Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence

Vasileios Charisopoulos* Yudong Chen† Damek Davis‡
Mateo Díaz§ Lijun Ding¶ Dmitriy Drusvyatskiy‖

Abstract

The task of recovering a low-rank matrix from its noisy linear measurements plays a central role in computational science. Smooth formulations of the problem often exhibit an undesirable phenomenon: the condition number, classically defined, scales poorly with the dimension of the ambient space. In contrast, we here show that in a variety of concrete circumstances, nonsmooth penalty formulations do not suffer from the same type of ill-conditioning. Consequently, standard algorithms for nonsmooth optimization, such as subgradient and prox-linear methods, converge at a rapid dimension-independent rate when initialized within constant relative error of the solution. Moreover, nonsmooth formulations are naturally robust against outliers. Our framework subsumes such important computational tasks as phase retrieval, blind deconvolution, quadratic sensing, matrix completion, and robust PCA. Numerical experiments on these problems illustrate the benefits of the proposed approach.

*School of ORIE, Cornell University, Ithaca, NY 14850, USA; people.orie.cornell.edu/vc333/
†School of ORIE, Cornell University, Ithaca, NY 14850, USA; people.orie.cornell.edu/yudong.chen/
‡School of ORIE, Cornell University, Ithaca, NY 14850, USA; people.orie.cornell.edu/dsd95/.
§CAM, Cornell University. Ithaca, NY 14850, USA; people.cam.cornell.edu/md825/
¶School of ORIE, Cornell University, Ithaca, NY 14850, USA; people.orie.cornell.edu/ld446/.
‖Department of Mathematics, U. Washington, Seattle, WA 98195; www.math.washington.edu/~ddrusv.
Research of Drusvyatskiy was supported by the NSF DMS 1651851 and CCF 1740551 awards.

Contents

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 [12, 42, 55], blind deconvolution [1, 38, 41, 57], matrix completion [9, 21, 56], covariance matrix estimation [18, 40], and robust principal component analysis [11, 15]. 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 [3, 32, 33], 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 work 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 paper 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. [17, 42, 46, 57]), 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

$$\min_{X \in \mathbf{R}^{d \times r}} \; f(X) := h\left(\mathcal{A}(XX^\top) - b\right) \qquad \text{subject to } X \in \mathcal{D}, \tag{1.1}$$

or when $M_\sharp$ is rectangular, one may instead use the formulation

$$\min_{X \in \mathbf{R}^{d_1 \times r},\; Y \in \mathbf{R}^{r \times d_2}} \; f(X, Y) := h\left(\mathcal{A}(XY) - b\right) \qquad \text{subject to } (X, Y) \in \mathcal{D}. \tag{1.2}$$

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 (1.1) and (1.2) 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. [42, 46, 57]). 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 [20] 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 (1.1) and (1.2) *are well-conditioned* when h is an appropriate *nonsmooth convex penalty*.

To explain what we mean by "good conditioning," let us treat (1.1) and (1.2) within the broader *convex composite* problem class:

$$\min_{x \in \mathcal{X}} \; f(x) := h(F(x)), \tag{1.3}$$

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 (1.3) have been well-studied in nonlinear programming [6, 7, 31], their computational promise in data science has only begun recently to emerge. For example, the papers [22, 26, 28] discuss stochastic and inexact algorithms on composite problems, while the papers [24, 27], [16], and [39] investigate applications to phase retrieval, blind deconvolution, and matrix sensing, respectively.

A number of algorithms are available for problems of the form (1.3), and hence for (1.1) and (1.2). Two most notable ones are the projected subgradient[1] method [23, 34]

$$x_{t+1} = \mathrm{proj}_{\mathcal{X}}(x_t - \alpha_t v_t) \qquad \text{with} \qquad v_t \in \partial f(x_t),$$

and the prox-linear algorithm [6, 25, 37]

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \; h\Big(F(x_t) + \nabla F(x_t)(x - x_t)\Big) + \frac{\beta}{2}\|x - x_t\|_2^2.$$

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 $\mathcal{X}^* := \operatorname{argmin}_{\mathcal{X}} f$, 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)** $|f(y) - f_x(y)| \leq \frac{\rho}{2}\|y - x\|_2^2$ for all $x, y \in \mathcal{X}$,
- **(sharpness)** $f(x) - \inf f \geq \mu \cdot \mathrm{dist}(x, \mathcal{X}^*)$ for all $x \in \mathcal{X}$.

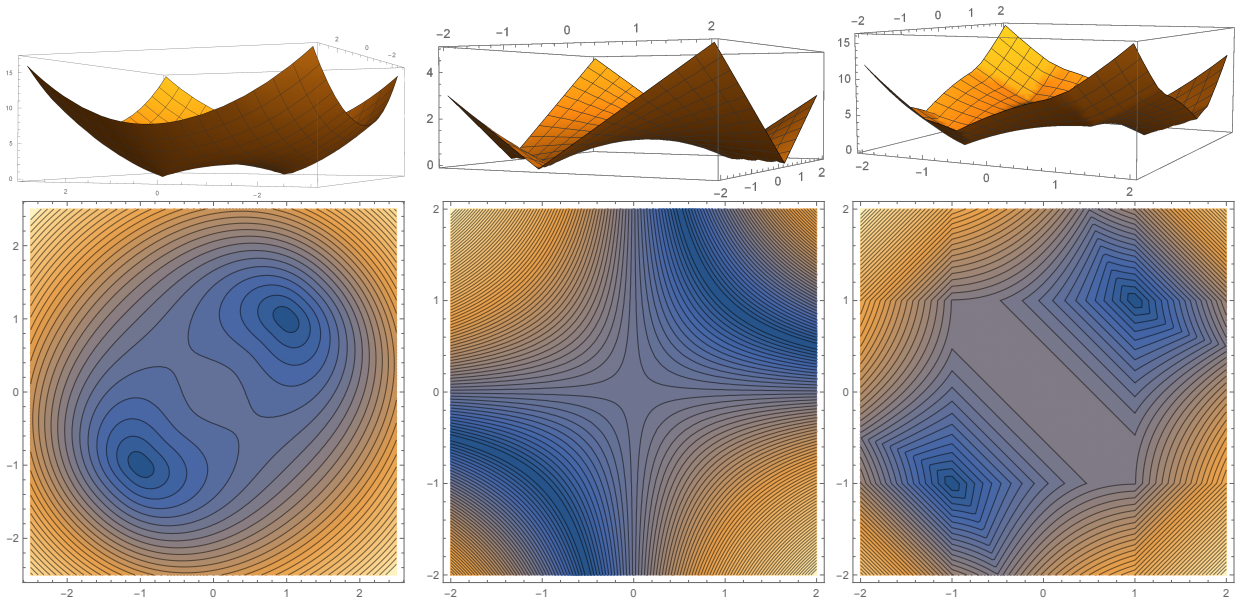
The approximation and sharpness properties have intuitive meanings. The former says that the nonconvex function $f(y)$ is well approximated by the convex model $f_x(y)$, with quality that degrades quadratically as y deviates from x. In particular, this property guarantees that the quadratically perturbed 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 amplitute 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.

Approximation and sharpness, taken together, guarantee rapid convergence of numerical methods when initialized within the tube:

$$\mathcal{T} = \Big\{x \in \mathcal{X} : \mathrm{dist}(x, \mathcal{X}^*) \leq \frac{\mu}{\rho}\Big\}.$$

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 refer the reader to the survey [20], and references therein, for details.

[1]Here, the subdifferential is formally obtained through the chain rule $\partial f(x) = \nabla F(x)^* \partial h(F(x))$, where $\partial h(\cdot)$ is the subdifferential in the sense of convex analysis.



(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)

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 $\mathcal{T}$.[2]

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 (1.1) and (1.2) are "well-conditioned", and subgradient and prox-linear methods converge rapidly when initialized within constant relative error of the optimal solution.

Approximation and sharpness via the Restricted Isometry Property

We begin verifying our thesis by showing that the composite problems, (1.1) and (1.2), are well-conditioned under the following Restricted Isometry Property (RIP): there exists a norm $|||\cdot|||$ and numerical constants $\kappa_1, \kappa_2 > 0$ so that

$$\kappa_1 \|W\|_F \leq |||\mathcal{A}(W)||| \leq \kappa_2 \|W\|_F, \tag{1.4}$$

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 $M_\sharp$:

$$\mu = 0.9\kappa_1\sqrt{\sigma_r(M_\sharp)}, \qquad \rho = \kappa_2, \qquad L = 0.9\kappa_1\sqrt{\sigma_r(M_\sharp)} + 2\kappa_2\sqrt{\sigma_1(M_\sharp)}.$$

[2]Both the parameters α_t and β must be properly chosen for these guarantees to take hold.

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. [57]). We note that similar results in the symmetric setting were independently obtained in the complimentary work [39], albeit with a looser estimate of L; the two treatments of the asymmetric setting are distinct, however.[3]

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 [13, 50] showed that if $\mathcal{A}(\cdot)$ arises from a Gaussian ensemble, then in 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 [27, 29], blind deconvolution [16], and quadratic sensing [18]; 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 [19].

The goal of the matrix completion problem [9] is to recover a low rank matrix $M_\sharp$ from its partially observed entries. We focus on the formulation

$$\operatorname*{argmin}_{X \in \mathcal{X}} \; f(X) = \|\Pi_\Omega(XX^\top) - \Pi_\Omega(M_\sharp)\|_2,$$

where Π_Ω is the projection onto the index set of observed entries Ω and

$$\mathcal{X} = \left\{ X \in \mathbb{R}^{d\times r} : \|X\|_{2,\infty} \leq \sqrt{\frac{\nu r \|M_\sharp\|_{\mathrm{op}}}{d}} \right\}$$

[3]The authors of [39] provide a bound on L that scales with the Frobenius norm $\sqrt{\|M_\sharp\|_F}$. We instead derive a sharper bound that scales as $\sqrt{\|M_\sharp\|_{\mathrm{op}}}$. 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 $\|M_\sharp\|_F/\sigma_r(M_\sharp)$.

is the set of incoherent matrices. 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 [19], 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:

$$X_{k+1} = \operatorname*{argmin}_{X \in \mathcal{X}} \; f_{X_k}(X) + \sqrt{1+\epsilon}\|X - X_k\|_F^2 + \sqrt{\epsilon}\|X - X_k\|_F.$$

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 $\operatorname{dist}(X_0, \mathcal{X}^*)$ are below some constant thresholds, both the subgradient and the modified prox-linear methods converge linearly with high probability:

$$\operatorname{dist}(X_k, \mathcal{X}^*) \lesssim \begin{cases} \left(1 - \frac{c}{\nu r}\right)^{k/2} & \text{subgradient} \\ 2^{-k} & \text{prox-linear} \end{cases}.$$

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 or problem parameters—an observation fully supported by our numerical experiments.

As the final example, we consider the problem of robust PCA [11, 15], 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:

$$\min_{(X,S) \in \mathcal{D}_1} \; F((X,S)) = \|XX^\top + S - W\|_F, \tag{1.5}$$

and

$$\min_{X \in \mathcal{D}_2} f(X) = \|XX^\top - W\|_1, \tag{1.6}$$

where $\mathcal{D}_1$ and $\mathcal{D}_2$ are appropriately defined convex regions. Under standard incoherence assumptions, we show that the formulation (1.5) is well-conditioned, and therefore subgradient and prox-linear methods are applicable. Still, formulation (1.5) 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 (1.6) as a more practical alternative.

The analysis of (1.6) is more challenging than that of (1.5). 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:

$$|f(Y) - f_X(Y)| \leq \|Y - X\|_{2,1}^2 \qquad \text{for all } X, Y$$

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.

Numerical experiments

We conclude with an experimental evaluation of our theoretical findings on quadratic and bilinear matrix sensing, matrix completion, and robust PCA problems. In the first set of experiments, we test the robustness of the proposed methods against varying combinations of rank/corruption level by reporting the empirical recovery rate across independent runs of synthetic problem instances. All the aforementioned model problems exhibit sharp phase transitions, yet our methods succeed for more than moderate levels of corruption (or unobserved entries in the case of matrix completion). For example, in the case of matrix sensing, we can corrupt almost half of the measurements $\mathcal{A}_i(M)$ and still retain perfect recovery rates. Interestingly, our experimental findings indicate that the prox-linear method can tolerate slightly higher levels of corruption compared to the subgradient method, making it the method of choice for small-to-moderate dimensions.

We then demonstrate that the convergence rate analysis is fully supported by empirical evidence. In particular, we test the subgradient and prox-linear methods for different rank/corruption configurations. In the case of quadratic/bilinear sensing and robust PCA, we observe that the subgradient method converges linearly and the prox-linear method converges quadratically, as expected. In particular, our numerical experiments appear to support our sharpness conjecture for the robust PCA problem. In the case of matrix completion, both algorithms converge linearly. The prox-linear method in particular, converges extremely quickly, reaching high accuracy solutions in under 25 iterations for reasonable values of p.

In the noiseless setting, we compare against gradient descent with constant step-size on smooth formulations of each problem (except for robust PCA). We notice that the Polyak subgradient method outperforms gradient descent in all cases. That being said, one can

heuristically equip gradient descent with the Polyak step-size as well. To the best of our knowledge, the gradient method with Polyak step-size has has not been investigated on smooth problem formulations we consider here. Experimentally, we see that the Polyak (sub)gradient methods on smooth and nonsmooth formulations perform comparably in the noiseless setting.

Outline of the paper

The outline of the paper is as follows. Section 2 records some basic notation we will use. Section 3 informally discusses the sharpness and approximation properties, and their impact on convergence of the subgradient and prox-linear methods. Section 4 analyzes the parameters μ, ρ, L under RIP. Section 5 rigorously discusses convergence guarantees of numerical methods under regularity conditions. Section 6 reviews examples of problems satisfying RIP and deduces convergence guarantees for subgradient and prox-linear algorithms. Sections 7 and 8 discuss the matrix completion and robust PCA problems, respectively. Section 9 discusses robust recovery up to a noise tolerance. The final Section 10 illustrates the developed theory and algorithms with numerical experiments on quadratic/bi-linear sensing, matrix completion, and robust PCA problems.

2 Preliminaries

In this section, we summarize the basic notation we will use throughout the paper. Henceforth, the symbol $\mathbf{E}$ will denote a Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\|_2 = \sqrt{\langle x, x \rangle}$. The closed unit ball in $\mathbf{E}$ will be denoted by $\mathbb{B}$, 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

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

respectively. 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 $\mathcal{A}^* \colon \mathbf{Y} \to \mathbf{E}$. 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 $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$. Similarly, we will equip the space of rectangular matrices $\mathbf{R}^{d_1 \times d_2}$ with the trace product $\langle X, Y \rangle = \operatorname{Tr}(X^\top Y)$ and the induced Frobenius norm $\|X\|_F = \sqrt{\operatorname{Tr}(X^\top X)}$. The operator norm of a matrix $X \in \mathbf{R}^{d_1 \times d_2}$ will be written as $\|X\|_{\text{op}}$. The symbol $\sigma(X)$ will denote the vector of singular values of a matrix X in nonincreasing order. We also define the row-wise matrix norms $\|X\|_{b,a} = \|(\|X_{1\cdot}\|_b, \|X_{2\cdot}\|_b \ldots, \|X_{d_1\cdot}\|_b)\|_a$. 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.

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 [4, 45, 52]. Consider a function $f\colon \mathbf{E} \to \mathbf{R} \cup \{+\infty\}$ 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, while for a differentiable function it consists only of the gradient: $\partial f(x) = \{\nabla f(x)\}$. 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 [52, Theorem 10.6]:

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

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.

We say a that a random vector X in $\mathbf{R}^d$ is *η-sub-gaussian* whenever $\mathbb{E} \exp\left(\frac{\langle u, X\rangle^2}{\eta^2}\right) \leq 2$ for all unit vectors $u \in \mathbf{R}^d$. The *sub-gaussian norm* of a real-valued random variable X is defined to be $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E} \exp\left(\frac{X^2}{t^2}\right) \leq 2\}$, while the *sub-exponential norm* is defined by $\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E} \exp\left(\frac{|X|}{t}\right) \leq 2\}$.

3 Regularity conditions and algorithms (informal)

As outlined in Section 1, we consider the low-rank matrix recovery problem within the framework of compositional optimization:

$$\min_{x \in \mathcal{X}} f(x) := h(F(x)), \tag{3.1}$$

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 (3.1) 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 A. Suppose that the following properties hold for the composite optimization problem (3.1) 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 $\mathcal{X}^* := \operatorname*{argmin}_{x \in \mathcal{X}} f(x)$ 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 3.1 (Subgradient bound and Lipschitz continuity [52, Theorem 9.13])**.** *Suppose a function $f \colon \mathbf{E} \to \mathbf{R}$ 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 for details. The following general principle has recently emerged [16,23,24,27]. Under Assumption A, 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.

Algorithm 1: 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 $x_{k+1} = \operatorname{proj}_{\mathcal{X}} \left(x_k - \dfrac{f(x_k) - \min_{\mathcal{X}} f}{\|\zeta_k\|_2^2} \zeta_k \right)$.

Algorithm 2: Subgradient method with geometrically 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 $x_{k+1} = \operatorname{proj}_{\mathcal{X}} \left(x_k - \alpha_k \frac{\zeta_k}{\|\zeta_k\|_2} \right)$.

Algorithm 3: Prox-linear algorithm

Data: Initial point $x_0 \in \mathbf{R}^d$, proximal parameter $\beta > 0$.

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

Set $x_{k+1} \leftarrow \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ h\left(F(x_k) + \nabla F(x_k)(x - x_k)\right) + \dfrac{\beta}{2} \|x - x_k\|_2^2 \right\}$.

Algorithm 1 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 (1.1) and (1.2) 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 2, 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 subgradient methods is dominated by a single subgradient evaluation and a projection onto $\mathcal{X}$. Under appropriate parameter settings, Assumption A 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 2), 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 (1.1) and (1.2), 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 A (or a close variant) holds with favorable parameters $\rho, \mu, L > 0$ for common low-rank matrix recovery problems.

4 Regularity under RIP

In this section, we consider the low-rank recovery problems (1.1) and (1.2), 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 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:

$$\{X \in \mathbf{R}^{d_1 \times r} : XX^\top = M_\sharp\} = \{X_\sharp R : R \in O(r)\}, \tag{4.1}$$

$$\{(X, Y) \in \mathbf{R}^{d_1 \times r} \times \mathbf{R}^{r \times d_2} : XY = M_\sharp\} = \{(X_\sharp A, A^{-1} Y_\sharp) : A \in GL(r)\}. \tag{4.2}$$

[4]The guarantees we develop in the symmetric setting are similar to those in the recent preprint [39], 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 (1.2), while [39] 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 $M_\sharp$.

Throughout, we will let $\mathcal{D}^*(M_\sharp)$ refer to the set (4.1) in the symmetric case and to (4.2) 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 B (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 B 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 6 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 1 and refer the reader to Section 6 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 1: Common problems satisfying ℓ_1/ℓ_2 RIP in Assumption B. 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 6 for details.

