Proof. We provide the proof under model **N1**. The proof under model **N2** is completely analogous. Combining Proposition 6.3.2, Proposition 6.3.3, and Theorem 6.3.4, we deduce that there exist constants c_0, c_1, c_2, c_3, C such that as long as $m \geq c_3(d_1 + d_2)$ and $p_{\text{fail}} < c_0$, then for any $t \in [0,1]$, with probability $1 - c_1 \exp(-c_2 m t)$, we have

$$\left|\frac{\widehat{\Phi}}{\Phi} - 1\right| \leq \delta \leq \frac{1}{2}, \tag{6.8}$$

and

$$\frac{\left\|w_0 x_0^\top - \bar{w}\bar{x}^\top\right\|_F}{\Phi} \leq C\sqrt{\frac{\max\{d_1, d_2\}}{m} + t}.$$

In particular, notice from (6.8) that $1 \leq \nu \leq \sqrt{3}$ and therefore the feasible region $\mathcal{S}_\nu$ contains an optimal solution of the original problem (6.2). Using Theorem 6.2.1, we have

$$\|w_0 x_0^\top - \bar{w}\bar{x}^\top\|_F \geq \frac{\sqrt{\Phi}}{2\sqrt{2}(\nu + 1)} \text{dist}((w_0, x_0), \mathcal{S}_\nu^*).$$

[3] In the case of model **N1**, one can set $c_0 = 1/10$.

Combining the estimates, we conclude

$$\frac{\operatorname{dist}((w_0, x_0), \mathcal{S}_\nu^*)}{\sqrt{\Phi}} \leq 2\sqrt{2}(\nu+1)\frac{\|w_0 x_0^\top - \bar{w}\bar{x}^\top\|_F}{\Phi} \leq 2\sqrt{2}(\nu+1)C\sqrt{\frac{\max\{d_1, d_2\}}{m} + t}.$$

Thus to ensure the relative error assumption (6.5), it suffices to ensure the inequality

$$2\sqrt{2}(\nu+1)C\sqrt{\frac{\max\{d_1, d_2\}}{m} + t} \leq \frac{c_6(1-2p_{\text{fail}})}{4\sqrt{2}c_5(\nu+1)},$$

where c_5, c_6 are the constants from Corollary 6.2.3. Using the bound $\nu \leq \sqrt{3}$, it suffices to set

$$t = \left(\frac{c_6(1-2p)}{16\sqrt{3}c_5 C}\right)^2 - \frac{\max\{d_1, d_2\}}{m}.$$

Thus the probability of the desired event becomes $1 - c_2(\exp(-c_3(c_4 m - \max\{d_1, d_2\}))$ for some constant c_4. Finally, using the bound $\max\{d_1, d_2\} \leq d_1 + d_2 \leq \frac{m}{c_3}$ and applying Corollary 6.2.3 completes the proof. □

6.4 Nonsmooth landscape

As we alluded in the previous section, initialization procedures are nontrivial to develop and are often computationally more expensive than the refinement stage. Thus, it is natural to wonder if the initialization is necessary. For the blind deconvolution problem, numerical experiments suggest that a *randomly initilized* subgradient method applied to the unconstraint version of the nonsmooth formulation 6.2,

$$\arg\min_{w,x} f_S(w, x) = \frac{1}{m}\sum_{i=1}^{m} |\langle \ell_i, w\rangle\langle x, r_i\rangle - y_i|, \tag{6.9}$$

recovers the signal exactly, see Figure 6.2. However the guarantees developed thus far do not explain this behavior.

There are several examples of smooth nonconvex problems where simple iterative methods with random initilization provably converge to minimizers [106, 103, 26, 57, 142]. Analysis of these methods are of two types: those based on studying the iterate sequence [135, 78, 248], and those based on characterizing the landscape of smooth loss functions [103, 227, 161].

In this section, we push theory towards understanding the global success of numerical methods applied to (6.9). To this end, we study the nonsmooth nonconvex landscape of this problem. Unlike the aforementioned works, the loss function we study is nonsmooth loss, which presents fundamentally different technical challenges.

We study the landscape of f_S under Model $\overline{\mathbf{M}}$, i.e., L and R are standard Gaussian random matrices. Following the line of ideas in [70], we think of f_S as the empirical average approximation of the *population objective*

$$f_P(w, x) := \mathbb{E} f_S(w, x) = \mathbb{E}\left(|\ell^\top (wx - \bar{w}\bar{x}) r^\top|\right),$$

where $\ell \in \mathbf{R}^{d_1}$ and $r \in \mathbf{R}^{d_2}$ are standard Gaussian vectors. From now on, we will refer to f_S as the *sample objective*. The rationale is simple: we will describe the stationary points of f_P, then we will prove that the graph of the subdifferential ∂f_S concentrates around the graph of ∂f_P and combine these to describe the landscape of f_S. This strategy allows us to show that the set of spurious stationary points converges to a codimension two subspace at a controlled rate. We remark that these results are geometrical and not computational.

Denote the set of solutions of (6.9) by

$$\mathcal{S} := \{(\alpha\bar{w}, \bar{x}/\alpha) \mid \alpha \in \mathbf{R} \setminus \{0\}\}.$$

We now highlight the main contributions of this section.

Population objective. Interestingly, the population objective only depends on (w, x) through the singular values of the rank two matrix $X := wx^\top - \bar{w}\bar{x}^\top$. We show this function can be written as

$$f_P(w, x) = \sigma_{\max}(X) \sum_{n=0}^{\infty} \left(\frac{(2n)!}{2^{2n}(n!)^2} \right)^2 \frac{\left(1 - \kappa^{-2}(X)\right)^n}{1 - 2n}$$

where $\kappa(X) = \sigma_{\max}(X)/\sigma_{\min}(X)$ is the condition number of X. We characterize the stationary points of a broad family of spectral functions, containing f_P. By specializing this characterization we find that the stationary points of the population objectives are exactly

$$\mathcal{S} \cup \{(w, x) \mid \langle w, \bar{w} \rangle = 0, \langle x, \bar{x} \rangle = 0, \text{and } wx^\top = 0\},$$

revealing that the set of extraneous critical points of f_P is the subspace $(\bar{w}, 0)^\perp \cap (0, \bar{x})^\perp$.

Sample objective. Equipped with a quantitative version of Attouch-Wets' convergence theorem proved in [70], we show that with high probability any stationary point of f_S in a bounded set satisfies at least one of the following

$$\|(w, x)\| \le \Delta \|(\bar{w}, \bar{x})\|, \quad \|wx^\top - \bar{w}\bar{x}^\top\| \le \Delta \|\bar{w}\bar{x}^\top\|, \quad \text{or} \quad \begin{cases} |\langle w, \bar{w} \rangle| & \le \Delta \|(w, x)\| \|\bar{w}\|, \\ |\langle x, \bar{x} \rangle| & \le \Delta \|(w, x)\| \|\bar{x}\|. \end{cases}$$

provided that $m \gtrsim d_1 + d_2$, where $\Delta = \widetilde{O}\left(\frac{d_1+d_2}{m}\right)^{\frac{1}{8}}$.[4] Intuitively this means, that as the ratio $(d_1 + d_2)/m$ goes to zero, the stationary points lie closer and closer to three sets: the singleton zero, the set of solutions $\mathcal{S}$, and the subspace $(\bar{w}, 0)^\perp \cap (0, \bar{x})^\perp$.

[4] Where $\widetilde{O}$ hides a logarithmic terms.

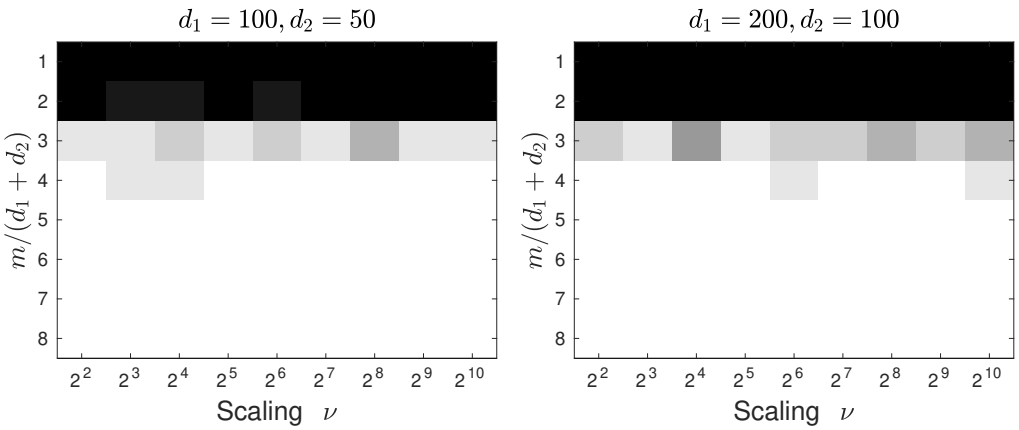


Figure 6.2: Empirical probability of recovery with random initailization on a cube $[-\nu, \nu]^{d_1+d_2}$. White denotes probability one and black denotes probability zero. Left and right images correspond to $(d_1, d_2) = (100, 50)$ and $(d_1, d_2) = (200, 100)$, respectively.

6.4.1 Interlude: Singular value functions

Singular value functions play a crucial role in our studies, we now take a moment to introduce them. For a pair of dimensions d_1, d_2 we will denote $d = \min\{d_1, d_2\}$. A function $f\colon \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$ is *symmetric* if $f(\pi x) = f(x)$ for any permutation matrix $\pi \in \{0,1\}^{d\times d}$. Additionally, a function f is *sign invariant* if $f(sx) = f(x)$ for any diagonal matrix $s \in \{-1, 0, 1\}^{d\times d}$ with diagonal entries in $\{\pm 1\}$. We say that $f_\sigma\colon \mathbf{R}^{d_1\times d_2} \to \mathbf{R} \cup \{\infty\}$ is a *singular value function* if it can be decomposed as $f_\sigma = (f \circ \sigma)$ for a symmetric sign invariant function f. A simple and illuminating example is the Frobenius norm, since $\|A\|_F = \|\sigma(A)\|_2$. This type of functions has been deeply studied in variational analysis [147, 150, 149].

A pair of matrices X and Y in $\mathbf{R}^{d_1\times d_2}$ have a *simultaneous ordered singular value decomposition* if there exist matrices $U \in O(d_1)$ and $V \in O(d_2)$ such that $X = U\mathrm{diag}(\sigma(X))V^\top$ and $Y = U\mathrm{diag}(\sigma(Y))V^\top$. We will make use of the following remarkable theorem.

Theorem 6.4.1 (Proposition 6.1 and Theorem 7.1 in [149])**.** *Let* $f_\sigma\colon \mathbf{R}^{d_1\times d_2} \to \mathbf{R} \cup$

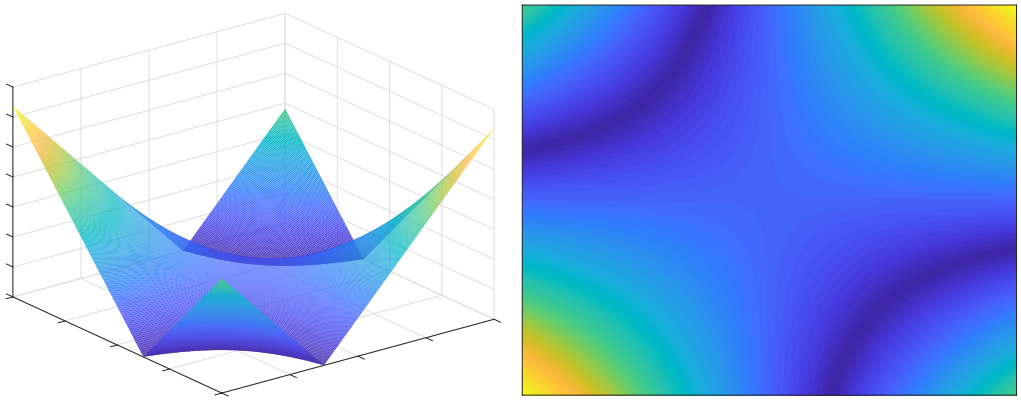


Figure 6.3: Population objective $d_1 = d_2 = 1$.

$\{\infty\}$ *be a singular value function with* $f_\sigma = f \circ \sigma$*. Then,* f_σ *is convex, if and only if,* f *is convex. Furthermore, the limiting subdifferential of* f_σ *at a matrix* $M \in \mathbf{R}^{d_1 \times d_2}$ *is given by*

$$\partial f_\sigma(M) = \{U\mathrm{diag}(\zeta)V^\top \mid \zeta \in \partial f(\sigma(M)) \text{ and } U\mathrm{diag}\,(\sigma(M))\,V^\top = M\}. \tag{6.10}$$

Hence M *and any of its subgradients have simultaneous ordered singular value decomposition.*

6.4.2 Population objective

In this subsection, we study the population objective f_P. A first important observation is that this function is a singular value function. Indeed, if we set $X = wx^\top - \bar{w}\bar{x}^\top$ then due to the orthogonal invariance of the Gaussian distribution we get

$$f_P(w, x) = \mathbb{E}|\ell^\top U\mathrm{diag}(\sigma(X))V^\top r| = \mathbb{E}|\sigma_1(X)\ell_1 r_1 + \sigma_2(X)\ell_2 r_2|, \tag{6.11}$$

where of course $U\mathrm{diag}(\sigma(X))V^\top$ is the singular value decomposition of X. This simple observation leads to our first result, a closed form characterization of

this function in terms $\sigma(X)$. We defer the proof to Section 6.6.3.

Proposition 6.4.2 (Population objective)**.** *The population objective can be written as*

$$f_P(w, x) = \sigma_{\max}(X) \sum_{n=0}^{\infty} \left(\frac{(2n)!}{2^{2n}(n!)^2} \right)^2 \frac{\left(1 - \kappa^{-2}(X)\right)^n}{1 - 2n} \tag{6.12}$$

where $\kappa(X) = \sigma_{\max}(X)/\sigma_{\min}(X)$ *is the condition number of* X*.*

When the signal $(\bar{w}, \bar{x})$ lives in $\mathbf{R}^2$ the landscape of the population objective is rather simple, the only critical points are the solutions and zero, see Figure 6.3. This is not the case in higher dimensions where an entire subspace of critical points appear. In the reminder of this section, we develop tools to describe the critical points of a broad class of functions and we then specialize these results to the blind deconvolution population objective (6.11).

Landscape analysis for a class of singular value functions

To characterize the critical points of f_P we will study a broader class of functions. We consider an arbitrary function $g : \mathbf{R}^{d_1} \times \mathbf{R}^{d_2} \to \mathbf{R}$ for which there exists a rank one matrix $\bar{w}\bar{x}^\top$ and a singular value function f_σ satisfying

$$g(w, x) = f_\sigma(wx^\top - \bar{w}\bar{x}^\top) = f \circ \sigma\left(wx^\top - \bar{w}\bar{x}^\top\right).$$

This gives us two useful characterizations of g that we will use throughout. In the following section we will see a way of recasting f_P in this form.

A simple application of the chain rule yields

$$\partial g(w, x) = \left\{ \begin{bmatrix} Yx \\ Y^\top w \end{bmatrix} \,\middle|\, Y \in \partial f_\sigma(X) \right\}. \tag{6.13}$$

Notice that we already have a description of $\partial f_\sigma(X)$ given by Theorem 6.4.1, that is $Y \in \partial f_\sigma(X)$ if and only if there exists matrices $U \in O(d_1)$ and $V \in O(d_2)$ satisfying

$$\sigma(Y) \in \partial f(\sigma(X)), \qquad Y = U\mathrm{diag}(\sigma(Y))V^\top, \qquad \text{and} \qquad X = U\mathrm{diag}(\sigma(X))V^\top. \tag{6.14}$$

Equipped with these tools we derive the following result regarding the critical points of g. We defer a proof to Section 6.6.3.

Theorem 6.4.3. *Let $g : \mathbf{R}^{d_1} \times \mathbf{R}^{d_2}$ be a function that can be decomposed as*

$$g(w, x) = f_\sigma(wx^\top - \bar{w}\bar{x}^\top) = f \circ \sigma\left(wx^\top - \bar{w}\bar{x}^\top\right),$$

where $f : \mathbf{R}^d \to \mathbf{R}$ is a symmetric sign invariant convex function.[5] *Suppose that (w, x) is a stationary point of g, meaning that $Yx = 0, Y^\top w = 0$ for some $Y \in \partial f_\mu(X)$. Then at least one of the following conditions hold:*

1. ***Small objective.*** $g(w, x) \le g(\bar{w}, \bar{x})$,
2. ***Zero.*** $(w, x) = 0$,
3. ***One zero component.*** $\langle w, \bar{w}\rangle = \langle x, \bar{x}\rangle = 0$, $wx^\top = 0$, *and (assuming that x is not zero) $Yx = 0$ (similarly for w).*
4. ***Small product norm.*** $\langle w, \bar{w}\rangle = \langle x, \bar{x}\rangle = 0$, $\mathrm{rank}(Y) = 1$, *and* $0 < \|wx^\top\| < \|\bar{w}\bar{x}^\top\|$.

Moreover, if $(\bar{w}, \bar{x})$ minimizes g, then (w, x) is a critical point if, and only if, it satisfies 1, 2, 3, or 4 for some $Y \in \partial f_\sigma(X)$.

[5] Recall $d = \min\{d_1, d_2\}$.

Landscape of the population objective

Our goal now is to apply Theorem 6.4.3 to describe the landscape of f_P. In order to do it we need to write $f_P(w, x) = f \circ \sigma(X)$ with $f\colon \mathbf{R}^d \to \mathbf{R}$ a symmetric sign-invariant convex function. An easy way to do this is to define

$$f(s_1, \dots, s_d) = \mathbb{E}\left(\left|\sum_{i=1}^{d} \ell_i r_i s_i\right|\right).$$

where $\ell, r \sim \mathrm{N}(0, I_d)$ The next lemma shows that the function is actually differentiable at every point but zero. We defer the proof of this result to Section 6.6.3.

Lemma 6.4.4. *For any nonzero vector $s \in \mathbf{R}^d_+ \setminus \{0\}$, the partial derivatives of f satisfy*

$$\frac{\partial f}{\partial s_j}(s) = \sqrt{\frac{2}{\pi}} s_j \, \mathbb{E}\left[\ell_j^2 \left(\sum_i (\ell_i s_i)^2\right)^{-\frac{1}{2}}\right]. \tag{6.15}$$

This lemma gives us the final tool to derive the main theorem regarding the landscape of f_P.

Theorem 6.4.5. *The set of critical points of the population objective g_P is exactly*

$$\{0\} \cup \{(w, x) \mid wx^\top = \bar{w}\bar{x}^\top\} \cup \{(w, x) \mid \langle w, \bar{w}\rangle = 0, \langle x, \bar{x}\rangle = 0, \textit{and } wx^\top = 0\}.$$

Proof. Notice that $(\bar{w}, \bar{x})$ minimizes the population objective f_P, therefore Theorem 6.4.3 gives a complete description of the critical points. Let us examine each one of the conditions in this theorem.

The points in $\{(w, x) \mid wx^\top = \bar{w}\bar{x}^\top\}$ and $\{0\}$ are contained in the set of stationary points because they satisfy the first and second condition, respectively.

Now, let $(w, x) \in \{\bar{w}\}^\perp \times \{\bar{x}\}^\perp$ such that $wx^\top = 0$. Thus, the matrix X is rank 1, and consequently (6.15) reveals that that any $Y \in \partial f_\sigma(X)$ satisfies $\sigma(Y) =$

$\nabla f(\sigma(X)) = (2/\pi, 0, \ldots, 0)$. Therefore, due to (6.14), we get $Y = \frac{2}{\pi}\bar{w}\bar{x}^\top/\|\bar{w}\|\|\bar{x}\|$. Without loss of generality, assume x is not zero. Then, $\|Yx\| = \frac{2}{\pi\|\bar{w}\|\|\bar{x}\|}|\langle x, \bar{x}\rangle| = 0$ and, consequently, (w, x) is stationary.

On the other hand, let $(w, x) \in \{\bar{w}\}^\perp \times \{\bar{x}\}^\perp$ such that $0 < \|wx^\top\| < \|\bar{w}\bar{x}^\top\|$. Therefore, the matrix X is rank 2 and so (6.15) gives that $\sigma_2(Y) > 0$ for all $Y \in \partial f_\sigma(X)$. Hence, (w, x) is not a stationary point, giving the result. □

6.4.3 Sample objective

In this section, we describe the approximate locations of the critical points of the sample objective. Unlike in the smooth case, nonsmooth losses do not exhibit point-wise concentration of the subgradients, or in other words, $\partial f_S(w, x)$ does not converges to $\partial f_P(w, x)$ as $m \to \infty$. To overcome this obstacle, we show that the graph of ∂f_S approaches that of ∂f_P at a quantifiable rate. Intuitively, if (w, x) is a critical point of f_S, then nearby there exists a point $(\widehat{w}, \widehat{x})$ with $\operatorname{dist}(0, \partial f_P(\widehat{w}, \widehat{x}))$ small.

The following result can be regarded as an analogous version of Theorem 6.4.5 for the sample objective. The proof of this result is more involved and will require us to study the location of epsilon critical points of the population. We defer the development of these arguments and the proof of the next result to Sections 6.6.3 and 6.6.3, respectively.

Theorem 6.4.6. *Consider the sample objective* (6.9) *generated with two Gaussian matrices L and R. For any fixed $\nu > 1$ there exist numerical constants $c_1, c_2, c_3 > 0$ such that if $m \geq c_1(d_1 + d_2 + 1)$, then with probability at least $1 - c_2 \exp(-c_3(d_1 + d_2 + 1))$, every stationary point (w, x) of f_S for which $\|(w, x)\| \leq \nu\|(\bar{w}, \bar{x})\|$ satisfies at least one of*

the following conditions

1. ***(Near zero)***

$$\|(w, x)\| \leq \|(\bar{w}, \bar{x})\|\Delta,$$

2. ***(Near a solution)***

$$\|wx^\top - \bar{w}\bar{x}^\top\| \lesssim (\nu^2 + 1)\|\bar{w}\bar{x}^\top\|\Delta,$$

3. ***(Near orthogonal)***

$$\begin{cases} |\langle w, \bar{w}\rangle| & \lesssim (\nu^2 + 1)\|(w, x)\|\|\bar{w}\|\Delta, \\ |\langle x, \bar{x}\rangle| & \lesssim (\nu^2 + 1)\|(w, x)\|\|\bar{x}\|\Delta. \end{cases}$$

where $\Delta = \left(\frac{d_1+d_2+1}{m} \log\left(\frac{m}{d_1+d_2+1}\right)\right)^{\frac{1}{2}}$.

6.5 Numerical Experiments

In this section we demonstrate the performance and stability of the prox-linear and subgradient methods, and the initialization procedure, when applied to real and artificial instances of Problem (6.2). All experiments were performed using the programming language `Julia` [22]. A reference implementation and code for the experiments is available in `https://github.com/COR-OPT/RobustBlindDeconv`.

Subgradient method implementation. Implementation of the subgradient method for Problem (6.2) is simple, and has low per-iteration cost. Indeed, one

may simply choose the subgradient

$$\frac{1}{m}\sum_{i=1}^{m}\operatorname{sign}(\langle \ell_i, w\rangle\langle x, r_i\rangle - y)\left(\langle x, r_i\rangle\begin{bmatrix}\ell_i\\0\end{bmatrix} + \langle \ell_i, w\rangle\begin{bmatrix}0\\r_i\end{bmatrix}\right) \in \partial f(w, x),$$

where $\operatorname{sign}(t)$ denotes the sign of t, with the convention $\operatorname{sign}(0) = 0$. The cost of computing this subgradient is on the order of four matrix multiplications. When applying Algorithm 3, choosing the correct parameters is important, since its convergence is especially sensitive to the value of the step-size decay q; the experiment described in Section 6.5.1 demonstrates this sensitivity. Setting $\lambda = 1.0$ sufficed for all the experiments depicted hereafter.

Prox-linear method implementation. Recall that the convex models used by the prox-linear method take the form:

$$f_{(w_k,x_k)}(w, x) = \frac{1}{m}\|\mathcal{A}(w_k x_k^\top + w_k(x - x_k)^\top + (w - w_k)x_k^\top) - y\|_1 \tag{6.16}$$

Equivalently, one may rewrite this expression as a Least Absolute Deviation (LAD) objective:

$$\begin{aligned} f_{(w_k,x_k)}(w, x) &= \frac{1}{m}\sum_{i=1}^{m}\Big|\underbrace{\left(\begin{array}{c|c}\langle x_k, r_i\rangle \ell_i^\top & \langle \ell_i, w_k\rangle r_i^\top\end{array}\right)}_{A_i}\underbrace{\begin{pmatrix}w - w_k\\ x - x_k\end{pmatrix}}_{z} - \underbrace{(b_i - \langle \ell_i, w_k\rangle\langle x_k, r_i\rangle)}_{\tilde{b}_i}\Big| \\ &= \frac{1}{m}\|Az - \tilde{y}\|_1 . \end{aligned}$$

Thus, each iteration of Algorithm 4 requires solving a strongly convex optimization problem:

$$z_{k+1} = \operatorname*{arg\,min}_{z\in\mathcal{S}_\nu}\left\{\frac{1}{m}\|Az - \tilde{y}\|_1 + \frac{1}{2\alpha}\|z\|_2^2\right\}.$$

Motivated by the work of [89] on robust phase retrieval, we solve this subproblem with the *graph splitting* variant of the Alternating Direction Method of Multipliers, as described in [200]. This iterative method applies to problems of the form

$$\begin{aligned} \min_{z \in \mathcal{X}} \quad & \frac{1}{m} \|t - \tilde{y}\|_1 + \frac{1}{2\alpha} \|z\|_2^2 \\ \text{s.t.} \quad & t = Az. \end{aligned}$$

The ADMM method takes the form:

$$\begin{aligned} z' &\leftarrow \arg\min_{z \in \mathcal{S}_\nu} \left\{ \frac{1}{2\alpha} \|z\|_2^2 + \frac{\rho}{2} \|z - (z_k - \lambda_k)\|_2^2 \right\} \\ t' &\leftarrow \arg\min_{t} \left\{ \frac{1}{m} \|t - \tilde{y}\|_1 + \frac{\rho}{2} \|t - (t_k - \nu_k)\|_2^2 \right\} \\ \begin{pmatrix} z_+ \\ t_+ \end{pmatrix} &\leftarrow \begin{bmatrix} I_{d_1+d_2} & A^\top \\ A & -I_m \end{bmatrix}^{-1} \begin{bmatrix} I_{d_1+d_2} & A^\top \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{pmatrix} z' + \lambda \\ t' + \nu \end{pmatrix} \\ \lambda_+ &\leftarrow \lambda + (z' - z_+), \nu_+ \leftarrow \nu + (t' - t_+), \end{aligned}$$

where $\lambda \in \mathbf{R}^{d_1+d_2}$ and $\nu \in \mathbf{R}^m$ are dual multipliers and $\rho > 0$ is a control parameter. Each above step may be computed analytically. We found in our experiments that choosing $\alpha = 1$ and $\rho \sim \frac{1}{m}$ yielded fast convergence. Our stopping criteria for this subproblem is considered met when the primal residual satisfies $\|(z_+, t_+) - (z, t)\| \leq \epsilon_k \cdot \left(\sqrt{d_1 + d_2} + \max\{\|z\|_2, \|t\|_2\} \right)$ and the dual residual satisfies $\|(\lambda_+, \nu_+) - (\lambda, \nu)\| \leq \epsilon_k \cdot \left(\sqrt{d_1 + d_2} + \max\{\|\lambda\|_2, \|\nu\|_2\} \right)$ with $\epsilon_k = 2^{-k}$.

6.5.1 Artificial Data

We first illustrate the performance of the prox-linear and subgradient methods under noise model **N1** with i.i.d. standard Gaussian noise ξ_i. Both methods are

initialized with Algorithm 6. We experimented with Gaussian noise of varying variances, and observed that higher levels did not adversely affect the performance of our algorithm. This is not surprising, since the theory suggests that both the objective and the initialization procedure are robust to gross outliers. We analyze the performance with problem dimensions $d_1 \in \{400, 1000\}$ and $d_2 = 500$ and with number of measurements $m = c \cdot (d_1 + d_2)$ with c varying from 1 to 8. In Figures 6.4 and 6.5, we have depicted how the quantity

$$\frac{\left\| w_k x_k^\top - \bar{w}\bar{x}^\top \right\|_F}{\|\bar{w}\bar{x}^\top\|_F}$$

changes per iteration for the prox-linear and subgradient methods. We conducted tests in both the moderate corruption ($p_{\text{fail}} = .25$) and high corruption ($p_{\text{fail}} = .45$) regimes. For both methods, under moderate corruption ($p_{\text{fail}} = .25$) we see that exact recovery is possible as long as $c \geq 5$. Likewise, even in high corruption regime ($p_{\text{fail}} = .45$) exact recovery is still possible as long as $c \geq 8$. We also illustrate the performance of Algorithm 2 when there is no corruption at all in Figure 6.4, which converges an order of magnitude faster than Algorithm 3.

In terms of algorithm performance, we see that the prox-linear method takes few outer iterations, approximately 15, to achieve very high accuracy, while the subgradient method requires a few hundred iterations. This behavior is expected as the prox-linear method converges quadratically and the subgradient method converges linearly. Although the number of iterations of the prox-linear method is small, we demonstrate in the sequel that its total run-time, including the cost of solving subproblems, can be higher than the subgradient method. Interestingly, Figure 6.5 shows how the performance of the prox-linear method stagnates for the first few iterations before dropping at a quadratic rate. This might indicate that for these choices of c the initialization procedure outputs a point slightly outside of the region of quadratic convergence. Another pos-

sibility is that the levels of accuracy set for solving the proximal subproblems, $\varepsilon_t := 2^{-t},\ t = 1, \ldots, T$, are not "fine" enough for the first few iterations.

Number of matrix-vector multiplications

Each iteration of the prox-linear method requires the numerical resolution of a convex optimization problem. We solve this subproblem using the *graph splitting* ADMM algorithm, as described in [200], the cost of which is dominated by the number of matrix vector products required to reach the target accuracy. The number of "inner iterations" of the prox-linear method and thus the number of matrix vector products is not determined a priori. The cost of each iteration of the subgradient method, on the other hand, is on the order of 4 matrix vector products. In the subsequent plots, we solve a sequence of synthetic problems for $d_1 = d_2 = 100$ and keep track of the total number of matrix-vector multiplications performed. We run both methods until we obtain $\frac{\left\|wx^\top - \bar{w}\bar{x}^\top\right\|_F}{\|\bar{w}\bar{x}^\top\|_F} \leq 10^{-5}$. Additionally, we keep track of the same statistics for the subgradient method. We present the results in Fig. 6.6. We observe that the number of matrix-vector multiplications required by the prox-linear method can be much greater than those required by the subgradient method. Additionally, they seem to be much more sensitive to the ratio $\frac{m}{d_1+d_2}$.

Choice of step size decay

Due to the sensitivity of Algorithm 3 to the step size decay q, we experiment with different choices of q in order to find an empirical range of values which yield acceptable performance. To that end, we generate synthetic problems of

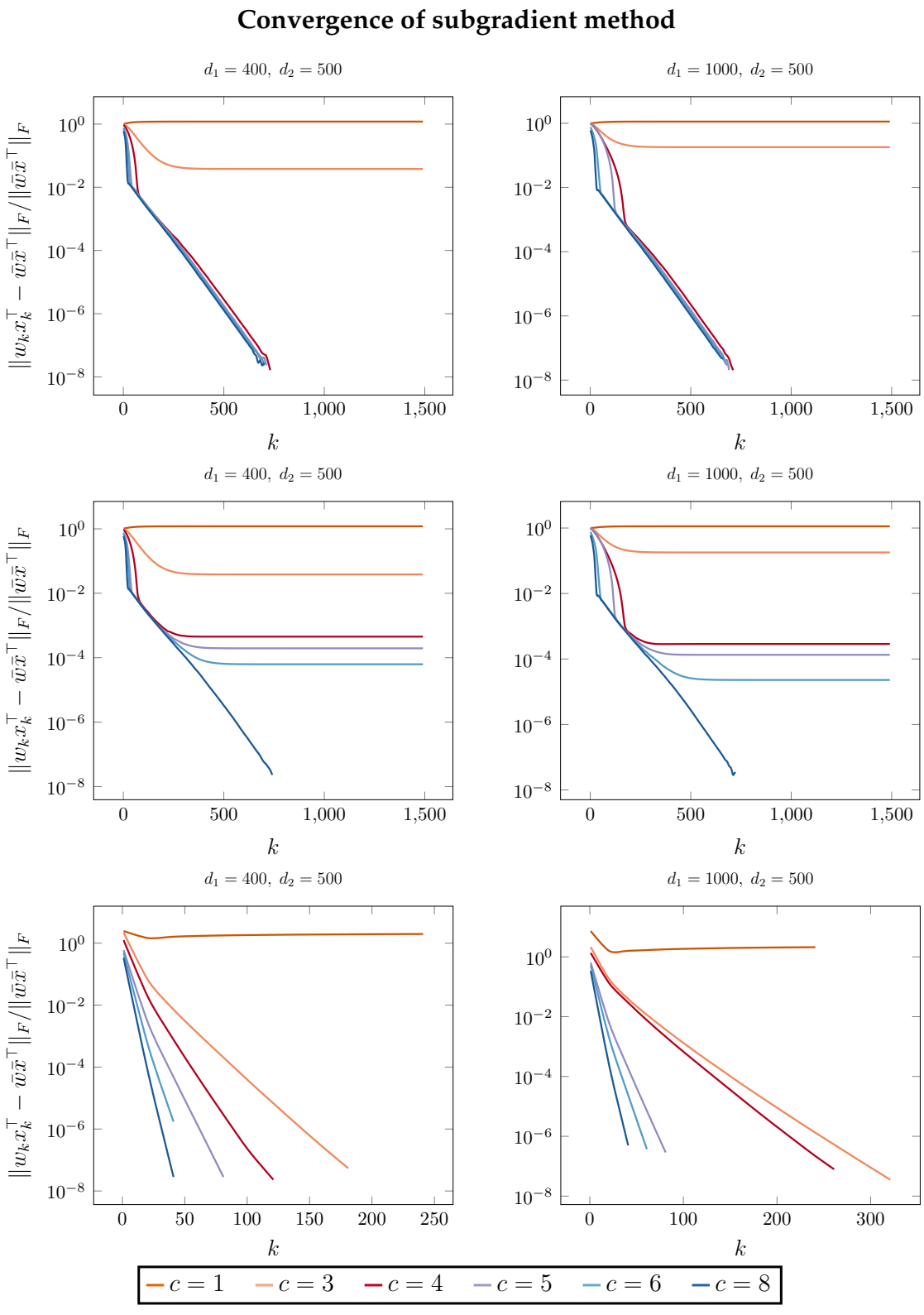


Figure 6.4: Error $\left\|w_k x_k^\top - \bar{w}\bar{x}^\top\right\|_F / \left\|\bar{w}\bar{x}^\top\right\|_F$ vs iteration count. Top row is using Algorithm 3 with $p_{\text{fail}} = 0.25$. Second row is using Algorithm 3 with $p_{\text{fail}} = 0.45$. Third row is using Algorithm 2 with $p_{\text{fail}} = 0$.

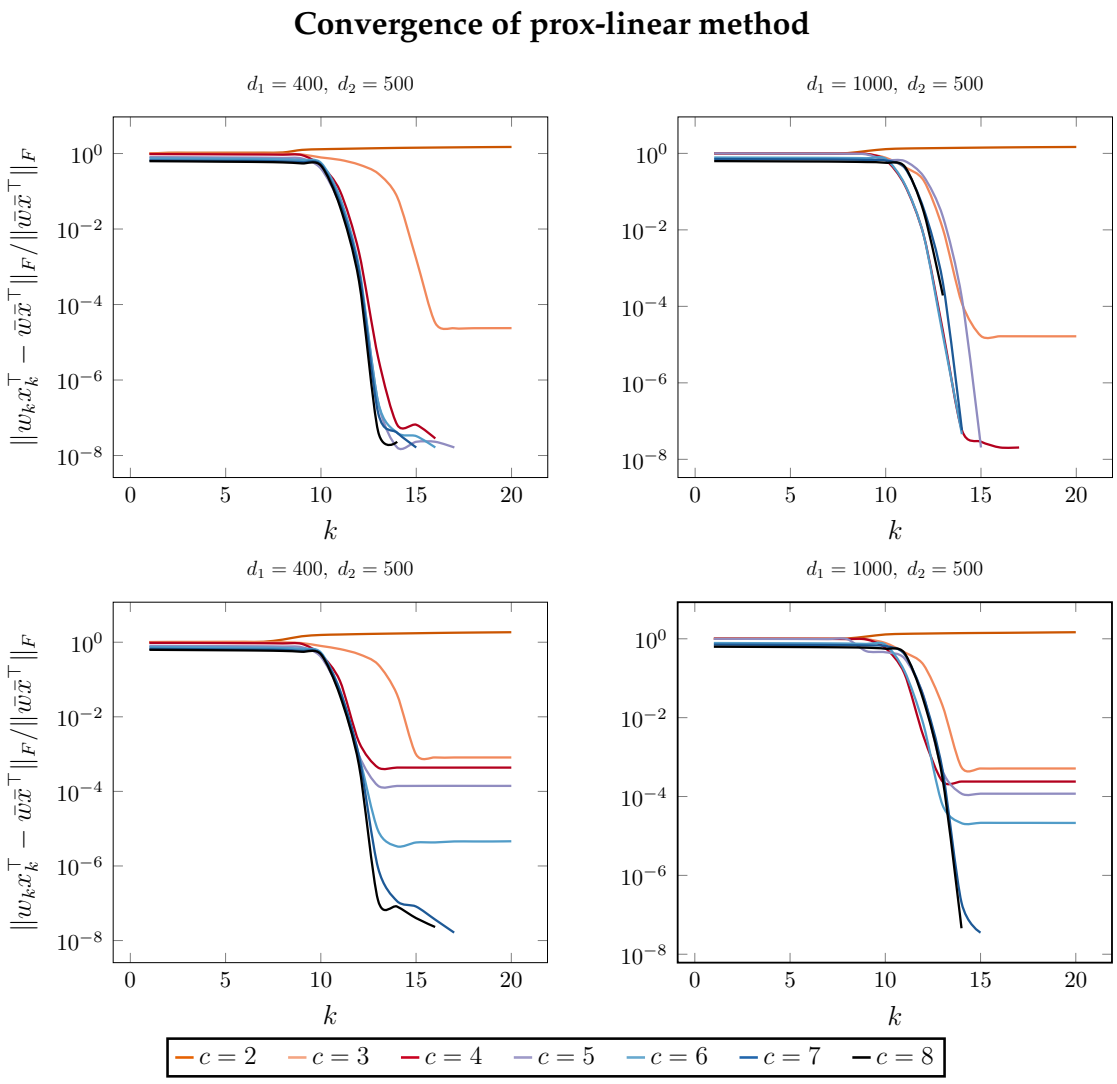


Figure 6.5: Error $\left\|w_k x_k^\top - \bar{w}\bar{x}^\top\right\|_F / \left\|\bar{w}\bar{x}^\top\right\|_F$ vs iteration count for an application of Algorithm 4 in the two settings: $p_{\text{fail}} = 0.25$ (top row) and $p_{\text{fail}} = 0.45$ (bottom row).

dimension 100×100 and choose $q \in \{0.90, 0.905, \ldots, 0.995\}$, and record the average error of the final iterate after 1000 iterations of the subgradient method for different choices of $m = c \cdot (d_1 + d_2)$. The average is taken over 50 test runs with $\lambda = 1.0$. We test both noisy and noiseless instances to see if corruption of entries significantly changes the effective range of q. Results are shown in Fig. 6.7.

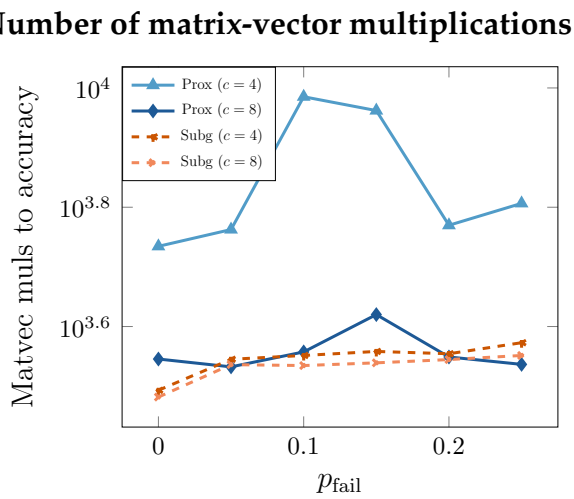


Figure 6.6: Matrix-vector multiplications to reach rel. accuracy of 10^{-5}.

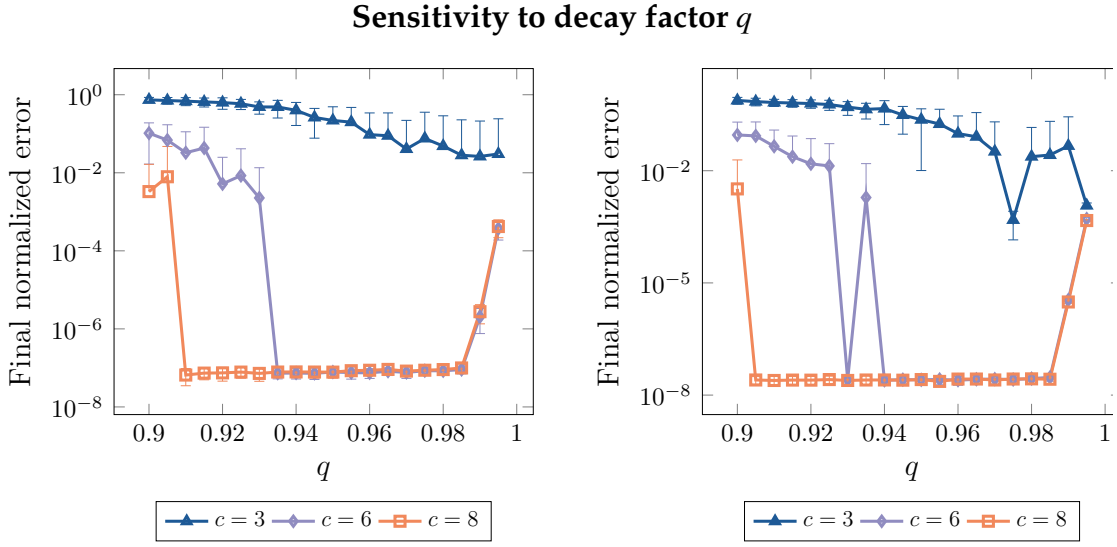


Figure 6.7: Final normalized error $\left\|w_k x_k^\top - \bar{w}\bar{x}^\top\right\|_F / \left\|\bar{w}\bar{x}^\top\right\|_F$ for Algorithm 3 with different choices of q, in the settings $p_{\text{fail}} = 0$ (left) and $p_{\text{fail}} = 0.25$ (right).

