Journal of Machine Learning Research 24 (2023) 1-55 Submitted 8/22; Revised 4/23; Published 4/23

Preconditioned Gradient Descent for Overparameterized Nonconvex Burer–Monteiro Factorization with Global Optimality Certification *

Gavin Zhang JIALUN2@ILLINOIS.EDU
Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Salar Fattahi FATTAHI@UMICH.EDU
Industrial and Operations Engineering
University of Michigan, Ann Arbor, MI 48109, USA

Richard Y. Zhang RYZ@ILLINOIS.EDU
Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Editor: Prateek Jain

Abstract

We consider using gradient descent to minimize the nonconvex function $f(X) = \phi(XX^T)$ over an $n \times r$ factor matrix X, in which ϕ is an underlying smooth convex cost function defined over $n \times n$ matrices. While only a second-order stationary point X can be provably found in reasonable time, if X is additionally *rank deficient*, then its rank deficiency certifies it as being globally optimal. This way of certifying global optimality necessarily requires the search rank r of the current iterate X to be *overparameterized* with respect to the rank $r^\star$ of the global minimizer $X^\star$. Unfortunately, overparameterization significantly slows down the convergence of gradient descent, from a linear rate with $r = r^\star$ to a sublinear rate when $r > r^\star$, even when ϕ is strongly convex. In this paper, we propose an inexpensive preconditioner that restores the convergence rate of gradient descent back to linear in the overparameterized case, while also making it agnostic to possible ill-conditioning in the global minimizer $X^\star$.

Keywords: Low-rank matrix recovery, Burer-Moneiro Factorization, Nonconvex Optimization, Global Optimality Certification

1. Introduction

Numerous state-of-the-art algorithms in statistical and machine learning can be viewed as gradient descent applied to the nonconvex Burer–Monteiro (Burer and Monteiro, 2003, 2005) problem

$$X^\star = \quad \text{minimize} \quad f(X) \stackrel{\text{def}}{=} \phi(XX^T) \text{ over } X \in \mathbb{R}^{n\times r}, \tag{BM}$$

in which ϕ is an underlying convex cost function defined over $n \times n$ matrices. Typically, the search rank $r \ll n$ is set significantly smaller than n, and an efficient gradient oracle

*. Financial support for this work was provided by NSF CAREER Award ECCS-2047462, NSF Award DMS-2152776, ONR Award N00014-22-1-2127.

$X \mapsto \nabla f(X)$ is available due to problem structure that costs $n \cdot \text{poly}(r)$ time per query. Under these two assumptions, each iteration of gradient descent $X_+ = X - \alpha \nabla f(X)$ costs $O(n)$ time and memory.

Gradient descent has become widely popular for problem (BM) because it is simple to implement but works exceptionally well in practice (Sun and Luo, 2016; Bhojanapalli et al., 2016a,b; Park et al., 2017; Chen and Candès, 2017; Park et al., 2018; Chen et al., 2019). Across a broad range of applications, gradient descent is consistently observed to converge from an arbitrary, possibly random initial guess X_0 to the global minimum $X^\star$, as if the function f were *convex*. In fact, in many cases, the observed convergence rate is even *linear*, meaning that gradient descent converges to ϵ global suboptimality in $O(\log(1/\epsilon))$ iterations, as if the function f were *strongly convex*. When this occurs, the resulting empirical complexity of ϵ-accuracy in $O(n \cdot \log(1/\epsilon))$ time matches the best figures achievable by algorithms for convex optimization.

However, due to the nonconvexity of f, it is always possible for gradient descent to fail by getting stuck at a *spurious* local minimum—a local minimum that is strictly worse than that of the global minimum. This is a particular concern for safety-critical applications like electricity grids (Zhang et al., 2019a) and robot navigation (Rosen et al., 2019, 2020), where mistaking a clearly suboptimal X for the globally optimal $X^\star$ could have serious rammifications. Recent authors have derived conditions under which f is guaranteed not to admit spurious local minima, but such *a priori* global optimality guarantees, which are valid for all initializations before running the algorithm, can be much stronger than what is needed for gradient descent to succeed in practice. For example, it may also be the case that spurious local minima do generally exist, but that gradient descent is frequently able to avoid them without any rigorous guarantees of doing so.

In this paper, we consider *overparameterizing* the search rank r, choosing it to be large enough so that $\text{rank}(X^\star) < r$ holds for all globally optimal $X^\star$. We are motivated by the ability to guarantee global optimality *a posteriori*, that is, *after* a candidate X has already been computed. To explain, it has been long suspected and recently rigorously shown Ge et al. (2015); Jin et al. (2017, 2021) that gradient descent can be made to converge to an approximate second-order stationary point X that satisfies

$$|\left\langle \nabla f(X), V \right\rangle| \leq \epsilon_g \|V\|_F, \quad \left\langle \nabla^2 f(X)[V], V \right\rangle \geq -\epsilon_H \|V\|_F^2 \quad \text{for all } V \in \mathbb{R}^{n \times r} \tag{1}$$

with arbitrarily small accuracy parameters $\epsilon_g, \epsilon_H > 0$. By evoking an argument first introduced by Burer and Monteiro (2005, Theorem 4.1) (see also Journée et al. (2010) and Boumal et al. (2016, 2020)) one can show that an X that satisfies (1) has global suboptimality:

$$f(X) - f(X^\star) \leq \underbrace{C_g \cdot \epsilon_g}_{\text{gradient norm}} + \underbrace{C_H \cdot \epsilon_H}_{\text{Hessian curvature}} + \underbrace{C_\lambda \cdot \lambda_{\min}(X^T X)}_{\text{rank deficiency}} \tag{2}$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue, and $C_g, C_H, C_\lambda > 0$ are absolute constants under standard assumptions. By overparameterizing the search rank so that $r > r^\star$ holds, where $r^\star$ denotes the *maximum rank* over all globally optimal $X^\star$, it follows from (2) that the global optimality of an X with $\epsilon_g \approx 0$ and $\epsilon_H \approx 0$ is conclusively determined by its rank deficiency term $\lambda_{\min}(X^T X)$:

1. (Near globally optimal) If $X \approx X^\star$, then X must be nearly rank deficient with $\lambda_{\min}(X^T X) \approx 0$. In this case, the near-global optimality of X can be rigorously *certified* by evoking (2) with $\epsilon_g \approx 0$ and $\epsilon_H \approx 0$ and $\lambda_{\min}(X^T X) \approx 0$.

2. (Stuck at spurious point) If $f(X) \gg f(X^\star)$, then by contradiction X must be nearly full-rank with $\lambda_{\min}(X^T X) \approx C_\lambda^{-1}(f(X) - f(X^\star)) \not\approx 0$ bounded away from zero.

As we describe in Section 3, the three parameters ϵ_g, ϵ_H, and $\lambda_{\min}(X^T X)$ for a given X can all be numerically evaluated in $O(n)$ time and memory, using a small number of calls to the gradient oracle $X \mapsto \nabla f(X)$. (In Section 3, we formally state and prove (2) as Proposition 10.)

Aside from the ability to certify global optimality, a second benefit of overparameterization is that f tends to admit fewer spurious local minima as the search rank r is increased beyond the maximum rank $r^\star \geq \operatorname{rank}(X^\star)$. Indeed, it is commonly observed in practice that any local optimization algorithm seem to globally solve problem (BM) as soon as r is slightly larger than $r^\star$; see (Burer and Monteiro, 2003; Journée et al., 2010; Rosen et al., 2019) for numerical examples of this behavior. Towards a rigorous explanation, Boumal et al. (2016, 2020) pointed out that if the search rank is overparameterized as $r \geq n$, then the function f is guaranteed to contain *no spurious local minima*, in the sense that every second-order stationary point Z satisfying $\nabla f(Z) = 0$ and $\nabla^2 f(Z) \succeq 0$ is guaranteed to be global optimal $f(Z) = f(X^\star)$. This result was recently sharpened by Zhang (2022), who proved that if the underlying convex cost ϕ is L-gradient Lipschitz and μ-strongly convex, then overparameterizing the search rank by a constant factor as $r > \max\{r^\star, \frac{1}{4}(L/\mu - 1)^2 r^\star\}$ is enough to guarantee that f contains no spurious local minima.

Unfortunately, overparameterization significantly slows down the convergence of gradient descent, both in theory and in practice. Under suitable strong convexity and optimality assumptions on ϕ, Zheng and Lafferty (2015b); Tu et al. (2016) showed that gradient descent $X_+ = X - \alpha \nabla f(X)$ locally converges as follows

$$f(X_+) - f(X^\star) \leq \left[1 - \alpha \cdot c \cdot \lambda_{\min}(X^T X)\right] \cdot [f(X) - f(X^\star)]$$

where $\alpha > 0$ is the corresponding step-size, and $c > 0$ is a constant (see also Section 5 for an alternative derivation). In the exactly parameterized regime $r = r^\star$, this inequality implies linear convergence, because $\lambda_{\min}(X^T X) > 0$ holds within a local neighborhood of the minimizer $X^\star$. In the overparameterized regime $r > r^\star$, however, the iterate X becomes increasingly singular $\lambda_{\min}(X^T X) \to 0$ as it makes progress towards the global minimizer $X^\star$, and the convergence quotient $Q = 1 - \alpha \cdot c \cdot \lambda_{\min}(X^T X)$ approaches 1. In practice, gradient descent slows down to *sublinear* convergence, now requiring $\operatorname{poly}(1/\epsilon)$ iterations to yield an ϵ suboptimal solution. This is a dramatic, exponential slow-down compared to the $O(\log(1/\epsilon))$ figure associated with linear convergence under exact rank parameterization $r = r^\star$.

For applications of gradient descent with very large values of n, this exponential slow-down suggests that the ability to certify global optimality via overparameterization can only come by dramatically worsening the quality of the computed solution. In most cases, it remains better to exactly parameterize the search rank as $r = r^\star$, in order to compute a high-quality solution without a rigorous proof of quality. For safety-critical applications for which a proof of quality is paramount, rank overparameterization $r > r^\star$ is used alongside

much more expensive trust-region methods (Rosen et al., 2019, 2020; Boumal et al., 2020). These methods can be made immune to the progressive ill-conditioning $\lambda_{\min}(X^TX) \to 0$ of the current iterate X, but have per-iteration costs of $O(n^3)$ time and $O(n^2)$ memory that limit n to modest values.

1.1 Summary of results

In this paper, we present an inexpensive *preconditioner* for gradient descent that restores the convergence rate of gradient descent back to linear in the overparameterized case, both in theory and in practice. We propose the following iterations

$$X_+ = X - \alpha \nabla f(X)(X^TX + \eta I)^{-1}, \tag{PrecGD}$$

where $\alpha \in (0, 1]$ is a fixed step-size, and $\eta \geq 0$ is a regularization parameter that may be changed from iteration to iteration. We call these iterations *preconditioned* gradient descent or PrecGD, because they can be viewed as gradient descent applied with a carefully chosen $r \times r$ preconditioner.

It is easy to see that PrecGD should maintain a similar per-iteration $O(n)$ cost to regular gradient descent in most applications where the Burer-Monteiro approach is used, where r is typically set orders of magnitude smaller than n. The method induces an additional cost of $O(r^3)$ each iteration to form and compute the preconditioner $(X^TX + \eta I)^{-1}$. But in applications with very large values of n and very small values of r, the small increase in the per-iteration cost, from $O(r)$ to $O(r^3)$, is completely offset by the exponential reduction in the number of iterations, from $O(1/\epsilon)$ to $O(\log(1/\epsilon))$. Therefore, PrecGD can serve as a plug-in replacement for gradient descent, in order to provide the ability to certify global optimality without sacrificing the high quality of the computed solution.

Our results are summarized as follows:

Local convergence. Starting within a neighborhood of the global minimizer $X^\star$, and under suitable strong convexity and optimality assumptions on ϕ, classical gradient descent converges to ϵ suboptimality in $O(1/\lambda_r \log(1/\epsilon))$ iterations where $\lambda_r = \lambda_{\min}(X^{\star T}X^\star)$ is the rank deficiency term of the global minimizer (Zheng and Lafferty, 2015b; Tu et al., 2016). This result breaks down in the overparameterized regime, where $r > r^\star = \text{rank}(X^\star)$ and $\lambda_r = 0$ holds by definition; instead, gradient descent now requires $\text{poly}(1/\epsilon)$ iterations to converge to ϵ suboptimality (Zhuo et al., 2021).

Under the same strong convexity and optimality assumptions, we prove that PrecGD with the parameter choice $\eta = \|\nabla f(X)(X^TX)^{-1/2}\|_F$ converges to ϵ global suboptimality in $O(\log(1/\epsilon))$ iterations, independent of $\lambda_r^\star = 0$. In fact, we prove that the convergence rate of PrecGD also becomes independent of the smallest nonzero singular value $\lambda_{r^\star} = \lambda_{r^\star}(X^{\star T}X^\star)$ of the global minimizer $X^\star$. In practice, this often allows PrecGD to converge faster in the overparameterized regime $r > r^\star$ than regular gradient descent in the exactly parameterized regime $r = r^\star$ (see Fig. 1). In our numerical results, we observe that the linear convergence rate of PrecGD for all values of $r \geq r^\star$ and $\lambda_{r^\star} > 0$ is the same as regular gradient descent with a perfectly conditioned global minimizer $X^\star$, i.e. with $r = r^\star$ and $\lambda_{r^\star} = \lambda_1(X^{\star T}X^\star)$. In fact, linear convergence was observed even for choices of ϕ that do not satisfy the notions of strong convexity considered in our theoretical results.

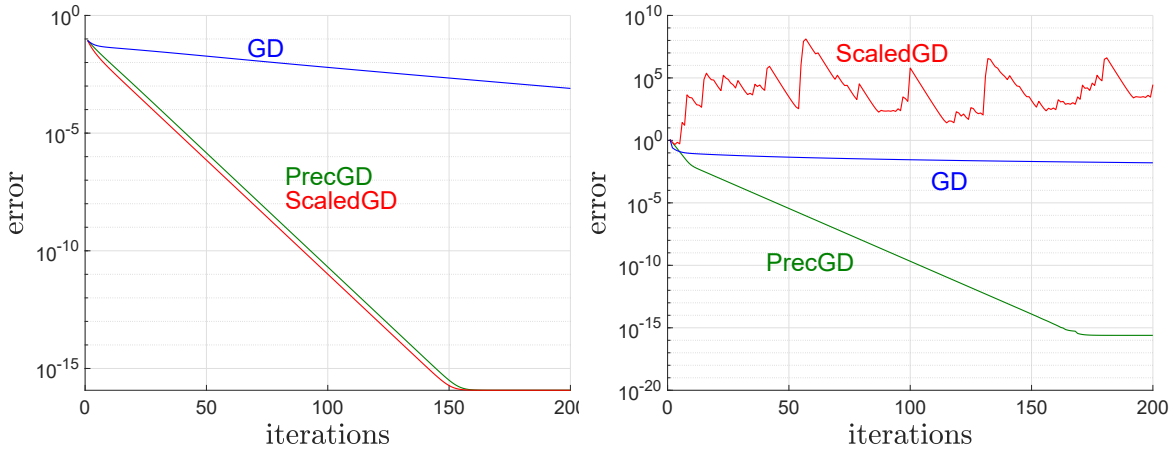

Figure 1: **PrecGD converges linearly in the overparameterized regime.** Comparison of PrecGD against regular gradient descent (GD), and the ScaledGD algorithm of Tong et al. (2020) for an instance of (BM) taken from (Zhang et al., 2018a, 2019b). The same initial points and the same step-size $\alpha = 2 \times 10^{-2}$ was used for all three algorithms. **(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 $r^* = 2$, overparameterization causes gradient descent to slow down to a sublinear rate. ScaledGD also behaves sporadically. Only PrecGD converges linearly to the global minimum.

Global convergence. If the function f can be assumed to admit no spurious local minima, then under a strict saddle assumption (Ge et al., 2015, 2017; Jin et al., 2017), classical gradient descent can be augmented with random perturbations to globally converge to ϵ suboptimality in $O(1/\lambda_r \log(1/\epsilon))$ iterations, starting from any arbitrary initial point. In the overparameterized regime, however, this global guarantee worsens by an exponential factor to poly$(1/\epsilon)$ iterations, due to the loss of local linear convergence.

Instead, under the same benign landscape assumptions on f, we show that PrecGD can be similarly augmented with random perturbations to globally converge to ϵ suboptimality in $O(\log(1/\epsilon))$ iterations, independent of $\lambda_r = 0$ and starting from any arbitrary initial point. A major difficulty here is the need to account for a preconditioner $(X^T X + \eta I)^{-1}$ that changes after each iteration. We prove an $\tilde{O}(1/\delta^2)$ iteration bound to δ approximate second-order stationarity for the perturbed version of PrecGD with a fixed $\eta = \eta_0$, by viewing the preconditioner as a local norm metric that is both well-conditioned and Lipschitz continuous.

Optimality certification. Finally, a crucial advantage of the overparameterizing the search rank $r > r^\star$ is that it allows *a posteriori* certification of convergence to a global minimum. We give a short proof that if X is ϵ suboptimal, then this fact can be explicitly verified by appealing to its second-order stationarity and its rank deficiency. Conversely, we

prove that if X is stuck at a spurious second-order critical point, then this fact can also be explicitly detected via its lack of rank deficiency.

1.2 Related work

Benign landscape. In recent years, there has been significant progress in developing rigorous guarantees on the global optimality of local optimization algorithms like gradient descent (Ge et al., 2016; Bhojanapalli et al., 2016a; Sun et al., 2016; Ge et al., 2017; Sun et al., 2018). For example, Bhojanapalli et al. (2016b) showed that if the underlying convex function ϕ is L-gradient Lipschitz and μ-strongly convex with a sufficiently small condition number L/μ, then f is guaranteed to have no spurious local minima and satisfy the strict saddle property of (Ge et al., 2015) (see also Ge et al. (2017) for an exposition of this result). Where these properties hold, Jin et al. (2017, 2021) showed that gradient descent is rigorously guaranteed (after minor modifications) to converge to ϵ global suboptimality in $O(\log(1/\epsilon))$ iterations, starting from any arbitrary initial point.

Unfortunately, *a priori* global optimality guarantees, which must hold for all initializations before running the algorithm, can often require assumptions that are too strong to be widely applicable in practice Ma and Fattahi (2022a,b). For example, Zhang et al. (2018b, 2019b) found for a global optimality guarantee to be possible, the underlying convex function ϕ must have a condition number of at most $L/\mu < 3$, or else the claim is false due to the existence of a counterexample. And while Zhang (2021) later extended this global optimality guarantee to arbitrarily large condition numbers L/μ by overparameterizing the search rank $r > \max\{r^\star, \frac{1}{4}(L/\mu - 1)^2 r^\star\}$, the result does require suitable strong convexity and optimality assumptions on ϕ. Once these assumptions are lifted, Waldspurger and Waters (2020) showed that a global optimality guarantee based on rank overparameterization would necessarily require $r \geq n$ in general; of course, with such a large search rank, gradient descent would no longer be efficient.

In this paper, we rigorously certify the global optimality of a point X after it has been computed. This kind of *a posteriori* global optimality guarantee may be more useful in practice, because it makes no assumptions on the landscape of the nonconvex function f, nor the algorithm used to compute X. In particular, f may admit many spurious local minima, but an *a posteriori* guarantee will continue to work so long as the algorithm is eventually able to compute a rank deficient second-order point $X^\star$, perhaps after many failures. Our numerical results find that PrecGD is able to broadly achieve an exponential speed-up over classical gradient descent, even when our theoretical assumptions do not hold.

Ill-conditioning and Over-parameterization When minimizing the function $\phi(XX^T)$, ill-conditioning in this problem can come from two separate sources: ill-conditioning of the ground truth $M^\star$ and ill-conditioning of the loss function ϕ. Both can cause gradient descent to slow down (Tu et al., 2016; Zhuo et al., 2021). In this work, we focus on the former kind of ill-conditioning, because it is usually the more serious issue in practice. In applications like matrix completion or matrix sensing, the condition number of the loss function ϕ is entirely driven by the number of samples that the practitioner has collected—the more samples, the better the condition number. Accordingly, any ill-conditioning in ϕ can usually be overcome by collecting more samples. On the other hand, the ill-conditioning in $M^\star$ is inherent to the underlying nature of the data. It cannot be resolved, for example, by collecting more

data. For these real-world applications, it was recently noted that the condition number of $M^\star$ can be as high as 10^8 (Cloninger et al., 2014). Indeed, if the rank of $M^\star$ is unknown or ill-defined, as in the over-parameterized case, the condition number is essentially infinite, and it was previously not known how to make gradient descent converge quickly.

ScaledGD. Our algorithm is closely related to the *scaled gradient descent* or *ScaledGD* algorithm of Tong et al. (2020), which uses a preconditioner of the form $(X^TX)^{-1}$. They prove that ScaledGD is able to maintain a constant-factor decrement after each iteration, even as $\lambda_r = \lambda_{\min}(X^{\star T}X^\star)$ becomes small and $X^\star$ becomes ill-conditioned. However, applying ScaledGD to the overparameterized problem with $\lambda_r = 0$ and a rank deficient $X^\star$ leads to sporadic and inconsistent behavior. The issue is that the admissible step-sizes needed to maintain a constant-factor decrement also shrinks to zero as λ_r goes to zero (we elaborate on this point in detail in Section 6). If we insist on using a constant step-size, then the method will on occasion *increment* after an iteration (see Fig. 1).

Our main result is that regularizing the preconditioner as $(X^TX + \eta I)^{-1}$ with an identity perturbation ηI on the same order of magnitude as the matrix error norm $\|XX^T - X^\star X^\star\|_F$ will maintain the constant-factor decrement of ScaledGD, while also keeping a constant admissible step-size. The resulting iterations, which we call PrecGD, is able to converge linearly, at a rate that is independent of the rank deficiency term λ_r, even as it goes to zero in the overparameterized regime.

Riemann Staircase. An alternative approach for certifying global optimality, often known in the literature as the *Riemann staircase* (Boumal, 2015; Boumal et al., 2016, 2020), is to progressively increase the search rank r only after a second order stationary point has been found. The essential idea is to keep the search rank exactly parameterized $r = r^\star$ during the local optimization phase, and to overparameterize only for the purpose of certifying global optimality. After a full-rank ϵ second order stationary point X has been found in as few as $O(\log(1/\epsilon))$ iterations, we attempt to certify it as ϵ globally suboptimal by increasing the search rank $r_+ = r + 1$ and augmenting $X_+ = [X, 0]$ with a column of zeros. If the augmented X_+ remains ϵ second order stationary under the new search rank r_+, then it is certifiably ϵ globally suboptimal. Otherwise, X_+ is a saddle point; we proceed to reestablish ϵ second order stationarity under the new search rank r_+ by performing another $O(\log(1/\epsilon))$ iterations.

The main issue with the Riemann staircase is that choices of f based on real data often admit global minimizers $X^\star$ whose singular values trail off slowly, for example like a power series $\sigma_i(X^\star) \approx 1/i$ for $i \in \{1, 2, \ldots, r\}$ (see e.g. Kosinski et al. (2013, Fig. S3) for a well-cited example). In this case, the search rank r is always exactly parameterized $r = r^\star$, but the corresponding ϵ second order stationary point X becomes progressively ill-conditioned as r is increased. In practice, ill-conditioning can cause a similarly dramatic slow-down to gradient descent as overparameterization, to the extent that it becomes indistinguishable from sublinear convergence. Indeed, existing implementations of the Riemann staircase are usually based on much more expensive trust-region methods; see e.g. Rosen et al. (2019, 2020).

Notations. We denote $\lambda_i(M)$ as the i-th eigenvalue of M in descending order, as in $\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_n(M)$. Similarly we use $\lambda_{\max}$ and $\lambda_{\min}$ to denote the largest and smallest singular value of a matrix. The matrix inner product is defined $\langle X, Y\rangle \stackrel{\text{def}}{=} \operatorname{tr}(X^TY)$,

and that it induces the Frobenius norm as $\|X\|_F = \sqrt{\langle X, X\rangle}$. The vectorization $\mathrm{vec}(X)$ is the usual column-stacking operation that turns a matrix into a column vector and $\otimes$ denote the Kronecker product. Moreover, we use $\|X\|$ to denote the spectral norm (i.e. the induced 2-norm) of a matrix. We use $\nabla f(X)$ to denote the gradient at X, which is itself a matrix of same dimensions as X. The Hessian $\nabla^2 f(X)$ is defined as the linear operator that satisfies $\nabla^2 f(X)[V] = \lim_{t\to 0} \frac{1}{t}[\nabla f(X+tV) - \nabla f(X)]$ for all V. The symbol $\mathbb{B}(d)$ shows the Euclidean ball of radius d centered at the origin. The notation $\tilde{O}(\cdot)$ is used to hide logarithmic terms in the usual big-O notation.

