Preconditioned Gradient Descent for Over-Parameterized Nonconvex Matrix Factorization

Gavin Zhang
University of Illinois at Urbana–Champaign
jialun2@illinois.edu

Salar Fattahi
University of Michigan
fattahi@umich.edu

Richard Y. Zhang
University of Illinois at Urbana–Champaign
ryz@illinois.edu

Abstract

In practical instances of nonconvex matrix factorization, the rank of the true solution $r^\star$ is often unknown, so the rank r of the model can be overspecified as $r > r^\star$. This over-parameterized regime of matrix factorization significantly slows down the convergence of local search algorithms, from a linear rate with $r = r^\star$ to a sublinear rate when $r > r^\star$. We propose an inexpensive preconditioner for the matrix sensing variant of nonconvex matrix factorization that restores the convergence rate of gradient descent back to linear, even in the over-parameterized case, while also making it agnostic to possible ill-conditioning in the ground truth. Classical gradient descent in a neighborhood of the solution slows down due to the need for the model matrix factor to become singular. Our key result is that this singularity can be corrected by ℓ_2 regularization with a specific range of values for the damping parameter. In fact, a good damping parameter can be inexpensively estimated from the current iterate. The resulting algorithm, which we call preconditioned gradient descent or PrecGD, is stable under noise, and converges linearly to an information theoretically optimal error bound. Our numerical experiments find that PrecGD works equally well in restoring the linear convergence of other variants of nonconvex matrix factorization in the over-parameterized regime.

1 Introduction

Numerous problems in machine learning can be reduced to the *matrix factorization* problem of recovering a low-rank positive semidefinite matrix $M^\star \succeq 0$, given a small number of potentially noisy observations [1–7]. In every case, the most common approach is to formulate an $n \times n$ candidate matrix $M = XX^T$ in factored form, and to minimize a *nonconvex* empirical loss $f(X)$ over its $n \times r$ low-rank factor X. But in most real applications of nonconvex matrix factorization, the rank of the ground truth $r^\star = \mathrm{rank}(M^\star)$ is unknown. It is reasonable to choose the rank r of the model XX^T conservatively, setting it to be potentially larger than $r^\star$, given that the ground truth can be exactly recovered so long as $r \geq r^\star$. In practice, this will often lead to an *over-parameterized* regime, in which $r > r^\star$, and we have specified more degrees of freedom in our model XX^T than exists in the underlying ground truth $M^\star$.

Zhuo et al. [8] recently pointed out that nonconvex matrix factorization becomes substantially less efficient in the over-parameterized regime. For the prototypical instance of matrix factorization known as *matrix sensing* (see Section 3 below for details) it is well-known that, if $r = r^\star$, then (classic) gradient descent or GD

$$X_{k+1} = X_k - \alpha \nabla f(X_k) \tag{GD}$$

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

converges at a linear rate, to an ϵ-accurate iterate in $O(\kappa \log(1/\epsilon))$ iterations, where $\kappa = \lambda_1(M^\star)/\lambda_{r^*}(M^\star)$ is the condition number of the ground truth [9, 10]. But in the case that $r > r^\star$, Zhuo et al. [8] proved that gradient descent slows down to a *sublinear* convergence rate, now requiring $\mathrm{poly}(1/\epsilon)$ iterations to yield a comparable ϵ-accurate solution. This is a dramatic, exponential slow-down: whereas 10 digits of accuracy can be expected in a just few hundred iterations when $r = r^\star$, tens of thousands of iterations might produce just 1-2 accurate digits once $r > r^\star$. The slow-down occurs even if r is just off by one, as in $r = r^\star + 1$.

It is helpful to understand this pheonomenon by viewing over-parameterization as a special, extreme case of ill-conditioning, where the condition number of the ground truth, κ, is taken to infinity. In this limit, the classic linear rate $O(\kappa \log(1/\epsilon))$ breaks down, and in reality, the convergence rate deterioriates to sublinear.

In this paper, we present an inexpensive *preconditioner* for gradient descent. The resulting algorithm, which we call PrecGD, corrects for both ill-conditioning and over-parameterization at the same time, without viewing them as distinct concepts. We prove, for the matrix sensing variant of nonconvex matrix factorization, that the preconditioner restores the convergence rate of gradient descent back to linear, even in the over-parameterized case, while also making it agnostic to possible ill-conditioning in the ground truth. Moreover, PrecGD maintains a similar per-iteration cost to regular gradient descent, is stable under noise, and converges linearly to an information theoretically optimal error bound.

We also perform numerical experiments on other variants of nonconvex matrix factorization, with different choices of the empirical loss function f. In particular, we consider different ℓ_p norms with $1 \le p < 2$, in order to gauge the effectiveness of PrecGD for increasingly nonsmooth loss functions. Our numerical experiments find that, if regular gradient descent is capable of converging quickly when the rank is known $r = r^\star$, then PrecGD restores this rapid converging behavior when $r > r^\star$. PrecGD is able to overcome ill-conditioning in the ground truth, and converge reliably without exhibiting sporadic behavior.

2 Proposed Algorithm: Preconditioned Gradient Descent

Our preconditioner is inspired by a recent work of Tong et al. [11] on matrix sensing with an ill-conditioned ground truth $M^\star$. Over-parameterization can be viewed as the limit of this regime, in which $\lambda_r(M^\star)$, the r-th largest eigenvalue of $M^\star$, is allowed to approach all the way to zero. For finite but potentially very small values of $\lambda_r(M^\star) > 0$, Tong et al. [11] suggests the following iterations, which they named *scaled* gradient descent or ScaledGD:

$$X_{k+1} = X_k - \alpha \nabla f(X_k)(X_k^T X_k)^{-1}. \tag{ScaledGD}$$

They prove that the scaling allows the iteration to make a large, constant amount of progress at every iteration, independent of the value of $\lambda_r(M^\star) > 0$. However, applying this same scheme to the over-parameterized case with $\lambda_r(M^\star) = 0$ results in an inconsistent, sporadic behavior.

The issues encountered by both regular GD and ScaledGD with over-parameterization $r > r^\star$ can be explained by the fact that our iterate X_k must necessarily become *singular* as our rank-r model $X_k X_k^T$ converges towards the rank-$r^\star$ ground truth $M^\star$. For GD, this singularity causes the per-iteration progress itself to decay, so that more and more iterations are required for each fixed amount of progress. ScaledGD corrects for this decay in per-iteration progress by suitably rescaling the search direction. However, the rescaling itself requires inverting a near-singular matrix, which causes algorithm to take on sporadic values.

A classical remedy to issues posed by singular matrices is ℓ_2 regularization, in which the singular matrix is made "less singular" by adding a small identity perturbation. Applying this idea to ScaledGD yields the following iterations

$$X_{k+1} = X_k - \alpha \nabla f(X_k)(X_k^T X_k + \eta_k I_r)^{-1}, \tag{PrecGD}$$

where $\eta_k \ge 0$ is the *damping* parameter specific to the k-th iteration. There are several interpretations to this scheme, but the most helpful is to view η as a parameter that allows us to interpolate between ScaledGD (with $\eta = 0$) and regular GD (in the limit $\eta \to \infty$). In this paper, we prove for matrix sensing that, if the k-th damping parameter η_k is chosen within a constant factor of the error

$$C_{\mathrm{lb}} \|X_k X_k^T - M^\star\|_F \le \eta_k \le C_{\mathrm{ub}} \|X_k X_k^T - M^\star\|_F, \quad \text{where } C_{\mathrm{lb}}, C_{\mathrm{ub}} > 0 \text{ are abs. const.} \tag{1}$$

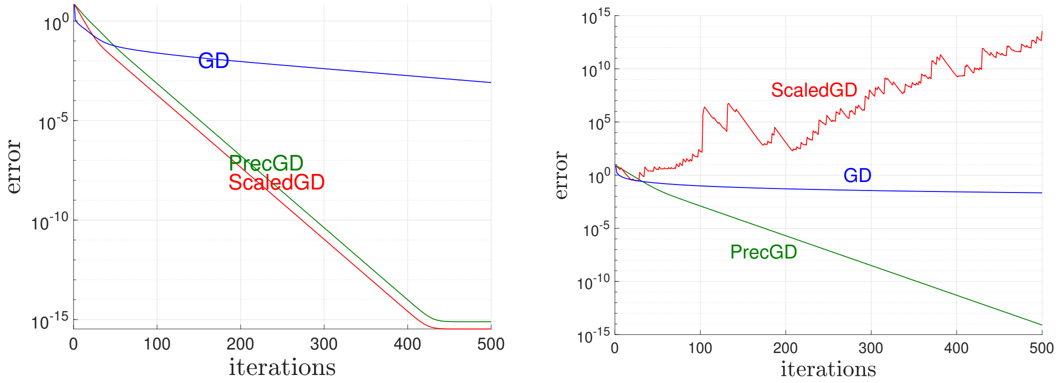

Figure 1: **PrecGD converges linearly in the overparameterized regime.** Convergence of regular gradient descent (GD), ScaledGD and PrecGD for noiseless matrix sensing (with data taken from [12, 13]) from the same initial points and using the same learning rate $\alpha = 2 \times 10^{-2}$. **(Left** $r = r^*$**)** Set $n = 4$ and $r^* = r = 2$. All three methods convergence at a linear rate, though GD converges at a slower rate due to ill-conditioning in the ground truth. **(Right** $r > r^*$**)** With $n = 4$, $r = 4$ and $r^* = 2$, over-parameterization causes gradient descent to slow down to a sublinear rate. ScaledGD also behaves sporadically. Only PrecGD converges linearly to the ground truth.

then the resulting iterations are guaranteed to converge linearly, at a rate that is independent of both over-parameterization and ill-conditioning in the ground truth $M^\star$. With noisy measurements, setting η_k to satisfy (1) will allow the iterations to converge to an error bound that is well-known to be minimax optimal up to logarithmic factors [14].

We refer to the resulting iterations (with a properly chosen η_k) as *preconditioned* gradient descent, or PrecGD for short. For matrix sensing with noiseless measurements, an optimal η_k that satisfies the condition (1) is obtained for free by setting $\eta_k = \sqrt{f(X_k)}$. In the case of noisy measurements, we show that a good choice of η_k is available based on an approximation of the noise variance.

3 Background and Related Work

Notations. We use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix and $\langle \cdot, \cdot \rangle$ is the corresponding inner product. We use $\gtrsim$ to denote an inequality that hides a constant factor. The big-O notation $\tilde{O}$ hides logarithimic factors. The gradient of the objective is denoted by $\nabla f(X) \in \mathbb{R}^{n \times r}$. The eigenvalues are assumed to be in decreasing order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$.

The symmetric, linear variant of matrix factorization known as matrix sensing aims to recover a positive semidefinite, rank-$r^\star$ ground truth matrix $M^\star$, from a small number m of possibly noisy measurements

$$y = \mathcal{A}(M^\star) + \epsilon, \qquad \text{where } \mathcal{A}(M^\star) = [\langle A_1, M^\star \rangle, \langle A_2, M^\star \rangle, \ldots, \langle A_m, M^\star \rangle]^T,$$

in which $\mathcal{A}$ is a linear measurement operator, and the length-m vector ϵ models the unknown measurement noise. A distinguishing feature of matrix sensing is that $\mathcal{A}$ is assumed to satisfy the *restricted isometry property* [14, 15]. Throughout this paper, we will always assume that $\mathcal{A}$ satisfies RIP with parameters $(2r, \delta)$.

Definition 1 (RIP)**.** The linear operator $\mathcal{A}$ satisfies RIP with parameters $(2r, \delta)$ if there exists constants $0 \leq \delta < 1$ and $m > 0$ such that, for every rank-$2r$ matrix M, we have

$$(1 - \delta)\|M\|_F^2 \leq \frac{1}{m}\|\mathcal{A}(M)\|^2 \leq (1 + \delta)\|M\|_F^2.$$

A common approach for matrix sensing is to use a simple algorithm like gradient descent to minimize the *nonconvex* loss function:

$$f(X) = \frac{1}{m}\left\|y - \mathcal{A}(XX^T)\right\| = \frac{1}{m}\left\|\mathcal{A}(M^\star - XX^T) + \epsilon\right\|^2. \tag{2}$$

Recent work has provided a theoretical explanation for the empirical success of this nonconvex approach. Two lines of work have emerged.

Local Guarantees. One line of work studies gradient descent initialized inside a neighborhood of the ground truth where $X_0 X_0^T \approx M^\star$ already holds [10, 16–19]. Such an initial point can be found using spectral initialization, see also [18, 20–23]. With exact rank $r = r^\star$, previous authors showed that gradient descent converges at a linear rate [9, 10]. In the over-parameterized regime, however, local restricted convexity no longer holds, so the linear convergence rate is lost. Zhuo et al. [8] showed that while spectral initialization continues to work under over-parameterization, gradient descent now slows down to a sublinear rate, but it still converges to a statistical error bound of $\tilde{\mathcal{O}}(\sigma^2 n r^\star / m)$, where σ denotes the noise variance. This is known to be minimax optimal up to logarithmic factors [14]. In this paper, we prove that PrecGD with a damping parameter η_k satisfying (1) also converges to an $\tilde{\mathcal{O}}(\sigma^2 n r^\star / m)$ statistical error bound.

Global Guarantees. A separate line of work [13, 24–31] established global properties of the landscapes of the nonconvex objective f in (2) and its variants and showed that local search methods can converge globally. With exact rank $r = r^\star$, Bhojanapalli et al. [24] proved that f has no spurious local minima, and that all saddles points have a strictly negative descent direction (strict saddle property [32], see also [28, 33]). In the over-parameterized regime, however, we are no longer guaranteed to recover the ground truth in polynomial time.

Other related work. Here we mention some other techniques can be use to solve matrix sensing in the over-parameterized regime. Classically, matrix factorization was solved via its convex SDP relaxation [14, 15, 34–36]. The resulting $\mathcal{O}(n^3)$ to $\mathcal{O}(n^6)$ time complexity [37] limits this technique to smaller problems, but these guarantees hold without prior knowledge on the true rank $r^\star$. First-order methods, such as ADMM [38–40] and soft-thresholding [41], can be used to solve these convex problems with a per-iteration complexity comparable to nonconvex gradient descent, but they likewise suffer from a sublinear convergence rate. Local recovery via spectral initialization was originally proposed for alternating minimization and other projection techniques [21, 23, 34, 42–45]. These also continue to work, though a drawback here is a higher per-iteration cost when compared to simple gradient methods. Finally, we mention a recent result of Li et al. [46], which showed in the over-parameterized regime that gradient descent with early termination enjoys an algorithmic regularization effect.

4 Sublinear Convergence of Gradient Descent

In order to understand how to improve gradient descent in the over-parameterized regime, we must first understand why existing methods fail. For an algorithm that moves in a search direction D with step-size α, it is a standard technique to measure the corresponding decrement in f with a Taylor-like expansion

$$f(X - \alpha D) \le f(X) - \alpha \underbrace{\langle \nabla f(X), D \rangle}_{\text{linear progress}} + \alpha^2 \underbrace{(L/2)\|D\|_F^2}_{\text{inverse step-size}} \tag{3}$$

in which L is the usual gradient Lipschitz constant (see e.g. Nocedal and Wright [47, Chapter 3]). A good search direction D is one that maximizes the linear progress $\langle \nabla f(X), D \rangle$ while also keeping the inverse step-size $(L/2)\|D\|_F^2$ sufficiently small in order to allow a reasonably large step to be taken. As we will show in this section, the main issue with gradient descent in the over-parameterized regime is the first term, namely, that the linear progress goes down to zero as the algorithm makes progress towards the solution.

Classical gradient descent uses the search direction $D = \nabla f(X)$. Here, a common technique is to bound the linear progress at each iteration by a condition known as *gradient dominance* (or the Polyak-Łojasiewicz or PL inequality), which is written as

$$\langle \nabla f(X), D \rangle = \|\nabla f(X)\|_F^2 \ge \mu (f(X) - f^\star) \quad \text{where } \mu > 0 \text{ and } f^\star = \min_X f(X). \tag{4}$$

Substituting the inequality (4) into the Taylor-like expansion (3) leads to

$$\begin{aligned} f(X - \alpha D) &\le f(X) - \alpha \|\nabla f(X)\|_F^2 + \alpha^2 (L/2) \|\nabla f(X)\|_F^2 \\ f(X - \alpha D) - f^\star &\le [1 - \mu\alpha(1 - \alpha L/2)] \cdot (f(X) - f^\star). \end{aligned} \tag{5}$$

Here, we can always pick a small enough step-size α to guarantee linear convergence:

$$Q = 1 - \mu\alpha + \mu\alpha^2 L/2 < 1 \implies f(X_k) - f^\star \leq Q^k[f(X_0) - f^\star]. \tag{6}$$

In particular, picking the optimal step-size $\alpha = 1/L$ minimizes the convergence quotient $Q = 1 - 1/(2\kappa)$, where $\kappa = L/\mu$ is the usual *condition number*. This shows that, with an optimal step-size, gradient descent needs at most $O(\kappa \log(1/\epsilon))$ iterations to find an ϵ-suboptimal X.