Robustness to noise

We now empirically validate the robustness of the prox-linear and subgradients algorithms to noise. In a setup familiar from other recent works [89, 8], we generate *phase transition plots*, where the x-axis varies with the level of corruption p_{fail}, the y-axis varies as the ratio $\frac{m}{d_1+d_2}$ changes, and the shade of each pixel represents the percentage of problem instances solved successfully. For every

configuration $(p_{\text{fail}}, m/(d_1 + d_2))$, we run 100 experiments.

Noise model N1 - independent noise Initially, we experiment with Gaussian random matrices and $(d_1, d_2) \in \{(100, 100), (200, 200)\}$, the results for which can be found in the top row of Fig. 6.8.

The phase transition plots are similar for both dimensionality choices, revealing that in the moderate independent noise regime ($p_{\text{fail}} \leq 25\%$), setting $m \geq 4(d_1 + d_2)$ suffices. On the other hand, for exact recovery in high noise regimes ($p_{\text{fail}} \simeq 45\%$), one may need to choose m as large as $8 \cdot (d_1 + d_2)$.

Noise model N2 - arbitrary noise We now repeat the previous experiments, but switch to noise model **N2**. In particular, we now adversarially hide a different signal in a subset of measurements, i.e., we set

$$b_i = \begin{cases} \langle \ell_i, \bar{w} \rangle \langle \bar{x}, r_i \rangle, & i \notin \mathcal{I}_{\text{in}}, \\ \langle \ell_i, \bar{w}_{\text{imp}} \rangle \langle \bar{x}_{\text{imp}}, r_i \rangle & i \in \mathcal{I}_{\text{out}}, \end{cases}$$

where in the above $(\bar{w}_{\text{imp}}, \bar{x}_{\text{imp}}) \in \mathbf{R}^{d_1} \times \mathbf{R}^{d_2}$ is an arbitrary pair of signals. Intuitively, this is a more challenging noise model than **N1**, since it allows an adversary try to trick the algorithm into recovering an entirely different signal. Our experiments confirm that this regime is indeed more difficult for the proposed algorithms, which is why we only depict the range $p_{\text{fail}} \in [0, 0.38]$ in the bottom row of Fig. 6.8 below.

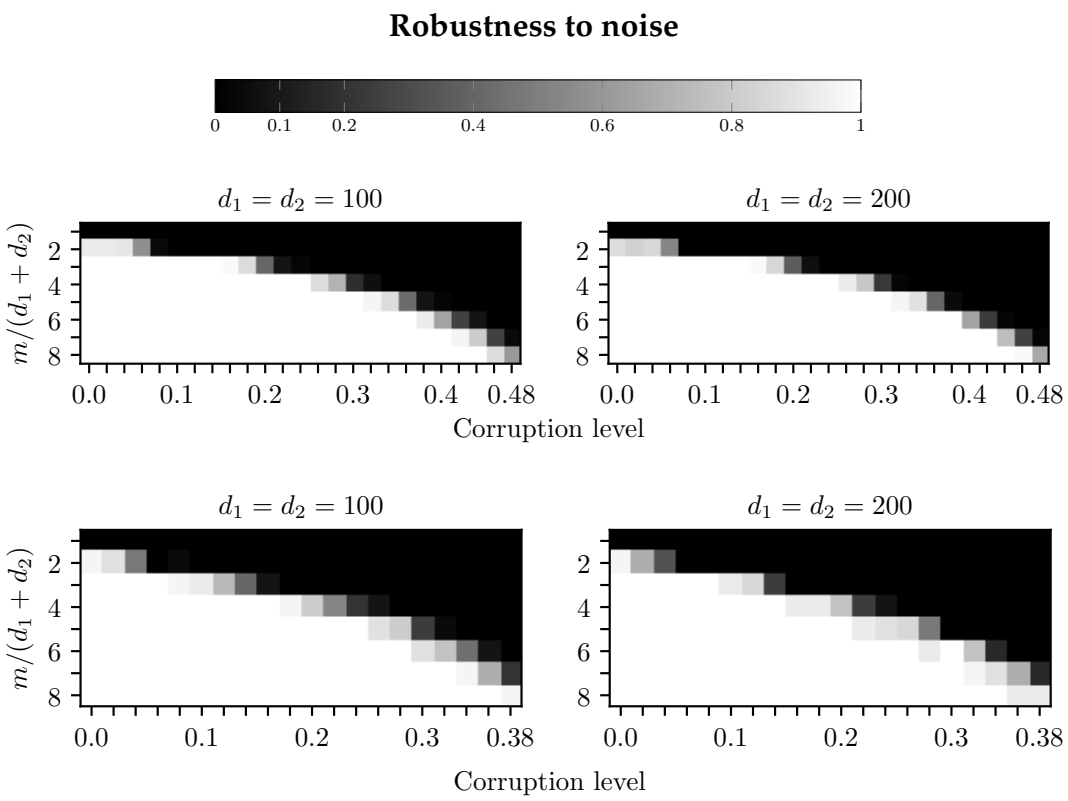


Figure 6.8: Empirical recovery probabilities for matrix model **M** and noise models **N1** (**top**) and **N2** (**bottom**) across 100 independent runs using Algorithm 3. Lighter cells imply higher recovery probability.

6.5.2 Performance of initialization on real data

We now demonstrate the proposed initialization strategy on real world images. Specifically, we set $\bar{w}$ and $\bar{x}$ to be two random digits from the training subset of the MNIST dataset [140]. In this experiment, the measurement matrices $L, R \in \mathbf{R}^{(16 \cdot 784) \times 784}$ have i.i.d. Gaussian entries, and the noise follows Model **N1** with $p_{\text{fail}} = 0.45$. We apply the initialization method and plot the resulting images (initial estimates) in Fig. 6.9. Evidently, the initial estimates of the images are visually similar to the true digits, up to sign; in other examples, the foreground appears to be switched with the background, which corresponds to the natural sign ambiguity. Finally, we plot the normalized error for the two recov-

ery methods (subgradient and prox-linear) in Fig. 6.10.

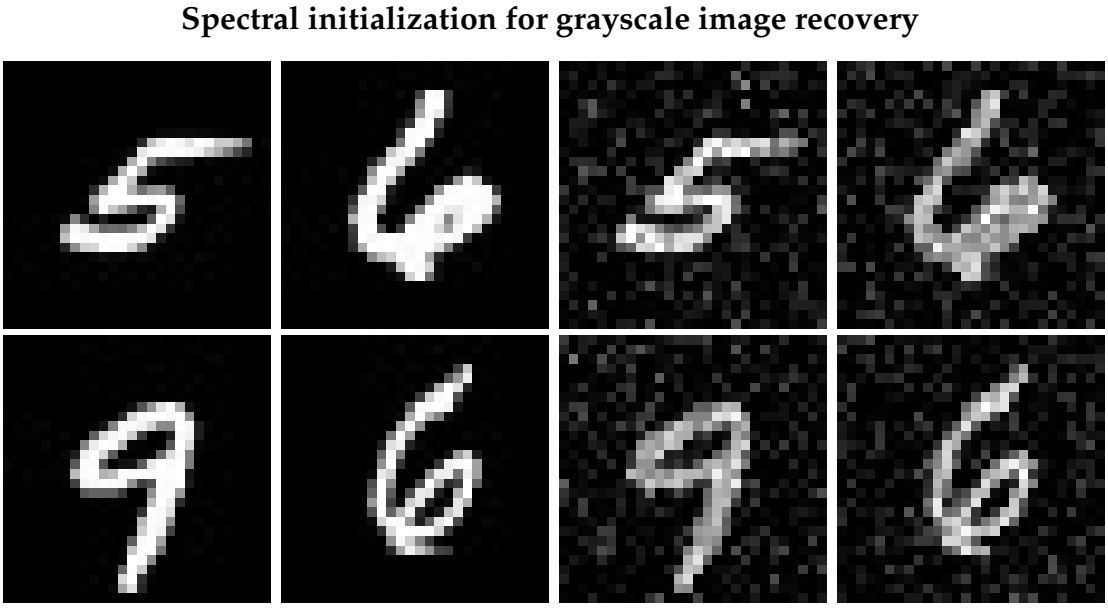


Figure 6.9: Digits 5, 6 (top) and 9, 6 (bottom). Original images are shown on the left, estimates from spectral initialization on the right. Parameters: $p_{\text{fail}} = 0.45, m = 16 \cdot 784$.

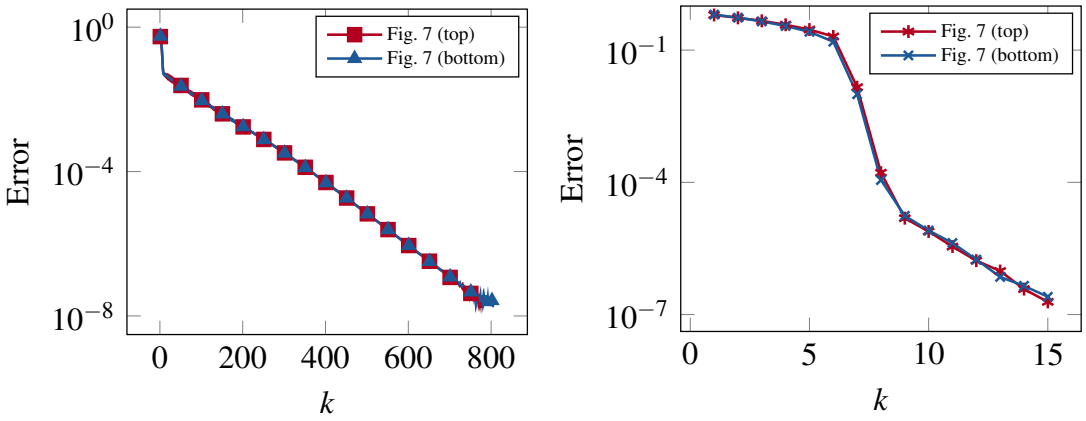


Figure 6.10: Relative error vs iteration count on `mnist` digits for subgradient method (left) and prox-linear method (right).

6.5.3 Experiments on Big Data

We apply the subgradient method for recovering large-scale real color images $W, X \in \mathbf{R}^{n\times n\times 3}$. In this setting, $p_{\text{fail}} = 0.0$ so using Algorithm 2 is applicable with $\min_X f = 0$. We "flatten" the matrices W, X into $3n^2$ dimensional vectors w, x. In contrast to the previous experiments, our sensing matrices are of the following form:

$$L = \begin{bmatrix} HS_1 \\ \vdots \\ HS_k \end{bmatrix}, \; R = \begin{bmatrix} HS_1' \\ \vdots \\ HS_k' \end{bmatrix},$$

where $H \in \{-1, 1\}^{d\times d} / \sqrt{d}$ is the $d \times d$ symmetric normalized Hadamard matrix and $S_i = \text{diag}(s_1, \ldots, s_d)$, where $s \sim_{\text{i.i.d}} \{-1, 1\}$, is a diagonal random sign matrix. The same holds for S_i'. Notice that we can perform the operations $w \mapsto Lw, \; x \mapsto Rx$ in $O(kd \log d)$ time: we first form the elementwise product between the signal and the random signs, and then take its Hadamard transform, which can be performed in $O(d \log d)$ flops. We can efficiently compute $p \mapsto L^\top p, \; q \mapsto R^\top q$, required for the subgradient method, in a similar fashion. We recover each channel separately, which means we essentially have to solve three similar minimization problems. Notice that this results in dimensionality $d_1 = d_2 = n^2, \; m = kn^2$ for each channel.

We observed that our initialization procedure (Algorithm 6) is extremely accurate in this setting. Therefore to better illustrate the performance of the local search algorithms, we perform the following heuristic initialization. For each channel, we first sample $\widehat{w}, \widehat{x} \sim \mathbb{S}^{d-1}$, rescale by the true magnitude of the signal, and run Algorithm 2 for one step to obtain our initial estimates w_0, x_0.

An example where we recover a pair of 512×512 color images using the

Polyak subgradient method (Algorithm 2) is shown below; Fig. 6.11 shows the progression of the estimates w_k, up until the 90-th iteration, while Fig.6.12 depicts the normalized error at each iteration for the different channels of the images.

Iterates of subgradient method in color image recovery

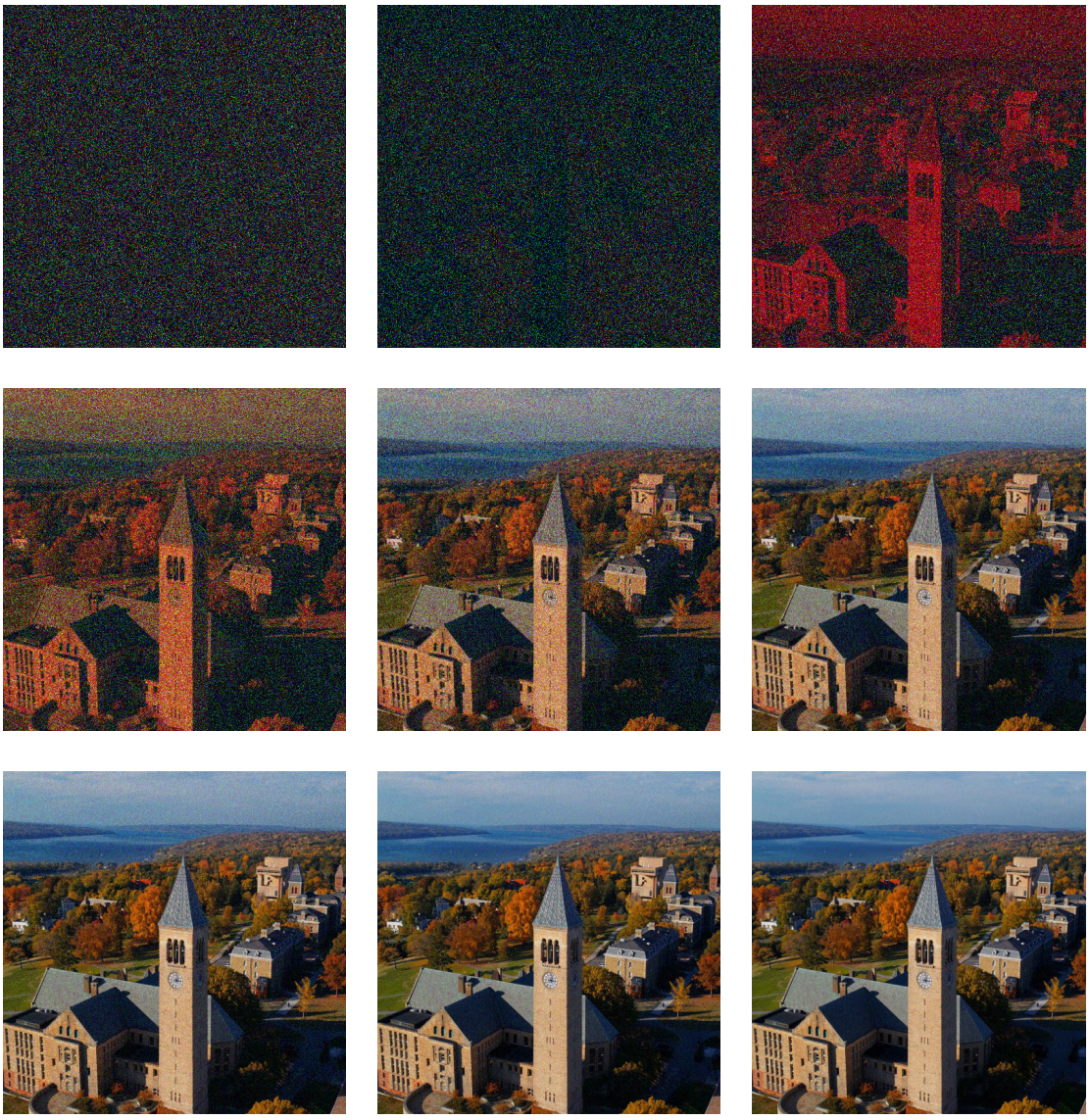


Figure 6.11: Iterates w_{10i}, $i = 1, \ldots, 9$. $(m, k, d, n) = (2^{22}, 16, 2^{18}, 512)$.

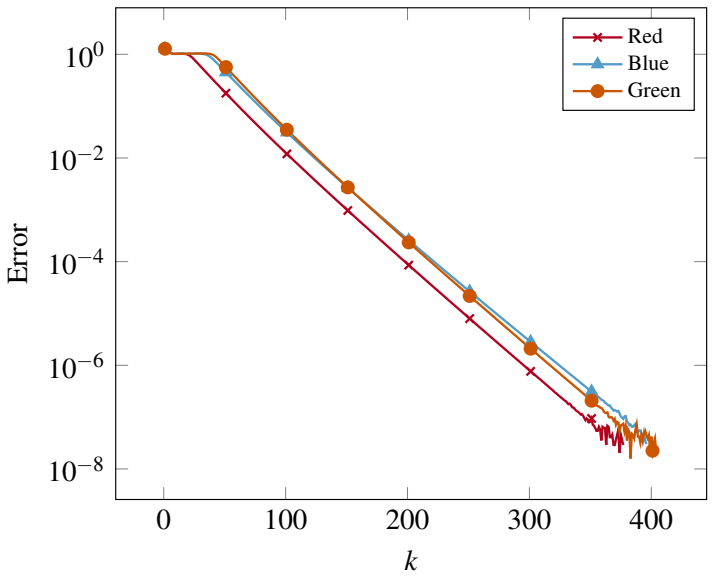


Figure 6.12: Normalized error for different channels in image recovery.

6.5.4 Experiments on complex blind deconvolution

In this section, we experiment on a realistic instance of the blind deconvolution problem, following [8, 153]. Throughout, the measurement vectors ℓ_i and r_i are complex and the vectors ℓ_i are moreover deterministic. Note that this setting is outside the scope of our guarantees, which require all the vectors ℓ_i and r_i to be stochastic; nonetheless, we will see that the proposed methods work well even in this setting.

Recall that the complex vector space $\mathbf{C}^n$ is endowed with the Hermitian inner product $\langle x, y\rangle := x^H y = \sum_{i=1}^n \bar{x}_i b_i$, which satisfies $\langle x, y\rangle = \overline{\langle y, x\rangle}$. In the space of matrices $\mathbf{C}^{m\times n}$, the inner product is defined in an analogous fashion, with $\langle A, B\rangle := \mathrm{Tr}(A^H B)$, with A^H denoting the Hermitian transpose of A. Additionally, we write $\Re(z), \Im(z)$ for the real and imaginary parts of z, understood to hold elementwise if z is a vector.

In the blind deconvolution problem, we observe the circular convolution of

two signals u and v, so that the measurements are

$$y_j = \sum_{i=1}^{m} u_i v_{(j-i+1) \bmod m}. \tag{6.17}$$

In (6.17), we assume that there is no observation noise for the sake of simplicity. A standard assumption is that u, v lie in known low-dimensional subspaces of $\mathbf{R}^m$ of dimensions d_1, d_2 respectively, so that

$$u = Bw_\sharp, \quad v = Cx_\sharp$$

To recast this problem as a bilinear sensing problem, we may pass to the Fourier domain. Denote by F_m the $m \times m$ DFT matrix, with elements

$$(F_m)_{ij} := \exp\left(-\iota 2\pi \frac{(i-1)(j-1)}{m}\right),$$

where we set $\iota = \sqrt{-1}$, and also define

$$L = \overline{F_m B} \in \mathbf{C}^{m \times d_1}, \quad R = F_m C \in \mathbf{C}^{m \times d_2}.$$

Then, following standard arguments (see e.g. [8]) the equivalent model to (6.17) in the Fourier domain is

$$\hat{y}_i = \langle \ell_i, w_\sharp \rangle \cdot \langle x_\sharp, r_i \rangle.$$

A common choice for B is the matrix $\begin{bmatrix} I_{d_1} \\ \mathbf{0} \end{bmatrix}$ [8, 153, 167], which leads to the partial DFT matrix $L \in \mathbf{R}^{m \times d_1}$ formed by taking the first d_1 columns of F_m and used in the experiments below. On the other hand, C is often assumed to have i.i.d. Gaussian entries, so that the entries of R are also i.i.d. and follow the complex Gaussian distribution. For simplicity, we relabel $\hat{y}$ to y in the sequel.

We therefore consider the nonsmooth formulation of the problem

$$\min_{\|w\|, \|x\| \le \nu \sqrt{\Phi}} \frac{1}{m} \sum_{i=1}^{m} \left| \langle \ell_i, w \rangle \langle x, r_i \rangle - y_i \right|, \tag{6.18}$$

where $|x|$ denotes the magnitude of the complex number x. With the help of Wirtinger calculus [133], we describe the extension of the subgradient method in the complex domain. The Wirtinger derivatives of a complex function $f(z)$ with $z = x + iy, (x, y) \in \mathbf{R}^n \times \mathbf{R}^n$ are given by

$$\frac{\partial f}{\partial z} = \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right)$$
$$\frac{\partial f}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right)$$

where, $x \mapsto \bar{x}$ denotes complex conjugation. The chain rule of Wirtinger calculus, summarized below, is useful in formally defining a subgradient of the nonsmooth objective:

$$\frac{\partial (f \circ g)}{\partial z} = \left(\frac{\partial f}{\partial z} \circ g\right)\frac{\partial g}{\partial z} + \left(\frac{\partial f}{\partial \bar{z}} \circ g\right)\frac{\partial \bar{g}}{\partial z} \tag{6.19}$$
$$\frac{\partial (f \circ g)}{\partial \bar{z}} = \left(\frac{\partial f}{\partial z} \circ g\right)\frac{\partial g}{\partial \bar{z}} + \left(\frac{\partial f}{\partial \bar{z}} \circ g\right)\frac{\partial \bar{g}}{\partial \bar{z}} \tag{6.20}$$

We now compute the Wirtinger derivative of the real-valued function

$$f_S(w, x) = \frac{1}{m}\sum_{i=1}^{m} |\langle \ell_i, w\rangle\langle x, r_i\rangle - y_i| = \frac{1}{m}\sum_{i=1}^{m} \left|\mathcal{A}(wx^H)_i - y_i\right|,$$

with $\mathcal{A}(X) = \left\{\ell_i^H X r_i\right\}_{i=1}^m$ the corresponding operator for the complex case. By the definition of the Wirtinger derivatives, it's easy to see that

$$\left.\frac{\partial |z|}{\partial \bar{z}}\right|_{z=z_k} = \begin{cases} 0, & z_k = \mathbf{0} + \mathbf{0}j \\ \frac{z_k}{2|z_k|}, & \text{otherwise} \end{cases}. \tag{6.21}$$

In this way, the linearization around z_k based on the Wirtinger gradient (see [133, pp. 20-21]) satisfies $|z| \geq |z_k| + 2\Re\left(\langle g(z_k), z - z_k\rangle\right)$, $g(z) := \frac{\partial |z|}{\partial \bar{z}}$, after elementary calculations, much like its $\mathbf{R}^n$-counterpart.

With this in hand, the application of the chain rule from Eq. (6.20) gives us

that

$$\partial f_S(w,x) \ni \zeta_k := \begin{pmatrix} \frac{\partial f}{\partial \bar{w}} \\ \\ \frac{\partial f}{\partial \bar{x}} \end{pmatrix} = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{|\langle \ell_i, w_k\rangle\langle x_k, r_i\rangle - y_i|} \begin{bmatrix} (\langle \ell_i, w\rangle\langle x_k, r_i\rangle - y_i)\langle r_i, x_k\rangle \ell_i \\ (\langle w_k, \ell_i\rangle\langle r_i, x_k\rangle - \bar{y}_i)\langle \ell_i, w_k\rangle r_i \end{bmatrix}$$

In the above, we make the convention that when $z = \mathbf{0}$, we set $\frac{z}{|z|} = \mathbf{0}$, as in (6.21).

Experiments. To avoid confusion due to the conjugate notation, in this section we will denote the ground truth signals by $w_\sharp$ and $x_\sharp$, respectively. We repeat the synthetic experiment under noise model **N1** for the case of complex measurements. Specifically, we form the left measurement matrix $L \in \mathbf{C}^{m\times d}$ by taking the first d columns of the (unnormalized) $m \times m$ discrete Fourier transform (DFT) matrix, with $L^H L = mI_d$. For the right measurement matrix $R \in \mathbf{C}^{m\times d}$, we set all entries equal to i.i.d. complex Gaussian random variables:

$$(R)_{i,k} = \sqrt{\frac{1}{2}}\left(X_{i,k} + jY_{i,k}\right), \quad X_{i,k}, Y_{i,k} \sim \mathrm{N}(0,1) \tag{6.22}$$

These are precisely the measurement matrices used in [153], the authors of which also provide a spectral initialization to find an ϵ-close initial estimate. However, this initialization requires $m \asymp d\log d$, so we opt for an artificial initialization as shown in (6.23), with $\delta := 0.25$:

$$w_0 := w_\sharp + \delta g_w, \quad x_0 := x_\sharp + \delta g_x, \quad g_x, g_w \sim \mathrm{Unif}(\mathbb{S}^{d-1}). \tag{6.23}$$

We apply the subgradient methods from Algorithms 2 and 3, with the subgradient now calculated using Wirtinger calculus, as illustrated above. In Figure 6.13, we generate synthetic instances with $\left\|w_\sharp\right\| = \left\|x_\sharp\right\| = 1$ and $p_{\text{fail}} \in \{0, 0.25, 0.45\}$ and evaluate the performance of our methods over a variety of measurement ratios $c := \frac{m}{d}$. We verify the linear rate of convergence of the projected subgradient

method, as well as the effect of p_{fail} on the number of measurements required to converge to a minimizer. We observe that the partial DFT setting requires us to set m as big as $10 \cdot d$ for the highest corruption levels.

Robustness to signal incoherence. For completeness, we evaluate the sensitivity of the nonsmooth formulation (6.2) to the *incoherence* between $w_\sharp$ and the rows of L, given by

$$\mu_h^2 := \frac{\left\|Lw_\sharp\right\|_\infty^2}{\left\|w_\sharp\right\|_2^2} \tag{6.24}$$

Intuitively, μ_h^2 captures the maximal correlation between rows of L and $\bar{w}$; in [153] the authors argue that signals with high μ_h^2 are the hardest to recover for smooth formulations. We generate noiseless instances where L is the partial $m \times d$ DFT matrix and R is a complex Gaussian matrix following (6.22), for a range of values of μ_h^2; for each such value, we set $x_\sharp \sim \text{Unif}(\mathbb{S}^{d-1})$ and $w_\sharp$ equal to a vector with μ_h^2 nonzero elements equal to 1 and all else equal to 0 (followed by normalization so that $w_\sharp \in \mathbb{S}^{d-1}$), which attains incoherence exactly μ_h^2 for this choice of L, following [153]. For simplicity, we set $w_0, x_0 \sim \text{Unif}(\mathbb{S}^{d-1})$, $d = 100$ and $m = 8 \cdot 2d$ and compare:

(i) the performance of Algorithm 2 applied to the objective (6.2),

(ii) the performance of gradient descent (using the Wirtinger gradient) with Polyak stepsize on the smooth counterpart of (6.2), where we replace the ℓ_1-norm with the squared ℓ_2 loss as:

$$f_{\text{smooth}}(w, x) := \frac{1}{m} \sum_{i=1}^{m} \left(\langle \ell_i, w\rangle\langle r_i, x\rangle - b_i\right)^2 \tag{6.25}$$

(iii) the performance of gradient descent applied to (6.25) with a fixed stepsize η chosen among $2^{-i},\ i \in 1, \ldots, 15$ so that the final iterate distance is

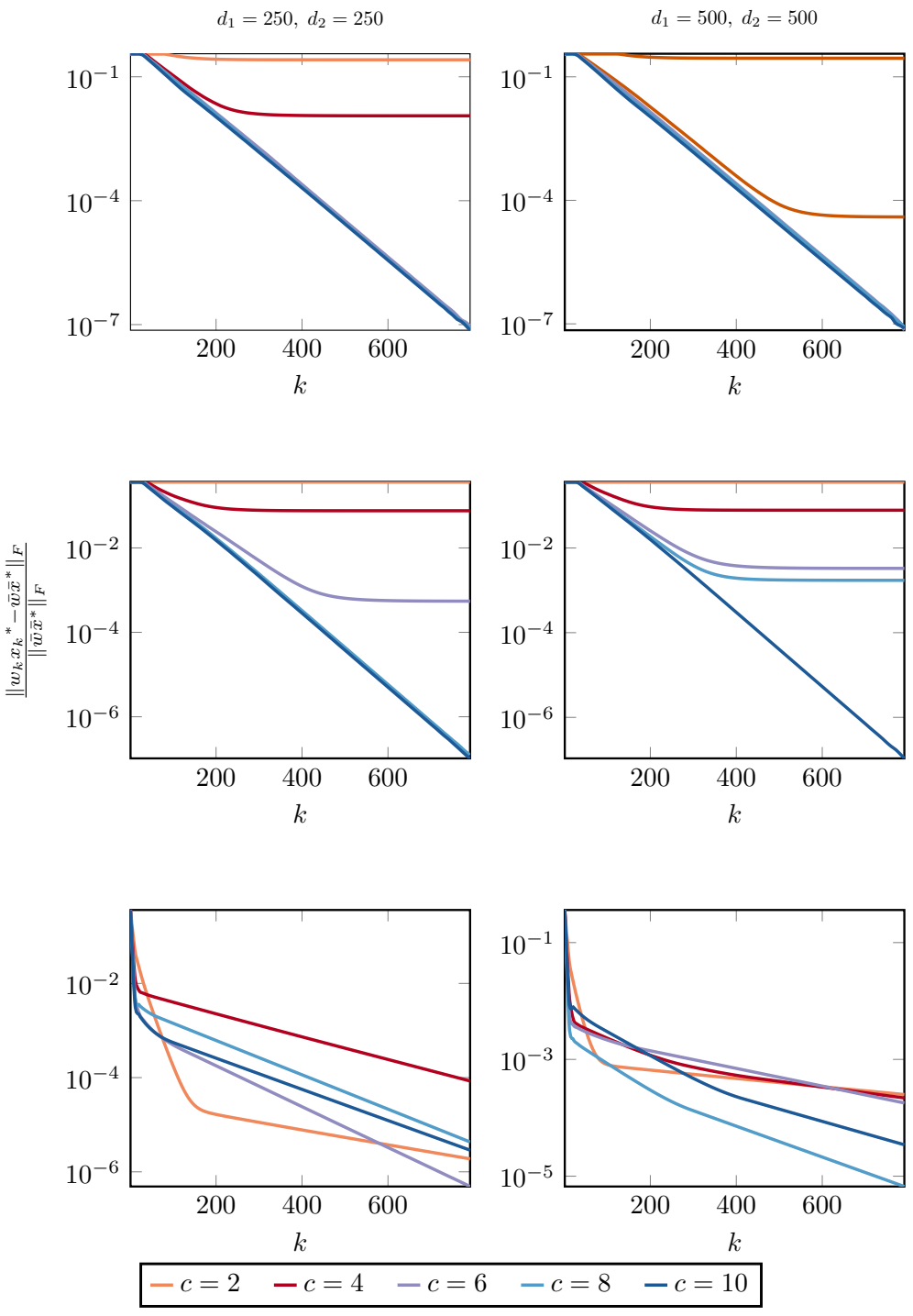


Figure 6.13: Convergence plot for synthetic instances in the complex domain, using $m = c \cdot d$. Left: $d = 250$. Right: $d = 500$. Top 2 rows: Algorithm 3 with $p_{\text{fail}} \in \{0.25, 0.45\}$. Bottom row: Algorithm 2 with $p_{\text{fail}} = 0$.

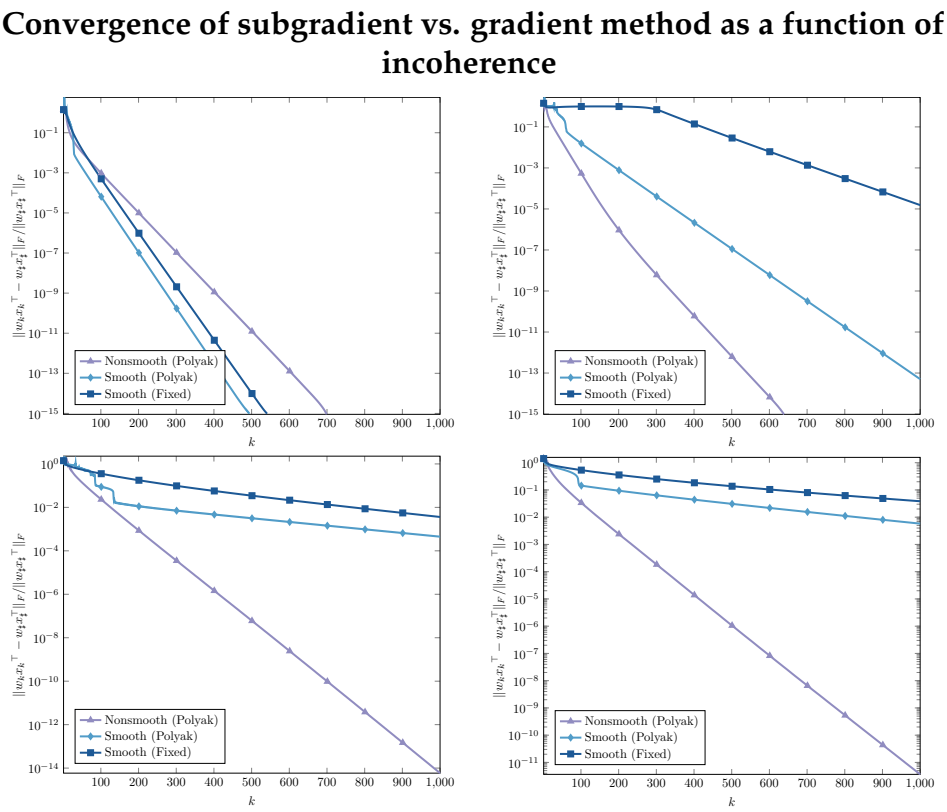


Figure 6.14: Convergence behavior of Algorithm 2 for minimizing (6.2) vs. gradient descent for minimizing its smooth counterpart for $d = 100$ and incoherences $\mu_h^2 \in \{12, 23, 89, 100\}$ (clockwise, starting from top left).

minimized.

Figure 6.14 illustrates that the nonsmooth objective is much more robust to variations on the incoherence of μ_h^2. As additional empirical evidence, Figure 6.15 shows the average ± one standard deviation of the number of iterations required to reach normalized distance 10^{-5} for the two formulations, minimized using the Polyak stepsize. Perhaps surprisingly, the nonsmooth version remains practically constant over all choices of μ_h^2.

Finally, we generate a few transition plots for $d \in \{100, 200\}$ that illustrate the effects of incoherence on the nonsmooth and smooth flavors of the recovery objective. Following the setting of [153], we choose 10 equispaced values for

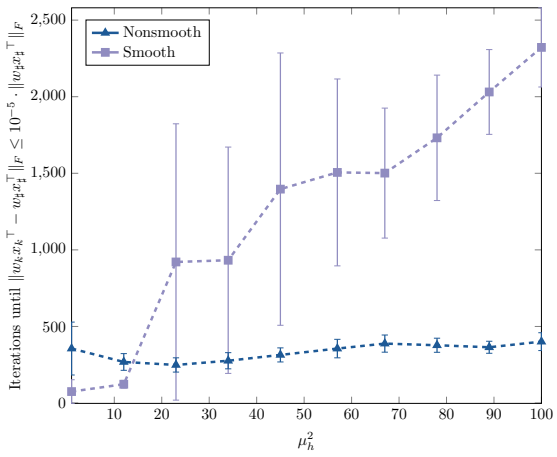


Figure 6.15: Average number of iterations to reach normalized distance 10^{-5} for Algorithm 2 applied to (6.2) vs. gradient descent with Polyak stepsize applied to objective (6.25). Dashed lines are average over 25 independent realizations with error bars indicating one standard deviation.

$\mu_h^2 \in [1, d]$ and plot the empirical probability of recovery over 50 independent runs for various ratios $\frac{m}{2d}$. We consider the result of a run successful if it satisfies $\frac{\left\|w_t x_t^\top - w_\sharp x_\sharp^\top\right\|_F}{\left\|w_\sharp x_\sharp^\top\right\|_F} \leq 10^{-5}$ after at most 1000 iterations. Figure 6.16 shows that the nonsmooth objective is far more robust to signal incoherence, but it also reveals that it is not entirely unaffected by it; in particular, we can see that we need a higher threshold $\frac{m}{2d}$ to recover signals with higher incoherence after fixing the dimension d.

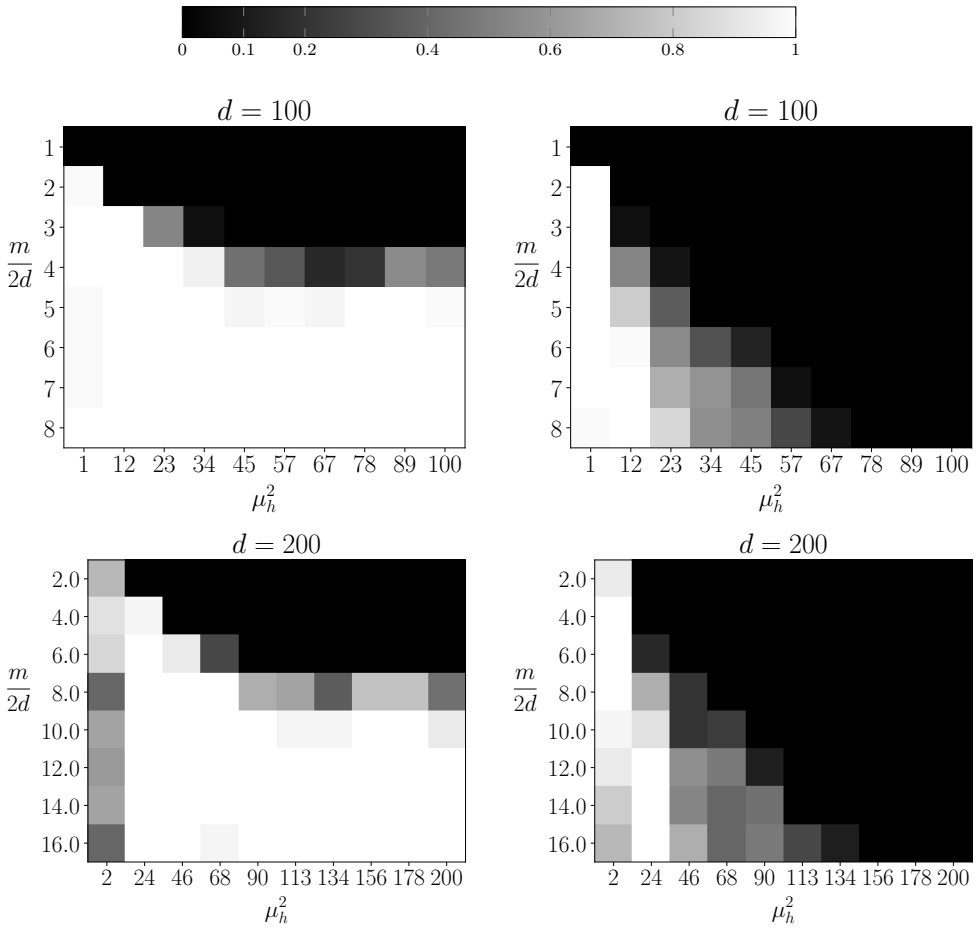


Figure 6.16: Empirical recovery probabilities for various values of $\left(\frac{m}{2d}, \mu_h^2\right)$ over 50 independent trials. Lighter cells imply higher recovery probability. **Left**: Algorithm 2. **Right**: gradient descent with Polyak stepsize minimizing (6.25).

6.6 Analysis

6.6.1 Proofs in Section 6.2

Proof of Proposition 6.2.1

Without loss of generality, we assume that $\Phi = 1$ (by rescaling) and that $\bar{w} = e_1 \in \mathbf{R}^{d_1}$ and $\bar{x} = e_1 \in \mathbf{R}^{d_2}$ (by rotation invariance). Recall that the distance to $\mathcal{S}_\nu^*$ may be written succinctly as

$$\operatorname{dist}((w,x),\mathcal{S}_\nu^*) = \sqrt{\inf_{(1/\nu)\le|\alpha|\le\nu}\left\{\|w-\alpha\bar{w}\|_2^2 + \|x-(1/\alpha)\bar{x}\,\|_2^2\right\}}.$$

Before we establish the general result, we first consider the simpler case, $d_1 = d_2 = 1$.

Claim 3. *The following bound holds:*

$$|wx-1| \ge \frac{1}{\sqrt{2}} \cdot \sqrt{\inf_{(1/\nu)\le|\alpha|\le\nu}\{|w-\alpha|^2 + |x-(1/\alpha)|^2\}},$$

for all $w, x \in [-\nu, \nu]$.

Proof of Claim. Consider a pair $(w,x) \in \mathbf{R}^2$ with $|w|, |x| \le \nu$. It is easy to see that without loss of generality, we may assume $w \ge |x|$. We then separate the proof into two cases, which are graphically depicted in Figure 6.17.