In light of Assumption B, it it natural to take the norm $|||\cdot|||$ as the penalty $h(\cdot)$ in (1.1) and (1.2) . Then the symmetric problem (1.1) becomes

$$\min_{X \in \mathbf{R}^{d \times r}} f(X) := |||\mathcal{A}(XX^\top) - b|||, \tag{4.3}$$

while the asymmetric formulation (1.2) becomes

$$\min_{X \in \mathbf{R}^{d_1 \times r},\ Y \in \mathbf{R}^{r \times d_2}} f(X, Y) := |||\mathcal{A}(XY) - b|||. \tag{4.4}$$

Our immediate goal is to show that under Assumption B, the problems (4.3) and (4.4) are well-conditioned in the sense of Assumption A. We note that the asymmetric setting is more nuanced that its symmetric counterpart because Assumption A can only be guaranteed to hold on bounded sets. Nonetheless, as we discuss in Section 5, a localized version of

[5] In the latter case, RIP also goes by the name of Restricted Uniform Boundedness (RUB) [8].

Assumption A 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 [39] 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.

4.1 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

$$\begin{aligned} f_X(Z) &= |||\mathcal{A}(XX^\top + X(Z-X)^\top + (Z-X)X^\top) - b|||, \\ f_{(X,Y)}(\widehat{X},\widehat{Y}) &= |||\mathcal{A}(XY + X(\widehat{Y}-Y) + (\widehat{X}-X)Y) - b|||. \end{aligned}$$

Proposition 4.1 (Approximation accuracy and Lipschitz continuity (symmetric))**.** *Suppose Assumption B holds. Then for all $X, Z \in \mathbf{R}^{d\times r}$ the following estimates hold:*

$$\begin{aligned} |f(Z) - f_X(Z)| &\le \kappa_2 \|Z-X\|_F^2, \\ |f(X) - f(Z)| &\le \kappa_2 \|X+Z\|_{op}\|X-Z\|_F. \end{aligned}$$

Proof. To see the first estimate, observe that

$$\begin{aligned} |f(Z) - f_X(Z)| &= |||\mathcal{A}(ZZ^\top) - b||| - |||\mathcal{A}(XX^\top + X(Z-X)^\top + (Z-X)X^\top) - b||| \\ &\le |||\mathcal{A}(ZZ^\top - XX^\top - X(Z-X)^\top - (Z-X)X^\top||| \qquad (4.5) \\ &= |||\mathcal{A}\big((Z-X)(Z-X)^\top\big)||| \\ &\le \kappa_2 \left\|(Z-X)(Z-X)^\top\right\|_F \qquad (4.6) \\ &\le \kappa_2 \|Z-X\|_F^2, \end{aligned}$$

where (4.5) follows from the reverse triangle inequality and (4.6) uses Assumption B. Next, for any $X, Z \in \mathcal{X}$ we successively compute:

$$\begin{aligned} |f(X) - f(Z)| &= \left| |||\mathcal{A}(XX^\top) - b||| - |||\mathcal{A}(ZZ^\top) - b||| \right| \\ &\le \left|||\mathcal{A}(XX^\top - ZZ^\top)||| \right| \qquad (4.7) \\ &\le \kappa_2 \|XX^\top - ZZ^\top\|_F \qquad (4.8) \\ &= \frac{\kappa_2}{2}\|(X+Z)(X-Z)^\top + (X-Z)(X+Z)^\top\|_F \\ &\le \kappa_2 \|(X+Z)(X-Z)\|_F \\ &\le \kappa_2 \|X+Z\|_{op}\|X-Z\|_F, \end{aligned}$$

where (4.7) follows from the reverse triangle inequality and (4.8) uses Assumption B. 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 4.2 (Approximation accuracy and Lipschitz continuity (asymmetric))**.** *Suppose Assumption B 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 following 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

$$\begin{aligned}
|f(\widehat{X},\widehat{Y}) - f_{(X,Y)}(\widehat{X},\widehat{Y})| &= \Big| |||\mathcal{A}(\widehat{X}\widehat{Y}) - b||| - |||\mathcal{A}(XY + X(\widehat{Y}-Y) + (\widehat{X}-X)Y) - b||| \Big| \\
&\le |||\mathcal{A}(\widehat{X}\widehat{Y} - XY - X(\widehat{Y}-Y) - (\widehat{X}-X)Y)||| \\
&= |||\mathcal{A}\big((X-\widehat{X})(Y-\widehat{Y})\big)||| \\
&\le \kappa_2 \Big\|(X-\widehat{X})(Y-\widehat{Y})\Big\|_F \\
&\le \frac{\kappa_2}{2}\left(\|X-\widehat{X}\|_F^2 + \|Y-\widehat{Y}\|_F^2\right),
\end{aligned}$$

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 $a, b \in \mathbf{R}$.

□

4.2 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 4.2.1, and then move on to the general case when the measurements are corrupted by outliers in Section 4.2.2.

4.2.1 Sharpness in the noiseless regime

We begin with with the symmetric setting in the noiseless case $b = \mathcal{A}(M_\sharp)$. By Assumption B, we have the estimate

$$f(X) = |||\mathcal{A}(XX^\top) - b||| = |||\mathcal{A}(XX^\top - X_\sharp X_\sharp^\top)||| \ge \kappa_1 \|XX^\top - X_\sharp X_\sharp^\top\|_F. \tag{4.9}$$

It follows that the set of minimizers $\operatorname{argmin}_{X \in \mathbf{R}^{d\times r}} f(X)$ 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 [57, Lemma 5.4].

Proposition 4.3 ([57, 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 B holds in the noiseless setting $b = \mathcal{A}(M_\sharp)$, then the bound holds:

$$f(X) \geq \kappa_1\sqrt{2(\sqrt{2}-1)\sigma_r(M_\sharp)} \cdot \operatorname{dist}(X, \mathcal{D}^*(M_\sharp)) \qquad \text{for all } X \in \mathbf{R}^{d\times r}.$$

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 $(X,Y) \mapsto \|XY - X_\sharp Y_\sharp\|_F$. Such a sharpness guarantee in the rank one case was recently shown in [16, Proposition 4.2].

Proposition 4.4 ([16, Proposition 4.2]). *Fix a rank 1 matrix $M_\sharp \in \mathbf{R}^{d_1\times d_2}$ and a constant $\nu \geq 1$. Then for any $x \in \mathbf{R}^{d_1}$ and $w \in \mathbf{R}^{d_2}$ satisfying*

$$\|w\|_2, \|x\|_2 \leq \nu\sqrt{\sigma_1(M_\sharp)},$$

the following estimate holds:

$$\|xw^\top - M_\sharp\|_F \geq \frac{\sqrt{\sigma_1(M_\sharp)}}{2\sqrt{2}(\nu+1)} \cdot \operatorname{dist}\big((x,w), \mathcal{D}^*(M_\sharp)\big).$$

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 corollary is a higher rank extension of Proposition 4.4.

Theorem 4.5 (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*

$$\begin{aligned} \max\{\|X - X_\sharp\|_F, \|Y - Y_\sharp\|_F\} &\leq \nu\sqrt{\sigma_r(M_\sharp)} \\ \operatorname{dist}((X,Y), \mathcal{D}^*(M_\sharp)) &\leq \frac{\sqrt{\sigma_r(M_\sharp)}}{1+2(1+\sqrt{2})\nu} \end{aligned}, \tag{4.10}$$

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 (4.10). Let $A \in GL(r)$ be an invertible matrix satisfying

$$A \in \operatorname*{argmin}_{A \in GL(r)} \left\{ \|X - X_\sharp A\|_F^2 + \|Y - A^{-1}Y_\sharp\|_F^2 \right\}. \tag{4.11}$$

As a first step, we successively compute

$$\begin{aligned}
&\|XY - X_\sharp Y_\sharp\|_F \\
&= \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp) + (X - X_\sharp A)(Y - A^{-1}Y_\sharp)\|_F \\
&\geq \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F - \|(X - X_\sharp A)(Y - A^{-1}Y_\sharp)\|_F \\
&\geq \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F - \|X - X_\sharp A\|_F \cdot \|Y - A^{-1}Y_\sharp\|_F \\
&\geq \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F - \frac{1}{2}(\|X - X_\sharp A\|_F^2 + \|Y - A^{-1}Y_\sharp\|_F^2) \\
&= \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F - \frac{1}{2}\mathrm{dist}^2((X,Y), \mathcal{D}^*(M_\sharp)) \\
&\geq \|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F - \frac{\delta\sqrt{\sigma_r(M_\sharp)}}{2} \cdot \mathrm{dist}((X,Y), \mathcal{D}^*(M_\sharp)).
\end{aligned} \tag{4.12}$$

We next aim to lower bound the first term on the right. To this end, observe

$$\begin{aligned}
&\|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F^2 \\
&\qquad = \|(X - X_\sharp A)(A^{-1}Y_\sharp)\|_F^2 + \|X_\sharp A(Y - A^{-1}Y_\sharp)\|_F^2 \\
&\qquad\quad + 2\mathrm{Tr}((X - X_\sharp A)(A^{-1}Y_\sharp)(Y - A^{-1}Y_\sharp)^\top (X_\sharp A)^\top).
\end{aligned} \tag{4.13}$$

We claim that the cross-term is non-negative. To see this, observe that first order optimality conditions in (4.11) 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 (4.13) we conclude that

$$\begin{aligned}
&\|(X - X_\sharp A)(A^{-1}Y_\sharp) + X_\sharp A(Y - A^{-1}Y_\sharp)\|_F \\
&\geq \sqrt{\|(X - X_\sharp A)(A^{-1}Y_\sharp)\|_F^2 + \|X_\sharp A(Y - A^{-1}Y_\sharp)\|_F^2} \\
&\geq \sqrt{\sigma_r(M_\sharp)} \cdot \min\{\sigma_r(A^{-1}), \sigma_r(A)\} \cdot \mathrm{dist}((X,Y), \mathcal{D}^*(M_\sharp)).
\end{aligned} \tag{4.14}$$

Combining (4.12) and (4.14), we obtain

$$\|XY - M_\sharp\|_F \geq \sqrt{\sigma_r(M_\sharp)} \cdot \left(\min\{\sigma_r(A^{-1}), \sigma_r(A)\} - \frac{\delta}{2} \right) \cdot \mathrm{dist}((X,Y), \mathcal{D}^*(M_\sharp)) \tag{4.15}$$

Finally, we estimate $\min\{\sigma_r(A^{-1}), \sigma_r(A)\}$. To this end, first note that

$$\begin{aligned}\|X_\sharp - X_\sharp A\|_F + \|Y_\sharp - A^{-1}Y_\sharp\|_F &\leq \|X_\sharp - X\|_F + \|Y_\sharp - Y\|_F + \sqrt{2}\cdot \mathrm{dist}((X,Y), \mathcal{D}^*(M_\sharp)) \\ &\leq 2\nu\sqrt{\sigma_r(M_\sharp)}\cdot(1+\sqrt{2}).\end{aligned} \tag{4.16}$$

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 (4.16), 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 (4.15) completes the proof. □

4.2.2 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 6 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 [27] on robust phase retrieval, we make the following assumption.

Assumption C ($\mathcal{I}$-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

$$\kappa_3\|W\|_F \leq \frac{1}{m}\|\mathcal{A}_{\mathcal{I}^c}(W)\|_1 - \frac{1}{m}\|\mathcal{A}_\mathcal{I}(W)\|_1. \tag{4.17}$$

The assumption is simple to interpret. To elucidate the bound (4.17), let us suppose that the restricted maps $\mathcal{A}_\mathcal{I}$ and $\mathcal{A}_{\mathcal{I}^c}$ satisfy Assumption B (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

$$\frac{1}{m}\|\mathcal{A}_{\mathcal{I}^c}(W)\|_1 - \frac{1}{m}\|\mathcal{A}_\mathcal{I}(W)\|_1 \geq \left((1-p_{\mathrm{fail}})\kappa_1 - p_{\mathrm{fail}}\hat{\kappa_2}\right)\|W\|_F,$$

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 $p_{\text{fail}} < \frac{\kappa_1}{\kappa_1+\kappa_2}$.

Combining Assumption C with Proposition 4.3 quickly yields sharpness of the objective even in the noisy setting.

Proposition 4.6 (Sharpness with outliers (symmetric))**.** *Suppose that Assumption C holds. Then*

$$f(X) - f(X_\sharp) \geq \kappa_3 \left(\sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp)\right) \operatorname{dist}\big(X, \mathcal{D}^*(M_\sharp)\big) \qquad \text{for all } X \in \mathbf{R}^{d\times r}.$$

Proof. Defining $\Delta := \mathcal{A}(X_\sharp X_\sharp^\top) - b$, we have the following bound:

$$\begin{aligned} m \cdot (f(X) - f(X_\sharp)) &= \|\mathcal{A}\left(XX^\top - X_\sharp X_\sharp^\top\right) + \Delta\|_1 - \|\Delta\|_1 \\ &= \|\mathcal{A}_{\mathcal{I}^c}(XX^\top - X_\sharp X_\sharp^\top)\|_1 + \sum_{i\in\mathcal{I}} \Big(\big|\left(\mathcal{A}(XX^\top - X_\sharp X_\sharp^\top)\right)_i + \Delta_i\big| - |\Delta_i|\Big) \\ &\geq \|\mathcal{A}_{\mathcal{I}^c}(XX^\top - X_\sharp X_\sharp^\top)\|_1 - \|\mathcal{A}_{\mathcal{I}}(XX^\top - X_\sharp X_\sharp^\top)\|_1 \\ &\geq \kappa_3 m \|XX^\top - X_\sharp X_\sharp^\top\|_F \geq \kappa_3 m \left(\sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp)\right) \operatorname{dist}\big(X, \mathcal{D}^*(M_\sharp)\big), \end{aligned}$$

where the first inequality follows by the reverse triangle inequality, the second inequality follows by Assumption (C2), and the final inequality follows from Proposition 4.3. The proof is complete. □

The argument in the asymmetric setting is completely analogous.

Proposition 4.7 (Sharpness with outliers (asymmetric))**.** *Suppose that Assumption C 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 General convergence guarantees for subgradient & prox-linear methods

In this section, we formally develop convergence guarantees for Algorithms 1, 2, and 3 under Assumption A, 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 (3.1). We say that a function $f\colon \mathbf{E} \to \mathbf{R}\cup\{+\infty\}$ is *ρ-weakly convex*[6] 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 [26, 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:

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

Setting the stage, we introduce the following assumption.

Assumption D. Consider the optimization problem,

$$\min_{x\in\mathcal{X}} \; f(x). \tag{5.1}$$

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 $\mathcal{X}^* := \operatorname*{argmin}_{x\in\mathcal{X}} f(x)$ 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 A implies Assumption D. 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 5.1 ([23, Lemma 3.1])**.** *Suppose that Assumption D holds. Then the problem* (5.1) *has no stationary points x satisfying*

$$0 < \operatorname{dist}(x; \mathcal{X}^*) < \frac{2\mu}{\rho}.$$

[6] 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 [2, 47, 49, 51, 53].

It is worthwhile to note that the estimate $\frac{2\mu}{\rho}$ of the radius in Lemma 5.1 is tight [16, Section 3]. Hence, let us define for any $\gamma > 0$ the tube

$$\mathcal{T}_\gamma := \left\{ z \in \mathcal{X} \colon \operatorname{dist}(z, \mathcal{X}^*) \leq \gamma \cdot \frac{\mu}{\rho} \right\}. \tag{5.2}$$

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 1 and 2) 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 [23, Lemma 3.2] shows $\tau \in [0, 1]$. The following theorem appears as [23, Theorem 4.1], while its application to phase retrieval was investigated in [24].

Theorem 5.2 (Polyak subgradient method)**.** *Suppose that Assumption D holds and fix a real number $\gamma \in (0, 1)$. Then Algorithm 1 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 [23, Theorem 6.1]. The convex version of the result dates back to Goffin [34].

Theorem 5.3 (Geometrically decaying subgradient method)**.** *Suppose that Assumption D 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 2. Then the iterates x_k generated by Algorithm 2, 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 A. Since Assumption A implies Assumption D, both subgradient Algorithms 1 and 2 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 [7, 25, 27].

Theorem 5.4 (Prox-linear algorithm)**.** *Suppose Assumption A holds. Choose any $\beta \geq \rho$ in Algorithm 3 and set $\gamma := \rho/\beta$. Then Algorithm 3 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 4. In particular, we have the following efficiency guarantees of the subgradient and prox-linear methods applied to this problem.

Corollary 5.5 (Convergence guarantees under RIP (symmetric))**.** *Suppose Assumptions B and C 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*

$$\frac{\operatorname{dist}(X_0, \mathcal{D}^*(M_\sharp))}{\sqrt{\sigma_r(M_\sharp)}} \le 0.2 \cdot \frac{\kappa_3}{\kappa_2}.$$

Define the condition number $\chi := \sigma_1(M_\sharp)/\sigma_r(M_\sharp)$*. Then the following are true.*

1. **(Polyak subgradient)** *Algorithm 1 initialized at* X_0 *produces iterates that converge linearly to* $\mathcal{D}^*(M_\sharp)$*, that is*

$$\frac{\operatorname{dist}^2(X_k, \mathcal{D}^*(M_\sharp))}{\sigma_r(M_\sharp)} \le \left(1 - \frac{0.2}{1 + \frac{4\kappa_2^2\chi}{\kappa_3^2}}\right)^k \cdot \frac{\kappa_3^2}{100\kappa_2^2} \qquad \forall k \ge 0.$$

2. **(geometric subgradient)** *Algorithm 2 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:*

$$\frac{\operatorname{dist}^2(X_k, \mathcal{D}^*(M_\sharp))}{\sigma_r(M_\sharp)} \le \left(1 - \frac{0.2}{1 + \frac{4\kappa_2^2\chi}{\kappa_3^2}}\right)^k \cdot \frac{\kappa_3^2}{100\kappa_2^2} \qquad \forall k \ge 0.$$

3. **(prox-linear)** *Algorithm 3 with* $\beta = \rho$ *and initialized at* X_0 *converges quadratically:*

$$\frac{\operatorname{dist}(X_k, \mathcal{D}^*(M_\sharp)))}{\sqrt{\sigma_r(M_\sharp)}} \le 2^{-2^k} \cdot \frac{0.45\kappa_3}{\kappa_2} \qquad \forall k \ge 0.$$

5.1 Guarantees under local regularity

As explained in Section 4, Assumptions A and D 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 [39], 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 5.6, 5.7, and 5.8) are deferred to Appendix A.

We begin with the following localized version of Assumption D.