Matrix sensing with exact rank $r = r^\star$ is easily shown to satisfy gradient dominance (4) by manipulating existing results on (restricted) local strong convexity. In the over-parameterized case $r > r^\star$, however, local strong convexity is lost, and gradient dominance can fail to hold. Indeed, consider the following instance of matrix sensing, with true rank $r^\star = 1$, search rank $r = 2$, and $\mathcal{A}$ set to the identity

$$f(X) = \|XX^T - zz^T\|_F^2 \text{ where } X = \begin{bmatrix} 1 & 0 \\ 0 & \xi \end{bmatrix} \text{ and } z = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{7}$$

We can verify that $\|\nabla f(X)\|^2 = 4\xi^2[f(X) - f^\star]$, and this suggests that f satisfies gradient dominance (4) with a constant of $\mu \leq 2\xi^2$. But ξ is itself a variable that goes to zero as the candidate XX^T approaches to ground truth zz^T. For every fixed $\mu > 0$ in the gradient dominance condition (4), we can find a counterexample X in (7) with $\xi < \sqrt{\mu}/2$. Therefore, we must conclude that gradient dominance fails to hold, because the inequality in (4) can only hold for $\mu = 0$.

In fact, this same example also shows why classical gradient descent slows down to a sublinear rate. Applying gradient descent $X_{k+1} = X_k - \alpha \nabla f(X_k)$ with fixed step-size α to (7) yields a sequence of iterates of the same form

$$X_0 = \begin{bmatrix} 1 & 0 \\ 0 & \xi_0 \end{bmatrix}, \qquad X_{k+1} = \begin{bmatrix} 1 & 0 \\ 0 & \xi_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \xi_k - \alpha\xi_k^3 \end{bmatrix},$$

from which we can verify that $f(X_{k+1}) = (1 - \alpha\xi_k^2)^4 \cdot f(X_k)$. As each k-th $X_k X_k^T$ approaches zz^T, the element ξ_k converges towards zero, and the convergence quotient $Q = (1 - \alpha\xi_k^2)^4$ approaches 1. We see a process of diminishing returns: every improvement to f worsens the quotient Q, thereby reducing the progress achievable in the subsequent step. This is precisely the notion that characterizes sublinear convergence.

5 Linear Convergence for the Noiseless Case

To understand how it is possible make gradient descent converge linearly in the over-parameterized regime, we begin by considering gradient method under a *change of metric*. Let $\mathbf{P}$ be a real symmetric, positive definite $nr \times nr$ matrix. We define a corresponding P-inner product, P-norm, and dual P-norm on $\mathbb{R}^{n\times r}$ as follows

$$\langle X, Y\rangle_P \overset{\text{def}}{=} \text{vec}(X)^T \mathbf{P} \text{vec}(Y), \quad \|X\|_P \overset{\text{def}}{=} \sqrt{\langle X, X\rangle_P}, \quad \|X\|_{P*} \overset{\text{def}}{=} \sqrt{\text{vec}(X)^T \mathbf{P}^{-1} \text{vec}(X)},$$

where $\text{vec} : \mathbb{R}^{n\times r} \to \mathbb{R}^{nr}$ is the usual column-stacking operation. Consider descending in the direction D satisfying $\text{vec}(D) = \mathbf{P}^{-1}\text{vec}(\nabla f(X))$; the resulting decrement in f can be quantified by a P-norm analog of the Taylor-like expansion (3)

$$\begin{aligned} f(X - \alpha D) &\leq f(X) - \alpha\langle \nabla f(X), D\rangle + \alpha^2 (L_P/2)\|D\|_P^2 & (8) \\ &= f(X) - \alpha(1 - \alpha(L_P/2))\|\nabla f(X)\|_{P*}^2 & (9) \end{aligned}$$

where L_P is a P-norm gradient Lipschitz constant. If we can demonstrate gradient dominance under the dual P-norm,

$$\|\nabla f(X)\|_{P*}^2 \geq \mu_P (f(X) - f^\star) \quad \text{ where } \mu_P > 0 \text{ and } f^\star = \min f(X), \tag{10}$$

then we have the desired linear convergence

$$\begin{aligned} f(X - \alpha D) - f^\star &\leq [1 - \mu_P \alpha(1 - \alpha L_P/2)] \cdot (f(X) - f^\star) & (11) \\ &= [1 - 1/(2\kappa_P)] \cdot (f(X) - f^\star) \text{ with } \alpha = 1/L_P, & (12) \end{aligned}$$

in which the condition number $\kappa_P = L_P/\mu_P$ should be upper-bounded. To make the most progress per iteration, we want to pick a metric $\mathbf{P}$ to make the condition number κ_P as small as possible.

The best choice of $\mathbf{P}$ for the fastest convergence is simply the Hessian $\nabla^2 f(X)$ itself, but this simply recovers Newton's method, which would force us to invert a large $nr \times nr$ matrix to compute the search direction D at every iteration. Instead, we look for a *preconditioner* $\mathbf{P}$ that is cheap to apply while still assuring a relatively small condition number κ_P. The following choice is particularly interesting (the Kronecker product $\otimes$ is defined to satisfy $\text{vec}(AXB^T) = (B \otimes A)\text{vec}(X)$)

$$\mathbf{P} = (X^T X + \eta I_r) \otimes I_n = X^T X \otimes I_n + \eta I_{nr},$$

because the resulting $D = \nabla f(X)(X^T X + \eta I)^{-1}$ allow us to *interpolate* between regular GD and the ScaledGD of Tong et al. [11]. Indeed, we recover regular GD in the limit $\eta \to \infty$, but as we saw in Section 4, gradient dominance (10) fails to hold, so the condition number $\kappa_P = L_P/\mu_P$ grows unbounded as $\mu_P \to 0$. Instead, setting $\eta = 0$ recovers ScaledGD. The key insight of Tong et al. [11] is that under this choice of $\mathbf{P}$, gradient dominance (10) is guaranteed to hold, with a large value of μ_P that is independent of the current iterate and the ground truth. But as we will now show, this change of metric can magnify the Lipschitz constant L_P by a factor of $\lambda_{\min}^{-1}(X^T X)$, so the condition number $\kappa_P = L_P/\mu_P$ becomes unbounded in the over-parameterized regime.

Lemma 2 (Lipschitz-like inequality)**.** *Let* $\|D\|_P = \|D(X^T X + \eta I_r)^{1/2}\|_F$. *Then we have*

$$f(X + D) \le f(X) + \langle \nabla f(X), D \rangle + \frac{1}{2} L_P(X, D) \|D\|_P^2$$

where

$$L_P(X, D) = 2(1 + \delta) \left[4 + \frac{2\|XX^T - M^\star\|_F + 4\|D\|_P}{\lambda_{\min}(X^T X) + \eta} + \left(\frac{\|D\|_P}{\lambda_{\min}(X^T X) + \eta} \right)^2 \right]$$

Lemma 3 (Bounded gradient)**.** *For the search direction* $D = \nabla f(X)(X^T X + \eta I)^{-1}$, *we have* $\|D\|_P^2 = \|\nabla f(X)\|_{P*}^2 \le 16(1 + \delta) f(X)$.

The proofs of Lemma 2 and Lemma 3 follows from straightforward linear algebra, and can be found in the Appendix. Substituting Lemma 3 into Lemma 2, we see for ScaledGD (with $\eta = 0$) that the Lipschitz-like constant is bounded as follows

$$L_P(X, D) \lesssim \left(\|XX^T - M^\star\|_F / \lambda_{\min}(X^T X) \right)^2. \tag{13}$$

In the exact rank case $r = r^\star$, the distance of X from singularity can be lower-bounded, within a "good" neighborhood of the ground truth, since $\lambda_{\min}(X^T X) = \lambda_r(X^T X)$ and

$$\|XX^T - M^\star\|_F \le \rho \lambda_r(M^\star), \quad \rho < 1 \implies \lambda_r(X^T X) \ge (1 - \rho)\lambda_r(M^\star) > 0. \tag{14}$$

Within this "good" neighborhood, substituting (14) into (13) yields a Lipschitz constant L_P that depends only on the radius ρ. The resulting iterations converge rapidly, independent of any ill-conditioning in the model XX^T nor in the ground-truth $M^\star$. In turn, ScaledGD can be initialized within the good neighborhood using spectral initialization (see Proposition 6 below).

In the over-parameterized case $r > r^\star$, however, the iterate X must become singular in order for XX^T to converge to $M^\star$, and the radius of the "good" neighorhood reduces to zero. The ScaledGD direction guarantees a large linear progress no matter how singular X may be, but the method may not be able to take a substantial step in this direction if X becomes singular too quickly. To illustrate: the algorithm would fail entirely if it lands at on a point where $\lambda_{\min}(X^T X) = 0$ but $XX^T \neq M^\star$.

While regular GD struggles to make the smallest eigenvalues of XX^T converge to zero, ScaledGD gets in trouble by making these eigenvalues converge quickly. In finding a good mix between these two methods, an intuitive idea is to use the damping parameter η to control the rate at which X becomes singular. More rigorously, we can pick an $\eta \approx \|XX^T - ZZ^T\|_F$ and use Lemma 2 to keep the Lipschitz constant L_P bounded. Substituting Lemma 3 into Lemma 2 and using RIP to upper-bound $f(X) \le (1 + \delta)\|XX^T - M^\star\|_F^2$ and $\delta \le 1$ yields

$$\eta \ge C_{\text{lb}} \|XX^T - ZZ^T\|_F \implies L_P(X, D) \le 16 + 136/C_{\text{lb}} + 256/C_{\text{lb}}^2. \tag{15}$$

However, the gradient dominance condition (10) will necessarily fail if η is set too large. Our main result in this paper is that keeping η within the same order of magnitude as the error norm $\|XX^T - ZZ^T\|_F$ is enough to maintain gradient dominance. The following is the noiseless version of this result.

Theorem 4 (Noiseless gradient dominance)**.** *Let* $\min_X f(X) = 0$ *for* $M^\star \neq 0$*. Suppose that* X *satisfies* $f(X) \leq \rho^2 \cdot (1-\delta)\lambda_{r^\star}^2(M^\star)$ *with radius* $\rho > 0$ *that satisfies* $\rho^2/(1-\rho^2) \leq (1-\delta^2)/2$*. Then, we have*

$$\eta \leq C_{\mathrm{ub}}\|XX^T - ZZ^T\|_F \quad \Longrightarrow \quad \|\nabla f(X)\|_{P*}^2 \geq 2\mu_P f(X)$$

where

$$\mu_P = \left(\sqrt{\frac{1+\delta^2}{2}} - \delta\right)^2 \cdot \min\left\{\left(\frac{C_{\mathrm{ub}}}{\sqrt{2}-1}\right)^{-1}, \left(1 + 3C_{\mathrm{ub}}\sqrt{\frac{(r-r^\star)}{1-\delta^2}}\right)^{-1}\right\}. \tag{16}$$

The proof of Theorem 4 is involved and we defer the details to the Appendix. In the noiseless case, we get a good estimate of η for free as a consequence of RIP:

$$\eta = \sqrt{f(X)} \implies \sqrt{1-\delta}\|XX^T - M^\star\|_F \leq \eta \leq \sqrt{1+\delta}\|XX^T - M^\star\|_F.$$

Repeating (8)-(12) with Lemma 2, (15) and (16) yields our main result below.

Corollary 5 (Linear convergence)**.** *Let* X *satisfy the same initial conditions as in Theorem 4. The search direction* $D = \nabla f(X)(X^TX + \eta I)^{-1}$ *with damping parameter* $\eta = \sqrt{f(X)}$ *and step-size* $\alpha \leq 1/L_P$ *yields*

$$f(X - \alpha D) \leq (1 - \alpha\mu_P/2)\, f(X)$$

where L_P *is as in (15) with* $C_{\mathrm{lb}} = \sqrt{1-\delta}$ *and* μ_P *is as in (16) with* $C_{\mathrm{ub}} = \sqrt{1+\delta}$*.*

For a fixed RIP constant δ, Corollary 5 says that PrecGD converges at a linear rate that is independent of the current iterate X, and also independent of possible ill-conditioning in the ground truth. However, it does require an initial point X_0 that satisfies

$$\|\mathcal{A}(X_0X_0^T - M^*)\|^2 < \rho^2(1-\delta)\lambda_{r^*}(M^\star)^2 \tag{17}$$

with a radius $\rho > 0$ satisfying $\rho^2/(1-\rho^2) \leq (1-\delta^2)/2$. Such an initial point can be found using spectral initialization, even if the measurements are tainted with noise. Concretely, we choose the initial point X_0 as

$$X_0 = \mathcal{P}_r\left(\frac{1}{m}\sum_{i=1}^m y_iA_i\right) \text{ where } \mathcal{P}_r(M) = \arg\min_{X\in\mathbb{R}^{n\times r}} \|XX^T - M\|_F, \tag{18}$$

where we recall that $y = \mathcal{A}(M^\star) + \epsilon$ are the m possibly noisy measurements collected of the ground truth, and that the rank-r projection operator can be efficiently implemented with a singular value decomposition. The proof of the following proposition can be found in the appendix.

Proposition 6 (Spectral Initialization)**.** *Suppose that* $\delta \leq (8\kappa\sqrt{r^*})^{-1}$ *and* $m \gtrsim \frac{1+\delta}{1-\delta}\frac{\sigma^2 rn\log n}{\rho^2\lambda_{r^\star}^2(M^\star)}$ *where* $\kappa = \lambda_1(M^\star)/\lambda_{r^\star}(M^\star)$*. Then, with high probability, the initial point* X_0 *produced by* (18) *satisfies the radius condition (17).*

However, if the measurements y are noisy, then $\sqrt{f(X)} = \|\mathcal{A}(XX^T - M^\star) + \varepsilon\|$ now gives a biased estimate of our desired damping parameter η. In the next section, we show that a good choice of η_k is available based on an approximation of the noise variance.

6 Extension to Noisy Setting

In this section, we extend our analysis to the matrix sensing with noisy measurements. Our main goal is to show that, with a proper choice of the damping coefficient η, the proposed algorithm converges linearly to an "optimal" estimation error.

Theorem 7 (Noisy measurements with optimal η)**.** *Suppose that the noise vector* $\epsilon \in \mathbb{R}^m$ *has sub-Gaussian entries with zero mean and variance* $\sigma^2 = \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\epsilon_i^2]$*. Moreover, suppose that* $\eta_k = \frac{1}{\sqrt{m}}\|\mathcal{A}(X_kX_k^T - M^*)\|$*, for* $k = 0, 1, \ldots, K$*, and that the initial point* X_0 *satisfies* $\|\mathcal{A}(X_0X_0^T - M^*)\|^2 < \rho^2(1-\delta)\lambda_{r^*}(M^\star)^2$*. Consider* $k^* = \arg\min_k \eta_k$*, and suppose that the step-size* $\alpha \leq 1/L$*, where* $L > 0$ *is a constant that only depends on* δ*. Then, with high probability, we have*

$$\|X_{k^*}X_{k^*}^T - M^\star\|_F^2 \lesssim \max\left\{\frac{1+\delta}{1-\delta}\left(1 - \alpha\frac{\mu_P}{2}\right)^K \|X_0X_0^T - M^*\|_F^2, \mathcal{E}_{stat}\right\}, \tag{19}$$

where $\mathcal{E}_{stat} := \frac{\sigma^2 nr\log n}{\mu_P(1-\delta)m}$*.*

Assuming fixed parameters for the problem, the above theorem shows that PrecGD outputs a solution with an estimation error of $\mathcal{O}\left(\mathcal{E}_{stat}\right)$ in $\mathcal{O}\left(\log\left(1/\mathcal{E}_{stat}\right)\right)$ iterations. Moreover, the error $\mathcal{O}\left(\mathcal{E}_{stat}\right)$ is minimax optimal (modulo logarithmic factors), and cannot be improved significantly. In particular, Candes and Plan [14] showed that *any* estimator $\widehat{X}$ must satisfy $\|\widehat{X}\widehat{X}^T - M^*\|_F^2 \gtrsim \sigma^2 nr/m$ with non-negligible probability. The classical methods for achieving this minimax rate suffer from computationally-prohibitive per iteration costs [15, 21, 48]. Regular gradient descent alleviates this issue at the expense of a slower convergence rate of $\mathcal{O}(\sqrt{1/\mathcal{E}_{stat}})$ [8]. Our proposed PrecGD achieves the best of both worlds: it converges to the minimax optimal error with cheap per-iteration complexity of $\mathcal{O}(nr^2 + r^3)$, while benefiting from an exponentially faster convergence rate than regular gradient descent in the over-parameterized regime.

Theorem 7 highlights the critical role of the damping coefficient η in the guaranteed linear convergence of the algorithm. In the noiseless regime, we showed in the previous section that an "optimal" choice $\eta = \sqrt{f(X)}$ is available for free. In the noisy setting, however, the same choice of η becomes biased by the noise variance, and is therefore no longer optimal. As is typically the case for regularized estimation methods [49–51], selecting the ideal parameter would amount to some kind of *resampling*, such as via cross-validation or bootstrapping [52–54], which is generally expensive to implement and use in practice. As an alternative approach, we show in our next theorem that a good choice of η is available based on an approximation of the noise variance σ^2.