Case 1: $w - x \le \frac{\nu^2-1}{\nu}$**.** In this case, we will traverse from (w,x) to the $\mathcal{S}_\nu^*$ in the direction $(1,1)$. See Figure 6.17. First, consider the equation

$$wx - \sqrt{2}(w+x)t + t^2/2 = 1,$$

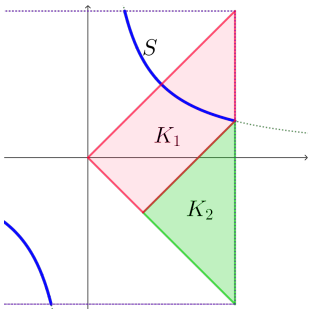


Figure 6.17: The regions K_1, K_2 correspond to cases 1 and 2 of the proof of Claim 3, respectively.

in the variable t and note the equality

$$wx - \sqrt{2}(w+x)t + t^2/2 = (w - t/\sqrt{2})(x - t/\sqrt{2}).$$

Using the quadratic formula to solve for t, we get

$$t = \sqrt{2}(w+x) - \sqrt{2(w+x)^2 - 2(wx-1)}.$$

Note that the discriminant is nonnegative since

$$(w+x)^2 - (wx-1) = w^2 + x^2 + xw + 1 \geq 1.$$

Set $\alpha = (w - t/\sqrt{2})$ and note the identity $1/\alpha = (x - t/\sqrt{2})$. Therefore,

$$\begin{aligned} |wx-1| &= |(1/\alpha)(w-\alpha) + \alpha(x - 1/\alpha) + (w-\alpha)(x-1/\alpha)| \\ &= |(x - t/\sqrt{2})(t/\sqrt{2}) + (w - t/\sqrt{2})(t/\sqrt{2}) + t^2/2| \\ &= \frac{|t|}{\sqrt{2}}|(w+x) - t/\sqrt{2}| = \frac{|t|}{2}\sqrt{2(w+x)^2 - 2(wx-1)} \geq \frac{|t|}{\sqrt{2}}. \end{aligned}$$

Observe now the equality

$$\frac{|t|}{\sqrt{2}} = \frac{1}{\sqrt{2}} \cdot (|w-\alpha|^2 + |x - 1/\alpha|^2)^{1/2}.$$

Hence it remains to bound α. First we note that $\alpha \geq 0, 1/\alpha \geq 0$, since

$$\begin{aligned} \alpha + 1/\alpha &= (w - t/\sqrt{2}) + (x - t/\sqrt{2}) \\ &= -(w+x) + 2\sqrt{(w+x)^2 - (wx-1)} \geq 0. \end{aligned}$$

In addition, since $w \geq x$, we have $\alpha = w - t/\sqrt{2} \geq x - t/\sqrt{2} = 1/\alpha$. Since α and $1/\alpha$ are positive, we must therefore have $\alpha \geq 1 \geq 1/\nu$. Thus, it remains to verify the bound $\alpha \leq \nu$. To that end, notice that

$$1/\alpha = x - t/\sqrt{2} \geq w - t/\sqrt{2} - \frac{\nu^2 - 1}{\nu} = \alpha - \frac{\nu^2 - 1}{\nu}.$$

Therefore, $\frac{\nu^2-1}{\nu} \geq \frac{\alpha^2-1}{\alpha}$. Since the function $t \mapsto \frac{t^2-1}{t}$ is increasing, we deduce $\alpha \leq \nu$.

Case 2: $w - x \geq \frac{\nu^2-1}{\nu}$**.** In this case, we will simply set $\alpha = \nu$. Define

$$t = \left((w-\nu)^2 + (x - 1/\nu)^2\right)^{1/2}, \qquad a = \frac{w-\nu}{t}, \qquad \text{and} \qquad b = \frac{x - 1/\nu}{t}.$$

Notice that proving the desired bound amounts to showing $|wx - 1| \geq \frac{t}{\sqrt{2}}$. Observe the following estimates

$$a, b \leq 0, \qquad b \leq a, \qquad a^2 + b^2 = 1, \qquad \text{and} \qquad t \leq -\frac{1}{(a+b)}\left(\frac{\nu^2 + \nu}{\nu}\right),$$

where the the first inequality follows from the bounds $w \leq \nu$ and $\nu \geq w \geq x + \nu - 1/\nu$, second inequality follows from the bound $w - x \geq (\nu^2 - 1)/\nu$, the equality follows from algebraic manipulations, and the third inequality follows from the estimate $w + x \geq 0$. Observe

$$|wx - 1| = |(\nu + ta)(1/\nu + tb) - 1| = |t^2 ab + t\nu b + ta/\nu|.$$

Thus, by dividing through by t, we need only show that

$$|tab + \nu b + a/\nu| \geq \frac{1}{\sqrt{2}}. \tag{6.26}$$

To prove this bound, note that since $2b^2 \geq a^2 + b^2 = 1$, we have the $-\nu b - a/\nu \geq -\nu b \geq 1/\sqrt{2}$. Therefore, in the particular case when $ab = 0$ the estimate 6.26 follows immediately. Define the linear function $p(s) := -(ab)s - \nu b - a/\nu$. Hence,

assume $ab \neq 0$. Notice $p(0) \geq 1/\sqrt{2}$. Thus it suffices to show that the solution s^* of the equation $p(s) = 1/\sqrt{2}$ satisfies $s^* \geq t$. To see this, we compute:

$$\begin{aligned}
s^* &= -\frac{1}{ab}\left(\nu b + a/\nu + \frac{1}{\sqrt{2}}\right) \\
&= -\frac{1}{(a+b)}(a+b)\left(\frac{\nu}{a} + \frac{1}{b\nu} + \frac{1}{\sqrt{2}ab}\right) \\
&= -\frac{1}{(a+b)}\left(\nu\left(1+\frac{b}{a}\right) + \frac{1}{\nu}\left(1+\frac{a}{b}\right) + \frac{1}{\sqrt{2}}\left(\frac{1}{a}+\frac{1}{b}\right)\right) \\
&\geq -\frac{1}{(a+b)}\left(\nu + \frac{1}{\nu}\left(1+\frac{a}{b}+\frac{b}{a}+\frac{1}{\sqrt{2}b}+\frac{1}{\sqrt{2}a}\right)\right) \\
&= -\frac{1}{(a+b)}\left(\nu + \frac{1}{\nu}\left(1+\frac{\sqrt{2}\left(a^2+b^2\right)-(|a|+|b|)}{\sqrt{2}ab}\right)\right) \\
&= -\frac{1}{(a+b)}\left(\nu + \frac{1}{\nu}\left(1+\frac{\sqrt{2}-(|a|+|b|)}{\sqrt{2}ab}\right)\right) \\
&\geq -\frac{1}{(a+b)}\left(\nu + \frac{1}{\nu}\right) \geq t,
\end{aligned}$$

where the first inequality follows since $\nu \geq 1$ and the second inequality follows since $a^2 + b^2 = 1$ and $\sqrt{2}\|(a,b)\|_2 \geq \|(a,b)\|_1$, as desired. □

Now we prove the general case. First suppose that $\|wx^\top - \bar{w}\bar{x}^\top\|_F \geq 1/2$. Since $\|w - \bar{w}\|_2 \leq (\nu+1)$ and $\|x - \bar{x}\|_2 \leq (\nu+1)$, we have

$$\operatorname{dist}((w,x), \mathcal{S}_\nu^*) \leq \sqrt{2}(\nu+1) \leq 2\sqrt{2}(\nu+1)\|wx^\top - \bar{w}\bar{x}^\top\|_F,$$

which proves the desired bound.

On the other hand, suppose that $\|wx^\top - \bar{w}\bar{x}^\top\|_F < 1/2$. Define the two vectors:

$$\tilde{w} = (w_1, 0, \ldots, 0)^\top \in \mathbf{R}^{d_1} \qquad \text{and} \qquad \tilde{x} = (x_1, 0, \ldots, 0)^\top \in \mathbf{R}^{d_2}.$$

With this notation, we find that by Claim 3, there exists an α satisfying $(1/\nu) \leq$

$|\alpha| \leq \nu$, such that the following holds:

$$\begin{aligned}
\|wx^\top - \bar{w}\bar{x}^\top\|_F^2 &= \|wx^\top - \tilde{w}\tilde{x}^\top + \tilde{w}\tilde{x}^\top - \bar{w}\bar{x}^\top\|_F^2 \\
&= \|wx^\top - \tilde{w}\tilde{x}^\top\|_F^2 + \|\tilde{w}\tilde{x}^\top - \bar{w}\bar{x}^\top\|_F^2 \\
&\geq \|wx^\top - \tilde{w}\tilde{x}^\top\|_F^2 + \frac{1}{2}\left(\|\tilde{w} - \alpha\bar{w}\|_F^2 + \|\tilde{x} - (1/\alpha)\bar{x}\|_F^2\right).
\end{aligned}$$

We now turn our attention to lower bounding the first term. Observe since $|w_1 x_1 - \bar{w}_1\bar{x}_1| \leq \|wx^T - \bar{w}\bar{x}^T\|_F < 1/2$, we have

$$|w_1 x_1| \geq |\bar{w}_1\bar{x}_1| - |w_1 x_1 - \bar{w}_1\bar{x}_1| \geq (1/2)|\bar{w}_1\bar{x}_1| = 1/2,$$

Moreover, note the estimates, $\nu|w_1| \geq |x_1||w_1| \geq 1/2$ and $\nu|x_1| \geq |x_1||w_1| \geq 1/2$, which imply that $|w_1| \geq 1/2\nu$ and $|x_1| \geq 1/2\nu$. Thus, we obtain the lower bound

$$\begin{aligned}
\|wx^\top - \tilde{w}\tilde{x}^\top\|_F^2 &= \|(w - \tilde{w})\tilde{x}^\top + \tilde{w}(x - \tilde{x})^\top + (w - \tilde{w})(x - \tilde{x})^\top\|_F^2 \\
&= |x_1|^2\|w - \tilde{w}\|^2 + |w_1|^2\|x - \tilde{x}\|_2^2 + \|(w - \tilde{w})(x - \tilde{x})^\top\|_F^2 \\
&\geq |x_1|^2\|w - \tilde{w}\|_2^2 + |w_1|^2\|x - \tilde{x}\|_2^2 \\
&\geq \left(\frac{1}{2\nu}\right)^2\left(\|w - \tilde{w}\|_2^2 + \|x - \tilde{x}\|_2^2\right).
\end{aligned}$$

Finally, we obtain the bound

$$\begin{aligned}
\|wx^\top - \bar{w}\bar{x}^\top\|_F^2 &\geq \|wx^\top - \tilde{w}\tilde{x}^\top\|_F^2 + \frac{1}{2}\left(\|\tilde{w} - \alpha\bar{w}\|_F^2 + \|\tilde{x} - (1/\alpha)\bar{x}\|_F^2\right) \\
&\geq \left(\frac{1}{2\nu}\right)^2\left(\|w - \tilde{w}\|_2^2 + \|x - \tilde{x}\|_2^2\right) + \frac{1}{2}\left(\|\tilde{w} - \alpha\bar{w}\|_2^2 + \|\tilde{x} - (1/\alpha)\bar{x}\|_2^2\right) \\
&\geq \min\left\{\frac{1}{2}, \left(\frac{1}{2\nu}\right)^2\right\}\left(\|w - \tilde{w}\|_2^2 + \|x - \tilde{x}\|_2^2 + \|\tilde{w} - \alpha\bar{w}\|_2^2 + \|\tilde{x} - (1/\alpha)\bar{x}\|_2^2\right) \\
&= \left(\frac{1}{2\nu}\right)^2 \cdot \mathrm{dist}^2((w, x), \mathcal{S}_\nu^*).
\end{aligned}$$

By recalling that $1/2\nu \geq 1/2\sqrt{2}(\nu + 1)$, the proof is complete.

6.6.2 Proofs in Section 6.3

Proof of Proposition 6.3.1

As stated in Section 6.3, we first verify that L^{init} and R^{init} are nearby matrices with minimal eigenvectors equal to $\bar{w}_\star$ and $\bar{x}_\star$. Then we apply the Davis-Kahan $\sin\theta$ theorem [65] to prove that the minimal eigenvectors of L^{init} and R^{init} must also be close to the optimal directions.

Throughout the rest of the proof, we define the sets of "selected" inliers and outliers:

$$\mathcal{I}_{\text{in}}^{\text{sel}} = \mathcal{I}_{\text{in}} \cap \mathcal{I}^{\text{sel}} \qquad \text{and} \qquad \mathcal{I}_{\text{out}}^{\text{sel}} = \mathcal{I}_{\text{out}} \cap \mathcal{I}^{\text{sel}}.$$

We record the relative size of these parameters as well, since they appear in the bounds that follow:

$$S_{\text{in}} := \frac{1}{m}\left|\mathcal{I}_{\text{in}}^{\text{sel}}\right| \qquad \text{and} \qquad S_{\text{out}} = \frac{1}{m}\left|\mathcal{I}_{\text{out}}^{\text{sel}}\right|.$$

Theorem 6.6.1. *There exist numerical constants $c_1, c_2, c_3, c_4, c_5 > 0$, so that for any $p_{\text{fail}} \in [0, 1/10]$ and $t \in [0, 1]$, with probability at least $1 - c_1(\exp(-c_2 mt)$ the following hold:*

1. *Under noise model* ***N1***

$$L^{\text{init}} = (S_{\text{in}} + S_{\text{out}})I_{d_1} - \gamma_1 \bar{w}_\star \bar{w}_\star^\top + \Delta_1, \quad R^{\text{init}} = (S_{\text{in}} + S_{\text{out}})I_{d_2} - \gamma_2 \bar{x}_\star \bar{x}_\star^\top + \Delta_2,$$

   where $\gamma_1 \geq c_3$ and $\gamma_2 \geq c_4$ and

$$\max\{\|\Delta_1\|_{\text{op}}, \|\Delta_2\|_{\text{op}}\} \leq c_5\left(\sqrt{\frac{\max\{d_1, d_2\}}{m}} + t\right).$$

2. *Under noise model* ***N2***

$$L^{\text{init}} = S_{\text{in}} I_{d_1} - \gamma_1 \bar{w}_\star \bar{w}_\star^\top + \Delta_1, \quad R^{\text{init}} = S_{\text{in}} I_{d_2} - \gamma_2 \bar{x}_\star \bar{x}_\star^\top + \Delta_2,$$

where $\gamma_1 \geq c_3$ and $\gamma_2 \geq c_4$ and

$$\max\{\|\Delta_1\|_{\mathrm{op}}, \|\Delta_2\|_{\mathrm{op}}\} \leq p_{\mathrm{fail}} + c_5\left(\sqrt{\frac{\max\{d_1, d_2\}}{m}} + t\right).$$

Proof. Without loss of generality, we only prove the result for L^{init}; the result for R^{init} follows by a symmetric argument.

Define the projection operators $P_{\bar{w}_\star} := \bar{w}_\star \bar{w}_\star^\top$ and let $P_{\bar{w}_\star}^\perp := I - \bar{w}_\star \bar{w}_\star^\top$. Then decompose L^{init} into the sums of four matrices Y_0, Y_1, Y_2, Y_3, as follows:

$$L^{\mathrm{init}} = \frac{1}{m}\Bigg(\underbrace{\sum_{i \in \mathcal{I}_{\mathrm{in}}^{\mathrm{sel}}} P_{\bar{w}_\star} \ell_i \ell_i^\top P_{\bar{w}_\star}}_{m \cdot Y_0} + \underbrace{\sum_{i \in \mathcal{I}_{\mathrm{in}}^{\mathrm{sel}}} \left(P_{\bar{w}_\star} \ell_i \ell_i^\top P_{\bar{w}_\star}^\perp + P_{\bar{w}_\star}^\perp \ell_i \ell_i^\top P_{\bar{w}_\star} \right)}_{m \cdot Y_1} + \underbrace{\sum_{i \in \mathcal{I}_{\mathrm{in}}^{\mathrm{sel}}} P_{\bar{w}_\star}^\perp \ell_i \ell_i^\top P_{\bar{w}_\star}^\perp}_{m \cdot Y_2} + \underbrace{\sum_{i \in \mathcal{I}_{\mathrm{out}}^{\mathrm{sel}}} \ell_i \ell_i^\top}_{m \cdot Y_3} \Bigg). \tag{6.27}$$

We will now study the properties of these four matrices under both noise models.

First, note that in either case we may write $Y_0 = y_0 \bar{w}_\star \bar{w}_\star^\top$, where

$$y_0 := \frac{1}{m} \sum_{i \in \mathcal{I}_{\mathrm{in}}^{\mathrm{sel}}} (\ell_i^\top \bar{w}_\star)^2.$$

In addition, we will present a series of Lemmas showing the following high probability deviation bounds:

$$\gamma_1 := S_{\mathrm{in}} - y_0 \gtrsim 1, \quad \|Y_1\|_{\mathrm{op}} \lesssim \sqrt{\frac{d_1}{m}}, \quad \text{and} \quad \|Y_2 - S_{\mathrm{in}}(I_{d_1} - \bar{w}_\star \bar{w}_\star^\top)\|_{\mathrm{op}} \lesssim \sqrt{\frac{d_1}{m}}.$$

Finally, our bounds on the term Y_3 as well as the definition of Δ_1 depend on the noise model under consideration. Thus, we separate this bound into two cases:

Noise model N1. Under this noise model, we have

$$\|Y_3 - S_{\mathrm{out}} I_{d_1}\|_{\mathrm{op}} \lesssim \sqrt{\frac{d_1}{m}}.$$

Thus, we set $\Delta_1 = Y_1 + (Y_2 - S_{\mathrm{in}}(I_{d_1} - \bar{w}_\star \bar{w}_\star^\top)) + (Y_3 - S_{\mathrm{out}} I_{d_1})$.

Noise model N2. Under this noise model, we have

$$\|Y_3\|_{\mathrm{op}} \lesssim p_{\mathrm{fail}} + \sqrt{\frac{d_1}{m}}.$$

Thus, we set $\Delta_1 = Y_1 + (Y_2 - S_{\mathrm{in}}(I_{d_1} - \bar{w}_\star \bar{w}_\star^\top)) + Y_3$.

Therefore, under either noise model, the result will follow immediately from the following four Lemmas. We defer the proofs for the moment.

Lemma 6.6.2. *There exist constants $c, c_1, c_2 > 0$ such that for any $p_{\mathrm{fail}} \in [0, 1/10]$ the following holds:*

$$\mathbb{P}\left(S_{\mathrm{in}} - y_0 \geq c\right) \geq 1 - c_1 \exp\left(-c_2 m\right).$$

Lemma 6.6.3. *For $t \geq 0$, we have*

$$\mathbb{P}\left(\|Y_1\|_{\mathrm{op}} \geq 2\sqrt{\frac{d_1 - 1}{m}} + t\right) \leq \exp\left(-\frac{mt^2}{8}\right) + \exp\left(-\frac{m}{2}\right).$$

Lemma 6.6.4. *There exist numerical constants $C, c > 0$ such that for any $t > 0$ we have*

$$\mathbb{P}\left(\left\|Y_2 - S_{\mathrm{in}}(I_{d_1} - \bar{w}_\star \bar{w}_\star^\top)\right\|_{\mathrm{op}} \geq C\sqrt{\frac{d_1}{m} + t}\right) \leq 2\exp(-cmt).$$

Lemma 6.6.5. *There exist constants $C_1, C_2, c_1, c_2 > 0$ such that for any $t > 0$ the following hold. Under the noise Model **N1**, we have the estimate*

$$\mathbb{P}\left(\left\|Y_3 - S_{\mathrm{out}} I_{d_1}\right\|_{\mathrm{op}} \geq c_3\sqrt{\frac{d_1}{m} + t}\right) \leq 2\exp(-c_4 mt),$$

*while under the noise model **N2** we have*

$$\mathbb{P}\left(\|Y_3\|_{\mathrm{op}} \geq p_{\mathrm{fail}} + c_1\sqrt{\frac{d_1}{m} + t}\right) \leq 2\exp(-c_2 mt).$$

The proof of the the theorem is complete. □

We now apply the Davis-Kahan $\sin\theta$ theorem [65] as stated in Lemma 6.6.15. Throughout we assume that we are in the event described in 6.6.1.

Proof of Proposition 6.3.1. We will use the notation from Theorem 6.6.1. We only prove the result under **N1**, since the proof under **N2** is completely analogous. Define matrices $V_1 = \gamma_1 \bar{w}_\star \bar{w}_\star^\top - (S_{\text{in}} + S_{\text{out}}) I_{d_1}$ and $V_2 = \gamma_2 \bar{x}_\star \bar{x}_\star^\top - (S_{\text{in}} + S_{\text{out}}) I_{d_2}$. Matrix V_1 has spectral gap γ_1 and top eigenvector $\bar{w}_\star$, while matrix V_2 has spectral gap γ_2 and top eigenvector $\bar{x}_\star$. Therefore, since $-L^{\text{init}} = V_1 - \Delta_1$ and $-R^{\text{init}} = V_2 - \Delta_2$, Lemma 6.6.15 implies that

$$\min_{s \in \{\pm 1\}} \|\widehat{w} - s\bar{w}_\star\|_2 \le \frac{\sqrt{2}\,\|\Delta_1\|_{\text{op}}}{\gamma_1} \qquad \text{and} \qquad \min_{s \in \{\pm 1\}} \|\widehat{x} - s\bar{x}_\star\|_2 \le \frac{\sqrt{2}\,\|\Delta_2\|_{\text{op}}}{\gamma_2}.$$

We will use these two inequalities to bound $\min_{s \in \{\pm 1\}} \|\widehat{w}\widehat{x}^\top - s\bar{w}_\star \bar{x}_\star^\top\|_F$. To do so, we need to analyze $s_1 = \arg\min_{s \in \{\pm 1\}} \|\widehat{w} - s\bar{w}_\star\|$ and $s_2 = \arg\min_{s \in \{\pm 1\}} \|\widehat{x} - s\bar{x}_\star\|$. We split the argument into two cases.

Suppose first $s_1 = s_2$. Then

$$\begin{aligned}\|\widehat{w}\widehat{x}^\top - \bar{w}_\star \bar{x}_\star^\top\|_F = \|\widehat{w}(\widehat{x} - s_2\bar{x}_\star)^\top - (\bar{w}_\star - s_1\widehat{w})\bar{x}_\star^\top\|_F &\le \|\widehat{x} - s_2\bar{x}_\star\|_2 + \|\bar{w}_\star - s_1\widehat{w}\|_2 \\ &\le \frac{2\sqrt{2}\max\{\|\Delta_1\|_{\text{op}}, \|\Delta_2\|_{\text{op}}\}}{\min\{\gamma_1, \gamma_2\}},\end{aligned}$$

as desired.

Suppose instead $s_1 = -s_2$. Then

$$\begin{aligned}\|\widehat{w}\widehat{x}^\top + \bar{w}_\star \bar{x}_\star^\top\|_F = \|\widehat{w}(\widehat{x} - s_2\bar{x}_\star)^\top + (\bar{w}_\star + s_2\widehat{w})\bar{x}_\star^\top\|_F &\le \|\widehat{x} - s_2\bar{x}_\star\|_2 + \|\bar{w}_\star - s_1\widehat{w}\|_2 \\ &\le \frac{2\sqrt{2}\max\{\|\Delta_1\|_{\text{op}}, \|\Delta_2\|_{\text{op}}\}}{\min\{\gamma_1, \gamma_2\}},\end{aligned}$$

as desired. Bounding $\max\{\|\Delta_1\|_{\text{op}}, \|\Delta_2\|_{\text{op}}\}$ using Theorem 6.6.1 completes the proof.

□

The next sections present the proof of Lemmas 6.6.2-6.6.5. We next set up the

notation. For any sequence of vectors $\{w_i\}_{i=1}^m$ in $\mathbf{R}^d$, we will use the symbol $w_{i,2:d}$ to denote the vector in $\mathbf{R}^{d-1}$ consisting of the last $d-1$ coordinates of w_i.

We will use the following two observations throughout. First, by rotation invariance we will assume, without loss of generality, that $\bar{w}_\star = e_1$ and $\bar{x}_\star = e_1$. Second, and crucially, this assumption implies that $\mathcal{I}_{\text{in}}^{\text{sel}}$ depends on $\{\ell_i\}_{i=1}^m$ only through the first component. In particular, we have that $\{\ell_{i,2:d_1}\}_{i=1}^m$ and $\mathcal{I}_{\text{in}}^{\text{sel}}$ are independent. Similarly, $\{r_{i,2:d_2}\}_{i=1}^m$ and $\mathcal{I}_{\text{in}}^{\text{sel}}$ are independent as well.

Proof of Lemma 6.6.2

Our goal is to lower bound the quantity

$$S_{\text{in}} - y_0 = \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} (1 - \ell_{i,1}^2).$$

To prove a lower bound, we need to control the random variables $\ell_{i,1}^2$ on the set $\mathcal{I}_{\text{in}}^{\text{sel}}$. Before proving the key claim, we first introduce some notation. First, define

$$q_{\text{fail}} := \frac{5 - 2p_{\text{fail}}}{8(1 - p_{\text{fail}})},$$

which is strictly less than one since $p_{\text{fail}} < 1/2$. Let $a, b \sim \text{N}(0,1)$ and define Q_{fail} to be the q_{fail}-quantile of the random variable $|ab|$. In particular, the following relationship holds

$$q_{\text{fail}} = \mathbb{P}\left(|ab| \leq Q_{\text{fail}}\right).$$

Additionally, define the conditional expected value

$$\omega_{\text{fail}} = \mathbb{E}\left[a^2 \mid |ab| \leq Q_{\text{fail}}\right].$$

Rather than analyzing $\mathcal{I}_{\text{in}}^{\text{sel}}$ directly, we introduce the following set $\mathcal{I}_{\text{in}}^{Q}$, which is simpler to analyze:

$$\mathcal{I}_{\text{in}}^{Q} := \left\{i \in \mathcal{I}_{\text{in}} \mid \left|\ell_i^\top \bar{w}_\star \bar{x}_\star^\top r_i\right| \leq Q_{\text{fail}}\right\}.$$

Then we prove the following claim.

Claim 4. *There exist numerical constants $c, K > 0$ such that for all $t \geq 0$ the following inequalities hold true:*

1. $\frac{|\mathcal{I}_{\text{in}}^{\text{sel}}|}{m} \geq \frac{1-2p_{\text{fail}}}{2}$.
2. $\mathbb{P}\left(\mathcal{I}_{\text{in}}^{Q} \supseteq \mathcal{I}_{\text{in}}^{\text{sel}}\right) \geq 1 - \exp\left(-\frac{3(1-2p_{\text{fail}})}{160}m\right)$,
3. $\mathbb{P}\left(|\mathcal{I}_{\text{in}}^{Q}| \geq \frac{6251m}{10000}\right) \leq \exp\left(-\frac{m}{2\cdot 10^8}\right)$,
4. $\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^{Q}|}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}} \ell_{i,1}^2 \geq \omega_{\text{fail}} + t\right) \leq \exp\left(-c\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\frac{m(1-2p_{\text{fail}})}{2}\right) + \exp\left(-\frac{3(1-2p_{\text{fail}})}{160}m\right)$.

Before we prove the claim, we show it leads to the conclusion of the lemma. Assuming we are in the event

$$\mathcal{E} = \left\{\mathcal{I}_{\text{in}}^{Q} \supseteq \mathcal{I}_{\text{in}}^{\text{sel}},\ \ |\mathcal{I}_{\text{in}}^{Q}| < \frac{6251m}{10000},\ \ \frac{1}{|\mathcal{I}_{\text{in}}^{Q}|}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}} \ell_{i,1}^2 \leq \frac{101}{100}\omega_{\text{fail}}\right\},$$

it follows that

$$\begin{aligned} S - y_0 = \frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}}(1-\ell_{i,1}^2) \geq \frac{|\mathcal{I}_{\text{in}}^{\text{sel}}|}{m} - \frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}}\ell_{i,1}^2 &\geq \frac{1-2p_{\text{fail}}}{2} - \frac{|\mathcal{I}_{\text{in}}^{Q}|}{m\left|\mathcal{I}_{\text{in}}^{Q}\right|}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}}\ell_{i,1}^2 \\ &\geq \frac{1-2p_{\text{fail}}}{2} - \frac{631351}{1000000}\omega_{\text{fail}} \geq 0.04644344. \end{aligned}$$

where the first three inequalities follow by the definition of the event $\mathcal{E}$. The fourth inequality follows by the definition of $\mathcal{E}$ and Lemma 6.6.23, which implies $\omega_{\text{fail}} \leq .56$ when $p_{\text{fail}} = .1$ and that the difference is minimized over $p_{\text{fail}} \in [0, .1]$ at the endpoint $p_{\text{fail}} = .1$. To get the claimed probabilities, we note that by Lemma 6.6.23, we have $\omega_{\text{fail}} \geq .5$ for any setting of p_{fail}.

Now we prove the claim.

Proof of the Claim. We separate the proof into four parts.

Part 1. By definition, we have

$$\frac{|\mathcal{I}_{\text{in}}^{\text{sel}}|}{m} = \frac{|\mathcal{I}_{\text{in}} \cap \mathcal{I}^{\text{sel}}|}{m} = \frac{|\mathcal{I}^{\text{sel}}| - |\mathcal{I}_{\text{out}} \cap \mathcal{I}^{\text{sel}}|}{m} \geq \frac{\frac{m}{2} - |\mathcal{I}_{\text{out}} \cap \mathcal{I}^{\text{sel}}|}{m} \geq \frac{\frac{m}{2} - mp_{\text{fail}}}{m} = \frac{1 - 2p_{\text{fail}}}{2}.$$

Part 2. By the definitions of $\mathcal{I}_{\text{in}}^{\text{sel}}$ and $\mathcal{I}_{\text{in}}^{Q}$, the result will follow once we show that

$$\mathbb{P}\left(\texttt{med}(\{|y_i|\}_i^m) \geq Q_{\text{fail}}\Phi\right) \leq \exp\left(-\frac{3(1-2p_{\text{fail}})}{160}m\right).$$

To that end, first note that

$$\begin{aligned}
\texttt{med}(\{|y_i|\}_i^m) &= \min\left\{|b_j| : j \in [m], \sum_{i=1}^m \mathbf{1}\{|y_i| \leq |y_j|\} \geq \frac{m}{2}\right\} \\
&= \min\left\{|y_j| : j \in [m], \sum_{i=1}^m \mathbf{1}\{|b_i| \leq |y_j|\} \geq \frac{|\mathcal{I}_{\text{in}}|}{2(1-p_{\text{fail}})}\right\} \\
&\leq \min\left\{|y_j| : j \in \mathcal{I}_{\text{in}}, \sum_{i=1}^m \mathbf{1}\{|y_i| \leq |y_j|\} \geq \frac{|\mathcal{I}_{\text{in}}|}{2(1-p_{\text{fail}})}\right\} \\
&\leq \min\left\{|y_j| : j \in \mathcal{I}_{\text{in}}, \sum_{i \in \mathcal{I}_{\text{in}}} \mathbf{1}\{|y_i| \leq |y_j|\} \geq \frac{|\mathcal{I}_{\text{in}}|}{2(1-p_{\text{fail}})}\right\} \\
&= \texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|\}_{i \in \mathcal{I}_{\text{in}}}\right),
\end{aligned}$$

where the first equality follows since $\frac{|\mathcal{I}_{\text{in}}|}{2(1-p_{\text{fail}})} = \frac{(1-p_{\text{fail}})m}{2(1-p_{\text{fail}})} = m/2$, the first inequality follows since the minimum is taken over a smaller set, and the second inequality follows since the sum is taken over a smaller set of indices. Therefore, we find that

$$\begin{aligned}
\mathbb{P}\left(\texttt{med}(\{|y_i|\}_i^m) \geq Q_{\text{fail}}\Phi\right) &\leq \mathbb{P}\left(\texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|\}_{i \in \mathcal{I}_{\text{in}}}\right) \geq Q_{\text{fail}}\Phi\right) \\
&= \mathbb{P}\left(\texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|/\Phi\}_{i \in \mathcal{I}_{\text{in}}}\right) \geq Q_{\text{fail}}\right),
\end{aligned}$$

and our remaining task is to bound this probability.

To bound this probability, we apply Lemma 6.6.19 to the i.i.d. sample $\{|y_i|/\Phi\colon i \in \mathcal{I}_{\text{in}}\}$, which is sampled from the distribution of $\mathcal{D}$ of $|ab|$ where $a, b \sim \mathrm{N}(0,1)$ and a, b are independent. Therefore, using the identities (for $i \in \mathcal{I}_{\text{in}}$)

$$q = \mathbb{P}\left(|y_i|/\Phi \le Q_{\text{fail}}\right) = q_{\text{fail}} = \frac{5 - 2p_{\text{fail}}}{8(1 - p_{\text{fail}})}$$

and choosing $p := (2(1 - p_{\text{fail}}))^{-1} < q$, we find that

$$\begin{aligned}
\mathbb{P}\left(\texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|/\Phi\}_{i\in\mathcal{I}_{\text{in}}}\right) \ge Q_{\text{fail}}\right) &\le \exp\left(\frac{m(q-p)^2}{2(q-p)/3 + 2q(1-q)}\right) \\
&= \exp\left(\frac{m(q-p)}{2/3 + 6q}\right) \\
&= \exp\left(-\frac{3(1-2p_{\text{fail}})m}{8(1-p_{\text{fail}})(2+18q)}\right) \\
&\le \exp\left(-\frac{3(1-2p_{\text{fail}})}{160}m\right),
\end{aligned}$$

where we have used the identity $q - p = \frac{1-2p_{\text{fail}}}{8(1-p_{\text{fail}})} = (1-q)/3$ in the first equality. This completes the bound and implies that $\mathcal{I}_{\text{in}}^{Q} \supseteq \mathcal{I}_{\text{in}}^{\text{sel}}$ with high probability, as desired.

Part 3. Since $\{|y_i|/\Phi\colon i \in \mathcal{I}_{\text{in}}\}$ is an i.i.d. sample from the distribution of $|ab|$ where $a, b \sim \mathrm{N}(0,1)$ are independent, we have for each $i \in \mathcal{I}_{\text{in}}$, that

$$\mathbb{P}\left(i \in \mathcal{I}_{\text{in}}^{Q}\right) = \mathbb{P}\left(|y_i|/\Phi \le Q_{\text{fail}}\right) = \mathbb{P}\left(|ab| \le Q_{\text{fail}}\right) = q_{\text{fail}}.$$

Therefore, $\mathbb{E}\left[|\mathcal{I}_{\text{in}}^{Q}|\right] = q_{\text{fail}}|\mathcal{I}_{\text{in}}| \le \frac{5-2p_{\text{fail}}}{8(1-p_{\text{fail}})}(1 - p_{\text{fail}})m \le \frac{5}{8}m$. Finally, we apply Hoeffding's inequality (Lemma 2.3.1) to the i.i.d. Bernoulli random variables $\mathbf{1}\{i \in \mathcal{I}_{\text{in}}^{q}\} - \mathbb{E}\left[\mathbf{1}\{i \in \mathcal{I}_{\text{in}}^{q}\}\right]$ $(i \in \mathcal{I}_{\text{in}})$ to deduce that

$$\begin{aligned}
\mathbb{P}\left(\frac{6251m}{10000} \le |\mathcal{I}_{\text{in}}^{Q}|\right) = \mathbb{P}\left(\frac{m}{10000} \le |\mathcal{I}_{\text{in}}^{Q}| - \frac{5m}{8}\right) &\le \mathbb{P}\left(\frac{m}{10000} \le |\mathcal{I}_{\text{in}}^{Q}| - \mathbb{E}|\mathcal{I}_{\text{in}}^{Q}|\right) \\
&\le \exp\left(-\frac{(1/10000)^2 m}{2(1-p_{\text{fail}})}\right) \le \exp\left(-\frac{m}{2\cdot 10^8}\right),
\end{aligned}$$

as desired.

Part 4. First write

$$\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^{Q}|}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}}\ell_{i,1}^{2}\geq\omega_{\text{fail}}+t\right)$$
$$=\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^{Q}|}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}}\ell_{i,1}^{2}\geq\omega_{\text{fail}}+t\text{ and }|\mathcal{I}_{\text{in}}^{Q}|\supseteq\mathcal{I}_{\text{in}}^{\text{sel}}\right)+\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^{Q}|}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}}\ell_{i,1}^{2}\geq\omega_{\text{fail}}+t\text{ and }\mathcal{I}_{\text{in}}^{Q}\not\supseteq\mathcal{I}_{\text{in}}^{\text{sel}}\right)$$
$$\leq\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^{Q}|}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}}\ell_{i,1}^{2}\geq\omega_{\text{fail}}+t\text{ and }|\mathcal{I}_{\text{in}}^{Q}|\geq\frac{m(1-2p_{\text{fail}})}{2}\right)+\exp\left(-\tfrac{3(1-2p_{\text{fail}})}{160}m\right),$$

where first inequality follows from Part 2 and the bound $\frac{|\mathcal{I}_{\text{in}}^{\text{sel}}|}{m}\geq\frac{1-2p_{\text{fail}}}{2}$. Thus, we focus on bounding the first term.

To that end, notice that

$$\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^{Q}|}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}}\ell_{i,1}^{2}\geq\omega_{\text{fail}}+t\text{ and }|\mathcal{I}_{\text{in}}^{Q}|\geq\frac{m(1-2p_{\text{fail}})}{2}\right)$$
$$=\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^{Q}|}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}}\ell_{i,1}^{2}\geq\omega_{\text{fail}}+t\middle||\mathcal{I}_{\text{in}}^{Q}|\geq\frac{m(1-2p_{\text{fail}})}{2}\right)\mathbb{P}\left(|\mathcal{I}_{\text{in}}^{Q}|\geq\frac{m(1-2p_{\text{fail}})}{2}\right).$$

Observe that for any index $i\in\mathcal{I}_{\text{in}}$ and $t\geq 0$, we have $\mathbb{P}\left(\ell_{i,1}^{2}\geq t\mid i\in\mathcal{I}_{\text{in}}^{Q}\right)=\mathbb{P}\left(a^{2}\geq t\mid|ab|\leq Q_{\text{fail}}\right)$, where $a,b\sim\mathrm{N}(0,1)$ are independent. In addition, we have $q_{\text{fail}}=P(|ab|\leq Q_{\text{fail}})=\frac{5-2p_{\text{fail}}}{8(1-p_{\text{fail}})}\geq 5/8>1/2$, where we have used the fact that q_{fail} is an increasing function of p_{fail}. Therefore, applying Lemma 6.6.20, we have the following bound:

$$\mathbb{P}\left(\ell_{i,1}^{2}\geq t\mid i\in\mathcal{I}_{\text{in}}^{Q}\right)\leq 2\exp(-t/2K_{1})\qquad\text{for all }t\geq 0\text{ and }i\in\mathcal{I}_{\text{in}},$$

where K_1 is a numerical constant. In particular, by Theorem 6.6.17 and the identity $\omega_{\text{fail}}=\mathbb{E}\left[a^{2}\geq t\mid|ab|\leq Q_{\text{fail}}\right]$, we have the following bound

$$\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^{Q}|}\sum_{i\in\mathcal{I}_{\text{in}}^{Q}}\ell_{i,1}^{2}\geq\omega_{\text{fail}}+t\middle||\mathcal{I}_{\text{in}}^{Q}|>\frac{m(1-2p_{\text{fail}})}{2}\right)\leq\exp\left(-c\min\left\{\frac{t^{2}}{K^{2}},\frac{t}{K}\right\}\frac{m(1-2p_{\text{fail}})}{2}\right)$$

for numerical constants c and K, as desired.

□

The proof is complete. □

Proof of Lemma 6.6.3

Our goal is to bound the operator norm of the following matrix:

$$Y_1 = \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \left(P_{\bar{w}_\star} \ell_i \ell_i^\top P_{\bar{w}_\star}^\perp + P_{\bar{w}_\star}^\perp \ell_i \ell_i^\top P_{\bar{w}_\star} \right) = \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \ell_{i,1} \left(e_1 \ell_{i,2:d}^\top + \ell_{i,2:d} e_1^\top \right).$$

Simplifying, we find that

$$Y_1 = \begin{bmatrix} 0 & \lambda_{2:d_1}^\top \\ \lambda_{2:d_1} & 0 \end{bmatrix} \qquad \text{for} \qquad \lambda := \begin{bmatrix} 0 \\ \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \ell_{i,1} \ell_{i,2:d_1} \end{bmatrix} \in \mathbf{R}^{d_1}.$$

Evidently, $\|Y_1\|_{\text{op}} \leq \left\| \lambda_{2:d_1} \right\|_2$, so our focus will be to bound this quantity. We will bound this quantity through the following claim, which is based on Gaussian concentration for Lipschitz functions.

Claim 5. *Consider the (random) function $F : \mathbf{R}^{m \times (d_1 - 1)} \to \mathbf{R}$, given by*

$$F(a_1, \ldots, a_m) = \left\| \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \ell_{i,1} a_i \right\|_2.$$

Then F is $\widehat{\eta} = \frac{1}{m} \sqrt{\sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \ell_{i,1}^2}$ Lipschitz continuous and

$$\mathbb{P}\left(F(\ell_{1,2:d}, \ldots, \ell_{m,2:d}) \geq 2 \sqrt{\frac{d_1 - 1}{m}} + t \,\middle|\, \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathcal{I}_{\text{in}}^{\text{sel}} \right) \leq \exp\left(-\frac{mt^2}{8} \right).$$

Moreover, the following bound holds:

$$\mathbb{P}\left(\widehat{\eta} \geq \frac{2}{\sqrt{m}} \right) \leq \exp\left(-\frac{m}{2} \right).$$

Proof of Claim. For any $A = \begin{bmatrix} a_1 & \dots & a_m \end{bmatrix} \in \mathbf{R}^{m\times(d_1-1)}$ and $B = \begin{bmatrix} b_1 & \dots & b_m \end{bmatrix} \in \mathbf{R}^{m\times(d_1-1)}$, we have

$$|F(A)-F(B)| \le \frac{1}{m}\|(A-B)(\ell_{i,1}\mathbf{1}\{i \in \mathcal{I}_{\text{in}}^{\text{sel}}\})_{i=1}^m\|_2 \le \frac{1}{m}\|(A-B)\|_{\text{op}}\|(\ell_{i,1}\mathbf{1}\{i \in \mathcal{I}_{\text{in}}^{\text{sel}}\})_{i=1}^m\|_2 \le \widehat{\eta}\|A-B\|_F,$$

which proves that F is $\widehat{\eta}$-Lipschitz. Therefore, since for all i the variables $\ell_{i,1}$ and $\ell_{i,2:d_1}$ are independent, standard results on Gaussian concentration for Lipschitz functions (applied conditionally), Theorem 6.6.18, imply that

$$\mathbb{P}\left(F(\ell_{1,2:d},\dots,\ell_{m,2:d}) - \mathbb{E}\left[F(\ell_{1,2:d},\dots,\ell_{m,2:d}) \,\middle|\, \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathcal{I}_{\text{in}}^{\text{sel}}\right] \ge t \,\middle|\, \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathcal{I}_{\text{in}}^{\text{sel}},\right)$$
$$\le \exp\left(-\frac{mt^2}{8}\right).$$

Thus, the first part of the claim is a consequence of the following bound:

$$\mathbb{E}\left[F(\ell_{1,2:d},\dots,\ell_{m,2:d}) \,\middle|\, \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathcal{I}_{\text{in}}^{\text{sel}}\right] \le \sqrt{\mathbb{E}\left[F(\ell_{1,2:d},\dots,\ell_{m,2:d})^2 \,\middle|\, \hat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathcal{I}_{\text{in}}^{\text{sel}}\right]}$$
$$= \sqrt{\frac{1}{m^2}\mathbb{E}\left[\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}} \ell_{i,1}^2(d_1-1) \,\middle|\, \hat{\eta} < \frac{2}{\sqrt{m}}\right]} \le 2\sqrt{\frac{d_1-1}{m}}.$$

We now turn our attention to the high probability bound on $\widehat{\eta}$.