We always use $\phi(\cdot)$ to denote the original convex objective and $f(X) = \phi(XX^T)$ to denote the factored objective function. We use $M^\star$ to denote the global minimizer of $\phi(\cdot)$. The dimension of $M^\star$ is n, and its rank is $r^\star$. Furthermore, the search rank is denoted by r, which means that X is $n \times r$. We always assume that $\phi(\cdot)$ Lipschitz gradients, and is (μ, r) restricted strongly convex (see next section for precise definition). When necessary, we will also assume that $\phi(\cdot)$ has L_2-Lipschitz Hessians.

2. Convergence Guarantees

2.1 Local convergence

Let $f(X) \overset{\text{def}}{=} \phi(XX^T)$ denote the a Burer–Monteiro cost function defined over $n \times r$ factor matrices X. Under gradient Lipschitz and strong convexity assumptions on ϕ, it is a basic result that convex gradient descent $M_+ = M - \alpha\nabla\phi(M)$ has a linear convergence rate. Under these same assumptions on ϕ, it was shown by Zheng and Lafferty (2015a); Tu et al. (2016) that nonconvex gradient descent $X_+ = X - \alpha\nabla f(X)$ also has a linear convergence rate within a neighborhood of the global minimizer $X^\star$, provided that the unique unconstrained minimizer $M^\star = \arg\min \phi$ is positive semidefinite $M^\star \succeq 0$, and has a rank $r^\star = \mathrm{rank}(M^\star) = r$ that matches the search rank.

Definition 1 (Gradient Lipschitz)**.** *The differentiable function $\phi : \mathbb{R}^{n\times n} \to \mathbb{R}$ is said to be* L_1-gradient Lipschitz *if*

$$\|\nabla\phi(M+E) - \nabla\phi(M)\|_F \le L_1 \cdot \|E\|_F$$

holds for all $M, E \in \mathbb{R}^{n\times n}$.

Definition 2 (Strong convexity)**.** *The twice differentiable function $\phi : \mathbb{R}^{n\times n} \to \mathbb{R}$ is said to be* μ-strongly convex *if*

$$\left\langle \nabla^2\phi(M)[E], E\right\rangle \ge \mu\|E\|_F^2$$

holds for all $M, E \in \mathbb{R}^{n\times n}$. It is said to be (μ, r)-restricted strongly convex *if the above holds for all matrices $M, E \in \mathbb{R}^{n\times n}$ with rank $\le r$.*

Remark 3. *Note that Zheng and Lafferty (2015a); Tu et al. (2016) actually assumed restricted strong convexity, which is a* milder *assumption than the usual notion of strong convexity. In particular, if ϕ is μ-strongly convex, then it is automatically (μ, r)-restricted strongly convex for all $r \le n$. In the context of low-rank matrix optimization, many guarantees made for a strongly convex ϕ can be trivially extended to a restricted strongly convex ϕ, because queries to $\phi(M)$ and its higher derivatives are only made with respect to a low-rank*

matrix argument $M = XX^T$. We also note that in the context of our work, the conditions in Definitions 1 and 2 actually only needs to be imposed on symmetric matrices E. However, for clarity we will follow the standard definition.

If $r^\star$, the rank of the unconstrained minimizer $M^\star \succeq 0$, is strictly less than the search rank r, however, nonconvex gradient descent slows down to a *sublinear* local convergence rate, both in theory and in practice. We emphasize that the sublinear rate manifests in spite of the strong convexity assumption on ϕ; it is purely a consequence of the fact that $r^\star < r$. In this paper, we prove that PrecGD $X_+ = X - \alpha \nabla f(X)(X^TX + \eta I)^{-1}$ has a *linear* local convergence rate, irrespective of the rank $r^\star \leq r$ of the minimizer $M^\star \succeq 0$. Note that a preliminary version of this result restricted to the nonlinear least-squares cost $f(X) = \|\mathcal{A}(XX^T) - b\|^2$ had appeared in a conference paper by the same authors (Zhang et al., 2021).

Theorem 4 (Linear convergence)**.** *Let ϕ be L_1-gradient Lipschitz and $(\mu, 2r)$-restricted strongly convex, and let $M^\star = \arg\min \phi$ satisfy $M^\star = X^\star X^{\star T}$ and $r^\star = \operatorname{rank}(M^\star) \leq r$. Define $f(X) \overset{def}{=} \phi(XX^T)$; if X is sufficiently close to global optimality*

$$f(X) - f(X^\star) \leq \frac{\mu}{2(1+\mu/L_1)} \cdot \frac{\lambda^2_{r^\star}(M^\star)}{2}$$

and if η is bounded from above and below by the distance to the global optimizer

$$C_{\text{lb}} \cdot \|XX^T - M^\star\|_F \leq \eta \leq C_{\text{ub}} \cdot \|XX^T - M^\star\|_F$$

then PrecGD $X_+ = X - \alpha \nabla f(X)(X^TX + \eta I)^{-1}$ converges linearly

$$f(X_+) - f(X^\star) \leq (1 - \alpha \cdot \tau)\,[f(X) - f(X^\star)] \;\; \text{for } \alpha \leq \min\{1, 1/\ell\},$$

with constants

$$\tau = \frac{\mu^2}{2L_1}\left(1 + C_{\text{ub}} \cdot \left(1 + \sqrt{2} + \frac{L_1 + \mu}{\sqrt{L_1 \mu}} \cdot \sqrt{r - r^\star}\right)\right)^{-1}$$
$$\ell = 4L_1 + (2L_1 + 8L_1^2) \cdot C_{\text{lb}}^{-1} + 4L_1^3 \cdot C_{\text{lb}}^{-2}.$$

Theorem 4 suggests choosing the size of the identity perturbation η in the preconditioner $(X^TX + \eta I)^{-1}$ to be within a constant factor of the error norm $\|XX^T - M^\star\|_F$. This condition is reminiscent of trust-region methods, which also requires a similar choice of η to ensure fast convergence towards an optimal point with a degenerate Hessian (see in particular Yamashita and Fukushima 2001, Assumption 2.2 and also Fan and Yuan 2005). The following provides an explicit choice of η that satisfies the condition in Theorem 4 in closed-form.

Corollary 5 (Optimal parameter)**.** *Under the same condition as Theorem 4, we have*

$$\frac{\mu}{\sqrt{2}} \cdot \|XX^T - M^\star\|_F \leq \|\nabla f(X)(X^TX)^{-1/2}\|_F \leq 2L_1 \cdot \|XX^T - M^\star\|_F$$

We provide a proof of Theorem 4 and Corollary 5 in Section 6.

Here, we point out that while PrecGD becomes immune to $\kappa = \lambda_1(M^\star)/\lambda(M^\star)$, the condition number of the ground truth $M^\star$, its dependence on $\chi = L_1/\mu$, the condition number of the convex loss function ϕ, is apparently much worse. Concretely, gradient descent is known to have an iteration complexity of $O(\kappa \cdot \chi \cdot \log(1/\epsilon))$, while Theorem 4 says that PrecGD has an iteration complexity of $O(\chi^4 \cdot \log(1/\epsilon))$. (Note that the constants C_{lb} and C_{ub} in Theorem 4 have "units" of μ and L_1 respectively, and therefore $\ell = O(L_1\chi^2)$.)

In our numerical experiments, however, both methods have the same dependence on χ. In other words, the iteration complexity of PrecGD is strictly better than GD in practice. Therefore, we believe that with a more refined analysis, our dependence on the latter can also be improved. However, as we discussed in the introduction, the ill-conditioning of $M^\star$ is usually much more serious in practice, so it is the main focus of our work.

2.2 Global convergence

If initialized from an arbitrary point X_0, gradient descent can become stuck at a suboptimal point X with small gradient norm $\|\nabla f(X)\|_F \le \delta$. A particularly simple way to escape such a point is to perturb the current iterate X by a small amount of random noise. Augmenting classical gradient descent with random perturbations in this manner, Jin et al. (2017, 2021) proved convergence to an δ approximate second-order stationary point X satisfying $\|\nabla f(X)\|_F \le \delta$ and $\nabla f(X) \succeq -\sqrt{\delta} \cdot I$ in at most $O(1/\delta^2 \log^4(nr/\delta))$ iterations, assuming that the convex function ϕ is gradient Lipschitz and also Hessian Lipschitz.

Definition 6 (Hessian Lipschitz)**.** *The twice-differentiable function $\phi : \mathbb{R}^{n\times n} \to \mathbb{R}$ is said to be L_2-*Hessian Lipschitz *if*

$$\|\nabla^2\phi(M)[E] - \nabla^2\phi(M')[E]\|_F \le L_2 \cdot \|E\|_F \cdot \|M - M'\|_F$$

holds for all $M, M', E \in \mathbb{R}^{n\times n}$.

It turns out that certain choices of f satisfy the property that every δ approximate second-order stationary point X lies poly(δ)-close to a global minimum (Sun et al., 2016, 2018; Bhojanapalli et al., 2016b). The following definition is adapted from Ge et al. (2015); see also (Ge et al., 2017; Jin et al., 2017).

Definition 7 (Strict saddle property)**.** *The function f is said to be* $(\epsilon_g, \epsilon_H, \rho)$-strict saddle *if at least one of the following holds for every X:*

- $\|\nabla f(X)\| \ge \epsilon_g$*;*
- $\lambda_{\min}(\nabla f(X)) \le -\epsilon_H$*;*
- *There exists Z satisfying $\nabla f(Z) = 0$ and $\nabla^2 f(Z) \succeq 0$ such that $\|X - Z\|_F \le \rho$.*

The function is said to be $(\epsilon_g, \epsilon_H, \rho)$-global strict saddle *if it is $(\epsilon_g, \epsilon_H, \rho)$-strict saddle, and that all Z that satisfy $\nabla f(Z) = 0$ and $\nabla^2 f(Z) \succeq 0$ also satisfy $f(Z) = f(X^\star)$.*

Assuming that $f(X) \overset{\text{def}}{=} \phi(XX^T)$ is $(\epsilon_g, \epsilon_H, \rho)$-global strict saddle, Jin et al. (2017) used perturbed gradient descent to arrive within a ρ-local neighborhood, after which it takes

gradient descent another $O(1/\lambda_r \log(\rho/\epsilon))$ iterations to converge to ϵ global suboptimality. Viewing $\epsilon_g, \epsilon_H, \rho$ as constants with respect to ϵ, the combined method globally converges to ϵ suboptimality in $O(1/\lambda_r \log(1/\epsilon))$ iterations, as if f were a smooth and strongly convex function.

If the rank $r^\star < r$ is strictly less than the search rank r, however, the global guarantee for gradient descent worsens by an exponential factor to $\mathrm{poly}(1/\epsilon)$ iterations, due to the loss of local linear convergence. Inspired by Jin et al. (2017), we consider augmenting PrecGD with random perturbations, in order to arrive within a local neighborhood for which our linear convergence result (Theorem 4) becomes valid. Concretely, we consider *perturbed* PrecGD or PPrecGD, defined as

$$X_{k+1} = X_k - \alpha[\nabla f(X_k)(X_k^T X_k + \eta I)^{-1} + \zeta_k], \tag{PPrecGD}$$

where we *fix* the value of the regularization parameter $\eta > 0$ and apply a random perturbation ζ_k whenever the gradient norm becomes small:

$$\begin{cases} \zeta_k \sim \mathbb{B}(\beta) & \|\nabla f(X_k)(X_k^T X_k + \eta_{\mathrm{fix}} I)^{-1/2}\|_F \le \epsilon, \text{ and } k \ge k_{\mathrm{last}} + \mathcal{T}, \\ \zeta_k = 0 & \text{otherwise.} \end{cases}$$

Here, k_{last} denotes the last iteration index for which a random perturbation was made. The condition $k \ge k_{\mathrm{last}} + \mathcal{T}$ ensures that the algorithm takes at least $\mathcal{T}$ iterations before making a new perturbation.

The algorithm parameters are the step-size $\alpha > 0$, the perturbation radius $\beta > 0$, the period of perturbation $\mathcal{T}$, the fixed regularization parameter $\eta > 0$, and the accuracy threshold $\epsilon > 0$. We show that PPrecGD is guaranteed to converge to an ϵ second-order stationary point, provided that the following sublevel set of ϕ (which contains all iterates $X_0, X_1, \ldots, X_k$) is bounded:

$$\mathcal{X} = \{X \in \mathbb{R}^{n\times r} : \phi(XX^T) \le \phi(X_0 X_0^T) + 2\sqrt{\|X_0\|_F^2 + \eta} \cdot \alpha\beta\epsilon\}.$$

Let $\Gamma = \max_{X\in\mathcal{X}} \|X\|_F$. Below, the notation $\tilde{O}(\cdot)$ hides polylogarithmic factors in the algorithm and function parameters η, L_1, L_2, Γ, the dimensionality n, r, the final accuracy $1/\epsilon$, and the initial suboptimality $f(X_0) - f(X^\star)$. The proof is given in Section 7.

Theorem 8 (Approximate second-order optimality)**.** *Let ϕ be L_1-gradient and L_2-Hessian Lipschitz. Define $f(X) \overset{def}{=} \phi(XX^T)$ and let $X^\star = \arg\min f$. For any $\epsilon = O(1/(L_d\sqrt{\Gamma^2+\eta}))$ and with an overwhelming probability, PPrecGD with parameters $\alpha = \eta/\ell_1$, $\beta = \tilde{O}(\epsilon/L_d)$, and $\mathcal{T} = \tilde{O}(L_1\Gamma^2/(\eta\sqrt{L_d\epsilon}))$ converges to a point X that satisfies*

$$\langle \nabla f(X), V\rangle \le \epsilon \cdot \|V\|_{X,\eta}, \quad \left\langle \nabla^2 f(X)[V], V\right\rangle \ge -\sqrt{L_d\epsilon} \cdot \|V\|_{X,\eta}^2 \quad \textit{for all } V, \tag{3}$$

where $\|V\|_{X,\eta} \overset{def}{=} \|V(X^TX + \eta I)^{1/2}\|_F$ in at most

$$\tilde{O}\left(\frac{\ell_1 \cdot [f(X_0) - f(X^\star)]}{\eta^2 \cdot \epsilon^2}\right) \textit{ iterations}$$

where $L_d = 5\max\{\ell_2, 2\Gamma\ell_1\sqrt{\Gamma^2+\eta}\}/\eta^{2.5}$, $\ell_1 = 9\Gamma^2 L_1$, $\ell_2 = (4\Gamma+2)L_1 + 4\Gamma^2 L_2$.

In Theorem 8, the total number of iterations it takes to converge to an approximate second-order stationary point depends additionally on Γ, the maximal radius of the iterates. The factor of Γ is largely an artifact of the proof technique, and does not appear in the practical performance of the algorithm. Previous work on naive gradient descent (see Theorem 8 of Jin et al. (2017)) also introduced a similar factor. Clearly, Γ is finite if $\phi(\cdot)$ is coercive, i.e., it diverges to ∞ as $\|X\|_F \to \infty$. (See Lemma 29.) In many statistical and machine learning applications, the loss function is purposely chosen to be coercive.

Assuming that f is $(\epsilon_g, \epsilon_H, \rho)$-global strict saddle, we use PPrecGD to arrive within a ρ-local neighborhood of the global minimum, and then switch to PrecGD for another $O(\log(\rho/\epsilon))$ iterations to converge to ϵ global suboptimality due to Theorem 4. (If the search rank is overparameterized $r > r^\star$, then the switching condition can be explicitly detected using Proposition 11 in the following section.) Below, we use $\tilde{O}(\cdot)$ to additionally hide polynomial factors in L_1, L_2, μ, Γ, while exposing all dependencies on final accuracy $1/\epsilon$ and the smallest nonzero eigenvalue $\lambda_{r^\star}(M^\star)$.

Corollary 9 (Global convergence)**.** *Let ϕ be L_1-gradient Lipschitz and L_2-Hessian Lipschitz and $(\mu, 2r)$-restricted strongly convex, and let $M^\star = \arg\min \phi$ satisfy $M^\star = X^\star X^{\star T}$ and $r^\star = \operatorname{rank}(M^\star) < r$. Suppose that $f(X) \overset{def}{=} \phi(XX^T)$ satisfies $(\epsilon_g, \epsilon_H, \rho)$-global strict saddle with*

$$\frac{1}{\Gamma^2} \cdot \epsilon_g + \epsilon_H + 4L_1 \cdot \rho^2 \leq \frac{\mu}{1 + \mu/L_1} \cdot \frac{\lambda^2_{r^\star}(M^\star)}{\operatorname{tr}(M^\star)},$$

Then, do the following:

1. *(Global phase) Run PPrecGD with a fixed $\eta = \eta_0 \leq \Gamma^2$ until $\|\nabla f(X_k)\|_F \leq \epsilon_g$, and $\lambda_{\min}(\nabla^2 f(X_k)) \geq -\epsilon_H$, and $\lambda_{\min}(X_k^T X_k) \leq \rho$;*
2. *(Local phase) Run PrecGD with $\eta = \|\nabla f(X)(X^T X)^{-1}\|_F$ and $\alpha = 1/\ell$.*

The combined algorithm arrives at a point X satisfying $f(X) - f(X^\star) \leq \epsilon$ in at most

$$\tilde{O}\left(\frac{f(X_0) - f(X^\star)}{\eta_0^2} \cdot \left(\frac{1}{\epsilon_g^2} + \frac{1}{\epsilon_H^4}\right) + \log\left(\frac{\lambda^2_{r^\star}(M^\star)}{\epsilon}\right)\right) \text{ iterations.}$$

Corollary 9 follows by running PPrecGD until it arrives at an (ϵ_g, ϵ_H)-second-order stationary point X_k (Theorem 8), and then using the global strict saddle property (Definition 7) to argue that X_k is also rank deficient, with $\lambda_{\min}(X_k^T X_k) \leq \rho^2$ via Weyl's inequality. It follows from second-order optimality and rank deficiency that X_k is sufficiently close to global optimality (Proposition 10 below), and therefore switching to PrecGD results in linear convergence (Theorem 4). Viewing $\epsilon_g, \epsilon_H, \rho$ as constants, the combined method globally converges to ϵ suboptimality in $O(\log(1/\epsilon))$ iterations, as if f were a smooth and strongly convex function, even in the overparameterized regime with $r > r^\star$.

Here, we point out that the strict saddle property (Definition 7) is usually defined for a second-order point measured in the Euclidean norm $\|\cdot\|_F$, but that Theorem 8 proves convergence to a second-order point measured in the local norm $\|\cdot\|_{X,\eta}$. Clearly, for a fixed $\eta = \eta_{\text{fix}}$, the two notions are equivalent up to a conversion factor. In deciding when to switch from PPrecGD to PrecGD, Corollary 9 uses the Euclidean norm (via Proposition 10 below) to remain consistent with the strict saddle property. In practice, however, it should

be less conservative to decide using the local norm (via Proposition 11 below), as this is the preferred norm that the algorithm tries to optimize.

3. Certifying Global Optimality via Rank Deficiency

We now turn to the problem of certifying the global optimality of an X computed using PrecGD by appealing to its rank deficiency. We begin by rigorously stating the global optimality guarantee previously quoted in (2). The core argument actually dates back to Burer and Monteiro (2005, Theorem 4.1) (and has also appeared in Journée et al. (2010) and Boumal et al. (2016, 2020)) but we restate it here with a shorter proof in order to convince the reader of its correctness.

Proposition 10 (Certificate of global optimality)**.** *Let ϕ be twice differentiable and convex and let $f(X) \overset{def}{=} \phi(XX^T)$. If X satisfies $\lambda_{\min}(X^TX) \leq \epsilon_\lambda$ and*

$$\langle \nabla f(X), V\rangle \leq \epsilon_g \cdot \|V\|_F, \quad \left\langle \nabla^2 f(X)[V], V\right\rangle \geq -\epsilon_H \cdot \|V\|_F^2 \quad \textit{for all } V,$$

where $\epsilon_g, \epsilon_H, \epsilon_\lambda \geq 0$, then X has suboptimality

$$f(X) - f(X^\star) \leq C_g \cdot \epsilon_g + C_H \cdot \epsilon_H + C_\lambda \cdot \epsilon_\lambda.$$

where $C_g = \frac{1}{2}\|X\|_F$ and $C_H = \frac{1}{2}\|X^\star\|_F^2$ and $C_\lambda = 2\|\nabla^2\phi(XX^T)\|\|X^\star\|_F^2$.

Proof Let (u_r, v_r, σ_r) the r-th singular value triple of X, i.e. we have $Xv_r = \sigma_r u_r$ with $\|v_r\| = \|u_r\| = 1$ and $\sigma_r^2 = \lambda_{\min}(X^TX)$. For $M = XX^T$ and $M^\star = X^\star X^{\star T}$, the convexity of ϕ implies $\phi(M^\star) \geq \phi(M) + \langle \nabla\phi(M), M^\star - M\rangle$ and therefore

$$f(X) - f(X^\star) = \phi(M) - \phi(M^\star) \leq \langle \nabla\phi(M), M\rangle - \lambda_{\min}[\nabla\phi(M)] \cdot \operatorname{tr}(M^\star).$$

Substituting $V = X$ into the first-order optimality conditions, as in

$$\langle \nabla f(X), V\rangle = 2\left\langle \nabla\phi(XX^T)X, X\right\rangle \leq \epsilon_g\|V\|_F = \epsilon_g \cdot 2C_g$$

yields $\langle \nabla\phi(M), M\rangle \leq C_g \cdot \epsilon$. Substituting $V = yv_r^T$ with an arbitrary $y \in \mathbb{R}^n$ with $\|y\| = 1$ into the second-order conditions yields

$$\begin{aligned}\left\langle \nabla^2 f(X)[V], V\right\rangle &\leq 2\left\langle \nabla\phi(XX^T), VV^T\right\rangle + \|\nabla^2\phi(XX^T)\| \cdot \|XV^T + VX^T\|_F^2 \\ &= 2y^T\nabla\phi(XX^T)y + \|\nabla^2\phi(XX^T)\| \cdot \sigma_r^2 \cdot \|u_r y^T + yu_r^T\|_F^2\end{aligned}$$

which combined with $\left\langle \nabla^2 f(X)[V], V\right\rangle \geq -\epsilon_H\|V\|_F^2 = -\epsilon_H$ gives

$$-y^T\nabla\phi(XX^T)y \leq \frac{1}{2}\epsilon_H + 2\|\nabla^2\phi(XX^T)\| \cdot \sigma_r^2$$

and therefore $-\lambda_{\min}[\nabla\phi(M)] \leq \frac{1}{2}\epsilon_H + C_\lambda \cdot \lambda_{\min}(X^TX)$. ■

Proposition 10 can also be rederived with respect to the local norm in Theorem 8. We omit the proof of the following as it is essentially identical to that of Proposition 10.

Proposition 11 (Global certificate in local norm)**.** *Let ϕ be twice differentiable and convex and let $f(X) \overset{def}{=} \phi(XX^T)$. If X satisfies $\lambda_{\min}(X^TX) \leq \epsilon_\lambda$ and*

$$\langle \nabla f(X), V\rangle \leq \epsilon_g \cdot \|V\|_{X,\eta}, \quad \left\langle \nabla^2 f(X)[V], V\right\rangle \geq -\epsilon_H \cdot \|V\|_{X,\eta}^2 \quad \textit{for all } V,$$

where $\|V\|_{X,\eta} \overset{def}{=} \|V(X^TX+\eta I)^{-1/2}\|_F$ and $\epsilon_g, \epsilon_H, \epsilon_\lambda \geq 0$, then X has suboptimality

$$f(X) - f(X^\star) \leq C_g \cdot \epsilon_g + C_H \cdot \epsilon_H \cdot (\epsilon_\lambda + \eta) + C_\lambda \cdot \epsilon_\lambda$$

where $C_g = \frac{1}{2}\sqrt{\|X^TX\|_F^2 + \eta\|X\|_F^2}$ and $C_H = \frac{1}{2}\|X^\star\|_F^2$ and $C_\lambda = 2\|\nabla^2\phi(XX^T)\|\|X^\star\|_F^2$.

Remark 12. *Under the same conditions as Theorem 8, it follows that $\|X\|_F, \|X^\star\|_F \leq \Gamma$ and $\|\nabla^2\phi(XX^T)\| \leq L_1$.*

Now let us explain how we can use PrecGD to solve an instance of (BM) to an X with provable global optimality via either Proposition 10 or Proposition 11. First, after overparameterizing the search rank $r > r^\star$, we run PPrecGD with a fixed parameter $\eta > 0$ until we reach the neighborhood of a global minimizer $X^\star$ where Theorem 4 holds. The following result says that this condition can always be detected by checking Proposition 10. Afterwards, we can switch to PrecGD with a variable parameter $\eta = \|\nabla f(X)(X^TX)^{-1}\|_F$ and expect linear convergence towards to global minimum.