Theorem 8 (Noisy measurements with variance proxy)**.** *Suppose that the noise vector $\epsilon \in \mathbb{R}^m$ has sub-Gaussian entries with zero mean and variance $\sigma^2 = \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\epsilon_i^2]$. Moreover, suppose that $\eta_k = \sqrt{|f(X_k) - \hat{\sigma}^2|}$ for $k = 0, 1, \dots, K$, where $\hat{\sigma}^2$ is an approximation of σ^2, and that the initial point X_0 satisfies $\|\mathcal{A}(X_0X_0^T - M^*)\|_F^2 < \rho^2(1-\delta)\lambda_{r^*}(M^\star)^2$. Consider $k^* = \arg\min_k \eta_k$, and suppose that the step-size $\alpha \le 1/L$, where $L > 0$ is a constant that only depends on δ. Then, with high probability, we have*

$$\|X_{k^*}X_{k^*}^T - M^*\|_F^2 \lesssim \max\left\{\frac{1+\delta}{1-\delta}\left(1 - \alpha\frac{\mu_P}{2}\right)^K \|X_0X_0^T - M^*\|_F^2, \mathcal{E}_{stat}, \mathcal{E}_{dev}, \mathcal{E}_{var}\right\}, \tag{20}$$

where

$$\mathcal{E}_{stat} := \frac{\sigma^2 nr\log n}{\mu_P(1-\delta)m}, \quad \mathcal{E}_{dev} := \frac{\sigma^2}{1-\delta}\sqrt{\frac{\log n}{m}}, \quad \mathcal{E}_{var} := |\sigma^2 - \hat{\sigma}^2|. \tag{21}$$

In the above theorem, $\mathcal{E}_{dev}$ captures the deviation of the empirical variance $\frac{1}{m}\sum_{i=1}^m \epsilon_i^2$ from its expectation σ^2. On the other hand, $\mathcal{E}_{var}$ captures the approximation error of the true variance. According to Theorem 8, it is possible to chose the damping factor η_k merely based on $f(X_k)$ and an approximation of σ^2, at the expense of a suboptimal estimation error rate. In particular, suppose that the noise variance is known precisely, i.e., $\hat{\sigma}^2 = \sigma^2$. Then, the above theorem implies that the estimation error is reduced to

$$\|X_{k^*}X_{k^*}^T - M^*\|_F^2 \lesssim \max\left\{\mathcal{E}_{stat}, \mathcal{E}_{dev}\right\} \quad \text{after} \quad \mathcal{O}\left(\log\left(\frac{1}{\max\left\{\mathcal{E}_{stat}, \mathcal{E}_{dev}\right\}}\right)\right) \text{ iterations.}$$

If m is not too large, i.e., $m \lesssim \sigma^2 n^2 r^2 \log n$, the estimation error can be improved to $\|X_{k^*}X_{k^*}^T - M^*\|_F^2 \lesssim \mathcal{E}_{stat}$, which is again optimal (modulo logarithmic factors). As m increases, the estimation error will become smaller, but the convergence rate will decrease. This suboptimal rate is due to the heavy tail phenomenon arising from the concentration of the noise variance. In particular, one can write

$$f(X) - \sigma^2 = \frac{1}{m}\|\mathcal{A}(XX^T - M^\star)\|^2 + \underbrace{\frac{1}{m}\|\epsilon\|^2 - \sigma^2}_{\text{variance deviation}} + \underbrace{\frac{2}{m}\langle\mathcal{A}(ZZ^T - XX^T), \epsilon\rangle}_{\text{cross-term}} \tag{22}$$

Evidently, $f(X) - \sigma^2$ is in the order of $\frac{1}{m}\|\mathcal{A}(XX^T - M^\star)\|^2$ if both variance deviation and cross-term are dominated by $\frac{1}{m}\|\mathcal{A}(XX^T - M^\star)\|^2$. In the proof of Theorem 8, we show that, with high probability, the variance deviation is upper bounded by $(1-\delta)\mathcal{E}_{dev}$ and it dominates the cross-term. This implies that the choice of $\eta = \sqrt{|f(X) - \sigma^2|}$ behaves similar to $\frac{1}{\sqrt{m}}\|\mathcal{A}(XX^T - M^\star)\|$, and hence, the result of Theorem 7 can be invoked, so long as

$$\frac{1}{m}\|\mathcal{A}(XX^T - M^\star)\|^2 \ge (1-\delta)\|XX^T - M^\star\|_F^2 \gtrsim (1-\delta)\mathcal{E}_{dev}.$$

7 Numerical Experiments

Finally, we numerically compare PrecGD on other matrix factorization problems that fall outside of the matrix sensing framework. We consider the ℓ_p empirical loss $f_p(X) = \sum_{i=1}^{m} |\langle A_i, XX^T - M^\star \rangle|^p$ for $1 \le p < 2$, in order to gauge the effectiveness of PrecGD for increasing nonsmooth loss functions. Here, we set the damping parameter $\eta_k = [f_p(X_k)]^{1/p}$ as a heuristic for the error $\|XX^T - M^\star\|_F$. The data matrices $A_1, \ldots, A_m$ were taken from [13, Example 12], the ground truth $M^\star = ZZ^T$ was constructed by sampling each column of $Z \in \mathbb{R}^{n \times r^\star}$ from the standard Gaussian, and then rescaling the last column to achieve a desired condition number.

The recent work of Tong et al. [55] showed that in the exactly-parameterized setting, ScaledGD works well for the ℓ_1 loss function. In particular, if the initial point is close to the ground truth, then with a Polyak stepsize $\alpha_k = f(X_k)/\|\nabla f(X_k)\|_P^*$, ScaledGD converges linearly to the ground truth. However, these theoretical guarantees no longer hold in the over-parameterized regime.

When $r > r^*$, our numerical experiments show that ScaledGD blows up due to singularity near the ground truth while PrecGD continues to converge linearly in this nonsmooth, over-parameterized setting. In Figure 2 we compare GD, ScaledGD and PrecGD in the exact and over-parameterized regimes for the ℓ_p norm, with $p = 1.1, 1.4$ and 1.7. For ScaledGD and PrecGD, we used a modified version of the Polyak step-size where $\alpha_k = f(X_k)^p/\|\nabla f(X_k)\|_P^*$. For GD we use a decaying stepsize. When $r = r^*$, we see that both ScaledGD and PrecGD converge linearly, but GD stagnates due to ill-conditioning of the ground truth. When $r > r^*$, GD still converges slowly and ScaledGD blows up very quickly, while PrecGD continues to converge reliably.

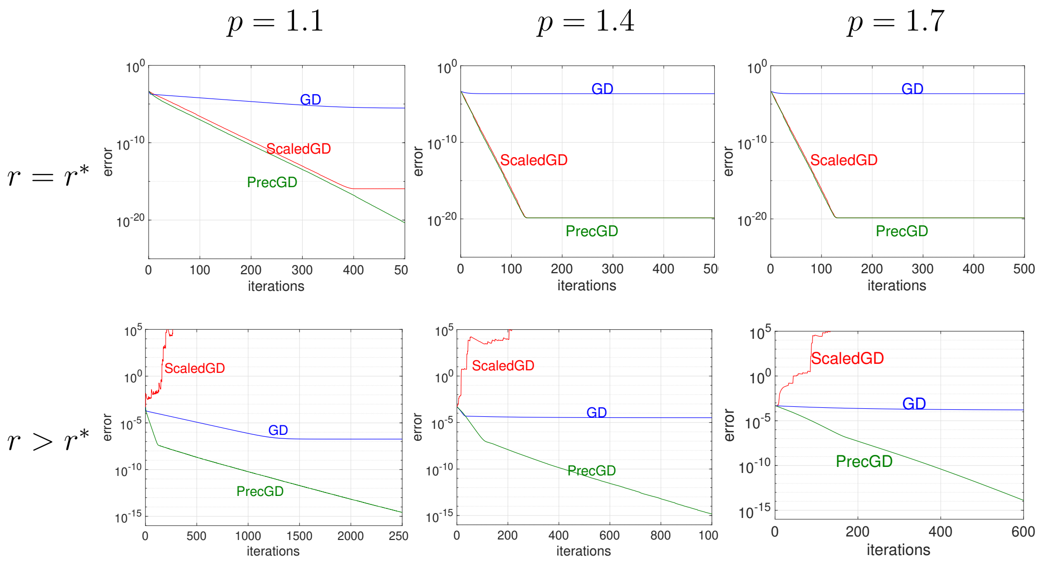

Figure 2: **Nonconvex matrix factorization with the ℓ_p empirical loss**. We compare ℓ_p matrix sensing with $n = 10$ and $r^\star = 2$ and $\mathcal{A}$ taken from [13]. The ground truth is chosen to be ill-conditioned ($\kappa = 10^2$). For ScaledGD and PrecGD, we use the Polyak step-size in [55]. For GD we use a decaying step-size. **(Top** $r = r^*$**)** For all three values of p, GD stagnates due to the ill-conditioning of the ground truth, while ScaledGD and PrecGD converge linearly in all three cases. **(Bottom** $r > r^*$**)** With $r = 4$, the problem is over-parameterized. GD again converges slowly and ScaledGD is sporadic due to near-singularity caused by over-parameterization. Once again we see PrecGD converge at a linear rate.

8 Conclusions

In this paper, we propose a *preconditioned* gradient descent or PrecGD for nonconvex matrix factorization with a comparable per-iteration cost to classical gradient descent. For over-parameterized matrix sensing, gradient descent slows down to a sublinear convergence rate, but PrecGD restores the convergence rate back to linear, while also making the iterations immune to ill-conditioning in the ground truth. While the thoeretical analysis in our paper uses some properties specific to RIP matrix sensing, our numerical experiments find that PrecGD works well for even for nonsmooth loss functions. We believe that these current results can be extended to similar problems such as matrix completion and robust PCA, where properties like incoherence can be used to select the damping parameter η_k with the desired properties, so that PrecGD converges linearly as well. It remains future work to provide rigorous justification for these observations.

Acknowledgements

G.Z. and R.Y.Z are supported by the NSF CAREER Award ECCS-2047462. S.F. is supported by MICDE Catalyst Grant and MIDAS PODS Grant. We also thank an anonymous reviewer who provided a simplified proof of Lemma 14 and made various insightful comments to help us improve an earlier version of this work.

References

[1] Kai Yu, Shenghuo Zhu, John Lafferty, and Yihong Gong. Fast nonparametric matrix factorization for large-scale collaborative filtering. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 211–218, 2009.

[2] Xin Luo, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics*, 10(2):1273–1284, 2014.

[3] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[4] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[5] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2013.

[6] Shuyang Ling and Thomas Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

[7] Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011.

[8] Jiacheng Zhuo, Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the computational and statistical complexity of over-parameterized matrix sensing. *arXiv preprint arXiv:2102.02756*, 2021.

[9] Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[10] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

[11] Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *arXiv preprint arXiv:2005.08898*, 2020.

[12] Richard Zhang, Cedric Josz, Somayeh Sojoudi, and Javad Lavaei. How much restricted isometry is needed in nonconvex matrix recovery? In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[13] Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *Journal of Machine Learning Research*, 20(114):1–34, 2019.

[14] Emmanuel J Candes and Yaniv 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.

[15] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[16] Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *arXiv preprint arXiv:1506.06081*, 2015.

[17] Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582. PMLR, 2016.

[18] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[19] Jianhao Ma and Salar Fattahi. Implicit regularization of sub-gradient method in robust matrix recovery: Don't be afraid of outliers. *arXiv preprint arXiv:2102.02969*, 2021.

[20] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[21] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[22] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[23] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. *arXiv preprint arXiv:1410.7660*, 2014.

[24] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.

[25] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019.

[26] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[27] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.

[28] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

[29] Ji Chen and Xiaodong Li. Memory-efficient kernel pca via partial matrix sampling and non-convex optimization: a model-free analysis of local minima. *arXiv preprint arXiv:1711.01742*, 2017.

[30] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

[31] Richard Y Zhang. Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime. *arXiv preprint arXiv:2104.10790*, 2021.

[32] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

[33] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

[34] Raghu Meka, Prateek Jain, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. *arXiv preprint arXiv:0909.5457*, 2009.

[35] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[36] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[37] Farid Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM journal on Optimization*, 5(1):13–51, 1995.

[38] Zaiwen Wen, Donald Goldfarb, and Wotao Yin. Alternating direction augmented lagrangian methods for semidefinite programming. *Mathematical Programming Computation*, 2(3-4): 203–230, 2010.

[39] Brendan O'donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

[40] Yang Zheng, Giovanni Fantuzzi, Antonis Papachristodoulou, Paul Goulart, and Andrew Wynn. Chordal decomposition in operator-splitting methods for sparse semidefinite programs. *Mathematical Programming*, 180(1):489–532, 2020.

[41] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

[42] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

[43] Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. In *Conference on learning theory*, pages 638–678. PMLR, 2014.

[44] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust pca via gradient descent. *arXiv preprint arXiv:1605.07784*, 2016.

[45] Mahdi Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *IEEE Transactions on Information Theory*, 65(4):2374–2400, 2019.

[46] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

[47] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[48] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[49] Ernesto De Vito, Andrea Caponnetto, and Lorenzo Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5(1): 59–85, 2005.

[50] Gavin C Cawley. Leave-one-out cross-validation based model selection criteria for weighted ls-svms. In *The 2006 IEEE international joint conference on neural network proceedings*, pages 1661–1668. IEEE, 2006.

[51] Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[52] Phillip I Good. *Resampling methods*. Springer, 2006.

[53] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[54] David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. CRC Press, 1979.

[55] Tian Tong, Cong Ma, and Yuejie Chi. Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. *IEEE Transactions on Signal Processing*, 69:2396–2409, 2021.

[56] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[57] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

A Preliminaries for the Noiseless Case

Recall that the matrix inner product is defined $\langle X, Y\rangle \overset{\text{def}}{=} \operatorname{Tr}\left(X^T Y\right)$, and that it induces the Frobenius norm as $\|X\|_F = \sqrt{\langle X, X\rangle}$. The vectorization $\operatorname{vec}(X)$ is the usual column-stacking operation that turns an $m \times n$ matrix into a length-mn vector; it preserves the matrix inner product $\langle X, Y\rangle = \mathrm{vec}(X)^T \mathrm{vec}(Y)$ and the Frobenius norm $\|\mathrm{vec}(X)\| = \|X\|_F$. The Kronecker product $\otimes$ is implicitly defined to satisfy $\mathrm{vec}(AXB^T) = (B \otimes A)\mathrm{vec}X$.

We denote $\lambda_i(M)$ and $\sigma_i(M)$ as the i-th eigenvalue and singular value of a symmetric matrix $M = M^T$, ordered from the most positive to the most negative. We will often write $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ to index the most positive and most negative eigenvalues, and $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ for the largest and smallest singular values.

We denote $\mathbf{A} = [\mathrm{vec}(A_1), \ldots, \mathrm{vec}(A_m)]^T$ as the matrix representation of $\mathcal{A}$, and note that $\mathcal{A}(X) = \mathbf{A}\,\mathrm{vec}(X)$. For fixed X and $M^\star$, we can rewrite f in terms of the error matrix E or its vectorization $\mathbf{e}$ as follows

$$f(X) = \|\mathcal{A}(E)\|^2 = \|\mathbf{A}\mathbf{e}\|^2 \text{ where } E = XX^T - M^\star, \quad \mathbf{e} = \mathrm{vec}(E). \tag{23}$$

The gradient satisfies for any matrix $D \in \mathbb{R}^{n\times r}$

$$\langle \nabla f(X), D\rangle = 2\left\langle \mathcal{A}\left(XD^T + DX^T\right), \mathcal{A}\left(E\right)\right\rangle. \tag{24}$$

Letting $\mathbf{J}$ denote the Jacobian of the vectorized error $\mathbf{e}$ with respect to X implicitly as the matrix that satisfies

$$\mathbf{J}\,\mathrm{vec}(Y) = \mathrm{vec}(XY^T + YX^T) \qquad \text{for all } Y \in \mathbb{R}^{n\times r}. \tag{25}$$

allows us to write the gradient exactly as $\mathrm{vec}(\nabla f(X)) = 2\mathbf{J}^T\mathbf{A}^T\mathbf{A}\mathbf{e}$. The noisy versions of (23) and (24) are obvious, though we will defer these to Section E.

Recall that $\mathcal{A}$ is assumed to satisfy RIP (Definition 1) with parameters $(2r, \delta)$. Here, we set $m = 1$ without loss of generality to avoid carrying the normalizing constant; the resulting RIP inequality reads

$$(1-\delta)\|M\|_F^2 \le \|\mathcal{A}(M)\|^2 \le (1+\delta)\|M\|_F^2 \text{ for all } M \text{ such that } \mathrm{rank}(M) \le 2r, \tag{26}$$

where we recall that $0 \le \delta < 1$. It is easy to see that RIP preserves the Cauchy–Schwarz identity for all rank-$2r$ matrices G and H:

$$\langle \mathcal{A}(G), \mathcal{A}(H)\rangle \le \|\mathcal{A}(G)\|\|\mathcal{A}(H)\| \le (1+\delta)\|G\|_F\|H\|_F. \tag{27}$$

As before, we introduce the preconditioner matrix P as

$$P \overset{\text{def}}{=} X^TX + \eta I_r, \qquad \mathbf{P} \overset{\text{def}}{=} P \otimes I_n = (X^TX + \eta I_r) \otimes I_n$$

and define a corresponding P-inner product, P-norm, and dual P-norm on $\mathbb{R}^{n\times r}$ as follows

$$\langle X, Y\rangle_P \overset{\text{def}}{=} \mathrm{vec}(X)^T\mathbf{P}\mathrm{vec}(Y) = \left\langle XP^{1/2}, YP^{1/2}\right\rangle = \mathrm{Tr}\left(XPY^T\right), \tag{28a}$$

$$\|X\|_P \overset{\text{def}}{=} \sqrt{\langle X, X\rangle_P} = \|\mathbf{P}^{1/2}\mathrm{vec}(X)\| = \|XP^{1/2}\|_F, \tag{28b}$$

$$\|X\|_{P*} \overset{\text{def}}{=} \max_{\|Y\|_P = 1}\langle Y, X\rangle = \|\mathbf{P}^{-1/2}\mathrm{vec}(X)\| = \|XP^{-1/2}\|_F. \tag{28c}$$

Finally, we will sometimes need to factorize the ground truth $M^\star = ZZ^T$ in terms of the low-rank factor $Z \in \mathbb{R}^{n\times r^\star}$.