To that end, notice that the (random) function $E\colon \mathbf{R}^m \to \mathbf{R}$ given by

$$E(a) = \frac{1}{m}\sqrt{\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}} a_i^2} = \frac{1}{m}\|(a_i\mathbf{1}\{i \in \mathcal{I}_{\text{in}}^{\text{sel}}\})_{i=1}^m\|_2.$$

is m^{-1}-Lipschitz continuous. Moreover, we have that $\mathbb{E}[E(\ell_{1,i},\dots,\ell_{1,d})] \le \frac{1}{m}\mathbb{E}\left[\|(\ell_{1,i})_{i=1}^m\|_2\right] \le m^{-1/2}$. Therefore, by Gaussian concentration we have

$$\mathbb{P}\left(\widehat{\eta} \ge \frac{2}{\sqrt{m}}\right) \ge \mathbb{P}\left(E(\ell_{1,i},\dots,\ell_{1,d}) - \mathbb{E}\left[E(\ell_{1,i},\dots,\ell_{1,d})\right] \ge \frac{1}{\sqrt{m}}\right) \le \exp\left(-\frac{m}{2}\right),$$

as desired. □

To complete the proof, observe that

$$\begin{aligned}
&\mathbb{P}\left(\left\|\lambda_{2:d_1}\right\|_2 \geq 2\sqrt{\frac{d_1-1}{m}}+t\right)\\
&=\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}}\ell_{i,1}\ell_{i,2:d_1}\right\|_2 \geq 2\sqrt{\frac{d_1-1}{m}}+t\right)\\
&\leq\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}}\ell_{i,1}\ell_{i,2:d_1}\right\|_2 \geq 2\sqrt{\frac{d_1-1}{m}}+t\,\middle|\,\widehat{\eta}<\frac{2}{\sqrt{m}}\right)\mathbb{P}\left(\widehat{\eta}<\frac{2}{\sqrt{m}}\right)+\mathbb{P}\left(\widehat{\eta}\geq\frac{2}{\sqrt{m}}\right)\\
&\leq\mathbb{P}\left(F(\ell_{1,2:d},\ldots,\ell_{m,2:d})\geq 2\sqrt{\frac{d_1-1}{m}}+t\,\middle|\,\widehat{\eta}<\frac{2}{\sqrt{m}}\right)+\exp\left(-\frac{m}{2}\right),
\end{aligned}$$

where the second inequality is due to Claim 5. Finally, by Claim 5, the conditional probability is bounded as follows

$$\begin{aligned}
&\mathbb{P}\left(F(\ell_{1,2:d},\ldots,\ell_{m,2:d})\geq 2\sqrt{\frac{d_1-1}{m}}+t\,\middle|\,\widehat{\eta}<\frac{2}{\sqrt{m}}\right)\\
&=\mathbb{E}_{\mathcal{I}_{\text{in}}^{\text{sel}},\{\ell_{i,1}\}_{i=1}^m}\left[\mathbb{P}\left(F(\ell_{1,2:d},\ldots,\ell_{m,2:d})\geq 2\sqrt{\frac{d_1-1}{m}}+t\,\middle|\,\widehat{\eta}<\frac{2}{\sqrt{m}},\{\ell_{1,i}\}_{i=1}^m,\mathcal{I}_{\text{in}}^{\text{sel}}\right)\right]\\
&\leq\exp\left(-\frac{mt^2}{8}\right),
\end{aligned}$$

which completes the proof.

Proof of Lemma 6.6.4

Observe the equality

$$Y_2=\frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}}\begin{bmatrix}0\\ \ell_{i,2:d_1}\end{bmatrix}\begin{bmatrix}0 & \ell_{i,2:d_1}^\top\end{bmatrix}.$$

Therefore, we seek to bound the following operator norm:

$$\left\|Y_2-S_{\text{in}}\left(I_{d_1}-e_1e_1^\top\right)\right\|_{\text{op}}=\left\|\frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}}\left(\ell_{i,2:d_1}\ell_{i,2:d_1}^\top-I_{d_1-1}\right)\right\|_{\text{op}}.$$

Using the tower rule for expectations and appealing to Corollary 6.6.22, we therefore deduce

$$\begin{aligned}
&\mathbb{P}\left(\left\|Y_2 - S_{\text{in}}\left(I_{d_1} - e_1 e_1^\top\right)\right\|_{\text{op}} \geq C\sqrt{\frac{d_1}{m} + t}\right) \\
&\leq \mathbb{E}_{\mathcal{I}_{\text{in}}^{\text{sel}}}\left[\mathbb{P}\left(\left\|Y_2 - S_{\text{in}}\left(I_{d_1} - e_1 e_1^\top\right)\right\|_{\text{op}} \geq C\sqrt{\frac{d_1}{m} + t} \;\middle|\; \mathcal{I}_{\text{in}}^{\text{sel}} = \mathcal{I}\right)\right] \leq 2\exp(-cmt),
\end{aligned}$$

as desired.

□

Proof of Lemma 6.6.5

Noise model N1 Under this noise model, we write

$$\left\|Y_3 - S_{\text{out}} I_{d_1}\right\|_{\text{op}} = \left\|\frac{1}{m}\sum_{i \in \mathcal{I}_{\text{out}}^{\text{sel}}} \ell_i \ell_i^\top - S_{\text{out}} I_{d_1}\right\|_{\text{op}}.$$

The proof follows by repeating the conditioning argument as in the proof of 6.6.4.

Noise model N2 Observe that

$$\begin{aligned}
\left\|\frac{1}{m}\sum_{i \in \mathcal{I}_{\text{out}}^{\text{sel}}} \ell_i \ell_i^\top\right\|_{\text{op}} \leq \left\|\frac{1}{m}\sum_{i \in \mathcal{I}_{\text{out}}} \ell_i \ell_i^\top\right\|_{\text{op}} &\leq \left\|\frac{1}{m}\sum_{i \in \mathcal{I}_{\text{out}}} (\ell_i \ell_i^\top - I_{d_1})\right\|_{\text{op}} + \left\|\frac{1}{m}\sum_{i \in \mathcal{I}_{\text{out}}} I_{d_1}\right\|_{\text{op}} \\
&= \left\|\frac{1}{m}\sum_{i \in \mathcal{I}_{\text{out}}} (\ell_i \ell_i^\top - I_{d_1})\right\|_{\text{op}} + p_{\text{fail}}.
\end{aligned}$$

Appealing to Corollary 6.6.22, the result follows immediately. □

Proof of Proposition 6.3.2

We will assume that $\|\hat{w}\hat{x}^\top - \bar{w}_\star \bar{x}_\star^\top\|_F \le \|\hat{w}\hat{x}^\top + \bar{w}_\star \bar{x}_\star^\top\|_F$. We will show that with high probability, $|\widehat{\Phi} - \Phi| \le \delta\Phi$, and moreover in this event if $\delta < 1$, we have $\widehat{\Phi} > 0$. The other setting $\|\hat{w}\hat{x}^T - \bar{w}_\star \bar{x}_\star^\top\|_F \ge \|\hat{w}\hat{x}^T + \bar{w}_\star \bar{x}_\star^\top\|_F$ can treated similarly.

We will use the guarantees of Theorem 5.5.8. In particular, there exist numerical constants $c_1, \dots, c_6 > 0$ so that as long as $m \ge \frac{c_1(d_1+d_2+1)}{(1-\frac{2|\mathcal{I}|}{m})^2} \ln\left(c_2 + \frac{1}{1-2|\mathcal{I}|/m}\right)$, then with probability at least $1 - 4\exp\left(-c_3(1-\frac{2|\mathcal{I}|}{m})^2 m\right)$ we have

$$c_4\|X\|_F \le \frac{1}{m}\|\mathcal{A}(X)\|_1 \le c_5\|X\|_F \qquad \text{for all rank} \le 2 \text{ matrices } X \in \mathbf{R}^{d_1 \times d_2},$$

and

$$c_6\,(1-2p_{\text{fail}})\,\|X\|_F \le \frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}} |\ell_i^\top X r_i| - \frac{1}{m}\sum_{i\in\mathcal{I}_{\text{out}}} |\ell_i^\top X r_i| \qquad \text{for all rank} \le 2 \text{ matrices } X \in \mathbf{R}^{d_1 \times d_2}.$$

Throughout the remainder of the proof, suppose we are in this event. Define the two univariate functions

$$\widehat{g}(a) := \frac{1}{m}\sum_{i=1}^m \left|b_i - (1+a)\Phi\ell_i^\top \widehat{w}\widehat{x}^\top r_i\right|,$$

$$g(a) := \frac{1}{m}\sum_{i=1}^m \left|b_i - (1+a)\Phi\ell_i^\top \bar{w}\bar{x}^\top r_i\right|$$

By construction, if $a^\star$ minimizes $\widehat{g}(\cdot)$ then $(1+a^\star)\Phi$ minimizes G. Thus, to prove the claim we need only show that any minimizer $a^\star$ of $\widehat{g}$ satisfies $-\delta \le a^\star \le \delta$.

To that end, first note that $g(0)$ and $\widehat{g}(0)$ are close:

$$|\widehat{g}(0) - g(0)| \le \frac{\Phi}{m}\sum_{i=1}^m |\ell_i^\top \widehat{w}\widehat{x}^\top r_i - \ell_i^\top \bar{w}_\star \bar{x}_\star^\top r_i| \le c_5\Phi\left\|\widehat{w}\widehat{x}^\top - \bar{w}_\star\bar{x}_\star^\top\right\|_F, \tag{6.28}$$

Therefore, setting $\mu_3 = c_6(1 - 2p_{\text{fail}})$, we obtain

$$\begin{aligned}
\hat{g}(a) &= \frac{1}{m}\sum_{i=1}^{m}\left|b_i - (1+a)\Phi\ell_i^\top\widehat{wx}^\top r_i\right| \\
&= \frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}}\left|\ell_i^\top\bar{w}\bar{x}^\top r_i - (1+a)\Phi\ell_i^\top\widehat{wx}^\top r_i\right| + \frac{1}{m}\sum_{i\in\mathcal{I}_{\text{out}}}\left|b_i - (1+a)\Phi\ell_i^\top\widehat{wx}^\top r_i\right| \\
&\geq \frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}}\left|\ell_i^\top\bar{w}\bar{x}^\top r_i - (1+a)\Phi\ell_i^\top\widehat{wx}^\top r_i\right| - \frac{1}{m}\sum_{i\in\mathcal{I}_{\text{out}}}\left|\ell_i^\top\bar{w}\bar{x}^\top r_i - (1+a)M\ell_i^\top\widehat{wx}^\top r_i\right| \\
&\quad + \frac{1}{m}\sum_{i\in\mathcal{I}_{\text{out}}}\left|b_i - \ell_i^\top\bar{w}\bar{x}^\top r_i\right| \\
&\geq g(0) + \mu_3\|(1+a)\Phi\widehat{wx}^\top - \bar{w}\bar{x}^\top\|_F \\
&\geq \widehat{g}(0) + \mu_3\|(1+a)\Phi\widehat{wx}^\top - \bar{w}\bar{x}^\top\|_F - c_5\Phi\left\|\widehat{wx}^\top - \bar{w}_\star\bar{x}_\star^\top\right\|_F \\
&\geq \widehat{g}(0) + \mu_3|a|\Phi - (\mu_3\Phi + c_5\Phi)\left\|\widehat{wx}^\top - \bar{w}_\star\bar{x}_\star^\top\right\|_F,
\end{aligned}$$

where the second inequality follows from Theorem 5.5.8, the third inequality follows from Equation (6.28), and the fourth follows from the reverse triangle inequality. Thus, any minimizer $a^\star$ of $\hat{g}$ must satisfy

$$|a^\star| \leq \left(1 + \frac{c_5}{\mu_3}\right)\left\|\widehat{wx}^\top - \bar{w}_\star\bar{x}_\star^\top\right\|_F = \delta,$$

as desired. Finally suppose $\delta < 1$. Then we deduce $\widehat{\Phi} = (1 + |a^\star|)\Phi \geq (1 - \delta)\Phi > 0$. The proof is complete.

6.6.3 Proofs in Section 6.4

Proof of Proposition 6.4.2

Recall that we defined the functions $f : \mathbf{R}_+^d \to \mathbf{R}$ and $f_\sigma : \mathbf{R}^{d\times n} \to \mathbf{R}$ to be such that $f_P(w, x) = f_\sigma(X) = f(\sigma(X))$. It is known that for constants $c_1, c_2 \in \mathbf{R}_+$ we have

that $c_1 r_1 + c_2 r_2 \overset{(d)}{=} \sqrt{c_1^2 + c_2^2} r_1$, where $\overset{(d)}{=}$ denotes equality in distribution. Then

$$\begin{aligned}
f(s_1, s_2, 0, \ldots, 0) &= \mathbb{E}(|s_1 \ell_1 r_1 + s_2 \ell_2 r_2|) \\
&= \mathbb{E}\left(\mathbb{E}(|s_1 \ell_1 r_1 + s_2 \ell_2 r_2| \mid \ell_1, \ell_2)\right) \\
&= \mathbb{E}\left(\mathbb{E}(\sqrt{(s_1 \ell_1)^2 + (s_2 \ell_2)^2} |r_1| \mid \ell_1, \ell_2)\right) \\
&= \sqrt{\frac{2}{\pi}} \mathbb{E} \sqrt{(s_1 \ell_1)^2 + (s_2 \ell_2)^2} \\
&= \frac{\sqrt{\pi}}{\sqrt{2}\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sqrt{(\ell_1 s_1)^2 + (\ell_2 s_2)^2} \exp\left(-\frac{\ell_1^2 + \ell_2^2}{2}\right) d\ell_1 d\ell_2 \\
&= 4 \frac{\sqrt{\pi}}{\sqrt{2}\pi^2} \int_0^{\infty} \int_0^{\infty} \sqrt{(\ell_1 s_1)^2 + (\ell_2 s_2)^2} \exp\left(-\frac{\ell_1^2 + \ell_2^2}{2}\right) d\ell_1 d\ell_2 \\
&= 2 \frac{\sqrt{2\pi}}{\pi^2} \int_0^{\infty} \int_0^{\pi/2} t^2 \sqrt{s_1^2 \cos^2\theta + s_2^2 \sin^2\theta} \exp\left(-\frac{t^2}{2}\right) d\theta dt \\
&= \frac{2}{\pi} \int_0^{\pi/2} \sqrt{s_1^2 \cos^2\theta + s_2^2 \sin^2\theta} d\theta \\
&= \frac{2 s_1}{\pi} \int_0^{\pi/2} \sqrt{\cos^2\theta + \frac{s_2^2}{s_1^2} \sin^2\theta} d\theta \\
&= \frac{2 s_1}{\pi} \int_0^{\pi/2} \sqrt{1 - \left(1 - \frac{s_2^2}{s_1^2}\right) \sin^2\theta} d\theta \\
&= \frac{2 s_1}{\pi} E\left(\left(1 - \frac{s_2^2}{s_1^2}\right)^{\frac{1}{2}}\right)
\end{aligned}$$

where $E(\cdot)$ is the complete elliptic integral of the second kind. Thus altogether we obtain

$$f_\sigma(X) = \sigma_{\max}(X) \sum_{n=0}^{\infty} \left(\frac{(2n)!}{2^{2n}(n!)^2}\right)^2 \frac{\left(1 - \kappa^{-2}(X)\right)^n}{1 - 2n}$$

where $\kappa(X) = \sigma_{\max}(X)/\sigma_{\min}(X)$ is the condition number of X.

Proof of Theorem 6.4.3

The proof of this result builds upon the next three lemmas. We will prove these lemmas before we dive into the proof. Recall that $U \in O(d_1)$ and $V \in O(d_2)$ are

any pair of matrices for which $X = U\sigma(X)V = \sum_i \sigma_i(X)U_iV_i^\top$.

Lemma 6.6.6. *The following are true.*

1. *(**Anticorrelation**) The next equalities hold*

$$\langle U_1, w\rangle\langle x, V_2\rangle = \langle U_1, \bar{w}\rangle\langle \bar{x}, V_2\rangle \qquad \text{and} \qquad \langle U_2, w\rangle\langle x, V_1\rangle = \langle U_2, \bar{w}\rangle\langle \bar{x}, V_1\rangle.$$

2. *(**Singular values**) The singular values of X satisfy*

$$\sigma_1(X) = \langle U_1, w\rangle\langle x, V_1\rangle - \langle U_1, \bar{w}\rangle\langle \bar{x}, V_1\rangle \geq 0,$$

$$\sigma_2(X) = \langle U_2, w\rangle\langle x, V_2\rangle - \langle U_2, \bar{w}\rangle\langle \bar{x}, V_2\rangle \geq 0.$$

3. *(**Correlation**) Assume that $\sigma_2(wx^\top - \bar{w}\bar{x}^\top) > 0$, then $\text{span}\{x, \bar{x}\} = \text{span}\{V_1, V_2\}$, $\text{span}\{w, \bar{w}\} = \text{span}\{U_1, U_2\}$, and consequently,*

$$\langle w, \bar{w}\rangle = \langle U_1, w\rangle\langle U_1, \bar{w}\rangle + \langle U_2, w\rangle\langle U_2, \bar{w}\rangle,$$

$$\langle x, \bar{x}\rangle = \langle V_1, x\rangle\langle V_1, \bar{x}\rangle + \langle V_2, x\rangle\langle V_2, \bar{x}\rangle.$$

Proof. The first equality in item one follows by observing that $U_1^\top XV_2 = 0$, expanding the expression on the left-hand-side gives the result. The same argument starting from $U_2^\top XV_1 = 0$ gives the other equality. The second item follows by definition.

To prove the last item note that

$$\langle U_i, w\rangle x - \langle U_i, \bar{w}\rangle\bar{x} = X^\top U_i = \sigma_i(X)V_i \qquad \forall i \in \{1, 2\}.$$

Dividing through by $\sigma_i(X)$ shows that $\text{span}\{x, \bar{x}\} = \text{span}\{V_1, V_2\}$. Therefore, we can write $x = \langle x, V_1\rangle V_1 + \langle x, V_2\rangle V_2$ and $\bar{x} = \langle \bar{x}, V_1\rangle V_1 + \langle \bar{x}, V_2\rangle V_2$. Hence,

$$\langle x, \bar{x}\rangle = \langle\langle x, V_1\rangle V_1 + \langle x, V_2\rangle V_2, \langle \bar{x}, V_1\rangle V_1 + \langle \bar{x}, V_2\rangle V_2\rangle = \langle V_1, x\rangle\langle V_1, \bar{x}\rangle + \langle V_2, x\rangle\langle V_2, \bar{x}\rangle$$

An analogous argument shows the statement for w and $\bar{w}$. □

Lemma 6.6.7. *The following inequalities hold true.*

1. ***(Maximum correlation)***

$$\begin{aligned} \max\{|\sigma_1(Y)\langle v_1, x\rangle|, |\sigma_2(Y)\langle v_2, x\rangle|\} &\leq \|Yx\|, \\ \max\{|\sigma_1(Y)\langle u_1, w\rangle|, |\sigma_2(Y)\langle u_2, w\rangle|\} &\leq \|Y^\top w\|. \end{aligned} \tag{6.29}$$

2. ***(Objective gap)***

$$g(w, x) - g(\bar{w}, \bar{x}) \leq \sigma_1(Y)\sigma_1(X) + \sigma_2(Y)\sigma_2(X). \tag{6.30}$$

Proof. Note that $\|Yx\| \geq \langle z, Yx\rangle$ for all $z \in \mathbb{S}^{d-1}$, then the very first claim follows by testing with $z \in \{\pm U_1, \pm U_2\}$. An analogous argument gives the statement for w. Recall that f is convex, consequently f_σ is convex and the subgradient inequality gives

$$g(w, x) - g(\bar{w}, \bar{x}) = f_\sigma(X) - f_\sigma(0) \leq \langle Y, X\rangle = \sigma_1(Y)\sigma_1(X) + \sigma_2(Y)\sigma_2(X).$$

□

Lemma 6.6.8. *Assume $\bar{w} \in \mathbf{R}^{d_1}$ and $\bar{x} \in \mathbf{R}^{d_2}$ are nonzero vectors. Set $X = wx^\top + \bar{w}\bar{x}^\top$, then X is a rank 1 matrix if, and only if, $w = \lambda\bar{w}$ or $x = \lambda\bar{x}$ for some $\lambda \in \mathbf{R}$.*

Proof. It is trivial to see that if the later holds then X is rank 1. Let us prove the other direction. Notice that if any of the vectors is zero we are done, so assume that none of them is. Recall that all the columns of X are spanned from one vector. Consider the case where x and $\bar{x}$ have different support (i.e. set of nonzero entries), then it is immediate that w and $\bar{w}$ have to be multiples of each other.

Now assume that this is not the case, without loss of generality assume that $w \notin \text{span}\{\bar{w}\}$ and x and $\bar{x}$ are nonzero and their first component is equal to one. Then the first column of X is equal to $w + \bar{w}$, furthermore the second column is equal to $x_2 w + \bar{x}_2 \bar{w}$ has to be a multiple of the first one. By assumption $w, \bar{w}$ are linearly independent therefore $x_2 = \bar{x}_2$. Using the same procedure for the rest of the entries we obtain $x = \bar{x}$. □

We are now in good shape to describe the landscape of the function g.

Proof of Theorem 6.4.3. To prove that at least one of the conditions hold we will show that if the first two don't hold then at least one of the other two have two hold. Assume that that the first two conditions are not satisfied, therefore $g(w, x) > g(\bar{w}, \bar{x})$ and $(w, x) \neq (0, 0)$. Let us furnished some facts before we prove this is the case. Notice that from (6.30) we can derive

$$0 < \sigma_1(Y)\sigma_1(X) + \sigma_2(Y)\sigma_1(X) \leq 2\sigma_1(Y)\sigma_1(X),$$

thus $\sigma_1(Y), \sigma_1(X) > 0$. On the other hand, since (w, x) is critical inequalities (6.29) immediately give

$$\sigma_1(Y)\langle V_1, x\rangle = \sigma_2(Y)\langle V_2, x\rangle = 0, \qquad \text{and} \qquad \sigma_1(Y)\langle U_1, w\rangle = \sigma_2(Y)\langle U_2, w\rangle = 0. \tag{6.31}$$

So $\langle V_1, x\rangle = 0$ and $\langle U_1, w\rangle = 0$, then the first claim in Lemma 6.6.6 gives. Additionally, this and the second claim in Lemma 6.6.6 imply that

$$\langle U_1, \bar{w}\rangle\langle \bar{x}, V_2\rangle = \langle U_2, \bar{w}\rangle\langle \bar{x}, V_1\rangle = 0, \qquad \text{and} \qquad -\langle U_1, \bar{w}\rangle\langle \bar{x}, V_1\rangle = \sigma_1(X) > 0.$$

Combining these two gives $\langle U_2, \bar{w}\rangle = \langle \bar{x}, V_2\rangle = 0$. Then by applying the second claim in Lemma 6.6.6 we get $\sigma_2(X) = \langle U_2, w\rangle\langle x, V_2\rangle$. Using Equations (6.31) we conclude that $\sigma_2(Y)\sigma_2(X) = 0$.

Now we will show that at least one of the conditions holds, depending on the value of $\sigma_2(X)$, let us consider two cases:

Case 1. Assume $\sigma_2(X) = 0$. This means that $X = wx^\top - \bar{w}\bar{x}^\top$ is a rank 1 matrix. By Lemma 6.6.8 we have that $w = \lambda\bar{w}$ or $x = \lambda\bar{x}$ for some $\lambda \in \mathbf{R}$. Note that if $w = \lambda\bar{w}$ then $U_1 = \pm\bar{w}/\|\bar{w}\|$, then using Equation 6.31 we get that $\lambda\|\bar{w}\| = 0$. Which implies that $\lambda = 0$, and consequently $wx^\top = 0$. An analogous argument applies when $x = \lambda\bar{x}$. By assumption we have that $Yx = 0$ and $Y^\top w = 0$. Additionally, since $X = -\bar{w}\bar{x}^\top$ we get that that $U_1 = \pm\bar{w}/\|\bar{w}\|$ and $V_1 = \pm\bar{x}/\|\bar{x}\|$. Recall that $Y = U\mathrm{diag}(\sigma(Y))V^\top$, then using the fact that (w, x) is critical we conclude $\langle w, \bar{w}\rangle = \langle x, \bar{x}\rangle = 0$. Implying that property three holds.

Case 2. Assume $\sigma_2(X) \neq 0$. This immediately implies that $\sigma_2(Y) = 0$. By the third part of Lemma 6.6.6 we get that

$$\langle x, \bar{x}\rangle = \langle V_1, x\rangle\langle V_1, \bar{x}\rangle + \langle V_2, x\rangle\langle V_2, \bar{x}\rangle = 0$$

and analogously $\langle w, \bar{w}\rangle = 0$. Moreover, since $w \perp \bar{w}$ and $x \perp \bar{w}$ (and none of them are zero by assumption) we get that $(w/\|w\|, x/\|x\|)$ and $(\bar{w}/\|\bar{w}\|, \bar{x}/\|\bar{x}\|)$ are pairs of left and right singular vectors, with associated singular values $w^\top Xx = \|wx^\top\|$ and $\bar{w}^\top X\bar{x} = \|\bar{w}\bar{x}^\top\|$, respectively. Assume that $\|wx^\top\| \geq \|\bar{w}\bar{x}^\top\|$, thus $0 = w^\top Yx = \|wx^\top\|\sigma_1(Y) > 0$, yielding a contradiction. Hence the condition four holds true.

Finally, we will prove that if $(\bar{w}, \bar{x})$ minimizes g, then (w, x) is a critical point if it satisfies at least one of the four conditions in the statement of the theorem for some $Y \in \partial f_\sigma(X)$.. Assume that $(\bar{w}, \bar{x})$ minimizes g. The set of points that satisfies the first conditions is the collection of minimizers so they are critical. Clearly $(w, x) = 0$ is always a stationary point, since $\|Y^\top w\| = \|Yx\| = 0$. Now let's construct a certificate $Y \in \partial f_\sigma(X)$ that ensures criticality for the remaining cases.

Suppose that the third condition is satisfied. That is, assume that (w, x) satisfies $w = 0$ and there exists $Y \in \partial f_\sigma(w, x)$ such that $Yx = 0$ and $\langle x, \bar{x}\rangle = 0$. Using (6.13), it is immediate that (w, x) is a stationary point. A similar argument follows when $x = 0$.

Suppose that the fourth condition is satisfied. Thus, assume that (w, x) is such that $0 < \|wx^\top\| < \|\bar{w}\bar{x}^\top\|$, $\langle w, \bar{w}\rangle = \langle x, \bar{x}\rangle = 0$ and there exists $Y \in \partial f_\sigma(X)$ with $\sigma_2(Y) = 0$. Since $w \perp \bar{w}$, $x \perp x$, $\|wx^\top\| < \|\bar{w}\bar{x}^\top\|$, using the same argument as in Case 2, we get that any pair of admissible matrices U, V satisfy $U_1 = \pm\bar{w}/\|\bar{w}\|$ and $V_1 = \pm\bar{x}/\|\bar{x}\|$. Therefore

$$Yx = (\sigma_1(Y)U_1V_1^\top)x = \pm\frac{\sigma_1(Y)}{\|\bar{x}\|}\langle\bar{x}, x\rangle U_1 = 0,$$

analogously $Y^\top w = 0$. □

Proof of Lemma 6.4.4

It is well-known that if $(\ell_1, \ell_2, \ldots, \ell_d)$ is a fixed vector, then

$$\sum_{i=1}^{d} \ell_i r_i s_i \overset{(d)}{=} \left(\sum_{i=1}^{d} (\ell_i s_i)^2\right)^{\frac{1}{2}} r$$

and b is a standard normal random variable independent of the rest of the data. Therefore

$$\begin{aligned} f(s_1, \ldots, s_d) &= \mathbb{E}\left(\left|\sum_{i=1}^{d} \ell_i r_i s_i\right|\right) = \mathbb{E}\left(\mathbb{E}\left(\left|\sum_{i=1}^{d} \ell_i r_i s_i\right| \middle| \ell_1, \ldots, \ell_d\right)\right) \\ &= \mathbb{E}\left(\left(\sum_{i=1}^{d} (\ell_i s_i)^2\right)^{\frac{1}{2}} \mathbb{E}\left(|r| \mid \ell_1, \ldots, \ell_d\right)\right) = \sqrt{\frac{2}{\pi}} \cdot \mathbb{E}\left(\sum_{i=1}^{d} (\ell_i s_i)^2\right)^{\frac{1}{2}}. \end{aligned}$$

Now, we need a technical tool in order to procede.

Theorem 6.6.9 (**Leibniz Integral Rule**, Theorem 5.4.12 in [208])**.** *Let U be an open subset of $\mathbf{R}^d$ and Ω be a measure space. Suppose that the function $h : U \times \Omega \to \mathbf{R}$ satisfies the following:*

1. *For all $x \in U$, the function $h(x, \cdot)$ is Lebesgue integrable.*
2. *For almost all $w \in \Omega$, if we define $h^\omega(\cdot) = f(\cdot, \omega)$ the partial derivatives $\frac{\partial h^\omega}{\partial x_i}(x)$ exists for all $x \in U$.*
3. *There is an integrable function $\Phi : \Omega \to \mathbf{R}$ such that $|\frac{\partial h^\omega}{\partial x_i}(x)| \leq \Phi(\omega)$ for all $x \in U$ and almost every $\omega \in \Omega$.*

Then, we have that for all $x \in U$

$$\frac{\partial}{\partial x_i} \int_\Omega h(x, \omega) d\omega = \int_\Omega \frac{\partial h^\omega}{\partial x_i}(x) d\omega.$$

This theorem tell us that we can swap partial derivatives and integrals provided that the function satisfies all the conditions above. Consider Ω to be the set $\mathbf{R}^d$ endowed with the Borel σ-algebra and the multivariate Gaussian measure. Define $h : \mathbf{R}^d \times \Omega \to \mathbf{R}$ to be given by

$$(s, \ell) \mapsto \left(\sum_{i=1}^d (\ell_i s_i)^2 \right)^{\frac{1}{2}}.$$

Take $s \in \mathbf{R}^d \setminus \{0\}$ to be an arbitary element, set $S = \{u \in \mathbf{R}^d \mid \text{supp}(s) \subseteq \text{supp}(u)\}$, and define $U = B_\epsilon(s)$ with ϵ small enough such that $U \subseteq S$ and $\inf_{u \in U} \min_{i \in \text{supp}(s)} |u_i| > 0$. Then it is easy to see that the first two conditions hold, in particular the second condition hold for all $a \neq 0$. Further, for any $x \in U$

$$\begin{aligned} \left| \frac{\partial h^\ell}{\partial s_j}(x) \right| = \left| \frac{\ell_j^2 x_j}{\left(\sum_i^d (\ell_i x_i)^2 \right)^{\frac{1}{2}}} \right| &\leq \frac{\sup_{u \in U} \|u\|_\infty}{\inf_{u \in U} \min_{i \in \text{supp}(s)} |u_i|} \frac{\sum_{i \in \text{supp}(s)} \ell_j^2}{\left(\sum_{i \in \text{supp}(s)} \ell_i^2 \right)^{\frac{1}{2}}} \\ &\leq \frac{\sup_{u \in U} \|u\|_\infty}{\inf_{u \in U} \min_i |u_i|} \left(\sum_{i \in \text{supp}(s)} \ell_i^2 \right)^{\frac{1}{2}}, \end{aligned}$$

where the last function is integrable with respect to the Gaussian measure. Thus, Theorem 6.6.9 ensures that the function f is differentiable at every nonzero point. Consequently, for all $s \in \mathbf{R}^d \setminus \{0\}$

$$\frac{\partial f}{\partial s_j}(s) = \sqrt{\frac{2}{\pi}} s_j \, \mathbb{E} \frac{\ell_j^2}{\left(\sum_i^d (\ell_i s_i)^2\right)^{\frac{1}{2}}}.$$

Approximate critical points of a spectral function family

In Section 6.4.2, we characterized the points for which $0 \in \partial f_P(w, x)$. In order to derive similar results for f_S we will need to understand ε-critical points of f_P, i.e. points (w, x) for which $\operatorname{dist}(0, \partial f(w, x)) \leq \varepsilon$. Just as before we adopt a more general viewpoint and consider spectral functions of the form $g(w, x) = f \circ \sigma(wx^\top - \bar{w}\bar{x}^\top)$.

The main result in this section is Theorem 6.6.12. Given the fact that we don't have second order information in the form of a Hessian, we need to appeal to a different kind of growth condition. Turns out that the natural condition for this problem is

$$g(w, x) - g(\bar{w}, \bar{x}) \geq \kappa \left\| wx^\top - \bar{w}\bar{x}^\top \right\|_F \qquad \forall (w, x) \in \mathbf{R}^{d_1} \times \mathbf{R}^{d_2}, \tag{6.32}$$

for some $\kappa > 0$. Intuitively this means that the function grows sharply away from minimizers.

Before we dive into the main theorem, let us provide some technical lemmas.

Lemma 6.6.10. *Suppose there exists a constant $\kappa > 0$ such that* (6.32) *holds. Then, for any point (w, x) such that $wx^\top \neq \bar{w}\bar{x}^\top$ we have $\sigma_1(Y) + \sigma_2(Y) \geq \kappa$.*

Proof. By definition $\sigma_2(X) \leq \sigma_1(X) \leq \left\|wx^\top - \bar{w}\bar{x}^\top\right\|_F$. Then, applying (6.30) gives

$$\begin{aligned}\kappa \left\|wx^\top - \bar{w}\bar{x}^\top\right\|_F \leq g(w,x) - g(\bar{w},\bar{x}) &\leq \sigma_1(Y)\sigma_1(X) + \sigma_2(Y)\sigma_2(X) \\ &\leq (\sigma_1(Y) + \sigma_2(Y)) \left\|wx^\top - \bar{w}\bar{x}^\top\right\|_F.\end{aligned}$$

□

Lemma 6.6.11. *Suppose there exists a constant $\kappa > 0$ such that* (6.32) *holds. Then any pair $(w,x) \in \mathbf{R}^{d_1+d_2} \setminus \{0\}$ satisfies*

$$\frac{1}{\min\{\|w\|, \|x\|\}} \left(\kappa \left\|wx^\top - \bar{w}\bar{x}^\top\right\| - (\sigma_1(Y) + \sigma_2(Y)) \left\|\bar{w}\bar{x}^\top\right\|\right) \leq \operatorname{dist}(0; \partial g(w,x)).$$

Proof. The result holds trivially if $wx^\top = \bar{w}\bar{x}^\top$. Assume this is not the case. Recall that $\partial g(w,x) = \partial f_\sigma(X)x \times (\partial f_\sigma(X))^\top w$. Pick $Y \in \partial g(w,x)$ such that $\operatorname{dist}(0, \partial g(w,x)) = \sqrt{\|Yx\|^2 + \|Y^\top w\|^2}$. Using the convexity of f_σ we get

$$\begin{aligned}\kappa\|wx^\top - \bar{w}\bar{x}\|_F \leq g(w,x) - g(\bar{w},\bar{x}) &= f_\sigma(X) - f_\sigma(0) \\ &\leq \langle Y, wx^\top - \bar{w}\bar{x}^\top \rangle \\ &\leq \|x\|\|Y^\top w\| + |w^\top Yx| \leq \|x\|\operatorname{dist}(0, \partial g(w,x)) + |w^\top Yx|,\end{aligned}$$

where the last inequality follows by Cauchy-Schwartz. Applying the same argument using $w^\top Yx \leq \|w\|\|Yx\|$ gives

$$g(w,x) - g(\bar{w},\bar{x}) \leq \min\{\|w\|, \|x\|\}\operatorname{dist}(0, \partial g(w,x)) + |w^\top Yx|.$$

Now, let's bound the second term on the right-hand-side. Note that

$$|\bar{w}^\top Y\bar{x}| = |\langle Y, wx^\top \rangle| \leq \|Y\|\|wx^\top\| \leq (\sigma_1(Y) + \sigma_2(Y)) \left\|\bar{w}\bar{x}^\top\right\|.$$

The result follows immediately. □

We now prove the main result of this section, a detailed location description of ε-critical points. This is a quantitative version of Corollary 6.4.3. Its proof is however more involved due to the inexactness of the assumptions.

Theorem 6.6.12. *Assume that* $\|\bar{w}\| = \|\bar{x}\|$ *and that there exists a constant* $\kappa > 0$ *such that* (6.32) *holds. Let* $Y \in \partial f_\sigma(wx^\top - \bar{w}\bar{x}^\top)$ *and suppose* $\sigma_1(Y)$ *is bounded by some numerical constant.*[6] *Let* $\zeta = (Yx, Y^\top w) \in \partial g(w, x)$*, and set* $\varepsilon = \|\zeta\|$*. Then if* $wx^\top = 0$ *we have that*

$$\max\{\|Yx\|, \|Y^\top w\|\} \leq \varepsilon, \qquad \text{and} \qquad \begin{cases} |\langle w, \bar{w}\rangle| & \lesssim \varepsilon\|\bar{w}\| \\ |\langle x, \bar{x}\rangle| & \lesssim \varepsilon\|\bar{x}\| \end{cases}.$$

On the other hand, if $wx^\top \neq 0$ *and* $\|(w, x)\| \leq \nu\|(\bar{w}, \bar{x})\|$ *for some fixed* $\nu > 1$*. There exists a constant*[7] $\gamma > 0$ *such that if* $\varepsilon \leq \gamma \max\{\|w\|, \|x\|\}$ *then* $\left\|wx^\top\right\| \lesssim \left\|\bar{w}\bar{x}^\top\right\|$ *and at least one of the following holds*

1. $$\max\{\|w\|, \|x\|\} \left\|wx^\top - \bar{w}\bar{x}^\top\right\| \lesssim \varepsilon \left\|\bar{w}\bar{x}^\top\right\|$$

2. $$\min\{\|w\|, \|x\|\} \lesssim \varepsilon \qquad \text{and} \qquad \begin{cases} |\langle w, \bar{w}\rangle| & \lesssim \nu^2\varepsilon\|\bar{w}\| \\ |\langle x, \bar{x}\rangle| & \lesssim \nu^2\varepsilon\|\bar{x}\| \end{cases}.$$

3. $$\sigma_2(Y) \lesssim \frac{\varepsilon}{\max\{\|w\|, \|x\|\}} \qquad \text{and} \qquad \begin{cases} |\langle w, \bar{w}\rangle| & \lesssim \nu^2\varepsilon\|\bar{w}\| \\ |\langle x, \bar{x}\rangle| & \lesssim \nu^2\varepsilon\|\bar{x}\| \end{cases}.$$

Proof. First assume that $wx^\top = 0$, then it is clear that $\max\{\|Yx\|, \|Y^\top w\|\} = \|\zeta\| \leq \varepsilon$. Without loss of generality assume that $x = 0$. Let $U\sigma(Y)V^\top$ be the singular value decomposition of Y. Since $X = -\bar{w}\bar{x}^\top$ then $U_1 = \pm\bar{w}/\|\bar{w}\|$ and $V_1 = \pm\bar{x}/\|\bar{x}\|$ and so

$$\varepsilon \geq \|Y^\top w\| = \left\|\frac{\sigma_1(Y)}{\|\bar{w}\|}\langle\bar{w}, w\rangle V_1 + z\right\| \geq \frac{\sigma_1(Y)}{\|\bar{w}\|}|\langle\bar{w}, w\rangle| \geq \frac{\kappa}{2\|\bar{w}\|}\langle\bar{w}, w\rangle \tag{6.33}$$

where z is orthogonal to V_1 and the second inequality follows by Lemma 6.6.10. This proves the first statement in the theorem.

[6] This is implied for example when f is Lipschitz.
[7] Independent of ν.

We know move to the next statement, assume $wx^\top \neq 0$ and $\|(w,x)\| \leq \nu\|(\bar{w},\bar{x})\|$. Notice that the result holds immediately if $(w,x) \in \{(\alpha\bar{w}, \bar{x}/\alpha) \mid \alpha \in \mathbf{R}\}$. Further, due to Theorem 6.4.3 it also holds when $\varepsilon = 0$. Let us assume that none of these two conditions are satisfied.

We will start by showing that $\left\|wx^\top\right\| \lesssim \left\|\bar{w}\bar{x}^\top\right\|$. Set

$$\delta = \frac{\sqrt{2}}{\kappa}(\sigma_1(Y) + \sigma_2(Y)) + 1. \tag{6.34}$$

We showed in Lemma 6.6.10 that $(\sigma_1(Y) + \sigma_2(Y)) \geq \kappa$ and thus $\delta > 1$.