Corollary 13 (Certifiability of near-global minimizers)**.** *Under the same condition as Theorem 4, let X satisfy $f(X) - f(X^\star) \leq \frac{1}{2}\mu\epsilon^2$. If $r > r^\star$, then X also satisfies*

$$\|\nabla f(X)\|_F \leq 2L_1\|X\|_F \cdot \epsilon, \quad \lambda_{\min}(\nabla^2 f(X)) \geq -L_1 \cdot \epsilon, \quad \lambda_{\min}(X^TX) \leq \epsilon.$$

Proof It follows immediately from $\frac{1}{2}\mu\epsilon^2 \geq f(X) - f(X^\star) \geq \frac{1}{2}\mu\|XX^T - M^\star\|_F$ in Lemma 16, which yields $\|\nabla\phi(XX^T)\|_F \leq L_1\epsilon$ via gradient Lipschitzness and $\lambda_{\min}(X^TX) = \lambda_r(XX^T) \leq \epsilon$ via Weyl's inequality. Finally, to see why the second statement holds, note that the Hessian of $f(X)$ can be written as

$$\left\langle V, \nabla^2 f(X)[V]\right\rangle = \left\langle V, \nabla\phi(XX^T)V\right\rangle + \left\langle V, \nabla^2\phi(XX^T)\left[XV^T + VX^T\right]X\right\rangle.$$

Since ϕ is convex, the second term is always non-negative. Thus the second statement follows from the fact that $\|\nabla\phi(XX^T)\|_F \leq L_1\epsilon$. ■

On the other hand, if PPrecGD becomes stuck within a neighborhood of a spurious local minimum or nonstrict saddle point Z, then this fact can also be explicitly detected by numerically evaluating the rank deficiency parameter $\epsilon_\lambda = \lambda_{\min}(X^TX)$. Note that if $\|X - Z\|_F \leq \rho$, then it follows from Weyl's inequality that $\lambda_{\min}^{1/2}(Z^TZ) \geq \lambda_{\min}^{1/2}(X^TX) - \rho$.

Corollary 14 (Spurious points have high rank)**.** *Under the same condition as Theorem 4, let Z satisfy $\nabla f(Z) = 0$ and $\nabla^2 f(Z) \succeq 0$. If $r > r^\star$, then we have*

$$f(Z) > f(X^\star) \quad \iff \quad \lambda_{\min}(Z^TZ) > \frac{\mu}{4\cdot(L_1+\mu)} \cdot \frac{\lambda_{r^\star}^2(M^\star)}{\operatorname{tr}(M^\star)}.$$

Proof It follows from Theorem 4 that any point Z that satisfies $\nabla f(Z) = 0$ within the neighborhood $f(Z) - f(X^\star) \leq R = \frac{\mu}{2\cdot(1+\mu/L_1)}\lambda^2_{r^\star}(M^\star)$ must actually be globally optimal $f(Z) = f(X^\star)$. Therefore, any suboptimal Z with $\nabla f(Z) = 0$ and $\nabla^2 f(Z) \succeq 0$ and $f(Z) > f(X^\star)$ must lie outside of this neighborhood, as in $f(Z) - f(X^\star) > R$. It follows from Proposition 10 that Z must satisfy:

$$R < f(Z) - f(X^\star) \leq C_\lambda \cdot \lambda_{\min}(Z^T Z) \leq 2L_1 \operatorname{tr}(M^\star) \cdot \lambda_{\min}(Z^T Z).$$

Conversely, if $r > r^\star = \operatorname{rank}(M^\star)$, then Z is globally optimal $f(Z) = f(X^\star)$ if and only if it is rank deficient, as in $\lambda_{\min}(Z^T Z) = 0$. ■

Finally, we turn to the practical problem of evaluating the parameters in Proposition 10. It is straightforward to see that it costs $O(nr^2 + r^3)$ time to compute the gradient norm term $\epsilon_g = \|\nabla f(X)\|_F$ and the rank deficiency term $\lambda_{\min}(X^T X)$, after computing the nonconvex gradient $\nabla f(X)$ in $n \cdot \operatorname{poly}(r)$ time via the gradient oracle. To compute the Hessian curvature $\epsilon_H = -\lambda_{\min}[\nabla^2 f(X)]$ without explicitly forming the $nr \times nr$ Hessian matrix, we suggest using a shifted power iteration

$$V_{k+1} = \tilde{V}_k / \|\tilde{V}_k\|_F \qquad \text{where } \tilde{V}_k = \lambda V_k - \nabla^2 f(X)[V_k],$$

where we roughly choose the shift parameter λ so that $\lambda \geq \lambda_{\max}[\nabla^2 f(X)]$ and approximate each Hessian matrix-vector product using finite differences

$$\nabla^2 f(X)[V] \approx \frac{1}{t}[\nabla f(X + tV) - \nabla f(X)].$$

The Rayleigh quotient converges linearly, achieving δ-accuracy in $O(\log(1/\delta))$ iterations (Kaniel, 1966; Paige, 1971; Saad, 1980). Each iteration requires a single nonconvex gradient evaluation $\nabla f(X + tV)$, which we have assumed to cost $n \cdot \operatorname{poly}(r)$ time. Technically, linear convergence to $\lambda_{\min}(\nabla^2 f(X))$ requires the eigenvalue to be simple and well separated. If instead the eigenvalue has multiplicity $b > 1$ (or lies within a well-separated cluster of b eigenvalues), then we use a block power iteration with block-size b to recover linear convergence to $\lambda_{\min}[\nabla^2 f(X)]$, with an increased per-iteration cost of $O(nb)$ time (Saad, 1980).

4. Preliminaries

Our analysis will assume that ϕ is L-gradient Lipschitz and $(\mu, 2r)$-restricted strongly convex, meaning that

$$\mu\|E\|_F^2 \leq \left\langle \nabla^2 \phi(M)[E], E \right\rangle \leq L\|E\|_F^2 \tag{4}$$

in which the lower-bound is restricted over matrices M, E whose $\operatorname{rank}(M) \leq 2r$ and $\operatorname{rank}(E) \leq 2r$. (See Definition 1 and Definition 2.) The purpose of these assumptions is to render the function ϕ well-conditioned, so that its suboptimality can serve as a good approximation for the matrix error norm

$$f(X) - f(X^\star) \approx \|XX^T - M^\star\|_F^2 \text{ up to a constant.}$$

In turn, we would also expect the nonconvex gradient $\nabla f(X)$ to be closely related to the gradient of the matrix error norm $\|XX^T - M^\star\|_F^2$ taken with respect to X. To make these arguments rigorous, we will need the following lemma from Li et al. (2019, Proposition 2.1). The proof is a straightforward extension of Candes (2008, Lemma 2.1).

Lemma 15 (Preservation of inner product)**.** *Let ϕ be L-gradient Lipschitz and (μ, r)-restricted strongly convex. Then, we have*

$$\left| \frac{2}{\mu + L} \left\langle \nabla^2 \phi(M)[E], F \right\rangle - \langle E, F \rangle \right| \leq \frac{L - \mu}{L + \mu} \|E\|_F \|F\|_F$$

for all $\operatorname{rank}(M) \leq r$ *and* $\operatorname{rank}(E + F) \leq r$.

Lemma 16 (Preservation of error norm)**.** *Let ϕ be L-gradient Lipschitz and $(\mu, 2r)$-restricted strongly convex. Let $M^\star = \arg\min \phi$ satisfy $M^\star \succeq 0$ and $\operatorname{rank}(M^\star) \leq r$. Define $f(X) \overset{def}{=} \phi(XX^T)$ and let $X^\star = \arg\min f$. Then f satisfies*

$$\frac{1}{2}\mu \|XX^T - M^\star\|_F^2 \leq f(X) - f(X^\star) \leq \frac{1}{2} L \|XX^T - M^\star\|_F^2$$

for all $\operatorname{rank}(M) \leq r$.

Lemma 17 (Preservation of error gradient)**.** *Under the same conditions as Lemma 16, we have*

$$\|\nabla f(X)\|_F \geq \nu \cdot \max_{\|Y\|_F = 1} \left[\left\langle E, XY^T + YX^T \right\rangle - \delta \|E\|_F \|XY^T + YX^T\|_F \right], \tag{5a}$$

where $\nu = \frac{1}{2}(\mu + L)$ and $\delta = \frac{L - \mu}{L + \mu}$ and $E = XX^T - M^\star$.

Proof Let $Y^\star$ denote a maximizer for the right-hand side of (5a), and let Π denote the orthogonal projector onto

$$\operatorname{range}(X) + \operatorname{range}(X^\star) = \{Xu + X^\star v : u, v \in \mathbb{R}^r\}.$$

(Explicitly, $\Pi = QQ^T$ where $Q = \operatorname{orth}([X, X^\star])$.) We claim that the projected matrix $Y = \Pi Y^\star$ is also a maximizer. Note that, by the definition of Π, we have $X = \Pi X$ and $E = \Pi E \Pi$. It follows that

$$\begin{aligned} \left\langle XY^T + YX^T, E \right\rangle &= \left\langle \Pi \left[XY^{\star T} + Y^\star X^T \right] \Pi, E \right\rangle \\ &= \left\langle XY^{\star T} + Y^\star X^T, \Pi E \Pi \right\rangle = \left\langle XY^{\star T} + Y^\star X^T, E \right\rangle, \end{aligned}$$

and

$$\|XY^T + YX^T\|_F = \|\Pi \left[XY^{\star T} + Y^\star X^T \right] \Pi\|_F \leq \|XY^{\star T} + Y^\star X^T\|_F,$$

and $\|Y\|_F = \|\Pi Y^\star\|_F \leq \|Y^\star\|_F \leq 1$. Therefore, we conclude that Y is feasible and achieves the same optimal value as the maximizer $Y^\star$.

Now, let $Y^\star = \Pi Y^\star$ without loss of generality due to the above. We evoke the lower-bound in Lemma 15 and $\nabla\phi(M^\star) = 0$ to obtain the following

$$\begin{aligned}\langle \nabla f(X), Y^\star\rangle &= \left\langle \nabla\phi(XX^T) - \nabla\phi(M^\star), XY^{\star T} + Y^\star X^T\right\rangle \\ &= \int_0^1 \left\langle \nabla^2\phi(M^\star + tE)[E], XY^{\star T} + Y^\star X^T\right\rangle \mathrm{d}t \\ &\geq \nu\cdot\left[\left\langle E, XY^{\star T} + Y^\star X^T\right\rangle - \delta\cdot\|E\|_F\cdot\|XY^{\star T} + Y^\star X^T\|_F\right]\end{aligned}$$

where we crucially note that $\operatorname{rank}(XY^{\star T} + Y^\star X^T \pm E) \leq 2r$ because $XY^{\star T} = \Pi XY^{\star T}\Pi$ and $E = \Pi E\Pi$ and $\operatorname{rank}(\Pi) \leq \operatorname{rank}(X) + \operatorname{rank}(X^\star) \leq 2r$. We conclude that (5a) is true, because $Y^\star$ is a maximizer for the right-hand side of (5a). ■

5. Local Sublinear Convergence of Gradient Descent

In order to explain why PrecGD is able to maintain linear convergence in the overparameterized regime $r > r^\star$, we must first understand why gradient descent slows down to sublinear convergence. In this paper, we focus on a property known as *gradient dominance* (Polyak, 1963; Nesterov and Polyak, 2006) or the *Polyak-Lojasiewicz* inequality (Lojasiewicz, 1963), which is a simple, well-known sufficient condition for linear convergence. Here, we use the degree-2 definition from Nesterov and Polyak (2006, Definition 3).

Definition 18. *A function f is said to satisfy* gradient dominance *(in the Euclidean norm) if it attains a global minimum $f^\star = f(X^\star)$ at some point $X^\star$ and we have*

$$f(X) - f^\star \leq R \quad\implies\quad \tau\cdot[f(X) - f^\star] \leq \frac{1}{2}\|\nabla f(X)\|_F^2 \tag{6}$$

for a radius constant $R > 0$ and dominance constant $\tau > 0$.

If the function f is additionally ℓ-gradient Lipschitz, as in

$$f(X + \alpha V) \leq f(X) + \alpha\langle\nabla f(X), V\rangle + \frac{\ell}{2}\alpha^2\|V\|_F^2,$$

then it follows that the amount of progress made by an iteration of gradient descent $X_+ = X - \alpha\nabla f(X)$ is proportional to the gradient norm squared:

$$\begin{aligned}f(X_+) &\leq f(X) - \alpha\langle\nabla f(X), \nabla f(X)\rangle + \frac{\ell}{2}\alpha^2\|\nabla f(X)\|_F^2 \\ &= f(X) - \alpha\left(1 - \frac{\ell}{2}\alpha\right)\|\nabla f(X)\|_F^2 \\ &\leq f(X) - \frac{\alpha}{2}\|\nabla f(X)\|_F^2 \qquad \text{with } \alpha \leq \frac{1}{\ell}.\end{aligned}$$

The purpose of gradient dominance (6), therefore, is to ensure that the gradient norm remains large enough for good progress to be made. Substituting (6) yields

$$f(X_+) - f^\star \leq (1 - \tau\alpha)\cdot(f(X) - f^\star) \qquad \text{with } \alpha \leq \frac{1}{\ell}. \tag{7}$$

Starting from an initial point X_0 within the radius $f(X_0) - f^\star \le R$, it follows that gradient descent $X_{k+1} = X_k - \frac{1}{\ell}\nabla f(X_k)$ converges to an ϵ-suboptimal point X_k that satisfies $f(X_k) - f^\star \le \epsilon$ in at most $k = O((\tau/\ell)\log(R/\epsilon))$ iterations.

The nonconvex objective $f(X) \stackrel{\text{def}}{=} \phi(XX^T)$ associated with a well-conditioned convex objective ϕ is easily shown to satisfy gradient dominance (6) in the exactly parameterized regime $r = r^\star$, for example by manipulating existing results on local strong convexity (Sun and Luo, 2016) (Chi et al., 2019, Lemma 4). In the overparameterized case $r > r^\star$, however, local strong convexity is lost, and gradient dominance can fail to hold.

The goal of this section is to elucidate this failure mechanism, in order to motivate the "fix" encompassed by PrecGD. We begin by considering a specific instance of the following nonconvex objective f_0, corresponding to a perfectly conditioned quadratic objective ϕ_0:

$$f_0(X) \stackrel{\text{def}}{=} \phi_0(XX^T) = f_0^\star + \frac{1}{2}\|XX^T - M^\star\|_F^2. \tag{8}$$

The associated gradient norm has a variational characterization

$$\|\nabla f_0(X)\|_F = \max_{\|Y\|_F=1} \langle \nabla f_0(X), Y\rangle = \max_{\|Y\|_F=1} \left\langle XX^T - M^\star, XY^T + YX^T\right\rangle, \tag{9}$$

which we can interpret as a *projection* from the error vector $XX^T - M^\star$ onto the linear subspace $\{XY^T + YX^T : Y \in \mathbb{R}^{n\times r}\}$, as in

$$\|\nabla f_0(X)\|_F = \|XX^T - M^\star\|_F \|XY^{\star T} + Y^\star X^T\|_F \cos\theta. \tag{10}$$

Here, the incidence angle θ is defined

$$\cos\theta = \max_{Y\in\mathbb{R}^{n\times r}} \frac{\left\langle XX^T - M^\star, XY^T + YX^T\right\rangle}{\|XX^T - M^\star\|_F \|XY^T + YX^T\|_F}, \tag{11}$$

and $Y^\star$ is a corresponding maximizer for (11) scaled to satisfy $\|Y^\star\|_F = 1$. Substituting the suboptimality $f_0(X) - f_0^\star$ in place of the error norm $\|XX^T - M^\star\|_F$ via Lemma 16 yields a critical identity:

$$\frac{1}{2}\|\nabla f_0(X)\|_F^2 = \|XY^{\star T} + Y^\star X^T\|_F^2 \cdot \cos^2\theta \cdot [f_0(X) - f_0^\star]. \tag{12}$$

The loss of gradient dominance implies that at least one of the two terms $\|XY^{\star T} + Y^\star X^T\|_F$ and $\cos\theta$ in (12) must decay to zero as gradient descent makes progress towards the solution.

The term $\cos\theta$ in (12) becomes small if the error $XX^T - M^\star$ becomes *poorly aligned* to the linear subspace $\{XY^T + YX^T : Y \in \mathbb{R}^{n\times r}\}$. In fact, this failure mechanism cannot occur within a sufficiently small neighborhood of the ground truth, due to the following key lemma. Its proof is technical, and is deferred to Appendix A.

Lemma 19 (Basis alignment)**.** *For $M^\star \in \mathbb{R}^{n\times n}$, $M^\star \succeq 0$, suppose that $X \in \mathbb{R}^{n\times r}$ satisfies $\|XX^T - M^\star\|_F \le \rho\lambda_{r^\star}(M^\star)$ with $r^\star = \operatorname{rank}(M^\star)$ and $\rho \le 1/\sqrt{2}$. Then the incidence angle θ defined in (11) satisfies*

$$\sin\theta = \frac{\|(I - XX^\dagger)M^\star(I - XX^\dagger)\|_F}{\|XX^T - M^\star\|_F} \le \frac{1}{\sqrt{2}}\frac{\rho}{\sqrt{1-\rho^2}} \tag{13}$$

where $\dagger$ denotes the pseudoinverse.

The term $\|XY^{\star T} + Y^\star X^T\|_F$ in (12) becomes small if the error vector $XX^T - M^\star$ concentrates itself within the *ill-conditioned directions* of $\{XY^T + YX^T : Y \in \mathbb{R}^{n\times r}\}$. In particular, if $XX^T - M^\star$ lies entirely with the subspace $\{u_r y^T + y u_r^T : y \in \mathbb{R}^n\}$ associated with the r-th eigenpair (λ_r, u_r) of the matrix XX^T, and if the corresponding eigenvalue $\lambda_r = \lambda_{\min}(X^T X)$ decays towards zero, then the term $\|XY^{\star T} + Y^\star X^T\|_F$ must also decay towards zero. The following lemma provides a lower-bound on $\|XY^{\star T} + Y^\star X^T\|_F$ by accounting for this mechanism.

Lemma 20 (Basis scaling)**.** *For any $H \in \mathbb{R}^{n\times n}$ and $X \in \mathbb{R}^{n\times r}$, there exists a choice of $Y^\star = \arg\max_Y \langle H, XY^T + YX^T\rangle$ such that*

$$\|XY^{\star T} + Y^\star X^T\|_F^2 \geq 2 \cdot \lambda_k(XX^T) \cdot \|Y^\star\|_F^2 \qquad \textit{where } k = \operatorname{rank}(X).$$

Proof Define $\mathcal{J} : \mathbb{R}^{n\times r} \to \mathcal{S}^n$ such that $\mathcal{J}(Y) = XY^T + YX^T$ for all Y. We observe that $Y^\star = \mathcal{J}^\dagger(H)$ where $\dagger$ denotes the pseudoinverse. Without loss of generality, let $X = [\Sigma; 0]$ where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ and $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$. Then, the minimum norm solution is written

$$\mathcal{J}^\dagger(H) = \arg\min_{Y=[Y_1;Y_2]} \left\| \begin{bmatrix} \Sigma Y_1^T + Y_1 \Sigma & \Sigma Y_2^T \\ Y_2 \Sigma & 0 \end{bmatrix} - \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} \right\|^2 = \begin{bmatrix} \frac{1}{2} H_{11} \\ H_{12}^T \end{bmatrix} \Sigma^\dagger,$$

where $\Sigma^\dagger = \operatorname{diag}(\sigma_1^{-1}, \ldots, \sigma_k^{-1}, 0, \ldots, 0)$. From this we see that the pseudoinverse $\mathcal{J}^\dagger$ has operator norm

$$\|\mathcal{J}^\dagger\|_{\text{op}} = \max_{\|H\|_F=1} \|\mathcal{J}^\dagger(H)\| = (\sqrt{2}\sigma_k)^{-1},$$

with maximizer $H_{11}^\star = 0$ and $H_{22}^\star = 0$ and $H_{12}^\star = \frac{1}{\sqrt{2}} e_k h^T$, where h is any unit vector with $\|h\| = 1$ and e_k is the k-th column of the identity matrix. The desired claim then follows from the fact that $Y^\star \in \operatorname{range}(\mathcal{J}^T)$ and therefore $Y^\star = \mathcal{J}^\dagger \mathcal{J}(Y^\star)$ and $\|Y^\star\|_F^2 \leq \|\mathcal{J}^\dagger\|_{\text{op}}^2 \cdot \|\mathcal{J}(Y^\star)\|_F^2$. ∎

Suppose that $\cos^2\theta \geq 1/2$ holds due to Lemma 19 within the neighborhood $f(X) - f^\star \leq R$ for some radius $R > 0$. Substituting Lemma 20 into (12) yields a *local* gradient dominance condition

$$\frac{1}{2}\|\nabla f_0(X)\|_F^2 \geq \lambda_{\min}(X^T X) \cdot [f_0(X) - f_0^\star]. \tag{14}$$

In the overparameterized case $r > r^\star$, however, (14) does not prove gradient dominance, because $\lambda_{\min}(X^T X)$ becomes arbitrarily small as it converges towards $\lambda_r(M^\star) = 0$. Indeed, the inequality (14) suggests a *sublinear* convergence rate, given that

$$f_0(X_+) - f_0^\star \leq \left(1 - \alpha \lambda_{\min}(X^T X)\right)\left(f_0(X) - f_0^\star\right), \tag{15}$$

has a linear convergence rate $1 - \alpha\lambda_r(XX^T)$ that itself converges to 1.

6. Local Linear Convergence of Preconditioned Gradient Descent

In the literature, *right preconditioning* is a technique frequently used to improve the condition number of a matrix (i.e. its *conditioning*) without affecting its column span (i.e. its

alignment); see e.g. Saad (2003, Section 9.3.4) or Greenbaum (1997, Chapter 10). In this section, we define a local norm and dual local norm based on right preconditioning with a positive definite preconditioner

$$\|U\|_{X,\eta} \stackrel{\text{def}}{=} \|UP_{X,\eta}^{1/2}\|_F, \qquad \|V\|_{X,\eta}^* \stackrel{\text{def}}{=} \|VP_{X,\eta}^{-1/2}\|_F, \qquad P_{X,\eta} \stackrel{\text{def}}{=} X^TX + \eta I.$$

If we can demonstrate gradient dominance under the dual local norm

$$f(X) - f^\star \le R \quad \Longrightarrow \quad \tau_{X,\eta} \cdot [f(X) - f^\star] \le \frac{1}{2}(\|\nabla f(X)\|_{X,\eta}^*)^2$$

for some radius constant $R > 0$ and dominance constant $\tau_P > 0$, and if the function f remains gradient Lipschitz under the local norm

$$f(X + \alpha V) \le f(X) + \alpha \langle \nabla f(X), V\rangle + \frac{\ell_{X,\eta}}{2}\alpha^2\|V\|_{X,\eta}^2,$$

for some new lipschitz constant $\ell_{X,\eta}$, then it follows from the same reasoning as before that the *right preconditioned* gradient descent iterations $X_+ = X - \alpha\nabla f(X)P_{X,\eta}^{-1}$ achieves linear convergence

$$f(X_+) - f^\star \le (1 - \alpha\cdot\tau_{X,\eta})\cdot(f(X) - f^\star) \qquad \text{with step size } \alpha \le \ell_{X,\eta}^{-1}.$$

Starting from an initial point X_0 within the radius $f(X_0) - f^\star \le R$, it follows that preconditioned gradient descent converges to an ϵ-suboptimal point X_k that satisfies $f(X_k) - f^\star \le \epsilon$ in at most $k = O((\tau_{X,\eta}/\ell_{X,\eta})\log(R/\epsilon))$ iterations.