B Proof of Lipschitz-like Inequality (Lemma 2)

In this section we give a proof of Lemma 2, which is a Lipschitz-like inequality under the P-norm. Recall that we proved linear convergence for PrecGD by lower-bounding the linear progress $\langle \nabla f(X), D\rangle$ and upper-bounding $\|D\|_P$.

Lemma 9 (Lipschitz-like inequality; Lemma 2 restated)**.** *Let $\|D\|_P = \|D(X^TX + \eta I)^{1/2}\|_F$. Then we have*

$$f(X + D) \le f(X) + \langle \nabla f(X), D\rangle + \frac{1}{2}L_P(X, D)\|D\|_P^2$$

where

$$L_P(X, D) = 2(1+\delta)\left[4 + \frac{2\|XX^T - M^\star\|_F + 4\|D\|_P}{\lambda_{\min}(X^TX) + \eta} + \left(\frac{\|D\|_P}{\lambda_{\min}(X^TX) + \eta}\right)^2\right]$$

Proof. Recall that $E = XX^T - M^\star$. We obtain a Taylor expansion of the quartic polynomial f by directly expanding the quadratic terms

$$\begin{aligned} f(X+D) &= \|\mathcal{A}((X+D)(X+D)^T - M^\star)\|^2 \\ &= \underbrace{\|\mathcal{A}(E)\|^2 + 2\langle \mathcal{A}(E), \mathcal{A}(XD^T + DX^T)\rangle}_{f(X)+\langle \nabla f(X), D\rangle} + \underbrace{2\langle \mathcal{A}(E), \mathcal{A}(DD^T)\rangle + \|\mathcal{A}(XD^T + DX^T)\|^2}_{\frac{1}{2}\langle \nabla^2 f(X)[D], D\rangle} \\ &\quad + \underbrace{2\langle \mathcal{A}(XD^T + DX^T), \mathcal{A}(DD^T)\rangle}_{\frac{1}{6}\langle \nabla^3 f(X)[D,D], D\rangle} + \underbrace{\|\mathcal{A}(DD^T)\|^2}_{\frac{1}{24}\langle \nabla^4 f(X)[D,D,D], D\rangle} \quad . \end{aligned}$$

We evoke RIP to preserve Cauchy–Schwarz as in (27), and then bound the second, third, and fourth order terms

$$\begin{aligned} T =& 2\langle \mathcal{A}(E), \mathcal{A}(DD^T)\rangle + \|\mathcal{A}(XD^T + DX^T)\|^2 + 2\langle \mathcal{A}(XD^T + DX^T), \mathcal{A}(DD^T)\rangle + \|\mathcal{A}(DD^T)\|^2 \\ \le& (1+\delta)\left(2\|E\|_F\|DD^T\|_F + \|XD^T + DX^T\|^2 + 2\|XD^T + DX^T\|_F\|DD^T\|_F + \|DD^T\|_F^2\right) \\ \le& (1+\delta)\left(2\|E\|_F\|D\|_F^2 + 4\|XD^T\|^2 + 4\|XD^T\|_F\|D\|_F^2 + \|D\|_F^4\right) \end{aligned} \tag{29}$$

where the third line uses $\|DD^T\|_F \le \|D\|_F^2$ and $\|XD^T + DX^T\|_F \le 2\|XD^T\|_F$. Now, write $d = \mathrm{vec}(D)$ and observe that

$$\|D\|_F^2 = d^T d = (d^T \mathbf{P}^{1/2})\mathbf{P}^{-1}(\mathbf{P}^{1/2} d) \le (d^T \mathbf{P} d)\lambda_{\max}(\mathbf{P}^{-1}) = \|D\|_P^2 / \lambda_{\min}(\mathbf{P}). \tag{30}$$

Similarly, we have

$$\|XD^T\|_F = \|XP^{-1/2}P^{1/2}D^T\|_F \le \sigma_{\max}(XP^{-1/2})\|P^{1/2}D^T\|_F \le \|D\|_P. \tag{31}$$

The final inequality uses $\|P^{1/2}D^T\|_F = \|DP^{1/2}\|_F = \|D\|_P$ and that

$$\sigma_{\max}(XP^{-1/2}) = \sigma_{\max}[X(X^T X + \eta I)^{-1/2}] = \sigma_{\max}(X)/\sqrt{\sigma_{\max}^2(X) + \eta} \le 1. \tag{32}$$

Substituting (30) and (31) into (29) yields

$$T \le (1+\delta)\left(2\|E\|_F \frac{\|D\|_P^2}{\lambda_{\min}(\mathbf{P})} + 4\|D\|_P^2 + \frac{4\|D\|_P^3}{\lambda_{\min}(\mathbf{P})} + \frac{\|D\|_P^4}{\lambda_{\min}^2(\mathbf{P})}\right) = \frac{1}{2} L_P(X,D)\|D\|_P^2$$

where we substitute $\lambda_{\min}(\mathbf{P}) = \lambda_{\min}(X^T X) + \eta$. □

C Proof of Bounded Gradient (Lemma 3)

In this section we prove Lemma 3, which shows that the gradient measured in the dual P-norm $\|\nabla f(X)\|_{P*}$ is controlled by the objective value as $\sqrt{f(X)}$.

Lemma 10 (Bounded Gradient; Lemma 3 restated)**.** *For the search direction $D = \nabla f(X)(X^T X + \eta I)^{-1}$, we have $\|D\|_P^2 = \|\nabla f(X)\|_{P*}^2 \le 16(1+\delta)f(X)$.*

Proof. We apply the variation definition of the dual P-norm in (28c) to the gradient in (24) to obtain

$$\begin{aligned} \|\nabla f(X)\|_{P^*} &= \max_{\|Y\|_P = 1} \langle \nabla f(X), Y\rangle = \max_{\|Y\|_P = 1} 2\left\langle \mathcal{A}(XY^T + YX^T), \mathcal{A}(E)\right\rangle \\ &\overset{\text{(a)}}{\le} 2\|\mathcal{A}(E)\| \max_{\|Y\|_P = 1} \|\mathcal{A}(XY^T + YX^T)\| \overset{\text{(b)}}{\le} 4\sqrt{(1+\delta)f(X)} \max_{\|Y\|_P = 1} \|XY^T\|_F \end{aligned}$$

Here (a) applies Cauchy–Schwarz; and (b) substitutes $f(X) = \|\mathcal{A}(E)\|^2$ and $\|\mathcal{A}(M)\| \le \sqrt{1+\delta}\|M\|_F$ for rank-$2r$ matrix M and $\|XY^T + YX^T\|_F \le 2\|XY^T\|_F$. Now, we bound the final term

$$\max_{\|Y\|_P = 1} \|XY^T\|_F = \max_{\|YP^{1/2}\|_F = 1} \|XY^T\|_F = \max_{\|\tilde{Y}\|_F = 1} \|XP^{-1/2}\tilde{Y}^T\|_F = \sigma_{\max}(XP^{-1/2}) \le 1$$

where the final inequality uses (32). □

D Proof of Gradient Dominance (Theorem 4)

In this section we prove our first main result: the gradient $\nabla f(X)$ satisfies gradient dominance the P-norm. This is the key insight that allowed us to establish the linear convergence rate of PrecGD in the main text. The theorem is restated below.

Theorem 11 (Gradient Dominance; Theorem 4 restated)**.** *Let $\min_X f(X) = 0$ for $M^\star \neq 0$. Suppose that X satisfies $f(X) \leq \rho^2 \cdot (1-\delta)\lambda_{r^\star}^2(M^\star)$ with radius $\rho > 0$ that satisfies $\rho^2/(1-\rho^2) \leq (1-\delta^2)/2$. Then, we have*

$$\eta \leq C_{\mathrm{ub}} \|XX^T - M^\star\|_F \quad \Longrightarrow \quad \|\nabla f(X)\|_{P*}^2 \geq \mu_P f(X)$$

where

$$\mu_P = \left(\sqrt{\frac{1+\delta^2}{2}} - \delta\right)^2 \cdot \min\left\{\left(1 + \frac{C_{\mathrm{ub}}}{\sqrt{2}-1}\right)^{-1}, \left(1 + 3C_{\mathrm{ub}}\sqrt{\frac{(r-r^\star)}{1-\delta^2}}\right)^{-1}\right\}. \tag{33}$$

The theorem is a consequence of the following lemma, which shows that the PL constant $\mu_P > 0$ is driven in part by the alignment between the model XX^T and the ground truth $M^\star$, and in part in the relationship between η and the singular values of X. We defer its proof to Section D.1 and first use it to prove Theorem 4.

Lemma 12 (Gradient lower bound)**.** *Let $XX^T = U\Lambda U^T$ where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$, $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$, and $U^T U = I_r$ denote the usual eigenvalue decomposition. Let U_k denote the first k columns of U. Then, we have*

$$\|\nabla f(X)\|_{P^*}^2 \geq \max_{k\in\{1,2,\ldots,r\}} \frac{2(\cos\theta_k - \delta)^2}{1+\eta/\lambda_k} \|XX^T - M^\star\|_F^2 \tag{34}$$

where each θ_k is defined

$$\sin\theta_k = \frac{\left\|\left(I - U_k U_k^T\right)\left(XX^T - M^\star\right)\left(I - U_k U_k^T\right)\right\|_F}{\left\|XX^T - M^\star\right\|_F}. \tag{35}$$

From Lemma 12, we see that deriving a PL constant μ_P requires balancing two goals: (1) ensuring that $\cos\theta_k$ is large with respect to the RIP constant δ; (2) ensuring that $\lambda_k(X^T X)$ is large with respect to the damping parameter η.

As we will soon show, in the case that $k = r$, the corresponding $\cos\theta_r$ is guaranteed to be large with respect to δ, once XX^T converges towards $M^\star$. At the same time, we have by Weyl's inequality

$$\lambda_k(X^T X) = \lambda_k(XX^T) \geq \lambda_k(M^\star) - \|XX^T - M^\star\|_F \text{ for all } k \in \{1, 2, \ldots, r\}.$$

Therefore, when $k = r^\star$ and XX^T is close to $M^\star$, the corresponding $\lambda_{r^\star}(X^T X)$ is guaranteed to be large with respect to η . However, in order to use Lemma 12 to derive a PL constant $\mu_P > 0$, we actually need $\cos\theta_k$ and $\lambda_k(X^T X)$ to both be large for the *same* value of k. It turns out that when $\eta \gtrsim \|XX^T - M^\star\|_F$, it is possible to prove this claim using an inductive argument.

Before we present the complete argument and prove Theorem 4, we state one more lemma that will be used in the proof.

Lemma 13 (Basis alignment)**.** *Define the $n \times k$ matrix U_k in terms of the first k eigenvectors of X as in Lemma 12. Let $Z \in \mathbb{R}^{n\times r^\star}$ satisfy $\lambda_{\min}(Z^T Z) > 0$ and suppose that $\|XX^T - ZZ^T\|_F \leq \rho\lambda_{\min}(Z^T Z)$ with $\rho \leq 1/\sqrt{2}$. Then,*

$$\frac{\|Z^T(I - U_k U_k^T)Z\|_F}{\|XX^T - ZZ^T\|_F} \leq \frac{1}{\sqrt{2}} \frac{\rho}{\sqrt{1-\rho^2}} \quad \textit{for all } k \geq r^\star. \tag{36}$$

Essentially, this lemma states that as the rank-r matrix XX^T converges to the rank-$r^\star$ matrix $M^\star$, the top $r^\star$ eigenvectors of XX^T must necessarily rotate into alignment with $M^\star$. In fact, this is easily verified to be sharp by considering the $r = r^\star = 1$ case; we defer its proof to Section D.2.

With Lemma 12 and Lemma 13, we are ready to prove Theorem 4.

Proof of Theorem 4. We pick some μ satisfying $\delta < \mu < 1$ and prove that $\frac{\rho^2}{1-\rho^2} \le 1-\mu^2$ implies $\|\nabla f(X)\|_{P*}^2 \ge \mu_P f(X)$ where

$$\mu_P = (\mu-\delta)^2 \cdot \min\left\{ \left(1+\frac{C_{\text{ub}}}{\sqrt{2}-1}\right)^{-1}, \left(1+3C_{\text{ub}}\sqrt{\frac{r-r^\star}{1-\mu^2}}\right)^{-1} \right\}. \tag{37}$$

Then, setting $1-\mu^2 = \frac{1}{2}(1-\delta^2)$ yields our desired claim.

To begin, note that the hypothesis $\frac{\rho^2}{1-\rho^2} \le 1-\mu^2 \le 1$ implies $\rho \le 1/\sqrt{2}$. Denote $E = XX^T - M^\star$. We have

$$\frac{\|\nabla f(X)\|_{P*}^2}{f(X)} \overset{\text{(a)}}{\ge} \frac{\|\nabla f(X)\|_{P*}^2}{(1+\delta)\|E\|_F^2} \overset{\text{(b)}}{\ge} \frac{2(\cos\theta_k - \delta)^2}{(1+\delta)(1+\eta/\lambda_k(X^TX))} \overset{\text{(c)}}{\ge} \frac{(\cos\theta_k-\delta)^2}{1+\eta/\lambda_k(X^TX)} \text{ for all } k \ge r^\star. \tag{38}$$

Step (a) follows from RIP; Step (b) applies Lemma 12; Step (c) applies $1+\delta \le 2$. Equation (38) proves gradient dominance if we can show that both $\lambda_k(X^TX)$ and $\cos\theta_k$ are large for the same k. We begin with $k = r^\star$. Here we have by RIP and by hypothesis

$$(1-\delta)\|XX^T - M^\star\|_F^2 \le f(X) \le \rho^2 \cdot (1-\delta)\lambda_{\min}^2(Z^TZ), \tag{39}$$

which by Weyl's inequality yields

$$\lambda_{r^\star}(X^TX) = \lambda_{r^\star}(XX^T) \ge \lambda_{r^\star}(M^\star) - \|XX^T - M^\star\|_F \ge (1-\rho)\lambda_{r^\star}(M^\star).$$

This, combined with (39) and our hypothesis $\eta \le C_{\text{ub}}\|XX^T - ZZ^T\|_F$ and $\rho \le 1/\sqrt{2}$ gives

$$\frac{\eta}{\lambda_{r^\star}(X^TX)} \le \frac{\rho C_{\text{ub}}\lambda_{r^\star}(M^\star)}{(1-\rho)\lambda_{r^\star}(M^\star)} = \frac{\rho C_{\text{ub}}}{1-\rho} \le \frac{C_{\text{ub}}}{\sqrt{2}-1}, \tag{40}$$

which shows that $\lambda_{r^\star}(X^TX)$ is large. If $\cos\theta_k \ge \mu$ is also large, then substituting (40) into (38) yields gradient dominance

$$\frac{\|\nabla f(X)\|_{P*}^2}{f(X)} \ge (\mu-\delta)^2\left(1+\frac{C_{\text{ub}}}{\sqrt{2}-1}\right)^{-1},$$

and this yields the first term in (37). If $\cos\theta_k < \mu$ is actually small, then $\sin^2\theta_k > 1-\mu^2$ is large. We will show that this lower bound on $\sin\theta_k$ actually implies that $\lambda_{k+1}(X^TX)$ will be large.

To see this, let us write $XX^T = U_k\Lambda_kU_k^T + R$ where the $n\times k$ matrix of eigenvectors U_k is defined as in Lemma 12, Λ_k is the corresponding $k\times k$ diagonal matrix of eigenvalues, and $U_k^TR = 0$. Denote $\Pi_k = I - U_kU_k^T$ and note that

$$\|\Pi_k(XX^T - M^\star)\Pi_k\|_F = \|\Pi_kXX^T\Pi_k - \Pi_kM^\star\Pi_k\|_F = \|R - \Pi_kM^\star\Pi_k\|_F.$$

By the subaddivity of the norm $\|R - \Pi_kM^\star\Pi_k\|_F \le \|R\|_F + \|\Pi_kM^\star\Pi_k\|_F$. Dividing both sides by $\|E\|_F$ yields

$$\sin\theta_k = \frac{\|R - \Pi_kM^\star\Pi_k\|_F}{\|E\|_F} \le \frac{\|\Pi_kM^\star\Pi_k\|_F}{\|E\|_F} + \frac{\|R\|_F}{\|E\|_F}.$$

Since $\rho \le 1/\sqrt{2}$ by assumption, Lemma 13 yields

$$\frac{\|\Pi_kM^\star\Pi_k\|_F}{\|E\|_F} \le \frac{1}{\sqrt{2}}\frac{\rho}{\sqrt{1-\rho^2}} \le \rho.$$

In addition,

$$\|R\|_F \le \|R\| \cdot \sqrt{\text{rank}(R)} = \lambda_{k+1}(XX^T)\cdot\sqrt{r-k}.$$

Combining the two inequalities above we get

$$\sqrt{1-\mu^2} \le \sin\theta_k \le \frac{1}{\sqrt{2}}\frac{\rho}{\sqrt{1-\rho^2}} + \sqrt{r-k}\cdot\frac{\lambda_{k+1}\left(X^TX\right)}{\|E\|_F}.$$

Rearranging, we get

$$\frac{\lambda_{k+1}\left(X^TX\right)}{\|E\|_F} \geq \frac{1}{\sqrt{r-k}}\left(\sqrt{1-\mu^2}-\frac{1}{\sqrt{2}}\frac{\rho}{\sqrt{1-\rho^2}}\right) \geq \left(1-\frac{1}{\sqrt{2}}\right)\sqrt{\frac{1-\mu^2}{r-k}}.$$

Note that the last inequality above follows from the assumption that $\frac{\rho^2}{1-\rho^2} \leq 1-\mu^2$. Now substituting $\eta \leq C_{\text{ub}}\|XX^T - M^\star\|_F$ and $r-k \leq r-r^\star$ and noting that $\left(1-\frac{1}{\sqrt{2}}\right) \leq 1/3$ we get

$$\frac{\eta}{\lambda_{k+1}(X^TX)} \leq C_{\text{ub}}\frac{\|XX^T-M^\star\|_F}{\lambda_{k+1}(X^TX)} \leq 3C_{\text{ub}}\sqrt{\frac{r-k}{1-\mu^2}} \leq 3C_{\text{ub}}\sqrt{\frac{r-r^\star}{1-\mu^2}}, \tag{41}$$

which shows that $\lambda_{k+1}(X^TX)$ is large.