Claim 6. *The inequality $\left\|wx^\top\right\| \leq \delta \left\|\bar{w}\bar{x}^\top\right\|$ holds true.*

Proof. Seeking contradiction assume that this is not the case. By Lemma 6.6.11,

$$\frac{\sqrt{2}}{\|wx^\top\|}\kappa \left\|wx^\top - \bar{w}\bar{x}^\top\right\| - \frac{\varepsilon}{\max\{\|w\|, \|x\|\}} \leq (\sigma_1(Y) + \sigma_2(Y))\frac{\left\|\bar{w}\bar{x}^\top\right\|}{\|wx^\top\|}. \tag{6.35}$$

Using $\delta\|\bar{w}\bar{x}^\top\| < \|wx^\top\|$, we get

$$\frac{\sqrt{2}}{\|wx^\top\|}\kappa \left\|wx^\top - \bar{w}\bar{x}^\top\right\| = \sqrt{2}\kappa \left\|\frac{wx^\top}{\|wx^\top\|} - \frac{\bar{w}\bar{x}^\top}{\|wx^\top\|}\right\| \geq \sqrt{2}\kappa \left|1 - \frac{1}{\delta}\right|.$$

We set γ small enough to ensure $\gamma < \frac{\sqrt{2}\kappa}{2}\left|1 - \frac{1}{\delta}\right|$, which implies $\frac{\varepsilon}{\max\{\|w\|,\|x\|\}} < \frac{\sqrt{2}\kappa}{2}\left|1 - \frac{1}{\delta}\right|$. Combining these inequality leads

$$\begin{aligned}\frac{\sqrt{2}\kappa\left|1 - \frac{1}{\delta}\right|}{2(\sigma_1(Y) + \sigma_2(Y))} &\leq \frac{1}{(\sigma_1(Y) + \sigma_2(Y))}\left(\frac{\sqrt{2}}{\|wx^\top\|}\kappa \left\|wx^\top - \bar{w}\bar{x}^\top\right\| - \frac{\varepsilon}{\max\{\|w\|, \|x\|\}}\right) \\ &\leq \frac{\left\|\bar{w}\bar{x}^\top\right\|}{\|wx^\top\|} < \frac{1}{\delta}.\end{aligned}$$

Rearranging we get

$$|\delta - 1| < \frac{\sqrt{2}}{\kappa}(\sigma_1(Y) + \sigma_2(Y)),$$

contradicting the definition of δ. This establishes the claim. □

We now move on to proving that at least one of the three conditions has to hold. To this end, define

$$\rho_1 := \frac{\max\{\|w\|, \|x\|\}}{\sqrt{2}} \qquad \text{and} \qquad \rho_2 := \frac{1}{\kappa}\max\{2\sqrt{2}(1+\delta), 4\sigma_1(Y)\}\frac{\left\|\bar{w}\bar{x}^\top\right\|}{\max\{\|w\|, \|x\|\}}.$$

Observe that if $\varepsilon\rho_2 \geq \left\|wx^\top - \bar{w}\bar{x}^\top\right\|$ then condition 1 holds and the result follows. Then, assume from now on that $\varepsilon\rho_2 < \left\|wx^\top - \bar{w}\bar{x}^\top\right\|$.

Our road map is as follows, we will start by assuming $\min\{\|w\|, \|x\|\} \leq 2\varepsilon/\kappa$ and we will show that this implies the second condition in item two. Then we will move to assume that $\min\{\|w\|, \|x\|\} > 2\varepsilon/\kappa$ and show that item three has to hold.

Now, let we list some facts that we will use later. By Lemma 6.6.7

$$\max\{\sigma_1(Y)|\langle V_1, x\rangle|, \sigma_2(Y)|\langle V_2, x\rangle|, \sigma_1(Y)|\langle U_1, w\rangle|, \sigma_2(Y)|\langle U_2, w\rangle|\} \leq \varepsilon \tag{6.36}$$

which together with $\sigma_1(Y) > \kappa/2$ implies that

$$\max\{|\langle U_1, w\rangle|, |\langle V_1, x\rangle|\} \leq \frac{\varepsilon}{\sigma_1(Y)} \leq \frac{2\varepsilon}{\kappa}. \tag{6.37}$$

Notice that this implies by Lemma 6.6.6

$$|\langle U_1, \bar{w}\rangle\langle \bar{x}, V_2\rangle| = |\langle U_1, w\rangle\langle x, V_2\rangle| \leq \frac{2\|x\|\varepsilon}{\kappa} \qquad \text{and} \qquad |\langle U_2, \bar{w}\rangle\langle \bar{x}, V_1\rangle| \leq \frac{2\|w\|\varepsilon}{\kappa}. \tag{6.38}$$

Observe that

$$\max\{\|w\|, \|x\|\} \leq \|(w, x)\| \leq \nu\|(\bar{w}, \bar{x})\| = \sqrt{2}\nu\min\{\|\bar{w}\|, \|\bar{x}\|\}. \tag{6.39}$$

We can now continue with the proof. We will now assume that $\min\{\|w\|, \|x\|\} \leq 4\delta\varepsilon/\kappa$ and prove that item two holds.

Claim 7. *Assume that* $\min\{\|w\|, \|x\|\} \leq 4\delta\varepsilon/\kappa$. *Then*

$$|\langle w, \bar{w}\rangle| \lesssim \nu^2\varepsilon\|\bar{w}\| \qquad \textit{and} \qquad |\langle x, \bar{x}\rangle| \lesssim \nu^2\varepsilon\|\bar{x}\|.$$

Proof. Notice

$$\left|\langle w, \frac{\bar{w}}{\|\bar{w}\|}\rangle\right| \leq \left|\langle w, \frac{\bar{w}}{\|\bar{w}\|} - U_1\rangle\right| + |\langle w, U_1\rangle| \leq \|w\| \left\|\frac{\bar{w}}{\|\bar{w}\|} - U_1\right\| + \frac{2\varepsilon}{\kappa}$$

where the last inequality follows by Cauchy-Schwartz and (6.37). A similar argument gives the same bound with $\|\bar{w}/\|\bar{w}\| + U_1\|$ instead.

By letting $A = -\bar{w}\bar{x}^\top$ and $\widehat{A} = wx^\top - \bar{w}\bar{x}^\top$ in the variant of Davis-Kahan $\sin\theta$ Theorem stated in Lemma 6.6.16 we get

$$\begin{aligned}
\min\left\{\left\|\frac{\bar{w}}{\|\bar{w}\|} + U_1\right\|, \left\|\frac{\bar{w}}{\|\bar{w}\|} - U_1\right\|\right\} &\leq \sqrt{2}\sin\left(\theta(\bar{w}/\|\bar{w}\|, U_1)\right) \\
&\leq 2\sqrt{2}\frac{(2\|\bar{w}\bar{x}^\top\| + \|wx^\top\|)}{\|\bar{w}\bar{x}^\top\|^2}\|wx^\top\| \\
&\leq 2\sqrt{2}(2+\delta)\frac{\|wx^\top\|}{\|\bar{w}\bar{x}^\top\|} \\
&\leq 2\sqrt{2}(2+\delta)\nu\frac{\varepsilon}{\|\bar{w}\|}
\end{aligned}$$

where the last inequality follows since $\|\bar{w}\| = \|\bar{x}\|$, $\|wx^\top\| \leq \varepsilon\max\{\|w\|, \|x\|\}$ and (6.39). Hence from the previous inequalities we derive

$$\left|\langle w, \frac{\bar{w}}{\|\bar{w}\|}\rangle\right| \leq \|w\|\left\|\frac{\bar{w}}{\|\bar{w}\|} - U_1\right\| + \frac{2\varepsilon}{\kappa} \leq 2\sqrt{2}(2+\delta)\nu\frac{\|w\|}{\|\bar{w}\|}\varepsilon + \frac{2\varepsilon}{\kappa} = \left(2\sqrt{2}(2+\delta)\nu^2 + \frac{2}{\kappa}\right)\varepsilon.$$

A completely analogous result holds for $|\langle x, \bar{x}\rangle|$. □

Suppose now that $\min\{\|w\|, \|x\|\} > 4\delta\varepsilon/\kappa$. In the remainder of the proof we will show that in this case, item three has to hold.

Claim 8. *The rank of $X = wx^\top - \bar{w}\bar{x}^\top$ is two.*

Proof. Assume $w = \lambda\bar{w}$, then $U_1 = \pm w/\|\bar{w}\|$. Then, (6.33) gives

$$\lambda\|\bar{w}\| \leq 2\varepsilon/\kappa \leq 4\delta\varepsilon/\kappa,$$

where we used that $\delta > 1$. This implies $\min\{\|w\|, \|x\|\} \leq 4\delta\varepsilon/\kappa$, yielding a contradiction. An analogous argument holds for $x = \lambda\bar{x}$. Thus, Lemma 6.6.8 implies that $\sigma_2(wx^\top - \bar{w}\bar{x}^\top) > 0$. □

Claim 9. $\sigma_2(Y) < \frac{\varepsilon}{\rho_1}$.

Proof. Without loss of generality suppose $\|w\| = \max\{\|w\|, \|x\|\}$. Assume seeking contradiction that this isn't true, thus $\sigma_2(Y) \geq \varepsilon/\rho_1$ then Inequality (6.36) gives $|\langle U_2, w\rangle| \leq \rho_1$. Furthermore, notice that due to Lemma 6.6.6 we have that $\|w\|^2 = \langle U_1, w\rangle^2 + \langle U_2, w\rangle^2$ and consequently $|\langle U_1, w\rangle| \geq \sqrt{\|w\|^2 - \rho_1^2}$. Again, due to (6.36)

$$\sigma_1(Y) \leq \frac{\varepsilon}{|\langle U_1, w\rangle|} \leq \frac{\varepsilon}{\sqrt{\|w\|^2 - \rho_1^2}}.$$

In turn this implies

$$\begin{aligned}
\kappa\varepsilon\rho_2 < \kappa\left\|wx^\top - \bar{w}\bar{x}^\top\right\| \leq g(w,x) - g(\bar{w},\bar{x}) &\leq \sigma_1(Y)\sigma_1(X) + \sigma_2(Y)\sigma_2(X) \\
&\leq 2\sigma_1(Y)\sigma_1(X) \\
&\leq 2\frac{\varepsilon}{\sqrt{\|w\|^2 - \rho_1^2}}|\langle U_1, w\rangle\langle x, V_1\rangle - \langle U_1, \bar{w}\rangle\langle \bar{x}, V_1\rangle| \\
&\leq 2\frac{\varepsilon}{\sqrt{\|w\|^2 - \rho_1^2}}\left(\left\|wx^\top\right\| + \left\|\bar{w}\bar{x}^\top\right\|\right) \\
&\leq \frac{2\sqrt{2}\varepsilon}{\|w\|}(1+\delta)\left\|\bar{w}\bar{x}^\top\right\|
\end{aligned}$$

Rearranging we get

$$\rho < \frac{2\sqrt{2}(1+\delta)}{\kappa}\frac{\left\|\bar{w}\bar{x}^\top\right\|}{\max\{\|w\|, \|x\|\}},$$

yielding a contradiction. □

We now need to prove an additional claim.

Claim 10. $|\langle U_2, \bar{w}\rangle| \leq |\langle U_1, \bar{w}\rangle|$ *and* $|\langle V_2, \bar{x}\rangle| \leq |\langle V_1, \bar{x}\rangle|$.

Proof. Seeking contradiction we assume the possible contrary cases.

Case 1. Assume $|\langle U_2, \bar{w}\rangle| > |\langle U_1, \bar{w}\rangle|$ and $|\langle V_2, \bar{x}\rangle| > |\langle U_1, \bar{x}\rangle|$, then (6.37) and (6.38) imply

$$\max\{|\langle U_1, w\rangle\langle V_1, x\rangle|, |\langle U_1, \bar{w}\rangle\langle V_1, \bar{x}\rangle|\} \leq \frac{2\min\{\|w\|, \|x\|\}\varepsilon}{\kappa}.$$

From which we derive

$$\kappa\varepsilon\rho_2 < g(w, x) - g(\bar{w}, \bar{x}) \leq 2\sigma_1(Y)\sigma_1(X) \leq 4\sigma_1(Y)\delta\frac{\left\|\bar{w}\bar{x}^\top\right\|}{\max\{\|w\|, \|x\|\}}\varepsilon.$$

contradicting the definition of ρ_2.

Case 2. Assume that $|\langle U_2, \bar{w}\rangle| \leq |\langle U_1, \bar{w}\rangle|$ and $|\langle V_2, \bar{x}\rangle| > |\langle V_1, \bar{x}\rangle|$. Notice that $\|\bar{w}\|^2 = \langle U_1, \bar{w}\rangle^2 + \langle U_2, \bar{w}\rangle^2$, hence $|\langle U_1, \bar{w}\rangle| \geq \|\bar{w}\|/\sqrt{2}$ and similarly $|\langle V_2, \bar{x}\rangle| > \|\bar{x}\|/\sqrt{2}$. Thus,

$$\frac{\|\bar{w}\|}{\sqrt{2}} \leq |\langle U_1, \bar{w}\rangle| \leq \frac{2\|x\|\varepsilon}{\kappa|\langle \bar{x}, V_2\rangle|} < \frac{2\sqrt{2}\|x\|\varepsilon}{\kappa\|\bar{x}\|}.$$

This implies that

$$\min\{\|w\|, \|x\|\} \leq \|w\| \leq \delta\frac{\left\|\bar{w}\bar{x}^\top\right\|}{\|x\|} < \frac{4\delta\varepsilon}{\kappa},$$

yielding a contradiction. □

Without loss of generality let us assume $\|w\| \leq \|x\|$.

Claim 11. $|\langle w, \bar{w}\rangle| \lesssim \varepsilon\|\bar{w}\|$ *and* $|\langle x, \bar{x}\rangle| \lesssim \varepsilon\|\bar{x}\|$.

Proof. By the previous claim and the fact that $\|\bar{w}\|^2 = \langle U_1, \bar{w}\rangle^2 + \langle U_2, \bar{w}\rangle^2$ we get that $|\langle U_1, \bar{w}\rangle| \geq \|w\|/\sqrt{2}$, combining this with (6.38) gives

$$|\langle \bar{x}, V_2\rangle| \leq \frac{2\sqrt{2}\|x\|\varepsilon}{\kappa\|\bar{w}\|} \leq \frac{4\delta}{\kappa}\nu\varepsilon$$

Then by Lemma 6.6.6

$$\begin{aligned}|\langle x, \bar{x}\rangle| = |\langle V_1, x\rangle\langle V_1, \bar{x}\rangle + \langle V_2, x\rangle\langle V_2, \bar{x}\rangle| &\leq |\langle V_1, x\rangle\langle V_1, \bar{x}\rangle| + |\langle V_2, x\rangle\langle V_2, \bar{x}\rangle| \\ &\leq \frac{2\varepsilon}{\kappa}\|\bar{x}\| + \|x\||\langle V_2, \bar{x}\rangle| \\ &\leq \left(\frac{2}{\kappa} + \frac{4\delta}{\kappa}\nu^2\right)\varepsilon\|\bar{x}\| \leq \left(\frac{2}{\kappa} + \frac{4\delta}{\kappa}\nu^2\right)\varepsilon\|\bar{x}\|.\end{aligned}$$

where we used (6.39). Notice that the same analysis gives

$$|\langle w, \bar{w}\rangle| \leq \left(\frac{2}{\kappa} + \frac{2\sqrt{2}\delta}{\kappa}\frac{\|w\|}{\|x\|}\right)\varepsilon\|\bar{w}\| \leq \left(\frac{2}{\kappa} + \frac{2\sqrt{2}\delta}{\kappa}\right)\varepsilon\|\bar{w}\|.$$

□

This last claim finishes the proof of the theorem. □

Proofs of Theorem 6.4.6

In order to prove the theorem we will apply three steps: we will show that the graphs of ∂f_S and ∂f_P are close, use Theorem 6.6.12 to give bounds on the ϵ-critical points of f_P, and then connect it back to the landscape of f_S by combining the previous two steps. The following two propositions handle the first part.

Proposition 6.6.13. *Fix two functions $f, g : \mathbf{R}^{d_1} \times \mathbf{R}^{d_2} \to \mathbf{R}$ such that g is ρ-weakly convex. Suppose that there exists a point $(\bar{w}, \bar{x})$ and a real $\delta > 0$ such that the inequality*

$$|f(w, x) - g(w, x)| \leq \delta \left\|wx^\top - \bar{w}\bar{x}^\top\right\|_F \qquad \text{holds for all } (w, x) \in \mathbf{R}^{d_1} \times \mathbf{R}^{d_2}.$$

Then for any stationary point (w, x) of g, there exists a point $(\widehat{w}, \widehat{x})$ satisfying

$$\begin{cases} \left\|(w, x) - (\widehat{w}, \widehat{x})\right\| & \leq 2\sqrt{\frac{\delta\|wx^\top - \bar{w}\bar{x}^\top\|}{\rho+\delta}} \\ \left\|\operatorname{dist}(0, \partial f(\widehat{w}, \widehat{x}))\right\| & \leq \left(\delta + \sqrt{2\delta(\rho+\delta)}\right)(\|(w, x)\| + \|(\bar{w}, \bar{x})\|)\,. \end{cases}$$

Proof. The proposition is a corollary of Theorem 6.1 of [70]. Recall that for a function $l : \mathbf{R}^d \to \overline{\mathbf{R}}$ the Lipschitz constant at $\bar{y} \in \mathbf{R}^d$ is given by

$$\operatorname{lip}(l, y) := \limsup_{y \to \bar{y}} \frac{|l(y) - l(\bar{y})|}{|y - \bar{y}|}.$$

Set $u(x) = \delta \left\| wx^\top - \bar{w}\bar{x}^\top \right\|$, and $l(x) = -\delta \left\| wx^\top - \bar{w}\bar{x}^\top \right\|$. It is easy to see that at differentiable points the gradient of $l(\cdot)$ is equal to

$$\nabla l(w, x) = -\frac{\delta}{\|wx^\top - \bar{w}\bar{x}^\top\|_F} \begin{bmatrix} (wx^\top - \bar{w}\bar{x}^\top)x \\ (wx^\top - \bar{w}\bar{x}^\top)^\top w \end{bmatrix} \implies \|\nabla l(w, x)\| \leq \delta \|(w, x)\|$$

Then, since $\operatorname{lip}(l; w, x) = \limsup_{(w', x') \to (w, x)} \|\nabla l(w', x')\|$, we can over estimate

$$\operatorname{lip}(l; w, x) \leq \delta \left(\|(w, x)\| + \|(\bar{w}, \bar{x})\| \right).$$

Thus applying Theorem 6.1 of [70] we get that for all $\gamma > 0$ there exists $(\widehat{w}, \widehat{x})$ such that $\|(w, x) - (\widehat{w}, \widehat{x})\| \leq 2\gamma$ and

$$\operatorname{dist}(0, \partial f(\widehat{w}, \widehat{x})) \leq 2\rho\gamma + 2\delta \frac{\left\| wx^\top - \bar{w}\bar{x}^\top \right\|}{\gamma} + \delta \left(\left\| (\widehat{w}, \widehat{x}) \right\| + \|(\bar{w}, \bar{x})\| \right)$$

By the triangular inequality we get $\left\| (\widehat{w}, \widehat{x}) \right\| \leq 2\gamma + \|(w, x)\|$ and therefore

$$\operatorname{dist}(0, \partial f(\widehat{w}, \widehat{x})) \leq 2(\rho + \delta)\gamma + 2\delta \frac{\left\| wx^\top - \bar{w}\bar{x}^\top \right\|}{\gamma} + \delta \|(w, x)\|.$$

Hence setting $\gamma = \sqrt{\frac{\delta \|wx^\top - \bar{w}\bar{x}^\top\|}{\rho + \delta}}$, gives

$$\begin{aligned}
\operatorname{dist}(0, \partial f(\widehat{w}, \widehat{x})) &\leq 2\sqrt{\delta(\rho + \delta) \|wx^\top - \bar{w}\bar{x}^\top\|} + \delta \left(\left\| (\widehat{w}, \widehat{x}) \right\| + \|(\bar{w}, \bar{x})\| \right) \\
&\leq 2\sqrt{\delta(\rho + \delta) \left(\|wx^\top\| + \|\bar{w}\bar{x}^\top\| \right)} + \delta \left(\left\| (\widehat{w}, \widehat{x}) \right\| + \|(\bar{w}, \bar{x})\| \right) \\
&\leq 2\sqrt{\delta(\rho + \delta)} \left(\sqrt{\|wx^\top\|} + \sqrt{\|\bar{w}\bar{x}^\top\|} \right) + \delta \left(\left\| (\widehat{w}, \widehat{x}) \right\| + \|(\bar{w}, \bar{x})\| \right) \\
&\leq \left(\delta + \sqrt{2\delta(\rho + \delta)} \right) \left(\|(w, x)\| + \|(\bar{w}, \bar{x})\| \right)
\end{aligned}$$

where we used that $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ and $ab \leq (a^2 + b^2)/2$. □

Proposition 6.6.14. *There exist numerical constants $c_1, c_2 > 0$ such that for all $(w, x) \in \mathbf{R}^{d_1 \times d_2}$ we have*

$$\left| f_S(w,x) - f_P(w,x) \right| \lesssim \left(\frac{d_1 + d_2 + 1}{m} \log\left(\frac{m}{d_1 + d_2 + 1} \right) \right)^{\frac{1}{2}} \|wx^\top - \bar{w}\bar{x}^\top\| \tag{6.40}$$

with probability at least $1 - 2\exp(-c_1(d_1 + d_2 + 1))$ provided $m \geq c_2(d_1 + d_2 + 1)$.

Proof. The proof of this proposition is almost entirely analogous to the proof of Proposition 5.10.2 with $\mathcal{I} = \emptyset$, $\beta(r) \asymp 1$, $c(m, r) \asymp m$, and $r = 1$ (note that this setting satisfies the assumptions thanks to Lemma 5.10.6). The proof is exactly the same up to (5.97). Instead of repeating the proof, we refer the reader to Section 5.10.2. Up to (5.97) we had proved the following result:

Claim 12. *There exists constants $c_1, c_2, c_3, c_4 > 0$ such that for any $t \in (0, c_4)$ the following uniform concentration bound holds*

$$\left| f_S(w,x) - f_P(w,x) \right| \leq \frac{3}{2} t \|wx^\top - \bar{w}\bar{x}^\top\|_F \quad \textit{for all } (w,x) \in \mathbf{R}^{d_1 \times d_2}$$

with probability at least $1 - 2\exp(c_1(d_1 + d_2 + 1)\log(c_2/t) - c_3 t^2 m)$.

This probability bound is at least $1 - 2\exp(-c_3 t^2 m/2)$ provided that

$$\frac{d_1 + d_2 + 1}{m} \leq \frac{c_3 t^2}{2c_1 \log(c_2/t)}. \tag{6.41}$$

Set $t = \max\left(\sqrt{\frac{2c_1}{c_3}}, c_2 \right) \left(\frac{d_1+d_2+1}{m} \log\left(\frac{m}{d_1+d_2+1} \right) \right)^{\frac{1}{2}}$. This choice ensures that (6.41) holds, since

$$\frac{d_1 + d_2 + 1}{m} \leq \frac{(d_1 + d_2 + 1)\log\left(\frac{m}{d_1+d_2+1} \right)}{m \log\left(\frac{m}{d_1+d_2+1} \log^{-1}\left(\frac{m}{d_1+d_2+1} \right) \right)} \leq \frac{c_3 t^2}{2c_1 \log(c_2/t)},$$

where we used that the function $\frac{\log(x)}{\log(x/\log(x))} \geq 1$ for all $x \geq e$. We guarantee that this holds for $x = m/(d_1+d_2+1)$ and that $t \in (0, c_4)$ by setting $m \geq C(d_1+d_2+1)$ with C sufficiently large. After relabeling the constants, this proves the result. □

We are finally in position to proof the theorem.

Proof of Theorem 6.4.6. Fix $v \geq 1$ and a fix point (w, x) satisfying $\|(w, x)\| \leq v\|(\bar{w}, \bar{x})\|$. Proposition 6.6.14 shows that there exist constants $c_1, c_2 > 0$ such that with probability at least $1 - 2\exp(-c_1(d_1 + d_2 + 1))$ we have

$$|f_S(w, x) - f_P(w, x)| \leq \tilde{O}\left(\left(\frac{d_1 + d_2 + 1}{m}\right)^{\frac{1}{2}}\right)\|wx^\top - \bar{w}\bar{x}^\top\|_F \qquad \forall (w, x) \in \mathbf{R}^{d_1} \times \mathbf{R}^{d_2}$$

provided that $m \geq c_2(d_1 + d_2 + 1)$. To ease the notation let us denote $\Delta := \tilde{O}\left(\left(\frac{d_1+d_2+1}{m}\right)^{\frac{1}{2}}\right)$. Assume that we are in the event in which this holds. As described in the introduction of this chapter, we showed in Chapter 5 that f_S is ρ-weakly and μ-sharp with high probability provided that $m \geq C(d_1 + d_2 + 1)$. Now, assume that m is big enough and we are in the intersection of this two events. This holds with probability $1 - c_3\exp(c_1(d_1 + d_2 + 1))$ (for some possibly different constants c_1, c_3). Hence by Proposition 6.6.13 there exits a point $(\widehat{w}, \widehat{x})$ such that

$$\|(w, x) - (\widehat{w}, \widehat{x})\| \leq \frac{2}{\sqrt{\rho}}\sqrt{\Delta}D_{wx} \qquad \text{and} \qquad \operatorname{dist}(0, \partial f(\bar{w}, \bar{x})) \leq C\sqrt{\Delta}D_{wx}$$

where $D_{wx} = \|(w, x)\| + \|(\bar{w}, \bar{x})\|$.

Notice that if $\|(w, x)\| \leq \Delta^{\frac{1}{4}}\|(\bar{w}, \bar{x})\|$ holds then the result holds immediately. So assume that this inequality is not satisfied. So we can lower bound

$$\|(\widehat{w}, \widehat{x})\| \geq \|(w, x)\| - \|(\widehat{w}, \widehat{x}) - (w, x)\| \geq \left(1 - 2\left(\frac{\bar{\Delta}}{\rho}\right)^{\frac{1}{2}}\left(1 + \bar{\Delta}^{-\frac{1}{4}}\right)\right)\|(w, x)\| \geq \frac{1}{2}\|(w, x)\|$$

where the first inequality follow by applying the triangle inequality and the last inequality follows for m sufficiently large, since we can ensure that for such m

the term in the parenthesis is bigger than $1/2$. Therefore,

$$\begin{aligned}\operatorname{dist}(0, \partial f(\widehat{w}, \widehat{x})) &\leq C\Delta^{\frac{1}{2}} \left(\|(w, x)\| + \|(\bar{w}, \bar{x})\|\right) \\ &\leq C\Delta^{\frac{1}{2}} \left(1 + \Delta^{-\frac{1}{4}}\right) \|(w, x)\| \\ &\leq 2C\Delta^{\frac{1}{2}} \left(1 + \Delta^{-\frac{1}{4}}\right) \|(\widehat{w}, \widehat{x})\| \\ &\leq 4C\Delta^{\frac{1}{4}} \|(\widehat{w}, \widehat{x})\|.\end{aligned}$$

Hence, by reducing Δ if necessary we can guarantee that $\operatorname{dist}(0, \partial f(\widehat{w}, \widehat{x})) \leq \gamma\|(\widehat{w}, \widehat{x})\|$ and consequently Theorem 6.6.12 gives that at least one of the following two holds

$$\max\{\|\widehat{w}\|, \|\widehat{x}\|\}\|\widehat{w}\widehat{x}^\top - \bar{w}\bar{x}^\top\| \lesssim \Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\bar{x}^\top\| \qquad \text{and} \qquad \begin{cases} |\langle w, \bar{w}\rangle| & \lesssim \nu^2 \Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\| \\ |\langle x, \bar{x}\rangle| & \lesssim \nu^2 \Delta^{\frac{1}{2}} D_{wx} \|\bar{x}\| \end{cases} \tag{6.42}$$

Let us prove that this implies the statement of the theorem.

Case 1. Assume that the second condition in (6.42) holds. Notice that due to $\|(w, x)\| \leq \Delta^{\frac{1}{4}} \|(\bar{w}, \bar{x})\|$ we have $\Delta^{\frac{1}{2}} D_{wx} \lesssim \Delta^{\frac{1}{4}} \|(w, x)\|$ for m big enough. This implies

$$|\langle w, \bar{w}\rangle| \leq \left|\langle \widehat{w}, \bar{w}\rangle\right| + \|\bar{w}\|\|\widehat{w} - w\| \lesssim (\nu^2 + 1)\Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\| \lesssim (\nu^2 + 1)\Delta^{\frac{1}{4}} \|(w, x)\|\|\bar{w}\|.$$

A similar argument yields the result for $|\langle w, x\rangle|$.

Case 2. On the other hand, if the first condition holds, there exist $e_w \in$

$\mathbf{R}^{d_1}, e_x \in \mathbf{R}^{d_2}$ such that $\widehat{w} = w + e_w$ and $\widehat{x} = x + e_x$ with $\|e_w\|, \|e_x\| \leq \Delta^{\frac{1}{2}} D_{wx}$. Then

$$\begin{aligned}
\|(w,x)\|\|wx^\top - \bar{w}\bar{x}^\top\| &\leq \|(w,x)\|\|wx^\top - \widehat{w}\widehat{x}^\top\| + \|(w,x)\|\|\widehat{w}\widehat{x}^\top - \bar{w}\bar{x}^\top\| \\
&\leq \|(w,x)\|\|wx^\top - \widehat{w}\widehat{x}^\top\| + 2\|(\widehat{w},\widehat{x})\|\|\widehat{w}\widehat{x}^\top - \bar{w}\bar{x}^\top\| \\
&\lesssim \|(w,x)\|\|wx^\top - (w+e_w)(x+e_x)^\top\| + \Delta^{\frac{1}{2}} D_{wx}\|\bar{w}\bar{x}^\top\| \\
&\leq \|(w,x)\|\left(\|we_x^\top\| + \|e_w x^\top\| + \|e_w e_x^\top\|\right) + \Delta^{\frac{1}{2}} D_{wx}\|\bar{w}\bar{x}^\top\| \\
&\leq \|(w,x)\|\Delta^{\frac{1}{2}} D_{wx}\left(\|w\| + \|x^\top\| + \Delta^{\frac{1}{2}} D_{wx}\right) + \Delta^{\frac{1}{2}} D_{wx}\|\bar{w}\bar{x}^\top\| \\
&\lesssim \|(w,x)\|\Delta^{\frac{1}{2}} D_{wx}\left(\|(w,x)\| + \Delta^{\frac{1}{4}}\|(w,x)\|\right) + \Delta^{\frac{1}{2}} D_{wx}\|\bar{w}\bar{x}^\top\| \\
&\lesssim \|(w,x)\|^2\Delta^{\frac{1}{2}} D_{wx} + \Delta^{\frac{1}{2}} D_{wx}\|\bar{w}\bar{x}^\top\| \\
&\lesssim (\nu^2+1)\|\bar{w}\bar{x}^\top\|\Delta^{\frac{1}{2}} D_{wx} \\
&\lesssim (\nu^2+1)\Delta^{\frac{1}{4}}\|(w,x)\|\|\bar{w}\bar{x}^\top\|.
\end{aligned}$$

Proving the desired result. □

6.6.4 Auxiliary results

This subsection presents technical lemmas we employed in our proofs. The first result we need is a special case of the celebrated Davis-Kahan $\sin\theta$ Theorem (see [65]). For any two unit vectors $u_1, v_1 \in \mathbb{S}^{d-1}$, define $\theta(u_1, v_1) = \cos^{-1}(|\langle u_1, v_1\rangle|)$.

Lemma 6.6.15. *Consider symmetric matrices $X, \Delta, Z \in \mathbf{R}^{n\times n}$, where $Z = X + \Delta$. Define δ to be the eigengap $\lambda_1(X) - \lambda_2(X)$, and denote the first eigenvectors of X, Z by u_1, v_1, respectively. Then*

$$\frac{1}{\sqrt{2}}\min\left\{\|u - v\|_2, \|u + v\|_2\right\} \leq \sqrt{1 - \langle u_1, v_1\rangle^2} = |\sin\theta(u_1, v_1)| \leq \frac{\|\Delta\|_{\mathrm{op}}}{\delta}.$$

We also need a version of this result for rectangular matrices.

Lemma 6.6.16. *Let $A, \widehat{A} \in \mathbf{R}^{d_1 \times d_2}$ with* $\operatorname{rank}(A) = 1$. *Let $A = U\sigma(A)V^\top$ and $\widehat{A} = \widehat{U}\sigma(\widehat{A})\widehat{V}^\top$ be their singular value decompositions. Then the*

$$\sin\theta(V_1, \widehat{V}_1) \leq \frac{2(2\sigma_1(A) + \|A - \widehat{A}\|_{\text{op}})}{\|A\|_{\text{op}}^2}\|A - \widehat{A}\|_F,$$

the same bound holds for $U_1, \widehat{U}_1$.

Proof. This is a corollary of Theorem 3 in [246]. For any pair of matrices $B, \widehat{B}$ said Theorem establishes the following upper bound

$$|\sin\theta(V_1(B), V_1(\widehat{B}))| \leq \frac{2(2\sigma_1(B) + \|B - \widehat{B}\|_{\text{op}})}{\sigma_1(B)^2 - \hat{\sigma}_2(B)}\|B - \widehat{B}\|_F \tag{6.43}$$

where $V_1(X)$ is the top right singular value of the matrix X; and $\hat{\sigma}_2(B) = \sigma_2(B)$ if $\operatorname{rank}(B) > 1$ and $\hat{\sigma}_2(B) = -\infty$ otherwise.

Fix U_2 and V_2 such that $U_2 \perp U_1$ and $V_2 \perp V_1$. For any n define

$$A_n := \sigma_1(A)U_1V_1^\top + \frac{\sigma_1(A)}{2n}U_2V_2^\top.$$

By construction $V_1(A) = V_1(A_n)$ for all n. Then, applying (6.43) we get

$$|\sin\theta(V_1, \widehat{V}_1)| \leq \frac{2(2\sigma_1(A) + \|A - \widehat{A}\|_{\text{op}})}{\sigma_1(A)^2}\|A - \widehat{A}\|_F + e_n$$

where $(e_n)_{n=1}^\infty$ is a sequence satisfying $\lim_{n\to\infty} e_n = 0$. Taking limits gives the desired bound. An analogous argument gives the result for $U, \hat{U}$. □

Now, we provide a few well-known concentration inequalities.

Theorem 6.6.17 (Corollary 2.8.3 in [235])**.** *Let $X_1, \ldots, X_m$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m X_i\right| \geq t\right) \leq 2\exp\left[-cm\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right]$$

where $c > 0$ is a numerical constant and $K := \max_i \|X_i\|_{\psi_1}$.

Theorem 6.6.18 (Theorem 5.6 in [25])**.** *Let $X = (X_1, \ldots, X_m)$ be a vector of n independent standard normal random variables. Let $f : \mathbf{R}^n \to \mathbf{R}$ denote an L-Lipschitz function. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(f(X) - \mathbb{E}f(X) \geq t\right) \leq \exp\left(-\frac{t^2}{2L^2}\right).$$

The following concentration inequalities deal with quantiles of distributions:

Lemma 6.6.19. *Let $X_1, \ldots, X_m$ be an i.i.d. sample with distribution $\mathcal{D}$, choose Q_q to be the q population quantile of the distribution $\mathcal{D}$, that is $q = \mathbb{P}\left(X_1 \leq Q_q\right)$, and let $p \in (0, 1)$ be any probability with $p < q$. Then,*

$$\mathbb{P}\left(\texttt{quant}_p(\{X_i\}_{i=1}^m) \geq Q_q\right) \leq \exp\left(\frac{m(q-p)^2}{2(q-p)/3 + 2q(1-q)}\right),$$

where $\texttt{quant}_p(\{X_i\}_{i=1}^m)$ denotes the p-th quantile of the sample $\{X_i\}$.

Proof. It is easy to see that the following holds, $\texttt{quant}_p(\{X_i\}_i^m) \geq Q_q$ if, and only if, $\frac{1}{m}\sum_{i=1}^m \mathbf{1}\{X_i \leq Q_q\} \leq p$. Notice that $\mathbf{1}\{X_i \leq Q_q\} \sim \mathrm{B}(q)$ are i.i.d. Bernoulli random variables and thus $\mathrm{Var}(\mathbf{1}\{X_i \leq Q_q\}) = q(1-q)$. Then, the result follows by applying Bernstein's inequality (Theorem 2.3.2) to $\frac{1}{m}\sum \mathbf{1}\{X_i \leq Q_q\} - q$. □

Lemma 6.6.20. *Let a, b be i.i.d. sub-gaussian random variables. For any $Q > 0$ such that $q := \mathbb{P}\left(|ab| \leq Q\right) > 1/2$, consider the random variable c^2 defined as a^2 conditioned on the event $|ab| \leq Q$, namely for all t*

$$\mathbb{P}\left(c^2 \leq t\right) = \mathbb{P}\left(a^2 \leq t \mid |ab| \leq Q\right).$$

Then, c^2 is a sub-exponential random variable, in other words for all $t \geq 0$ we have that

$$\mathbb{P}\left(c^2 \geq t\right) \leq 2\exp(-t/2K)$$

where K is the minimum scalar such that $\mathbb{P}\left(a^2 \geq t\right) \leq 2\exp(-t/K)$.

Proof. Let us consider two cases. Suppose first $t \leq 2K \log 2$. Then we have that $1 \leq 2\exp(-t/2K)$ and therefore the stated inequality is trivial.

Suppose now $t \geq 2K \log 2$. Then we have that

$$\frac{t}{2K} \geq \log 2 \iff \exp(t/K - t/2K) \geq 2 \iff \exp(-t/2K) \geq 2\exp(-t/K).$$

With this we can bound the probability

$$\begin{aligned}\mathbb{P}\left(c^2 \geq t\right) = \frac{1}{q}\mathbb{P}\left(a^2\mathbf{1}\{|ab| \leq Q\} \geq t\right) \leq \frac{1}{q}\mathbb{P}\left(a^2 \geq t\right) \leq \frac{2}{q}\exp(-t/K) \\ \leq 4\exp(-t/K) \leq 2\exp(-t/2K),\end{aligned}$$

as claimed. □

The following Theorem from [234] is especially useful in bounding the operator norm of random matrices:

Theorem 6.6.21 (Operator norm of random matrices)**.** *Consider an $m \times n$ matrix A whose rows A_i are independent, sub-gaussian, isotropic random vectors in $\mathbf{R}^n$. Then, for every $t \geq 0$, one has*

$$\mathbb{P}\left(\left\|\frac{1}{m}AA^\top - I_n\right\|_{\text{op}} \leq C\sqrt{\frac{n}{m} + t}\right) \geq 1 - 2\exp\left(-cmt\right),$$

where C depends only on $K := \max_i \|A_i\|_{\psi_2}$.

Proof. The Theorem is a direct Corollary of [234, Theorem 5.39]. Specifically, the concavity of the square root gives us $\sqrt{a} + \sqrt{b} \leq \sqrt{2}\sqrt{a+b}$, implying that

$$C\sqrt{\frac{n}{m}} + \sqrt{\frac{t}{m}} \leq C\sqrt{2}\sqrt{\frac{n}{m} + \frac{t}{m}}.$$

Additionally, [234, Theorem 5.39] gives us that

$$\mathbb{P}\left(\left\|\frac{1}{m}AA^\top - I_n\right\|_{\text{op}} \leq C\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\right) \geq 1 - 2\exp\left(-ct^2\right).$$

Setting $t' = C\sqrt{mt}$ and a bit of relabeling, along with the square root inequality, gives us the desired inequality. □

Let us record the following elementary consequence.

Corollary 6.6.22. *Let $a_1, \dots, a_m \in \mathbf{R}^d$ be independent, sub-gaussian, isotropic random vectors in $\mathbf{R}^n$ and let $\mathcal{I} \subset \{1, \dots, m\}$ be an arbitrary set. Then, for every $t \geq 0$, one has*

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i\in\mathcal{I}}(a_i a_i^\top - I_d)\right\|_{\mathrm{op}} \leq C\sqrt{\frac{d}{m} + t}\right) \geq 1 - 2\exp\left(-cmt\right),$$

where C depends only on $K := \max_i \|A_i\|_{\psi_2}$.

Proof. Consider the matrix $A \in \mathbf{R}^{|\mathcal{I}|\times d}$ whose rows are the vectors a_i for $i \in \mathcal{I}$. Then we deduce

$$\left\|\frac{1}{m}\sum_{i\in\mathcal{I}}(a_i a_i^\top - I_d)\right\|_{\mathrm{op}} = \frac{|\mathcal{I}|}{m}\left\|\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} a_i a_i^\top - I_d\right\|_{\mathrm{op}} = \frac{|\mathcal{I}|}{m}\left\|\frac{1}{|\mathcal{I}|}AA^\top - I_d\right\|_{\mathrm{op}}.$$

Appealing to 6.6.21, we therefore deduce for any $\gamma > 0$ the estimate

$$\left\|\frac{1}{m}\sum_{i\in\mathcal{I}}(a_i a_i^\top - I_d)\right\|_{\mathrm{op}} \leq \frac{|\mathcal{I}|}{m}\sqrt{\frac{d}{|\mathcal{I}|} + \gamma} \leq C\sqrt{\frac{d|\mathcal{I}|}{m^2} + \frac{\gamma|\mathcal{I}|^2}{m^2}},$$

holds with probability $1 - 2\exp(-c|\mathcal{I}|\gamma)$. Now for any $t > 0$, choose γ such that, $\frac{d|\mathcal{I}|}{m^2} + \frac{\gamma|\mathcal{I}|^2}{m^2} = \frac{d}{m} + t$, namely $\gamma = \frac{m^2}{|\mathcal{I}|^2}[\frac{d}{m}(1 - \frac{|\mathcal{I}|}{m}) + t]$. Noting

$$|\mathcal{I}|\gamma = m \cdot \frac{m}{|\mathcal{I}|}\left[\frac{d}{m}\left(1 - \frac{|\mathcal{I}|}{m}\right) + t\right] \geq mt,$$

completes the proof. □

Recall that we defined the functions $q_{\mathrm{fail}}(p_{\mathrm{fail}}) = \frac{5-2p_{\mathrm{fail}}}{8(1-p_{\mathrm{fail}})}$ and $Q_{\mathrm{fail}}(q_{\mathrm{fail}})$ given as the q_{fail}-quantile of $|ab|$ where a, b are i.i.d. standard normal. Furthermore we defined $\omega_{\mathrm{fail}} = \mathbb{E}[a^2 \mid |ab| \leq Q_{\mathrm{fail}}]$.