In order to motivate our choice of preconditioner $P_{X,\eta}$, we return to the perfectly conditioned function $f_0(X) = f_0^\star + \frac{1}{2}\|XX^T - M^\star\|_F^2$ considered in the previous section. Repeating the derivation of (10) results in the following

$$\begin{aligned}\|\nabla f_0(X)\|_{X,\eta}^* &= \max_{\|Y\|_{X,\eta}=1} \langle \nabla f_0(X), Y\rangle = \max_{\|Y\|_P=1} \left\langle XX^T - M^\star, XY^T + YX^T\right\rangle \\ &= \|XX^T - M^\star\|_F\|XY^{\star T} + Y^\star X^T\|_F\cos\theta, \end{aligned} \tag{16}$$

in which the incidence angle θ coincides with the one previously defined in (11), but the corresponding maximizer $Y^\star$ is rescaled so that $\|Y^\star\|_{X,\eta} = 1$. Suppose that $\cos^2\theta \ge 1/2$ holds due to Lemma 19 within the neighborhood $f(X) - f^\star \le R$ for some radius $R > 0$. Evoking Lemma 20 with $Y \leftarrow YP_{X,\eta}^{+1/2}$ and $X \leftarrow XP_{X,\eta}^{-1/2}$ to lower-bound $\|XY^{\star T} + Y^\star X^T\|_F$ yields:

$$\frac{1}{2}(\|\nabla f_0(X)\|_{X,\eta}^*)^2 \ge \lambda_{\min}(P_{X,\eta}^{-1/2}X^TXP_{X,\eta}^{-1/2})\cdot[f_0(X) - f_0^\star]. \tag{17}$$

While right preconditioning does not affect the term $\cos\theta$, which captures the *alignment* between the column span of X and the ground truth $M^\star$, it can substantially improve the *conditioning* of the subspace $\{XY^T + YX^T : Y \in \mathbb{R}^{n\times r}\}$.

In particular, choosing $\eta = 0$ sets $P_{X,0} = X^TX$ and $\lambda_{\min}(P_{X,\eta}^{-1/2}X^TXP_{X,\eta}^{-1/2}) = 1$. While f_0 fails to satisfy gradient dominance under the Euclidean norm, this derivation shows that gradient dominance does indeed hold after a change of norm. The following is a specialization of Lemma 24 that we prove later in this section.

Corollary 21 (Gradient dominance with $\eta = 0$)**.** *Let ϕ be (μ, r)-restricted strongly convex and L-gradient Lipschitz, and let $M^\star \succeq 0$ satisfy $\nabla\phi(M^\star) = 0$ and $r^\star = \operatorname{rank}(M^\star) \leq r$. Then, $f(X) \overset{def}{=} \phi(XX^T)$ satisfies gradient dominance*

$$f(X) - f^\star \leq \frac{\mu \cdot \lambda_{r^\star}^2(M^\star)}{2(1 + L/\mu)} \quad \Longrightarrow \quad \frac{\mu^2}{2L} \cdot [f(X) - f^\star] \leq \frac{1}{2}(\|\nabla f(X)\|_{X,0}^*)^2. \tag{18}$$

In fact, the resulting iterations $X_+ = X - \alpha\nabla f(X)(X^TX)^{-1}$ coincide with the ScaledGD of Tong et al. (2020). One might speculate that gradient dominance in this case would readily imply linear convergence, given that

$$f(X_+) - f^\star \leq (1 - \alpha\tau_{X,0})(f(X) - f^\star) \qquad \text{with } \alpha \leq \max\{1, \ell_{X,0}^{-1}\}.$$

However, the Lipschitz parameter $\ell_{X,\eta}$ may diverge to infinity as $\eta \to 0$, and this causes the admissible step-size $\alpha \leq \max\{1, \ell_{X,\eta}^{-1}\}$ to shrink to zero. Conversely, if we insist on using a fixed step-size $\alpha > 0$, then the objective function may on occasion *increase* after an iteration, as in $f(X_+) > f(X)$. Indeed, this possible increment explains the apparently sporadic behavior exhibited by ScaledGD.

Measured under the Euclidean norm, the function f is gradient Lipschitz but not gradient dominant. Measured under a right-preconditioned P-norm with $P = X^TX$, the function f is gradient dominant but not gradient Lipschitz. Viewing the Euclidean norm as simply a right-preconditioned norm with $P = I$, a natural idea is to *interpolate* between these two norms, by choosing the preconditioner $P_{X,\eta} = X^TX + \eta I$. It is not difficult to show that keeping η sufficiently *large* with respect to the error norm $\|XX^T - M^\star\|_F$ is enough to ensure that f continues to satisfy gradient Lipschitzness under the local norm. The proof of Lemma 22 and Lemma 23 below follows from straightforward linear algebra, and are deferred to Appendix B and Appendix C respectively.

Lemma 22 (Gradient Lipschitz)**.** *Let ϕ be L-gradient Lipschitz. Let $M^\star = \arg\min \phi$ satisfy $M^\star \succeq 0$. Then $f(X) \overset{def}{=} \phi(XX^T)$ satisfies*

$$f(X + V) \leq f(X) + \langle \nabla f(X), V\rangle + \frac{\ell_{X,\eta}}{2}\|V\|_{X,\eta}^2$$
$$\text{where } \ell_{X,\eta} = L \cdot \left[4 + \frac{2\|XX^T - M^\star\|_F + 4\|V\|_{X,\eta}}{\lambda_{\min}(X^TX) + \eta} + \left(\frac{\|V\|_{X,\eta}}{\lambda_{\min}(X^TX) + \eta}\right)^2\right].$$

Lemma 23 (Bounded gradient)**.** *Under the same conditions as Lemma 22, the search direction $V = \nabla f(X)(X^TX + \eta I)^{-1}$ satisfies $\|V\|_{X,\eta} = \|\nabla f(X)\|_{X,\eta}^* \leq 2L\|XX^T - M^\star\|_F$.*

Substituting $X_+ = X - \alpha\nabla f(X)(X^TX + \eta I)^{-1}$ into Lemma 22 yields the usual form of the Lipschitz gradient decrement

$$f(X_+) \leq f(X) - \alpha \cdot (\|\nabla f(X)\|_{X,\eta}^*)^2 + \alpha^2 \cdot \frac{\ell_{X,\eta}}{2}(\|\nabla f(X)\|_{X,\eta}^*)^2 \tag{19a}$$

in which the *local* Lipschitz term $\ell_{X,\eta}$ is bounded by Lemma 23 as

$$\ell_{X,\eta} \leq 4L + (2L + 8L^2) \cdot \frac{\|XX^T - M^\star\|_F}{\lambda_{\min}(X^TX) + \eta} + 4L^3 \cdot \left(\frac{\|XX^T - M^\star\|_F}{\lambda_{\min}(X^TX) + \eta}\right)^2. \tag{19b}$$

By keeping η sufficiently *large* with respect to the error norm $\|XX^T - M^\star\|_F$, it follows that $\ell_{X,\eta}$ can be replaced by a *global* Lipschitz constant $\ell \geq \ell_{X,\eta}$ that is independent of X.

Our main result in this paper is that keeping η sufficiently *small* with respect to the error norm $\|XX^T - M^\star\|_F$ is enough to ensure that f satisfies gradient dominance, even in the overparameterized regime where $r > r^\star$.

Lemma 24 (Gradient dominance)**.** *Let ϕ be L-gradient Lipschitz and $(\mu, 2r)$-restricted strongly convex. Let $M^\star = \arg\min \phi$ satisfy $M^\star \succeq 0$ and $r^\star = \operatorname{rank}(M^\star) \leq r$. Then, $f(X) \stackrel{def}{=} \phi(XX^T)$ satisfies*

$$f(X) - f^\star \leq \frac{\mu}{2(1+L/\mu)} \cdot \lambda^2_{r^\star}(M^\star) \qquad \Longrightarrow$$
$$\frac{\mu}{\sqrt{2}} \left(1 + \eta \cdot \frac{c_0 + c_1 \cdot \sqrt{r - r^\star}}{\|XX^T - M^\star\|_F}\right)^{-1/2} \leq \frac{\|\nabla f(X)\|^*_{X,\eta}}{\|XX^T - M^\star\|_F}$$

where $c_0 = 1 + \sqrt{2}$ and $c_1 = (L+\mu)/\sqrt{\mu L}$.

Substituting $f(X) - f^\star \leq \frac{L}{2}\|XX^T - M^\star\|^2_F$ from Lemma 16 into Lemma 24 recovers the usual form of gradient dominance

$$\tau_{X,\eta} \cdot [f(X) - f^\star] \leq \frac{1}{2}(\|\nabla f(X)\|^*_{X,\eta})^2 \tag{20a}$$

in which the *local* dominance term $\tau_{X,\eta}$ reads

$$\tau_{X,\eta} = \frac{\mu^2}{2L} \left(1 + \eta \cdot \frac{c_0 + c_1 \cdot \sqrt{r - r^\star}}{\|XX^T - M^\star\|_F}\right)^{-1} > 0. \tag{20b}$$

By keeping η sufficiently *small* with respect to the error norm $\|XX^T - M^\star\|_F$, it follows that $\tau_{X,\eta}$ can be replaced by a *global* dominance constant $\tau \leq \tau_{X,\eta}$ that is independent of X. Finally, substituting the global Lipschitz constant $\ell \geq \ell_{X,\eta}$ and the global dominance constant $\tau \leq \tau_{X,\eta}$ into (19) and (20) yields a proof of linear convergence in Theorem 4.

Proof [Proof of Theorem 4]It follows from (19b) that

$$\eta \geq C_{\text{lb}} \cdot \|XX^T - M^\star\|_F \quad \Longrightarrow \quad \ell_{X,\eta} \leq 4L + \frac{2L + 8L^2}{C_{\text{lb}}} + \frac{4L^3}{C^2_{\text{lb}}} \stackrel{\text{def}}{=} \ell.$$

Substituting $\ell \geq \ell_{X,\eta}$ into (19a) yields a guaranteed gradient decrement

$$f(X_+) - f(X) \leq -\frac{\alpha}{2}(\|\nabla f(X)\|^*_{X,\eta})^2 \leq 0 \text{ for } \alpha \leq \min\{1, \ell^{-1}\}, \tag{21}$$

for a fixed step-size $\alpha > 0$. It follows from (20b) that

$$\eta \leq C_{\text{ub}} \cdot \|XX^T - M^\star\|_F \quad \Longrightarrow \quad \tau_{X,\eta} \geq \frac{\mu^2}{2L} \left(1 + \frac{c_0 + c_1 \cdot \sqrt{r - r^\star}}{C^{-1}_{\text{ub}}}\right)^{-1} \stackrel{\text{def}}{=} \tau.$$

Substituting $\tau \leq \tau_{X,\eta}$ and gradient dominance (20a) into the decrement in (21) yields linear convergence

$$f(X_+) - f^\star \leq (1 - \alpha \cdot \tau) \cdot (f(X) - f^\star) \text{ for } \alpha \leq \min\{1, \ell^{-1}\},$$

which is exactly the claim in Theorem 4. ■

Proof [Corollary 5] Within the neighborhood stated in Lemma 24 where f is gradient dominant, it follows immediately from Lemma 23 and Lemma 24 that the choice of $\eta = \|\nabla f(X)\|^*_{X,0}$ satisfies

$$\frac{\mu}{\sqrt{2}} \cdot \|XX^T - M^\star\|_F \leq \|\nabla f(X)\|^*_{X,0} \leq 2L \cdot \|XX^T - M^\star\|_F,$$

which is exactly the claim in Corollary 5. ■

We now turn our attention to the proof of Lemma 24. Previously, in motivating our proof for gradient dominance under the Euclidean norm, we derived a bound like

$$\begin{aligned}\|\nabla f_0(X)\|^*_F &= \max_{\|Y\|_F=1} \left\langle XX^T - M^\star, XY^T + YX^T \right\rangle \\ &= \|XX^T - M^\star\|_F \|XY^{\star T} + Y^\star X^T\|_F \cos\theta \end{aligned} \tag{22}$$

where $Y^\star$ is a maximizer such that $\|Y^\star\|_F = 1$. We found that $\cos\theta$ is always large, because the error $XX^T - M^\star$ is guaranteed to align well with the linear subspace $\{XY^T + YX^T : Y \in \mathbb{R}^{n\times r}\}$, but that the term $\|XY^{\star T} + Y^\star X^T\|_F$ can decay to zero if the error concentrates within the degenerate directions of the subspace.

In our initial experiments with PrecGD, we observed that small values of η caused the error to preferrably align towards the *well-conditioned* directions of the subspace. Suppose that X contains k large, well-conditioned singular values, and $r-k$ near-zero singular values. Let X_k denote the rank-k approximation of X, constructed by setting the $r-k$ near-zero singular values of X as exactly zero:

$$X_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T = \arg\min_{\tilde{X}\in\mathbb{R}^{n\times r}} \left\{ \|\tilde{X} - X\| : \operatorname{rank}(\tilde{X}) \leq k \right\}.$$

Then, our observation is that small values of η tend to concentrate the error $XX^T - M^\star$ within the well-conditioned subspace $\{X_kY^T + YX_k^T : Y \in \mathbb{R}^{n\times r}\}$.

In order to sharpen the bound (22) to reflect the possibility that $\{XY^T + YX^T : Y \in \mathbb{R}^{n\times r}\}$ may contain degenerate directions that do not significantly align with the error vector $XX^T - M^\star$, we suggest the following refinement

$$\begin{aligned}\|\nabla f_0(X)\|_F &\geq \max_{\|Y\|_F=1} \left\{ \left\langle XX^T - M^\star, XY^T + YX^T \right\rangle : Yv_i = 0 \text{ for } i > k \right\} \\ &= \max_{\|Y\|_F=1} \left\langle XX^T - M^\star, X_kY^T + YX_k^T \right\rangle \\ &= \|XX^T - M^\star\|_F \|X_kY_k^{\star T} + Y_k^\star X_k^T\|_F \cos\theta_k, \end{aligned}$$

where each $\cos\theta_k$ measures the alignment between the error $XX^T - M^\star$ and the well-conditioned subspace $\{X_kY^T + YX_k^T : Y \in \mathbb{R}^{n\times r}\}$. While $\cos\theta_k$ must be necessarily be worse than $\cos\theta$, given that the well-conditioned subspace is a subset of the whole subspace, our hope is that eliminating the degenerate directions will allow the term $\|X_kY_k^{\star T} + Y_k^\star X_k^T\|_F$ to be significantly improved from $\|XY^{\star T} + Y^\star X^T\|_F$.

Lemma 25 (Alignment lower-bound). *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 its singular value decomposition. Under the same conditions as Lemma 24, we have*

$$\frac{\|\nabla f(X)\|_{X,\eta}^*}{\|XX^T - M^\star\|_F} \geq \max_{k\in\{1,2,\ldots,r\}} \frac{\mu + L}{\sqrt{2}} \cdot \frac{\cos\theta_k - \delta}{\sqrt{1 + \eta/\lambda_k(XX^T)}} \tag{23}$$

where $\delta = \frac{L-\mu}{L+\mu}$ and each θ_k is defined

$$\cos\theta_k = \max_{Y\in\mathbb{R}^{n\times r}} \frac{\left\langle XX^T - M^\star, X_k Y^T + Y X_k^T \right\rangle}{\|XX^T - M^\star\|_F \|X_k Y^T + Y X_k^T\|_F}, \qquad X_k = \sum_{i=1}^k \sigma_i u_i v_i^T. \tag{24}$$

Proof Let $E = XX^T - M^\star$ and $\mathcal{J}(Y) = XY^T + YX^T$ and $\mathcal{J}_k = X_kY^T + YX_k^T$. Repeating the proof of Lemma 17 yields the following corollary

$$\|\nabla f(X)\|_{X,\eta}^* \geq \nu \cdot \left\{ \max_{\|Y\|_{X,\eta}=1} \langle E, \mathcal{J}(Y)\rangle - \delta\|E\|_F \|\mathcal{J}(Y)\|_F \right\}$$

where $\nu = \frac{1}{2}(\mu + L)$ and $\delta = \frac{L-\mu}{L+\mu}$. For any $k \in \{1, 2, \ldots, r\}$, we can restrict this problem so that

$$\begin{aligned}
\|\nabla f(X)\|_{X,\eta}^* &\geq \nu \cdot \left\{ \max_{\|Y\|_{X,\eta}=1} \langle E, \mathcal{J}(Y)\rangle - \delta\|E\|_F \|\mathcal{J}(Y)\|_F : Yv_i = 0 \text{ for } i > k \right\} \\
&= \nu \cdot \left\{ \max_{\|Y\|_{X,\eta}=1} \langle E, \mathcal{J}_k(Y)\rangle - \delta\|E\|_F \|\mathcal{J}_k(Y)\|_F \right\} \\
&\geq \nu \cdot \|E\|_F \|\mathcal{J}_k(Y_k^\star)\|_F (\cos\theta_k - \delta)
\end{aligned}$$

where $Y_k^\star = \mathcal{J}_k^\dagger(E)$ denotes the solution of (24) rescaled so that $\|Y\|_{X,\eta} = 1$. Let $P = X^TX + \eta I$ and observe that $\|Y_k^\star\|_{X,\eta}^2 = \|Y_k^\star P^{1/2}\|_F^2$. It follows from Lemma 20 with $X \leftarrow X_k P^{-1/2}$ and $Y \leftarrow Y_k^\star P^{1/2}$ that

$$\|X_k Y_k^{\star T} + Y_k^\star X_k^T\|_F^2 \geq 2 \cdot \lambda_{\min}(P^{-1/2} X_k^T X_k P^{-1/2}) \cdot \|Y_k^\star\|_{X,\eta}^2.$$

In turn, we have $P = \sum(\sigma_i^2 + \eta) v_i v_i^T$ and therefore

$$\lambda_{\min}(P^{-1/2} X_k^T X_k P^{-1/2}) = \min_{i\leq k} \left\{ \frac{\sigma_i^2}{\eta + \sigma_i^2} \right\} = \frac{\sigma_k^2}{\eta + \sigma_k^2} = \frac{1}{1 + \eta/\sigma_k^2}.$$

Substituting these together yields

$$\begin{aligned}
\|\nabla f(X)\|_{X,\eta}^* &\geq \nu \cdot \|E\|_F \cdot \|\mathcal{J}_k(Y_k^\star)\|_F \cdot (\cos\theta_k - \delta) \\
&\geq \frac{\mu + L}{2} \cdot \|E\|_F \cdot \frac{\sqrt{2}}{\sqrt{1 + \eta/\sigma_k^2}} \cdot (\cos\theta_k - \delta).
\end{aligned}$$

■

From Lemma 25, we see that gradient dominance holds if the subspace $\{X_kY^T + YX_k^T : Y \in \mathbb{R}^{n\times r}\}$ induced by the rank-k approximation of X is *well-conditioned*, and if the error vector $XX^T - M^\star$ is *well-aligned* with it. Specifically, this is to require both $\lambda_k(XX^T)$ and $\cos\theta_k$ to remain sufficiently large for the *same* value of k. Within a neighborhood of the ground truth, it follows from Weyl's inequalty that $\lambda_k(XX^T)$ will remain sufficiently large for $k = r^\star$; see Lemma 26 below. In the overparameterized regime $r > r^\star$, however, it is not necessarily true that $\cos\theta_k \to 1$ for $k = r^\star$. Instead, we use an induction argument: if $\cos\theta_k$ is too small to prove gradient dominance, then the smallness of $\cos\theta_k$ provides a lower-bound on $\lambda_{k+1}(XX^T)$ via Lemma 27 below. Inductively repeating this argument for $k = r^\star, r^\star + 1, \dots$ arrives at a lower-bound on $\lambda_r(XX^T)$. At this point, Lemma 19 guarantees that $\cos\theta_r$ is large, and therefore, we conclude that gradient dominance must hold.

Lemma 26 (Base case)**.** *Under the same conditions as Lemma 24, let* $f(X) - f(X^\star) \le \frac{\mu}{2(1+L/\mu)} \cdot \lambda_{r^\star}^2(M^\star)$*. Then,*

$$\lambda_{r^\star}(XX^T) \ge (\sqrt{1+L/\mu} - 1) \cdot \|XX^T - M^\star\|_F.$$

Proof By our choice of neighborhood, we have

$$\|XX^T - M^\star\|_F^2 \le \frac{2}{\mu} \cdot [f(X) - f(X^\star)] \le \frac{1}{1+L/\mu} \cdot \lambda_{r^\star}^2(M^\star).$$

The desired claim follows from Weyl's inequality:

$$\begin{aligned}\lambda_{r^\star}(XX^T) &= \lambda_{r^\star}(M^\star + XX^T - M^\star)\\ &\ge \lambda_{r^\star}(M^\star) - \|XX^T - M^\star\|_F\\ &\ge (\sqrt{1+L/\mu} - 1) \cdot \|XX^T - M^\star\|_F.\end{aligned}$$

■

Lemma 27 (Induction step)**.** *Under the same conditions as Lemma 24, let* $f(X) - f(X^\star) \le \frac{\mu}{2(1+L/\mu)} \cdot \lambda_{r^\star}^2(M^\star)$*. Then,* $\cos\theta_k$ *defined in (24) gives the following lower-bound on* $\lambda_{k+1}(XX^T)$*:*

$$\frac{\lambda_{k+1}^2(XX^T)\cdot(r-k)}{\|XX^T - M^\star\|_F^2} - \frac{\mu L}{(L+\mu)^2} \ge \left(\frac{L}{L+\mu}\right)^2 - \cos^2\theta_k.$$

Proof For $k \le r$, we have

$$\begin{aligned}&\|(I - X_kX_k^\dagger)(XX^T - M^\star)(I - X_kX_k^\dagger)\|_F^2\\ \le&\|(I - X_kX_k^\dagger)XX^T(I - X_kX_k^\dagger)\|_F^2 + \|(I - X_kX_k^\dagger)M^\star(I - X_kX_k^\dagger)\|_F^2\\ \le&\lambda_{k+1}^2(XX^T)(r-k) + \|(I - XX^\dagger)M^\star(I - XX^\dagger)\|_F^2\end{aligned}$$

and therefore

$$\sin^2\theta_k \le \frac{\lambda_{k+1}^2(XX^T)(r-k)}{\|XX^T-M^\star\|_F^2} + \frac{\|(I-XX^\dagger)M^\star(I-XX^\dagger)\|_F^2}{\|XX^T-M^\star\|_F^2}$$

By the choice of the neighborhood, we have

$$\|XX^T-M^\star\|_F^2 \le \frac{2}{\mu}\cdot[f(X)-f(X^\star)] \le \frac{1}{1+L/\mu}\cdot\lambda_{r^\star}^2(M^\star).$$

Substituting $\rho = \frac{1}{\sqrt{1+L/\mu}} \le \frac{1}{\sqrt{2}}$ into Lemma 19 proves that

$$\sin^2\theta_r = \frac{\|(I-XX^\dagger)M^\star(I-XX^\dagger)\|_F^2}{\|XX^T-M^\star\|_F^2} \le \frac{1}{2}\frac{\rho}{1-\rho^2} = \frac{\mu}{2L}.$$

Splitting

$$\begin{aligned}1-\frac{\mu}{2L} &= \left(\frac{L}{L+\mu}\right)^2 + \frac{\mu}{2L}\frac{(3L^2-\mu^2)}{(L+\mu)^2}\\ &\ge \left(\frac{L}{L+\mu}\right)^2 + \frac{\mu L}{(L+\mu)^2}\end{aligned}$$

and bounding $\cos^2\theta_k \ge 1-\sin^2\theta_k$ yields the desired bound. ■

Rigorously repeating this induction results in a proof of Lemma 24.