If $\cos\theta_{k+1} \geq \mu$ is also large, then substituting (41) into (38) yields gradient dominance

$$\frac{\|\nabla f(X)\|_{P^*}^2}{f(X)} \geq \frac{(\cos\theta_{k+1}-\delta)^2}{1+\eta/\lambda_{k+1}^2(X)} \geq (\mu-\delta)^2\left(1+3C_{\text{ub}}\sqrt{\frac{r-r^\star}{1-\mu^2}}\right)^{-1}, \tag{42}$$

and this yields the second term in (37) so we are done. If $\cos\theta_{k+1} < \mu$ then we can simply repeat the argument above to show that $\lambda_{k+1}(X^TX)$ is large. We can repeat this process until $k+1 = r$. At this point, we have

$$\cos^2\theta_r = 1-\sin^2\theta_r \geq 1-\frac{1}{2}\frac{\rho^2}{1-\rho^2} \geq \mu^2$$

where we used our hypothesis $1-\mu^2 \geq \frac{\rho^2}{1-\rho^2} \geq \frac{1}{2}\frac{\rho^2}{1-\rho^2}$, and substituting (41) into (38) again yields gradient dominance in (42). □

D.1 Proof of Gradient Lower Bound (Lemma 12)

In this section we prove Lemma 12, where we prove gradient dominance $\|\nabla f(X)\|_{P^*}^2 \geq \mu_P f(X)$ with a PL constant μ_P that is proportional to $\cos\theta_k - \delta$ and to $\lambda_k(X^TX)/\eta$. We first prove the following result which will be useful in the proof of Lemma 12.

Lemma 14. *Let $\mathcal{A}$ satisfy RIP with parameters (ζ, δ), where $\zeta = \operatorname{rank}([X, Z])$. Then, we have*

$$\|\nabla f(X)\|_{P*} \geq \max_{\|Y\|_P \leq 1}\langle XY^T + YX^T, E\rangle - \delta\|XY^T+YX^T\|_F\|E\|_F \tag{43}$$

Proof. Let Y maximize the right-hand side of (43) and let W be the matrix corrresponding to the orthogonal projection onto $\operatorname{range}(X)+\operatorname{range}(Y)$. Set $\tilde{Y} = WY$, then

$$\langle X\tilde{Y}^T + \tilde{Y}X^T, E\rangle = \langle XY^T, EW\rangle + \langle YX^T, WE\rangle = \langle XY^T+YX^T, E\rangle.$$

On the other hand, we have

$$\|X\tilde{Y}^T+\tilde{Y}X^T\|_F = \|W\left(XY^T+YX^T\right)W\|_F \leq \|XY^T+YX^T\|_F$$

and

$$\|\tilde{Y}\|_P = \|WYP^{1/2}\|_F \leq \|YP^{1/2}\|_F = \|Y\|_P.$$

This means that $\tilde{Y}$ is feasible and makes the right-hand side at least as large as Y. Since Y is the maximizer by definition, we conclude that $\tilde{Y}$ also maximizes the right-hand side of (43).

By definition, $\operatorname{range}(\tilde{Y}) \subset \operatorname{range}(X)+\operatorname{range}(Z)$, so $(2r,\delta)$-RIP implies

$$|\langle A(X\tilde{Y}^T+\tilde{Y}X^T), A(E)\rangle - \langle X\tilde{Y}^T+\tilde{Y}X^T, E\rangle| \leq \delta\|X\tilde{Y}^T+\tilde{Y}X^T\|_F\|E\|_F.$$

Now we have

$$\begin{aligned}\|\nabla f(X)\|_{P*} &= \max_{\|Y\|_P\leq 1}\langle\mathcal{A}(XY^T+YX^T), \mathcal{A}(E)\rangle\\ &\geq \langle\mathcal{A}(X\tilde{Y}^T+\tilde{Y}X^T), \mathcal{A}(E)\rangle\\ &\geq \langle X\tilde{Y}^T+\tilde{Y}X^T, E\rangle - \delta\|X\tilde{Y}^T+\tilde{Y}X^T\|_F\|E\|_F\\ &= \max_{\|Y\|_P\leq 1}\langle XY^T+YX^T, E\rangle - \delta\|XY^T+YX^T\|_F\|E\|_F.\end{aligned}$$

This completes the proof. □

Proof of Lemma 12. Let $X = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ with $\|u_i\| = \|v_i\| = 1$ and $\sigma_1 \geq \cdots \geq \sigma_r$ denote the usual singular value decomposition. Observe that the preconditioned Jacobian $\mathbf{J}\mathbf{P}^{-1/2}$ satisfies

$$\mathbf{J}\mathbf{P}^{-1/2}\mathrm{vec}(Y) = \mathrm{vec}(XP^{-1/2}Y^T + YP^{-1/2}X^T) = \mathrm{vec}\left(\sum_{i=1}^{r} \frac{u_i y_i^T + y_i u_i^T}{\sqrt{1+\eta/\sigma_i^2}}\right)$$

where $y_i = Yv_i$. This motivates the following family of singular value decompositions

$$\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T\mathrm{vec}(Y) = \mathrm{vec}\left(\sum_{i=1}^{k} \frac{u_i y_i^T + y_i u_i^T}{\sqrt{1+\eta/\sigma_i^2}}\right) \text{ for all } k \in \{1,2,\dots,r\}, \quad \mathbf{J}\mathbf{P}^{-1/2} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T. \tag{44}$$

Here, the $n^2 \times \zeta_k$ matrix $\mathbf{U}_k$ and the $nr \times \zeta_k$ matrix $\mathbf{V}_k$ have orthonormal columns, and the rank can be verified as $\zeta_k = nk - k(k-1)/2 < nr \leq n^2$. Now, we rewrite Lemma 14 by vectorizing $y = \mathrm{vec}(Y)$ and writing

$$\begin{aligned}\|\nabla f(X)\|_{P*} &\geq \max_{\|\mathbf{P}^{1/2}y\|\leq 1}\left(\frac{\mathbf{e}^T\mathbf{J}y}{\|\mathbf{e}\|\|\mathbf{J}y\|} - \delta\right)\|\mathbf{e}\|\|\mathbf{J}y\| \overset{\text{(a)}}{=} \max_{\|y'\|\leq 1}\left(\frac{\mathbf{e}^T\mathbf{J}\mathbf{P}^{-1/2}y}{\|\mathbf{e}\|\|\mathbf{J}\mathbf{P}^{-1/2}y\|} - \delta\right)\|\mathbf{e}\|\|\mathbf{J}\mathbf{P}^{-1/2}y\| \\ &\overset{\text{(b)}}{=} \max_{\|y'\|\leq 1}\left(\frac{\mathbf{e}^T\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T y}{\|\mathbf{e}\|\|\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T y\|} - \delta\right)\|\mathbf{e}\|\|\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T y\| \\ &\overset{\text{(c)}}{\geq} \left(\frac{\mathbf{e}^T\mathbf{U}_k\mathbf{U}_k^T\mathbf{e}}{\|\mathbf{e}\|\|\mathbf{U}_k^T\mathbf{e}\|} - \delta\right)\|\mathbf{e}\|\frac{\|\mathbf{U}_k^T\mathbf{e}\|}{\|\mathbf{\Sigma}_k^{-1}\mathbf{U}_k^T\mathbf{e}\|} \overset{\text{(d)}}{\geq} \left(\frac{\|\mathbf{U}_k^T\mathbf{e}\|}{\|\mathbf{e}\|} - \delta\right)\|\mathbf{e}\|\lambda_{\min}(\mathbf{\Sigma}_k).\end{aligned}$$

Step (a) makes a change of variables $y \leftarrow \mathbf{P}^{1/2}y$; Step (b) substitutes (44); Step (c) substitutes the heuristic choice $y = d/\|d\|$ where $d = \mathbf{V}_k\mathbf{\Sigma}_k^{-1}\mathbf{U}_k^T\mathbf{e}$; Step (d) notes that $\mathbf{e}^T\mathbf{U}_k\mathbf{U}_k^T\mathbf{e} = \|\mathbf{U}_k^T\mathbf{e}\|^2$ and that $\|\mathbf{\Sigma}_k^{-1}\mathbf{U}_k^T\mathbf{e}\| \leq \|\mathbf{U}_k^T\mathbf{e}\| \cdot \lambda_{\max}(\mathbf{\Sigma}_k^{-1}) = \|\mathbf{U}_k^T\mathbf{e}\|/\lambda_{\min}(\mathbf{\Sigma}_k)$. Finally, we can mechanically verify from (44) that

$$\cos^2\theta_k \overset{\text{def}}{=} \frac{\|\mathbf{U}_k^T\mathbf{e}\|^2}{\|\mathbf{e}\|^2} = 1 - \frac{\|(I - \mathbf{U}_k^T\mathbf{U}_k^T)\mathbf{e}\|^2}{\|\mathbf{e}\|^2} = 1 - \frac{\|(I - U_kU_k^T)E(I - U_kU_k^T)\|_F^2}{\|E\|_F^2}$$

where $U_k = [u_1, \dots, u_k]$, and that

$$\lambda_{\min}^2(\mathbf{\Sigma}_k) = \min_{\|y_k\|=1}\left\|\frac{u_k y_k^T + y_k u_k^T}{\sqrt{1+\eta/\sigma_k^2}}\right\|_F^2 = \min_{\|y_k\|=1}\frac{2\|u_k\|^2\|y_k\|^2 + 2(u_k^Ty_k)^2}{1+\eta/\sigma_k^2} = \frac{2}{1+\eta/\sigma_k^2}.$$

□

D.2 Proof of Basis Alignment (Lemma 13)

Before we prove this lemma, we make two observations that simplifies the proof. First, even though our goal is to prove the inequality (36) for all $k \geq r^*$, it actually suffices to consider the case $k = r^*$. This is because the numerator $\|Z^T(I - U_kU_k^T)Z\|_F$ decreases monotonically as k increases. Indeed, for any $k \geq r^\star$, define VV^T as below

$$I - U_kU_k^T = I - U_{r^\star}U_{r^\star}^T - VV^T = (I - U_{r^\star}U_{r^\star}^T)(I - VV^T) = (I - VV^T)(I - U_{r^\star}U_{r^\star}^T).$$

Then, we have

$$\begin{aligned}\|Z^T(I - U_kU_k^T)Z\|_F &= \|(I - U_kU_k^T)ZZ^T(I - U_kU_k^T)\|_F \\ &= \|(I - VV^T)(I - U_{r^\star}U_{r^\star}^T)ZZ^T(I - U_{r^\star}U_{r^\star}^T)(I - VV^T)\|_F \\ &\leq \|(I - U_{r^\star}U_{r^\star}^T)ZZ^T(I - U_{r^\star}U_{r^\star}^T)\|_F.\end{aligned}$$

Second, due to the rotational invariance of this problem, we can assume without loss of generality that X, Z are of the form

$$X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}, \; Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}. \tag{45}$$

where $X_1 \in \mathbb{R}^{k\times k}$, $Z_1 \in \mathbb{R}^{k\times r^\star}$ and $\sigma_{\min}(X_1) \geq \sigma_{\max}(X_2)$. (Concretely, we compute the singular value decomposition $X = USV^T$ with $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{r\times r}$, and then set $X \leftarrow U^TXV$ and $Z \leftarrow U^TZ$.) We first need to show that as XX^T approaches ZZ^T, the dominant directions of X must align with Z in a way as to make the Z_2 portion of Z go to zero.

Lemma 15. *Suppose that X, Z are in the form in (45), and $k \geq r^\star$. If $\|XX^T - ZZ^T\|_F \leq \rho\lambda_{\min}(Z^TZ)$ and $\rho^2 < 1/2$, then $\lambda_{\min}(Z_1^TZ_1) \geq \lambda_{\max}(Z_2^TZ_2)$.*

Proof. Denote $\gamma = \lambda_{\min}(Z_1^TZ_1)$ and $\beta = \lambda_{\max}(Z_2^TZ_2)$. We will assume $\gamma < \beta$ and prove that $\rho^2 \geq 1/2$, which contradicts our hypothesis. The claim is invariant to scaling of X and Z, so we assume without loss of generality that $\lambda_{\min}(Z^TZ) = 1$. Our radius hypothesis then reads

$$\begin{aligned}\|XX^T - ZZ^T\|_F^2 &= \left\| \begin{bmatrix} X_1X_1^T - Z_1Z_1^T & -Z_1Z_2^T \\ -Z_2Z_1^T & X_2X_2^T - Z_2Z_2^T \end{bmatrix} \right\|_F^2 \\ &= \|X_1X_1^T - Z_1Z_1^T\|_F^2 + 2\langle Z_1^TZ_1, Z_2^TZ_2\rangle + \|X_2X_2^T - Z_2Z_2^T\|_F^2 \leq \rho^2.\end{aligned}$$

Now, we optimize over X_1 and X_2 to minimize the left-hand side. Recall by construction in (45) we restricted $\sigma_{\min}(X_1) \geq \sigma_{\max}(X_2)$. Accordingly, we consider

$$\min_{X_1,X_2} \left\{ \|X_1X_1^T - Z_1Z_1^T\|_F^2 + \|X_2X_2^T - Z_2Z_2^T\|_F^2 : \lambda_{\min}(X_1X_1^T) \geq \lambda_{\max}(X_2X_2^T) \right\}. \tag{46}$$

We relax $X_1X_1^T$ and $X_2X_2^T$ into positive semidefinite matrices

$$(46) \geq \min_{S_1 \succeq 0, S_2 \succeq 0} \{\|S_1 - Z_1Z_1^T\|_F^2 + \|S_2 - Z_2Z_2^T\|_F^2 : \lambda_{\min}(S_1) \geq \lambda_{\max}(S_2)\} \tag{47}$$

The equation above is invariant to a change of basis for both S_1 and S_2, so we change the basis of S_1 and S_2 into the eigenbases of $Z_1Z_1^T$ and $Z_2Z_2^T$ to yield

$$(47) = \min_{s_1 \geq 0, s_2 \geq 0} \{\|s_1 - \lambda(Z_1Z_1^T)\|^2 + \|s_2 - \lambda(Z_2Z_2^T)\|^2 : \min(s_1) \geq \max(s_2)\} \tag{48}$$

where $\lambda(Z_1Z_1^T) \geq 0$ and $\lambda(Z_2Z_2^T) \geq 0$ are the vector of eigenvalues. We lower-bound (48) by dropping all the terms in the sum of squares except the one associated with $\lambda_{\min}(Z_1^TZ_1)$ and $\lambda_{\max}(Z_2Z_2^T)$ to obtain

$$\begin{aligned}(48) &\geq \min_{d_1,d_2 \in \mathbb{R}_+} \{[d_1 - \lambda_{\min}(Z_1^TZ_1)]^2 + [d_2 - \lambda_{\max}(Z_2Z_2^T)]^2 : d_1 \geq d_2\} &(49)\\ &= \min_{d_1,d_2 \in \mathbb{R}_+} \{[d_1 - \gamma]^2 + [d_2 - \beta]^2 : d_1 \geq d_2\} = (\gamma - \beta)^2/2, &(50)\end{aligned}$$

where we use the fact that $\gamma < \beta$ to argue that $d_1 = d_2$ at optimality. Now we have

$$\begin{aligned}\rho^2 &\geq \|X_1X_1^T - Z_1Z_1^T\|_F^2 + \|X_2X_2^T - Z_2Z_2^T\|_F^2 + 2\langle Z_1^TZ_1, Z_2^TZ_2\rangle \\ &\geq \|X_1X_1^T - Z_1Z_1^T\|_F^2 + \|X_2X_2^T - Z_2Z_2^T\|_F^2 + 2\lambda_{\min}(Z_1^TZ_1)\lambda_{\max}(Z_2^TZ_2) \\ &\geq \min_{d_1,d_2 \in \mathbb{R}_+} \{[d_1 - \gamma]^2 + [d_2 - \beta]^2 : d_1 \geq d_2\} + 2\gamma\beta \\ &\geq \frac{(\gamma - \beta)^2}{2} + 2\gamma\beta = \frac{1}{2}(\gamma + \beta)^2.\end{aligned}$$

Finally, note that

$$\gamma + \beta = \lambda_{\min}(Z_1^TZ_1) + \lambda_{\max}(Z_2^TZ_2) \geq \lambda_{\min}(Z_1^TZ_1 + Z_2^TZ_2) = \lambda_{\min}(Z^TZ) = 1.$$

Therefore, we have $\rho^2 \geq 1/2$, a contradiction. This completes the proof. □

Now we are ready to prove Lemma 13.