Assumption E. Consider the optimization problem,

$$\min_{x\in\mathcal{X}} \; f(x). \tag{5.3}$$

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 E. 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 5.6 (Polyak subgradient method (local regularity))**.** *Suppose Assumption E 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 1 initialized at x_0 produces iterates x_k that always lie in $B_\epsilon(\bar{x})$ and satisfy

$$\operatorname{dist}^2(x_{k+1}, \mathcal{X}^*) \leq \left(1 - \tfrac{1}{2}\tau^2\right)\operatorname{dist}^2(x_k, \mathcal{X}^*), \qquad \text{for all } k \geq 0. \tag{5.4}$$

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 5.7 (Geometrically decaying subgradient method (local regularity))**.** *Suppose that Assumption E 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 2 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*

$$\operatorname{dist}^2(x_k; \mathcal{X}^*) \leq \frac{\gamma^2\mu^2}{\rho^2}\left(1 - (1-\gamma)\tau^2\right)^k \qquad \text{for all } k \geq 0. \tag{5.5}$$

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 A.

Assumption F. 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 F.

Theorem 5.8 (Prox-linear (local))**.** *Suppose Assumption F 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 3 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 [39]. 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 5.9 (Convergence guarantees under RIP (asymmetric))**.** *Suppose Assumptions B and C are valid*[7] *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 1 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 2 initialized at* (X_0, Y_0) *satisfying* $\frac{\|(X_0,Y_0)-(X_\sharp,Y_\sharp)\|_F}{\sqrt{\sigma_r(M_\sharp)}} \lesssim \min\{1, \frac{\kappa_3}{\kappa_2\chi}\}$*, will generate an iterate sequence that converges at the linear rate:*

$$\frac{\operatorname{dist}((X_k,Y_k),\mathcal{D}^*(M_\sharp))}{\sqrt{\sigma_r(M_\sharp)}} \le \delta \qquad \text{after} \qquad k \gtrsim \frac{\kappa_2^2\chi^2}{\kappa_3^2}\cdot\ln\left(\frac{\eta}{\delta}\right) \qquad \text{iterations.}$$

3. **(prox-linear)** *Algorithm 3 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:*

$$\frac{\operatorname{dist}((X_k,Y_k),\mathcal{D}^*(M_\sharp))}{\sqrt{\sigma_r(M_\sharp)}} \lesssim \frac{\kappa_3}{\kappa_2}\cdot 2^{-2^k} \qquad \text{for all } k \ge 0.$$

6 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 5.5 and 5.9 immediately apply. Most of the RIP results in this section were previously known (albeit under more restrictive assumptions); we provide self-contained arguments in the Appendix B 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.

[7] with $|||\cdot||| = \frac{1}{m}\|\cdot\|_1$

Definition 6.1 (**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

$$b_i = \begin{cases} \mathcal{A}(M_\sharp) & \text{if } i \notin \mathcal{I}, \text{ and} \\ \eta_i & \text{if } i \in \mathcal{I}, \end{cases} \tag{6.1}$$

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:

$$\mathcal{A}(M_\sharp)_i := \langle P_i, M_\sharp \rangle \qquad \text{for } i = 1, \dots, m \tag{6.2}$$

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:

$$\mathcal{A}(M_\sharp)_i = p_i^\top M_\sharp p_i = \|X_\sharp^\top p_i\|_2^2 \qquad \text{for } i = 1, \dots, m, \tag{6.3}$$

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:

$$\mathcal{A}(M_\sharp)_i = p_i^\top M_\sharp p_i - \tilde{p}_i^\top M_\sharp \tilde{p}_i = \|X_\sharp^\top p_i\|_2^2 - \|X_\sharp^\top \tilde{p}_i\|_2^2 \qquad \text{for } i = 1, \dots, m, \tag{6.4}$$

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:

$$\mathcal{A}(M_\sharp)_i = p_i^\top M_\sharp q_i \qquad \text{for } i = 1, \dots, m, \tag{6.5}$$

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 [13, 30, 50]. Quadratic sensing in its full generality appeared in [18] and is a higher-rank generalization of the much older (real) phase retrieval problem [10, 14, 35]. Besides phase retrieval, quadratic sensing has applications to covariance sketching, shallow neural networks, and quantum state tomography; see for example [40] for a discussion. Bilinear sensing is a natural modification of quadratic sensing and is a higher-rank generalization of the blind deconvolution problem [1]; it was first proposed and studied in [8].

The reader is reminded that once ℓ_1/ℓ_2 RIP guarantees, in particular Assumptions B and C, are established for the above four operators, the guarantees of Corollaries 5.5 and Corollary 5.9 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.

6.1 Warm-up: ℓ_2/ℓ_2 RIP for matrix sensing with Gaussian design

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 [13, 50].

Theorem 6.2 (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 B is satisfied.

Following the general recipe of the paper, we see that the nonsmooth formulation

$$\min_{X \in \mathbf{R}^{d_1 \times r},\; Y \in \mathbf{R}^{r \times d_2}} \frac{1}{\sqrt{m}} \|\mathcal{A}(XY) - b\|_2 = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(\mathrm{Tr}(Y P_i^\top X) - b_i\right)^2} \tag{6.6}$$

is immediately amenable to subgradient and prox-linear algorithms in the noiseless setting $\mathcal{I} = \emptyset$. In particular, a direct analogue of Corollary 5.9, which was stated for the penalty function $h = \frac{1}{m}\|\cdot\|_1$, holds; we omit the straightforward details.

6.2 The ℓ_1/ℓ_2 RIP and $\mathcal{I}$-outlier bounds: quadratic and bilinear sensing

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 6.2. 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 G (Matrix Sensing)**.** The matrices $\{P_i\}$ are i.i.d. realizations of an η-sub-Gaussian random matrix[8] $P \in \mathbf{R}^{d_1 \times d_2}$. Furthermore, there exists a numerical constant $\alpha > 0$ such that

$$\inf_{\substack{M:\ \mathrm{Rank}\ M \leq 2r \\ \|M\|_F = 1}} \mathbb{E}|\langle P, M \rangle| \geq \alpha. \tag{6.7}$$

Assumption H (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

$$\inf_{\substack{M \in \mathcal{S}^d:\ \mathrm{Rank}\ M \leq 2r \\ \|M\|_F = 1}} \mathbb{E}|p^\top M p| \geq \alpha. \tag{6.8}$$

Assumption I (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

$$\inf_{\substack{M \in \mathcal{S}^d:\ \mathrm{Rank}\ M \leq 2r \\ \|M\|_F = 1}} \mathbb{E}|p^\top M p - \tilde{p}^\top M \tilde{p}| \geq \alpha. \tag{6.9}$$

Assumption J (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

$$\inf_{\substack{M:\ \mathrm{Rank}\ M \leq 2r \\ \|M\|_F = 1}} \mathbb{E}|p^\top M q| \geq \alpha. \tag{6.10}$$

The Assumptions G-J are all valid for i.i.d. Gaussian realizations with independent identity covariance, as the following lemma shows. We defer its proof to Appendix B.1.

Lemma 6.3. *Assumption G holds for matrices P with i.i.d. standard Gaussian entries. Assumptions H and I hold for vectors $p, \tilde{p}$ with i.i.d. standard Gaussian entries. Assumption J 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 6.1. Then, we wish to establish the inequalities

$$\kappa_1 \|M\|_F \leq \frac{1}{m} \|\mathcal{A}(M)\|_1 \leq \kappa_2 \|M\|_F \tag{6.11}$$

and

$$\kappa_3 \|M\|_F \leq \frac{1}{m} \big(\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 \big), \tag{6.12}$$

and, hence, Assumptions B and C, respectively, for certain constants κ_1, κ_2, and κ_3. We defer the proof of this theorem to Appendix B.2.

Theorem 6.4 (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* (6.2)*,* (6.3)*,* (6.4)*, and* (6.5)*, respectively*

[8] By this we mean that the vectorized matrix $\mathbf{vec}(P)$ is a η-sub-gaussian random vector.

1. **(Matrix sensing)** *Suppose Assumption G 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* (6.11) *and* (6.12) *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 H 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* (6.11) *and* (6.12) *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 I 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* (6.11) *and* (6.12) *with constants* $\kappa_1 = c_4, \kappa_2 = c_5$ *and* $\kappa_3 = c_6(1-2p_{\text{fail}})$.

4. **(Bilinear sensing)** *Suppose Assumption J 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* (6.11) *and* (6.12) *with constants* $\kappa_1 = c_4, \kappa_2 = c_5$ *and* $\kappa_3 = c_6(1-2p_{\text{fail}})$.

The guarantees of Theorem 6.4 were previously known under stronger assumptions. In particular, item (1) generalizes the results in [39] for the pure Gaussian setting. The case $r = 1$ of item (2) can be found, in a sightly different form, in [27, 29]. Item (3) sharpens slightly the analogous guarantee in [18] by weakening the assumptions on the moments of the measuring vectors to the uniform lower bound (6.9). Special cases of item (4) were established in [16], for the case $r = 1$, and [8], 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 (6.2) is that md_1d_2 scalars are required to represent the map $\mathcal{A}$. In contrast, all other measurement operators may be represented with only $m(d_1 + d_2)$ scalars.

7 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 8, 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 [19].

Lemma 7.1 (Sharpness [19])**.** *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 7.2 (Lemma 5 in [19])**.** *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 7.3 (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*

$$\begin{aligned}\frac{1}{\sqrt{p}}|f(X) - f_Y(X)| &\leq \sqrt{(1+\epsilon)}\|X-Y\|_F^2 + \sqrt{\epsilon}\|X-Y\|_F, \\ |f(X) - f(Y)| &\leq \sqrt{p\nu r}\|X-Y\|_F,\end{aligned}$$

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 7.2. 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 7.2. □

Note that the approximation bound in Lemma 7.2 is not in terms of the square Euclidean norm. Therefore the results in Section 5 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 the appendix.

Lemma 7.4. *Suppose that Assumption A 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:

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f_{x_k}(x) + a\|x - x_k\|_2^2 + b\|x - x_k\|_2 \right\}.$$

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 7.4 with our estimates of the sharpness and approximation accuracy, we deduce the following convergence guarantee for matrix completion.

Corollary 7.5 (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*

$$X_{k+1} = \operatorname*{argmin}_{X \in \mathcal{X}} \left\{ f_{X_k}(X) + \sqrt{p(1+\epsilon)} \cdot \|X - X_k\|_2^2 + \sqrt{p\epsilon}\|X - X_k\|_2 \right\} \tag{7.1}$$

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 4.3 and Lemmas 7.1 and 7.3 we may appeal to Lemma 7.4 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 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 [23, Theorem 3.1], and appears in the appendix.

Theorem 7.6. *Suppose that Assumption A 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 1. 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 7.6 with our estimates of the sharpness and approximation accuracy, we deduce the following convergence guarantee for matrix completion.

Corollary 7.7 (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 1 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 $L \leq \sqrt{p\nu r}$ by Lemma 7.3. By invoking Proposition 4.3 and Lemmas 7.1 and 7.3 we may appeal to Theorem 7.6 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 10. The caveat is that the per iteration cost of the prox-linear method is significantly higher than that of the subgradient method.

8 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 [11,15,59]. 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.

8.1 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 $\|X_\sharp\|_{\mathrm{op}} \leq 1$ and $\|X_\sharp\|_{2,\infty} \leq \sqrt{\frac{\nu r}{d}}$.[9]

[9] Recall that $\|X\|_{2,\infty} = \max_{i \in [d]} \|X_{i\cdot}\|_2$ is the maximum row norm.

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:

$$\min_{X\in\mathcal{X},S\in\mathcal{S}} F\big((X,S)\big) = \|XX^\top + S - W\|_F, \tag{8.1}$$

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 [19, 32]. 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 Appendix D.1.

Lemma 8.1. *For all $X \in \mathcal{X}$ and $S \in \mathcal{S}$, the estimate holds:*

$$|\langle S - S_\sharp, XX^\top - X_\sharp X_\sharp^\top\rangle| \le 10\sqrt{\frac{\nu r k}{d}} \cdot \|S - S_\sharp\|_F \cdot \|X - X_\sharp\|_F.$$

Equipped with the above lemma, we can estimate the sharpness and approximation parameters μ, ρ for the formulation (8.1).

Lemma 8.2 (Regularity constants)**.** *For all $X \in \mathcal{X}$ and $S \in \mathcal{S}$, the estimates hold:*

$$F((X,S))^2 \ge \left(\frac{1}{2}\sigma_r^2(X_\sharp) - 10\sqrt{\frac{\nu r k}{d}}\right) \cdot \left(\operatorname{dist}(X, \mathcal{D}^*(M_\sharp))^2 + \|S - S_\sharp\|_F^2\right) \tag{8.2}$$

and

$$|F((X,S)) - F_Y((X,S))| \le \|X - Y\|_F^2. \tag{8.3}$$

Moreover, for any $X_1, X_2 \in \mathcal{X}$ and $S_1, S_2 \in \mathcal{S}$, the Lipschitz bounds holds:

$$|F((X_1,S_1)) - F((X_2,S_2))| \le 2\sqrt{\nu r}\|X_1 - X_2\|_F + \|S_1 - S_2\|_F.$$

Proof. Let $X_\sharp \in \operatorname{proj}_{\mathcal{D}^*(M_\sharp)}(X)$. To establish the bound (8.2), we observe that

$$\begin{aligned}\|XX^\top + S - W\|_F^2 &= \|XX^\top - M_\sharp\|_F^2 + 2\langle S - S_\sharp, XX^\top - M_\sharp\rangle + \|S - S_\sharp\|_F^2 \\ &\ge \frac{1}{2}\sigma_r^2(X_\sharp)\|X - X_\sharp\|_F^2 - 20\sqrt{\frac{\nu r k}{d}}\|S - S_\sharp\|_F\|X - X_\sharp\|_F + \|S - S_\sharp\|_F^2,\end{aligned}$$

where the first inequality follows from Proposition 4.3 and Lemma 8.1. 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

$$2abc \le a\varepsilon b^2 + (a/\varepsilon)c^2.$$

Thus, for any $\varepsilon > 0$, we have

$$\|XX^\top + S - W\|_F^2 \ge sb^2 - 2abc + c^2 \ge (s - a\varepsilon)b^2 + (1 - a/\varepsilon)c^2.$$

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 $s - a\varepsilon \ge \frac{1}{2}\sigma_r^2(X_\sharp) - 10\sqrt{\nu r k/d}$, the claimed bound (8.2) follows immediately. The bound (8.3) follows from the reverse triangle inequality:

$$\begin{aligned}
|F((X,S)) - F_Y((X,S))| &\le \|XX^\top - YY^\top - (X-Y)Y^\top - Y^\top(X-Y)\|_F \\
&= \|XX^\top - XY^\top - YX^\top + YY^\top\|_F \\
&= \|(X-Y)(X-Y)^\top\|_F \\
&\le \|X-Y\|_F^2.
\end{aligned}$$

Finally observe

$$\begin{aligned}
|F((X_1,S_1)) - F((X_2,S_2))| &\le \|X_1X_1^\top - X_2X_2^\top\|_F + \|S_1 - S_2\|_F \\
&\le \|X_1 + X_2\|_{\mathrm{op}}\|X_1 - X_2\|_F + \|S_1 - S_2\|_F \\
&\le 2\sqrt{\nu r}\|X_1 - X_2\|_F + \|S_1 - S_2\|_F,
\end{aligned}$$

where we use the bound $\|X_i\|_{\mathrm{op}} \le \sqrt{d}\|X_i\|_{2,\infty} \le \sqrt{\nu r}$ 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

$$\sqrt{\frac{\nu r k}{d}} \le c_0\sigma_r^2(X_\sharp),$$

the Polyak subgradient method will converge at the linear rate,

$$\operatorname{dist}(X_k, \mathcal{D}^*(M_\sharp)) \le \left(1 - \frac{c_1\sigma_r^2(X_\sharp)}{\nu r}\right)^{\frac{k}{2}} \cdot c_2\mu,$$

and the prox-linear method will converge quadratically when initialized at $X_0 \in \mathcal{X}$ satisfying $\operatorname{dist}(X_0, \mathcal{D}^*(M_\sharp)) < c_2\sigma_r(X_\sharp)$.

8.2 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

$$\min_{X\in\mathcal{X}} f(X) = \|XX^\top - W\|_1 \quad \text{where } \mathcal{X} = \{X \in \mathbf{R}^{d\times r} \mid \|X\|_{2,\infty} \le C\|X_\sharp\|_{2,\infty}\}, \tag{8.4}$$

for a constant $C > 1$. Unlike Section 8.1, 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 $\|X_\sharp\|_{2,\infty} \leq \sqrt{\frac{\nu r}{d}}\|X_\sharp\|_{\mathrm{op}}$.

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 8.3 (Approximation accuracy)**.** *The estimate holds:*

$$|f(Z) - f_X(Z)| \leq \|Z - X\|_{2,1}^2 \qquad \textit{for all } X, Z \in \mathbf{R}^{d\times r}.$$

Proof. As in the proof of Proposition 4.1, we compute

$$\begin{aligned}
|f(Z) - f_X(Z)| &= \Big|\|ZZ^\top - W\|_1 - \|XX - W + X(Z-X)^\top + (Z-X)X^\top\|_1\Big| \\
&\leq \|(Z-X)(Z-X)^\top\|_1 = \sum_{i,j} |e_i^\top(Z-X)(e_j^\top(Z-X))^\top| \\
&\leq \sum_{i,j} \|e_i^\top(Z-X)\|_2 \cdot \|e_j^\top(Z-X)\|_2 = \|Z-X\|_{2,1}^2,
\end{aligned}$$

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 (8.4), 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 5.8.

Theorem 8.4. *Suppose that Assumption A holds where $\|\cdot\|$ is replaced by an arbitrary norm $|||\cdot|||$. Choose any $\beta \geq \rho$ and set $\gamma := \rho/\beta$ in Algorithm 3. Then Algorithm 3 initialized at any point x_0 satisfying $\mathrm{dist}_{|||\cdot|||}(x_0, \mathcal{X}^*) < \frac{\mu}{\rho}$ converges quadratically:*

$$\mathrm{dist}_{|||\cdot|||}(x_{k+1}, \mathcal{X}^*) \leq \tfrac{\rho}{\mu} \cdot \mathrm{dist}^2_{|||\cdot|||}(x_k, \mathcal{X}^*) \qquad \forall k \geq 0.$$

To apply the above generic convergence guarantees for the prox-linear method, it remains to show that the objective function f in (8.4) 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 4.3 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 8.5 (Sharpness (rank one))**.** *Consider two vectors $x, \bar{x} \in \mathbf{R}^d$ satisfying*

$$\operatorname{dist}_{\|\cdot\|_1}(x, \{\pm\bar{x}\}) \leq (\sqrt{2}-1)\|\bar{x}\|_1.$$

Then the estimate holds:

$$\|xx^\top - \bar{x}\bar{x}^\top\|_1 \geq (\sqrt{2}-1) \cdot \|\bar{x}\|_1 \cdot \operatorname{dist}_{\|\cdot\|_1}(x, \{\pm\bar{x}\}).$$

The proof of this result appears in Appendix D.2. We leave as an intriguing open question to determine if an analogous result holds in the higher rank setting.

Conjecture 8.6 (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:*

$$\|XX^\top - M\|_1 \geq c \cdot \operatorname{dist}_{\|\cdot\|_{2,1}}(X, \mathcal{D}^*),$$

for all $X \in \mathcal{X}$ satisfying $\operatorname{dist}_{\|\cdot\|_{2,1}}(X, \mathcal{D}^*) \leq \gamma$.