Lemma 6.6.23. *The function* $\omega : [0,1] \to \mathbf{R}_+$ *given by*

$$p_{\text{fail}} \mapsto \mathbb{E}[a^2 \mid |ab| \leq Q_{\text{fail}}]$$

is nondecreasing. In particular, there exist numerical constants $c_1, c_2 > 0$ *such that for any* $0 \leq p_{\text{fail}} \leq 0.1$ *we have*

$$c_1 \leq \omega_{\text{fail}} \leq c_2,$$

where the tightest constants are given by $c_1 = \omega(0) \geq 0.5$ *and* $c_2 = \omega(0.1) \leq 0.56$.

Proof. The bulk of this result is contained in the following claim.

Claim 13. *Let* $0 \leq Q \leq Q'$ *be arbitrary numbers, then*

$$\mathbb{P}(a^2 \geq t \mid |ab| \leq Q) \leq \mathbb{P}(a^2 \geq t \mid |ab| \leq Q') \qquad \forall t \in \mathbf{R}.$$

We defer the proof of the claim and show how it implies the lemma. Observe that the functions $p_{\text{fail}} \mapsto q_{\text{fail}}$ and $q_{\text{fail}} \mapsto Q_{\text{fail}}$ are nondecreasing, thus it suffices to show that the function $Q \mapsto \mathbb{E}[a^2 \mid |ab| \leq Q]$ is nondecreasing. Let $0 \leq Q \leq Q'$

$$\begin{aligned}
\mathbb{E}[a^2 \mid |ab| \leq Q] &= \int_0^\infty \mathbb{P}(a^2 \geq t \mid |ab| \leq Q) dt \\
&\leq \int_0^\infty \mathbb{P}(a^2 \geq t \mid |ab| \leq Q') dt \\
&= \mathbb{E}[a^2 \mid |ab| \leq Q'],
\end{aligned}$$

where the inequality follows from the claim and the equalities follow from the identity $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t) dt$ for nonnegative random variables X. Hence ω is a nondecreasing function.

The above implies that for any $p_{\text{fail}} \in [0, 0.1]$ we have $\omega(0) \leq \omega_{\text{fail}} \leq \omega(0.1)$. Note that $\omega(0)$ is positive since it is defined by a positive integrand on a set of non-negligible measure. The bounds on $\omega(0)$ and $\omega(0.1)$ follow by a numerical

computation. In particular we obtain that with $Q = 0.6$ the probability $\mathbb{P}(|ab| \leq Q) \geq 0.6679 \geq 2/3 = q_{\text{fail}}(0.1)$. Then computing numerically (with precision set to 32 digits) we obtain $\omega(0.1) \leq \mathbb{E}[a^2 \mid |ab| \leq Q] \leq 0.56$. Similarly we find that if we set $Q = 0.5$ we get $\mathbb{P}(|ab| \leq Q) \leq 0.5903 \leq 5/8 = q_{\text{fail}}(0)$. Then evaluating we find $\omega(0) \geq \mathbb{E}[a^2 \mid |ab| \leq Q] \geq 0.5$.

Proof of the claim. The statement of the claim is equivalent to having that for any $t \in \mathbf{R}_+$ the function $h_t : \mathbf{R}_+ \to \mathbf{R}$ given by

$$Q \mapsto \frac{\mathbb{P}(a^2 \leq t; |ab| \leq Q)}{\mathbb{P}(|ab| \leq Q)}$$

is nonincreasing. Our goal is to show that $h_t' \leq 0$. In order to prove this result we proceed as follows. Define

$$g(Q) := \frac{\pi}{2}\mathbb{P}(|ab| \leq Q) = \int_0^\infty \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx,$$

and

$$f_t(Q) := \frac{\pi}{2}\mathbb{P}(a^2 \leq t; |ab| \leq Q) = \int_0^{\sqrt{t}} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx.$$

Observe $h_t = f_t/g$. Thus it suffices to show $f_t' g - f_t g' \leq 0$. Invoking Leibniz rule we get

$$\begin{aligned} f_t'(Q) &= \frac{\partial}{\partial Q} \int_0^{\sqrt{t}} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx \\ &= \int_0^{\sqrt{t}} \frac{\partial}{\partial Q} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx \\ &= \int_0^{\sqrt{t}} \frac{1}{x} \exp(-(x^2 + Q^2/x^2)/2)\, dx. \end{aligned}$$

Repeating the same procedure we get $g'(Q) = \int_0^\infty \frac{1}{x} \exp(-(x^2 + Q^2/x^2)/2)\, dx$. Some algebra reveals we want to show

$$\xi(t) := \frac{\left(\int_0^{\sqrt{t}} \frac{1}{x} \exp(-(x^2 + Q^2/x^2)/2)\, dx\right)}{\left(\int_0^{\sqrt{t}} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx\right)} \leq \frac{\left(\int_0^{\infty} \frac{1}{x} \exp(-(x^2 + Q^2/x^2)/2)\right)}{\left(\int_0^{\infty} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx\right)}.$$

It is enough to show that the function $\xi(t)$ is monotonically increasing. Define

$$\zeta_Q(t) = \int_0^{\sqrt{t}} \frac{1}{x} \exp(-(x^2+Q^2/x^2)/2)\, dx \qquad \text{and} \qquad \psi_Q(t) = \int_0^{\sqrt{t}} \int_0^{Q/x} \exp(-(x^2+y^2)/2)\, dy\, dx,$$

Thus we have

$$\zeta'_Q(t) = \frac{1}{2t} \exp(-(t + Q^2/t)/2) \qquad \text{and} \qquad \psi'_Q(t) = \frac{1}{2\sqrt{t}} \int_0^{Q/\sqrt{t}} \exp(-(t + y^2)/2) dy.$$

Again, $\xi(t) = \zeta_Q(t)/\psi_Q(t)$, hence we need to show $\zeta'_Q\psi_Q \geq \zeta_Q\psi'_Q$. After some algebra, this amounts to proving

$$\left(\int_0^{\sqrt{t}} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx \right) \geq \left(\int_0^{\sqrt{t}} \frac{\sqrt{t}}{x} \exp(-(Q^2/x^2 - Q^2/t)/2) \int_0^{Q/\sqrt{t}} \exp(-(x^2 + y^2)/2) dy\, dx \right).$$

The inequality is true if in particular the same holds for the integrands, i.e.

$$\int_0^{Q/x} \exp(-y^2/2)\, dy \geq \frac{\sqrt{t}}{x} \exp\left(-\left(\frac{Q^2}{x^2} - \frac{Q^2}{t} \right)/2 \right) \int_0^{Q/\sqrt{t}} \exp(-y^2/2)\, dy.$$

Since $x \leq \sqrt{t}$, the previous inequality holds if

$$x \mapsto \frac{1}{x} \frac{\exp\left(-\frac{Q^2}{2x^2}\right)}{\int_0^{Q/x} \exp(-y^2/2) dy}$$

is increasing. By taking derivatives and reordering terms we see that this is equivalent to

$$\frac{Q - x^2}{Qx} \int_0^{Q/x} \exp(-y^2/2) dy + \exp(-Q^2/2x^2) \geq 0.$$

Since $\exp(-y^2/2)$ is decreasing, we have

$$\frac{Q - x^2}{qx} \int_0^{Q/x} \exp(-y^2/2) dy \geq \frac{Q - x^2}{Qx} \frac{Q}{x} \exp(-Q^2/2x^2) \geq -\exp(-Q^2/2x^2)$$

proving the claim. □

Thus the proof is complete. □

7

ESCAPING STRICT SADDLE POINTS OF WEAKLY-CONVEX FUNCTIONS EFFICIENTLY

"... y alcanzó a decirle con el último aliento:

- Sólo Dios sabe cuánto te quise."

— Gabriel García Márquez, *El amor en los tiempos del cólera*

7.1 Introduction

Though nonconvex optimization problems are NP-hard in general, simple nonconvex optimization techniques, e.g., gradient descent, are broadly used and often highly successful in high-dimensional statistical estimation and machine learning problems. A common explanation for their success is that *smooth* nonconvex functions $g\colon \mathbf{R}^d \to \mathbf{R}$ found in machine learning have amenable geometry: all local minima are (nearly) global minima and all saddle points are strict (i.e., have a direction of negative curvature).

This explanation is well grounded: several important estimation and learning problems have amenable geometry [106, 226, 23, 105, 227, 240], and simple randomly initialized iterative methods, such as gradient descent, asymptotically avoid strict saddle points [142, 141]. Moreover, "randomly perturbed" variants [123] "efficiently" converge to $(\varepsilon_1, \varepsilon_2)$-*approximate second-order critical points*, meaning those satisfying

$$\|\nabla g(x)\| \le \varepsilon_1 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 g(x)) \ge -\varepsilon_2. \tag{7.1}$$

Recent work furthermore extends these results to C^2 *smooth* manifold constrained optimization [58, 229, 104]. Other extensions to *nonsmooth* convex constraint sets have proposed *second-order* methods for avoiding saddle points, but such methods must *at every step* minimize a nonconvex quadratic over a convex set (an NP hard problem in general) [111, 178, 191].

While impressive, the aforementioned works crucially rely on smoothness of objective functions or constraint sets. This is not an artifact of their proof techniques: there are simple C^1 functions for which randomly initialized gradient descent with constant probability converges to points that admit directions of second order descent [67, Figure 1]. Despite this example, recent work [67] shows that randomly initialized *proximal methods* avoid certain "active" strict saddle points of (nonsmooth) *weakly convex* functions. The class of weakly convex functions is broad, capturing, for example those formed by composing convex functions h with smooth nonlinear maps c, which often appear in statistical recovery problems. They moreover show that for "generic" semialgebraic problems, every critical point is either a local minimizer or an active strict saddle. A key limitation of [67], however, is that the result is asymptotic, and in fact pure proximal methods may take exponentially many iterations to find local minimizers [87]. Motivated by [67], the recent work [119] develops efficiency estimates for certain randomly perturbed proximal methods. The work [119] has two limitations: its measure of complexity appears to be algorithmically dependent and the results do not extend to subgradient methods.

The purpose of this Chapter is to develop "efficient" methods for escaping saddle points of weakly convex functions. Much like [119], our approach is based on [67], but the resulting algorithms and their convergence guarantees

are distinct from those in [119]. We begin with a useful observation from [67]: near active strict saddle points $\bar{x}$, a certain C^1 smoothing, called the *Moreau envelope*, is C^2 and has a strict saddle point at $\bar{x}$. If one could *exactly* execute the perturbed gradient method of [123], efficiency guarantees would then immediately follow. While this is not possible in general, it is possible to *inexactly* evaluate the gradient of the Moreau envelope by solving a strongly convex optimization problem. Leveraging this idea, we extend the work [123] to allow for inexact gradient evaluations, proving similar efficiency guarantees.

Setting the stage, we consider a minimization problem

$$\text{minimize}_{x\in\mathbf{R}^d} f(x) \tag{7.2}$$

where $f\colon \mathbf{R}^d \to \mathbf{R}\cup\{+\infty\}$ is closed and ρ-weakly convex, meaning the mapping $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex. Although such functions are nonsmooth in general, they admit a global C^1 smoothing furnished by the Moreau envelope. For all $\mu < \rho^{-1}$, the *Moreau envelope* and the *proximal mapping* are defined to be

$$f_\mu(x) = \min_{y\in\mathbf{R}^d}\ f(y) + \frac{1}{2\mu}\|y-x\|^2 \quad \text{and} \quad \mathbf{prox}_{\mu f}(x) = \arg\min_{y\in\mathbf{R}^d} f(y) + \frac{1}{2\mu}\|y-x\|^2, \tag{7.3}$$

respectively. The minimizing properties of f and f_μ are moreover closely aligned, for example, their first-order critical points and local/global minimizers coincide. Inspired by this relationship, this work thus seeks $(\varepsilon_1, \varepsilon_2)$-*approximate second-order critical points* x of f_μ, satisfying:

$$\|\nabla f_\mu(x)\| \le \varepsilon_1 \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 f_\mu(x)) \ge -\varepsilon_2. \tag{7.4}$$

An immediate difficulty is that f_μ is not C^2 in general. Indeed, the seminal work [146] shows f_μ is C^2-smooth *globally*, if and only if, f is C^2-smooth globally. Therefore assuming that f_μ is C^2 globally is meaningless for nonsmooth optimization. Nevertheless, known results in [84] imply that for "generic" semialgebraic functions, f_μ is locally C^2 near x whenever $\|\nabla f_\mu(x)\|$ is sufficiently small.

Turning to algorithm design, a natural strategy is to apply a "saddle escaping" gradient method [123] directly to f_μ. This strategy fails in general, since it is not possible to evaluate the gradient

$$\nabla f_\mu(x) = \frac{1}{\mu}(x - \mathbf{prox}_{\mu f}(x)) \tag{7.5}$$

in closed form. Somewhat expectedly, however, our **first contribution** is to show that one may extend the results of [123] to allow for *inexact* evaluations $G(x) \approx \nabla f_\mu(x)$ satisfying

$$\|G(x) - \nabla f_\mu(x)\| \le a\|\nabla f_\mu(x)\| + b \qquad \text{for all } x \in \mathbf{R}^d,$$

for appropriately small $a, b \ge 0$. The algorithm (Algorithm 7) returns a point x satisfying (7.4), with $\tilde{O}(\max\{\varepsilon_1^{-2}, \varepsilon_2^{-4}\})$ evaluations of G, matching the complexity of [123].

Our **second contribution** constructs approximate oracles $G(x)$, tailored to common problem structures. Each oracle satisfies

$$G(x) = \mu^{-1}\left(x - \textsc{ProxOracle}_{\mu f}(x)\right),$$

where $\textsc{ProxOracle}_{\mu f}$ is an approximate minimizer of the *strongly convex* subproblem defining $\mathbf{prox}_{\mu f}(x)$. Since the subproblem is strongly convex, we construct $\textsc{ProxOracle}_{\mu f}$ from K iterations of off-the-shelf first-order methods for convex optimization. We focus in particular on the class of *model-based methods* [66]. Starting from initial point $x_0 = x$, these methods attempt to minimize $f(y) + \frac{1}{2\mu}\|y - x\|^2$ by iterating

$$x_{k+1} = \operatorname*{arg\,min}_{y \in \mathbf{R}^d} \left\{ f_{x_k}(y) + \frac{1}{2\mu}\|y - x\|^2 + \frac{\theta_k}{2}\|y - x_k\|^2 \right\}, \tag{7.6}$$

where $\theta_k > 0$ is a control sequence and for all $z \in \mathbf{R}^d$, the function $f_z \colon \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ is a local weakly convex model of f. In Table 7.1, we show three

Algorithm	Objective	Model function $f_z(y)$
Prox-Subgradient [66]	$l(y) + r(y)$	$l(z) + \langle v_z, y - z\rangle + r(y)$
Prox-gradient	$F(y) + r(y)$	$F(z) + \langle \nabla F(y), y - z\rangle + r(y)$
Prox-linear [99]	$h(c(x)) + r(x)$	$h(c(x) + \nabla c(x)(y - x)) + r(y)$

Table 7.1: The three algorithms with the update (7.6); we assume h is convex and Lipschitz, r is weakly convex and possibly infinite valued, both F and c are smooth, and l is Lipschitz and weakly convex on dom$r$ with $v_z \in \partial l(z)$.

models, adapted to possible decompositions of f. In Table 7.2, we show how the model function f_z influences the total complexity $\tilde{O}(K \times \max\{\varepsilon_1^{-2}, \varepsilon_2^{-4}\})$ of finding a second order stationary point of f_μ (7.4). In short, prox-gradient and prox-linear methods require $\tilde{O}(\max\{\varepsilon_1^{-2}, \varepsilon_2^{-4}\})$ iterations of (7.6), while prox-subgradient methods require $\tilde{O}(d \max\{\varepsilon_1^{-6}\varepsilon_2^{-6}, \varepsilon_2^{-18}\})$. The efficiency of the prox-gradient method directly matches the analogous guarantees for the perturbed gradient method in the smooth setting [123]. The convergence guarantee of the prox-subgradient method has no direct analogue in the literature. Extensions for stochastic variants of these algorithms follow trivially, when the proximal subproblem (7.6) can be approximately solved with high probability (e.g. using [114, 115, 125, 207]). The rates for the prox-gradient and prox-linear method are analogous to those in [119], which uses an algorithm-dependent measure of stationarity. Although the algorithms and the results in this work and in [119] are mostly of theoretical interest, they do suggest that efficiently escaping from saddle points is possible in nonsmooth optimization.

Related work. We highlight several approaches for finding second-order critical points. Asymptotic guarantees have been developed in deterministic [142, 141, 67] and stochastic settings [202]. Other approaches explicitly leverage second order information about the objective function, such as full Hessian or Hessian vector products computations [189, 60, 3, 42, 4, 216, 217, 199, 59].

Algorithm to Evaluate $g(x)$	Overall Algorithm Complexity
Prox-Subgradient [66]	$\tilde{O}(d \max\{\varepsilon_1^{-6}\varepsilon_2^{-6}, \varepsilon_2^{-18}\})$
Prox-gradient	$\tilde{O}(\max\{\varepsilon_1^{-2}, \varepsilon_2^{-4}\})$
Prox-linear [99]	$\tilde{O}(\max\{\varepsilon_1^{-2}, \varepsilon_2^{-4}\})$

Table 7.2: The overall complexity of the proposed algorithm $\tilde{O}(K \times \max\{\varepsilon_1^{-2}, \varepsilon_2^{-4}\})$, where K is the number of steps of (7.6) required to evaluate $g(x)$. The rate for Prox-subgradient holds in the regime $\varepsilon_1 = O(\varepsilon_2)$.

Several methods exploit only first-order information combined with random perturbations [104, 123, 61, 124, 122]. The work [122] also studies saddle avoiding methods with inexact gradient oracles G; a key difference: the oracle of [122] is the gradient of a smooth function $G = \nabla g$. Several existing works have developed methods that find second-order stationary points of manifold [58, 229], convex [178, 191, 164, 244], and low-rank matrix constrained problems [249, 197].

Road map. Section 7.2 presents a result for finding second-order stationary points with inexact gradient evaluations. Section 7.3 develops several oracle mappings that approximately evaluate the gradient of the Moreau Envelope and derives the complexity estimates of Table 7.2.

7.2 Escaping saddle points with inexact gradients

In this section, we analyze an inexact gradient method on smooth functions, focusing on convergence to second-order stationary points. The consequences for nonsmooth optimization, which will follow from a smoothing technique, will be explored in Section 7.2.

We begin with the following standard assumption, which asserts that the function f in question has a globally Lipchitz continuous gradient.

Assumption 7.2.1 (Globally Lipschitz gradient)**.** *Fix a function $g\colon \mathbf{R}^d \to \mathbf{R}$ that is bounded from below and whose gradient is globally Lipschitz continuous with constant L_1, meaning*

$$\|\nabla g(x) - \nabla g(y)\| \leq L_1 \|x - y\| \qquad \text{for all } x, y \in \mathbf{R}^d.$$

The next assumption is more subtle: it requires the Hessian $\nabla^2 g$ to be Lipschitz continuous on a neighborhood of any point where the gradient is sufficiently small. When we discuss consequences for nonsmooth optimization in the later sections, the fact that f is assumed to be C^2-smooth only locally will be crucial to our analysis.

Assumption 7.2.2 (Locally Lipschitz Hessian)**.** *Fix a function $g\colon \mathbf{R}^d \to \mathbf{R}$ and assume that there exist positive constants α, β, L_2 satisfying the following: For any point $\bar{x}$ with $\|\nabla g(\bar{x})\| \leq \alpha$, the function g is C^2-smooth on $\mathbf{B}_\beta(\bar{x})$ and satisfies the Lipschitz condition:*

$$\|\nabla^2 g(x) - \nabla^2 g(y)\| \leq L_2 \|x - y\| \qquad \text{for all } x, y \in B_\beta(\bar{x}).$$

We aim to analyze an inexact gradient method for minimizing the function f under Assumptions 7.2.1 and 7.2.2. The type of inexactness we allow is summarized by the following oracle model.

Definition 7.2.3 (Inexact oracle)**.** *A map $G\colon \mathbf{R}^d \to \mathbf{R}^d$ is an (a, b)*-inexact gradient oracle *for f if it satisfies*

$$\|\nabla g(x) - G(x)\| \leq a \cdot \|\nabla g(x)\| + b \qquad \forall x \in \mathbf{R}^d. \tag{7.7}$$

Turning to algorithm design, the method we introduce (Algorithm 7) directly extends the perturbed gradient method introduced in [123] to inexact gradient

Algorithm 7: Perturbed inexact gradient descent

Data: $x_0 \in \mathbf{R}^d$, $T \in \mathbf{N}$, and $\eta, r, \varepsilon_1, M > 0$
Set $t_{\text{pert}} = -M$
Step $t = 0, \ldots, T$**:**
  Set $u_t = 0$
  If $\|G(x_t)\| \le \varepsilon_1/2$ **and** $t - t_{\text{pert}} \ge M$:
    Update $t_{\text{pert}} = t$
    Draw perturbation $u_t \sim \text{Unif}(r\mathbf{B})$
  Set $x_{t+1} \leftarrow x_t - \eta \cdot (G(x_t) + u_t)$.

oracles in the sense of Definition 7.2.3. The convergence guarantees for the algorithm will be based on the following explicit setting of parameters. Fix target accuracies $\varepsilon_1, \varepsilon_2 > 0$ and choose any $\Delta_g \ge g(x_0) - \inf g$. We first define the *auxiliary parameters*:

$$\phi := 2^{24} \max\left\{1, 5\frac{L_2\varepsilon_1}{L_1\varepsilon_2}\right\} \frac{L_1^2}{\delta}\sqrt{d}\left(\Delta_g \max\left\{\frac{L_2^2}{\varepsilon_2^5}, \frac{1}{\varepsilon_1^2\varepsilon_2^1}\right\} + \frac{1}{\varepsilon_2^2}\right) \quad \text{and} \quad \gamma := \log_2(\phi \log_2(\phi)^8), \tag{7.8}$$

and

$$F = \frac{1}{800\gamma^3}\frac{1-a}{(1+a)^2}\frac{\varepsilon_2^3}{L_2^2} \qquad \text{and} \qquad R = \frac{1}{4\gamma}\frac{\varepsilon_2}{L_2}.$$

The parameters required by the algorithm are then set as

$$\eta = \frac{1-a}{(1+a)^2}\frac{1}{L_1}, \quad r = \frac{\varepsilon_2^2}{400L_2\gamma^3}\min\left\{1, \frac{L_1\varepsilon_2}{5\varepsilon_1 L_2}\right\}, \quad M = \frac{(1+a)^2}{(1-a)}\frac{L_1}{\varepsilon_2}\gamma. \tag{7.9}$$

The following is the main result of the section. The proof follows closely the argument in [123] and therefore appears in Appendix 7.4.1.

Theorem 7.2.4 (Perturbed inexact gradient descent)**.** *Suppose that $g\colon \mathbf{R}^d \to \mathbf{R}$ is a function satisfying Assumptions 7.2.1 and 7.2.2 and $G\colon \mathbf{R}^d \to \mathbf{R}^d$ is an (a, b)-inexact gradient oracle for g. Let $\delta \in (0, 1)$, $\varepsilon_1 \in (0, \alpha)$, $\varepsilon_2 \in (0, \min\{4\gamma\beta L_2, L_1, L_1^2\})$, and suppose that*

$$a \le \min\left\{\frac{1}{20}, \frac{1}{L_1\eta M 2^{\gamma+2}}, \frac{R}{\varepsilon_1\eta M 2^{\gamma+2}}\right\} \quad \textit{and}$$
$$b \le \min\left\{\frac{\varepsilon_1}{64}, \left(\frac{F}{40\eta M}\right)^{1/2}, \left(\frac{L_1 F}{M(5L_1+1)}\right)^{1/2}, \frac{R}{M\eta 2^{(\gamma+2)}}\right\}.$$

Then with probability at least $1 - \delta$*, at least one iterate generated by Algorithm 7 with parameters* (7.9) *is a* $(\varepsilon_1, \varepsilon_2)$*-second-order critical point of* g *after*

$$T = 8\Delta_g \max\left\{2\frac{M}{F}, \frac{256}{(1-a)\eta\varepsilon_1^2}\right\} + 4M = \tilde{O}\left(L_1\Delta_g \max\left\{\frac{L_2^2}{\varepsilon_2^4}, \frac{1}{\varepsilon_1^2}\right\}\right) \quad \textit{iterations.} \tag{7.10}$$

The necessary bounds for a and b can be estimated as

$$\begin{aligned} a &\lesssim \frac{\delta}{L_1^3\Delta_g} \cdot d^{-1/2} \cdot \min\left\{\frac{\varepsilon_2^6}{L_2^2}, \varepsilon_1^2\varepsilon_2^2\right\} \cdot \min\left\{1, \frac{L_1\varepsilon_2}{L_2\varepsilon_1}\right\}^2 \quad \text{and} \\ b &\lesssim \frac{\delta}{L_1^2L_2\Delta_g} \cdot d^{-1/2} \cdot \min\left\{\frac{\varepsilon_2^7}{L_2^2}, \varepsilon_1^2\varepsilon_2^3\right\} \cdot \min\left\{1, \frac{L_1\varepsilon_2}{L_2\varepsilon_1}\right\}, \end{aligned} \tag{7.11}$$

where the symbol "$\lesssim$" denotes inequality up to polylogarithmic factors. Thus, Algorithm 7 is guaranteed to find a second order stationary point efficiently, provided that the gradient oracles are highly accurate. In particular, when $a = b = 0$ we recover the known rates from [123].

7.3 Escaping saddle points of the Moreau envelope

In this section, we apply Algorithm 7 to the Moreau Envelope (7.3) of the weakly convex optimization problem (7.2) in order to find a second order stationary point of f_μ (7.4). We will see that a variety of standard algorithms for nonsmooth convex optimization can be used as inexact gradient oracles for the Moreau envelope. Before developing those algorithms, we summarize our main assumptions on f_μ, describe why approximate second order stationary points of f_μ are meaningful for f, and show that Assumption 7.2.2, while not automatic for general f_μ, holds for a large class of semialgebraic functions.

As stated in the introduction, for $\mu < \rho^{-1}$, the Moreau envelope is an everywhere C^1 smooth with Lipschitz continuous gradient. In particular,

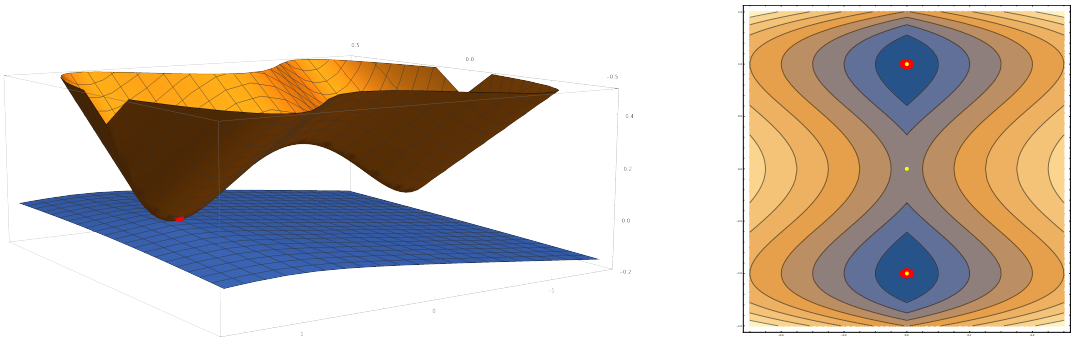


Figure 7.1: Critical poins of f in (7.12). We use $\varepsilon_1 = \varepsilon_2 = 0.04$. On the left: The function, a point $(x, f(x))$ with x an $(\varepsilon_1, \varepsilon_2)$-second-order critical point of f_μ and its corresponding quadratic $q(\cdot)$. On the right: The set of first-order critical points of f (yellow) and the set of $(\varepsilon_1, \varepsilon_2)$-second-order critical points of f_μ (red).

Assumption 7.2.1 holds automatically for f_μ with $L_1 = \max\left\{\mu^{-1}, \frac{\rho}{1-\mu\rho}\right\}$.

See for example [67] for a short proof. Assumption 7.2.2, however, is not automatic, so we impose the following assumption throughout.

Assumption 7.3.1. *Let $f\colon \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$ be a closed ρ-weakly convex function whose Moreau envelope f_μ satisfies Assumption 7.2.2 with constants α, β, L_2.*

Turning to stationarity conditions, a natural question is whether the second order condition (7.4) is meaningful for f. The next proposition shows that the condition (7.4) implies the existence of an approximate quadratic minorant of f with small slope and curvature at a nearby point.

We defer the proof to Section 7.4.3.

Proposition 7.3.2. *Consider $f\colon \mathbf{R}^d \to \bar{\mathbb{R}}$ satisfying Assumption 7.3.1. Assume that $x \in \mathbf{R}^d$ is an $(\varepsilon_1, \varepsilon_2)$-second order critical point of f_μ with $\varepsilon_1 \le \min\{\alpha, \frac{\varepsilon_2}{2L_2\mu}\}$ and let $\hat{x} := \mathbf{prox}_{\mu f}(x)$. Then there exists a quadratic function $q\colon \mathbf{R}^d \to \mathbf{R}$ and a neighborhood $\mathcal{U} = B_{3\varepsilon_2/4L_2}(\widehat{x})$ of x for which the following hold.*

1. *(**Nearby point**) The point $\widehat{x}$ is close to x: $\|x - \widehat{x}\| \le \mu \cdot \varepsilon_1$.*

2. *(**Minorant**) For any $y \in \mathcal{U}$, we have $q(y) \le f(y)$.*

3. *(**Small subgradient**) The quadratic has a small gradient at $\widehat{x}$:*

$$\|\nabla q(\widehat{x})\| \le \varepsilon_1.$$

4. *(**Small negative curvature**) The quadratic has small negative curvature:*

$$\nabla^2 q(\widehat{x}) \succeq -3\varepsilon_2.$$

5. *(**Approximate match**) The quadratic almost matches the function at $\widehat{x}$:*

$$f(\widehat{x}) - q(\widehat{x}) \le \frac{\mu}{2}(1 + 3\mu\varepsilon_2)\,\varepsilon_1^2.$$

In Figure 7.1, we illustrate the proposition with the following nonsmooth function:

$$f(x, y) = |x| + \frac{1}{4}(y^2 - 1)^2. \tag{7.12}$$

The Moreau envelope of this function has three first-order critical points: a strict saddle point $(0, 0)$ and two global minima $(-1, 0)$, and $(1, 0)$. As shown in the right plot of Figure 7.1, approximate second-order critical points of f_μ cluster around minimizers of f. In addition, the left plot of Figure 7.1 shows the lower bounding quadratic from Proposition 7.3.2.

Finally, we complete this section by showing that Assumption 7.3.1 is reasonable: it holds for generic semialgebraic functions.[1]

Theorem 7.3.3. *Let $f\colon \mathbf{R}^d \to \bar{\mathbb{R}}$ be a semi-algebraic ρ-weakly-convex function. Then, the set of vectors $v \in \mathbf{R}^d$ for which the tilted function $g(x; v) = f(x) + \langle v, x\rangle$ satisfies Assumption 7.3.1 has full Lebesgue measure.*

[1] A set is semialgebraic if its graph can be written as a finite union of sets each defined by finitely many polynomial inequalities.

The proof appears in Section 7.4.4, and is a small modification of the argument in [67].

7.3.1 Inexact Oracles for the Moreau Envelope

In this section, we develop inexact gradient oracles for $\nabla f_\mu = \mu^{-1}(x - \mathbf{prox}_{\mu f}(x))$. Leveraging this expression, our oracles will satisfy

$$G(x) = \mu^{-1}\left(x - \textsc{ProxOracle}_{\mu f}(x)\right), \tag{7.13}$$

where $\textsc{ProxOracle}_{\mu f}$ is the output of a numerical scheme that solves (7.3). To ensure G meets the conditions of Definition 7.2.3, we require that

$$\|\textsc{ProxOracle}_{\mu f}(x) - \mathbf{prox}_{\mu f}(x)\| \leq a \cdot \|x - \mathbf{prox}_{\mu f}(x)\| + \mu \cdot b.$$

for some constants $a \in (0, 1)$ and $b > 0$.

Since f is ρ-weakly convex, evaluating $\mathbf{prox}_{\mu f}(x_k)$ amounts to minimizing the $(\mu^{-1} - \rho)$-strongly convex function $f(x) + \frac{1}{2\mu}\|x - x_k\|^2$. We now use this strong convexity to derive efficient proximal oracles via a class of algorithms called *model-based methods* [66], which we now briefly summarize. Given a minimization problem $\min_{x \in \mathbf{R}^d} g(x)$, where g is strongly convex, a *model-based method* is an algorithm that recursively updates

$$x_{k+1} \leftarrow \arg\min_x \; g_{x_k}(x) + \frac{\theta_t}{2}\|x - x_k\|^2, \tag{7.14}$$

where $g_{x_k} \colon \mathbf{R}^d \to \bar{\mathbb{R}}$ is a function that approximates g near x_k. Returning to the proximal subproblem, say we wish to compute $\mathbf{prox}_{\mu f}(x_0)$ for some given x_0. We consider an inner loop update of the form

$$x_{k+1} \leftarrow \arg\min_{x \in \mathbf{R}^d} \; f_{x_k}(x) + \frac{1}{2\mu}\|x - x_0\|^2 + \frac{\theta_k}{2}\|x - x_k\|^2, \tag{7.15}$$

Algorithm 8: PROXORACLE$^K_{\mu f}$

Data: Initial point $x_0 \in \mathbf{R}^d$.
Parameters: Stepsize $\theta_k > 0$, Flag `one_sided`.
Output: Approximation of $\mathbf{prox}_{\mu f}(x_0)$.
Step k ($k \leq K+1$)**:**
$x_{k+1} \leftarrow \arg\min_{x \in \mathbf{R}^d} f_{x_k}(x) + \frac{1+\theta_k\mu}{2\mu} \left\| x - \frac{(x_0+\theta_k\mu \cdot x_k)}{1+\theta_k\mu} \right\|^2$
If `one_sided` **:**
$\bar{x}_K = \frac{2}{(K+2)(K+3)-2} \sum_{k=1}^{K+1} (t+1) x_k$
return $\bar{x}_K$
Else:
return x_K

where $f_{x_k} \colon \mathbf{R}^d \to \bar{\mathbb{R}}$ is a function that locally approximates f (see Table 7.1 for three examples). Completing the square, this update can be equivalently written as a proximal step on f_{x_k}, where the reference point is a weighted average of x_0 and x_k as summarized in Algorithm 8. Turning to complexity, we note that the approximation quality of a model governs the speed at which iteration (7.15) converges. In what follows, we will present two families of models with different approximation properties, namely one- and two-sided models. We will see that models with double-sided accuracy require fewer iterations to approximate $\mathbf{prox}_{\mu f}(x_0)$.

One-sided models

We start by studying models that globally lower bound the function and agree with it at the reference point. Subgradient-type models are the canonical examples, and we will discuss them shortly.

Assumption 7.3.4 (One-sided model)**.** *Let $f = l + r$, where $r \colon \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ is a closed function and $l \colon \mathbf{R}^d \to \mathbf{R}$ is locally Lipschitz. Assume there exists $\tau > 0$ and a family of models $l_x \colon \mathbf{R}^d \to \mathbf{R}$, defined for each $x \in \mathbf{R}^d$, such that the following hold: For*

all $x \in \mathbf{R}^d$, l_x *is L-Lipschitz on* dom$r$ *and satisfies*

$$l_x(x) = l(x) \qquad \text{and} \qquad l_x(y) - l(y) \leq \tau \|y - x\|^2 \qquad \text{for all } y \in \mathbf{R}^d. \tag{7.16}$$

In addition, for all $x \in \mathbf{R}^d$, *the model*

$$f_x := l_x + r$$

is ρ*-weakly convex.*

Now we bound the number of iterations that are needed for Algorithm 8 to obtain a (a, b)-inexact proximal point oracle with one-sided models. The algorithm outputs an average of the iterates with nonuniform weights that improves the convergence speed.

Theorem 7.3.5. *Fix* $a, b > 0$ *and let* $f \colon \mathbf{R}^d \to \bar{\mathbb{R}}$ *be a* ρ*-weakly-convex function and let* $f_x \colon \mathbf{R}^d \to \bar{\mathbb{R}}$ *be a family of models that satisfy Assumption 7.3.4 for* $\tau = 0$. *Let* $\mu^{-1} > \rho$ *be a constant, and set* $\theta_k = \frac{(\mu^{-1} - \rho)}{2}(k+1)$ *then Algorithm 8 with flag* `one_sided = true` *outputs an a point* $\bar{x}_K$ *such that*

$$\|\bar{x}_K - \mathbf{prox}_{\mu f}(x_0)\|_2 \leq a \cdot \|x_0 - \mathbf{prox}_{\mu f}(x_0)\|_2 + \mu \cdot b,$$

provided the number of iterations is at least $K \geq \frac{4}{a} + \frac{16L^2}{(1-\mu\rho)^2 b^2}$.

The proof of this result follows easily from Theorem 4.5 in [66] and thus, we omit it. By exploiting this rate, we derive a complexity guarantee with one-sided models.

Theorem 7.3.6 (One-sided model-based method). *Consider an* L_f*-Lipschitz* ρ*-weakly-convex function* $f \colon \mathbf{R}^d \to \bar{\mathbb{R}}$ *that satisfies Assumption 7.3.1 and a family of models* f_x *satisfying Assumption 7.3.4. Then, for all sufficiently small* $\varepsilon_1 > 0$, *and any* $\varepsilon_2 > 0$, $\delta \in (0, 1)$ *there exists a parameter configuration* (η, r, M) *that ensures that*

with probability at least $1-\delta$ *one of the first* T *iterates generated by Algorithm 7 with gradient oracle*

$$g(x) = \mu^{-1}\left(x - \textsc{ProxOracle}^{K}_{\mu f}(x)\right) \qquad \textit{(Algorithm 8)}$$

is an $(\varepsilon_1, \varepsilon_2)$*-second-order critical point of* f_μ *provided that the inner and outer iterations satisfy*

$$\begin{aligned} K &= \tilde{O}\left((1-\mu\rho)^{-2} L_f^2 L_1^4 L_2^2 \Delta_f^2 \cdot \frac{d}{\delta} \cdot \max\left\{\frac{L_2^4}{\varepsilon_2^{14}}, \frac{1}{\varepsilon_1^4 \varepsilon_2^6}\right\} \cdot \max\left\{\frac{L_2^2 \varepsilon_1^2}{L_1^2 \varepsilon_2^2}, 1\right\}\right) \quad \textit{and} \\ T &= \tilde{O}\left(L_1 \Delta_f \max\left\{\frac{L_2^2}{\varepsilon_2^4}, \frac{1}{\varepsilon_1^2}\right\}\right) \end{aligned} \tag{7.17}$$

where $L_1 := \max\left\{\frac{1}{\mu}, \frac{\rho}{1-\mu\rho}\right\}$ *and* $\Delta_f = f(x_0) - \inf f$.

Proof. This result is a corollary of Theorem 7.3.5 and Theorem 7.2.4. By [67, Lemma 2.5] and Assumption 7.3.1 we conclude that the Moreau envelope satisfies the hypothesis of Theorem 7.2.4. Hence, the result follows from this theorem provided that we show that the gradient oracle is accurate enough. By Theorem 7.3.5 if we set the number of iterations according to (7.17) we get an inexact oracle that matches the assumptions of Theorem 7.2.4 □

The rate from Table 7.2 follows by noting that $\max\left\{\frac{L_2^2\varepsilon_1^2}{L_1^2\varepsilon_2^2}, 1\right\} = 1$ when $\varepsilon_1 \leq \frac{L_1}{L_2}\varepsilon_2$.

Example: proximal subgradient method. Consider the setting of Assumption 7.3.4, where $f = l + r$. Assuming that l is τ-weakly convex, it possesses an affine model:

$$l_x(y) = l(x) + \langle v, y - x\rangle, \qquad \text{where } v \in \partial l(x).$$

By weak convexity, $f_x = l_x + r$ satisfies Assumption 7.3.4. Moreover, the resulting

update (7.15) reduces to the following proximal subgradient method:

$$x_{k+1} = \mathbf{prox}_{\frac{\mu}{1+\theta_k\mu} r}\left(\frac{1}{1+\theta_k\mu}\left(x_0 + \theta_k\mu \cdot x_k - \mu \cdot v\right)\right).$$

Theorem 7.3.6 applied to this setting thus implies the rate in Table 7.2.

Two-sided models

The slow convergence of one-sided model-based algorithms motivates stronger approximation assumptions. In this section we study models that satisfy the following assumption.

Assumption 7.3.7 (Two-sided model)**.** *Assume that for any $x \in \mathbf{R}^d$, the function $f_x \colon \mathbf{R}^d \to \bar{\mathbb{R}}$ is ρ-weakly convex and satisfies*

$$|f_x(y) - f(y)| \leq \frac{q}{2}\|y - x\|^2 \qquad \text{for all } y \in \mathbf{R}^d. \tag{7.18}$$

When equipped with double-sided models, model-based algorithms for the proximal subproblem converge linearly.

Theorem 7.3.8. *Suppose that $f \colon \mathbf{R}^d \to \bar{\mathbb{R}}$ is a ρ-weakly-convex function, let f_x be a family models satisfying Assumption 7.3.7. Fix an accuracy level a. Set $\mu^{-1} > \rho + q$ and the stepsizes to $\theta_t = \theta > q$, then Algorithm 8 with flag* `one_sided = false` *outputs a point x_K such that*

$$\|x_K - \mathbf{prox}_{\mu f}(x_0)\|_2 \leq a \cdot \|x_0 - \mathbf{prox}_{\mu f}(x_0)\|_2,$$

provided that $K \geq 2\log(a^{-1}) \log\left(\frac{\mu^{-1}-\rho+\theta}{q+\theta}\right)^{-1}$.