Proof [Lemma 24] Lemma 25 proves gradient dominance if we can show that both $\cos\theta_k$ and $\lambda_k(XX^T)$ remain large for the same value of k. By Lemma 26 we have

$$\frac{\lambda_{r^\star}(XX^T)}{\|XX^T-M^\star\|_F} \ge \sqrt{1+L/\mu}-1 \ge \sqrt{2}-1 = \frac{1}{1+\sqrt{2}}. \tag{25}$$

If $\cos\theta_{r^\star} \ge \frac{L}{L+\mu}$, then substituting (25) into Lemma 25 with $k=r^\star$ yields gradient dominance:

$$\begin{aligned}\frac{\|\nabla f(X)\|_{X,\eta}^*}{\|XX^T-M^\star\|} &\ge \frac{(\mu+L)}{\sqrt{2}}\cdot\left(\frac{L}{L+\mu}-\frac{L-\mu}{L+\mu}\right)\left(1+\frac{\eta}{\lambda_{r^\star}(XX^T)}\right)^{-1/2}\\ &\ge \frac{\mu}{\sqrt{2}}\left(1+\eta\cdot\frac{1+\sqrt{2}}{\|XX^T-M^\star\|_F}\right)^{-1/2}.\end{aligned} \tag{26}$$

Otherwise, if $\cos\theta_{r^\star} < \frac{L}{L+\mu}$, then we proceed with an induction argument. Beginning at the base case $k=r^\star$, we evoke Lemma 27 and use $\cos\theta_k < \frac{L}{L+\mu}$ to lower-bound $\lambda_{k+1}(XX^T)$ by a constant:

$$\begin{aligned}&\frac{\lambda_{k+1}^2(XX^T)\cdot(r-k)}{\|XX^T-M^\star\|_F^2} - \frac{\mu L}{(L+\mu)^2} \ge \left(\frac{L}{L+\mu}\right)^2 - \cos^2\theta_k > 0,\\ &\implies \quad \frac{\lambda_{k+1}(XX^T)}{\|XX^T-M^\star\|_F} > \frac{1}{\sqrt{r-r^\star}}\cdot\frac{\sqrt{\mu L}}{L+\mu}.\end{aligned} \tag{27}$$

If $\cos\theta_{k+1} \geq \frac{L}{L+\mu}$, then substituting (27) into Lemma 25 yields gradient dominance:

$$\begin{aligned}\frac{\|\nabla f(X)\|_{X,\eta}^*}{\|XX^T - M^\star\|} &\geq \frac{(\mu+L)}{\sqrt{2}}\left(\frac{L}{L+\mu} - \frac{L-\mu}{L+\mu}\right)\left(1 + \frac{\eta}{\lambda_{k+1}(XX^T)}\right)^{-1/2} \\ &\geq \frac{\mu}{\sqrt{2}}\left(1 + \eta \cdot \frac{\sqrt{r-r^\star}\cdot\frac{L+\mu}{\sqrt{\mu L}}}{\|XX^T - M^\star\|_F}\right)^{-1/2}. \end{aligned} \tag{28}$$

Otherwise, if $\cos\theta_{k+1} < \frac{L}{L+\mu}$, then we repeat the same argument in (27) with $k \leftarrow k+1$, until we arrive at $k = r$. At this point, Lemma 27 guarantees $\cos\theta_r \geq \frac{L}{L+\mu}$, since

$$0 \geq \underbrace{\frac{\lambda_{k+1}^2(XX^T)\cdot(r-k)}{\|XX^T - M^\star\|_F^2}}_{=0 \text{ because } k=r} - \frac{\mu L}{(L+\mu)^2} \geq \left(\frac{L}{L+\mu}\right)^2 - \cos^2\theta_k,$$

so the induction terminates with (28). Finally, lower-bounding the two bounds (26) and (28) via $\min\{a^{-1}, b^{-1}\} \geq (a+b)^{-1}$ yields the desired Lemma 24. ■

7. Global Convergence

In this section, we study the global convergence of perturbed PrecGD or PPrecGD from an arbitrary initial point to an approximate second order stationary point. To establish the global convergence of PPrecGD, we study a slightly more general variant of gradient descent, which we call *perturbed metric gradient descent*.

7.1 Perturbed Metric Gradient Descent

Let $P : \mathbb{R}^d \to \mathcal{S}_{++}^d$ denote an arbitrary metric function, which we use to define the following two local norms

$$\|v\|_x \overset{\text{def}}{=} \sqrt{v^T P(x) v}, \qquad \|v\|_x^* \overset{\text{def}}{=} \sqrt{v^T P(x)^{-1} v}$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ denote an arbitrary ℓ_1-gradient Lipschitz and ℓ_2-Hessian Lipschitz function. We consider solving the general minimization problem $f^\star = \min_x f(x)$ via *perturbed metric gradient descent*, defined as

$$x_{k+1} = x_k - \alpha P(x_k)^{-1}\nabla f(x_k) + \alpha\zeta_k, \tag{PMGD}$$

in which the random perturbation ζ_k is chosen as

$$\begin{cases} \zeta_k \sim \mathbb{B}(\beta) & \text{if } \|\nabla f(x)\|_x^* \leq \epsilon, \text{ and it has been at least } \mathcal{T} \text{ iters since last perturbation} \\ \zeta_k = 0 & \text{otherwise.} \end{cases}$$

Indeed, PPrecGD is a special case of PMGD after choosing $x = \mathrm{vec}(X)$ and $P(x) = (X^TX + \eta_0 I_n) \otimes I_r$. Our main result in this section is to show that PMGD converges to ϵ-second order stationary point measured in the metric norm

$$\|\nabla f(x)\|_x^* \leq \epsilon, \quad \nabla^2 f(x) \succeq -\sqrt{L_d\epsilon}\cdot P(x),$$

for some constant L_d to be defined later, in at most $\tilde{O}(\epsilon^{-2})$ iterations under two assumptions:

- The metric P should be *well-conditioned*:

$$p_{\text{lb}} I \preceq P(x) \preceq p_{\text{ub}} I \qquad \text{for all } x \in \mathbb{R}^d. \tag{29}$$

  for some $p_{\text{ub}} \geq p_{\text{lb}} > 0$.

- The metric P should be *Lipschitz continuous*:

$$\|P(x) - P(y)\| \leq L_P \cdot \|x - y\| \qquad \text{for all } x, y \in \mathbb{R}^d, \tag{30}$$

  for some $L_P > 0$.

Comparison to Perturbed Gradient Descent. Jin et al. (2021) showed that the episodic injection of isotropic noise to gradient descent enables it to escape strict saddle points efficiently. Indeed, this algorithm, called perturbed gradient descent or PGD for short, can be regarded as an special case of PMGD, with a crucial simplification that the metric function P is the identity mapping throughout the iterations of the algorithm. As will be explained later, such simplification enables PGD to behave almost like the power method within the vicinity of a strict saddle point, thereby steering the iterations away from it at an exponential rate along the negative curvature of the function. Extending this result to PMGD with a more general metric function that changes along the solution trajectory requires a more intricate analysis, which will be provided next. Our main theorem shows that PMGD can also escape strict saddle points, so long as the metric function $P(x)$ remains well-conditioned and Lipschitz continuous.

Theorem 28 (Global Convergence of PMGD)**.** *Let f be ℓ_1-gradient and ℓ_2-Hessian Lipschitz, and let P satisfy $p_{\text{lb}} I \preceq P(x) \preceq p_{\text{ub}} I$ and $\|P(x) - P(y)\| \leq L_P \cdot \|x - y\|$. Then, with an overwhelming probability and for any $\epsilon = \tilde{O}(1/(L_d p_{\text{ub}}))$, PMGD with perturbation radius $\beta = \tilde{\mathcal{O}}(\epsilon/L_d)$ and time interval $\mathcal{T} = \tilde{\mathcal{O}}(\ell_1/(p_{\text{lb}}\sqrt{L_d \epsilon}))$ converges to a point x that satisfies*

$$\|\nabla f(x)\|_x^* \leq \epsilon, \quad \textit{and} \quad \nabla^2 f(x) \succeq -\sqrt{L_d \epsilon} \cdot P(x),$$

in at most $\tilde{O}(\mathcal{C}(f(x) - f^*)/\epsilon^2)$ iterations, where f^* is the optimal objective value, and $L_d = 5 \max\{\ell_2, L_P \ell_1 \sqrt{p_{\text{ub}}}\}/p_{\text{lb}}^{2.5}$, $\mathcal{C} = \ell_1/p_{\text{lb}}^2$.

Before providing the proof for Theorem 28, we first show how it can be invoked to prove Theorem 8. To apply Theorem 28, we need to show that: (i) $f(X) = \phi(XX^T)$ is gradient and Hessian Lipschitz; and (ii) $P = (XX^\top + \eta I)$ is well-conditioned. However, the function $f(X) = \phi(XX^T)$ may neither be gradient nor Hessian Lipschitz, even if these properties hold for ϕ. To see this, consider $\phi(M) = \|M - M^\star\|_F^2$. Evidently, $\phi(M)$ is 2-gradient Lipschitz with constant Hessian. However, $f(X) = \phi(XX^\top) = \|XX^\top - M^\star\|_F^2$ is neither gradient- nor Hessian-Lipschitz since it is a quartic function of X. To alleviate this hurdle, we show that, under a mild condition on the coercivity of ϕ, the iterations of PPrecGD reside in a bounded region, within which $f(X)$ is both gradient and Hessian Lipschitz.

Lemma 29. *Suppose that ϕ is coercive. Let Γ_F and Γ_2 be defined as*

$$\Gamma_F = \max\left\{ \|X\|_F : \phi(XX^T) \leq \phi(X_0 X_0^T) + 2\sqrt{\|X_0\|_F^2 + \eta} \cdot \alpha r \epsilon \right\} \tag{31}$$

$$\Gamma_2 = \max\left\{ \|X\|_2 : \phi(XX^T) \leq \phi(X_0 X_0^T) + 2\sqrt{\|X_0\|_F^2 + \eta} \cdot \alpha r \epsilon \right\}. \tag{32}$$

Then, the following statements hold:

- *Every iteration of PPrecGD satisfies* $\|X_t\|_F \le \Gamma_F$ *and* $\|X_t\| \le \Gamma_2$.
- *The function* $f(X)$ *is* $9\Gamma_F^2 L_1$*-gradient Lipschitz within the ball* $\{M : \|M\|_F \le \Gamma_F\}$.
- *The function* $f(X)$ *is* $\big((4\Gamma_F + 2)L_1 + 4\Gamma_F^2 L_2\big)$*-Hessian Lipschitz within the ball* $\{M : \|M\|_F \le \Gamma_F\}$.
- *For* $\mathbf{P}_{X,\eta} = (X^T X + \eta I_n) \otimes I_r$*, we have* $\eta I \preceq \mathbf{P}_{X,\eta} \preceq (\Gamma_2^2 + \eta) I$ *and* $\|\mathbf{P}_{X,\eta} - \mathbf{P}_{Y,\eta}\| \le 2\Gamma_2 \|X - Y\|$ *within the ball* $\{M : \|M\| \le \Gamma_2\}$.

Equipped with Lemma 29 and Theorem 28, we are ready to present the global convergence result for PPrecGD.

Proof [Theorem 8] Due to our definition of Γ_F and Γ_2, every iteration of PPrecGD belongs to the set $\mathcal{D} = \{X : \|X\|_F \le \Gamma_F, \|X\| \le \Gamma_2\}$. On the other hand, Lemma 29 implies that $f(X)$ is $9L_1\Gamma_F^2$-gradient Lipschitz and $((4\Gamma_F + 2)L_1 + 4\Gamma_F^2 L_2)$-Hessian Lipschitz within $\mathcal{D}$. Moreover, $\eta I \preceq \mathbf{P}_{X,\eta} \preceq (\Gamma_2^2 + \eta) I$ and $\|\mathbf{P}_{X,\eta} - \mathbf{P}_{Y,\eta}\| \le 2\Gamma_2 \|X - Y\|$ within $\mathcal{D}$. Therefore, invoking Theorem 28 with parameters $\ell_1 = 9L_1\Gamma_F^2$, $\ell_2 = ((4\Gamma_F + 2)L_1 + 4\Gamma_F^2 L_2)$, $p_{\text{lb}} = \eta$, $p_{\text{ub}} = \Gamma_2^2 + \eta$, and $L_P = 2\Gamma_2$ completes the proof. ■

7.2 Proof of Theorem 28

To prove Theorem 28, we follow the main idea of Jin et al. (2021) and split the iterations into two parts:

- **Large gradient in local norm:** Suppose that $\|\nabla f(x)\|_x^* > \epsilon$ for some $\epsilon > 0$. Then, we show in Lemma 30 that a single iteration of PPrecGD without perturbation reduces the objective function by $\Omega(\epsilon^2)$.
- **Large negative curvature in local norm:** Suppose that $\|\nabla f(x)\|_x^* \ge \epsilon$ and x is not an ϵ-second order stationary point in local norm, i.e., $\nabla^2 f(x) \not\succeq -\sqrt{L_d \epsilon} P(x)$. We show in Lemma 31 that perturbing x with an isotropic noise followed by $\tilde{\mathcal{O}}(\epsilon^{-1/2})$ iterations of PPrecGD reduces the the objective function by $\tilde{\Omega}(\epsilon^{3/2})$.

Combining the above two scenarios, we show that PMGD decreases the objective value by $\tilde{\Omega}(\epsilon^2)$ per iteration (on average). Therefore, it takes at most $\tilde{\mathcal{O}}((f(x_0) - f^*)\epsilon^{-2})$ iterations to reach a ϵ-second order point in local norm.

Lemma 30 (Large gradient $\implies$ large decrement)**.** *Let* f *is* ℓ_1*-gradient Lipschitz, and let* $p_{\text{lb}} I \preceq P(x) \preceq p_{\text{ub}} I$ *for every* x*. Suppose that* x *satisfies* $\|\nabla f(x)\|_x^* > \epsilon$*, and define*

$$x_+ = x - \alpha P(x)^{-1} \nabla f(x)$$

with step-size $\alpha = p_{\text{lb}}/(2\ell_1)$*. Then, we have*

$$f(x_+) - f(x) < -\frac{p_{\text{lb}}}{4\ell_1}\epsilon^2.$$

Proof Due to the gradient Lipschitz continuity of $f(x)$, we have:

$$f(x_+) \leq f(x) + \alpha \left\langle \nabla f(x), -P(x)^{-1}\nabla f(x) \right\rangle + \frac{\ell_1}{2p_{\rm lb}}\alpha^2 \|P(x)^{-1}\nabla f(x)\|_x^2 \tag{33}$$

$$= f(x) - \alpha(\|\nabla f(x)\|_x^*)^2 \left(1 - \frac{\ell_1}{2p_{\rm lb}}\alpha\right), \tag{34}$$

$$\leq f(x) - \frac{\alpha}{2}(\|\nabla f(x)\|_x^*)^2 \tag{35}$$

$$\leq f(x) - \frac{p_{\rm lb}}{4\ell_1}\epsilon^2, \tag{36}$$

where in the last inequality we used the optimal step-size $\alpha = p_{\rm lb}/(2\ell_1)$ and the assumption $\|\nabla f(x)\|_x^* > \epsilon$. ■

Lemma 31 (Escape from saddle point)**.** *Let f be ℓ_1-gradient and ℓ_2-Hessian Lipschitz. Moreover, let $p_{\rm lb} I \preceq P(x) \preceq p_{\rm ub} I$ and $\|P(x) - P(y)\| \leq L_P\|x - y\|$ for every x and y. Suppose that $\bar{x}$ satisfies $\|\nabla f(\bar{x})\|_{\bar{x}}^* \leq \epsilon$ and $\nabla^2 f(\bar{x}) \not\succeq -\sqrt{L_d\epsilon}\cdot P(\bar{x})$. Then, the PMGD defined as*

$$x_{k+1} = x_k - \alpha P(x_k)^{-1}\nabla f(x_k), \qquad \textit{starting at} \qquad x_0 = \bar{x} + \alpha\cdot\xi,$$

with step-size $\alpha = p_{\rm lb}/(2\ell_1)$ and initial perturbation $\xi \sim \mathbb{B}(\beta)$ with $\beta = \epsilon/(400L_d\iota^3)$ achieves the following decrement with probability of at least $1-\delta$

$$f(x_t) - f(\bar{x}) \leq -\underbrace{\frac{1}{50\iota^3}\sqrt{\frac{\epsilon^3}{L_d}}}_{:=\mathcal{F}} \qquad \textit{after } \mathcal{T} = \frac{\ell_1}{p_{\rm lb}\sqrt{L_d\epsilon}}\iota \textit{ iterations,}$$

where $L_d = 5\max\{\ell_2, L_P\ell_1\sqrt{p_{\rm ub}}\}/p_{\rm lb}^{2.5}$ and $\iota = c\cdot\log(p_{\rm ub}d\ell_1(f(x_0)-f^*)/(p_{\rm lb}\ell_2\epsilon\delta))$ for some absolute constant c.

Before presenting the sketch of the proof for Lemma 31, we complete the proof of Theorem 28 based on Lemmas 30 and 31.

Proof [Theorem 28.] Let us define $T = 2\left(\mathcal{T}/\mathcal{F} + 4\ell_1/(p_{\rm lb}\epsilon^2)\right)\cdot(f(x_0) - f^*)$. By contradiction, suppose that x_t is *not* a ϵ-second order stationary point in local norm for any $t \leq T$. This implies that we either have $\|\nabla f(x_t)\|_{x_t}^* > \epsilon$, or $\|\nabla f(x_t)\|_{x_t}^* \leq \epsilon$ and $\nabla^2 f(x_t) \not\succeq -\sqrt{L_d\epsilon}\cdot P(x_t)$ for every $t \leq T$. Define T_1 as the number of iterations that satisfy $\|\nabla f(x_t)\|_{x_t}^* > \epsilon$. Similarly, define T_2 as the number of iterations that satisfy $\|\nabla f(x_t)\|_{x_t}^* \leq \epsilon$ and $\nabla^2 f(x_t) \not\succeq -\sqrt{L_d\epsilon}\cdot P(x_t)$. Evidently, we have $T_1 + T_2 = T$. We divide our analysis into two parts:

- Due to the definition of T_2, and in light of Lemma 31, we perturb the metric gradient at least $T_2/\mathcal{T}$ times. After each perturbation followed by $\mathcal{T}$ iterations, the PMGD reduces the objective function by at least $\mathcal{F}$ with probability $1-\delta$. Since the objective value cannot be less than f^*, we have

$$f(x_0) - (\mathcal{F}/\mathcal{T})T_2 \geq f^* \implies T_2 \leq (\mathcal{T}/\mathcal{F})\cdot(f(x_0) - f^*),$$

which holds with probability of at least

$$1-(T_2/\mathcal{T})\delta \geq 1-\frac{f(x_0)-f^*}{\mathcal{F}}\cdot\delta = 1-\delta',$$

where δ is chosen to be small enough so that $\delta' = (f(x_0)-f^*)\delta/\mathcal{F} < 1$.

- Excluding T_2 iterations that are within $\mathcal{T}$ steps after adding the perturbation, we are left with T_1 iterations with $\|\nabla f(x_t)\|_{x_t}^* > \epsilon$. According to Lemma 30, the PMGD reduces the objective by $p_{\mathrm{lb}}\epsilon^2/(4\ell_1)$ at every iteration x_t that satisfies $\|\nabla f(x_t)\|_{x_t}^* > \epsilon$. Therefore, we have

$$f(x_0)-(p_{\mathrm{lb}}\epsilon^2/(4\ell_1))T_1 \geq f^* \implies T_1 \leq (4\ell_1/(p_{\mathrm{lb}}\epsilon^2))\cdot(f(x_0)-f^*)$$

Recalling the definition of T, the above two cases imply that $T_1+T_2 < T$ with probability of at least $1-\delta'$, which is a contradiction. Therefore, $\|\nabla f(x_t)\|_{x_t}^* \leq \epsilon$ and $\nabla^2 f(x_t) \succeq -\sqrt{L_d\epsilon}\cdot P(x_t)$ after at most $2\left(\mathcal{T}/\mathcal{F}+4\ell_1/(p_{\mathrm{lb}}\epsilon^2)\right)\cdot(f(x_0)-f^*)$ iterations. Finally, note that

$$\begin{aligned} T &= \mathcal{O}\left(\left(\frac{\ell_1}{p_{\mathrm{lb}}^2\epsilon^2}+\frac{\ell_1}{p_{\mathrm{lb}}\epsilon^2}\right)\cdot(f(x_0)-f^*)\cdot\iota^4\right) \\ &= \tilde{\mathcal{O}}\left(\mathcal{C}\cdot\frac{(f(x_0)-f^*)}{\epsilon^2}\right). \end{aligned}$$

This completes the proof. ■

Finally, we explain the proof of Lemma 31. To streamline the presentation, we only provide a sketch of the proof and defer the detailed arguments to the appendix.

Sketch of the proof for Lemma 31. The proof of this lemma follows that of (Jin et al., 2021, Lemma 5.3) with a few key differences to account for the metric function $P(x)$, which will be explained below.

Consider two sequences $\{x_t\}_{t=0}^{T}$ and $\{y_t\}_{t=0}^{T}$ generated by PMGD and initialized at $x_0 = \bar{x}+\alpha\zeta_1$ and $y_0 = \bar{x}+\alpha\zeta_1$, for some $\zeta_1,\zeta_2 \in \mathbb{B}(\beta)$. By contradiction, suppose that Lemma 31 does not hold for either sequences $\{x_t\}_{t=0}^{\mathcal{T}}$ and $\{y_t\}_{t=0}^{\mathcal{T}}$, i.e., $f(x_t)-f(\bar{x}) \geq -\mathcal{F}$ and $f(y_t)-f(\bar{x}) \geq -\mathcal{F}$ for $t \leq \mathcal{T}$. That is, PMGD did not make sufficient decrement along $\{x_t\}_{t=0}^{\mathcal{T}}$ and $\{y_t\}_{t=0}^{\mathcal{T}}$. As a critical step in our proof, we show in the appendix (see Lemma 36) that such small decrement in the objective function implies that both sequences $\{x_t\}_{t=0}^{\mathcal{T}}$ and $\{y_t\}_{t=0}^{\mathcal{T}}$ must remain close to $\bar{x}$. The closeness of $\{x_t\}_{t=0}^{T}$ and $\{y_t\}_{t=0}^{T}$ to $\bar{x}$ implies that their differences can be modeled as a quadratic function with a small deviation term:

$$\begin{aligned} &P(\bar{x})^{1/2}(x_{t+1}-y_{t+1}) \\ =&P(\bar{x})^{1/2}(x_t-y_t)-\alpha P(\bar{x})^{1/2}(P(x)^{-1}\nabla f(x_t)-P(y)^{-1}\nabla f(y_t)) \\ =&\left(I-\alpha P(\bar{x})^{-1/2}\nabla^2 f(\bar{x})P(\bar{x})^{-1/2}\right)P(\bar{x})^{1/2}(x_t-y_t)+\alpha\xi(\bar{x},x_t,y_t) \end{aligned} \tag{37}$$

where the deviation term $\xi(\bar{x},x_t,y_t)$ is defined as

$$\xi(\bar{x},x_t,y_t) = P(\bar{x})^{1/2}\left(P(\bar{x})^{-1}\nabla^2 f(\bar{x})(x_t-y_t)-\left(P(x_t)^{-1}\nabla f(x_t)-P(y_t)^{-1}\nabla f(y_t)\right)\right)$$

Jin et al. (2021) showed that, when $P(x_t) = P(y_t) = P(\bar{x}) = I$, the deviation term $\xi(\bar{x}, x_t, y_t)$ remains small. However, such argument cannot be readily applied to the PMGD, since the metric function $P(x)$ can change drastically throughout the iterations. The crucial step in our proof is to show that the rate of change in $P(x)$ slows down drastically around stationary points.

Lemma 32 (Informal). *Let f be ℓ_1-gradient and ℓ_2-Hessian Lipschitz. Let $P(x)$ be well-conditioned and Lipschitz continuous. Suppose that u satisfies $\|\nabla f(u)\|_u^* \leq \epsilon$. Then, we have*

$$\|\zeta(u,x,y)\| \leq C_1 \max\{\|x-u\|, \|y-u\|\} \left\|P(u)^{1/2}(x-y)\right\| + C_2\epsilon \left\|P(u)^{1/2}(x-y)\right\|,$$

for constants C_1 and C_2 that only depend on $\ell_1, \ell_2, L_P, p_{\rm lb}, p_{\rm ub}$.

The formal version of Lemma 32 can be found in appendix (see Lemma 35). Recall that due to our assumption, the term $\max\{\|x_t - \bar{x}\|, \|y_t - \bar{x}\|\}$ must remain small for every $t \leq \mathcal{T}$. Therefore, applying Lemma 32 with $u = \bar{x}$, $x = x_t$, and $y = y_t$ implies that $\|\zeta(\bar{x}, x_t, y_t)\| = o(\|x_t - y_t\|)$. Therefore, (37) can be further approximated as

$$P(\bar{x})^{1/2}(x_t - y_t) \approx \left(I - \alpha P(\bar{x})^{-1/2}\nabla^2 f(\bar{x}) P(\bar{x})^{-1/2}\right)^t P(\bar{x})^{1/2}(x_0 - y_0) \tag{38}$$

Indeed, the above approximation enables us to argue that $P_{\bar{x}}^{1/2}(x_t - y_t)$ evolves according to a power iteration. In particular, suppose that x_0 and y_0 are picked such that $P(\bar{x})^{1/2}(x_0 - y_0) = cv$, where v is the eigenvector corresponding the smallest eigenvalue $\lambda_{\min}\left(P(\bar{x})^{-1/2}\nabla^2 f(\bar{x}) P(\bar{x})^{-1/2}\right) = -\gamma \leq -\sqrt{L_d\epsilon} < 0$. With this assumption, (38) can be approximated as the following power iteration:

$$P(\bar{x})^{1/2}(x_t - y_t) \approx c(1+\alpha\gamma)^t v \tag{39}$$

Suppose that $\{x_t\}_{t=0}^{\mathcal{T}}$ does not escape the strict saddle point, i.e., $f(x_t) - f(x_0) \geq -\mathcal{F}$. This implies that $\{x_t\}_{t=0}^{\mathcal{T}}$ remains close to $\bar{x}$. On the other hand, (39) implies that the sequence $\{y_t\}_{t=0}^{\mathcal{T}}$ must diverge from $\bar{x}$ *exponentially fast*, which is in direct contradiction with our initial assumption that $\{y_t\}_{t=0}^{\mathcal{T}}$ and $\bar{x}$ remain close. This in turn implies that $f(x_t) - f(\bar{x}) < -\mathcal{F}$ or $f(y_t) - f(\bar{x}) < -\mathcal{F}$. In other words, at least one of the sequences $\{x_t\}_{t=0}^{\mathcal{T}}$ and $\{y_t\}_{t=0}^{\mathcal{T}}$ must escape the strict saddle point. Considering $\{y_t\}_{t=0}^{\mathcal{T}}$ as a perturbed variant of $\{x_t\}_{t=0}^{\mathcal{T}}$, the above argument implies that it takes an *exponentially small* perturbation in the direction of v for PMGD to escape the strict saddle point $\bar{x}$. The sketch of the proof is completed by noting that such perturbation in the direction v is guaranteed to occur (with high probability) with a random isotropic perturbation of x_0.