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 Appendix D.3. 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 8.7. *Assume Conjecture 8.6. 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:*

$$\sum_{i,j=1}^{d} \delta_{ij} |\langle X_i, X_j\rangle - \langle (X_\sharp)_i, (X_\sharp)_j\rangle| \leq \left(\tau + \frac{c_3 C\sqrt{\tau\nu r \log d}}{c}\|X_\sharp\|_{\mathrm{op}}\right)\|XX^\top - X_\sharp X_\sharp^\top\|_1$$

for all $X \in \mathcal{X}$ satisfying $\operatorname{dist}_{\|\cdot\|_{2,1}}(X, \mathcal{D}^*) \leq \gamma$.

We remark that we expect c to scale with $\|X_\sharp\|_{\mathrm{op}}$ in the above bound, yielding a ratio $\|X_\sharp\|_{\mathrm{op}}/c$ dependent on the conditioning of $X_\sharp$. Given the above lemma, sharpness of f quickly follows.

Lemma 8.8 (Sharpness of Non-Euclidean Robust PCA)**.** *Assume Conjecture 8.6. 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:*

$$f(X) - f(X_\sharp) \geq c \cdot \left(1 - 2\tau - \frac{2c_3 C\sqrt{\tau\nu r \log d}}{c}\|X_\sharp\|_{op}\right) \cdot \operatorname{dist}_{\|\cdot\|_{2,1}}(X, \mathcal{D}^*(M_\sharp))$$

for all $X \in \mathcal{X}$ satisfying and $\operatorname{dist}_{\|\cdot\|_{2,1}}(X, \mathcal{D}^*(M_\sharp)) \leq \gamma$.

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 8.7. □

Combining Lemma 8.8 and Theorem 8.4, we deduce the following convergence guarantee.

Theorem 8.9 (Convergence for non-Euclidean Robust PCA)**.** *Assume Conjecture 8.6. Then there exist constants $c_1, c_2, c_3 > 0$ so that for all τ satisfying $1 - 2\tau - 2c_3 C\sqrt{\tau\nu r \log d}\|X_\sharp\|_{op}/c > 0$ 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*

$$X_{k+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left\{ f_{X_k}(X) + \frac{1}{2\gamma}\|X - X_k\|_{2,1}^2 \right\} \tag{8.5}$$

satisfy

$$\operatorname{dist}_{\|\cdot\|_{2,1}}(X_{k+1}, \mathcal{D}^*(M_\sharp)) \leq \frac{2}{c \cdot \left(1 - 2\tau - \frac{2c_3 C\sqrt{\tau\nu r \log d}}{c}\|X_\sharp\|_{op}\right)} \cdot \operatorname{dist}^2_{\|\cdot\|_{2,1}}(X_k, \mathcal{D}^*(M_\sharp)), \qquad \forall k \geq 0.$$

In particular, the iterates converge quadratically as long as the initial iterate $X_0 \in \mathcal{X}$ satisfies

$$\operatorname{dist}_{\|\cdot\|_{2,1}}(X_0, \mathcal{D}^*(M_\sharp)) < \min\left\{ (1/2)c \cdot \left(1 - 2\tau - \frac{2c_3 C\sqrt{\tau\nu r \log d}}{c}\|X_\sharp\|_{op}\right), \gamma \right\}.$$

9 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

$$\min_{x\in\mathcal{X}} \ \tilde{f}(x) := h(F(x) + e). \tag{9.1}$$

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

$$\operatorname{dist}(x^*, \mathcal{X}^*) \leq (2\eta/\mu)|||e|||, \tag{9.2}$$

where as before we set $\mathcal{X}^* = \operatorname{argmin}_{\mathcal{X}} f$. In this section, we show that subgradient and prox-linear algorithms applied to the perturbed problem (9.1) 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 4.2.2, 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 9.1.

Throughout the rest of the section, let us define the noise level:

$$\varepsilon := \eta|||e|||.$$

Mirroring the discussion in Section 5, define the annulus:

$$\widetilde{\mathcal{T}}_\gamma := \left\{ z \in \mathcal{X} : \frac{14\varepsilon}{\mu} < \operatorname{dist}(z, \mathcal{X}^*) < \frac{\gamma\mu}{4\rho} \right\}, \tag{9.3}$$

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 the paper, it is possible to show directly that $\tilde{L} \leq L$ as defined in Assumption D. 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 (5.2); the proof appears in the appendix.

Lemma 9.1. *Suppose $\epsilon < \frac{\mu^2}{56\rho}$ so that $\widetilde{\mathcal{T}}_1$ is nonempty. Then the following bound holds:*

$$\tilde{L} \le \sup\left\{\|\zeta\|_2 : \zeta \in \partial f(x), \operatorname{dist}(x, \mathcal{X}^*) \le \frac{\mu}{\rho}, \operatorname{dist}(x, \mathcal{X}) \le 2\sqrt{\frac{\varepsilon}{\rho}}\right\} + 2\sqrt{8\rho\varepsilon}.$$

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 9.2 (Approximate sharpness)**.** *We have the following bound:*

$$\tilde{f}(x) - \inf_{\mathcal{X}} f \ge \mu \cdot \operatorname{dist}(x, \mathcal{X}^*) - \varepsilon \qquad \forall x \in \mathcal{X}. \tag{9.4}$$

Proof. For any $x \in \mathcal{X}$, observe $\tilde{f}(x) - \inf f \ge f(x) - \inf f - \varepsilon \ge \mu \cdot \operatorname{dist}(x, \mathcal{X}^*) - \varepsilon$, as claimed. □

Next, we show that $\tilde{f}$ satisfies the following approximate subgradient inequality.

Lemma 9.3 (Approximate subgradient inequality)**.** *The following bound holds:*

$$f(y) \ge \tilde{f}(x) + \langle \zeta, y - x\rangle - \frac{\rho}{2}\|x - y\|^2 - 3\varepsilon \qquad \forall x, y \text{ and } \zeta \in \partial\tilde{f}(x).$$

Proof. First notice that $|f_x(y) - \tilde{f}_x(y)| \le \varepsilon$ 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

$$\begin{aligned}
\langle \zeta, y - x\rangle &\le \tilde{f}_x(y) - \tilde{f}_x(x) \\
&\le f_x(y) - f_x(x) + 2\eta|||e||| \\
&\le f(y) - f(x) + \frac{\rho}{2}\|x - y\|^2 + 2\varepsilon \\
&\le f(y) - \tilde{f}(x) + \frac{\rho}{2}\|x - y\|^2 + 3\varepsilon,
\end{aligned}$$

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 the appendix.

Algorithm 4: Modified Polyak Subgradient Method

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

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

Choose $\zeta_k \in \partial\tilde{f}(x_k)$. **If** $\zeta_k = 0$, then exit algorithm.

Set $x_{k+1} = \operatorname{proj}_{\mathcal{X}}\left(x_k - \dfrac{\tilde{f}(x_k) - \min_{\mathcal{X}} f}{\|\zeta_k\|^2}\zeta_k\right)$.

Theorem 9.4 (Polyak subgradient method)**.** *Suppose that Assumption D holds and suppose that $\varepsilon \leq \mu^2/14\rho$. Then Algorithm 4 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*

$$\operatorname{dist}^2(x_{k+1}, \mathcal{X}^*) \leq \left(1 - \frac{13\tilde{\tau}^2}{56}\right) \operatorname{dist}^2(x_k, \mathcal{X}^*) \qquad \forall k \geq 0 \text{ with } \operatorname{dist}(x_k, \mathcal{X}^*) \geq 14\varepsilon/\mu.$$

Next we provide theoretical guarantees for Algorithm 5.3, where one does not know the optimal value $\min_{\mathcal{X}} f$. The proof of this result is a straightforward modification of [23, Theorem 6.1] based on the Lemmas 9.2 and 9.3, and therefore we omit it.

Theorem 9.5 (Geometrically decaying subgradient method)**.** *Suppose that Assumption D 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 2. Then the iterates x_k generated by Algorithm 2 on the perturbed problem (9.1), initialized at a point $x_0 \in \tilde{\mathcal{T}}_\gamma$, satisfy:*

$$\operatorname{dist}^2(x_k; \mathcal{X}^*) \leq \frac{\gamma^2\mu^2}{16\rho^2}\left(1-(1-\gamma)\tilde{\tau}^2\right)^k \qquad \forall k \geq 0 \text{ with } \operatorname{dist}(x_k, \mathcal{X}^*) \geq 14\varepsilon/\mu.$$

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 the appendix.

Theorem 9.6 (Prox-linear algorithm)**.** *Suppose Assumptions A holds. Choose any $\beta \geq \rho$ in Algorithm 3 applied to the perturbed problem (9.1) and set $\gamma := \rho/\beta$. Suppose moreover $\epsilon < \frac{\mu^2\gamma}{56\rho}$ so that $\tilde{\mathcal{T}}_\gamma$ is nonempty. Then Algorithm 3 initialized at any point $x_0 \in \widetilde{\mathcal{T}}_\gamma$ converges quadratically up to tolerance $14\varepsilon/\mu$:*

$$\operatorname{dist}(x_{k+1}, \mathcal{X}^*) \leq \tfrac{7\beta}{6\mu} \cdot \operatorname{dist}^2(x_k, \mathcal{X}^*) \qquad \forall k \geq 0 \text{ with } \operatorname{dist}(x_{k+1}, \mathcal{X}^*) \geq 14\varepsilon/\mu.$$

9.1 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 4.2.2, in particular Assumption C, to the following observation model.

Assumption K ($\mathcal{I}$-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) $b = \mathcal{A}(M_\natural) + \Delta + e$.

(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 9.7. *Suppose Assumption B and K are valid. Then the following hold:*

1. **(Sharpness)** *We have*

$$f(X) - f(X_\sharp) \geq \mu \cdot \operatorname{dist}\big(X, \mathcal{D}^*(M_\sharp)\big) \qquad \text{for all } X \in \mathbf{R}^{d\times r} \text{ and } \mu := \kappa_3\sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp),$$

2. **(Weak Convexity)** *The function f is $\rho := 2\kappa_2$-weakly convex.*

3. **(Minimizers)** *All minimizers of $\tilde{f}$ satisfy*

$$\operatorname{dist}(X^*, \mathcal{X}^*) \leq \frac{2\frac{1}{m}\|e\|_1}{\kappa_3\sqrt{2(\sqrt{2}-1)}\sigma_r(X_\sharp)} \qquad \forall X^* \in \operatorname*{argmin}_{\mathcal{X}} \tilde{f}.$$

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 4.6, while weak convexity follows from Proposition 4.1. The minimizer bound follows from (9.2). Finally, due to Lemma 3.1, the argument given in Proposition (4.1), 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 9.7 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 9.8 (Convergence guarantees under RIP with sparse outliers and dense noise (symmetric))**.** *Suppose Assumptions B is and K 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 1 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 2 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:*

$$\frac{\operatorname{dist}^2(X_k, \mathcal{D}^*(M_\sharp))}{\sigma_r(M_\sharp)} \leq \left(1 - \frac{c_2}{1 + \frac{c_3\kappa_2^2\chi}{\kappa_3^2}}\right)^k \cdot \frac{c_4\kappa_3^2}{\kappa_2^2} \qquad \forall k \geq 0 \text{ with } \operatorname{dist}(X_k, \mathcal{X}^*) \geq \delta.$$

3. **(prox-linear)** *Algorithm 3 with $\beta = \rho$ and initialized at X_0 converges quadratically:*

$$\frac{\operatorname{dist}(X_k, \mathcal{D}^*(M_\sharp)))}{\sqrt{\sigma_r(M_\sharp)}} \leq 2^{-2^k} \cdot \frac{c_6\kappa_3}{\kappa_2} \qquad \forall k \geq 0 \text{ with } \operatorname{dist}(X_k, \mathcal{X}^*) \geq \delta.$$

10 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.

10.1 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

$$X_0 := X_\sharp + \delta \cdot \|X_\sharp\|_F \cdot \Delta, \quad \text{where } \Delta = \frac{G}{\|G\|_F},\ G_{ij} \sim_{\text{i.i.d.}} N(0, I).$$

   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 step-size, 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 2 and 3 depict the phase transition plots for bilinear (6.5) and symmetrized quadratic (6.4) 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.

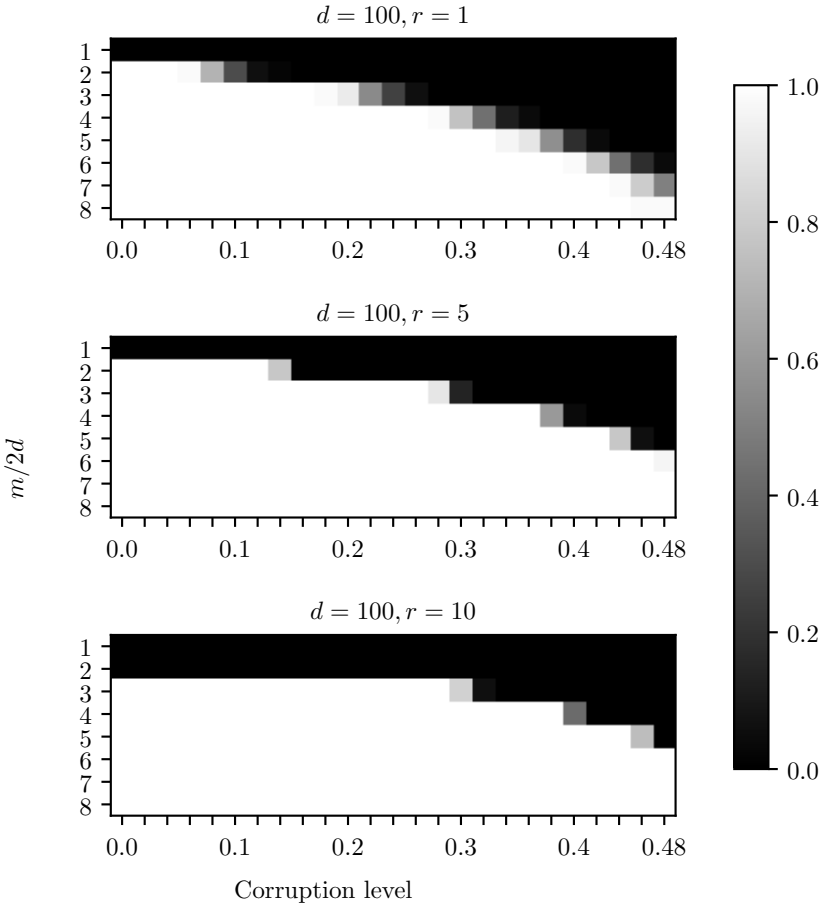


Figure 2: Bilinear sensing with $d_1 = d_2 = d = 100$ using Algorithm 2.

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 8.2, 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 3, and the prox-linear algorithm (8.5). 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 [48]. We set tolerance $\epsilon_k = \frac{10^{-4}}{2k}$ for the proximal subproblems, which we continue solve for at most 500 iterations. We choose $\gamma = 10$ in all subproblems. The phase transition plots are shown in Figure 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 depicts the empirical recovery rate using

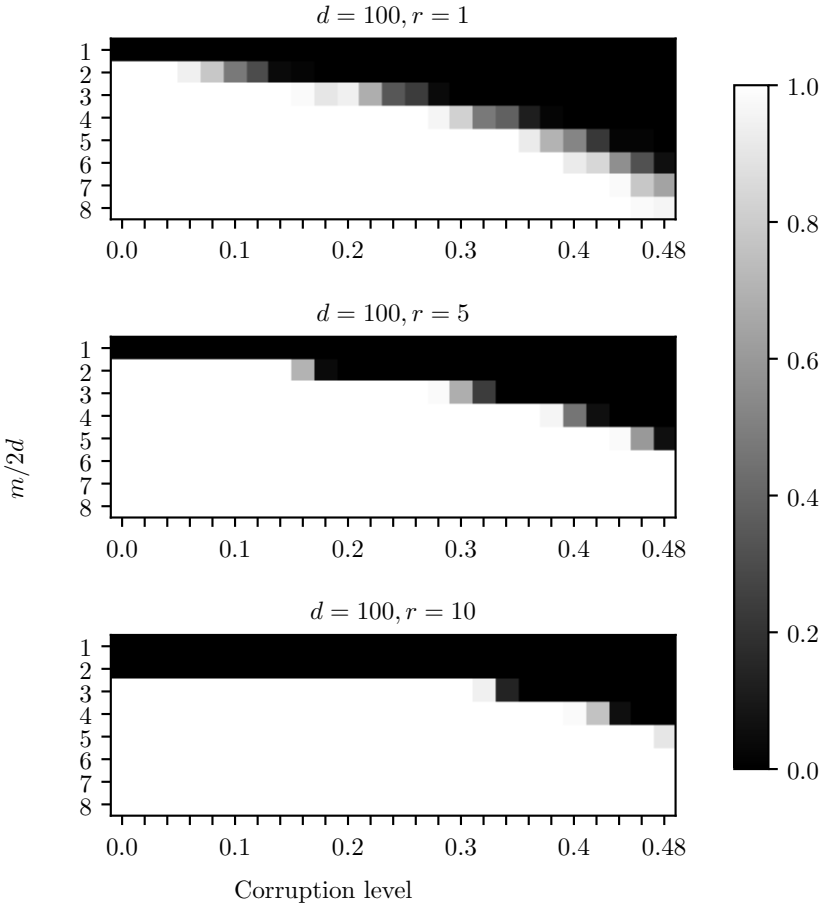


Figure 3: Quadratic sensing with symmetrized measurements using Algorithm 2.

the Polyak subgradient method and the modified prox-linear algorithm (7.1). Similarly to the quadratic/bilinear sensing problems, low-rank matrix completion exhibits a sharp phase transition. As predicted in Section 7, 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.

10.2 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.

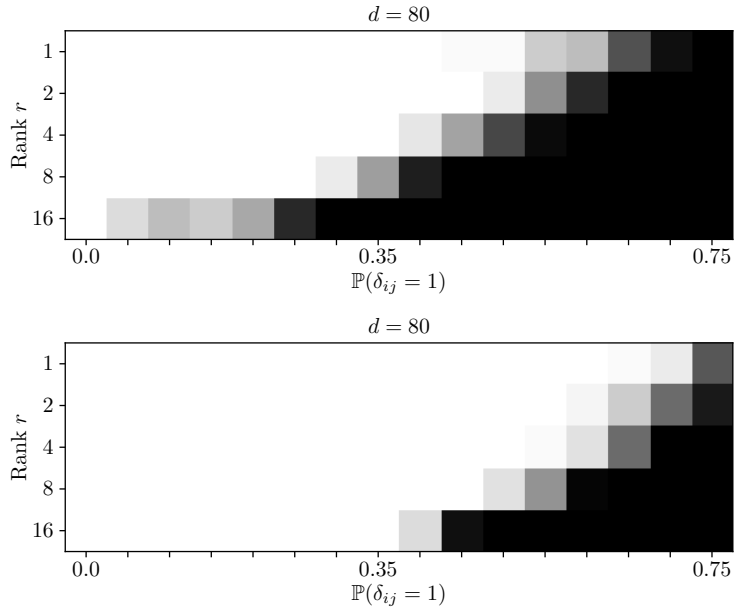


Figure 4: Robust PCA using the subgradient method, Algorithm 2, (top) and the modified prox-linear method (8.5) (bottom).