Proof. As before, assume with out loss of generality that X, Z are of the form (45). From the proof of Lemma 15 we already know

$$\|XX^T - ZZ^T\|_F^2 = \|X_1X_1^T - Z_1Z_1^T\|_F^2 + 2\langle Z_1^TZ_1, Z_2^TZ_2\rangle + \|X_2X_2^T - Z_2Z_2^T\|_F^2.$$

Moreoever, we can compute

$$\|Z^T(I - U_kU_k^T)Z\|_F = \left\| \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}^T \left(I - \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} \right) \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \right\|_F = \|Z_2^TZ_2\|_F = \|Z_2Z_2^T\|_F. \tag{51}$$

We will show that in the neighborhood $\|XX^T - ZZ^T\| \le \rho\lambda_{\min}(Z^TZ)$ that

$$\rho \le 1/\sqrt{2} \implies \sin\phi \overset{\text{def}}{=} \|(I - U_kU_k^T)Z\|_F/\sigma_k(Z) = \|Z_2\|_F/\sigma_{r^\star}(Z) \le \rho. \tag{52}$$

Then we obtain

$$\begin{aligned}\frac{\|Z_2Z_2^T\|_F^2}{\|XX^T - ZZ^T\|^2} &\overset{\text{(a)}}{\le} \frac{\|Z_2\|_F^4}{2\langle Z_1^TZ_1, Z_2^TZ_2\rangle} \overset{\text{(b)}}{\le} \frac{\|Z_2\|_F^4}{2\lambda_{\min}(Z_1^TZ_1)\|Z_2\|_F^2}\\ &\overset{\text{(c)}}{\le} \frac{\|Z_2\|_F^2}{2[\lambda_{\min}(Z^TZ) - \|Z_2\|_F^2]} = \frac{\sin^2\phi}{2[1-\sin^2\phi]} \qquad (53)\\ &\le \frac{1}{2}\frac{\rho^2}{1-\rho^2}. \qquad (54)\end{aligned}$$

Step (a) bounds the numerator as $\|Z_2Z_2^T\|_F \le \|Z_2\|_F^2$ and uses the fact that the denominator is greater than $2\langle Z_1^TZ_1, Z_2^TZ_2\rangle$. Step (b) follows from the inequality $\langle Z_1^TZ_1, Z_2^TZ_2\rangle \ge \lambda_{\min}(Z_1^TZ_1)\|Z_2Z_2^T\|_F$. Finally, step (c) bounds the minimum eigenvalue of $Z_1^TZ_1$ by noting that

$$\begin{aligned}\lambda_{\min}(Z_1^TZ_1) &= \lambda_{\min}(Z_1^TZ_1 + Z_2^TZ_2 - Z_2^TZ_2)\\ &\ge \lambda_{\min}(Z_1^TZ_1 + Z_2^TZ_2) - \lambda_{\max}(Z_2^TZ_2)\\ &\ge \lambda_{\min}(Z^TZ) - \|Z_2\|_F^2, \qquad (55)\end{aligned}$$

where the last line bounds the operator norm of Z_2 with the Frobenius norm.

To prove (52), we know from Lemma 15 that $\rho \le 1/\sqrt{2}$ implies that $\lambda_{\min}(Z_1^TZ_1) \ge \lambda_{\max}(Z_2^TZ_2)$. This implies $\lambda_{\min}(Z_1^TZ_1) \ge \frac{1}{2}\lambda_{\min}(Z^TZ)$, since

$$2\lambda_{\min}(Z_1^TZ_1) \ge \lambda_{\min}(Z_1^TZ_1) + \lambda_{\max}(Z_2^TZ_2) \ge \lambda_{\min}(Z^TZ)$$

This implies the following

$$\begin{aligned}\|XX^T - ZZ^T\|_F^2 &= \|X_1X_1^T - Z_1Z_1^T\|_F^2 + 2\langle Z_1^TZ_1, Z_2^TZ_2\rangle + \|X_2X_2^T - Z_2Z_2^T\|_F^2\\ &\ge 2\langle Z_1^TZ_1, Z_2^TZ_2\rangle \ge 2\lambda_{\min}(Z_1^TZ_1)\|Z\|_F^2 \ge \lambda_{\min}(Z^TZ)\|Z\|_F^2\end{aligned}$$

and we have therefore

$$\rho^2\lambda_{\min}^2(Z^TZ) \ge \|XX^T - ZZ^T\|_F^2 \ge \lambda_{\min}(Z^TZ)\|Z\|_F^2 \ge \lambda_{\min}(Z^TZ)\|Z_2\|_F^2$$

which this proves $\sin^2\phi = \|Z_2\|_F^2/\lambda_{\min}(Z^TZ) \le \rho^2$ as desired. □

E Preliminaries for the Noisy Case

E.1 Notations

In the following sections, we extend our proofs to the noisy setting. As before, we denote by $M^\star = ZZ^T \in \mathbb{R}^{n\times n}$ our ground truth. Our measurements are of the form $y = \mathcal{A}(ZZ^T) + \epsilon \in \mathbb{R}^m$. We make the standard assumption that the noise vector $\epsilon \in \mathbb{R}^m$ has sub-Gaussian entries with zero mean and variance $\sigma^2 = \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\epsilon_i^2]$.

In this case, the objective function can be written as

$$f(X) = \frac{1}{m}\|\mathcal{A}(XX^T) - y\|^2 = f_c(X) + \frac{1}{m}\|\epsilon\|^2 - \frac{2}{m}\langle\mathcal{A}(XX^T - M^\star), \epsilon\rangle,$$

where $f_c(X) = \frac{1}{m}\|\mathcal{A}(XX^T - M^\star)\|^2$ is the objective function with clean measurements that are not corrupted with noise. Note that compared to the noiseless case, we have rescaled our objective by a factor of $1/m$ to emphasize the number of measurements m.

Moreover, we say that an event $\mathcal{E}$ happens with overwhelming or high probability, if its probability of occurrence is at least $1 - cn^{-c'}$, for some $0 < c, c' < \infty$. Moreover, to streamline the presentation, we omit the statement "with high or overwhelming probabily" if it is implied by the context.

We make a few simplifications on notations. As before, we will use α to denote the step-size and D to denote the local search direction. We will use lower case letters x and d to refer to $\mathrm{vec}(X)$ and $\mathrm{vec}(D)$ respectively.

Similarly, we will write $f(x) \in \mathbb{R}^{nr}$ and $\nabla f(x) \in R^{nr}$ as the vectorized versions of $f(X)$ and its gradient. This notation is also used for $f_c(X)$. As before, we define $P = X^T X + \eta I_r$ and $\mathbf{P} = (X^T X + \eta I_r) \otimes I_n$. For the vectorized version of the gradient, we simply define its P-norm (and P^*-norm) to be the same as the matrix version, that is,

$$\|\nabla f(x)\|_P = \|\nabla f(X)\|_P, \qquad \|\nabla f(x)\|_{P^*} = \|\nabla f(X)\|_{P^*}.$$

We drop the iteration index k from our subsequent analysis, and refer to x_{k+1} and x_k as $\tilde{x}$ and x, respectively. Thus, with noisy measurements, the iterations of PrecGD take the form

$$X_{k+1} = X_k - \alpha \nabla f(X_k)(X_k^T X_k)^{-1}.$$

The *vectorized* version of the gradient update above can be written as $\tilde{x} = x - \alpha d$, where

$$\begin{aligned} d = \mathrm{vec}(\nabla f(X) P^{-1}) &= \mathrm{vec}\left(f_c(X) + \frac{1}{m}\|\epsilon\|^2 - \frac{2}{m}\langle \mathcal{A}(XX^T - M^\star), \epsilon\rangle \right) \\ &= \mathbf{P}^{-1}\nabla f_c(x) - \frac{2}{m}\mathbf{P}^{-1}\left(I_r \otimes \sum_{i=1}^m \epsilon_i A_i \right) x. \end{aligned} \tag{56}$$

Inspired by the variational representation of the Frobenius norm, for any matrix $H \in \mathbb{R}^{n\times n}$ we define its *restricted Frobenius norm* as

$$\|H\|_{F,r} = \arg \max_{Y \in S_n^+, \mathrm{rank}(Y) \le r} \langle H, Y \rangle, \tag{57}$$

where S_n^+ is the set of $n \times n$ positive semidefinite matrices. It is easy to verify that $\|H\|_F = \|H\|_{F,n}$ and $\|H\|_{F,r} = \sqrt{\sum_{i=1}^r \sigma_i(H)^2}$.

For any two real numbers $a, b \in R$, we say that $a \asymp b$ if there exists some constant C_1, C_2 such that $C_1 b \le a \le C_2 b$. Through out the section we will use one symbol C to denote constants that might differ.

Finally, we also recall that μ_P, which is used repeatedly in this section, is the constant defined in (33).

E.2 Auxiliary Lemmas

Now we present a few auxiliary lemmas that we will use for the proof of the noisy case. At the core of our subsequent proofs is the following standard concentration bound.

Lemma 16. *Suppose that the number of measurements satisfies $m \gtrsim \sigma n \log n$. Then, with high probability, we have*

$$\frac{1}{m}\left\| \sum_{i=1}^m A_i \epsilon_i \right\|_2 \lesssim \sqrt{\frac{\sigma^2 n \log n}{m}},$$

where $\|\cdot\|_2$ denotes the operator norm of a matrix.

Lemma 16 will be used extensively in the proofs of Proposition 6, and Theorems 7 and 8.

Our strategy for establishing linear convergence is similar to that of the noiseless case. Essentially, our goal is to show that with an appropriate step-size, there is sufficient decrement in the objective value in terms of $\|\nabla f_c(X)\|_{P^*}$. Then applying Theorem 4 will result in the desired convergence rate.

In the noiseless case, we proved a Lipschitz-like inequality (Lemma 2) and bounded the Lipschitz constant above in a neighborhood around the ground truth. Similar results hold in the noisy case. However, because of the noise, it will be easier to directly work with the quartic polynomial $f_c(X - \alpha D)$ instead. In particular, we have the following lemma that characterizes how much progress we make by taking a step in the direction D.

Lemma 17. *For any descent direction $D \in \mathbb{R}^{n\times r}$ and step-size $\alpha > 0$ we have*

$$f_c(X - \alpha D) \le f_c(X) - \alpha \nabla f_c(X)^T D + \frac{\alpha^2}{2} D^T \nabla^2 f_c(X) D \tag{58}$$

$$+ \frac{(1+\delta)\alpha^3}{m}\|D\|_F^2 \left(2\|DX^T + XD^T\|_F + \alpha \|D\|_F^2 \right). \tag{59}$$

Proof. Directly expanding the quadratic $f_c(X-\alpha D)$, we get

$$\begin{aligned}
f_c(X-\alpha D) &= \frac{1}{m}\|\mathcal{A}((X-\alpha D)(X-\alpha D)^T - M^\star)\|^2 \\
&= \frac{1}{m}\|\mathcal{A}(XX^T - M^\star)\|^2 - \frac{2\alpha}{m}\langle \mathcal{A}(XX^T - M^\star), \mathcal{A}(XD^T + DX^T)\rangle \\
&\quad + \frac{\alpha^2}{m}\left[2\langle \mathcal{A}(XX^T - M^\star), \mathcal{A}(DD^T)\rangle + \|\mathcal{A}(XD^T + DX^T)\|^2\right] \\
&\quad - \frac{2\alpha^3}{m}\langle \mathcal{A}(XD^T + DX^T), \mathcal{A}(DD^T)\rangle + \frac{\alpha^4}{m}\|\mathcal{A}(DD^T)\|^2.
\end{aligned}$$

We bound the third- and fourth- order terms

$$\begin{aligned}
|\langle \mathcal{A}(XD^T + DX^T), \mathcal{A}(DD^T)\rangle| &\overset{\text{(a)}}{\leq} \|\mathcal{A}(XD^T + DX^T)\|\|\mathcal{A}(DD^T)\rangle\| \\
&\overset{\text{(b)}}{\leq} (1+\delta)\|XD^T + DX^T\|_F\|DD^T\|_F \\
&\overset{\text{(c)}}{\leq} (1+\delta)\|XD^T + DX^T\|_F\|D\|_F^2
\end{aligned}$$

and

$$\|\mathcal{A}(DD^T)\|^2 \overset{\text{(b)}}{\leq} (1+\delta)\|DD^T\|_F^2 \overset{\text{(c)}}{\leq} (1+\delta)\|D\|_F^4,$$

Step (a) uses the Cauchy–Schwarz inequality; Step (b) applies $(\delta, 2r)$-RIP; Step (c) bounds $\|DD^T\|_F \leq \|D\|_F^2$. Summing up these inequalities we get the desired result. □

It turns out that in our proofs it will be easier to work with the *vectorized* version of (59), which we can write as

$$f_c(x-\alpha d) \leq f_c(x) - \alpha\nabla f_c(x)^T d + \frac{\alpha^2}{2}d^T\nabla^2 f_c(x)d + \frac{(1+\delta)\alpha^3}{m}\|d\|^2\left(2\|\mathbf{J}_X d\| + \alpha\|d\|^2\right), \quad (60)$$

where we recall that $J_X : \mathbb{R}^{nr} \to \mathbb{R}^{n^2}$ is the linear operator that satisfies $J_X d = \mathrm{vec}(XD^T + DX^T)$.

Now we proceed to bound the higher-order terms in the Taylor-like expansion above.

Lemma 18 (Second-order term)**.** *We have*

$$\sigma_{\max}(\mathbf{P}^{-1/2}\nabla^2 f_c(x)\mathbf{P}^{-1/2}) \leq \frac{2(1+\delta)}{m}\left(\frac{8\sigma_r^2(X) + \|XX^T - ZZ^T\|_F}{\sigma_r^2(X)+\eta}\right).$$

Proof. For any $v \in \mathbb{R}^{nr}$ where $v = \mathrm{vec}(V)$, we have

$$\begin{aligned}
m \cdot v^T\nabla^2 f_c(x)v &= 4\langle \mathcal{A}(XX^T - ZZ^T), \mathcal{A}(VV^T) + 2\|\mathcal{A}(XV^T + VX^T)\|^2 \\
&\leq 4\|\mathcal{A}(XX^T - ZZ^T)\|\|\mathcal{A}(VV^T)\| + 2\|\mathcal{A}(XV^T + VX^T)\|^2 \\
&\leq 2(1+\delta)\left(\|XX^T - ZZ^T\|_F\|VV^T\|_F + 2\|XV^T + VX^T\|_F^2\right)
\end{aligned}$$

Now, let $v = \mathbf{P}^{-1/2}u$ for $u = \mathrm{vec}(U)$. Then, $V = UP^{-1/2}$ and

$$\|VV^T\|_F = \|UP^{-1}U^T\|_F \leq \sigma_{\max}(P^{-1})\|U\|_F^2 = \frac{\|U\|_F^2}{\sigma_r^2(X)+\eta}.$$

Also, $\|XV^T + VX^T\|_F \leq 2\|XV^T\|_F$ and

$$\|XV^T\| = \|XP^{-1/2}U^T\| \leq \sigma_{\max}(XP^{-1/2})\|U\|_F = \left(\frac{\sigma_r^2(X)}{\sigma_r^2(X)+\eta}\right)^{1/2}\|U\|_F.$$

Since $\|u\| = \|U\|_F$, it follows that

$$u^T\mathbf{P}^{-1/2}\nabla^2 f_c(x)\mathbf{P}^{-1/2}u \leq \frac{2(1+\delta)}{m}\left(\frac{8\sigma_r^2(X) + \|XX^T - ZZ^T\|}{\sigma_r^2(X)+\eta}\right)\|u\|^2,$$

which gives the desired bound on the largest singular value. □

The following lemma gives a bound on the third- and fourth-order terms in (60).

Lemma 19. *Set $d = \mathbf{P}^{-1}\nabla f_c(x)$, then we have $\|\mathbf{J}d\|^2 \le 8m^2\|\nabla f_c(x)\|_{P^*}^2$ and $\|d\|^2 \le \|\nabla f_c(x)\|_{P^*}^2/\eta$.*

Proof. We have

$$\begin{aligned}\|\mathbf{J}_X d\|^2 &= \|\mathcal{A}(XD^T + DX^T)\|^2 \le (1+\delta)\|XD^T + DX^T\|^2\\ &= (1+\delta)\|\mathbf{J}_X d\|^2 = m^2(1+\delta)\|\mathbf{J}\mathbf{P}^{-1}\nabla f_c(x)\|^2\\ &\le m^2(1+\delta)\sigma_{\max}^2(\mathbf{J}\mathbf{P}^{-1/2})\|\mathbf{P}^{-1/2}\nabla f_c(x)\|^2\\ &= 4m^2(1+\delta)\frac{\sigma_r^2}{\sigma_r^2+\eta}\|\nabla f_c(x)\|_{P^*}^2 \le 8m^2\|\nabla f_c(x)\|_{P^*}^2\end{aligned}$$

and

$$\begin{aligned}\|d\|^2 = \|\mathbf{P}^{-1}\nabla f_c(x)\|^2 &\le \sigma_{\max}(\mathbf{P}^{-1})\|\mathbf{P}^{-1/2}\nabla f_c(x)\|^2\\ &= \frac{1}{\sigma_r^2+\eta}\|\nabla f(x)\|_{P^*}^2 \le \|\nabla f(x)\|_{P^*}^2/\eta.\end{aligned}$$

□

F Proof of Noisy Case with Optimal Damping Parameter

Now we are ready to prove Theorem 7, which we restate below for convenience.