8. Numerical Experiments

In this section, we provide numerical experiments to illustrate our theoretical findings. All experiments are implemented using MATLAB 2020b, and performed with a 2.6 Ghz 6-Core Intel Core i7 CPU with 16 GB of RAM.

We begin with numerical simulations for our local convergence result, Theorem 4, where we proved that PrecGD with an appropriately chosen regularizer η converges linearly towards the optimal solution $M^\star$, at a rate that is independent of both ill-conditioning and overparameterization. In contrast, gradient descent is slowed down significantly by both ill-conditioning and overparameterization.

We plot the convergence of GD and PrecGD for three choices of $\phi(\cdot)$, which correspond to the problems of low-rank matrix recovery (Candes and Plan, 2011), 1-bit matrix sensing (Davenport et al., 2014), and phase retrieval (Candes et al., 2013) respectively:

- **Low-rank matrix recovery with ℓ_2 loss**. Our goal is to find a low-rank matrix $M^\star \succeq 0$ that satisfies $\mathcal{A}(M^\star) = b$, where $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ is a linear operator and $b \in \mathbb{R}^m$ is given. To find $M^\star$, we minimize the objective

$$\phi(M) = \|\mathcal{A}(M) - b\|^2 \tag{40}$$

  subject to the constraint that M is low-rank. The Burer-Montiero formulation of this problem then becomes: minimize $f(X) = \|\mathcal{A}(XX^T) - b\|^2$ with $X \in \mathbb{R}^{n\times r}$.

- **1-bit matrix sensing**. The goal of 1-bit matrix recovery is to recovery a low-rank matrix $M^\star \succeq 0$ through 1-bit measurements of each entry M_{ij}. Each measurement on the M_{ij} is equal to 1 with probability $\sigma(M_{ij})$ and 0 with probability $1 - \sigma(M_{ij})$, where $\sigma(\cdot)$ is the sigmoid function. After a number of measurements have been taken, let α_{ij} denote the percentage of measurements on the (i, j)-entry that is equal to 1. To recover $M^\star$, we want to find the maximum likelihood estimator for $M^\star$ by minimizing

$$\phi(M) = \sum_{i=1}^{n}\sum_{j=1}^{n}(\log(1 + e^{M_{ij}}) - \alpha_{ij}M_{ij}). \tag{41}$$

  subject to the constraint $\operatorname{rank}(M) \le r$.

- **Phase retrieval**. The goal of phase retrieval is to recover a vector $z \in \mathbb{R}^d$ from m measurements of the form $y_i = |\langle a_i, z\rangle|^2$, where $a_i, i = 1, \dots, m$ are measurement vectors in $\mathbb{C}^d$. Equivalently, we can view this problem as recovering a rank-1 matrix zz^* from m linear measurements of the form $y_i = \langle a_i a_i^*, zz^*\rangle$. To find zz^* we minimize the ℓ_2 loss

$$\phi(M) = \sum_{i=1}^{m}\left(\langle a_i a_i^*, M\rangle - y_i\right)^2 \tag{42}$$

  subject to the constraint that M is rank-1.

One can readily check that both low-rank matrix recovery (40) and 1-bit matrix sensing (43) satisfy (μ, r)-restricted strong convexity (see Li et al. 2019 for details), so our theoretical results predict that PrecGD will converge linearly. While phase retrieval (42) *does not* satisfy restricted strong convexity, we will nevertheless see that PrecGD continues to converge linearly for phase retrieval. This indicates that PrecGD will continue to work well for more general optimization problems that present a low-rank structure. We leave the theoretical justifications of these numerical results as future work.

8.1 Low-rank matrix recovery with ℓ_2 loss

In this problem we assume that there is a $n \times n$, rank $r^\star$ matrix $M^\star \succeq 0$, which we call the ground truth, that we cannot observe directly. However, we have access to linear measurements of $M^\star$ in the form $b = A(M^\star)$. Here the linear operator $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ is defined as $\mathcal{A}(M^\star) = [\langle A_1, M^\star\rangle, \dots \langle A_m, M^\star\rangle]$, where each A_i is a fixed matrix of size $n \times n$. The goal is to recovery $M^\star$ from b, potentially with $m \ll n^2$ measurements by exploiting the low-rank structure of $M^\star$. This problem has numerous applications in areas such as colloborative filtering (Rennie and Srebro, 2005), quantum state tomography (Gross et al., 2010), power state estimation (Zhang et al., 2019a).

To recover $M^\star$, we minimize the objective $\phi(M)$ in (40) by solving the unconstrained problem

$$\min_{X\in\mathbb{R}^{n\times r}} f(X) = \phi(XX^T) = \|\mathcal{A}(XX^T) - b\|^2$$

using both GD and PrecGD. To gauge the effects of ill-conditioning, in our experiments we consider two choices of $M^\star$: one well-conditioned and one ill-conditioned. In the well-conditioned case, the ground truth is a rank-2 ($r^* = 2$) positive semidefinite matrix of size 100×100, where both of the non-zero eigenvalues are 1. To generate $M^\star$, we compute $M^\star = Q^T\Lambda Q$, where $\Lambda = \text{diag}(1, 1, 0, \dots, 0)$ and Q is a random orthogonal matrix of size $n \times n$ (sampled uniformly from the orthogonal group). In the ill-conditioned case, we set $M^\star = Q^T\Lambda Q$, where $\Lambda = \text{diag}(1, 1/5, 0, \dots, 0)$.

For each $M^\star$ we perform two set of experiments: the exactly-parameterized case with $r = 2$ and the overparameterized case where $r = 4$. The step-size is set to 2×10^{-6} for both GD and PrecGD in the first case and to 1×10^{-5} in the latter case. For PrecGD, the regularization parameter is set to $\eta = \|\nabla f(X)\|^*_{X,0}$. Both methods are initialized near the ground truth. In particular, we compute $M^\star = ZZ^T$ with $Z \in \mathbb{R}^{n\times r}$ and choose the initial point as $X_0 = Z + 10^{-2}w$, where w is a $n \times r$ random matrix with standard Gaussian entries. In practice, the closeness of a initial point to the ground truth can be guaranteed via spectral initialization (Chi et al. 2019; Tu et al. 2016). Finally, to ensure that $\phi(M) = \|A(M) - b\|^2$ satisfies restricted strong convexity, we set the linear operator $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ to be $\mathcal{A}(M^\star) = [\langle A_1, M^\star\rangle, \dots \langle A_m, M^\star\rangle]$, where $m = 3nr$ and each A_i is a standard Gaussian matrix with i.i.d. entries (Recht et al. 2010).

We plot the error $\|XX^T - M^\star\|_F$ versus the number of iterations for both GD and PrecGD. The results of our experiments for a well-conditioned $M^\star$ are shown on the first row of Figure 2. We see here that if $M^\star$ is well-conditioned and $r = r^*$, then GD converges at a linear rate and reaches machine precision quickly. The performance of PrecGD is almost identical. However once the search rank r exceeds the true rank r^*, then GD slows down significantly, as we can see from the figure on the right. In contrast, PrecGD continue to converge at a linear rate, obtaining machine accuracy within a few hundred iterations.

The results of our experiments for an ill-conditioned $M^\star$ are shown on the second row of Figure 2. We can see that ill-conditioning causes GD to slow down significantly, even in the exactly-parameterized case. In the overparameterized case, GD becomes even slower. On the other hand, PrecGD is fully agnostic to ill-conditioning and overparameterization. In fact, as Theorem 1 shows, the convergence rate of PrecGD is unaffected by both ill-conditioning and overparameterization.

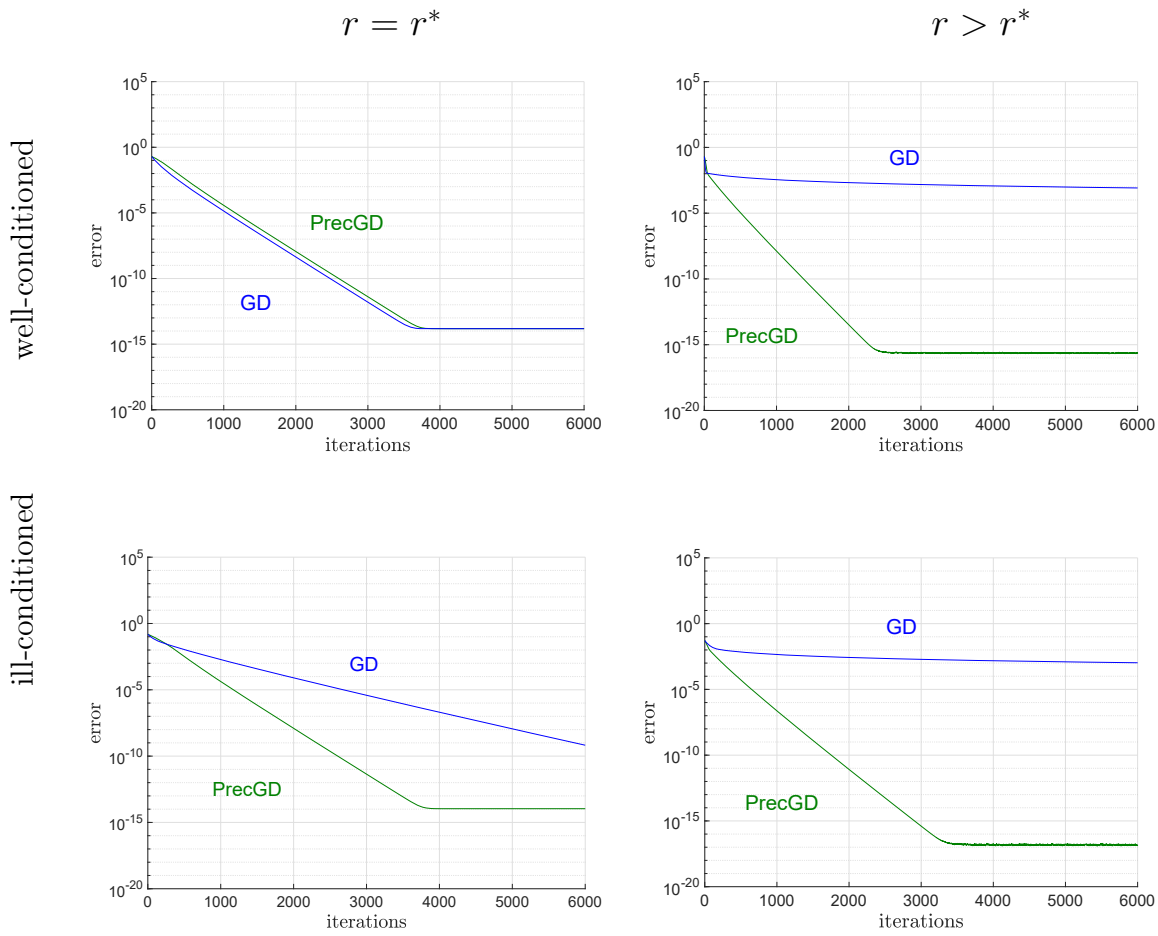

Figure 2: Low-rank matrix recovery with ℓ_2 loss. **First row**: Well-conditioned ($\kappa = 1$), rank-2 ground truth of size 100×100. The left panel shows the performance of GD and PrecGD for $r = r^* = 2$. Both algorithms converge linearly to machine error. The right panel shows the performance of GD and PrecGD for $r = 4$. The overparameterized GD converges sublinearly, while PrecGD maintains the same converge rate. **Second row**: Ill-conditioned ($\kappa = 5$), rank-2 ground truth of size 100×100. The left panel shows the performance of GD and PrecGD for $r = r^* = 2$. GD stagnates due to ill-conditioning while PrecGD converges linearly. The right panel shows the performance of GD and PrecGD for $r = 4$. The overparameterized GD continues to stagnate, while PrecGD maintains the same linear convergence rate.

In Figure 3 , we also plot a comparison of PrecGD, ScaledGD Tong et al. (2020) and GD with all three methods initialized at a random initial point and using the same step-size. Here, we see that both GD and PrecGD was able to converge towards the global solution, while ScaledGD behaves sporadically and diverges.

8.2 1-bit Matrix Sensing

Similar to low-rank matrix recovery, in 1-bit matrix sensing we also assume that there is a low rank matrix $M^\star \succeq 0$, which we call the ground truth, that we cannot observe directly,

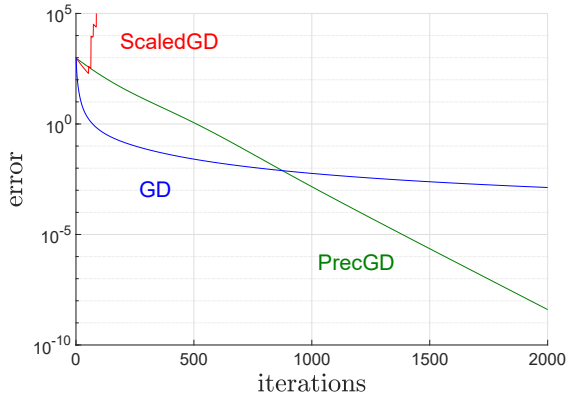

Figure 3: **PrecGD, ScaledGD and GD with random initialization.** Comparison of PrecGD against regular gradient descent (GD), and the ScaledGD algorithm. All three methods uses the same global Gaussian random initialization. The same step-size $\alpha = 2 \times 10^{-3}$ was used for all three algorithms. With $n = 4$, $r = 4$ and $r^* = 2$, overparameterization causes gradient descent to slow down to a sublinear rate. ScaledGD behaves sporadically and diverges. Only PrecGD converges linearly to the global minimum.

but have access to a total number of m 1-bit measurements of $M^\star$. Each measurement of M_{ij} is 1 with probability $\sigma(M_{ij})$ and 0 with probability $1 - \sigma(M_{ij})$, where $\sigma(\cdot)$ is the sigmoid function. This problem is a variant of the classical matrix completion problem and appears in applications where only quantized observations are available; see (Singer, 2008; Gross et al., 2010) for instance.

Let α_{ij} denote the percentage of measurements on the (i, j)-entry that is equal to 1. Then the MLE estimator can formulated as the minimizer of

$$\phi(M) = \sum_{i=1}^{n} \sum_{j=1}^{n} (\log(1 + e^{M_{ij}}) - \alpha_{ij} M_{ij}). \tag{43}$$

It is easy to check that $\nabla^2 \phi(M)$ is positive definite with bounded eigenvalues (see Li et al. 2019), so $\phi(M)$ satisfies the restricted strong convexity, which is required by Theorem 4.

To find the minimizer, we solve the problem $\min_{X \in \mathbb{R}^{n \times r}} \phi(XX^T)$ using GD and PrecGD. For presentation, we assume that the number of measurements m is large enough so that $\alpha_{ij} = \sigma(M_{ij})$. In this case the optimal solution of (43) is $M^\star$ and the error $\|XX^T - M^\star\|$ will go to zero when GD or PrecGD converges.

In our experiments, we use exactly the same choices of well- and ill-conditioned $M^\star$ as in Section 8.1. The rest of the experimental set up is also the same. We perform two set of experiments: (1) the exactly-parameterized case with $r = 2$ and (2) the overparameterized case where $r = 4$. Moreover, we use the same initialization scheme and same regularization parameter $\eta = \|\nabla f(X)\|^*_{X,0}$ for PrecGD. The step-size is chosen to be 0.5 in all four plots.

Our experiments are shown in Figure 4. We observe almost identical results as those of low-rank matrix recovery in Figure 2. In short, for 1-bit matrix sensing, both ill-conditioning

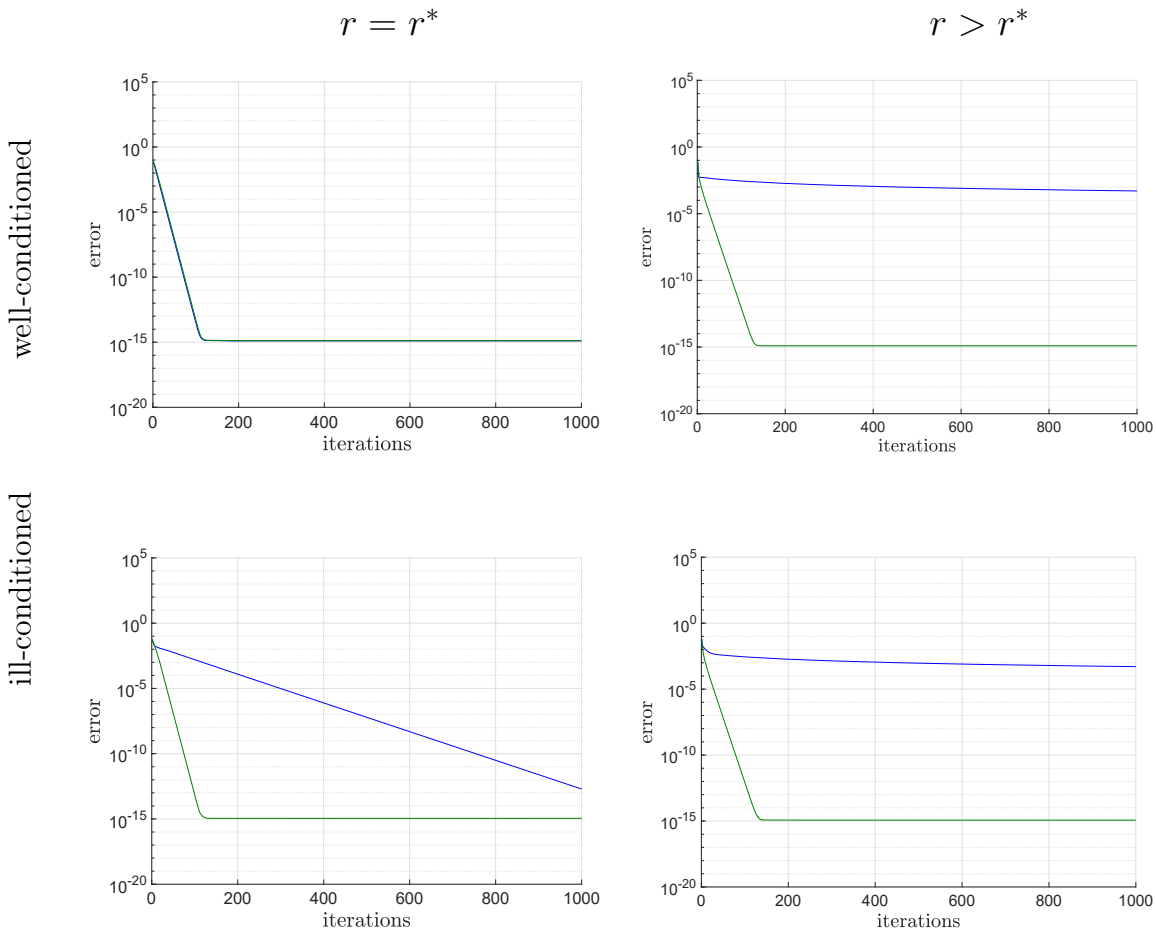

Figure 4: 1-bit matrix sensing. **First row**: Well-conditioned ($\kappa = 1$), rank-2 ground truth of size 100×100. The left panel shows the performance of GD and PrecGD for $r = r^* = 2$. Both algorithms converge linearly to machine error. The right panel shows the performance of GD and PrecGD for $r = 4$. The overparameterized GD converges sublinearly, while PrecGD maintains the same converge rate. **Second row**: Ill-conditioned ($\kappa = 10$), rank-2 ground truth of size 100×100. The left panel shows the performance of GD and PrecGD for $r = r^* = 2$. GD stagnates due to ill-conditioning while PrecGD converges linearly. The right panel shows the performance of GD and PrecGD for $r = 4$. The overparameterized GD continues to stagnate, while PrecGD maintains the same linear convergence rate.

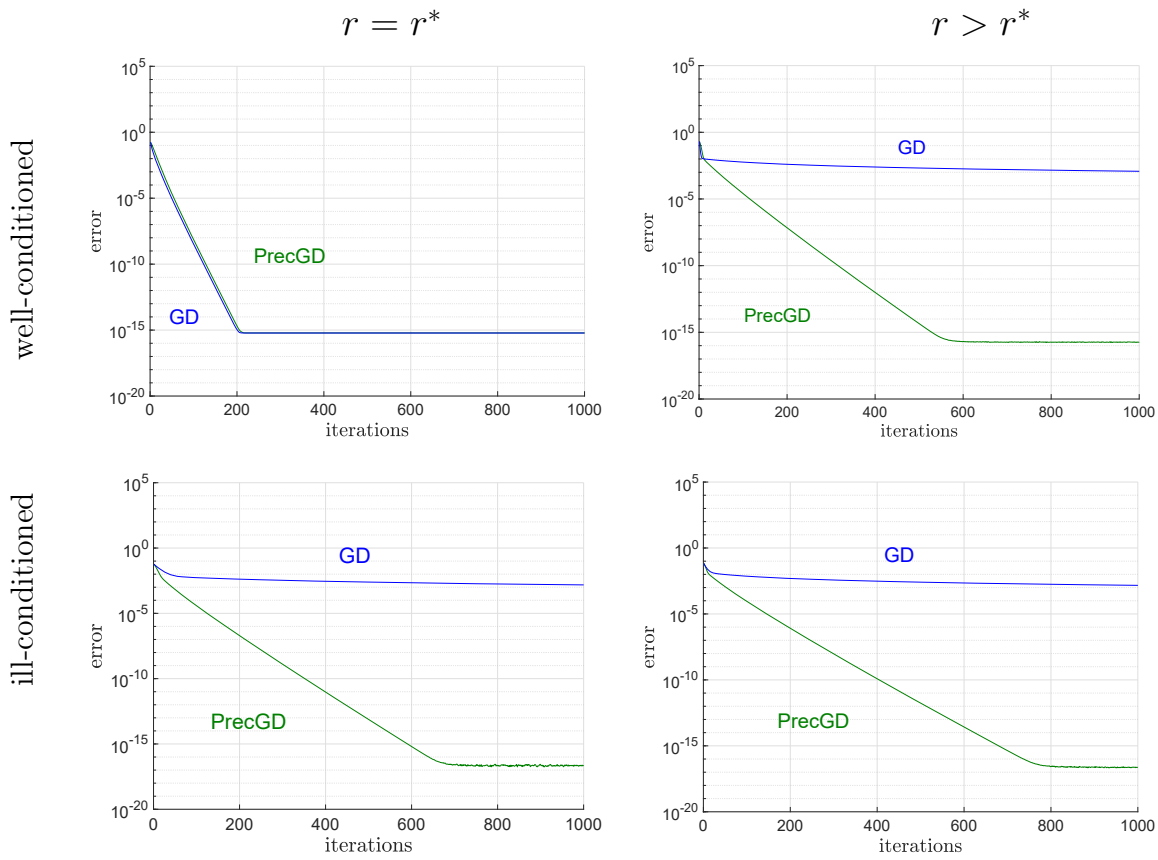

Figure 5: Phase retrieval. **First row**: Well-conditioned ($\kappa = 1$), rank-2 ground truth of size 100×100. The left panel shows the performance of GD and PrecGD for $r = r^* = 2$. Both algorithms converge linearly to machine error. The right panel shows the performance of GD and PrecGD for $r = 4$. The overparameterized GD converges sublinearly, while PrecGD maintains the same converge rate. **Second row**: Ill-conditioned ($\kappa = 5$), rank-2 ground truth of size 100×100. The left panel shows the performance of GD and PrecGD for $r = r^* = 2$. GD stagnates due to ill-conditioning while PrecGD converges linearly. The right panel shows the performance of GD and PrecGD for $r = 4$. The overparameterized GD continues to stagnate, while PrecGD maintains the same linear convergence rate.

and overparameterization causes gradient descent to slow down significantly, while PrecGD maintains a linear convergence rate independent of both.