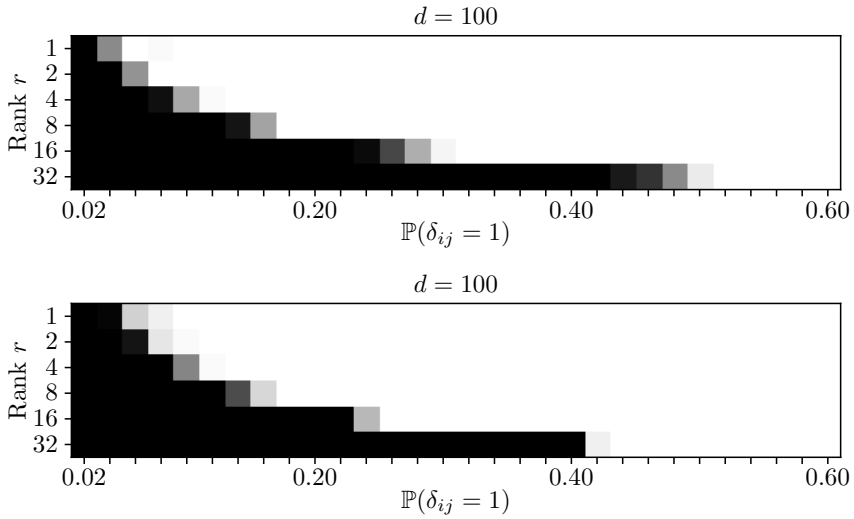


Figure 5: Low-rank matrix completion using the subgradient method, Algorithm 1 (top), and the modified prox-linear method (7.1) (bottom).

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 6, we plot the performance of Algorithm 2 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 7 depicts the behavior of Algorithm 1 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 3 for both problems in Figure 8, where we solve the proximal subproblems approximately.

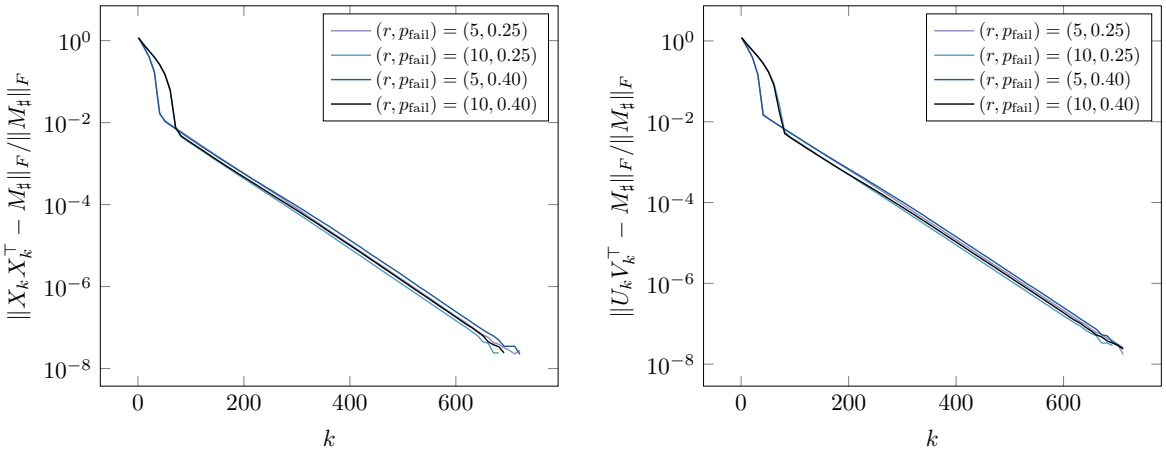


Figure 6: Quadratic (left) and bilinear (right) matrix sensing with $d = 200, m = 8 \cdot rd$, using the subgradient method, Algorithm 2.

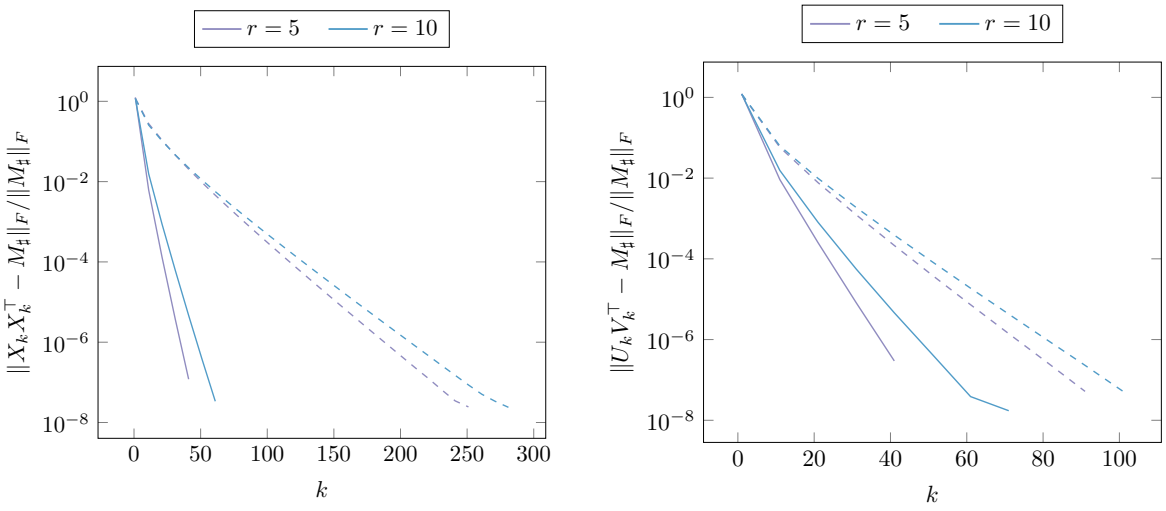


Figure 7: Algorithm 1 (solid lines) against gradient descent (dashed lines) with step size η. Left: quadratic sensing, $\eta = 10^{-4}$. Right: bilinear sensing, $\eta = 10^{-3}$.

Matrix completion. In our comparison with smooth methods, we apply gradient descent on the following minimization problem:

$$\min_{X \in \mathbf{R}^{d \times r} : \|X\|_{2,\infty} \le C} \left\| \Pi_\Omega(XX^\top) - \Pi_\Omega(M) \right\|_F^2 . \tag{10.1}$$

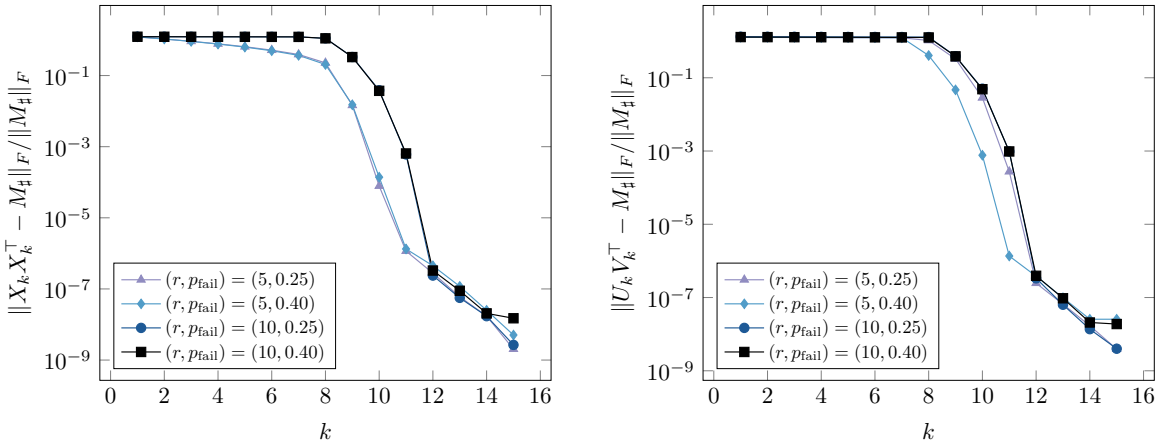


Figure 8: Quadratic (left) and bilinear (right) matrix sensing with $d = 100, m = 8 \cdot rd$, using the prox-linear method, Algorithm 3.

Figure 9 depicts the convergence behavior of Algorithm 1 (solid lines) versus gradient descent applied to Problem (10.1) 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 descent in all regimes.

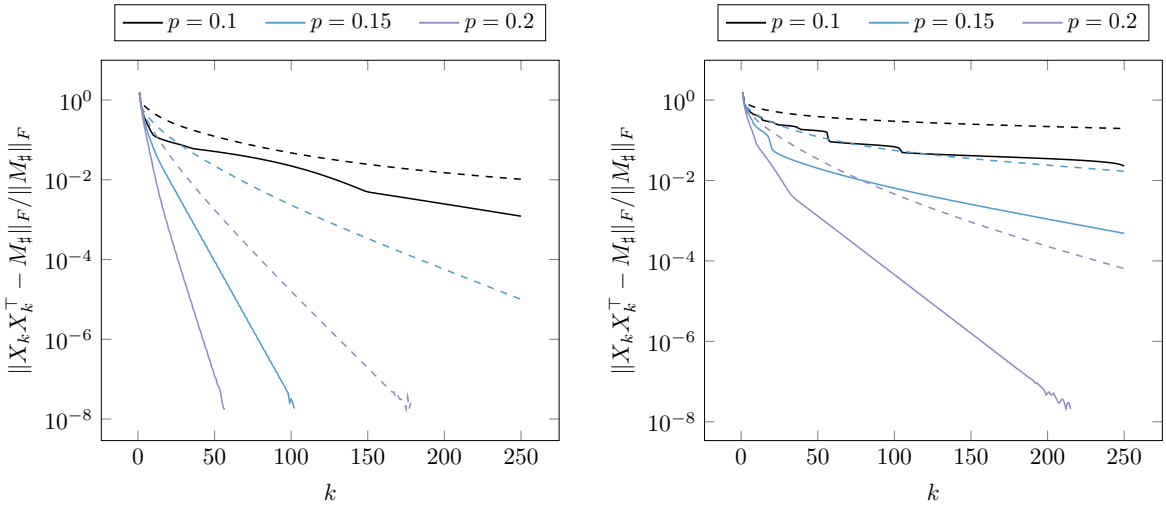


Figure 9: Low rank matrix completion with $d = 100$. Left: $r = 4$, right: $r = 8$. Solid lines use Algorithm 1, dashed lines use gradient descent with step $\eta = 0.004$.

Figure 10 depicts the performance of the modified prox-linear method (7.1) in the same setting as Figure 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 [48].

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

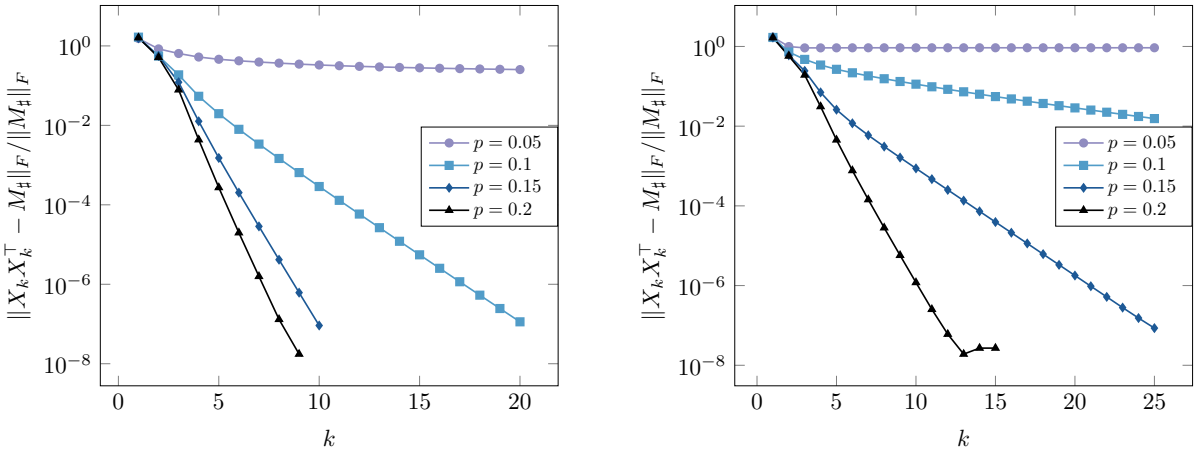


Figure 10: Low rank matrix completion with $d = 100$ using the modified prox-linear Algorithm (7.1). Left: $r = 4$, right: $r = 8$.

prox-linear methods, using the non-Euclidean formulation of Section 8.2. 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 11. The parameters of the prox-linear method are chosen in the same way reported in Section 10.1. In particular, our numerical experiments appear to support our sharpness Conjecture 8.6 for the robust PCA problem.

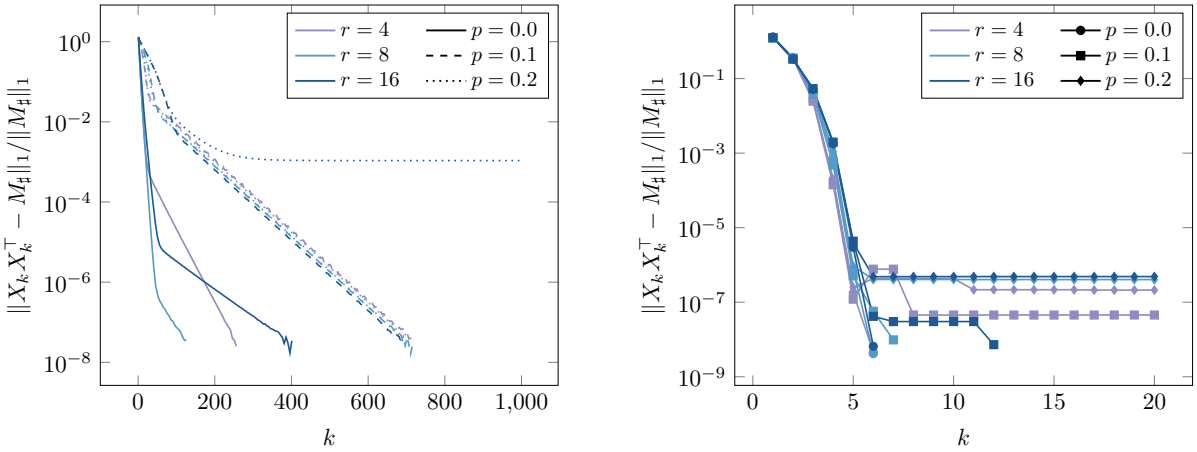


Figure 11: ℓ_1-robust PCA with $d = 100$ and $p := \mathbb{P}(\delta_{ij} = 1)$. Left: Algorithm 2, right: Algorithm (7.1).

10.2.1 Recovery up to tolerance

In this last section, we test the performance of the prox-linear method and the modified Polyak subgradient method (Algorithm 4) for the quadratic sensing and matrix completion problems, under a dense noise model of Section 9. 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 $\delta := 10^{-k}\sigma_r(X_\sharp),\ k \in \{1,\dots,4\}$. The relevant plots can be found in Figures 12 and 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 (7.1) is more robust to additive noise for the matrix completion problem, with Algorithm 4 exhibiting heavy fluctuations and failing to converge for the highest level of dense noise.

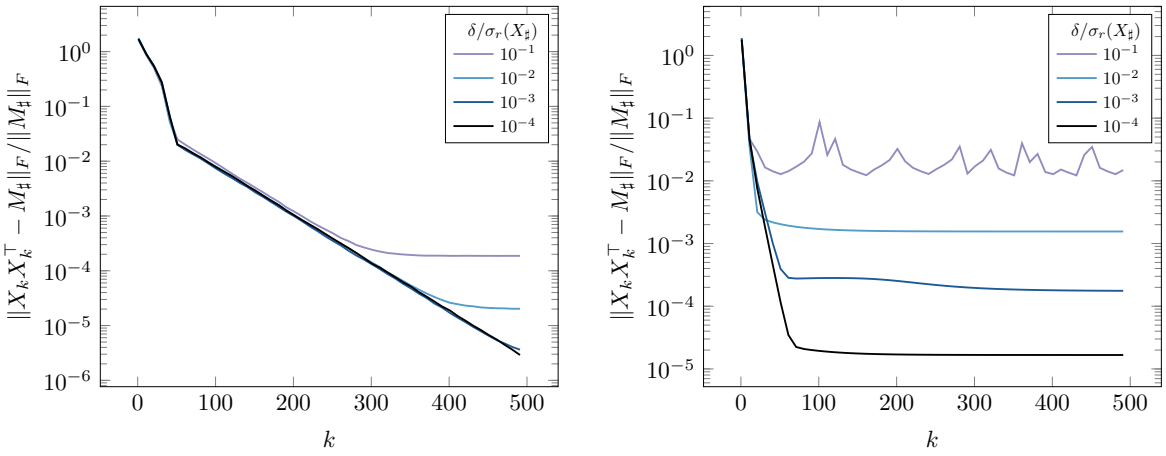


Figure 12: Quadratic sensing with $r = 5$ (left) and matrix completion with $r = 8$ (right), $d = 100$, using Algorithm 4.

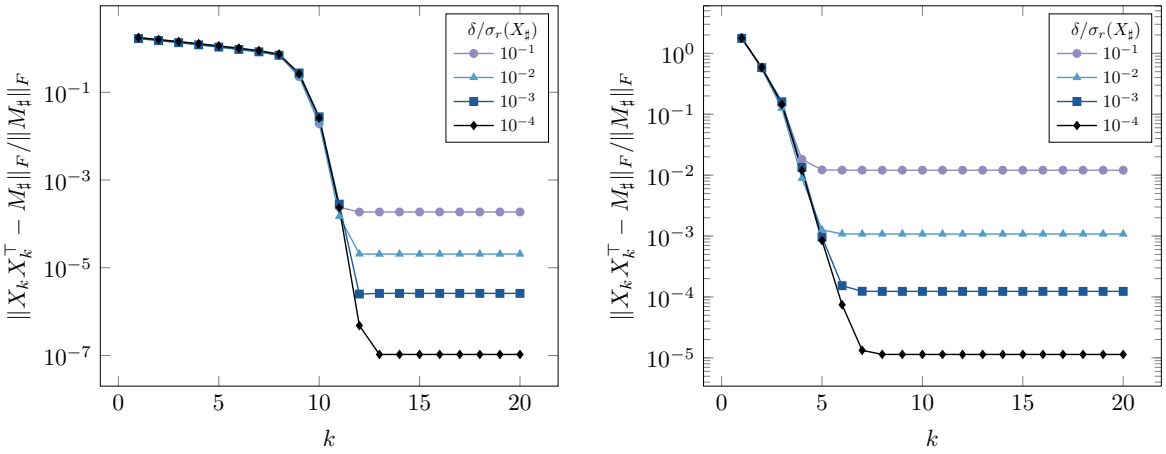


Figure 13: Quadratic sensing with $r = 5$ (left) and matrix completion with $r = 8$ (right), $d = 100$, using Algorithm (7.1).