We defer the proof of this result to Section 7.4.5. Given this guarantee for two-sided models, we derive the following theorem. The proof is analogous to

that of Theorem 7.3.6: the only difference is that we use Theorem 7.3.8 instead of Theorem 7.3.5. Thus we omit the proof.

Theorem 7.3.9 (**Two-sided model-based method**). *Consider a weakly convex function $f\colon \mathbf{R}^d \to \bar{\mathbb{R}}$ that satisfies Assumption 7.3.1 and a family of models f_x satisfying Assumption 7.3.7. Then for any $\delta \in (0,1)$ and sufficiently small $\varepsilon_1 > 0$, there exists a parameter configuration (η, r, M) such that with probability at least $1-\delta$ one of the first T iterates generated by Algorithm 7 with inexact oracle*

$$g(x) = \mu^{-1}\left(x - \textsc{ProxOracle}^K_{\mu f}(x)\right) \qquad \textit{(Algorithm 8)}$$

is an $(\varepsilon_1, \varepsilon_2)$-second-order critical point of f_μ provided that the inner and outer iterations satisfy

$$K = \tilde{O}(1) \quad \textit{and} \quad T = \tilde{O}\left(\max\left\{\frac{1}{\mu}, \frac{\rho}{1-\mu\rho}\right\}(f(x_0) - \inf f)\min\left\{L_2^2\varepsilon_1^{-4}, \varepsilon_1^{-2}\right\}\right).$$

We close this section with two examples of two-sided models.

Example: Prox-gradient method. Suppose that

$$f = F + r$$

where $r\colon \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ is closed and ρ-weakly convex and F is C^1 with q-Lipschitz continuous derivative on $\mathrm{dom} r$. Then due to the classical inequality

$$|F(y) - F(x) - \langle \nabla F(x), y - x\rangle| \le \frac{q}{2}\|y - x\|^2 \qquad \text{for all } x, y \in \mathrm{dom} r,$$

the model

$$f_x(y) = F(x) + \langle \nabla F(x), y - x\rangle + r(x),$$

satisfies Assumption 7.3.7. Moreover, the resulting update (7.15) reduces to the following proximal gradient method:

$$x_{k+1} = \mathbf{prox}_{\frac{\mu}{1+\theta_k\mu} r}\left(\frac{1}{1+\theta_k\mu}\left(x_0 + \theta_k\mu \cdot x_k - \mu \cdot \nabla F(x_k)\right)\right).$$

Theorem 7.3.9 applied to this setting thus implies the rate in Table 7.2.

Example: Prox-linear method. Suppose that

$$f = h \circ c + r$$

where $r\colon \mathbf{R}^d \to \mathbf{R} \cup \{+\infty\}$ is closed and ρ-weakly convex, h is L-Lipschitz and convex on $\mathrm{dom} r$, and c is C^1 with β-Lipschitz Jacobian on $\mathrm{dom} r$. Then due to the classical inequality $\|c(y) - c(x) - \nabla c(x), y - x\| \leq \frac{\beta}{2}\|y - x\|^2$, we have

$$|h(c(y)) - h(c(x) + \nabla c(x)(y - x))| \leq \frac{\beta L}{2}\|x - y\|^2, \qquad \text{for all } x, y \in \mathrm{dom} r.$$

Consequently, the model

$$f_x(y) = h(c(x) + \nabla c(x)(y - x)) + r(x),$$

satisfies Assumption 7.3.7 with $q = \beta L$. Moreover, the resulting update (7.15) reduces to the following prox-linear method [99]:

$$x_{k+1} = \operatorname*{arg\,min}_{y \in \mathbf{R}^d} h(c(x_k) + \nabla c(x_k)(y - x_k)) + r(x) + \frac{1 + \theta_k \mu}{2\mu} \left\| x - \frac{x_0 + \theta_k \mu \cdot x_k}{1 + \theta_k \mu} \right\|^2.$$

Theorem 7.3.9 applied to this setting thus implies the rate in Table 7.2.

7.4 Analysis

7.4.1 Proof of Theorem 7.2.4

Throughout this section, we assume the setting of Theorem 7.2.4. We begin by recording some inequalities that we will use later on.

Lemma 7.4.1. *The following inequalities hold.*

1. ***(Radius)***
$$\sqrt{32\eta\frac{(1+a)^2}{(1-a)}MF} + \eta r < R.$$

2. ***(Function value)***
$$\varepsilon_1\eta r + L_1\eta^2 r^2/2 \le F/2.$$

3. ***(Probability)***
$$p := \frac{TL_1\frac{(1+a)^2}{(1-a)}\frac{\sqrt{d}}{\varepsilon_2}\gamma^2\max\left\{1, 5\frac{L_2\varepsilon_1}{L_1\varepsilon_2}\right\}2^9}{2^\gamma} \le \delta.$$

Proof. We start with the first inequality, observe that

$$32\eta\frac{(1+a)^2}{(1-a)} \le 32\frac{1}{L_1} \quad \text{and} \quad FM = \frac{1}{800\gamma^3}\frac{1-a}{(1+a)^2}\frac{\varepsilon_2^3}{L_2^2}\cdot\frac{(1+a)^2}{(1-a)}\frac{L_1}{\varepsilon_2}\gamma = \frac{\varepsilon_2^2 L_1}{800L_2^2\gamma^2}.$$

Therefore, since

$$\eta \le \frac{1}{L_1} \qquad \text{and} \qquad r = \frac{\varepsilon_2^2}{400L_2\gamma^3}\min\left\{1, \frac{L_1\varepsilon_2}{5\varepsilon_1 L_2}\right\} \le \frac{\varepsilon_2^2}{400L_2\gamma^3},$$

we have

$$\begin{aligned}\sqrt{32\eta\frac{(1+a)^2}{(1-a)}MF} + \eta r &\le \frac{1}{5\gamma}\frac{\varepsilon_2}{L_2} + \frac{\varepsilon_2^2}{400L_1L_2\gamma} \\ &\le \frac{1}{5\gamma}\frac{\varepsilon_2}{L_2} + \frac{1}{400\gamma}\frac{\varepsilon_2}{L_2} < \frac{1}{4\gamma}\frac{\varepsilon_2}{L_2} = R.\end{aligned}$$

where the third inequality follows from $L_1/\varepsilon_2 \ge 1$.

Now, we prove the second statement: $\varepsilon_1\eta r + L_1\eta^2 r^2/2 \le F/2$. Indeed, first recall the definition of r above and that $\eta = \frac{1-a}{(1+a)^2}\frac{1}{L_1}$, $F = \frac{1}{800\gamma^3}\frac{1-a}{(1+a)^2}\frac{\varepsilon_2^3}{L_2^2}$. Thus, we bound the first term:

$$\varepsilon_1\cdot\eta\cdot r \le \varepsilon_1\cdot\frac{1-a}{(1+a)^2}\frac{1}{L_1}\cdot\frac{\varepsilon_2^2}{400L_2\gamma^3}\frac{L_1\varepsilon_2}{5\varepsilon_1 L_2} \le \frac{1-a}{(1+a)^2}\frac{\varepsilon_2^3}{2000L_2^2\gamma^3} \le \frac{2}{5}F.$$

Next, we bound the second term:

$$\begin{aligned}\frac{L_1 \cdot \eta^2 \cdot r^2}{2} &= \frac{1}{2}L_1 \cdot \left(\frac{1-a}{(1+a)^2}\frac{1}{L_1}\right)^2 \cdot \left(\frac{\varepsilon_2^2}{400L_2\gamma^3}\right)^2 \\ &= \frac{\varepsilon_2}{L_1}\frac{1-a}{(1+a)^2}\frac{1}{400\gamma^3}\frac{1}{800\gamma^3}\frac{1-a}{(1+a)^2}\frac{\varepsilon_2^3}{L_2^2} \\ &\leq \frac{1}{400\gamma^3}\frac{1}{800\gamma^3}\frac{1-a}{(1+a)^2}\frac{\varepsilon_2^3}{L_2^2} \leq \frac{F}{10}\end{aligned}$$

where we used $(1-a)/(1+a)^2 \leq 1$, $\varepsilon_2 \leq L_1$ and the simple inequality $1/400\gamma^3 \leq 1/10$.

Finally, we show that $p \leq \delta$. Recall that by definition,

$$T = 8\Delta_g \max\left\{\frac{M}{F}, \frac{256}{\eta\varepsilon_1^2}\right\} + 4M.$$

We upper bound T using $F = \frac{1}{800\gamma^3}\frac{1-a}{(1+a)^2}\frac{\varepsilon_2^3}{L_2^2}$, $M = \frac{(1+a)^2}{(1-a)}\frac{L_1}{\varepsilon_2}\gamma$, and $\eta = \frac{1-a}{(1+a)^2}\frac{1}{L_1}$:

$$\begin{aligned}T &= 2^4\frac{(1+a)^2}{1-a}\Delta_g L_1 \max\left\{800\gamma^4\frac{(1+a)^2}{1-a}\frac{L_2^2}{\varepsilon_2^4}, \frac{256}{\varepsilon_1^2}\right\} + 4\frac{(1+a)^2}{(1-a)}\frac{L_1}{\varepsilon_2}\gamma \\ &\leq 2^4 \cdot 800\left(\frac{(1+a)^2}{1-a}\right)^2 \cdot L_1\gamma^4 \cdot \left(\Delta_g \max\left\{\frac{L_2^2}{\varepsilon_2^4}, \frac{1}{\varepsilon_1^2}\right\} + \frac{1}{\varepsilon_2}\right).\end{aligned}$$

This yields:

$$p \leq \frac{2^{13} \cdot 800\left(\frac{(1+a)^2}{1-a}\right)^3 \cdot L_1^2\gamma^6\sqrt{d} \cdot \max\left\{1, 5\frac{L_2\varepsilon_1}{L_1\varepsilon_2}\right\}\left(\Delta_g \max\left\{\frac{L_2^2}{\varepsilon_2^5}, \frac{1}{\varepsilon_1^2\varepsilon_2}\right\} + \frac{1}{\varepsilon_2^2}\right)}{2^\gamma}.$$

Next, recall that $2^\gamma = \phi \cdot \log_2(\phi)^8$, where

$$\phi := 2^{24}\frac{L_1^2}{\delta}\sqrt{d}\max\left\{1, 5\frac{L_2\varepsilon_1}{L_1\varepsilon_2}\right\}\left(\Delta_g \max\left\{\frac{L_2^2}{\varepsilon_2^5}, \frac{1}{\varepsilon_1^2\varepsilon_2}\right\} + \frac{1}{\varepsilon_2^2}\right).$$

Note that $\phi \geq 2^{24}\frac{L_1^2}{\varepsilon_2^2} \geq 2^{24}$ since $\varepsilon_2 \leq L_1$. Therefore,

$$p \leq 2^{13} \cdot 800\left(\frac{(1+a)^2}{1-a}\right)^3 \frac{\gamma^6}{2^{24}\log_2^8(\phi)}\delta \leq \delta$$

where the final inequality follows from $\log_2(x\log_2(x)^8)^6 \leq \log_2(x)^8$ for any $x \geq 2^{24}$ and $2^{13} \times 800 \times \left(\frac{(1+a)^2}{1-a}\right)^3 \leq 2^{24}$ since $a \leq 1/20$. □

We assume that G is an (a, b)-inexact gradient oracle for g. We derive two simple consequences of Definition 7.2.3.

Lemma 7.4.2. *Then we have that for any $x \in \mathbf{R}^d$ the following inequalities hold:*

1. ***(Norm similarity)*** $|\|G(x)\| - \|\nabla g(x)\|| \le a\|\nabla g(x)\| + b$.
2. ***(Correlation)*** $\langle \nabla g(x), G(x) \rangle \ge (7/8)(1-a)\|\nabla g(x)\|^2 - 2b^2$.

Proof. Throughout the proof we let $v = \nabla g(x)$ and $u = G(x)$ and use that $\|u - v\| \le a\|v\| + b$. The first part of the theorem is then a consequence of the triangle inequality. The second part follows since $\|u\|^2 \ge (1-a)^2\|v\|^2 - 2b(1-a)\|v\| + b^2$ and

$$\|u\|^2 - 2\langle u, v\rangle + \|v\|^2 = \|u - v\|^2 \le a^2\|v\|^2 + 2ab\|v\| + b^2,$$

which implies the following:

$$\begin{aligned}
2\langle u, v\rangle &\ge (1-a)^2\|v\|^2 + (1-a^2)\|v\|^2 - 2(1-2a)b\|v\| \\
&= 2(1-a)\|v\|^2 - 2(1-2a)b\|v\| \\
&\ge 2(1-a)(1-c)\|v\|^2 - \frac{(1-2a)^2}{2(1-a)c}b^2 \\
&\ge 2(1-a)(1-c)\|v\|^2 - \frac{1}{2c}b^2
\end{aligned}$$

where the third inequality uses $a \le 1/2$ and the second inequality follows from Young's inequality: $2 \cdot ((1-2a)b \cdot \|v\|) \le ((1-2a)b)^2/(2c(1-a)) + 2c(1-a)\|v\|^2$. To complete the result, set $c = 1/8$. □

As a consequence of this Lemma, we prove that the function g decreases along the inexact gradient descent sequences with oracle G.

Lemma 7.4.3 (Descent lemma). *Given $y_0 \in \mathbf{R}^d$, consider the inexact gradient descent sequence: $y_{t+1} \leftarrow y_t - \eta \cdot G_t(y_t)$. Then for all $t \ge 0$, we have*

$$g(y_t) - g(y_0) \le -\frac{\eta}{8}(1-a)\sum_{i=0}^{t-1}\|\nabla g(y_i)\|^2 + 5t\eta b^2. \tag{7.19}$$

Proof. Since the function g has L_1-Lipschitz gradients we have

$$\begin{aligned} g(y_{t+1}) &\le g(y_t) - \eta\langle \nabla g(y_t), G(y_t)\rangle + \frac{L_1\eta^2}{2}\|G(y_t)\|^2 \\ &\le g(y_t) - \eta\frac{7(1-a)}{8}\|\nabla g(y_t)\|^2 + 2\eta b^2 + \frac{L_1\eta^2}{2}\left((1+a)\|\nabla g(y_t)\| + b\right)^2 \\ &\le g(y_t) - \eta\frac{7(1-a)}{8}\|\nabla g(y_t)\|^2 + 2\eta b^2 \\ &\qquad + \frac{L_1\eta^2}{2}\left(\frac{6}{5}(1+a)^2\|\nabla g(y_t)\|^2 + 6b^2\right). \end{aligned}$$

Here the second inequality follows from Lemma 7.4.2 and the third follows from Young's inequality: $2(1+a)\|\nabla g(y_t)\|b \le \frac{1}{5}(1+a)^2\|\nabla g(y_t)\|^2 + 5b^2$. Next, observe that

$$\begin{aligned} &-\eta\frac{7(1-a)}{8}\|\nabla g(y_t)\|^2 + 2\eta b^2 + \frac{L_1\eta^2}{2}\left(\frac{6}{5}(1+a)^2\|\nabla g(y_t)\|^2 + 6b^2\right) \\ &\le -\eta\left(\frac{7(1-a)}{8} - \frac{6}{10}(1+a)^2\right)\|\nabla g(y_t)\|^2 + (2+3)\,\eta b^2 \\ &\le -\frac{\eta(1-a)}{8}\|\nabla g(y_t)\|^2 + 5\eta b^2, \end{aligned}$$

where the second line follows since $\eta \le 1/L_1$ and the last inequality follows from $(6/10)(1+a)^2 \le (3/4)(1-a)$ for $a \le 1/20$. Thus, we have shown that

$$g(y_t) - g(y_0) \le -\frac{\eta(1-a)}{8}\|\nabla g(y_t)\|^2 + 5\eta b^2,$$

which implies (7.19). □

As a consequence of the above Lemma, we now show that inexact gradient descent sequences $\{y_t\}$ either (a) significantly decrease g or (b) remain close to y_0.

Lemma 7.4.4 (Improve or localize). *Given $y_0 \in \mathbf{R}^d$, consider the inexact gradient descent sequence: $y_{t+1} \leftarrow y_t - \eta \cdot G_t(y_t)$. Then, for all $\tau \le t$, we have*

$$\|y_\tau - y_0\|^2 \le 16\eta t\frac{(1+a)^2}{(1-a)}\left(g(y_0) - g(y_t) + (5+\eta)\,tb^2\right). \tag{7.20}$$

Proof. By Lemma 7.4.2, we have

$$\|y_\tau - y_0\|^2 = \eta^2 \left\| \sum_{i=0}^{\tau-1} G(y_i) \right\|^2 \le \eta^2 \left(\sum_{i=0}^{t-1} (1+a) \|\nabla g(y_i)\| + tb \right)^2$$
$$\le 2 \left(t\eta^2 \sum_{i=0}^{t-1} (1+a)^2 \|\nabla g(y_i)\|^2 + \eta^2 t^2 b^2 \right),$$

where the last inequality follows from Jensen's inequality. Next apply Lemma 7.4.3, to bound $\eta^2 \sum_{i=0}^{t-1} \|\nabla g(y_i)\|^2 \le \frac{8\eta}{(1-a)}(g(y_0) - g(y_t) + 5b^2)$. Plugging this bound into the above inequality, we have

$$\|y_\tau - y_0\|^2 \le 2 \left(8\eta t \frac{(1+a)^2}{(1-a)} \left(g(y_0) - g(y_t) + 5b^2 t \right) + \eta^2 t^2 b^2 \right)$$
$$\le 16\eta t \frac{(1+a)^2}{(1-a)} \left(g(y_0) - g(y_t) + (5+\eta)\, tb^2 \right).$$

This concludes the proof. □

In the next two Lemmas, we show that, when randomly initialized near a critical point with negative curvature, inexact gradient descent sequences decrease the objective g with high probability. The first result (Lemma 7.4.5) will help us estimate the failure probability.

Lemma 7.4.5. *Fix a point $\tilde{y}$ satisfying $\|\nabla g(\tilde{y})\| \le \varepsilon_1$ and $\lambda_{\min}(\nabla^2 g(\tilde{y})) \le -\varepsilon_2$ and let e_0 denote an eigenvector associated to the smallest eigenvalue of $\nabla^2 g(\tilde{y})$. Consider two points y_0 and y_0' with*

$$y_0 = y_0' + \eta r_0 e_0 \quad \textit{and} \quad \max\{\|y_0 - \tilde{y}\|, \|y_0' - \tilde{y}\|\} \le \eta r,$$

where $r_0 \ge \omega := \frac{1}{\eta} 2^{3-\gamma} R$. Let $\{y_t\}, \{y_t'\}$ be two inexact gradient descent sequences, initialized at y_0 and y_0', respectively:

$$y_{t+1} = y_t - \eta G(y_t) \qquad \textit{and} \qquad y_{t+1}' = y_t' - \eta G(y_t').$$

Then $\min\{g(y_M) - g(y_0), g(y_M') - g(y_0')\} \le -F$.

Proof. We argue by contradiction. Suppose that

$$\max\{g(y_0) - g(y_M), g(y'_0) - g(y'_M)\} < F.$$

Then by Lemma 7.4.4, the iterates of both sequences remain close to their initializers:

$$\begin{aligned} \max\{\|y_t - y_0\|, \|y'_t - y'_0\|\} &\leq \sqrt{16\eta \frac{(1+a)^2}{(1-a)} M \left(F + (5+\eta) Mb^2\right)} \\ &\leq \sqrt{32\eta \frac{(1+a)^2}{(1-a)} MF}, \qquad \text{for all } t \leq M. \end{aligned} \tag{7.21}$$

where the second inequality follows from two upper bound: $\eta \leq 1/L_1$ and $b^2 \leq \frac{L_1 F}{M(5L_1+1)}$. We now use (7.21) to show for all $t \leq M$, iterates y_t and y'_t remain close to $\tilde{y}$. By Lemma 7.4.1, we get

$$\begin{aligned} \max\{\|y_t - \tilde{y}\|, \left\|y'_t - \tilde{y}\right\|\} &\leq \max\{\|y_t - y_0\|, \|y'_t - y'_0\|\} + \max\{\|y_0 - \tilde{y}\|, \|y'_0 - \tilde{y}\|\} \\ &\leq \sqrt{32\eta \frac{(1+a)^2}{(1-a)} MF} + \eta r < R. \end{aligned} \tag{7.22}$$

In the remainder of the proof, we will argue that inequality (7.22) cannot hold. In particular, we will show that negative curvature of g implies the sequences y_t and y'_t must rapidly diverge from each other.

To leverage negative curvature, we first claim that g is C^2 with L_2-Lipschitz Hessian in $\mathbf{B}_R(\tilde{y})$, which contains y_t and y'_t for $t \leq M$. Indeed, since $\tilde{y}$ satisfies $\|\nabla g(\tilde{y})\| \leq \varepsilon_1 \leq \alpha$, Assumption 7.2.2 ensures $\nabla^2 g(y)$ is defined and L_2-Lipschitz through $B_\beta(\tilde{y})$. The claim then follows since $R = \frac{1}{4\gamma}\frac{\varepsilon_2}{L_2} \leq \beta$, which follows from the assumption $\varepsilon_2 \leq 4\gamma\beta L_2$

Now observe that $\{y'_t + s(y_t - y'_t) \mid s \in [0,1]\} \subseteq \mathbf{B}_R(\tilde{y})$ for all $t \leq M$. Therefore, defining $\mathcal{H} := \nabla^2 g(\tilde{y})$, $v_t := \nabla g(y_t) - G(y_t)$, $v'_t := \nabla g(y'_t) - G(y'_t)$, and $\widehat{y_t} := y_t - y'_t$, we

have for all $t \leq M-1$

$$\begin{aligned}
\hat{y}_{t+1} &= \hat{y}_t - \eta(\nabla g(y_{t+1}) - \nabla g(y'_{t+1})) - \eta(v_t - v'_t) \\
&= (I - \eta\mathcal{H})\hat{y}_t - \eta\left[\int_0^1 \left(\nabla^2 g(y'_t + s(y_t - y'_t)) - \mathcal{H}\right) ds\right]\hat{y}_t - \eta(v_t - v'_t) \\
&= \underbrace{(I - \eta\mathcal{H})^{t+1}\hat{y}_0}_{=:p(t+1)} - \underbrace{\eta\sum^{t}(I - \eta\mathcal{H})^{t-\tau}\left[\int_0^1 \left(\nabla^2 g(y'_\tau + s(y_\tau - y'_\tau)) - \mathcal{H}\right) ds\right]\hat{y}_\tau}_{=:q(t+1)} \\
&\qquad - \underbrace{\eta\sum_{\tau=0}^{t}(I - \eta\mathcal{H})^{t-\tau}(v_\tau - v'_\tau)}_{=:n(t+1)}
\end{aligned}$$

where the last equality follows from the recursive definition of y_t and y'_t. In what follows we will argue that $p(t)$ diverges exponentially and dominates $q(t)$ and $n(t)$.

Beginning with exponential growth, notice that $\widehat{y_0}$ is an eigenvector of $\mathcal{H}$ with eigenvalue $\lambda := -\lambda_{\min}(\mathcal{H})$. Therefore,

$$\|p(t)\| = (1 + \eta\lambda)^t |\widehat{y_0}\| = (1 + \eta\lambda)^t \eta r_0. \tag{7.23}$$

Consequently, if $\max\{\|q(t)\|, 2\|n(t)\|\} \leq \frac{\|p(t)\|}{2}$, then the following bound would hold:

$$\begin{aligned}
\max\{\|y_M - \tilde{y}\|, \|y'_M - \tilde{y}\|\} &\geq \frac{\|\hat{y}_M\|}{2} \\
&\geq \frac{1}{2}\left(\|p(M)\| - \|q(M)\| - \|n(M)\|\right) \\
&\geq \frac{1}{8}\|p(M)\| \\
&= \frac{(1 + \eta\lambda)^M \eta r_0}{8} \\
&\geq 2^{\gamma-3}\eta r_0 \geq R,
\end{aligned}$$

where the fourth inequality follows since $M = \gamma/\eta\varepsilon_2$, $(1 + \eta\lambda) \geq (1 + \eta\varepsilon_2)$ and $(1+x)^{1/x} \geq 2$ for all $x \in (0, 1)$, while the final inequality follows since $r_0 \geq \omega = \frac{R}{2^{\gamma-3}\eta}$.

Thus, by proving the following claim, we will contradict (7.22) and prove the result.

Claim 14. *For all $t \le M$, we have* $\max\{\|q(t)\|, 2\|n(t)\|\} \le \frac{\|p(t)\|}{2}$.

The proof of the claim follows by induction on t and the following bound

$$\|I - \eta\mathcal{H}\| \le (1 + \eta\lambda),$$

which holds since η is small enough that $I - \eta\mathcal{H} \succcurlyeq 0$.

Turning to the inductive proof, we note that the base case holds since

$$2n(0) = q(0) = 0 \le \|\hat{y}_0\|/4.$$

Now assume the claim holds for all $\tau \le t$. Then for all $\tau \le t$ we have

$$\|\hat{y}_\tau\| \le \|p(\tau)\| + \|q(\tau)\| + \|n(\tau)\| \le 2\|p(\tau)\| \le 2(1 + \eta\lambda)^\tau \eta r_0,$$

where the final inequality follows from (7.23). Consequently, we may bound $\|q(t + 1)\|$ as follows:

$$\begin{aligned}
\|q(t+1)\| &\le \eta \sum_{\tau=0}^{t} \|I - \eta\mathcal{H}\|^{t-\tau} \left\| \int_0^1 \left(\nabla^2 g(y'_\tau + s(y_\tau - y'_\tau)) - \mathcal{H}\right) ds \right\| \|\hat{y}_\tau\| \\
&\le \eta L_2 \sum_{\tau=0}^{t} \|I - \eta\mathcal{H}\|^{t-\tau} \max\{\|y_t - \tilde{y}\|, \left\|y'_t - \tilde{y}\right\|\} \|\hat{y}_\tau\| \\
&\le \eta L_2 R \sum_{\tau=0}^{t} \|I - \eta\mathcal{H}\|^{t} \eta r_0 \\
&= \eta L_2 R M \|I - \eta\mathcal{H}\|^{t} \eta r_0 \\
&\le 2\eta L_2 R M \|p(t+1)\| \\
&\le \frac{\|p(t+1)\|}{2},
\end{aligned}$$

where the second inequality follows from L_2-Lipschitz continuity of $\nabla^2 g$ on $B_R(\tilde{y})$, the third inequality follows from the inclusions $y_t, y'_t \in B_R(\tilde{y})$, the fourth in-

equality follows from (7.23), and the fifth inequality follow from $2\eta L_2 RM \leq 1/2$. This proves half of the inductive step.

To prove the other half of the inductive step, we bound $\|n(t+1)\|$ as follows:

$$\begin{aligned}
\|n(t+1)\| &\leq \eta \sum_{\tau=0}^{t} \|I - \eta\mathcal{H}\|^{t-\tau} \left\|v_\tau - v'_\tau\right\| \\
&\leq \eta \sum_{\tau=0}^{t} \|I - \eta\mathcal{H}\|^{t-\tau} \left[a\left(\|\nabla g(y_\tau)\| + \left\|\nabla g(y'_\tau)\right\|\right) + 2b\right] \\
&\leq 2\eta \sum_{\tau=0}^{t} \|I - \eta\mathcal{H}\|^{t-\tau} \left[a\left(L_1 R + \varepsilon_1\right) + b\right] \\
&\leq 2\eta(1+\eta\lambda)^t \Big[Ma\left(L_1 R + \varepsilon_1\right) + Mb\Big]
\end{aligned}$$

where the third inequality follows from L_1 Lipschitz continuity of ∇g, the inclusions $y_t, y'_t \in B_R(\tilde{y})$, and the bound $\|\nabla g(\tilde{y})\| \leq \varepsilon_1$; and the fourth inequality follows from the bound $\|I - \eta\mathcal{H}\|^{t-\tau} \leq (1+\eta\lambda)^t$. To complete the proof, we recall that three inequalities: $b \leq \frac{R}{M\eta 2^{(\gamma+2)}}$, $a \leq \frac{1}{\eta M 2^{\gamma+2}} \min\{\frac{1}{L_1}, \frac{R}{\varepsilon_1}\}$, and $r_0 \geq \omega = \frac{R}{2^{\gamma-3}\eta}$. Then, we find that

$$\begin{aligned}
\|n(t+1)\| &\leq 2\eta(1+\eta\lambda)^t \Big[Ma\left(L_1 R + \varepsilon_1\right) + Mb\Big] \\
&\leq \frac{3(1+\eta\lambda)^t R}{2^{\gamma+1}} \\
&\leq \frac{3(1+\eta\lambda)^t \eta r_0}{16} \\
&\leq \|p(t+1)\|/4.
\end{aligned}$$

This concludes the proof of the claim. Consequently, the proof of the Lemma is complete. □

Using the Lemma 7.4.5, the following Lemma proves that inexact gradient descent will decrease the objective value by a large amount if it is randomly initialized near a point with negative curvature.

Lemma 7.4.6 (Descent with negative curvature). *Fix a point $\tilde{y}$ satisfying $\|\nabla g(\tilde{y})\| \leq \varepsilon_1$ and $\lambda_{\min}(\nabla^2 g(\tilde{y})) \leq -\varepsilon_2$. Consider an initial point $y_0 := \tilde{y} + \eta \cdot u$ with $u \sim Unif(r\mathbf{B})$. Let $\{y_t\}$ be an inexact gradient descent sequence, initialized at y_0:*

$$y_{t+1} = y_t - \eta G(y_t).$$

Then with probability at least

$$p := 1 - L_1 \frac{(1+a)^2}{(1-a)} \frac{\sqrt{d}}{\varepsilon_2} \gamma^2 \max\left\{1, 5\frac{L_2\varepsilon_1}{L_1\varepsilon_2}\right\} 2^{9-\gamma}, \tag{7.24}$$

we have $g(y_M) - g(\tilde{y}) \leq -F/2$

Proof. We show that the bound $g(y_M) - g(\tilde{y}) \leq -F/2$ follows from the inequality $g(y_M) - g(y_0) \leq -F$. To that end, first observe that.

$$g(y_0) - g(\tilde{y}) \leq \langle \nabla g(\tilde{y}), y_0 - \tilde{y} \rangle + \frac{L_1\eta^2}{2}\|y_0 - \tilde{y}\|^2 \leq \varepsilon_1 \eta r + \frac{L_1\eta^2 r^2}{2} \leq -F/2$$

where the last inequality follows by Lemma 7.4.1. Consequently,

$$g(y_M) - g(\tilde{y}) \leq g(y_M) - g(y_0) + g(y_0) - g(\tilde{y}) \leq -F/2.$$

This shows that it is sufficient to study $g(y_M) - g(y_0) \leq -F$ as desired.

In the remainder of the proof, we show the event $\{g(y_M) - g(y_0) \leq -F\}$ holds with the claimed probability in (7.24). To that end, given any $y_0' \in \mathbf{R}^d$, let us define $T_M(y_0') = y_M'$, where $y_{t+1}' = y_t' - \eta G(y_t')$ for all $t \geq 0$. Consider the set of points $y \in \mathbf{B}_{\eta r}(\tilde{y})$, for which M steps of the inexact gradient method with oracle G fail to decrease the g significantly:

$$\mathcal{X}_{\text{stuck}} = \{y \in \mathbf{B}_{\eta r}(\tilde{y}) \mid g(T_M(y)) - g(y_0) > -F\}.$$

We now show that $P(y_0 \in \mathcal{X}_{\text{stuck}}) \leq 1 - p$. Indeed, Lemma 7.4.5 shows that there exists $e_0 \in \mathbb{S}^{d-1}$ such that width of $\mathcal{X}_{\text{stuck}}$ along e_0 is upper bounded by $\eta\omega$. Thus

the volume of $\mathcal{X}_{\text{stuck}}$ is bounded by the volume of the cylinder $[0, \omega] \times \mathbf{B}_{\eta r}^{d-1}(0)$, which yields the result:

$$\begin{aligned}
\mathbb{P}(y_0 \in \mathcal{X}_{\text{stuck}}) = \frac{\text{Vol}(\mathcal{X}_{\text{stuck}})}{\text{Vol}(B_{\eta r}^d(0))} &\leq \frac{\eta\omega \cdot \text{Vol}(\eta r \mathbf{B}^{d-1})}{\text{Vol}(\eta r \mathbf{B}^d)} \\
&\leq \frac{\omega \cdot \Gamma\left(\frac{d+1}{2} + \frac{1}{2}\right)}{r\sqrt{\pi}\Gamma\left(\frac{d+1}{2}\right)} \\
&\leq \frac{\omega}{r} \cdot \sqrt{\frac{d}{\pi}} \\
&\leq \frac{2^{3-\gamma}R}{\eta r} \cdot \sqrt{\frac{d}{\pi}} \\
&\leq L_1 \frac{(1+a)^2}{(1-a)} \frac{\sqrt{d}}{\varepsilon_2} \gamma^2 \max\left\{1, 5\frac{L_2\varepsilon_1}{L_1\varepsilon_2}\right\} 2^{9-\gamma}.
\end{aligned}$$

where the second inequality follows from the identity $\text{Vol}(\eta r \mathbf{B}^d) = (\eta r)^d \pi^{d/2} / \Gamma(\frac{d}{2} + 1)$; the third inequality follows from the bound $\Gamma(x + \frac{1}{2})/\Gamma(x) \leq \sqrt{x}$ for any $x \geq 0$ [121]; the fourth inequality follows from the definition $\omega = \frac{R}{2^{\gamma-3}\eta}$; and the fifth inequality follows from the definitions $\eta = (1-a)/L_1(1+a)^2$, $R = \frac{1}{4\gamma}\frac{\varepsilon_2}{L_2}$, and $r = \frac{\varepsilon_2^2}{400 L_2 \gamma^3} \min\left\{1, \frac{L_1\varepsilon_2}{5\varepsilon_1 L_2}\right\}$, as well as the bound $400 \cdot 2^3/(4\sqrt{\pi}) \leq 2^9$. This concludes the proof. □

To conclude this section, we now combine all the Lemmas to prove Theorem 7.2.4.

Proof of Theorem 7.2.4. Set the number of iterations to

$$T = 8\Delta_g \max\left\{\frac{M}{F}, \frac{256}{\eta\varepsilon_1^2}\right\} + 4M.$$

Then, we will prove the slightly stronger claim that there is at least one $(\varepsilon_1/4, \varepsilon_2)$-second-order critical point. Let $\{x_t\}_{t=0}^T$ be the sequence generated by Algorithm 7. We partition this sequence into three disjoint sets:

1. The set of $(\varepsilon_1/4, \varepsilon_2)$-second-order critical points, denoted $\mathcal{S}_2$.

2. The set of $(\varepsilon_1/4)$-first-order critical points that are not in $\mathcal{S}_2$, denoted $\mathcal{S}_1$.

3. All the other points $\mathcal{S}_3 = \{x_t\}_{t=0}^T \setminus (\mathcal{S}_1 \cup \mathcal{S}_2)$.

We first prove that $|\mathcal{S}_3| \le T/4$:

$$\begin{aligned} g(x_T) - g(x_0) &= \sum_{t=0}^{T-1} (g(x_{t+1}) - g(x_t)) \\ &\le -\eta \frac{(1-a)}{8} \sum_{t=0}^{T-1} \|\nabla g(x_t)\|^2 + 5\eta T b^2 \\ &\le -\eta \frac{(1-a)}{8} \sum_{t \in \mathcal{S}_3} \|\nabla g(x_t)\|^2 + 5\eta T b^2 \\ &< -\eta |\mathcal{S}_3| \varepsilon_1^2 (1-a) \frac{1}{128} + 5\eta T b^2 \end{aligned}$$

Rearranging, and applying $b^2 \le \frac{\varepsilon_1^2}{4096}$, we find

$$|\mathcal{S}_3| \le \frac{g(x_0) - g(x_T)}{\eta \varepsilon_1^2 (1-a)\frac{1}{128}} + \frac{5Tb^2}{\varepsilon_1^2 (1-a)\frac{1}{128}} \le \frac{T}{(1-a)16} + \frac{640T}{(1-a)4096} \le T/4,$$

since $a \le 1/20$.

Now suppose for the sake of contradiction that $|\mathcal{S}_2|$ is empty. Define $\Gamma \subset [T]$ be the set of iteration numbers where Algorithm 7 adds a perturbation to the iterate:

$$\Gamma := \{t \in [T] \mid \|G(x_t)\| \le \varepsilon_1/2 \text{ and } t - t_{\text{pert}} \ge M\}.$$

Every x_t with $t \in \Gamma$ is first-order stationary, since

$$\|\nabla g(x_t)\| \le \frac{1}{1-a} (\|G(x_t)\| + b) \le \frac{1}{1-a} \left(\frac{\varepsilon_1}{2} + b\right) \le \frac{20}{19} \left(\frac{\varepsilon_1}{2} + \frac{\varepsilon_1}{64}\right) \le \varepsilon_1.$$

Moreover, since $|\mathcal{S}_2|$ is empty, such x_t satisfy $\lambda_{\min}(\nabla^2 g(x_t)) < -\varepsilon_2$. Therefore, by Lemma 7.4.6 and a union bound, the following event

$$\mathcal{E} = \left\{ g(x_{t+M}) - g(x_t) \le -\frac{F}{2} \quad \text{for all } t \in \Gamma \right\}$$

does not happen with probability at most

$$\mathbb{P}(\mathcal{E}^c) \leq \frac{TL_1 \frac{(1+a)^2}{(1-a)} \frac{\sqrt{d}}{\varepsilon_2} \gamma^2 \max\left\{1, 5\frac{L_2\varepsilon_1}{L_1\varepsilon_2}\right\} 2^9}{2^\gamma}. \tag{7.25}$$

By Lemma 7.4.1, this probability is upper bounded by δ. Therefore, throughout the remainder of the proof, we suppose the event $\mathcal{E}$ happens. In this event we will show that we will show that $g(x_t) < \inf g$ for some t, which yields the desired contradiction.

To that end, recall that by Lemma 7.4.3, g cannot increase by much at each iteration:

$$g(x_{t+1}) - g(x_t) \leq 5\eta b^2 \qquad \text{for all } t \in [T].$$

Thus, defining $t_{\text{last}} := \max\{t \mid t + M < T\}$ and we find that

$$\begin{aligned}
g(x_{t_{\text{last}}+M+1}) - g(x_0) &= \sum_{t=0}^{t_{\text{last}}+M} (g(x_{t+1}) - g(x_t)) \\
&\leq \sum_{\substack{k\in\Gamma \\ k\leq t_{\text{last}}}} \sum_{t\in[k,k+M-1]} (g(x_{t+1}) - g(x_t)) + 5\eta b^2|T| \\
&= \sum_{\substack{k\in\Gamma \\ k\leq t_{\text{last}}}} (g(x_{t+M}) - g(x_t)) + 5\eta b^2|T| \\
&\leq -(|\Gamma| - 1)F/2 + 5\eta b^2|T|
\end{aligned}$$

To arrive at the desired contradiction, we will show that $|\Gamma|$ is large. In particular, we claim that

$$|\Gamma| \geq \frac{3T}{4M}.$$

To prove this claim, first observe that the definition of Algorithm 7 ensures that $\{x_t \mid \|G(x_t)\| \leq \varepsilon_1/2\} \subseteq \bigcup_{k\in\Gamma}\{k, \ldots, k+M\}$. Moreover, $\mathcal{S}_1 \subseteq \{x_t \mid \|G(x_t)\| \leq \varepsilon_1/2\}$ by Lemma 7.4.2:

$$\|\nabla g(x_t)\| \leq \varepsilon_1/4 \implies \|G(x)\| \leq (1+a)\frac{\varepsilon_1}{4} + b \leq \frac{21}{20}\frac{\varepsilon_1}{4} + \frac{\varepsilon_1}{64} \leq \frac{\varepsilon_1}{2},$$

since $a \leq 1/20$ and $b \leq \varepsilon_1/64$. Therefore, since $|\mathcal{S}_1| = T - |\mathcal{S}_3| \geq 3T/4$, we have $(3T/4) \leq |\mathcal{S}_1| \leq |\Gamma|M$, as desired.

Finally, we find

$$\begin{aligned}
&g(x_{t_{\text{last}}+M+1}) - g(x_0) \\
&\leq -(|\Gamma| - 1)F/2 + 5\eta b^2|T| \\
&\leq -\left(\frac{3T}{4M} - 1\right)\frac{F}{2} + 5\eta b^2|T| \\
&\leq -\frac{TF}{4M} + 5\eta b^2|T| \\
&\leq -\frac{TF}{8M} < \inf g - g(x_0),
\end{aligned}$$

where the third inequality follows since $T \geq 4M$ and the fourth inequality follows since $b^2 \leq \frac{1}{40\eta}\frac{F}{M}$. Thus, yielding a contradiction. This completes the proof. □

7.4.2 Proof of Proposition 7.3.2

To prove Part 1, recall that $\|\nabla f_\mu(x)\| = \mu^{-1}(x - \hat{x})$, so

$$\|x - \hat{x}\| \leq \mu\|\nabla f_\mu(x)\| \leq \mu\varepsilon_1,$$

as desired. Note that this implies $x \in \mathcal{U} = B_{3\varepsilon_2/4L_2}(\hat{x})$ since $\varepsilon_1 \leq \frac{\varepsilon_2}{2L_2\mu}$.