8.3 Phase Retrieval

For our final set of experiments we consider the problem of recovering a low matrix $M^\star \succeq 0$ from *quadratic* measurements of the form $y_i = a_i^T M^\star a_i$ where $a_i \in \mathbb{R}^n$ are the measurement vectors. In general, the measurement vectors a_i can be complex, but for illustration purposes we focus on the case where the measurements are real. Suppose that we have a total of m

measurements, then our objective is

$$\min_{X\in\mathbb{R}^{n\times r}} f(X) = \sum_{i=1}^{m}(\|a_i^T X\|_F^2 - y_i)^2. \tag{44}$$

In the special case where $M^\star$ is rank-1, this problem is known as phase retrieval, which arises in a wide range of problems including crystallography (Harrison, 1993; Millane, 1990), diffraction and array imaging (Bunk et al., 2007), quantum mechanics (Corbett, 2006) and so on.

To gauge the effects of ill-conditioning in $M^\star$, we focus on the case where $M^\star$ is rank-2 instead. As before, we consider two choices of $M^\star$, one well-conditioned and one ill-conditioned, generated exactly the same way as the previous two problems. The measurement vectors a_i are chosen to be random vectors with standard Gaussian entries.

We perform two set of experiments: (1) the exactly-parameterized case with $r = 2$ and (2) the overparameterized case where $r = 4$. In the case $r = 2$, the step-size is set to 4×10^{-4} and in the case $r = 4$, the step-size is set to 10^{-4}. As before, both methods are initialized near the ground truth: we compute $M^\star = ZZ^T$ with $Z \in \mathbb{R}^{n\times r}$ and set the initial point $X_0 = Z + 10^{-2}w$, where w is a $n \times r$ random matrix with standard Gaussian entries.

Our experiments are shown in Figure 4. Even though the objective for phase retrieval no longer satisfies restricted strong convexity, we still observe the same results as before. Both ill-conditioning and overparameterization causes gradient descent to slow down significantly, while PrecGD maintains a linear convergence rate independent of both.

8.4 Certification of optimality

A key advantage of overparameterization is that it allows us to *certify* the optimality of a point X computed using local search methods. As we proved in Proposition 10, the suboptimality of a point X can be bounded as

$$f(X) - f(X^\star) \le C_g \cdot \epsilon_g + C_H \cdot \epsilon_H + C_\lambda \cdot \epsilon_\lambda. \tag{45}$$

Here we recall that $\langle \nabla f(X), V\rangle \le \epsilon_g \cdot \|V\|_F, \left\langle \nabla^2 f(X)[V], V\right\rangle \ge -\epsilon_H \cdot \|V\|_F^2$ for all V, and $\lambda_{\min}(X^T X) \le \epsilon_\lambda$. To evaluate the effectiveness of this optimality certificate, we consider three problems as before: matrix sensing with ℓ_2 loss, 1-bit matrix sensing, and phase retrieval. The experimental setup is the same as before. For each problem, we plot the function value $f(X) - f(X^\star)$ as the number of iterations increases, where $X^\star$ is the global minimizer of $f(\cdot)$. Additionally, we also compute the suboptimality as given by (45). The constants in (45) can be computed efficiently in linear time. For ϵ_H in particular, we apply the shifted power-iteration as described in Section 3.

The results are shown in Figures 6 and 7, for matrix sensing, phase retrieval, and 1-bit matrix sensing, respectively. We see that in each case, the upper bound in (45) indeed bounds the suboptimality $f(X) - f(X^\star)$. Moreover, this upper bound also converges linearly, albeit at a different rate. This slower rate is due to the fact that ϵ_g, the norm of the gradient, typically scales as $\sqrt{\epsilon_H}$ (Jin et al. 2017; Nesterov and Polyak 2006), hence it converges to 0 slower (by a square root). As a result, we see in all three plots that the upper bound converges slower roughly by a factor of a square root. In practice, this mean that if we want to certify n digits of accuracy within optimality, we would need our iterate to be accurate up to roughly $2n$ digits.

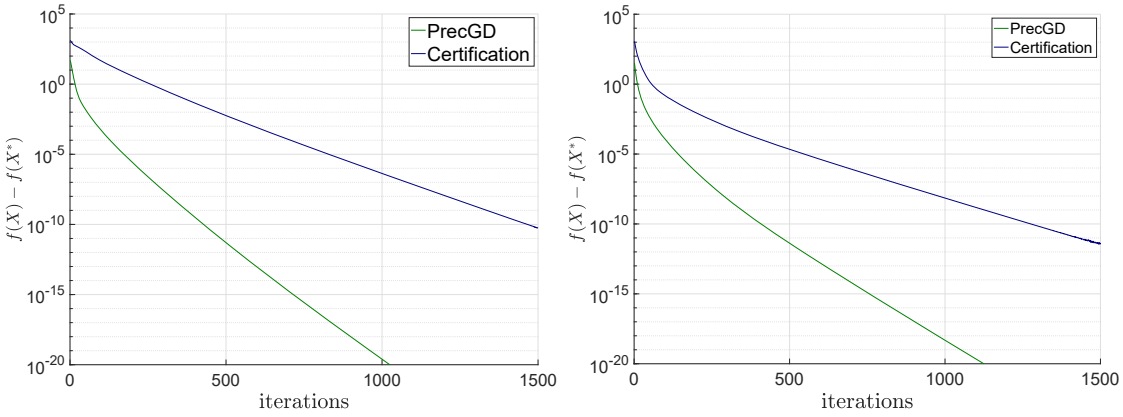

Figure 6: Certificate of global optimality. **Left**: Matrix sensing with a well-conditioned ($\kappa = 1$), rank-2 ground truth of size 100×100. The search rank is set to $r = 4$ and the algorithm is initialized within a neighborhood of radius 10^{-2} around the ground truth. The stepsize is set to 5×10^{-5}. **Right**: Phase retrieval with a well-conditioned ($\kappa = 1$), rank-2 ground truth of size 100×100. The search rank is set to $r = 4$ and the algorithm is initialized within a neighborhood of radius 10^{-2} around the ground truth. The step-size is set to 3×10^{-5}.

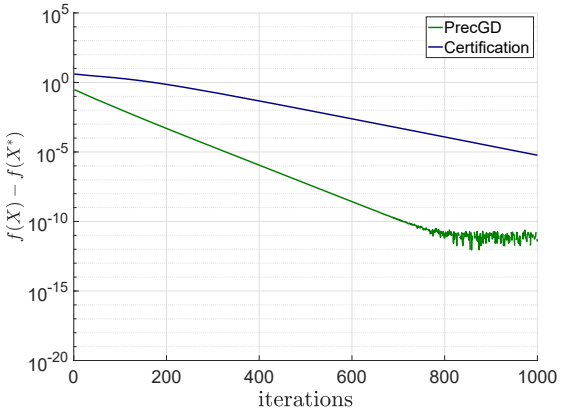

Figure 7: Certificate of global optimality for 1-bit matrix sensing with a well-conditioned ($\kappa = 1$), rank-2 ground truth of size 100×100. The search rank is set to $r = 4$, and the algorithm is initialized within a neighborhood of radius 10^{-2} around the ground truth. The step-size is set to 3×10^{-2}.

9. Conclusions

In this work, we consider the problem of minimizing a smooth convex function ϕ over a positive semidefinite matrix M. The Burer-Monteiro approach eliminates the large $n \times n$ positive semidefinite matrix by reformulating the problem as minimizing the nonconvex function $f(X) = \phi(XX^T)$ over an $n \times r$ factor matrix X. Here, we overparameterize the search rank $r > r^\star$ with respect to the true rank $r^\star$ of the solution $X^\star$, because the rank deficiency of a second-order stationary point X allows us to *certify* that X is globally optimal.

Unfortunately, gradient descent becomes extremely slow once the problem is overparameterized. Instead, we propose a method known as PrecGD, which enjoys a similar per-iteration cost as gradient descent, but speeds up the convergence rate of gradient descent exponentially in overparameterized case. In particular, we prove that within a neighborhood around the ground truth, PrecGD converges linearly towards the ground truth, at a rate independent of both ill-conditioning in the ground truth and overparameterization. We also prove that, similar to gradient descent, a perturbed version of PrecGD converges globally, from any initial point.

Our numerical experiments find that preconditioned gradient descent works equally well in restoring the linear convergence of gradient descent in the overparmeterized regime for choices of ϕ that do not satisfy restricted strong convexity. We leave the justification of these results for future work.

Acknowledgments

Financial support for this work was provided by NSF CAREER Award ECCS-2047462, NSF Award DMS-2152776, ONR Award N00014-22-1-2127.

Appendix A. Proof of Basis Alignment (Lemma 19)

For $X \in \mathbb{R}^{n\times r}$ and $Z \in \mathbb{R}^{n\times r^\star}$, suppose that X satisfies

$$\rho \overset{\text{def}}{=} \frac{\|XX^T - ZZ^T\|_F}{\lambda_{\min}(Z^TZ)} < \frac{1}{\sqrt{2}}. \tag{46}$$

In this section, we prove that the incidence angle θ defined as

$$\cos\theta = \max_{Y\in\mathbb{R}^{n\times r}} \frac{\langle XX^T - ZZ^T, XY^T + YX^T\rangle}{\|XX^T - ZZ^T\|_F\|XY^T + YX^T\|_F},$$

satisfies

$$\sin\theta = \frac{\|(I - XX^\dagger)ZZ^T(I - XX^\dagger)\|_F}{\|XX^T - ZZ^T\|_F} \le \frac{1}{\sqrt{2}}\frac{\rho}{\sqrt{1-\rho^2}} \tag{47}$$

where $\dagger$ denotes the pseudoinverse.

First, note that an X that satisfies (46) must have $\operatorname{rank}(X) \ge r^\star$. This follows from Weyl's inequality

$$\lambda_{r^\star}(XX^T) \ge \lambda_{r^\star}(ZZ^T) - \|XX^T - ZZ^T\|_F \ge \left(1 - \frac{1}{\sqrt{2}}\right)\cdot\lambda_{r^\star}(ZZ^T).$$

Next, due to the rotational invariance of this problem, we can assume without loss of generality[1] that X, Z are of the form

$$X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}, \quad Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}, \quad \sigma_{\min}(X_1) \geq \sigma_{\max}(X_2) \tag{48}$$

where $X_1 \in \mathbb{R}^{k\times k}$, $Z_1 \in \mathbb{R}^{k\times r^\star}$. For $k \geq r^\star$, the fact that $\text{rank}(X) \geq r^\star$ additionally implies that $\sigma_{\min}(X_1) > 0$.

The equality in (47) immediately follows by setting $k = \text{rank}(X)$ and solving the projection problem

$$\begin{aligned} \|E\|_F \sin\theta &= \min_Y \|(XY^T + YX^T) - (XX - ZZ^T)\|_F \\ &= \min_{Y_1,Y_2} \left\| \begin{bmatrix} X_1Y_1^T + Y_1X_1^T & X_1Y_2^T \\ Y_2X_1^T & 0 \end{bmatrix} - \begin{bmatrix} X_1X_1^T - Z_1Z_1^T & Z_2Z_1^T \\ Z_1Z_2^T & -Z_2Z_2^T \end{bmatrix} \right\|_F \\ &= \|Z_2Z_2^T\|_F = \|(I - XX^\dagger)ZZ^T(I - XX^\dagger)\|_F. \end{aligned}$$

Before we prove the inequality in (47), we state and prove a technical lemma that will be used in the proof.

Lemma 33. *Suppose that X, Z are in the form in (48), and $k \geq r^\star$. If ρ defined in (46) satisfies $\rho < 1/\sqrt{2}$, we have $\lambda_{\min}(Z_1^TZ_1) \geq \lambda_{\max}(Z_2^TZ_2)$.*

Proof Denote $\gamma_1 = \lambda_{\min}(Z_1^TZ_1)$ and $\gamma_2 = \lambda_{\max}(Z_2^TZ_2)$. By contradiction, we will prove that $\gamma_1 < \gamma_2$ implies $\rho \geq 1/\sqrt{2}$. This claim is invariant to scaling of X and Z, so we assume without loss of generality that $\lambda_{\min}(Z^TZ) = 1$. Under (48), our radius hypothesis $\rho \geq \|XX^T - ZZ^T\|_F$ reads

$$\begin{aligned} \rho^2 &\geq \|X_1X_1^T - Z_1Z_1^T\|_F^2 + 2\|Z_1Z_2^T\|_F^2 + \|X_2X_2^T - Z_2Z_2^T\|_F^2. \\ &\geq \|X_1X_1^T - Z_1Z_1^T\|_F^2 + \|X_2X_2^T - Z_2Z_2^T\|_F^2 + 2\lambda_{\min}(Z_1^TZ_1)\lambda_{\max}(Z_2^TZ_2). \end{aligned}$$

Below, we will prove that X_1, X_2 that satisfy $\sigma_{\min}(X_1) \geq \sigma_{\max}(X_2)$ also satisfies

$$\begin{aligned} &\|X_1X_1^T - Z_1Z_1^T\|_F^2 + \|X_2X_2^T - Z_2Z_2^T\|_F^2 \\ \geq &\min_{d_1,d_2\in\mathbb{R}_+} \{[d_1 - \gamma_1]^2 + [d_2 - \gamma_2]^2 : d_1 \geq d_2\} \end{aligned} \tag{49}$$

If $\gamma_1 < \gamma_2$, then $d_1 = d_2$ holds at optimality, so the minimum value is $\frac{1}{2}(\gamma_1 - \gamma_2)^2$. Substituting $\gamma_1 = \lambda_{\min}(Z_1^TZ_1)$ and $\gamma_2 = \lambda_{\max}(Z_2^TZ_2)$ then proves that

$$\rho^2 \geq \frac{(\gamma_1 - \gamma_2)^2}{2} + 2\gamma_1\gamma_2 = \frac{1}{2}(\gamma_1 + \gamma_2)^2.$$

But we also have

$$\gamma_1 + \gamma_2 = \lambda_{\min}(Z_1^TZ_1) + \lambda_{\max}(Z_2^TZ_2) \geq \lambda_{\min}(Z_1^TZ_1 + Z_2^TZ_2) = \lambda_{\min}(Z^TZ) = 1$$

1. 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 $Z \leftarrow U^TZ$.

and this implies $\rho^2 \geq 1/2$, a contradiction.

We now prove (49). Consider the following optimization problem

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

We relax $X_1X_1^T$ into $S_1 \succeq 0$ and $X_2X_2^T$ into $S_2 \succeq 0$ to yield a lower-bound

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

The problem is invariant to a change of basis, so we change into the eigenbases of $Z_1Z_1^T$ and $Z_2Z_2^T$ to yield

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

where $\lambda(Z_1Z_1^T) \geq 0$ and $\lambda(Z_2Z_2^T) \geq 0$ denote the vector of eigenvalues. We lower-bound this problem by dropping all the terms in the sum of squares except the one associated with $\gamma_1 = \lambda_{\min}(Z_1^TZ_1)$ and $\gamma_2 = \lambda_{\max}(Z_2Z_2^T)$ to obtain

$$\geq \min_{d_1,d_2\in\mathbb{R}_+} \{[d_1 - \gamma_1]^2 + [d_2 - \gamma_2]^2 : d_1 \geq d_2\} \tag{50}$$

which is exactly (49) as desired. ■

Now we are ready to prove the inequality in (47).

Proof For any $k \geq r^\star$ and within the radius $\rho < 1/\sqrt{2}$, we begin by proving that the following incidence angle between X and Z satisfies

$$\sin\phi \overset{\text{def}}{=} \frac{\|(I - XX^\dagger)Z\|_F}{\sigma_{r^\star}(Z)} \leq \frac{\|Z_2\|_F}{\sqrt{\lambda_{\min}(Z^TZ)}} \leq \frac{\|XX^T - ZZ^T\|_F}{\lambda_{\min}(Z^TZ)} = \rho. \tag{51}$$

This follows from the following chain of inequalities

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

where we use $\lambda_{\min}(Z_1^TZ_1) \geq \frac{1}{2}\lambda_{\min}(Z^TZ)$ because

$$\lambda_{\min}(Z^TZ) = \lambda_{\min}(Z_1^TZ_1 + Z_2^TZ_2) \leq \lambda_{\min}(Z_1^TZ_1) + \lambda_{\max}(Z_2^TZ_2) \leq 2\lambda_{\min}(Z_1^TZ_1) \tag{52}$$

where we used $\lambda_{\min}(Z_1^TZ_1) \geq \lambda_{\max}(Z_2^TZ_2)$ via Lemma 33. Moreover, for any $k \geq r^\star$, we have $\sigma_{\min}(X_1) > 0$ and therefore

$$\|(I - XX^\dagger)ZZ^T(I - XX^\dagger)\|_F = \|(I - X_2X_2^\dagger)Z_2Z_2^T(I - X_2X_2^\dagger)\|_F \leq \|Z_2Z_2^T\|_F.$$

Then, (47) is true because

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

Step (a) bounds $\|Z_2 Z_2^T\|_F \le \|Z_2\|_F^2$ and $2\langle Z_1^T Z_1, Z_2^T Z_2\rangle \le \|XX^T - ZZ^T\|_F^2$. Step (b) bounds $\langle Z_1^T Z_1, Z_2^T Z_2\rangle \ge \lambda_{\min}(Z_1^T Z_1)\cdot \mathrm{tr}(Z_2 Z_2^T)$. Step (c) uses (52) and $\|Z_2\|_F^2 \ge \lambda_{\max}(Z_2^T Z_2)$. Finally, step (d) substitutes (51). ■

Appendix B. Proof of Gradient Lipschitz (Lemma 22)

Proof Let ϕ be L-gradient Lipschitz. Let $M^\star = \arg\min \phi$ satisfy $M^\star \succeq 0$. In this section, we prove that $f(X) \overset{\text{def}}{=} \phi(XX^T)$ satisfies

$$f(X+V) \le f(X) + \langle \nabla f(X), V\rangle + \frac{L}{2}\gamma_{X,\eta}\|V\|_{X,\eta}^2$$
$$\text{where } \gamma_{X,\eta} = 4 + \frac{2\|E\|_F + 4\|V\|_{X,\eta}}{\lambda_{\min}+\eta} + \frac{\|V\|_{X,\eta}^2}{(\lambda_{\min}+\eta)^2}$$

where $\|V\|_{X,\eta} = \|V(X^TX+\eta I)^{-1/2}\|_F$ and $\lambda_{\min} \equiv \lambda_{\min}(X^TX)$ and $E = XX^T - M^\star$. First, it follows from the L-gradient Lipschitz property of ϕ that

$$\underbrace{\phi((X+V)(X+V)^T)}_{f(X+V)} = \underbrace{\phi(XX^T)}_{f(X)} + \underbrace{\left\langle \nabla\phi(XX^T), XV^T + VX^T\right\rangle}_{\langle \nabla f(X), V\rangle} + \left\langle \nabla\phi(XX^T), VV^T\right\rangle + \frac{L}{2}\|XV^T + VX^T + VV^T\|_F^2.$$

Substituting the following

$$\|VX^T\|_F \le \|(X^TX+\eta I)^{-1/2}X^T\|\cdot\|V\|_{X,\eta} \le \|V\|_{X,\eta}$$
$$\|VV^T\|_F \le \|(X^TX+\eta I)^{-1}\|\cdot\|V\|_{X,\eta}^2 = [\lambda_{\min}(X^TX)+\eta]^{-1}\|V\|_{X,\eta}^2$$
$$\|\nabla\phi(XX^T)\|_F = \|\nabla\phi(XX^T) - \nabla\phi(M^\star)\|_F \le L\|E\|_F,$$

bounds the error term

$$\begin{aligned}
&\left\langle \nabla\phi(XX^T), VV^T\right\rangle + \frac{L}{2}\|XV^T + VX^T + VV^T\|_F^2\\
=&\frac{\|\nabla\phi(XX^T)\|_F\|V\|_{X,\eta}^2}{\lambda_{\min}+\eta} + \frac{L}{2}\left(4\|V\|_{X,\eta}^2 + \frac{4\|V\|_{X,\eta}^3}{\lambda_{\min}+\eta} + \frac{\|V\|_{X,\eta}^4}{(\lambda_{\min}+\eta)^2}\right)\\
\le&\frac{L\cdot\|V\|_{X,\eta}^2}{2}\left(4 + \frac{2\|E\|_F + 4\|V\|_{X,\eta}}{\lambda_{\min}+\eta} + \frac{\|V\|_{X,\eta}^2}{(\lambda_{\min}+\eta)^2}\right).
\end{aligned}$$

This completes the proof. ■

Appendix C. Proof of Bounded Gradient (Lemma 23)

Proof Let ϕ be L-gradient Lipschitz, and let $f(X) \overset{\text{def}}{=} \phi(XX^T)$. Let $M^\star = \arg\min\phi$ satisfy $M^\star \succeq 0$. In this section, we prove that $V = \nabla f(X)(X^TX+\eta I)^{-1}$ satisfies

$$\|V\|_{X,\eta} = \|\nabla f(X)\|_{X,\eta}^* \le 2L\|XX^T - M^\star\|_F,$$

where $\|V\|_{X,\eta} = \|VP_{X,\eta}^{1/2}\|_F$ and $\|\nabla f(X)\|_{X,\eta}^* = \|\nabla f(X)P_{X,\eta}^{-1/2}\|_F$ and $P_{X,\eta} = X^TX + \eta I$. Indeed, $\|V\|_{X,\eta} = \|\nabla f(X)\|_{X,\eta}^*$ can be verified by inspection. We have

$$\begin{aligned}\|\nabla f(X)\|_{X,\eta}^* &= \max_{\|Y\|_{X,\eta}=1} \left\langle \nabla\phi(XX^T), XY^T + YX^T \right\rangle \\ &\le \|\nabla\phi(XX^T)\|_F \|XY^{\star T} + Y^\star X^T\|_F \\ &\le \|\nabla\phi(XX^T) - \nabla\phi(M^\star)\|_F \left(2\|X(X^TX+\eta I)^{-1/2}\| \cdot \|Y^\star\|_{X,\eta}\right) \\ &\le L\|XX^T - M^\star\|_F \cdot 2.\end{aligned}$$

This completes the proof. ■

Appendix D. Proofs of Global Convergence

In this section, we provide the proofs of Lemmas 31 and 29 which play critical roles in proving the global convergence of PMGD (Theorem 28) and PPrecGD (Corollary 8).

D.1 Proof of Lemma 31

To proceed with the proof of Lemma 31, we define the following quantities to streamline our presentation:

$$\alpha = \frac{p_{\rm lb}}{2\ell_1}, \quad \beta = \frac{\epsilon}{400L_d\cdot\iota^3}, \quad \mathcal{T} = \frac{\ell_1}{p_{\rm lb}\sqrt{L_d\epsilon}}\cdot\iota, \quad \mathcal{F} = \frac{p_{\rm lb}}{50\iota^3}\sqrt{\frac{\epsilon^3}{L_d}}, \quad \mathcal{S} := \frac{1}{5\iota}\sqrt{\frac{\epsilon}{L_d}} \tag{54}$$

$$L_d = \frac{5\max\{\ell_2, L_P\ell_1\sqrt{p_{\rm ub}}\}}{p_{\rm lb}^{2.5}}, \quad \iota = c\cdot\log\left(\frac{p_{\rm ub}d\ell_1(f(x_0)-f^*)}{p_{\rm lb}\ell_2\epsilon\delta}\right) \tag{55}$$

for some absolute constant c. Once Lemma 34 below is established, the proof of Lemma 31 follows by identically repeating the arguments in the proof of (Jin et al., 2021, Lemma 5.3).