References

[1] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

[2] 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.

[3] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[4] 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.

[5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[6] J.V. Burke. Descent methods for composite nondifferentiable optimization problems. *Math. Programming*, 33(3):260–279, 1985.

[7] J.V. Burke and M.C. Ferris. A Gauss-Newton method for convex composite optimization. *Math. Programming*, 71(2, Ser. A):179–194, 1995.

[8] T.T. Cai and A. Zhang. ROP: matrix recovery via rank-one projections. *Ann. Statist.*, 43(1):102–138, 2015.

[9] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[10] 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.

[11] E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):Art. 11, 37, 2011.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy. Composite optimization for robust blind deconvolution. *arXiv:1901.01624*, 2019.

[17] 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.

[18] 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.

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

[20] Y. Chi, Y.M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv:1809.09573*, 2018.

[21] 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.

[22] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[23] 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.

[24] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *To appear in IMA J. Numer. Anal., arXiv:1711.03247*, 2017.

[25] D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*, 43(3):919–948, 2018.

[26] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Math. Prog.*, pages 1–56, 2018.

[27] 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.

[28] J.C. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM J. Optim.*, 28(4):3229–3259, 2018.

[29] Y.C. Eldar and S. Mendelson. Phase retrieval: stability and recovery guarantees. *Appl. Comput. Harmon. Anal.*, 36(3):473–494, 2014.

[30] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[31] 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).

[32] 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.

[33] 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.

[34] J.L. Goffin. On convergence rates of subgradient optimization methods. *Math. Programming*, 13(3):329–347, 1977.

[35] T. Goldstein and C. Studer. Phasemax: Convex phase retrieval via basis pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, April 2018.

[36] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

[37] A.S. Lewis and S.J. Wright. A proximal method for composite minimization. *Math. Program.*, 158(1-2, Ser. A):501–546, 2016.

[38] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv:1606.04933*, 2016.

[39] X. Li, Z. Zhu, A.M.-C. So, and R. Vidal. Nonconvex robust low-rank matrix recovery. *arXiv:1809.09237*, 2018.

[40] Y. Li, C. Ma, Y. Chen, and Y. Chi. Nonconvex matrix factorization from rank-one measurements. *arXiv:1802.06286*, 2018.

[41] S. Ling and T. Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 31, 2015.

[42] 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, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[43] 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.

[44] S. Mendelson. Learning without concentration. *J. ACM*, 62(3):Art. 21, 25, 2015.

[45] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Grundlehren der mathematischen Wissenschaften, Vol 330, Springer, Berlin, 2006.

[46] 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.

[47] E.A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

[48] N. Parikh and S. Boyd. Block splitting for distributed optimization. *Mathematical Programming Computation*, 6(1):77–102, 2014.

[49] R.A. Poliquin and R.T. Rockafellar. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.*, 348:1805–1838, 1996.

[50] 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.

[51] 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.

[52] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[53] 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.

[54] M. Rudelson and R. Vershynin. Small ball probabilities for linear images of high-dimensional distributions. *International Mathematics Research Notices*, 2015(19):9594–9617, 2014.

[55] 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.

[56] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inform. Theory*, 62(11):6535–6579, 2016.

[57] 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.

[58] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[59] 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.

A Proofs in Section 5

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 5.1.

Lemma A.1. *Suppose Assumption E 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. □

A.1 Proof of Theorem 5.6

Let k be the first index (possibly infinite) such that $x_k \notin B_{\epsilon/2}(\bar{x})$. We claim that (5.4) holds for all $i < k$. We show this by induction. To this end, suppose (5.4) 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

$$\begin{aligned}
\|x_{i+1} - x^*\|^2 &= \left\|\mathrm{proj}_{\mathcal{X}}\left(x_i - \tfrac{f(x_i)-\min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right) - \mathrm{proj}_{\mathcal{X}}(x^*)\right\|^2 \\
&\leq \left\|(x_i - x^*) - \tfrac{f(x_i)-\min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right\|^2 \qquad \text{(A.1)} \\
&= \|x_i - x^*\|^2 + \frac{2(f(x_i) - \min_{\mathcal{X}} f)}{\|\zeta_i\|^2} \cdot \langle \zeta_i, x^* - x_i\rangle + \frac{(f(x_i) - f(x^*))^2}{\|\zeta_i\|^2} \\
&\leq \|x_i - x^*\|^2 + \frac{2(f(x_i) - \min f)}{\|\zeta_i\|^2}\left(f(x^*) - f(x_i) + \frac{\rho}{2}\|x_i - x^*\|^2\right) \\
&\qquad + \frac{(f(x_i) - f(x^*))^2}{\|\zeta_i\|^2} \qquad \text{(A.2)} \\
&= \|x_i - x^*\|^2 + \frac{f(x_i) - \min f}{\|\zeta_i\|^2}\left(\rho\|x_i - x^*\|^2 - (f(x_i) - f(x^*))\right) \\
&\leq \|x_i - x^*\|^2 + \frac{f(x_i) - \min f}{\|\zeta_i\|^2}\left(\rho\|x_i - x^*\|^2 - \mu\|x_i - x^*\|\right) \qquad \text{(A.3)} \\
&= \|x_i - x^*\|^2 + \frac{\rho(f(x_i) - \min f)}{\|\zeta_i\|^2}\left(\|x_i - x^*\| - \frac{\mu}{\rho}\right)\|x_i - x^*\| \\
&\leq \|x_i - x^*\|^2 - \frac{\mu(f(x_i) - \min f)}{2\|\zeta_i\|^2} \cdot \|x_i - x^*\| \qquad \text{(A.4)} \\
&\leq \left(1 - \frac{\mu^2}{2\|\zeta_i\|^2}\right)\|x_i - x^*\|^2. \qquad \text{(A.5)}
\end{aligned}$$

Here, the estimate (A.1) follows from the fact that the projection $\mathrm{proj}_Q(\cdot)$ is nonexpansive, (A.2) uses local weak convexity, (A.4) follow from the estimate $\mathrm{dist}(x_i, \mathcal{X}^*) \leq \frac{\mu}{2\rho}$, while (A.3) and (A.5) use local sharpness. We therefore deduce

$$\mathrm{dist}^2(x_{i+1}; \mathcal{X}^*) \leq \|x_{i+1} - x^*\|^2 \leq \left(1 - \frac{\mu^2}{2L^2}\right)\mathrm{dist}^2(x_i, \mathcal{X}^*). \qquad \text{(A.6)}$$

Thus (5.4) holds for all indices up to $k-1$. We next show that k is infinite. To this end, observe

$$\begin{aligned}
\|x_k - x_0\| \le \sum_{i=0}^{k-1} \|x_{i+1} - x_i\| &= \sum_{i=0}^{k-1} \left\| \mathrm{proj}_{\mathcal{X}} \left(x_i - \tfrac{f(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2} \zeta_i \right) - \mathrm{proj}_{\mathcal{X}}(x_i) \right\| \\
&\le \sum_{i=0}^{k-1} \frac{f(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|} \\
&\le \sum_{i=0}^{k-1} \left\langle \tfrac{\zeta_i}{\|\zeta_i\|}, x_i - \mathrm{proj}_{\mathcal{X}^*}(x_i) \right\rangle + \frac{\rho}{2\|\zeta_i\|} \|x_i - \mathrm{proj}_{\mathcal{X}^*}(x_i)\|^2 \\
&\le \sum_{i=0}^{k-1} \mathrm{dist}(x_i, \mathcal{X}^*) + \frac{2\rho}{3\mu} \mathrm{dist}^2(x_i, \mathcal{X}^*) \qquad \text{(A.7)} \\
&\le \frac{4}{3} \cdot \sum_{i=0}^{k-1} \mathrm{dist}(x_i, \mathcal{X}^*) \qquad \text{(A.8)} \\
&\le \frac{4}{3} \cdot \mathrm{dist}(x_0, \mathcal{X}^*) \cdot \sum_{i=0}^{k-1} \left(1 - \frac{\mu^2}{2L^2} \right)^{\frac{i}{2}} \qquad \text{(A.9)} \\
&\le \frac{16L^2}{3\mu^2} \cdot \mathrm{dist}(x_0, \mathcal{X}^*) \le \frac{\epsilon}{4},
\end{aligned}$$

where (A.7) follows by Lemma A.1 with $\gamma = 1/2$, the bound in (A.8) follows by (A.6) and the assumption on $\mathrm{dist}(x_0, \mathcal{X}^*)$, finally (A.9) holds thanks to (A.6). Thus applying the triangle inequality we get the contradiction $\|x_k - \bar{x}\| \le \epsilon/2$. Consequently all the iterates x_k for $k = 0, 1, \ldots, \infty$ lie in $B_{\epsilon/2}(\bar{x})$ and satisfy (5.4).

Finally, let x_∞ be any limit point of the sequence $\{x_i\}$. We then successively compute

$$\begin{aligned}
\|x_k - x_\infty\| \le \sum_{i=k}^{\infty} \|x_{i+1} - x_i\| &\le \sum_{i=k}^{\infty} \frac{f(x_i) - \min f}{\|\zeta_i\|} \\
&\le \frac{4L}{3\mu} \cdot \sum_{i=k}^{\infty} \mathrm{dist}(x_i, \mathcal{X}^*) \\
&\le \frac{4L}{3\mu} \cdot \mathrm{dist}(x_0, \mathcal{X}^*) \cdot \sum_{i=k}^{\infty} \left(1 - \frac{\mu^2}{2L^2} \right)^{\frac{i}{2}} \\
&\le \frac{16L^3}{3\mu^3} \cdot \mathrm{dist}(x_0, \mathcal{X}^*) \cdot \left(1 - \frac{\mu^2}{2L^2} \right)^{\frac{k}{2}}.
\end{aligned}$$

This completes the proof.

A.2 Proof of Theorem 5.7

Fix an arbitrary index k and observe

$$\|x_{k+1} - x_k\| = \left\| \mathrm{proj}_Q(x_k) - \mathrm{proj}_Q \left(x_k - \alpha_k \frac{\xi_k}{\|\xi_k\|} \right) \right\| \le \alpha_k = \lambda \cdot q^k.$$

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 [23, Theorem 5.1].

A.3 Proof of Theorem 5.8

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

$$\left(f_{x_i}(x_{i+1}) + \frac{\beta}{2}\|x_{i+1} - x_i\|^2\right) + \frac{\beta}{2}\|x_{i+1} - x\|^2 \leq f_{x_i}(x) + \frac{\beta}{2}\|x - x_i\|^2. \tag{A.10}$$

Setting $x = x_i$ and appealing to approximation accuracy, we obtain the descent guarantee

$$\|x_{i+1} - x_i\|^2 \leq \frac{2}{\beta}(f(x_i) - f(x_{i+1})). \tag{A.11}$$

In particular, the function values are decreasing along the iterate sequence. Next choosing any $x^* \in \operatorname{proj}_{\mathcal{X}^*}(x_i)$ and setting $x = x^*$ in (A.10) 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

$$f(x_{i+1}) \leq f(x^*) + \beta\|x^* - x_i\|^2. \tag{A.12}$$

Using sharpness we deduce the contraction guarantee

$$\begin{aligned} f(x_{i+1}) - f(x^*) &\leq \beta \cdot \operatorname{dist}^2(x_i, \mathcal{X}^*) \\ &\leq \frac{\beta}{\mu^2}(f(x_i) - f(x^*))^2 \\ &\leq \frac{\beta(f(x_i) - f(x^*))}{\mu^2} \cdot (f(x_i) - f(x^*)) \leq \frac{1}{2} \cdot (f(x_i) - f(x^*)), \end{aligned} \tag{A.13}$$

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

$$\begin{aligned}
\|x_k - x_0\| \leq \sum_{i=0}^{k-1} \|x_{i+1} - x_i\| &\leq \sqrt{\frac{2}{\beta}} \cdot \sum_{i=0}^{k-1} \sqrt{f(x_i) - f(x_{i+1})} && \text{(A.14)} \\
&\leq \sqrt{\frac{2}{\beta}} \cdot \sum_{i=0}^{k-1} \sqrt{f(x_i) - f(x^*)} \\
&\leq \sqrt{\frac{2}{\beta}} \cdot \sqrt{f(x_0) - f(x^*)} \cdot \sum_{i=0}^{k-1} \left(\frac{1}{2}\right)^{\frac{i}{2}} && \text{(A.15)} \\
&\leq \frac{1}{\sqrt{2}-1} \sqrt{\frac{f(x_0) - f(x^*)}{\beta}} \leq \epsilon/2,
\end{aligned}$$

where (A.14) follows from (A.11) and (A.15) follows from (A.13). 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 (A.12) 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

$$\begin{aligned}
\|x_k - x_\infty\| \leq \sum_{i=k}^{\infty} \|x_{i+1} - x_i\| &\leq \sqrt{\frac{2}{\beta}} \cdot \sum_{i=k}^{\infty} \sqrt{f(x_i) - f(x_{i+1})} \\
&\leq \sqrt{\frac{2}{\beta}} \cdot \sum_{i=k}^{\infty} \sqrt{f(x_i) - \min_{\mathcal{X}} f} \\
&\leq \frac{\mu\sqrt{2}}{\beta} \cdot \sum_{i=k}^{\infty} \left(\frac{\beta}{\mu^2}(f(x_0) - \min f)\right)^{2^{i-1}} && \text{(A.16)} \\
&\leq \frac{\mu\sqrt{2}}{\beta} \cdot \sum_{i=k}^{\infty} \left(\frac{1}{2}\right)^{2^{i-1}} \\
&\leq \frac{\mu\sqrt{2}}{\beta} \sum_{j=0}^{\infty} \left(\frac{1}{2}\right)^{2^{k-1}+j} \leq \frac{2\sqrt{2}\mu}{\beta} \cdot \left(\frac{1}{2}\right)^{2^{k-1}},
\end{aligned}$$

where (A.16) follows from (A.13). The theorem is proved.

B Proofs in Section 6

B.1 Proof of Lemma 6.3

In order to prove that the assumption in each case, we will prove a stronger "small-ball condition" [43, 44], 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)**
$$\inf_{\substack{M\colon \text{Rank } M\le 2r\\ \|M\|_F=1}} \mathbb{P}(|\langle P, M\rangle| \ge \mu_0) \ge p_0,$$

2. **(Quadratic Sensing I)**
$$\inf_{\substack{M\in\mathcal{S}^d\colon \text{Rank } M\le 2r\\ \|M\|_F=1}} \mathbb{P}(|p^\top M p| \ge \mu_0) \ge p_0,$$

3. **(Quadratic Sensing II)**
$$\inf_{\substack{M\in\mathcal{S}^d\colon \text{Rank } M\le 2r\\ \|M\|_F=1}} \mathbb{P}\big(|p^\top M p - \tilde{p}^\top M \tilde{p}| \ge \mu_0\big) \ge p_0,$$

4. **(Bilinear Sensing)**
$$\inf_{\substack{M\colon \text{Rank } M\le 2r\\ \|M\|_F=1}} \mathbb{P}(|p^\top M q| \ge \mu_0) \ge p_0.$$

These conditions immediately imply Assumptions G-J. 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
$$\inf_{\substack{M\colon \text{Rank } M\le 2r\\ \|M\|_F=1}} \mathbb{P}(|\langle P, M\rangle| \ge \mu_0) = \mathbb{P}(|g| \ge \mu_0) \ge p_0.$$

Quadratic Sensing I. Fix a matrix M with 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 F.2. 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 F.2 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 H 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

$$p^\top Mp - \tilde{p}^\top M\tilde{p} \stackrel{d}{=} p^\top Dp - \tilde{p}^\top D\tilde{p} = \sum_{k=1}^{2r} \lambda_k \left(p_k^2 - \tilde{p}_k^2\right) \stackrel{d}{=} 2\sum_{k=1}^{2r} \lambda_k p_k \tilde{p}_k,$$

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 F.2. 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

$$\phi(t) = \sup_{u \in \mathbf{R}} \mathbb{P}(|2z\tilde{z} - u| \le t)$$

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 F.2, we get

$$\mathbb{P}\left(\left|2\sum_{k=1}^{2r} \sigma_k p_k \tilde{p}_k\right| \le \mu_0\right) \le \sup_{u \in \mathbf{R}} \mathbb{P}\left(\left|2\sum_{k=1}^{2r} \sigma_k p_k \tilde{p}_k - u\right| \le \mu_0\right) \le C\phi(\mu_0) \le \frac{1}{2}.$$

By taking the supremum of both sides of the inequality we conclude that Assumption I holds with μ_0 and $p_0 = 1/2$.

We omit the details for the bilinear case, which follow by similar arguments.

B.2 Proof of Theorem 6.4

The proofs in this section rely on the following proposition, which shows that that pointwise concentration imply uniform concentration. We defer the proof to Appendix B.3.

Proposition B.1. *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 $\mathcal{I} \subseteq \{1, \dots, m\}$ satisfying $|\mathcal{I}| < m/2$, the following hold:*

(1) *The measurements $\mathcal{A}(M)_1, \dots, \mathcal{A}(M)_m$ are i.i.d.*

(2) *RIP holds in expected value:*

$$\alpha \le \mathbb{E}|\mathcal{A}(M)_i| \le \beta(r) \qquad \text{for all } i \in \{1, \dots, m\} \tag{B.1}$$

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*

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1\right]\right| \le t \tag{B.2}$$

holds with probability at least $1 - 2\exp(-t^2 c(m, r))$.

Then, there exist universal constants $c_1, \dots, c_6 > 0$ depending only on α and K such that if $\mathcal{I} \subseteq \{1, \dots, m\}$ 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

$$c_4\|M\|_F \leq \frac{1}{m}\|\mathcal{A}(M)\|_1 \leq c_5\beta(r)\|M\|_F, \tag{B.3}$$

and

$$c_6\left(1 - \frac{2|\mathcal{I}|}{m}\right)\|M\|_F \leq \frac{1}{m}\left(\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}M\|_1\right). \tag{B.4}$$

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.

B.2.1 Part 1 of Theorem 6.4 (Matrix sensing)

Lemma B.2. *The random variable $|\langle P, M\rangle|$ is sub-gaussian with parameter $C\eta$. Consequently,*

$$\alpha \leq \mathbb{E}|\langle P, M\rangle| \lesssim \eta. \tag{B.5}$$

Moreover, there exists a universal constant $c > 0$ such that for any $t \in [0, \infty)$ the deviation bound

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1\right]\right| \leq t \tag{B.6}$$

holds with probability at least $1 - 2\exp\left(-\frac{ct^2}{\eta^2}m\right)$.

Proof. Assumption G immediately implies the lower bound in (B.5). 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 $\mathbb{E}|\langle P, M\rangle| \lesssim \eta$ [58, Proposition 2.5.2]. Thus, we have proved (B.5).