To prove the remaining statements, we recall the following consequence of the L_2-Lipschitz continuity of $\nabla^2 f_\mu$ on the ball $\mathbf{B}_\beta(x)$ [186, Lemma 1.2.4]: for all $y \in \mathbf{B}_\beta(x)$

$$f_\mu(x) + \langle \nabla f_\mu(x), y - x\rangle + \frac{1}{2}\langle \nabla^2 f_\mu(x)(y - x), y - x\rangle - \frac{L_2}{6}\|y - x\|^3 \leq f_\mu(y).$$

Since x is an $(\varepsilon_1, \varepsilon_2)$-second order critical point, we may lower bound the left hand side by a simple quadratic: letting $r = 3\varepsilon_2/2L_2$, we have

$$q_0(y) := f_\mu(x) + \langle \nabla f_\mu(x), y - x\rangle - \frac{3}{4}\varepsilon_2\|y - x\|^2 \le f_\mu(y) \quad \text{for all } y \in \mathbf{B}_r(x) \tag{7.26}$$

Now, define the quadratic

$$q(y) := f(\widehat{x}) - \frac{\mu}{2}(1 + 3\mu\varepsilon_2)\,\varepsilon_1^2 + \langle \nabla f_\mu(x), y - \widehat{x}\rangle - \frac{3\varepsilon_2}{2}\|y - \hat{x}\|^2$$

We claim that $q(y) \le q_0(y)$.

Indeed, first observe that by $\nabla f_\mu(x) = \mu(x - \hat{x})$, we have

$$f_\mu(x) + \langle \nabla f_\mu(x), y - x\rangle = f(\widehat{x}) - \frac{1}{2\mu}\|x - \widehat{x}\|^2 + \langle \nabla f_\mu(x), y - \widehat{x}\rangle.$$

Next, we may recenter the quadratic up to a small error:

$$\|y - x\|^2 \le 2\|y - \hat{x}\|^2 + 2\|x - \hat{x}\|^2$$

Therefore, we have

$$\begin{aligned} q_0(y) &= f(\widehat{x}) - \frac{1}{2\mu}\|x - \widehat{x}\|^2 + \langle \nabla f_\mu(x), y - \widehat{x}\rangle - \frac{3\varepsilon_2}{4}\|y - x\|^2 \\ &\ge f(\widehat{x}) - \frac{1}{2}\left(\mu^{-1} + 3\varepsilon_2\right)\|x - \widehat{x}\|^2 + \langle \nabla f_\mu(x), y - \widehat{x}\rangle - \frac{3\varepsilon_2}{2}\|y - \hat{x}\|^2 \ge q(y), \end{aligned}$$

where the third inequality follows from the bound $\|\hat{x} - x\|^2 \le \mu^2\varepsilon_1^2$. This proves the claim.

We now prove the remaining parts of the claim. First, Part 2 follows from (7.27) since $\mathcal{U} \subseteq \mathbf{B}_r(x)$ and $q(y) \le q_0(y) \le f_\mu(y) \le f(y)$ for all $y \in \mathbf{B}_r(x)$. Second, Part 3 follows since $\nabla q(\hat{x}) = \nabla f_\mu(x)$. Finally Parts 4 and 5 follow by direct computation.

7.4.3 Proof of Proposition 7.3.2

To prove Part 1, recall that $\|\nabla f_\mu(x)\| = \mu^{-1}(x - \hat{x})$, so

$$\|x - \hat{x}\| \le \mu\|\nabla f_\mu(x)\| \le \mu\varepsilon_1,$$

as desired. Note that this implies $x \in \mathcal{U} = B_{3\varepsilon_2/4L_2}(\hat{x})$ since $\varepsilon_1 \le \frac{\varepsilon_2}{2L_2\mu}$.

To prove the remaining statements, we recall the following consequence of the L_2-Lipschitz continuity of $\nabla^2 f_\mu$ on the ball $\mathbf{B}_\beta(x)$ [186, Lemma 1.2.4]: for all $y \in \mathbf{B}_\beta(x)$

$$f_\mu(x) + \langle \nabla f_\mu(x), y - x\rangle + \frac{1}{2}\langle \nabla^2 f_\mu(x)(y - x), y - x\rangle - \frac{L_2}{6}\|y - x\|^3 \le f_\mu(y).$$

Since x is an $(\varepsilon_1, \varepsilon_2)$-second order critical point, we may lower bound the left hand side by a simple quadratic: letting $r = 3\varepsilon_2/2L_2$, we have

$$q_0(y) := f_\mu(x) + \langle \nabla f_\mu(x), y - x\rangle - \frac{3}{4}\varepsilon_2\|y - x\|^2 \le f_\mu(y) \quad \text{for all } y \in \mathbf{B}_r(x) \tag{7.27}$$

Now, define the quadratic

$$q(y) := f(\widehat{x}) - \frac{\mu}{2}(1 + 3\mu\varepsilon_2)\,\varepsilon_1^2 + \langle \nabla f_\mu(x), y - \widehat{x}\rangle - \frac{3\varepsilon_2}{2}\|y - \hat{x}\|^2$$

We claim that $q(y) \le q_0(y)$.

Indeed, first observe that by $\nabla f_\mu(x) = \mu(x - \hat{x})$, we have

$$f_\mu(x) + \langle \nabla f_\mu(x), y - x\rangle = f(\widehat{x}) - \frac{1}{2\mu}\|x - \widehat{x}\|^2 + \langle \nabla f_\mu(x), y - \widehat{x}\rangle.$$

Next, we may recenter the quadratic up to a small error:

$$\|y - x\|^2 \le 2\|y - \hat{x}\|^2 + 2\|x - \hat{x}\|^2$$

Therefore, we have

$$\begin{aligned} q_0(y) &= f(\widehat{x}) - \frac{1}{2\mu}\|x - \widehat{x}\|^2 + \langle \nabla f_\mu(x), y - \widehat{x}\rangle - \frac{3\varepsilon_2}{4}\|y - x\|^2 \\ &\ge f(\widehat{x}) - \frac{1}{2}\left(\mu^{-1} + 3\varepsilon_2\right)\|x - \widehat{x}\|^2 + \langle \nabla f_\mu(x), y - \widehat{x}\rangle - \frac{3\varepsilon_2}{2}\|y - \hat{x}\|^2 \ge q(y), \end{aligned}$$

where the third inequality follows from the bound $\|\hat{x} - x\|^2 \leq \mu^2 \varepsilon_1^2$. This proves the claim.

We now prove the remaining parts of the claim. First, Part 2 follows from (7.27) since $\mathcal{U} \subseteq \mathbf{B}_r(x)$ and $q(y) \leq q_0(y) \leq f_\mu(y) \leq f(y)$ for all $y \in \mathbf{B}_r(x)$. Second, Part 3 follows since $\nabla q(\hat{x}) = \nabla f_\mu(x)$. Finally Parts 4 and 5 follow by direct computation.

7.4.4 Proof of Theorem 7.3.3

By [84, Theorem 3.7], there exist disjoint open sets $\{V_1, \ldots, V_k\}$ in $\mathbf{R}^d$, whose union has full measure in $\mathbf{R}^d$, and such that for each $i = 1, \ldots, k$, there exist finitely many smooth maps $g_1, \ldots, g_m$ satisfying

$$(\partial f)^{-1}(v) = \{g_1(v), \ldots, g_m(v)\} \qquad \forall v \in V_i.$$

In particular, since g_i are locally Lipschitz continuous, for every $v \in V_i$, there exists a constant ℓ satisfying

$$(\partial f)^{-1}(\mathbf{B}_\epsilon(v)) \subset \bigcup_{j=m}^{k} \mathbf{B}_{\ell\epsilon}(g_j(v)), \tag{7.28}$$

for all small $\epsilon > 0$. Moreover, by [84, Corollary 4.8] we may assume that for every point v in V_i and for sufficiently small $\epsilon > 0$ the set $g_j(\mathbf{B}_\epsilon(v))$ is an active manifold around $g_j(v)$ for the tilted function $f(\cdot; v) = f(\cdot) - \langle v, \cdot \rangle$. Taking into account [67, Theorem 3.1], we may also assume that the Moreau envelope $f_\mu(\cdot; v)$ of $f(\cdot; v)$ is C^p-smooth on a neighborhood of each point $g_j(v)$.

Fix now a set V_i a point $v \in V_i$. Clearly, then there exist constants $r, \beta, L_2 > 0$, such that for any point y with $\operatorname{dist}(y, (\partial f)^{-1}(v)) \leq r$, the Hessian $\nabla^2 f_\mu(\cdot; v)$ is L_2-Lipschitz on the ball $\mathbf{B}_\beta(y)$. It remains to show that for all sufficiently small

$\alpha > 0$, any point y satisfying $\|\nabla f_\mu(y; v)\| \leq \alpha$ also satisfies $\operatorname{dist}(y, (\partial f)^{-1}(v)) \leq r$. To this end, consider a point y with $\|\nabla f_\mu(y; v)\| \leq \alpha$ for some $\alpha > 0$. Note the proximal point $\hat{y}$ of $f_\mu(\cdot; v)$ at y then satisfies

$$\operatorname{dist}(v, \partial f(\hat{y})) \leq \alpha \qquad \text{and} \qquad \|\hat{y} - y\| \leq \mu \cdot \alpha.$$

Therefore we deduce, $\hat{y} \in (\partial f)^{-1}(\mathbf{B}_\alpha(v))$ and $\operatorname{dist}(y, (\partial f)^{-1}(\mathbf{B}_\alpha(v)) \leq \mu \cdot \alpha$. Thus, using (7.28) we deduce that for sufficiently small $\alpha > 0$, we have

$$\operatorname{dist}(y, (\partial f)^{-1}(v)) \leq (\mu + \ell) \cdot \alpha.$$

Choosing $\alpha < r/(\mu + \ell)$ completes the proof.

7.4.5 Proof of Theorem 7.3.8

The proof of the theorem is a consequence of the following Lemma.

Lemma 7.4.7. *Assume that $g\colon \mathbf{R}^d \to \mathbf{R} \cup +\infty$ is α-strongly convex with minimizer $x^\star$. Let $g_x\colon \mathbf{R}^d \to \bar{\mathbb{R}}$ be a family of* convex models *satisfying Assumption 7.3.7. Let $x_0 \in \mathbf{R}^d$, let $\theta > q$, and consider the following sequence:*

$$x_{k+1} \leftarrow \operatorname*{arg\,min}_{x \in \mathbf{R}^d} \left\{ g_{x_k}(x) + \frac{\theta}{2}\|x - x_k\|^2 \right\}$$

Then

$$\|x_{k+1} - x^\star\| \leq \left(\frac{\theta + q}{\alpha + \theta}\right)^{\frac{k+1}{2}} \|x_0 - x^\star\|. \tag{7.29}$$

Proof. By θ-strong convexity and quadratic accuracy, we have

$$\begin{aligned}\left(g_{x_k}(x_{k+1}) + \frac{\theta}{2}\|x_k - x_{k+1}\|^2\right) + \frac{\theta}{2}\|x^* - x_{k+1}\|^2 &\leq g_{x_k}(x^*) + \frac{\theta}{2}\|x^* - x_k\|^2 \\ &\leq g(x^*) + \frac{\theta + q}{2}\|x^* - x_k\|^2.\end{aligned}$$

From $g(x_{k+1}) \leq g_{x_k}(x_{k+1}) + \frac{\theta}{2}\|x_k - x_{k+1}\|^2$ and and the above inequality, we have

$$g(x_{k+1}) + \frac{\theta}{2}\|x^* - x_{k+1}\|^2 \leq g(x^*) + \frac{\theta + q}{2}\|x^* - x_k\|^2$$

Subtract $g(x^*)$ from both sides and use $g(x_{k+1}) - g(x^*) \geq \frac{\alpha}{2}\|x_{k+1} - x^*\|^2$ to get the result. □

To complete the proof notice that both the function $g(y) = f + \frac{1}{2\mu}\|y - x_0\|^2$ and the models $g_x = f_x + \frac{1}{2\mu}\|y - x_0\|^2$ are $\alpha = (\mu^{-1} - \rho)$-strongly convex. Therefore, Theorem 7.3.8 follows from an application of Lemma 7.4.7.

BIBLIOGRAPHY

[1] Libsvm data: Classification (binary class). `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`. Accessed: 2021-05-31.

[2] Tobias Achterberg. Conflict analysis in mixed integer programming. *Discrete Optimization*, 4(1):4 – 20, 2007. Mixed Integer Programming.

[3] Naman Agarwal, Nicolas Boumal, Brian Bullins, and Coralia Cartis. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, pages 1–50.

[4] Naman Agarwal, Zeyuan Allen Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In Hamed Hatami, Pierre McKenzie, and Valerie King, editors, *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 1195–1199. ACM, 2017.

[5] Alireza Aghasi, Ali Ahmed, and Paul Hand. Branchhull: Convex bilinear inversion from the entrywise product of signals with known signs. *arXiv:1702.04342*, 2017.

[6] Alireza Aghasi, Ali Ahmed, Paul Hand, and Babhru Joshi. A convex program for bilinear inversion of sparse vectors. *arXiv:1809.08359*, 2018.

[7] Alireza Aghasi, Ali Ahmed, Paul Hand, and Babhru Joshi. Bilinear compressed sensing under known signs via convex programming. *IEEE Transactions on Signal Processing*, 68:6366–6379, 2020.

[8] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

[9] Ali Ahmed, Alireza Aghasi, and Paul Hand. Blind deconvolutional phase retrieval via convex programming. *arXiv:1806.08091*, 2018.

[10] F. Al-Khayyal and J. Kyparisis. Finite convergence of algorithms for nonlinear programs and variational inequalities. *J. Optim. Theory Appl.*, 70(2):319–332, 1991.

[11] P. Albano and P. Cannarsa. Singularities of semiconcave functions in Banach spaces. In *Stochastic analysis, control, optimization and applications*, Systems Control Found. Appl., pages 171–190. Birkhäuser Boston, Boston, MA, 1999.

[12] Pierre Apkarian, Dominikus Noll, and Olivier Prot. A Proximity Control Algorithm to Minimize Nonsmooth and Nonconvex Semi-infinite Maximum Eigenvalue Functions. *J. Convex Anal.*, 16(3-4):641–666, 2009.

[13] David Applegate, Mateo Díaz, Haihao Lu, and Miles Lubin. Infeasibility detection with primal-dual hybrid gradient for large-scale linear programming. *arXiv preprint arXiv:2102.04592*, 2021.

[14] David Applegate, Mateo Díaz, Oliver Hinder, Haihao Lu, Miles Lubin, Brendan O'Donoghue, and Warren Schudy. Practical large-scale linear programming using primal-dual hybrid gradient, 2021.

[15] JB Bailion, Ronald E Bruck, and Simeon Reich. On the asymptotic behavior of nonexpansive mappings and semigroups in banach spaces. *Houston Journal of Mathematics*, 4(1):1–9, 1978.

[16] Goran Banjac, Paul Goulart, Bartolomeo Stellato, and Stephen Boyd. Infeasibility detection in the alternating direction method of multipliers for convex optimization. *Journal of Optimization Theory and Applications*, 183(2):490–519, 2019.

[17] Kinjal Basu, Amol Ghoting, Rahul Mazumder, and Yao Pan. ECLIPSE: An extreme-scale linear program solver for web-applications. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 704–714, Virtual, 13–18 Jul 2020. PMLR.

[18] Heinz H Bauschke, Patrick L Combettes, and D Russell Luke. Finding best approximation pairs relative to two closed convex sets in hilbert spaces. *Journal of Approximation theory*, 127(2):178–192, 2004.

[19] Heinz H Bauschke and Walaa M Moursi. The Douglas–Rachford algorithm for two (not necessarily intersecting) affine subspaces. *SIAM Journal on Optimization*, 26(2):968–985, 2016.

[20] Heinz H Bauschke, Xianfu Wang, and Liangjin Yao. General resolvents for monotone operators: characterization and extension. *arXiv preprint arXiv:0810.3905*, 2008.

[21] Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[22] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

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

[24] J.M. Borwein and A.S. Lewis. *Convex analysis and nonlinear optimization*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 3. Springer-Verlag, New York, 2000. Theory and examples.

[25] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[26] Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.

[27] Oliver Bunk, Ana Diaz, Franz Pfeiffer, Christian David, Bernd Schmitt, Dillip K Satapathy, and J Friso Van Der Veen. Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. *Acta Crystallographica Section A: Foundations of Crystallography*, 63(4):306–314, 2007.

[28] J.V. Burke. Descent methods for composite nondifferentiable optimization problems. *Math. Programming*, 33(3):260–279, 1985.

[29] J.V. Burke. On the identification of active constraints. II. The nonconvex case. *SIAM J. Numer. Anal.*, 27(4):1081–1103, 1990.

[30] J.V. Burke and M.C. Ferris. Weak sharp minima in mathematical programming. *SIAM J. Control Optim.*, 31(5):1340–1359, 1993.

[31] J.V. Burke and M.C. Ferris. A Gauss-Newton method for convex composite optimization. *Math. Programming*, 71(2, Ser. A):179–194, 1995.

[32] J.V. Burke and J.J. Moré. On the identification of active constraints. *SIAM J. Numer. Anal.*, 25(5):1197–1211, 1988.

[33] T.T. Cai and A. Zhang. ROP: matrix recovery via rank-one projections. *Ann. Statist.*, 43(1):102–138, 2015.

[34] P.H. Calamai and J.J. Moré. Projected gradient methods for linearly constrained problems. *Math. Prog.*, 39(1):93–116, 1987.

[35] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[36] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[37] E.J. Candès, Y.C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM J. Imaging Sci.*, 6(1):199–225, 2013.

[38] E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):Art. 11, 37, 2011.

[39] E.J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015.

[40] E.J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[41] E.J. Candes, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[42] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

[43] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[44] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[45] V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, and A.S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21(2):572–596, 2011.

[46] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[47] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy. Composite optimization for robust blind deconvolution. *arXiv:1901.01624*, 2019.

[48] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, pages 1–89, 2021.

[49] Vasileios Charisopoulos, Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Composite optimization for robust rank one bilinear sensing. *Information and Inference: A Journal of the IMA*, 2020.

[50] Y. Chen and E.J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Comm. Pure Appl. Math.*, 70(5):822–883, 2017.

[51] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, Feb 2019.

[52] Y. Chen, Y. Chi, and A.J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Trans. Inform. Theory*, 61(7):4034–4059, 2015.

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

[54] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[55] John W. Chinneck. Computer codes for the analysis of infeasible linear programs. *The Journal of the Operational Research Society*, 47(1):61–72, 1996.

[56] Sunav Choudhary and Urbashi Mitra. Sparse blind deconvolution: What cannot be done. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 3002–3006. IEEE, 2014.

[57] Diego Cifuentes. Burer-monteiro guarantees for general semidefinite programs. *arXiv preprint arXiv:1904.07147*, 2019.

[58] Chris Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. In Wallach et al. [237], pages 5985–5995.

[59] Frank E Curtis, Daniel P Robinson, Clément W Royer, and Stephen J Wright. Trust-region newton-cg with strong second-order complexity guarantees for nonconvex optimization. *SIAM Journal on Optimization*, 31(1):518–544, 2021.

[60] Frank E. Curtis, Daniel P. Robinson, and Mohammadreza Samadi. A trust region algorithm with a worst-case iteration complexity of $O(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, 162(1-2):1–32, May 2016.

[61] Hadi Daneshmand, Jonas Moritz Kohler, Aurélien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1163–1172. PMLR, 2018.

[62] G.B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, Princeton, 1963.

[63] George B Dantzig. Origins of the simplex method. In *A history of scientific computing*, pages 141–151. Association for Computing Machinery, New York, NY, USA, 1990.

[64] M.A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, June 2016.

[65] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[66] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of

weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[67] D. Davis and D. Drusvyatskiy. Proximal methods avoid active strict saddles of weakly convex functions. *To appear in Foundations of Computational Mathematics*, 2021.

[68] D. Davis, D. Drusvyatskiy, K.J. MacPhee, and C. Paquette. Subgradient methods for sharp weakly convex functions. *J. Optim. Theory Appl.*, 179(3):962–982, 2018.

[69] Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Escaping strict saddle points of the Moreau envelope in nonsmooth optimization. 2021.

[70] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, 2020.

[71] Damek Davis and Wotao Yin. Convergence rate analysis of several splitting schemes. In *Splitting methods in communication, imaging, science, and engineering*, pages 115–163. Springer, 2016.

[72] Welington de Oliveira. Proximal bundle methods for nonsmooth DC programming. *J. Glob. Optim.*, 75(2):523–563, 2019.

[73] Welington de Oliveira, Claudia A. Sagastizábal, and Claude Lemaréchal. Convex proximal bundle methods in depth: a unified analysis for inexact oracles. *Math. Program.*, 148(1-2):241–277, 2014.

[74] Welington de Oliveira and Mikhail Solodov. *Bundle Methods for Inexact Data*, pages 417–459. Springer International Publishing, Cham, 2020.

[75] Mateo Díaz. The nonsmooth landscape of blind deconvolution. *Workshop on Optimization for Machine Learning*, 2019.

[76] Mateo Díaz and Benjamin Grimmer. Optimal convergence rates for the proximal bundle method. *arXiv preprint arXiv:2105.07874*, 2021.

[77] Mateo Díaz, Mauricio Junca, Felipe Rincón, and Mauricio Velasco. Compressed sensing of data with a known distribution. *Applied and Computational Harmonic Analysis*, 45(3):486–504, 2018.

[78] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 66(11):7274–7301, 2020.

[79] Lijun Ding and Benjamin Grimmer. Revisit of spectral bundle methods: Primal-dual (sub)linear convergence rates, 2020.

[80] D.L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[81] Jim Douglas and Henry H Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439, 1956.

[82] D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*, 43(3):919–948, 2018.

[83] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Math. Prog.*, pages 1–56, 2018.

[84] Dmitriy Drusvyatskiy, Alexander D Ioffe, and Adrian S Lewis. Generic minimizing behavior in semialgebraic optimization. *SIAM Journal on Optimization*, 26(1):513–534, 2016.

[85] Dmitriy Drusvyatskiy, Alexander D. Ioffe, and Adrian S. Lewis. Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria. *Math. Program.*, 185(1-2):357–383, 2021.

[86] Dmitriy Drusvyatskiy and Adrian S Lewis. Optimality, identifiability, and sensitivity. *Mathematical Programming*, 147(1-2):467–498, 2014.

[87] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Barnabás Póczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. *Advances in Neural Information Processing Systems*, 2017:1068–1078, 2017.

[88] Yu Du and Andrzej Ruszczyński. Rate of Convergence of the Bundle Method. *J. Optim. Theory Appl.*, 173(3):908–922, June 2017.

[89] J.C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities:

composite optimization for robust phase retrieval. *IMA J. Information and Inference, doi:10.1093/imaiai/iay015*, 2018.

[90] J.C. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM J. Optim.*, 28(4):3229–3259, 2018.

[91] J.C. Dunn. On the convergence of projected gradient processes to singular critical points. *J. Optim. Theory Appl.*, 55(2):203–216, 1987.

[92] Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

[93] Jonathan Eckstein and Gyorgy Matyasfalvi. Efficient distributed-memory parallel matrix-vector multiplication with wide or tall unstructured sparse matrices. *CoRR*, abs/1812.00904, 2018.

[94] Y.C. Eldar and S. Mendelson. Phase retrieval: stability and recovery guarantees. *Appl. Comput. Harmon. Anal.*, 36(3):473–494, 2014.

[95] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[96] M.C. Ferris. Finite termination of the proximal point algorithm. *Math. Program.*, 50(3, (Ser. A)):359–366, 1991.

[97] S.D. Flm. On finite convergence and constraint identification of subgradient projection methods. *Math. Program.*, 57:427–437, 1992.

[98] R. Fletcher. A model algorithm for composite nondifferentiable optimization problems. *Math. Programming Stud.*, (17):67–76, 1982. Nondifferential and variational techniques in optimization (Lexington, Ky., 1980).

[99] R Fletcher. A model algorithm for composite nondifferentiable optimization problems. In *Nondifferential and Variational Techniques in Optimization*, pages 67–76. Springer, 1982.

[100] Antonio Frangioni. *Standard Bundle Methods: Untrusted Models and Duality*, pages 61–116. Springer International Publishing, Cham, 2020.

[101] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution

of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40, 1976.

[102] David M. Gay. *Netlib: infeasible linear programming test problems*, 2013 (accessed May 31, 2021). `http://netlib.org/lp/infeas/index.html`.

[103] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1233–1242. JMLR. org, 2017.

[104] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 797–842. JMLR.org, 2015.

[105] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

[106] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2973–2981. Curran Associates, Inc., 2016.

[107] Andrew Gilpin, Javier Peña, and Tuomas Sandholm. First-order algorithm with $\tilde{O}(\ln(1/\epsilon))$ convergence for ϵ-equilibrium in two-person zero-sum games. *Mathematical Programming*, 133(1):279–298, Jun 2012.

[108] Roland Glowinski and A Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76, 1975.

[109] J.L. Goffin. On convergence rates of subgradient optimization methods. *Math. Programming*, 13(3):329–347, 1977.

[110] T. Goldstein and C. Studer. Phasemax: Convex phase retrieval via basis

pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, April 2018.

[111] Nadav Hallak and Marc Teboulle. Finding second-order stationary points in constrained minimization: A feasible direction approach. *Journal of Optimization Theory and Applications*, 186(2):480–503, 2020.

[112] Warren Hare and Claudia Sagastizábal. A Redistributed Proximal Bundle Method for Nonconvex Optimization. *SIAM J. Optim.*, 20(5):2442–2473, 2010.

[113] Warren Hare, Claudia Sagastizábal, and Mikhail Solodov. A Proximal Bundle Method for Nonsmooth Nonconvex Functions with Inexact Information. *Computational Optimization and Applications*, 63(1):1–28, Jan 2016.

[114] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

[115] Nicholas JA Harvey, Christopher Liaw, and Sikander Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*, 2019.

[116] Bingsheng He, Yanfei You, and Xiaoming Yuan. On the convergence of primal-dual hybrid gradient algorithm. *SIAM Journal on Imaging Sciences*, 7(4):2526–2537, 2014.

[117] C. Helmberg and F. Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696, 2000.

[118] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Acceleration of the Cutting-Plane Algorithm: Primal Forms of Bundle Methods*, pages 275–330. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993.

[119] Minhui Huang. Escaping saddle points for nonsmooth weakly convex functions via perturbed proximal algorithms. *arXiv preprint arXiv:2102.02837*, 2021.

[120] Wen Huang and Paul Hand. Blind deconvolution by steepest descent algorithm on a quotient manifold. *arXiv:1710.03309v2*, 2018.

[121] GJO Jameson. Inequalities for gamma function ratios. *The American Mathematical Monthly*, 120(10):936–940, 2013.

[122] Chi Jin, Lydia T Liu, Rong Ge, and Michael I Jordan. On the local minima of the empirical risk. *Advances in neural information processing systems*, 2018.

[123] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *J. ACM*, 68(2), February 2021.

[124] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1042–1085. PMLR, 2018.

[125] Sham M Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *NIPS*, pages 801–808, 2008.

[126] Michael Kech and Felix Krahmer. Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems. *SIAM Journal on Applied Algebra and Geometry*, 1(1):20–37, 2017.

[127] Krzysztof C. Kiwiel. An Aggregate Subgradient Method for Nonsmooth Convex Minimization. *Math. Program.*, 27(3):320–341, October 1983.

[128] Krzysztof C. Kiwiel. A Linearization Algorithm for Nonsmooth Minimization. *Mathematics of Operations Research*, 10(2):185–194, 1985.

[129] Krzysztof C. Kiwiel. *Methods of Descent for Nondifferentiable Optimization.* Springer, Berlin, 1985.

[130] Krzysztof C. Kiwiel. Proximal Level Bundle Methods for Convex Nondifferentiable Optimization, Saddle-point Problems and Variational Inequalities. *Math. Program.*, 69(1-3):89–109, July 1995.

[131] Krzysztof C. Kiwiel. Efficiency of Proximal Bundle Methods. *Journal of Optimization Theory and Applications*, 104(3):589–603, Mar 2000.

[132] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

[133] Ken Kreutz-Delgado. The complex gradient operator and the cr-calculus. *arXiv:0906.4835*, 2009.

[134] Han-Wen Kuo, Yenson Lau, Yuqian Zhang, and John Wright. Geometry and symmetry in short-and-sparse deconvolution. *arXiv:1901.00256*, 2019.

[135] Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global convergence of the em algorithm for mixtures of two component linear regression. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2055–2110, Phoenix, USA, 25–28 Jun 2019. PMLR.

[136] Jourdain Lamperski, Robert M. Freund, and Michael J. Todd. An oblivious ellipsoid algorithm for solving a system of (in)feasible linear inequalities, 2020.

[137] Guanghui Lan. Bundle-Level Type Methods Uniformly Optimal For Smooth And Nonsmooth Convex Optimization. *Mathematical Programming*, 149(1):1–45, Feb 2015.

[138] Guanghui Lan, Zhaosong Lu, and Renato D. C. Monteiro. Primal-dual first-order methods with $O(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, Jan 2011.

[139] Yenson Lau, Qing Qu, Han-Wen Kuo, Pengcheng Zhou, Yuqian Zhang, and John Wright. Short and sparse deconvolution - A geometric approach. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[140] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[141] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1–2):311–337, July 2019.

[142] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR, 2016.

[143] Claude Lemarechal. *An Extension of Davidon Methods to Nondifferentiable Problems*, pages 95–109. Springer Berlin Heidelberg, Berlin, Heidelberg, 1975.

[144] Claude Lemaréchal. *Lagrangian Relaxation*, pages 112–156. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[145] Claude Lemaréchal, Arkadii Nemirovskii, and Yurii Nesterov. New Variants of Bundle Methods. *Math. Program.*, 69(1-3):111–147, July 1995.

[146] Claude Lemaréchal and Claudia Sagastizábal. Practical aspects of the moreau–yosida regularization: Theoretical preliminaries. *SIAM Journal on Optimization*, 7(2):367–385, 1997.

[147] A. S. Lewis. Derivatives of spectral functions. *Math. Oper. Res.*, 21(3):576–588, 1996.

[148] Adrian S Lewis and Jingwei Liang. Partial smoothness and constant rank. *arXiv preprint arXiv:1807.03134*, 2018.

[149] Adrian S Lewis and Hristo S Sendov. Nonsmooth analysis of singular values. part i: Theory. *Set-Valued Analysis*, 13(3):213–241, 2005.

[150] A.S. Lewis. Nonsmooth analysis of eigenvalues. *Math. Program.*, 84(1, Ser. A):1–24, 1999.

[151] A.S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM J. Optim.*, 13(3):702–725 (electronic) (2003), 2002.

[152] A.S. Lewis and S.J. Wright. A proximal method for composite minimization. *Math. Program.*, 158(1-2, Ser. A):501–546, 2016.

[153] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv:1606.04933*, 2016.

[154] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020.

[155] Y. Li, C. Ma, Y. Chen, and Y. Chi. Nonconvex matrix factorization from rank-one measurements. *arXiv:1802.06286*, 2018.

[156] Yanjun Li, Kiryung Lee, and Yoram Bresler. Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Transactions on Information Theory*, 62(7):4266–4275, 2016.

[157] Jiaming Liang and Renato D. C. Monteiro. A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes, 2021.

[158] Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local convergence properties of Douglas–Rachford and alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 172(3):874–913, 2017.

[159] Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local linear convergence analysis of primal–dual splitting methods. *Optimization*, 67(6):821–853, 2018.

[160] S. Ling and T. Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 31, 2015.

[161] Shuyang Ling, Ruitu Xu, and Afonso S Bandeira. On the landscape of synchronization networks: A perspective from nonconvex optimization. *SIAM Journal on Optimization*, 29(3):1879–1907, 2019.

[162] Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[163] Yanli Liu, Ernest K Ryu, and Wotao Yin. A new use of Douglas–Rachford splitting for identifying infeasible, unbounded, and pathological conic programs. *Mathematical Programming*, 177(1-2):225–253, 2019.

[164] Songtao Lu, Meisam Razaviyayn, Bo Yang, Kejun Huang, and Mingyi Hong. Finding second-order stationary points efficiently in smooth nonconvex linearly constrained optimization problems. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[165] Jian Lv, Li-Ping Pang, and Fan-Yun Meng. A proximal bundle method

for constrained nonsmooth nonconvex optimization with inexact information. *J. Glob. Optim.*, 70(3):517–549, 2018.

[166] Jian Lv, Li-Ping Pang, Na Xu, and Ze-Hao Xiao. An infeasible bundle method for nonconvex constrained optimization with application to semi-infinite programming problems. *Numer. Algorithms*, 80(2):397–427, 2019.

[167] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3345–3354, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[168] Yura Malitsky. The primal-dual hybrid gradient method reduces to a primal method for linearly constrained optimization problems, 2019.

[169] István Maros. *Computational Techniques of the Simplex Method*. Kluwer Academic Publishers, 2003.

[170] Richard Kipp Martin. *Large Scale Linear and Integer Optimization: A Unified Approach*. Springer US, Boston, MA, 1999.

[171] S. Mendelson. A remark on the diameter of random sections of convex bodies. In *Geometric aspects of functional analysis*, volume 2116 of *Lecture Notes in Math.*, pages 395–404. Springer, Cham, 2014.

[172] S. Mendelson. Learning without concentration. *J. ACM*, 62(3):Art. 21, 25, 2015.

[173] Jianwei Miao, Tetsuya Ishikawa, Qun Shen, and Thomas Earnest. Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annu. Rev. Phys. Chem.*, 59:387–410, 2008.

[174] R. Mifflin and C. Sagastizábal. Primal-dual gradient structured functions: second-order results; links to epi-derivatives and partly smooth functions. *SIAM J. Optim.*, 13(4):1174–1194 (electronic), 2003.

[175] R. Mifflin and C. Sagastizábal. $\mathcal{VU}$-smoothness and proximal point results for some nonconvex functions. *Optim. Methods Softw.*, 19(5):463–478, 2004.

[176] R. Mifflin and C. Sagastizábal. A $\mathcal{VU}$-algorithm for convex minimization. *Math. Program.*, 104(2-3, Ser. B):583–608, 2005.

[177] Robert Mifflin. *A Modification and an Extension of Lemarechal's Algorithm for Nonsmooth Minimization*, pages 77–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 1982.

[178] Aryan Mokhtari, Asuman Ozdaglar, and Ali Jadbabaie. Escaping saddle points in constrained optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3633–3643, 2018.

[179] Cesare Molinari, Jingwei Liang, and Jalal Fadili. Convergence rates of forward–douglas–rachford splitting method. *Journal of Optimization Theory and Applications*, 182(2):606–639, 2019.

[180] Najmeh Hoseini Monjezi and S. Nobakhtian. A new infeasible proximal bundle algorithm for nonsmooth nonconvex constrained optimization. *Comput. Optim. Appl.*, 74(2):443–480, 2019.

[181] Najmeh Hoseini Monjezi and S. Nobakhtian. A filter proximal bundle method for nonsmooth nonconvex constrained optimization. *J. Glob. Optim.*, 79(1):1–37, 2021.

[182] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Grundlehren der mathematischen Wissenschaften, Vol 330, Springer, Berlin, 2006.

[183] Walaa M Moursi. *The Douglas–Rachford operator in the possibly inconsistent case: static properties and dynamic behaviour*. PhD thesis, University of British Columbia, 2016.

[184] S.N. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 2012.

[185] Arkadi Nemirovski. Prox-method with rate of convergence O(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[186] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Ap-*

plied Optimization. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[187] Yu. Nesterov. Subgradient methods for huge-scale optimization problems. *Mathematical Programming*, 146(1):275–297, Aug 2014.

[188] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

[189] Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[190] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015.

[191] Maher Nouiehed, Jason D Lee, and Meisam Razaviyayn. Convergence to second-order stationarity for constrained non-convex optimization. *arXiv preprint arXiv:1810.02024*, 2018.

[192] E.A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

[193] Peter Ochs, Jalal Fadili, and Thomas Brox. Non-smooth non-convex bregman minimization: Unification and new algorithms. *J. Optim. Theory Appl.*, 181(1):244–278, 2019.

[194] Daniel O'Connor and Lieven Vandenberghe. On the equivalence of the primal-dual hybrid gradient method and Douglas–Rachford splitting. *Mathematical Programming*, 179(1):85–108, Jan 2020.

[195] Brendan O'Donoghue. Operator splitting for a homogeneous embedding of the monotone linear complementarity problem. *arXiv preprint arXiv:2004.02177*, 2020.

[196] Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

[197] M. O'Neill and Stephen J. Wright. A line-search descent algorithm for

strict saddle functions with complexity guarantees. *arXiv: Optimization and Control*, 2020.

[198] François Oustry. A second-order bundle method to minimize the maximum eigenvalue function. *Math. Program.*, 89(1):1–33, 2000.

[199] Michael O'Neill and Stephen J Wright. A log-barrier newton-cg method for bound constrained optimization with complexity guarantees. *IMA Journal of Numerical Analysis*, 41(1):84–121, Apr 2020.

[200] N. Parikh and S. Boyd. Block splitting for distributed optimization. *Mathematical Programming Computation*, 6(1):77–102, 2014.

[201] A Pazy. Asymptotic behavior of contractions in hilbert space. *Israel Journal of Mathematics*, 9(2):235–240, 1971.

[202] Robin Pemantle et al. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, 18(2):698–712, 1990.

[203] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *2011 International Conference on Computer Vision*, pages 1762–1769, 2011.

[204] R.A. Poliquin and R.T. Rockafellar. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.*, 348:1805–1838, 1996.

[205] Michael JD Powell. Algorithms for nonlinear constraints that use lagrangian functions. *Mathematical programming*, 14(1):224–248, 1978.

[206] Arvind U Raghunathan and Stefano Di Cairano. Infeasibility detection in alternating direction method of multipliers for convex quadratic programs. In *53rd IEEE Conference on Decision and Control*, pages 5819–5824. IEEE, 2014.

[207] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, 2012.

[208] Inder K. Rana. *An introduction to measure and integration*, volume 45 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2002.

[209] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[210] James Renegar. *A mathematical view of interior-point methods in convex optimization*. SIAM, 2001.

[211] James Renegar. Accelerated first-order methods for hyperbolic programming. *Mathematical Programming*, 173(1):1–35, Jan 2019.

[212] James Renegar and Benjamin Grimmer. A Simple Nearly-Optimal Restart Scheme For Speeding-Up First Order Methods. *Foundations of Computational Mathematics*, 2021.

[213] R.T. Rockafellar. Favorable classes of Lipschitz-continuous functions in subgradient optimization. In *Progress in nondifferentiable optimization*, volume 8 of *IIASA Collaborative Proc. Ser. CP-82*, pages 125–143. Int. Inst. Appl. Sys. Anal., Laxenburg, 1982.

[214] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[215] S. Rolewicz. On paraconvex multifunctions. In *Third Symposium on Operations Research (Univ. Mannheim, Mannheim, 1978), Section I*, volume 31 of *Operations Res. Verfahren*, pages 539–546. Hain, Königstein/Ts., 1979.

[216] Clément W Royer and Stephen J Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477, 2018.

[217] Clément W. Royer, Michael O'Neill, and Stephen J. Wright. A newton-cg algorithm with complexity guarantees for smooth unconstrained optimization. *Mathematical Programming*, 180(1-2):451–488, Jan 2019.

[218] M. Rudelson and R. Vershynin. Small ball probabilities for linear images of high-dimensional distributions. *International Mathematics Research Notices*, 2015(19):9594–9617, 2014.

[219] Andrzej Ruszczynski. *Nonlinear Optimization*. Princeton University Press, Princeton, NJ, USA, 2006.

[220] Claudia Sagastizábal. Divide to Conquer: Decomposition Methods for

Energy Optimization. *Mathematical Programming*, 134(1):187–222, Aug 2012.

[221] Claudia Sagastizábal and Mikhail Solodov. An Infeasible Bundle Method for Nonsmooth Convex Constrained Optimization without a Penalty Function or a Filter. *SIAM Journal on Optimization*, 16(1):146–169, 2005.

[222] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[223] Y. Shechtman, Y.C. Eldar, O. Cohen, H.N. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, May 2015.

[224] Shu-Chung Shi. Semi-continuités génériques de multi-applications. *C.R. Acad. Sci. Paris (Série A)*, 293:27–29, 1981.

[225] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020.

[226] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *CoRR*, abs/1510.06096, 2015.

[227] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[228] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inform. Theory*, 62(11):6535–6579, 2016.

[229] Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on riemannian manifolds. In Wallach et al. [237], pages 7274–7284.

[230] M.J. Todd. Detecting infeasibility. In Felipe Cucker, Ron DeVore, Peter Olver, and Endre Süli, editors, *Foundations of Computational Mathematics, Minneapolis 2002*, London Mathematical Society Lecture Note Series, page 157–192. Cambridge University Press, 2004.

[231] Joel A Tropp and Stephen J Wright. Computational methods for sparse

solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

[232] Yuhsiang M. Tsai, Terry Cojean, and Hartwig Anzt. Sparse linear algebra on AMD and NVIDIA GPUs – the race is on. In Ponnuswamy Sadayappan, Bradford L. Chamberlain, Guido Juckeland, and Hatem Ltaief, editors, *High Performance Computing*, pages 309–327, Cham, 2020. Springer International Publishing.

[233] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 964–973. JMLR.org, 2016.

[234] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

[235] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[236] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[237] Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

[238] Adriaan Walther. The question of phase retrieval in optics. *Optica Acta: International Journal of Optics*, 10(1):41–49, 1963.

[239] G. Wang, G.B. Giannakis, and Y.C. Eldar. Solving systems of random quadratic equations via a truncated amplitude flow. *arXiv:1605.08285*, 2016.

[240] Kaizheng Wang, Yuling Yan, and Mateo Díaz. Efficient clustering for stretched mixtures: Landscape and optimality. *Advances in Neural Information Processing Systems*, 33, 2020.

[241] James D Watson and Francis HC Crick. Molecular structure of nucleic

acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[242] Philip Wolfe. *A Method of Conjugate Subgradients for Minimizing Nondifferentiable Functions*, pages 145–173. Springer Berlin Heidelberg, Berlin, Heidelberg, 1975.

[243] S.J. Wright. Identifiable surfaces in constrained optimization. *SIAM J. Control Optim.*, 31:1063–1079, July 1993.

[244] Yue Xie and Stephen J Wright. Complexity of projected newton methods for bound-constrained optimization. *arXiv preprint arXiv:2103.15989*, 2021.

[245] X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust pca via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

[246] Y Yu, T. Wang, and R. J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 2015.

[247] Yuqian Zhang, Yenson Lau, Han-wen Kuo, Sky Cheung, Abhay Pasupathy, and John Wright. On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4902, 2017.

[248] Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

[249] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of low-rank matrix optimization. *IEEE Transactions on Information Theory*, 67(2):1308–1331, 2021.