Theorem 20 (Noisy measurements with optimal η)**.** *Suppose that the noise vector $\epsilon \in \mathbb{R}^m$ has sub-Gaussian entries with zero mean and variance $\sigma^2 = \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\epsilon_i^2]$. Moreover, suppose that $\eta_k = \frac{1}{\sqrt{m}}\|\mathcal{A}(X_kX_k^T - M^*)\|$, for $k = 0, 1, \dots, K$, and that the initial point X_0 satisfies $\|\mathcal{A}(X_0X_0^T - M^*)\|^2 < \rho^2(1-\delta)\lambda_{r^*}(M^\star)^2$. Consider $k^* = \arg\min_k \eta_k$, and suppose that $\alpha \le 1/L$, where $L > 0$ is a constant that only depends on δ. Then, with high probability, we have*

$$\|X_{k^*}X_{k^*}^T - M^\star\|_F^2 \lesssim \max\left\{\frac{1+\delta}{1-\delta}\left(1-\alpha\frac{\mu_P}{2}\right)^K\|X_0X_0^T - M^*\|_F^2, \mathcal{E}_{stat}\right\}, \tag{61}$$

where $\mathcal{E}_{stat} := \frac{\sigma^2 nr\log n}{\mu_P(1-\delta)m}$.

Proof. **Step I. Using Lemma 17 to establish sufficient decrement.**

First, we write out the vectorized version of Lemma 60:

$$f_c(x-\alpha d) \le f_c(x) - \alpha\nabla f_c(x)^T d + \frac{\alpha^2}{2}d^T\nabla^2 f_c(x)d + \frac{(1+\delta)\alpha^3}{m}\|d\|^2\left(2\|\mathbf{J}_X d\| + \alpha\|d\|^2\right). \tag{62}$$

To simplify notation, we define the error term $\mathbb{E}(x) = \frac{2}{m}\left(I_r \otimes \sum_{i=1}^m \epsilon_i A_i\right)x$, so that the search direction (56) can be rewritten as $d = \mathbf{P}^{-1}(\nabla f_c(x) - \mathbb{E}(x))$.

Now plugging this d into (62) yields

$$f_c(x-\alpha d) \le f_c(x) - \alpha\|\nabla f_c(x)\|_{P*}^2 + T_1 + T_2 + T_3$$

where

$$\begin{aligned}T_1 =& \alpha\nabla f_c(x)^T\mathbf{P}^{-1}\mathbb{E}(x)\\ T_2 =& \frac{\alpha^2}{2}\Big(\nabla f_c(x)^T\mathbf{P}^{-1}\nabla^2 f_c(x)\mathbf{P}^{-1}\nabla f_c(x) + \mathbb{E}(x)^T\mathbf{P}^{-1}\nabla^2 f_c(x)\mathbf{P}^{-1}\mathbb{E}(x)\\ &- 2\nabla f_c(x)^T\mathbf{P}^{-1}\nabla^2 f_c(x)\mathbf{P}^{-1}\mathbb{E}(x)\Big)\\ T_3 =& (1+\delta)\alpha^3\left(\|\mathbf{P}^{-1}\nabla f_c(x) - \mathbf{P}^{-1}\mathbb{E}(x)\|^2\right)\left(2\|\mathbf{J}\mathbf{P}^{-1}\nabla f_c(x)\| + 2\|\mathbf{J}\mathbf{P}^{-1}\mathbb{E}(x)\|\right.\\ &\left.+ \alpha\|\mathbf{P}^{-1}\nabla f_c(x) - \mathbf{P}^{-1}\mathbb{E}(x)\|^2\right).\end{aligned}$$

II. Bounding T_1, T_2 and T_3.

We control each term in the above expression individually. First, we have

$$T_1 = \alpha \nabla f_c(x)^T \mathbf{P}^{-1}\mathbb{E}(x) \le \alpha\|\mathbf{P}^{-1}\nabla f_c(x)\|_P \|\mathbb{E}(x)\|_{P^*} = \alpha\|\nabla f_c(x)\|_{P^*}\|\mathbb{E}(x)\|_{P^*}.$$

To bound T_2, first we note that for any vectors $x, y \in \mathbb{R}^n$ and any positive semidefinite matrix $P \in S_+^n$, we always have $(x+y)^T P(x+y) \le 2(x^T Px + y^T Py)$. Therefore we can bound

$$T_2 \le \alpha^2 \left(\nabla f_c(x)^T \mathbf{P}^{-1}\nabla^2 f_c(x)\mathbf{P}^{-1}\nabla f_c(x) + \mathbb{E}(x)^T\mathbf{P}^{-1}\nabla^2 f_c(x)\mathbf{P}^{-1}\mathbb{E}(x)\right).$$

Next, we apply Lemma 18 to arrive at

$$\frac{1}{2}\sigma_{\max}(\mathbf{P}^{-1/2}\nabla^2 f_c(x)\mathbf{P}^{-1/2}) \le \frac{1+\delta}{m}\left(\frac{8\sigma_r^2(X) + \|XX^T - M^\star\|}{\sigma_r^2(X)+\eta}\right) \overset{def}{\le} L_\delta,$$

where L_δ is a constant that only depends on δ and m. Note that the last inequality follows from the fact that $\eta = O(\|XX^T - M^\star\|)$.

Now based on the above inequality, we have

$$\alpha^2\left(\nabla f_c(x)^T\mathbf{P}^{-1}\nabla^2 f_c(x)\mathbf{P}^{-1}\nabla f_c(x)\right) \le 2\alpha^2 L_\delta \|\nabla f_c(x)\|_{P^*}^2$$
$$\alpha^2\left(\mathbb{E}(x)^T\mathbf{P}^{-1}\nabla^2 f_c(x)\mathbf{P}^{-1}\mathbb{E}(x)\right) \le 2\alpha^2 L_\delta\|\mathbb{E}(x)\|_{P^*}^2,$$

which implies

$$T_2 \le 2\alpha^2 L_\delta\|\nabla f_c(x)\|_{P^*}^2 + 2\alpha^2 L_\delta \|\mathbb{E}(x)\|_{P^*}^2$$

Finally, to bound T_3, we first write

$$\|\mathbf{P}^{-1}\nabla f_c(x) - \mathbf{P}^{-1}\mathbb{E}(x)\|^2 \le 2\|\mathbf{P}^{-1}\nabla f_c(x)\|^2 + 2\|\mathbf{P}^{-1}\mathbb{E}(x)\|^2.$$

Moreover, invoking Lemma 19 leads to the following inequalities

$$\|\mathbf{P}^{-1}\nabla f_c(x)\|^2 \le \frac{\|\nabla f_c(x)\|_{P^*}^2}{\eta}, \qquad \|\mathbf{P}^{-1}\mathbb{E}(x)\|^2 \le \frac{\|\mathbb{E}(x)\|_{P^*}^2}{\eta}.$$
$$\|\mathbf{J}\mathbf{P}^{-1/2}\nabla f_c(x)\| \le 2\sqrt{2}\|\nabla f_c(x)\|_{P^*}, \qquad \|\mathbf{J}\mathbf{P}^{-1/2}\mathbb{E}(x)\| \le 2\sqrt{2}\|\mathbb{E}(x)\|_{P^*}.$$

Combining the above inequalities with the definition of T_3 leads to:

$$\begin{aligned} T_3 \le &\frac{4(1+\delta)\alpha^3}{\eta}\left(\|\nabla f_c(x)\|_{P^*}^2 + \|\mathbb{E}(x)\|_{P^*}^2\right) \\ &\times\left(2\sqrt{2}\|\nabla f_c(x)\|_{P^*} + 2\sqrt{2}\|\nabla\mathbb{E}(x)\|_{P^*} + \frac{\alpha}{\eta}\|\nabla f_c(x)\|_{P^*}^2 + \frac{\alpha}{\eta}\|\mathbb{E}(x)\|_{P^*}^2\right).\end{aligned}$$

III. Bounding the Error Term

Next, we provide an upper bound on $\|\mathbb{E}(x)\|_{P^*}$. The following chain of inequalities hold with high probability:

$$\begin{aligned}\|\mathbb{E}(x)\|_{P^*}^2 = \mathbb{E}(x)^T\mathbf{P}^{-1}\mathbb{E}(x) &= \left\|\left(\frac{2}{m}\sum_{i=1}^m \epsilon_i A_i\right)X(X^TX+\eta I)^{-1/2}\right\|_F^2 \\ &\le \left\|\left(\frac{2}{m}\sum_{i=1}^m \epsilon_i A_i\right)\right\|_2^2 \left\|X(X^TX+\eta I)^{-1/2}\right\|_F^2 \\ {}^{(a)} &\le C\frac{\sigma^2 n\log n}{m}\left(\sum_{i=1}^r \frac{\sigma_i^2(X)}{\sigma_i(X)^2+\eta}\right) \\ &\le C\frac{\sigma^2 rn\log n}{m},\end{aligned}$$

where C is an absolute constant and (a) follows from Lemma 16.

IV. Bounding all the terms using $\|\nabla f_c(x)\|_{P*}$

Combining the upper bound on $\|\mathbb{E}(X)\|_{P^*}$ with the previous bounds for T_1, T_2, T_3 and denoting $\Delta = \|\nabla f_c(x)\|_{P*}$, we have

$$T_1 \leq \alpha\Delta\sqrt{\frac{C\sigma^2 rn\log n}{m}},$$

$$T_2 \leq 2\alpha^2 L_\delta \Delta^2 + 2\alpha^2 L_\delta \frac{\sigma^2 rn\log n}{m}$$

$$T_3 \leq \frac{4(1+\delta)\alpha^3}{\eta}\left(\Delta^2 + \frac{C\sigma^2 rn\log n}{m}\right)\left(\frac{\alpha\Delta^2}{\eta} + \frac{\alpha C\sigma^2 rn\log n}{\eta m} + 2\sqrt{2}\Delta + 2\sqrt{2}\sqrt{\frac{C\sigma^2 rn\log n}{m}}\right)$$

Now, combining the upper bounds for T_1, T_2 and T_3 with (62) yields

$$\begin{aligned} f_c(x-\alpha d) \leq f_c(x) &- \alpha\Delta^2 + \alpha\Delta\sqrt{\frac{C\sigma^2 rn\log n}{m}} + 2\alpha^2 L_\delta\Delta^2 + 2C\alpha^2 L_\delta\frac{\sigma^2 rn\log n}{m} \\ &+ \frac{4(1+\delta)\alpha^3}{\eta}\left(\Delta^2 + \frac{C\sigma^2 rn\log n}{m}\right)\left(\frac{\alpha\Delta^2}{\eta} + \frac{\alpha C\sigma^2 rn\log n}{\eta m} + 2\sqrt{2}\Delta + 2\sqrt{2}\sqrt{\frac{C\sigma^2 rn\log n}{m}}\right). \end{aligned} \tag{63}$$

The above inequality holds with high probability for every iteration of PrecGD.

V. Two cases

Now, we consider two cases. First, suppose that $\eta \leq 2\sqrt{\frac{C\sigma^2 nr\log n}{\mu_P m}}$. This implies that $\min_k \eta_k \leq 2\sqrt{\frac{C\sigma^2 nr\log n}{\mu_P m}}$, and hence,

$$\|X_{k^*}X_{k^*}^T - M^\star\|_F^2 \lesssim \frac{1}{1-\delta}\frac{1}{m}\|\mathcal{A}(X_{k^*}X_{k^*}^T - M^\star)\|^2 \lesssim \mathcal{E}_{stat}$$

which completes the proof.

Otherwise, suppose that $\eta > 2\sqrt{\frac{C\sigma^2 nr\log n}{\mu_P m}}$. Due to Theorem 4, we have $\Delta \geq 2\sqrt{\frac{C\sigma^2 rn\log n}{m}}$, which leads to the following inequalities:

$$-\alpha\Delta^2 + \alpha\Delta\sqrt{\frac{C\sigma^2 rn\log n}{m}} \leq -\frac{\alpha}{2}\Delta^2, \qquad 2\alpha^2 L_\delta\Delta^2 + 2C\alpha^2 L_\delta\frac{\sigma^2 rn\log n}{m} \leq \frac{5}{2}\alpha^2 L_\delta\Delta^2.$$

Similarly, we have

$$\Delta^2 + \frac{C\sigma^2 rn\log n}{m} \leq \frac{5}{4}\Delta^2, \quad 2\sqrt{2}\Delta + 2\sqrt{2}\sqrt{\frac{C\sigma^2 rn\log n}{m}} \leq 3\sqrt{2}\Delta,$$

and

$$\frac{\alpha\Delta^2}{\eta} + \frac{\alpha}{\eta}\frac{C\sigma^2 rn\log n}{m} \leq \frac{5}{4}\frac{\alpha\Delta^2}{\eta}.$$

Combined with (63), we have

$$\begin{aligned} f_c(x-\alpha d) &\leq f_c(x) - \frac{\alpha}{2}\Delta^2 + \frac{5}{2}\alpha^2 L_\delta\Delta^2 + \frac{4(1+\delta)\alpha^3}{\eta}\left(\frac{5}{4}\Delta^2\right)\left(3\sqrt{2}\Delta + \frac{5}{4}\frac{\alpha\Delta^2}{\eta}\right) \\ &\leq f_c(x) - \frac{\alpha}{2}\Delta^2\left(1 - \frac{5}{2}L_\delta\alpha - 60\sqrt{2}\frac{\alpha^2\Delta}{\eta} - 25\alpha^3\left(\frac{\Delta}{\eta}\right)^2\right). \end{aligned}$$

Similar to the noiseless case, we can bound the ratio $\frac{\Delta}{\eta}$ as

$$\frac{\Delta}{\eta} = \frac{\|\nabla f_c(x)\|_{P*}}{\eta} \leq \frac{(1+\delta)\sigma_{\max}(\mathbf{J}\mathbf{P}^{-1/2})\|\mathbf{e}\|}{\|\mathbf{e}\|} = (1+\delta)\frac{\sigma_{\max}^2(X)}{\sigma_{\max}^2(X)+\eta} \leq 1+\delta,$$

which in turn leads to

$$f_c(x-\alpha d) \leq f_c(x) - \frac{\alpha}{2}\Delta^2\left(1 - \frac{5}{2}L_\delta\alpha - 60\sqrt{2}\alpha^2(1+\delta) - 25\alpha^3(1+\delta)^2\right).$$

Now, assuming that the step-size satisfies $\alpha \leq \min\left\{\frac{L_\delta}{60\sqrt{2}(1+\delta)+25(1+\delta)^2}, \frac{1}{7L_\delta}\right\}$. Since L_δ is a constant, we can simply write the condition above as $\alpha \leq 1/L$ where $L = \max\left\{\frac{60\sqrt{2}(1+\delta)+25(1+\delta)^2}{L_\delta}, 7L_\delta\right\}$. Now note that

$$\frac{5}{2}L_\delta + 60\sqrt{2}(1+\delta)\alpha + 25(1+\delta)^2\alpha^2 \leq \frac{7}{2}L_\delta$$
$$\implies 1 - \frac{5}{2}L_\delta\alpha - 60\sqrt{2}(1+\delta)\alpha^2 - 25(1+\delta)^2\alpha^3 \geq 1 - \frac{7}{2}L_\delta\alpha \geq \frac{1}{2}.$$

This implies that

$$f_c(x - \alpha d) \leq f_c(x) - \frac{t\Delta^2}{4} \leq \left(1 - \frac{\alpha\mu_P}{4}\right) f_c(x),$$

where in the last inequality, we used $\Delta^2 \geq \mu_P f_c(x)$, which is just the PL-inequality in Theorem 4. Finally, since $f_c(x)$ satisfies the RIP condition, combining the two cases above we get

$$\|X_{k^*}X_{k^*}^T - M^\star\|_F^2 \lesssim \max\left\{\frac{1+\delta}{1-\delta}\left(1 - \alpha\frac{\mu_P}{2}\right)^k \|X_0X_0^T - M^*\|_F^2, \mathcal{E}_{stat}\right\}, \tag{64}$$

as desired. □

G Proof of Noisy Case with Variance Proxy (Theorem 8)

In this section we prove Theorem 8, which we restate below for convenience. The only difference between this theorem and Theorem 7 is that we de not assume that we have access to the optimal choice of η. Instead, we only assume that we have some proxy $\hat{\sigma}^2$ of the true variance of the noise. For convenience we restate our result below.