To prove the deviation bound (B.6) 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 [58, Lemma 2.6.8]. Hence, Hoeffding's inequality for sub-gaussian random variables [58, 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 B.1 with $\beta(r) \asymp \eta$ and $c(m,r) \asymp m/\eta^2$ now yields the result. □

B.2.2 Part 2 of Theorem 6.4 (Quadratic sensing I)

Lemma B.3. *The random variable $|p^\top Mp|$ is sub-exponential with parameter $\sqrt{2r}\eta^2$. Consequently,*

$$\alpha \le \mathbb{E}|p^\top Mp| \lesssim \sqrt{2r}\eta^2. \tag{B.7}$$

Moreover, there exists a universal constant $c > 0$ such that for any $t \in [0, \sqrt{2r}\eta]$ the deviation bound

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}\big[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1\big]\right| \le t \tag{B.8}$$

holds with probability at least $1 - 2\exp\left(-\frac{ct^2}{\eta^4}m/r\right)$.

Proof. Assumption H gives the lower bound in (B.7). 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

$$\begin{aligned}\|p^\top Mp\|_{\psi_1} &= \left\|p^\top\left(\sum_{k=1}^{2r}\sigma_k u_k u_k^\top\right)p\right\|_{\psi_1} = \left\|\sum_{k=1}^{2r}\sigma_k\langle p, u_k\rangle^2\right\|_{\psi_1} \\ &\le \sum_{k=1}^{2r}\sigma_k\left\|\langle p, u_k\rangle^2\right\|_{\psi_1} \le \sum_{k=1}^{2r}\sigma_k\left\|\langle p, u_k\rangle\right\|_{\psi_2}^2 = \eta^2\sum_{k=1}^{2r}\sigma_k \le \sqrt{2r}\eta^2,\end{aligned}$$

where the first inequality follows since $\|\cdot\|_{\psi_1}$ is a norm, the second one follows since $\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$ [58, Lemma 2.7.7], and the third inequality holds since $\|\sigma\|_1 \le \sqrt{2r}\|\sigma\|_2$. This bound has two consequences, first $p^\top Mp$ is a sub-exponential random variable with parameter $\sqrt{r}\eta^2$ and second $\mathbb{E}p^\top Mp \le \sqrt{2r}\eta^2$ [58, Exercise 2.7.2]. Thus, we have proved (B.7).

To prove the deviation bound (B.8) we introduce the random variables

$$Y_i = \begin{cases} p_i^\top Mp_i - \mathbb{E}p_i^\top Mp_i & \text{if } i \notin \mathcal{I}, \text{ and} \\ -\left(p_i^\top Mp_i - \mathbb{E}p_i^\top Mp_i\right) & \text{otherwise.}\end{cases}$$

Since $p^\top Mp$ is sub-exponential, we have $\|Y_i\|_{\psi_1} \lesssim \sqrt{r}\eta^2$ for all i, see [58, Exercise 2.7.10]. Hence, Bernstein inequality for sub-exponential random variables [58, Theorem 2.8.2] gives the desired upper bound on $\mathbb{P}\left(\frac{1}{m}\left|\sum_{i=1}^m Y_i\right| \ge t\right)$. □

Applying Proposition B.1 with with $\beta(r) \asymp \sqrt{r}\eta^2$ and $c(m, r) \asymp m/\eta^4 r$ now yields the result. □

B.2.3 Part 3 of Theorem 6.4 (Quadratic sensing II)

Lemma B.4. *The random variable $|p^\top Mp - \tilde{p}^\top M\tilde{p}|$ is sub-exponential with parameter $C\eta^2$. Consequently,*

$$\alpha \le \mathbb{E}|p^\top Mp - \tilde{p}^\top M\tilde{p}| \lesssim \eta^2. \tag{B.9}$$

Moreover, there exists a universal constant $c > 0$ such that for any $t \in [0, \eta^2]$ the deviation bound

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}\big[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1\big]\right| \le t \tag{B.10}$$

holds with probability at least $1 - 2\exp\left(-\frac{ct^2}{\eta^4}m\right)$.

Proof. Assumption I implies the lower bound in (B.9). 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

$$p^\top Mp - \tilde{p}^\top M\tilde{p}^\top \stackrel{d}{=} \sum_{k=1}^{2r} \sigma_k \left(\langle u_k, p\rangle^2 - \langle u_k, \tilde{p}\rangle^2 \right). \tag{B.11}$$

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

$$\|X\|_{\psi_1} \leq Q \iff \mathbb{E}\exp(\lambda X) \leq \exp(\tilde{Q}^2\lambda^2) \text{ for all } \lambda \text{ satisfying } |\lambda| \leq 1/\tilde{Q}^2 \tag{B.12}$$

where the $Q \asymp \tilde{Q}$, see [58, Proposition 2.7.1]. To prove that the random variable (B.11) 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

$$\mathbb{E}\exp(\lambda\langle u_k, p\rangle^2) \leq \exp\left(c\eta^4\lambda^2\right) \qquad \text{for all } k \in [2r] \text{ and } |\lambda| \leq 1/c\eta^4 \tag{B.13}$$

holds. Let λ be an arbitrary scalar with $|\lambda| \leq 1/c\eta^4$, then by expanding the moment generating function of (B.11) 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 (B.13) and the last relation follows since σ is unit norm. Combining this with (B.12) 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 $\mathbb{E}|p^\top Mp - \tilde{p}^\top M\tilde{p}^\top| \leq C\eta^2$ [58, Exercise 2.7.2]. Thus, we have proved (B.9).

To prove the deviation bound (B.10) 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 [58, Exercise 2.7.10]. Hence, Bernstein inequality for sub-exponential random variables [58, 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 B.1 with $\beta(r) \asymp \eta^2$ and $c(m,r) \asymp m/\eta^4$ now yields the result. □

B.2.4 Part 4 of Theorem 6.4 (Bilinear sensing)

Lemma B.5. *The random variable $|p^\top M q|$ is sub-exponential with parameter $C\eta^2$. Consequently,*

$$\alpha \leq \mathbb{E}|p^\top M q| \lesssim \eta^2. \tag{B.14}$$

Moreover, there exists a universal constant $c > 0$ such that for any $t \in [0, \eta^2]$ the deviation bound

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1\right]\right| \leq t \tag{B.15}$$

holds with probability at least $1 - 2\exp\left(-\frac{ct^2}{\eta^4}m\right)$.

Proof. As before the lower bound in (B.14) is implied by Assumption J. To prove the upper bound, we will show that $\||p^\top M q|\|_{\psi_1} \leq \eta^2$. 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

$$p^\top M q \stackrel{d}{=} \sum_{k=1}^{2r} \sigma_k \langle p, u_k\rangle\langle v_k, q\rangle. \tag{B.16}$$

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 (B.16) is sub-exponential we will again use (B.12). 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

$$\mathbb{E}\exp(\lambda\langle p, u_k\rangle\langle v_k, q\rangle) \leq \exp\left(c\eta^4\lambda^2\right) \qquad \text{for all } k \in [2r] \text{ and } |\lambda| \leq 1/c\eta^4 \tag{B.17}$$

holds. Let λ be an arbitrary scalar with $|\lambda| \leq 1/c\eta^4$, then by expanding the moment generating function of (B.16) 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 (B.17) and the last relation follows since σ is unitary. Combining this with (B.12) gives

$$\||p^\top M q|\|_{\psi_1} \lesssim \eta^2.$$

Thus, we have proved (B.14).

Once again, to show the deviation bound (B.15) 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 [58, 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 B.1 with $\beta(r) \asymp \eta^2$ and $c(m,r) \asymp m/\eta^4$ now yields the result. □

B.3 Proof of Proposition B.1

Choose $\epsilon \in (0, \sqrt{2})$ and let $\mathcal{N}$ be the $(\epsilon/\sqrt{2})$-net guaranteed by Lemma F.1. Pick some $t \in (0, K]$ so that (B.2) 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 $M \in \mathcal{N}$:

$$\frac{1}{m}\Big|\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\mathcal{I}^c}(M)\|_1 - \|\mathcal{A}_{\mathcal{I}}(M)\|_1\right]\Big| \leq t, \tag{B.18}$$

$$\frac{1}{m}\Big|\|\mathcal{A}(M)\|_1 - \mathbb{E}\left[\|\mathcal{A}(M)\|_1\right]\Big| \leq t. \tag{B.19}$$

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

$$\begin{aligned} \frac{1}{m}\|\mathcal{A}(M)\|_1 &\leq \frac{1}{m}\|\mathcal{A}(M_\star)\|_1 + \frac{1}{m}\|\mathcal{A}(M - M_\star)\|_1 \\ &\leq \frac{1}{m}\mathbb{E}\|\mathcal{A}(M_\star)\|_1 + t + \frac{1}{m}\|\mathcal{A}(M - M_\star)\|_1 \qquad \text{(B.20)} \\ &\leq \frac{1}{m}\mathbb{E}\|\mathcal{A}(M)\|_1 + t + \frac{1}{m}\left(\mathbb{E}\|\mathcal{A}(M - M_\star)\|_1 + \|\mathcal{A}(M - M_\star)\|_1\right), \qquad \text{(B.21)} \end{aligned}$$

where (B.20) follows from (B.19) and (B.21) follows from the triangle inequality. To simplify the third term in (B.21), 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

$$\begin{aligned} \frac{1}{m}\|\mathcal{A}(M - M_\star)\|_1 &\leq \frac{1}{m}\|\mathcal{A}(M_1)\|_1 + \frac{1}{m}\|\mathcal{A}(M_2)\|_1 \\ &\leq c_2(\|M_1\|_F + \|M_2\|_F) \leq \sqrt{2}c_2\|M - M_\star\|_F \leq c_2\epsilon, \end{aligned} \tag{B.22}$$

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

$$\frac{1}{m}\|\mathcal{A}(M)\|_1 \leq \frac{1}{m}\mathbb{E}\|\mathcal{A}(M)\|_1 + t + 2c_2\epsilon. \tag{B.23}$$

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

$$c_2 \leq \bar{\sigma} := \frac{2}{m}\sup_{M\in S_{2r}} \mathbb{E}\|\mathcal{A}(M)\|_1 + 2t \leq 2\beta(r) + 2t, \tag{B.24}$$

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,

$$\begin{aligned}
&\frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1 \\
&= \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M_\star) + \mathcal{A}_{\hat{\mathcal{I}}^c}(M - M_\star)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(M_\star) + \mathcal{A}_{\hat{\mathcal{I}}}(M - M_\star)\|_1 \\
&\geq \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M_\star)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(M_\star)\|_1 - \frac{1}{m}\|\mathcal{A}(M - M_\star)\|_1 && \text{(B.25)} \\
&\geq \frac{1}{m}\mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M_\star)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M_\star)\|_1\right] - t - \frac{1}{m}\|\mathcal{A}(M - M_\star)\|_1 && \text{(B.26)} \\
&\geq \frac{1}{m}\mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1\right] - t - \frac{1}{m}\left(\mathbb{E}\|\mathcal{A}(M - M_\star)\|_1 + \|\mathcal{A}(M - M_\star)\|_1\right) && \text{(B.27)} \\
&\geq \frac{1}{m}\mathbb{E}\left[|\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1\right] - t - 2\bar{\sigma}\epsilon, && \text{(B.28)}
\end{aligned}$$

where (B.25) uses the forward and reverse triangle inequalities, (B.26) follows from (B.18), the estimate (B.27) follows from the forward and reverse triangle inequalities, and (B.28) follows from (B.22) and (B.24). 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

$$\frac{1}{m}\sup_{M\in S_{2r}} \left|\|\mathcal{A}(M)\|_1 - \mathbb{E}\left[\|\mathcal{A}(M)\|_1\right]\right| \leq \frac{3t}{2}$$

and therefore using (B.1), we conclude the RIP property

$$\alpha - \frac{3t}{2} \leq \frac{1}{m}\|\mathcal{A}(M)\|_1 \lesssim \beta(r) + \frac{3t}{2}, \qquad \forall X \in S_{2r}. \tag{B.29}$$

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

$$\frac{1}{m}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(M)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(M)\|_1\right] \geq \left(1 - \frac{2|\mathcal{I}|}{m}\right)\alpha - \frac{3t}{2}. \tag{B.30}$$

Setting $t = \frac{2}{3}\min\{\alpha, \alpha(1-2|\mathcal{I}|/m)/2\} = \frac{1}{3}\alpha(1-2|\mathcal{I}|/m)$ in (B.29) and (B.30), we deduce the claimed estimates (B.3) and (B.4). Finally, let us estimate the probability of $\mathcal{E}$. Using the union bound and Lemma F.1 yields

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^c) &\leq \sum_{M \in \mathcal{N}} \mathbb{P}\big\{\text{(B.18) or (B.19) fails at } M\big\} \\
&\leq 4|\mathcal{N}| \exp\left(-t^2 c(m,r)\right) \\
&\leq 4\left(\frac{9}{\epsilon}\right)^{2(d_1+d_2+1)r} \exp\left(-t^2 c(m,r)\right) \\
&= 4\exp\left(2(d_1+d_2+1)r\ln(9/\epsilon) - t^2 c(m,r)\right)
\end{aligned}$$

where $c(m,r)$ is the function guaranteed by assumption (3).

By (B.1) we get $1/\epsilon = 4\bar{\sigma}/t \lesssim 2 + \beta(r)/(1-2|\mathcal{I}|/m)$. Then we deduce

$$\mathbb{P}(\mathcal{E}^c) \leq 4\exp\left(c_1(d_2+d_2+1)r\ln\left(c_2 + \frac{c_2\beta(r)}{1-2|\mathcal{I}|/m}\right) - \frac{\alpha^2}{9}(1-\frac{2|\mathcal{I}|}{m})^2 c(m,r)\right).$$

Hence as long as $c(m,r) \geq \frac{9c_1(d_1+d_2+1)r^2\ln\left(c_2+\frac{c_2\beta(r)}{1-2|\mathcal{I}|/m}\right)}{\alpha^2\left(1-\frac{2|\mathcal{I}|}{m}\right)^2}$, we can be sure

$$\mathbb{P}(\mathcal{E}^c) \leq 4\exp\left(-\frac{\alpha^2}{18}\left(1-\frac{2|\mathcal{I}|}{m}\right)^2 c(m,r)\right).$$

Proving the desired result. □

C Proof in Section 7

C.1 Proof of Lemma 7.4

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:

$$f(x_{k+1}) \leq f_{x_k}(x_{k+1}) + P(x_{k+1}, x_k) \leq f_{x_k}(x^*) + P(x^*, x_k) \leq f(x^*) + 2P(x^*, x_k).$$

Rearranging and using the sharpness and approximation accuracy assumptions, we deduce

$$\mu \cdot \mathrm{dist}(x_{k+1}, \mathcal{X}^*) \leq 2(a \cdot \mathrm{dist}^2(x, \mathcal{X}^*) + b \cdot \mathrm{dist}(x, \mathcal{X}^*)) = 2(b + a\mathrm{dist}(x, \mathcal{X}^*))\mathrm{dist}(x, \mathcal{X}^*).$$

The result follows.

C.2 Proof of Theorem 7.6

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

$$f(y) + \langle v, x - y\rangle = f_y(y) + \langle v, x - y\rangle \le f_y(x) \le f(x) + a\|x - y\|_F^2 + b\|x - y\|_F. \tag{C.1}$$

Consequently, given that $\operatorname{dist}(x_i, \mathcal{X}^*) \le \gamma \cdot \frac{\mu - 2b}{2a}$, we have

$$\begin{aligned}
\|x_{i+1} - x^*\|^2 &= \left\|\operatorname{proj}_{\mathcal{X}}\left(x_i - \tfrac{f(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right) - \operatorname{proj}_{\mathcal{X}}(x^*)\right\|^2 \\
&\le \left\|(x_i - x^*) - \tfrac{f(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right\|^2 \qquad \text{(C.2)} \\
&= \|x_i - x^*\|^2 + \frac{2(f(x_i) - \min_{\mathcal{X}} f)}{\|\zeta_i\|^2} \cdot \langle \zeta_i, x^* - x_i\rangle + \frac{(f(x_i) - f(x^*))^2}{\|\zeta_i\|^2} \\
&\le \|x_i - x^*\|^2 + \frac{2(f(x_i) - \min f)}{\|\zeta_i\|^2}\left(f(x^*) - f(x_i) + a\|x_i - x^*\|^2 + b\|x_i - x^*\|\right) \\
&\qquad + \frac{(f(x_i) - f(x^*))^2}{\|\zeta_i\|^2} \qquad \text{(C.3)} \\
&= \|x_i - x^*\|^2 + \frac{f(x_i) - \min f}{\|\zeta_i\|^2}\left(2a\|x_i - x^*\|^2 + 2b\|x_i - x^*\| - (f(x_i) - f(x^*))\right) \\
&\le \|x_i - x^*\|^2 + \frac{f(x_i) - \min f}{\|\zeta_i\|^2}\left(a\|x_i - x^*\|^2 - (\mu - 2b)\|x_i - x^*\|\right) \qquad \text{(C.4)} \\
&= \|x_i - x^*\|^2 + \frac{2a(f(x_i) - \min f)}{\|\zeta_i\|^2}\left(\|x_i - x^*\| - \frac{\mu - 2b}{2a}\right)\|x_i - x^*\| \\
&\le \|x_i - x^*\|^2 - \frac{(1 - \gamma)(\mu - 2b)(f(x_i) - \min f)}{\|\zeta_i\|^2} \cdot \|x_i - x^*\| \qquad \text{(C.5)} \\
&\le \left(1 - \frac{(1 - \gamma)\mu(\mu - 2b)}{\|\zeta_i\|^2}\right)\|x_i - x^*\|^2. \qquad \text{(C.6)}
\end{aligned}$$

Here, the estimate (C.2) follows from the fact that the projection $\operatorname{proj}_{\mathcal{X}}(\cdot)$ is nonexpansive, (C.3) uses the bound in (C.1), (C.5) follow from the estimate $\operatorname{dist}(x_i, \mathcal{X}^*) \le \gamma \cdot \frac{\mu - 2b}{2a}$, while (C.4) and (C.6) use local sharpness. The result then follows by the upper bound $\|\zeta_i\| \le L$.

D Proofs in Section 8

D.1 Proof of Lemma 8.1

The inequality can be established using an argument similar to that for bounding the T_3 term in [19, 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

$$\begin{aligned}
|\langle S - S_\sharp, XX^\top - X_\sharp X_\sharp^\top\rangle| &= |2\langle \Delta_S, \Delta_X X_\sharp^\top\rangle + \langle \Delta_S, \Delta_X \Delta_X^\top\rangle| \\
&\le \left(2\|X_\sharp^\top \Delta_S\|_F + \|\Delta_X^\top \Delta_S\|_F\right) \cdot \|\Delta_X\|_F.
\end{aligned}$$

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 $v := Se_j$, $v^* := S_\sharp e_j$, and $T := \text{support}(v^*)$. 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

$$\begin{aligned}
\|X_\sharp^\top \Delta_S\|_F = \sqrt{\sum_{j \in [d]} \|X_\sharp^\top \Delta_S e_j\|_2^2} &\leq \sqrt{\sum_{j \in [d]} \|X_\sharp\|_{2,\infty}^2 \|\Delta_S e_j\|_1^2} \\
&\leq \|X_\sharp\|_{2,\infty} \sqrt{\sum_{j \in [d]} 4k\|\Delta_S e_j\|_2^2} \leq 2\sqrt{\frac{\nu r k}{d}}\|\Delta_S\|_F.
\end{aligned}$$

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.