Lemma 34 (Coupling Sequence). *Suppose that $\bar{x}$ satisfies $\|\nabla f(\tilde{x})\|_{\tilde{x}}^* \le \epsilon$ and $\nabla^2 f(\tilde{x}) \not\succeq -\sqrt{L_d\epsilon}\cdot P(\bar{x})$. Let $\{x_t\}_{t=0}^{\mathcal{T}}$ and $\{y_t\}_{t=0}^{\mathcal{T}}$ be two sequences generated by PMGD initialized respectively at x_0 and y_0 which satisfy: (1) $\max\{\|x_0-\bar{x}\|, \|y_0-\bar{x}\|\} \le \alpha\beta$; and (2) $P(\bar{x})^{1/2}(x_0-y_0) = \alpha\omega\cdot v$, where v is the eigenvector corresponding to the minimum eigenvalue of $P(\bar{x})^{-1/2}\nabla^2 f(\bar{x})P(\bar{x})^{-1/2}$ and $\omega > \bar{\omega} := 2^{3-\iota/4}\cdot\mathcal{S}$. Then:*

$$\min\{f(x_{\mathcal{T}}) - f(x_0), f(y_{\mathcal{T}}) - f(y_0)\} \le -2\mathcal{F}.$$

We point out that the Lemma 34 is *not* a direct consequence of (Jin et al., 2021, Lemma 5.5) which shows a similar result but for the perturbed gradient descent. The key difference in our analysis is the precise control of the general metric function as a preconditioner for the gradients. In particular, we show that, while in general $P(x)$ and $P(y)$ can be drastically different for different values of x and y, they can essentially be treated as constant matrices in the vicinity of strict saddle points. More precisely, according to (37), the term $P_{\bar{x}}^{1/2}(x_{t+1}-y_{t+1})$ can be written as

$$P_{\bar{x}}^{1/2}(x_{t+1}-y_{t+1}) = \left(I - \alpha P(\bar{x})^{-1/2}\nabla^2 f(\bar{x})P(^{-1/2}\right)P(\bar{x})^{1/2}(x_t-y_t) + \xi(\bar{x},x_t,y_t)$$

where $\xi(\bar{x}, x_t, y_t)$ is a deviation term defined as

$$\xi(\bar{x}, x_t, y_t) = \alpha P_{\bar{x}}^{1/2} \left(P_{\bar{x}}^{-1} \nabla^2 f(\bar{x})(x_t - y_t) - \left(P_{x_t}^{-1} \nabla f(x_t) - P_{y_t}^{-1} \nabla f(y_t) \right) \right)$$

Our goal is to show that $\xi(\bar{x}, x_t, y_t)$ remains small for every $t \leq \mathcal{T}$.

Lemma 35. *Let f be ℓ_1-gradient and ℓ_2-Hessian Lipschitz. Let $p_{\rm lb} I \preceq P(x) \preceq p_{\rm ub} I$ and $\|P(x) - P(y)\| \leq L_P \|x - y\|$ for every x and y. Suppose that u satisfies $\|\nabla f(u)\|_u^* \leq \epsilon$. Moreover, suppose that x and y satisfy $\max\{\|x - u\|, \|y - u\|\} \leq \mathcal{S}$ and $\epsilon \leq \mathcal{S}/\sqrt{p_{\rm ub}}$ for some $\mathcal{S} \geq 0$. Then, we have*

$$\|\zeta(u, x, y)\| \leq L_d \mathcal{S} \left\| P(u)^{1/2} (x - y) \right\|.$$

Proof Note that we can write

$$\begin{aligned}
\xi(u, x, y) =& - P(u)^{1/2} P(x)^{-1} \nabla f(x) + P(u)^{1/2} P(y)^{-1} \nabla f(y) + P(u)^{-1/2} \nabla^2 f(u)(x - y) \\
=& \underbrace{-P(u)^{-1/2} \nabla f(x) + P(u)^{-1/2} \nabla f(y) + P(u)^{-1/2} \nabla^2 f(u)(x - y)}_{T_1} \\
& - \underbrace{P(u)^{1/2} \left((P(x)^{-1} \nabla f(x) - P(u)^{-1} \nabla f(x)) + (P(u)^{-1} \nabla f(y) - P(y)^{-1} \nabla f(y)) \right)}_{T_2}.
\end{aligned}$$

We bound the norm of T_1 and T_2 separately. First, we have

$$\begin{aligned}
\|T_1\| &= \|P(u)^{-1/2} \left(\nabla f(x) - \nabla f(y) - \nabla^2 f(u)(x - y) \right) \| \\
&\leq \frac{1}{\sqrt{p_{\rm lb}}} \|\nabla f(x) - \nabla f(y) - \nabla^2 f(u)(x - y)\| \\
&= \frac{1}{\sqrt{p_{\rm lb}}} \left\| \int_0^1 \left(\nabla^2 f(y + t(x - y)) - \nabla^2 f(u) \right) \, dt \cdot (x - y) \right\| \\
&\leq \frac{\ell_2}{\sqrt{p_{\rm lb}}} \cdot \max\{\|x - u\|, \|y - u\|\} \cdot \|x - y\|,
\end{aligned}$$

where the last inequality follows from the assumption that the Hessian is Lipscthiz. On the other hand, we have

$$\begin{aligned}
\|T_2\| &\leq \sqrt{p_{\rm ub}} \left\| P(x)^{-1} \nabla f(x) - P(u)^{-1} \nabla f(x) + P(u)^{-1} \nabla f(y) - P(y)^{-1} \nabla f(y) \right\| \\
&= \left\| (P(x)^{-1} - P(u)^{-1})(\nabla f(x) - \nabla f(y)) + (P(x)^{-1} - P(y)^{-1}) \nabla f(y) \right\| \\
&\leq \frac{L_P \sqrt{p_{\rm ub}}}{p_{\rm lb}^2} \|x - u\| \cdot \ell_1 \|x - y\| + \frac{L_P \sqrt{p_{\rm ub}}}{p_{\rm lb}^2} \|x - y\| \|\nabla f(y)\|.
\end{aligned}$$

Since the gradient is ℓ_1-Lipschitz, we have

$$\|\nabla f(y)\| \leq \|\nabla f(u)\| + \ell_1 \|y - u\| \leq p_{\rm ub}^{1/2} \cdot \|\nabla f(u)\|_u^* + \ell_1 \|y - u\| \leq p_{\rm ub}^{1/2} \epsilon + \ell_1 \|y - u\|.$$

As a result, we get

$$\|T_2\| \leq \frac{L_P \sqrt{p_{\rm ub}}}{p_{\rm lb}^2} \|x - u\| \cdot \ell_1 \|x - y\| + \frac{L_P \sqrt{p_{\rm ub}}}{p_{\rm lb}^2} \|x - y\| \left(\sqrt{p_{\rm ub}} \epsilon + \ell_1 \|y - u\| \right)$$

Combining the derived upper bounds for T_1 and T_2 leads to

$$\|\zeta(u,x,y)\| \le \frac{2\max\{\ell_2, L_P\ell_1\sqrt{p_{\rm ub}}\}}{p_{\rm lb}^2}\left(\|x-u\|+\|y-u\|\right)\|x-y\| + \frac{L_P p_{\rm ub}}{p_{\rm lb}^2}\epsilon\|x-y\|.$$

Invoking the assumptions $\max\{\|x-u\|, \|y-u\|\} \le \mathcal{S}$ and $\epsilon \le \mathcal{S}/\sqrt{p_{\rm ub}}$ yields

$$\begin{aligned}\|\zeta(u,x,y)\| &\le \frac{5\max\{\ell_2, L_P\ell_1\sqrt{p_{\rm ub}}\}}{p_{\rm lb}^2}\cdot\mathcal{S}\|x-y\| \\ &\le \frac{5\max\{\ell_2, L_P\ell_1\sqrt{p_{\rm ub}}\}}{p_{\rm lb}^{2.5}}\cdot\mathcal{S}\|P(u)^{1/2}(x-y)\|\end{aligned}$$

This completes the proof. ■

To prove Lemma 34, we also need the following lemma, which shows that, for $t \le \mathcal{T}$, the iterations $\{x_t\}$ remain close to the initial point x_0 if $f(x_0) - f(x_t)$ is small. The proof is almost identical to that of (Jin et al., 2021, Lemma 5.4) and is omitted for brevity.

Lemma 36 (Improve or Localize)**.** *Under the assumptions of Lemma 31, we have for every $t \le \mathcal{T}$:*

$$\|x_t - x_0\| \le \frac{1}{\sqrt{p_{\rm lb}}}\sqrt{2\alpha t(f(x_0)-f(x_t))}.$$

Proof [Lemma 34.] By contradiction, suppose that

$$\min\{f(x_{\mathcal{T}}) - f(x_0), f(y_{\mathcal{T}}) - f(y_0)\} > -2\mathcal{F}.$$

Given this assumption, we can invoke Lemma 36 to show that both sequences remain close to $\bar{x}$, i.e., for any $t \le \mathcal{T}$:

$$\max\{\|x_t-\bar{x}\|, \|y_t-\bar{x}\|\} \le \frac{1}{\sqrt{p_{\rm lb}}}\sqrt{4\alpha\mathcal{T}\mathcal{F}} = \sqrt{\frac{p_{\rm lb}\epsilon}{25L_d^2\iota^2\ell_2}} \le \frac{1}{5\iota}\sqrt{\frac{\epsilon}{L_d}} := \mathcal{S} \tag{56}$$

where the first equality follows from our choice of α, $\mathcal{T}$, $\mathcal{F}$, and r. Upon defining $z_t = P(\bar{x})^{1/2}(x_t - y_t)$, we have

$$\begin{aligned}z_{t+1} &= z_t - \alpha P(\bar{x})^{-1/2}[P(x_t)^{-1}\nabla f(x_t) - P(y_t)^{-1}\nabla f(y_t)] \\ &= (I-\alpha H)z_t - \alpha\xi(\bar{x}, x_t, y_t) \\ &= \underbrace{(I-\alpha H)^{t+1}z_0}_{p(t+1)} - \underbrace{\alpha\sum_{\tau=0}^{t}(I-\alpha H)^{t-\tau}\xi(\bar{x}, x_\tau, y_\tau)}_{q(t+1)},\end{aligned}$$

where $H = P(\bar{x})^{-1/2}\nabla^2 f(\bar{x})P(\bar{x})^{-1/2}$, and

$$\xi(\bar{x}, x_t, y_t) = \alpha P(\bar{x})^{1/2}\left(P(\bar{x})^{-1}\nabla^2 f(\bar{x})(x_t - y_t) - \left(P(x_t)^{-1}\nabla f(x_t) - P(y_t)^{-1}\nabla f(y_t)\right)\right)$$

In the dynamic of z_{t+1}, the term $p(t+1)$ captures the effect of the difference in the initial points of the sequences $\{x_t\}_{t=0}^T$ and $\{y_t\}_{t=0}^T$. Moreover, the term $q(t+1)$ is due to the fact that the function f is not quadratic and the metric function $P(x)$ changes along the solution trajectory. We now use induction to show that the error term $q(t)$ remains smaller than the leading term $p(t)$. In particular, we show

$$\|q(t)\| \le \|p(t)\|/2, \qquad t \in \mathcal{T}.$$

The claim is true for the base case $t = 0$ as $\|q(0)\| = 0 \le \|z_0\|/2 = \|p(0)\|/2$. Now suppose the induction hypothesis is true up to t. Denote $\lambda_{\min}(H) = -\gamma$ with $\gamma \ge \sqrt{L_d\epsilon}$. Note that z_0 lies in the direction of the minimum eigenvector of H. Thus, for any $\tau \le t$, we have

$$\|z_\tau\| \le \|p(\tau)\| + \|q(\tau)\| \le 2\|p(\tau)\| = 2\|(I-\alpha H)^\tau z_0\| = 2(1+\alpha\gamma)^\tau \alpha\omega. \tag{57}$$

On the other hand, we have

$$\begin{aligned}
\|q(t+1)\| = \left\|\alpha\sum_{\tau=0}^{t}(I-\eta H)^{t-\tau}\zeta(\bar{x},x_t,y_t)\right\| &\le \alpha\sum_{\tau=0}^{t}\|(I-\eta H)^{t-\tau}\|\cdot L_d\mathcal{S}\|z_\tau\| \\
&\le \alpha\sum_{\tau=0}^{t}\|(I-\eta H)^{t-\tau}\|\cdot L_d\mathcal{S}\cdot(2(1+\alpha\gamma)^\tau\alpha\omega) \le 2\alpha L_d\mathcal{S}\sum_{\tau=0}^{t}(1+\alpha\gamma)^t\alpha\omega \\
&\le 2\alpha L_d\mathcal{S}\mathcal{T}p(t+1)
\end{aligned}$$

where in the first inequality we used Lemma 35 to bound $\|\zeta(\bar{x},x_t,y_t)\|$. Moreover, in the second and last inequalities we used (57), $t \le \mathcal{T}$, and $(1+\alpha\gamma)^t\alpha\omega \le p(t+1)$. Due to our choice of α, L_d, $\mathcal{S}$, and $\mathcal{T}$, it is easy to see that $2\alpha L_d\mathcal{S}\mathcal{T} = 1/5$. This leads to $\|q(t+1)\| \le \|p(t+1)\|/5$, thereby completing our inductive argument. Based on this inequality, we have

$$\begin{aligned}
\max\{\|x_\mathcal{T}-x_0\|,\|y_\mathcal{T}-x_0\|\} &\ge \frac{1}{2\sqrt{p_{\rm ub}}}\|z_\mathcal{T}\| \ge \frac{1}{2\sqrt{p_{\rm ub}}}(\|p_\mathcal{T}\|-\|q_\mathcal{T}\|) \ge \frac{1}{4\sqrt{p_{\rm ub}}}\|p_\mathcal{T}\| \\
&\ge \frac{(1+\alpha\gamma)^\mathcal{T}\alpha\omega}{4\sqrt{p_{\rm ub}}} \overset{(a)}{\ge} \frac{2^{\iota/2-3}p_{\rm lb}}{\sqrt{p_{\rm ub}}\ell_1}\bar{\omega} \overset{(b)}{>} \mathcal{S}.
\end{aligned}$$

where in (a), we used the inequality $(1+x)^{1/x} \ge 2$ for every $0 < x \le 1$ and in (b), we used the definition of $\bar{\omega}$. The above inequality contradicts with (56) and therefore completes our proof. ■

D.2 Proof of Lemma 29

To prove Lemma 29, first we provide an upper bound on $f(X_t)$.

Lemma 37. *For every iteration X_t of PPrecGD, we have*

$$f(X_t) \le f(X_0) + 2\sqrt{\|X_0\|_F^2+\eta}\cdot\alpha\beta\epsilon$$

Proof Note that $f(X_{t+1}) \leq f(X_t)$, except for when X_t is perturbed with a random perturbation. Moreover, we have already shown that, each perturbation followed by $\mathcal{T}$ iterations of PMGD strictly reduces the objective function. Therefore, $f(X)$ takes its maximum value when it is perturbed at the initial point. This can only happen if X_0 is close to a strict saddle point, i.e.,

$$\|\nabla f(X_0)\|^*_{X_0,\eta} \leq \epsilon, \quad \text{and} \quad \nabla^2 f(X_0) \not\succeq -\sqrt{L_d \epsilon} \cdot \mathbf{P}_{X_0,\eta},$$

Therefore, we have $X_1 = X_0 + \alpha\zeta$, where $\zeta \sim \mathbb{B}(\beta)$. This implies that

$$\begin{aligned} f(X_1) - f(X_0) &\leq \alpha\langle \nabla f(X_0), \zeta\rangle + \frac{\alpha^2 L_1}{2}\|\zeta\|_F^2 \leq \alpha \left\|\mathbf{P}^{1/2}_{X_0,\eta}\right\|_F \|\nabla f(X_0)\|^*_{X_0,\eta} \|\zeta\|_F + \frac{\alpha^2 L_1}{2}\|\zeta\|_F^2 \\ &\overset{(a)}{\leq} \sqrt{\|X_0\|_F^2 + \eta} \cdot \epsilon\alpha\beta + \frac{L_1}{2}\alpha^2\beta^2 \overset{(b)}{\leq} 2\sqrt{\|X_0\|_F^2 + \eta} \cdot \epsilon\alpha\beta \end{aligned}$$

where (a) follows from our assumption $\|\nabla f(X_0)\|^*_{X_0,\eta} \leq \epsilon$, and (b) is due to our choice of α and β. This implies that $f(X_t) \leq f(X_1) \leq f(X_0) + 2\sqrt{\|X_0\|_F^2 + \eta} \cdot \epsilon\alpha\beta$, thereby completing the proof. ■

The above lemma combined with the coercivity of ϕ implies that

$$X_t \in \mathcal{M}\left(\phi(X_0 X_0^T) + 2\sqrt{\|X_0\|_F^2 + \eta} \cdot \alpha\beta\epsilon\right)$$

for every iteration X_t of PPrecGD. Now, we proceed with the proof of Lemma 29.

Proof [Lemma 29.] First, we prove the gradient lipschitzness of $f(X)$. Due to the definition of Γ_F and Lemma 37, every iteration of PPrecGD belongs to the ball $\{M : \|M\|_F \leq \Gamma_F\}$. For every $X, Y \in \{M : \|M\|_F \leq \Gamma_F\}$, we have

$$\begin{aligned} \|\nabla f(X) - \nabla f(Y)\|_F &= 2\left\|\nabla\phi(XX^\top)X - \nabla\phi(YY^\top)Y\right\|_F \\ &\leq \left\|\nabla\phi(XX^\top) - \nabla\phi(YY^\top)\right\|_F \|X\|_F + 2\left\|\phi(YY^\top)\right\|_F \|X - Y\|_F \\ &\leq 2L_1\Gamma_F \left\|XX^\top - YY^\top\right\|_F + 5L_1\Gamma_F^2 \|X - Y\|_F \\ &\leq 2L_1\Gamma_F \left\|X(X-Y)^\top - (Y-X)Y^\top\right\|_F + 5L_1\Gamma_F^2 \|X - Y\|_F \\ &\leq 9L_1\Gamma_F^2 \|X - Y\|_F \end{aligned}$$

which shows that $f(X)$ is $9L_1\Gamma_F^2$-gradient Lipschitz within the ball $\{M : \|M\|_F \leq \Gamma_F\}$. Next, we prove the Hessian lipschitzness of $f(X)$. For any arbitrary V with $\|V\|_F = 1$, we

have

$$
\begin{aligned}
&\left|\left\langle \nabla^2 f(X)[V], V\right\rangle - \left\langle \nabla^2 f(Y)[V], V\right\rangle\right| \\
=&2\left|\left\langle \nabla\phi(XX^\top) - \nabla\phi(YY^\top), VV^\top\right\rangle\right| \\
&+\left|\left\langle \nabla^2\phi(XX^\top), XV^\top + VX^\top\right\rangle - \left\langle \nabla^2\phi(YY^\top), YV^\top + VY^\top\right\rangle\right| \\
\le&2\left\|\nabla\phi(XX^\top) - \nabla\phi(YY^\top)\right\|_F \\
&+\left|\left\langle \nabla^2\phi(XX^\top), XV^\top + VX^\top\right\rangle - \left\langle \nabla^2\phi(YY^\top), XV^\top + VX^\top\right\rangle\right| \\
&+\left|\left\langle \nabla^2\phi(YY^\top), YV^\top + VY^\top\right\rangle - \left\langle \nabla^2\phi(YY^\top), XV^\top + VX^\top\right\rangle\right| \\
\le&2L_1\left\|XX^\top - YY^\top\right\|_F \\
&+\left\|\nabla^2\phi(XX^\top) - \nabla^2\phi(YY^\top)\right\|\left\|XV^\top + VX^\top\right\|_F \\
&+\left\|\nabla^2\phi(YY^\top)\right\|\left\|(Y-X)V^\top + V(Y-X)^\top\right\|_F \\
\le&4L_1\Gamma_F\left\|X-Y\right\|_F + 4L_2\Gamma_F^2\left\|X-Y\right\|_F + 2L_1\left\|X-Y\right\|_F \\
=&((4\Gamma_F+2)L_1 + 4\Gamma_F^2 L_2)\left\|X-Y\right\|_F.
\end{aligned}
$$

Therefore, $f(X)$ is $((4\Gamma_F+2)L_1 + 4\Gamma_F^2 L_2)$-Hessian Lipschitz within the ball $\{M : \|M\|_F \le \Gamma_F\}$. Finally, it is easy to verify that the eigenvalues of $\mathbf{P}_{X,\eta} = (X^TX + \eta I_n)\otimes I_r$ are between η and $\Gamma_2^2 + \eta$ within the ball $\{M : \|M\| \le \Gamma_2\}$. Moreover, $\|\mathbf{P}_{X,\eta} - \mathbf{P}_{Y,\eta}\| \le \left\|X^\top X - Y^\top Y\right\| \le 2\Gamma_2\left\|X-Y\right\|$. This completes the proof of this lemma. ■

References

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

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

Nicolas Boumal. A riemannian low-rank method for optimization over semidefinite matrices with block-diagonal constraints. *arXiv preprint arXiv:1506.00575*, 2015.

Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. *arXiv preprint arXiv:1606.04970*, 2016.

Nicolas Boumal, Vladislav Voroninski, and Afonso S Bandeira. Deterministic guarantees for burer-monteiro factorizations of smooth semidefinite programs. *Communications on Pure and Applied Mathematics*, 73(3):581–608, 2020.

Oliver Bunk, Ana Diaz, Franz Pfeiffer, Christian David, Bernd Schmitt, Dillip K Satapathy, and J Friso Van Der Veen. Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. *Acta Crystallographica Section A: Foundations of Crystallography*, 63(4):306–314, 2007.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2): 329–357, 2003.

Samuel Burer and Renato DC Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

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.

Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

Yuxin Chen and Emmanuel J Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on pure and applied mathematics*, 70(5):822–883, 2017.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37, 2019.

Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Alexander Cloninger, Wojciech Czaja, Ruiliang Bai, and Peter J Basser. Solving 2d fredholm integral from incomplete measurements using compressive sensing. *SIAM journal on imaging sciences*, 7(3):1775–1798, 2014.

John V Corbett. The pauli problem, state reconstruction and quantum-real numbers. *Reports on Mathematical Physics*, 57(1):53–68, 2006.

Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

Jin-yan Fan and Ya-xiang Yuan. On the quadratic convergence of the levenberg-marquardt method without nonsingularity assumption. *Computing*, 74(1):23–39, 2005.

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.

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

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.

Anne Greenbaum. *Iterative methods for solving linear systems*. SIAM, 1997.

David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

Robert W Harrison. Phase problem in crystallography. *JOSA a*, 10(5):1046–1055, 1993.

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.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

Michel Journée, Francis Bach, P-A Absil, and Rodolphe Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5): 2327–2351, 2010.

Shmuel Kaniel. Estimates for some computational techniques in linear algebra. *Mathematics of Computation*, 20(95):369–378, 1966.

Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805, 2013.

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.

Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

Jianhao Ma and Salar Fattahi. Blessing of nonconvexity in deep linear models: Depth flattens the optimization landscape around the true solution. *arXiv preprint arXiv:2207.07612*, 2022a.

Jianhao Ma and Salar Fattahi. Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *arXiv preprint arXiv:2202.08788*, 2022b.

Rick P Millane. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Christopher Conway Paige. *The computation of eigenvalues and eigenvectors of very large sparse matrices.* PhD thesis, University of London, 1971.

Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74. PMLR, 2017.

Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably. *SIAM Journal on Imaging Sciences*, 11(4):2165–2204, 2018.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

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.

Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719, 2005.

David M Rosen, Luca Carlone, Afonso S Bandeira, and John J Leonard. Se-sync: A certifiably correct algorithm for synchronization over the special euclidean group. *The International Journal of Robotics Research*, 38(2-3):95–125, 2019.

David M Rosen, Luca Carlone, Afonso S Bandeira, and John J Leonard. A certifiably correct algorithm for synchronization over the special euclidean group. In *Algorithmic Foundations of Robotics XII*, pages 64–79. Springer, 2020.

Yousef Saad. On the rates of convergence of the lanczos and the block-lanczos methods. *SIAM Journal on Numerical Analysis*, 17(5):687–706, 1980.

Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

Amit Singer. A remark on global positioning from local distances. *Proceedings of the National Academy of Sciences*, 105(28):9507–9511, 2008.

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.

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

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

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

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.

Irene Waldspurger and Alden Waters. Rank optimality for the burer–monteiro factorization. *SIAM journal on Optimization*, 30(3):2577–2602, 2020.

Nobuo Yamashita and Masao Fukushima. On the rate of convergence of the levenberg-marquardt method. In *Topics in numerical analysis*, pages 239–249. Springer, 2001.

Jialun Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for over-parameterized nonconvex matrix factorization. *Advances in Neural Information Processing Systems*, 34:5985–5996, 2021.

Richard Zhang, Cedric Josz, Somayeh Sojoudi, and Javad Lavaei. How much restricted isometry is needed in nonconvex matrix recovery? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a. URL `https://proceedings.neurips.cc/paper/2018/file/f8da71e562ff44a2bc7edf3578c593da-Paper.pdf`.

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

Richard Y Zhang. Improved global guarantees for the nonconvex burer–monteiro factorization via rank overparameterization. *arXiv preprint arXiv:2207.01789*, 2022.

Richard Y Zhang, Cédric Josz, Somayeh Sojoudi, and Javad Lavaei. How much restricted isometry is needed in nonconvex matrix recovery? *arXiv preprint arXiv:1805.10251*, 2018b.

Richard Y Zhang, Javad Lavaei, and Ross Baldick. Spurious local minima in power system state estimation. *IEEE transactions on control of network systems*, 6(3):1086–1096, 2019a.

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, 2019b.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015a. URL `https://proceedings.neurips.cc/paper/2015/file/32bb90e8976aab5298d5da10fe66f21d-Paper.pdf`.

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*, 2015b.

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.