Theorem 21 (Noisy measurements with variance proxy)**.** *Suppose that the noise vector $\epsilon \in \mathbb{R}^m$ has sub-Gaussian entries with zero mean and variance $\sigma^2 = \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\epsilon_i^2]$. Moreover, suppose that $\eta_k = \sqrt{|f(X_k) - \hat{\sigma}^2|}$ for $k = 0, 1, \ldots, K$, where $\hat{\sigma}^2$ is an approximation of σ^2, and that the initial point X_0 satisfies $\|\mathcal{A}(X_0X_0^T - M^*)\|_F^2 < \rho^2(1-\delta)\lambda_{r^*}(M^\star)^2$. Consider $k^* = \arg\min_k \eta_k$, and suppose that $\alpha \leq 1/L$, where $L > 0$ is a constant that only depends on δ. Then, with high probability, we have*

$$\|X_{k^*}X_{k^*}^T - M^*\|_F^2 \lesssim \max\left\{\frac{1+\delta}{1-\delta}\left(1 - \alpha\frac{\mu_P}{2}\right)^K \|X_0X_0^T - M^*\|_F^2, \mathcal{E}_{stat}, \mathcal{E}_{dev}, \mathcal{E}_{var}\right\}, \tag{65}$$

where

$$\mathcal{E}_{stat} := \frac{\sigma^2 nr\log n}{\mu_P(1-\delta)m}, \quad \mathcal{E}_{dev} := \frac{\sigma^2}{1-\delta}\sqrt{\frac{\log n}{m}}, \quad \mathcal{E}_{var} := |\sigma^2 - \hat{\sigma}^2|. \tag{66}$$

The proof of Theorem 8 is similar to that of Theorem 7, with a key difference that $\eta_k = \frac{1}{\sqrt{m}}\|\mathcal{A}(X_kX_k^T - M^\star)\|$ is replaced with $\eta_k = \sqrt{|f(x_k) - \hat{\sigma}^2|}$. Our next lemma shows that this alternative choice of damping parameter remains close to $\frac{1}{\sqrt{m}}\|\mathcal{A}(X_kX_k^T - M^\star)\|$, provided that the error exceeds a certain threshold.

Lemma 22. *Set $\eta = \sqrt{|f(x) - \hat{\sigma}^2|}$. Then, with high probability, we have*

$$\sqrt{\frac{1/4-\delta}{1+\delta}}\frac{1}{\sqrt{m}}\left\|\mathcal{A}(XX^T - M^\star)\right\| \leq \eta \leq \sqrt{\frac{7/4+\delta}{1-\delta}}\frac{1}{\sqrt{m}}\left\|\mathcal{A}(XX^T - M^\star)\right\|$$

provided that

$$\|XX^T - M^\star\|_F^2 \gtrsim \max\left\{\frac{\sigma^2 rn\log n}{m}, \sqrt{\frac{\sigma^2\log n}{m}}, |\sigma^2 - \hat{\sigma}^2|\right\}.$$

Proof. One can write

$$\begin{aligned} f(x) &= \frac{1}{m}\|y-\mathcal{A}(XX^T)\|^2 = \frac{1}{m}\|\mathcal{A}(M^\star - XX^T)+\epsilon\|^2 \\ &= \frac{1}{m}\|\mathcal{A}(M^\star - XX^T)\|^2 + \frac{1}{m}\|\epsilon\|^2 + \frac{2}{m}\left\langle \mathcal{A}(M^\star - XX^T), \epsilon\right\rangle . \end{aligned}$$

Due to the definition of the restricted Frobenius norm (57), we have

$$|\left\langle \mathcal{A}(M^\star - XX^T), \epsilon\right\rangle| \leq \|M^\star - XX^T\|_F \left\|\frac{1}{m}\sum_{i=1}^m A_i\epsilon_i\right\|_{F,2r} .$$

Therefore, we have

$$\left|\frac{1}{m}\|\mathcal{A}(M^\star - XX^T)\|^2 + \frac{1}{m}\|\epsilon\|^2 - \hat{\sigma}^2 - 2\|M^\star - XX^T\|_F\left\|\frac{1}{m}\sum_{i=1}^m A_i\epsilon_i\right\|_{F,2r}\right| \leq \eta^2 \tag{67}$$

$$\left|\frac{1}{m}\|\mathcal{A}(M^\star - XX^T)\|^2 + \frac{1}{m}\|\epsilon\|^2 - \hat{\sigma}^2 + 2\|M^\star - XX^T\|_F\left\|\frac{1}{m}\sum_{i=1}^m A_i\epsilon_i\right\|_{F,2r}\right| \geq \eta^2. \tag{68}$$

Since the error ϵ_i is sub-Gaussian with parameter σ, the random variable ϵ_i^2 is sub-exponential with parameter 16σ. Therefore,

$$\mathbb{P}\left(\left|\frac{1}{m}\|\epsilon\|^2 - \sigma^2\right| \geq t\right) \leq 2\exp\left(-\frac{Cmt^2}{\sigma^2}\right).$$

Now, upon setting $t = \sqrt{\frac{\sigma^2\log n}{m}}$, we have

$$\left|\frac{1}{m}\|\epsilon\|^2 - \sigma^2\right| \leq \sqrt{\frac{\sigma^2\log n}{m}},$$

Moreover, we have

$$\left\|\frac{1}{m}\sum_{i=1}^m A_i\epsilon_i\right\|_{F,2r} \leq \sqrt{2r}\left\|\frac{1}{m}\sum_{i=1}^m A_i\epsilon_i\right\|_2 \lesssim \sqrt{\frac{\sigma^2 rn\log n}{m}}. \tag{69}$$

Combining the above two inequalities with (67) leads to

$$\begin{aligned} \eta^2 &\geq \frac{1}{m}\|\mathcal{A}(M^\star - XX^T)\|^2 - C\|M^\star - XX^T\|_F\sqrt{\frac{\sigma^2 rn\log n}{m}} - \sqrt{\frac{\sigma^2\log n}{m}} - |\sigma^2 - \hat{\sigma}^2| \\ &\geq (1-\delta)\|XX^T - M^\star\|_F^2 - C\|XX^T - M^\star\|_F\sqrt{\frac{\sigma^2 rn\log n}{m}} - \sqrt{\frac{\sigma^2\log n}{m}} - |\sigma^2 - \hat{\sigma}^2|. \end{aligned} \tag{70}$$

Now assuming that

$$\|XX^T - M^\star\|_F^2 \geq \max\left\{16C^2\frac{\sigma^2 rn\log n}{m}, 4\sqrt{\frac{\sigma^2\log n}{m}}, 4|\sigma^2 - \hat{\sigma}^2|\right\},$$

the inequality (70) can be further lower bounded as

$$\eta^2 \geq (1/4-\delta)\|XX^T - M^\star\|_F^2 \geq \frac{1/4-\delta}{1+\delta}\frac{1}{m}\|\mathcal{A}(XX^T - M^\star)\|,$$

which completes the proof for the lower bound. The upper bound on η^2 can be established in a similar fashion. □

Now we are ready to prove Theorem 8.

Proof. We consider two cases. First, suppose that

$$\min_k \eta_k \lesssim \max\left\{\frac{\sigma^2 rn\log n}{m}, \sqrt{\frac{\sigma^2\log n}{m}}, |\sigma^2-\hat{\sigma}^2|\right\}.$$

Combined with (70), this implies that

$$\begin{aligned}&(1-\delta)\|X_{k^*}X_{k^*}^T - M^\star\|_F^2 - C\|X_{k^*}X_{k^*}^T - M^\star\|_F\sqrt{\frac{\sigma^2 rn\log n}{m}}\\ &\lesssim \max\left\{\frac{\sigma^2 rn\log n}{m}, \sqrt{\frac{\sigma^2\log n}{m}}, |\sigma^2-\hat{\sigma}^2|\right\}.\end{aligned} \tag{71}$$

Now, if $\|X_{k^*}X_{k^*}^T - M^\star\|_F \le 2C\sqrt{\frac{\sigma^2 rn\log n}{m}}$, then the proof is complete. Therefore, suppose that $\|X_{k^*}X_{k^*}^T - M^\star\|_F > 2C\sqrt{\frac{\sigma^2 rn\log n}{m}}$. This together with (71) leads to

$$\|X_{k^*}X_{k^*}^T - M^\star\|_F^2 \lesssim \frac{1}{1/2-\delta}\max\left\{\frac{\sigma^2 rn\log n}{m}, \sqrt{\frac{\sigma^2\log n}{m}}, |\sigma^2-\hat{\sigma}^2|\right\},$$

which again completes the proof. Finally, suppose that

$$\min_k \eta_k \gtrsim \max\left\{\frac{\sigma^2 rn\log n}{m}, \sqrt{\frac{\sigma^2\log n}{m}}, |\sigma^2-\hat{\sigma}^2|\right\}.$$

This combined with (67) implies that

$$\begin{aligned}&(1+\delta)\|X_{k^*}X_{k^*}^T - M^\star\|_F^2 + C\|X_{k^*}X_{k^*}^T - M^\star\|_F\sqrt{\frac{\sigma^2 rn\log n}{m}}\\ &\gtrsim \max\left\{\frac{\sigma^2 rn\log n}{m}, \sqrt{\frac{\sigma^2\log n}{m}}, |\sigma^2-\hat{\sigma}^2|\right\},\end{aligned}$$

for every $k = 0, 1, \dots, K$. If $\|X_{k^*}X_{k^*}^T - M^\star\|_F \le 2C\sqrt{\frac{\sigma^2 rn\log n}{m}}$, then the proof is complete. Therefore, suppose that $\|X_{k^*}X_{k^*}^T - M^\star\|_F > 2C\sqrt{\frac{\sigma^2 rn\log n}{m}}$. This together with the above inequality results in

$$\begin{aligned}\|X_k X_k^T - M^\star\|_F^2 &\gtrsim \frac{1}{3/2+\delta}\max\left\{\frac{\sigma^2 rn\log n}{m}, \sqrt{\frac{\sigma^2\log n}{m}}, |\sigma^2-\hat{\sigma}^2|\right\}\\ &\gtrsim \max\left\{\frac{\sigma^2 rn\log n}{m}, \sqrt{\frac{\sigma^2\log n}{m}}, |\sigma^2-\hat{\sigma}^2|\right\}\end{aligned}$$

for every $k = 0, 1, \dots, K$. Therefore, Lemma 22 can be invoked to show that

$$\eta_k \asymp \frac{1}{\sqrt{m}}\|\mathcal{A}(X_k X_k^T - M^\star)\|.$$

With this choice of η_k, the rest of the proof is identical to that of Theorem 7, and omitted for brevity. □

H Proof for Spectral Initialization (Proposition 6)

In this section we prove that spectral initialization is able to generate a sufficiently good initial point so that PrecGD achieves a linear convergence rate, even in the noisy case. For convenience we restate our result below.

Proposition 23 (Spectral Initialization)**.** *Suppose that $\delta \le (8\kappa\sqrt{r^*})^{-1}$ and $m \gtrsim \frac{1+\delta}{1-\delta}\frac{\sigma^2 rn\log n}{\rho^2\lambda_{r^\star}^2(M^\star)}$ where $\kappa = \lambda_1(M^\star)/\lambda_{r^\star}(M^\star)$. Then, with high probability, the initial point X_0 produced by (18) satisfies the radius condition (17).*

Proof. Let $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^{n\times n}$ be the dual of the linear operator $\mathcal{A}(\cdot)$, defined as $\mathcal{A}^*(y) = \sum_{i=1}^m y_i A_i$. Based on this definition, the initial point $X_0 \in \mathbb{R}^{n\times r}$ satisfies $X_0 = \mathcal{P}_r\left(\frac{1}{m}\mathcal{A}^*(y)\right)$, where we recall that

$$\mathcal{P}_r(M) = \arg\min_{X\in\mathbb{R}^{n\times r}} \|XX^T - M\|_F.$$

Define $E = X_0X_0^T - M^\star$, and note that $\operatorname{rank}(E) \le 2r$. It follows that

$$\begin{aligned}
\|E\|_F &= \sqrt{\sum_{i=1}^{r}\sigma_i(E)^2 + \sum_{i=r+1}^{2r}\sigma_i(E)^2} \le \sqrt{2}\|E\|_{F,2r}\\
&\le \sqrt{2}\left\|X_0X_0^T - \frac{1}{m}\mathcal{A}^*(y)\right\|_{F,2r} + \sqrt{2}\left\|\frac{1}{m}\mathcal{A}^*(y) - M^\star\right\|_{F,2r}\\
&\le 2\sqrt{2}\left\|\frac{1}{m}\mathcal{A}^*(y) - M^\star\right\|_{F,2r}\\
&\le 2\sqrt{2}\left\|\frac{1}{m}\mathcal{A}^*(\mathcal{A}(M^\star)) - M^\star\right\|_{F,2r} + 2\sqrt{2}\left\|\frac{1}{m}A_i\epsilon_i\right\|_{F,2r}\\
&\le 2\sqrt{2}\delta\|M^\star\|_F + 2\sqrt{2}\left\|\frac{1}{m}A_i\epsilon_i\right\|_{F,2r}.
\end{aligned}$$

Now, note that $\|M^\star\|_F \le \sqrt{r^*}\kappa\lambda_{r^*}(M^\star)$. Moreover, due to Lemma 16, we have

$$2\sqrt{2}\left\|\frac{1}{m}A_i\epsilon_i\right\|_{F,2r} \le 2\sqrt{2}\sqrt{2r}\left\|\frac{1}{m}A_i\epsilon_i\right\|_2 \lesssim \sqrt{\frac{\sigma^2 rn\log n}{m}}. \tag{72}$$

This implies that

$$\frac{1}{m}\|\mathcal{A}(X_0X_0^T - M^\star)\|^2 \le 16(1+\delta)r^*\kappa^2\lambda_{r^*}(M^\star)^2\delta^2 + C\frac{\sigma^2 rn\log n}{m}$$

Therefore, upon choosing $\delta \le \frac{\rho}{8\sqrt{r^*}\kappa}$ and $m \gtrsim \frac{1+\delta}{1-\delta}\frac{\sigma^2 rn\log n}{\rho^2\lambda_{r^*}^2(M^\star)}$, we have

$$\frac{1}{m}\|\mathcal{A}(XX^T - M^*)\|^2 \le \rho^2(1-\delta)\lambda_{r^*}(M^\star)^2 \tag{73}$$

This completes the proof. □

I Proof of Lemma 16

First we state a standard concentration inequality. A proof of this result can be found in Tropp [56].

Lemma 24 (Matrix Bernstein's inequality)**.** *Suppose that $\{W_i\}_{i=1}^m$ are matrix-valued random variables such that $\mathbb{E}[W_i] = 0$ and $\|W_i\|_2 \le R^2$ for all $i = 1,\dots,m$. Then*

$$\mathbb{P}\left(\left\|\sum_{i=1}^m W_i\right\| \ge t\right) \le n\exp\left(\frac{-t^2}{2\left\|\sum_{i=1}^m \mathbb{E}\left[W_i^2\right]\right\|_2 + \frac{2R^2}{3}t}\right).$$

We also state a standard concentration bound for the operator norm of Gaussian ensembles. A simple proof can be found in Wainwright [57].

Lemma 25. *Let $A \in \mathbb{R}^{n\times n}$ be a standard Gaussian ensemble with i.i.d. entries. Then the largest singular value of A (or equivalently, the operator norm) satisfies*

$$\sigma_{\max}(A) \le (2+c)\sqrt{n}$$

with probability at least $1 - 2\exp(-nc^2/2)$.

For simplicity, we assume that the measurement matrices $A_i, i = 1,\dots m$ are fixed and all satisfy $\|A_i\| \le C\sqrt{n}$. Due to Lemma 25, this assumption holds with high probability for Gaussian measurement ensembles. Next, we provide the proof of Lemma 16.

Proof of Lemma 16. First, note that $\|A_i\varepsilon_i\|_2 \leq \|A_i\| \cdot |\varepsilon_i|$. The assumption $\|A_i\| \lesssim \sqrt{n}$ implies that $\|A_i\varepsilon_i\|$ is sub-Gaussian with parameter $C\sqrt{n}\sigma$. Therefore, we have $\mathbb{P}(\|A_i\varepsilon\| \gtrsim \sqrt{n}t) \geq 1 - 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$. Applying the union bound yields

$$\mathbb{P}(\max_{i=1,\dots,m} \|A_i\varepsilon\| \geq \sqrt{n}t) \geq 1 - 2m\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Moreover, one can write

$$\left\|\sum_{i=1}^{m} \mathbb{E}[(A_i\varepsilon_i)^2]\right\| \leq \sum_{i=1}^{m} \|A_i\|^2 \mathbb{E}[\varepsilon_i^2] \lesssim \sigma^2 mn \tag{74}$$

Using Matrix Bernstein's inequality, we get

$$\mathbb{P}\left(\frac{1}{m}\left\|\sum_{i=1}^{m} A_i\varepsilon\right\| \leq t\right) \geq 1 - n\exp\left(-\frac{t^2m^2}{2C\sigma^2mn + \frac{2}{3}C'\sqrt{n}mt}\right) - 2m\exp\left(-\frac{t^2}{2}\right).$$

Using $t \asymp \sqrt{\frac{\sigma^2 n\log n}{m}}$ in the above inequality leads to

$$\begin{aligned}\mathbb{P}\left(\frac{1}{m}\left\|\sum_{i=1}^{m} A_i\varepsilon\right\| \lesssim \sqrt{\frac{\sigma^2 n\log n}{m}}\right) &\geq 1 - n^{-C} - 2m\exp\left(-\frac{t^2}{2}\right) \\ &\gtrsim 1 - 3n^{-C},\end{aligned}$$

where the last inequality follows from the assumption $m \gtrsim \sigma n\log n$. This completes the proof. □