D.2 Proof of Theorem 8.5

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

$$|\mathrm{sign}(x-\bar{x})^\top \bar{x}| \geq a\|\bar{x}\|_1.$$

Then set $\bar{V} = \mathrm{sign}(\mathrm{sign}(x-\bar{x})^\top \bar{x}) \cdot \mathrm{sign}(x-\bar{x})\mathrm{sign}(x-\bar{x})^\top$, to get

$$\begin{aligned}
\|\bar{x}(x-\bar{x})^\top + (x-\bar{x})\bar{x}^\top\|_1 &\geq 2\mathrm{Tr}(\bar{x}(x-\bar{x})^\top \bar{V}) \\
&= 2\mathrm{sign}(\mathrm{sign}(x-\bar{x})^\top \bar{x}) \cdot \mathrm{Tr}((x-\bar{x})^\top \mathrm{sign}(x-\bar{x})\mathrm{sign}(x-\bar{x})^\top \bar{x}) \\
&= 2\|x-\bar{x}\|_1 \mathrm{sign}(\mathrm{sign}(x-\bar{x})^\top \bar{x}) \cdot \mathrm{sign}(x-\bar{x})^\top \bar{x} \\
&\geq 2a\|\bar{x}\|_1\|x-\bar{x}\|_1
\end{aligned}$$

Case 3: Suppose that

$$|(x-\bar{x})^\top \mathrm{sign}(\bar{x})| \leq a\|x-\bar{x}\|_1 \qquad \text{and} \qquad |\mathrm{sign}(x-\bar{x})^\top \bar{x}| \leq a\|\bar{x}\|_1$$

Define $\bar{V} = \frac{1}{2}(\mathrm{sign}(\bar{x}(x-\bar{x})^\top) + \mathrm{sign}((x-\bar{x})\bar{x}^\top))$. Observe that

$$\begin{aligned}\mathrm{Tr}(\bar{x}(x-\bar{x})^\top \mathrm{sign}(\bar{x}(x-\bar{x})^\top)) &= (x-\bar{x})^\top \mathrm{sign}(\bar{x})\mathrm{sign}(x-\bar{x})^\top \bar{x} \\ &\geq -a^2\|\bar{x}\|_1\|x-\bar{x}\|_1\end{aligned}$$

and

$$\begin{aligned}\mathrm{Tr}(\bar{x}(x-\bar{x})^\top \mathrm{sign}((x-\bar{x})\bar{x}^\top)) &= \mathrm{Tr}(\bar{x}(x-\bar{x})^\top \mathrm{sign}(x-\bar{x})\mathrm{sign}(\bar{x}^\top)) \\ &= \|\bar{x}\|_1\|x-\bar{x}\|_1.\end{aligned}$$

Putting these two bounds together, we find that

$$\|\bar{x}(x-\bar{x})^\top + (x-\bar{x})\bar{x}^\top\|_1 \geq 2\mathrm{Tr}(\bar{x}(x-\bar{x})^\top \bar{V}) = (1-a^2)\|\bar{x}\|_1\|x-\bar{x}\|_1.$$

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.

D.3 Proof of Lemma 8.7

We start by stating a claim we will use to prove the lemma. Let us introduce some notation. Consider the set

$$S = \left\{(\Delta_+, \Delta_-) \in \mathbf{R}^{d\times r} \times \mathbf{R}^{d\times r} \mid \|\Delta_+\|_{2,\infty} \leq (1+C)\sqrt{\frac{\nu r}{d}}\|X_\natural\|_{op}, \|\Delta_-\|_{2,1} \neq 0\right\}.$$

Define the random variable

$$\begin{aligned}Z = \sup_{(\Delta_+,\Delta_-)\in S} \Bigg| &\frac{1}{\|\Delta_-\|_{2,1}} \sum_{i,j=1}^d \delta_{ij}|\langle \Delta_{-,i}, \Delta_{+,j}\rangle + \langle \Delta_{+,i}, \Delta_{-,j}\rangle| \\ &- \mathbb{E}\frac{1}{\|\Delta_-\|_{2,1}} \sum_{i,j=1}^d \delta_{ij}|\langle \Delta_{-,i}, \Delta_{+,j}\rangle + \langle \Delta_{+,i}, \Delta_{-,j}\rangle|\Bigg|.\end{aligned}$$

Claim 1. *There exist constants $c_2, c_3 > 0$ such that with probability at least $1 - \exp(-c_2 \log d)$*

$$Z \leq c_3 C\sqrt{\tau\nu r \log d}\,\|X_\natural\|_{op}.$$

Before proving this claim, let us show how it implies the theorem. Let

$$R \in \operatorname*{argmin}_{\hat{R}^\top \hat{R} = I} \|X - X_\sharp \hat{R}\|_{2,1}.$$

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

$$\begin{aligned}\sum_{i,j=1}^d \delta_{ij} |\langle X_i, X_j\rangle - \langle (X_\sharp)_i, (X_\sharp)_j\rangle| &\le \tau \|XX^\top - X_\sharp X_\sharp^\top\|_1 + c_3 C\sqrt{\tau\nu r \log d}\|X_\sharp\|_{op}\|X - X_\sharp R\|_{2,1} \\ &\le \left(\tau + \frac{c_3 C\sqrt{\tau\nu r \log d}}{c}\|X_\sharp\|_{op}\right)\|XX^\top - X_\sharp X_\sharp^\top\|_1,\end{aligned}$$

where the last line follows by Conjecture 8.6. 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 [5, Lemma 11.4] to bound the expectation $\mathbb{E}Z$ as follows:

$$\begin{aligned}\mathbb{E}Z &\le 2\mathbb{E} \sup_{(\Delta_+,\Delta_-)\in S} \left| \frac{1}{\|\Delta_-\|_{2,1}} \sum_{i,j=1}^d \varepsilon_{ij}\delta_{ij} |\langle \Delta_{-,i}, \Delta_{+,j}\rangle + \langle \Delta_{+,i}, \Delta_{-,j}\rangle| \right| \\ &\le \underbrace{2\mathbb{E} \sup_{(\Delta_+,\Delta_-)\in S} \left| \frac{1}{\|\Delta_-\|_{2,1}} \sum_{i,j=1}^d \varepsilon_{ij}\delta_{ij} |\langle \Delta_{-,i}, \Delta_{+,j}\rangle| \right|}_{T_1} + \underbrace{2\mathbb{E} \sup_{(\Delta_+,\Delta_-)\in S} \left| \frac{1}{\|\Delta_-\|_{2,1}} \sum_{i,j=1}^d \varepsilon_{ij}\delta_{ij} |\langle \Delta_{+,i}, \Delta_{-,j}\rangle| \right|}_{T_2}.\end{aligned}$$

Observing that T_1 and T_2 can both be bounded by

$$\begin{aligned}\max\{T_1, T_2\} &\le 2 \sup_{(\Delta_+,\Delta_-)\in S} \frac{1}{\|\Delta_-\|_{2,1}} \|\Delta_+\Delta_-^\top\|_{2,\infty} \mathbb{E} \max_j \left| \sum_{i=1}^d \varepsilon_{ij}\delta_{ij} \right| \\ &\le 2 \sup_{(\Delta_+,\Delta_-)\in S} \|\Delta_+\|_{2,\infty} \mathbb{E} \max_j \left| \sum_{i=1}^d \varepsilon_{ij}\delta_{ij} \right| \\ &\le 2(1+C)\sqrt{\frac{\nu r}{d}}\|X_\sharp\|_{op} \mathbb{E} \max_j \left| \sum_{i=1}^d \varepsilon_{ij}\delta_{ij} \right| \\ &\lesssim C\sqrt{\frac{\nu r}{d}}\|X_\sharp\|_{op}(\sqrt{\tau d \log d} + \log d),\end{aligned}$$

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 F.3. 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

$$\begin{aligned} b = \sup_{ij,s\in\mathcal{S}}|Z_{ij,s}| &\le \sup_{ij,(\Delta_+,\Delta_-)\in S}\left|\frac{(\delta_{ij}-\tau)}{\|\Delta_-\|_{2,1}}(\|\Delta_{-,i}\|_F\|\Delta_{+,j}\|_F + \|\Delta_{+,i}\|_F\|\Delta_{-,j}\|_F)\right| \\ &\le (1+C)\sqrt{\frac{\nu r}{d}}\|X_\sharp\|_{op}\sup_{ij,(\Delta_+,\Delta_-)\in S}\left|\frac{1}{\|\Delta_-\|_{2,1}}(\|\Delta_{-,i}\|_F + \|\Delta_{-,j}\|_F)\right| \le 2C\sqrt{\frac{\nu r}{d}}\|X_\sharp\|_{op} \end{aligned}$$

and

$$\begin{aligned} \sigma^2 &= \sup_{(\Delta_+,\Delta_-)\in S}\mathbb{E}\frac{1}{\|\Delta_-\|_{2,1}^2}\sum_{ij=1}^d(\delta_{ij}-\tau)^2|\langle\Delta_{-,i},\Delta_{+,j}\rangle + \langle\Delta_{+,i},\Delta_{-,j}\rangle|^2 \\ &\le \tau\sup_{(\Delta_+,\Delta_-)\in S}\frac{1}{\|\Delta_-\|_{2,1}^2}\sum_{ij=1}^d(\|\Delta_{-,i}\|_F\|\Delta_{+,j}\|_F + \|\Delta_{+,i}\|_F\|\Delta_{-,j}\|_F)^2 \\ &\le \tau\sup_{(\Delta_+,\Delta_-)\in S}\frac{4}{\|\Delta_-\|_{2,1}^2}\sum_{ij=1}^d\|\Delta_{-,i}\|_F^2\|\Delta_{+,j}\|_F^2 \\ &\le \tau\frac{4(1+C)^2\nu r}{d}\|X_\sharp\|_{op}^2\sup_{(\Delta_+,\Delta_-)\in S}\frac{2}{\|\Delta_-\|_{2,1}^2}\sum_{ij=1}^d\|\Delta_{-,i}\|_F^2 \\ &\le \tau\frac{4(1+C)^2\nu r}{d}\|X_\sharp\|_{op}^2\sup_{(\Delta_+,\Delta_-)\in S}\frac{2d\|\Delta_-\|_F^2}{\|\Delta_-\|_{2,1}^2} \\ &\le 16\tau C^2\nu r\|X_\sharp\|_{op}^2. \end{aligned}$$

Therefore, due to Theorem F.3 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

$$
\begin{aligned}
Z &\le \mathbb{E}Z + \sqrt{8\left(2b\mathbb{E}Z + \sigma^2\right)t} + 8bt \\
&\le c_1 C\sqrt{\frac{\nu r}{d}}\|X_\sharp\|_{op}(\sqrt{\tau d \log d} + \log d) \\
&\quad + \sqrt{8c_2\left(\frac{c_1^2 C^2 \nu r}{d}\|X_\sharp\|_{op}^2(\sqrt{\tau d \log d} + \log d) + 16\tau C^2 \nu r \|X_\sharp\|_{op}^2\right)\log d} + 16c_2 C\sqrt{\frac{\nu r}{d}}\|X_\sharp\|_{op}\log(d) \\
&\le C\sqrt{\nu r \log d}\|X_\sharp\|_{op}\left(c_1\sqrt{\tau} + c_1\sqrt{\frac{\log d}{d}} + \sqrt{8c_2}\sqrt{c_1^2\sqrt{\frac{\tau\log d}{d}} + c_1^2\frac{\log d}{d} + 16\tau} + 16c_2\sqrt{\frac{\log d}{d}}\right) \\
&\le c_3 C\sqrt{\tau\nu r \log d}\|X_\sharp\|_{op}.
\end{aligned}
$$

where the last line follows since by assumption $\log d/d \lesssim \tau$. □

E Proofs in Section 9

E.1 Proof of Lemma 9.1

The proof follows the same strategy as [24, 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 9.3, that

$$f(y) \ge \tilde{f}(x) + \langle \zeta, y - x\rangle - \frac{\rho}{2}\|x - y\|_2^2 - 3\varepsilon.$$

Therefore, the function

$$g(y) := f(y) - \langle \zeta, y - x\rangle + \frac{\rho}{2}\|x - y\|_2^2 + 3\varepsilon$$

satisfies

$$g(x) - \inf g \le f(x) - \tilde{f}(x) + 3\varepsilon \le 4\varepsilon.$$

Now, for some $\gamma > 0$ to be determined momentarily, define

$$\hat{x} = \operatorname{argmin}\left\{g(x) + \frac{\varepsilon}{\gamma^2}\|x - y\|_2^2\right\}.$$

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,

$$\operatorname{dist}(\zeta, \partial f(\hat{x})) \le \left(\frac{2\varepsilon}{\gamma^2} + \rho\right)\|x - \hat{x}\|_2.$$

Now we estimate $\|x - \hat{x}\|_2$. Indeed, from the definition of $\hat{x}$ we have

$$\frac{\varepsilon}{\gamma^2}\|\hat{x} - x\|^2 \le g(x) - g(\hat{x}) \le g(x) - \inf g \le 4\varepsilon.$$

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.

E.2 Proof of Theorem 9.4

Let $i \geq 0$, suppose $x_i \in \widetilde{\mathcal{T}}_1$, and let $x^* \in \operatorname{proj}_{\mathcal{X}^*}(x_i)$. Notice that Lemma 9.2 implies $\tilde{f}(x_i) - \min_{\mathcal{X}} f > 0$. We successively compute

$$\begin{aligned}
\|x_{i+1} - x^*\|^2 &= \left\|\operatorname{proj}_{\mathcal{X}}\left(x_i - \tfrac{\tilde{f}(x_i)-\min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right) - \operatorname{proj}_{\mathcal{X}}(x^*)\right\|^2 \\
&\leq \left\|(x_i - x^*) - \tfrac{\tilde{f}(x_i)-\min_{\mathcal{X}} f}{\|\zeta_i\|^2}\zeta_i\right\|^2 && \text{(E.1)} \\
&= \|x_i - x^*\|^2 + \frac{2(\tilde{f}(x_i) - \min_{\mathcal{X}} f)}{\|\zeta_i\|^2} \cdot \langle \zeta_i, x^* - x_i\rangle + \frac{(\tilde{f}(x_i) - \min_{\mathcal{X}} f)^2}{\|\zeta_i\|^2} \\
&\leq \|x_i - x^*\|^2 + \frac{2(\tilde{f}(x_i) - \min_{\mathcal{X}} f)}{\|\zeta_i\|^2}\left(\min_{\mathcal{X}} f - \tilde{f}(x_i) + \frac{\rho}{2}\|x_i - x^*\|^2 + 3\varepsilon\right) \\
&\qquad + \frac{(\tilde{f}(x_i) - \min_{\mathcal{X}} f)^2}{\|\zeta_i\|^2} && \text{(E.2)} \\
&= \|x_i - x^*\|^2 + \frac{\tilde{f}(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2}\left(\rho\|x_i - x^*\|^2 - (\tilde{f}(x_i) - \min_{\mathcal{X}} f) + 6\varepsilon\right) \\
&\leq \|x_i - x^*\|^2 + \frac{\tilde{f}(x_i) - \min_{\mathcal{X}} f}{\|\zeta_i\|^2}\left(\rho\|x_i - x^*\|^2 - \mu\|x_i - x^*\| + 7\varepsilon\right) && \text{(E.3)} \\
&\leq \|x_i - x^*\|^2 + \frac{\rho(\tilde{f}(x_i) - \min_{\mathcal{X}} f)}{\|\zeta_i\|^2}\left(\|x_i - x^*\| - \frac{\mu}{2\rho}\right)\|x_i - x^*\| && \text{(E.4)} \\
&\leq \|x_i - x^*\|^2 - \frac{\mu(\tilde{f}(x_i) - \min_{\mathcal{X}} f)}{4\|\zeta_i\|^2} \cdot \|x_i - x^*\| && \text{(E.5)} \\
&\leq \|x_i - x^*\|^2 - \frac{\mu(\mu\|x_i - x^*\| - \varepsilon)}{4\|\zeta_i\|^2} \cdot \|x_i - x^*\| && \text{(E.6)} \\
&\leq \left(1 - \frac{13\mu^2}{56\|\zeta_i\|^2}\right)\|x_i - x^*\|^2.
\end{aligned}$$

Here, the estimate (E.1) follows from the fact that the projection $\mathrm{proj}_Q(\cdot)$ is nonexpansive, (E.2) uses Lemma 9.3, the estimate (E.4) follows from the assumption $\epsilon < \frac{\mu}{14}\|x_k - x^*\|$, the estimate (E.5) follows from the estimate $\|x_i - x^*\| \leq \frac{\mu}{4\rho}$, while (E.3) and (E.6) use Lemma 9.2. 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.

E.3 Proof of Theorem 9.6

Let $i \geq 0$, suppose $x_i \in \mathcal{T}_\gamma$, and let $x^* \in \mathrm{proj}_{\mathcal{X}^*}(x_i)$. Then

$$\begin{aligned}
\mu \mathrm{dist}(x_{i+1}, \mathcal{X}^*) \leq f(x_{i+1}) - \inf_{\mathcal{X}} f &\leq f_x(x_{i+1}) - \inf_{\mathcal{X}} f + \frac{\rho}{2}\|x_{i+1} - x_i\|^2 \\
&\leq \tilde{f}_x(x_{i+1}) - \inf_{\mathcal{X}} f + \frac{\rho}{2}\|x_{i+1} - x_i\|^2 + \varepsilon \\
&\leq \tilde{f}_x(x^*) - \inf_{\mathcal{X}} f + \frac{\beta}{2}\|x_i - x^*\|^2 + \varepsilon \\
&\leq f_x(x^*) - \inf_{\mathcal{X}} f + \frac{\beta}{2}\|x_i - x^*\|^2 + 2\varepsilon \\
&\leq f(x^*) - \inf_{\mathcal{X}} f + \beta\|x_i - x^*\|^2 + 2\varepsilon \\
&= \beta \mathrm{dist}^2(x_i, \mathcal{X}^*) + 2\varepsilon.
\end{aligned}$$

Rearranging yields the result.

F Auxiliary lemmas

Lemma F.1 (Lemma 3.1 in [13]). *Let $S_r := \left\{X \in \mathbf{R}^{d_1 \times d_2} \mid \mathit{Rank}\ (X) \leq r, \|X\|_F = 1\right\}$. 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 F.2 (Corollary 1.4 in [54]). *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 F.3 (Talagrand's Functional Bernstein for non-identically distributed variables [36, 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 $|Z_{i,s}| \leq b$. 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

$$P\left(Z - \mathbb{E}Z \geq \sqrt{8\left(2b\mathbb{E}Z + \sigma^2\right)t} + 8bt\right) \leq e^{-t}.$$