JMLR: Workshop and Conference Proceedings vol 49:1–53, 2016

Dropping Convexity for Faster Semi-definite Optimization

Srinadh Bhojanapalli SRINADH@TTIC.EDU
Toyota Technological Institute at Chicago

Anastasios Kyrillidis ANASTASIOS@UTEXAS.EDU

Sujay Sanghavi SANGHAVI@MAIL.UTEXAS.EDU
University of Texas at Austin

Abstract

We study the minimization of a convex function $f(X)$ over the set of $n \times n$ positive semi-definite matrices, but when the problem is recast as $\min_U g(U) := f(UU^\top)$, with $U \in \mathbb{R}^{n\times r}$ and $r \leq n$. We study the performance of gradient descent on g—which we refer to as Factored Gradient Descent (FGD)—under standard assumptions on the *original* function f.

We provide a rule for selecting the step size and, with this choice, show that the *local* convergence rate of FGD mirrors that of standard gradient descent on the original f: *i.e.*, after k steps, the error is $O(1/k)$ for smooth f, and exponentially small in k when f is (restricted) strongly convex. In addition, we provide a procedure to initialize FGD for (restricted) strongly convex objectives and when one only has access to f via a first-order oracle; for several problem instances, such proper initialization leads to *global* convergence guarantees.

FGD and similar procedures are widely used in practice for problems that can be posed as matrix factorization. To the best of our knowledge, this is the first paper to provide precise convergence rate guarantees for general convex functions under standard convex assumptions.

Keywords: Non-convex analysis and optimization, semi-definite matrix, rank minimization

1. Introduction

Consider the following standard convex semi-definite optimization problem:

$$\underset{X\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad f(X) \quad \text{subject to} \quad X \succeq 0, \tag{1}$$

where $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a convex and differentiable function, and $X \succeq 0$ denotes the convex set over positive semi-definite matrices in $\mathbb{R}^{n\times n}$. Let $X^\star$ be an optimum of (1) with $\text{rank}(X^\star) = r^\star \leq n$. This problem can be remodeled as a non-convex problem, by writing $X = UU^\top$ where U is an $n\times r$ matrix. Specifically, define $g(U) := f(UU^\top)$ and[1] consider direct optimization of the transformed problem, *i.e.*,

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad g(U) \quad \text{where } r \leq n. \tag{2}$$

Problems (1) and (2) will have the same optimum when $r = r^\star$. However, the recast problem is *unconstrained* and leads to computational gains in practice: *e.g.*, iterative update schemes, like gradient descent, do not need to do eigen-decompositions to satisfy semi-definite constraints at every iteration.

1. While g is a non-convex function, we note that it is a very specific kind of non-convexity, arising "only" due to the recasting of an originally convex function.

In this paper, we also consider the case of $r < r^\star$, which often occurs in applications. The reasons of such a choice chould be three-fold: (i) it might model better an underlying task (*e.g.*, f may have arisen from a relaxation of a rank constraint in the first place), (ii) it leads to computational gains, since smaller r means fewer variables to maintain and optimize, (iii) it leads to statistical "gains", as it might prevent over-fitting in machine learning or inference problems.

Such recasting of matrix optimization problems is empirically widely popular, especially as the size of problem instances increases. Some applications in modern machine learning includes matrix completion Candès and Recht (2009); Jain et al. (2013); Kyrillidis and Cevher (2014); Chen et al. (2014), affine rank minimization Recht et al. (2010); Jain et al. (2010); Becker et al. (2013), covariance / inverse covariance selection Hsieh et al. (2011); Kyrillidis et al. (2014), phase retrieval Netrapalli et al. (2013); Candes et al. (2015a); White et al. (2015); Sun et al. (2016), Euclidean distance matrix completion Mishra et al. (2011), finding the square root of a PSD matrix Jain et al. (2015), and sparse PCA d'Aspremont et al. (2007), just to name a few. Typically, one can solve (2) via simple, first-order methods on U like gradient descent. Unfortunately, such procedures have no guarantees on convergence to the optima of the original f, or on the rate thereof. Our goal in this paper is to provide such analytical guarantees, by using—simply and transparently—standard convexity properties of the original f.

Overview of our results. In this paper, we prove that updating U via gradient descent in (2) converges (fast) to optimal (or near-optimal) solutions. While there are some recent and very interesting works that consider using such non-convex parametrization Jain et al. (2013); Netrapalli et al. (2013); Tu et al. (2015); Zheng and Lafferty (2015); Sun and Luo (2014); Zhao et al. (2015), their results only apply to specific examples. To the best of our knowledge, this is the first paper that solves the re-parametrized problem with attractive convergence rate guarantees for *general convex functions* f and under common convex assumptions. Moreover, we achieve the above by assuming the *first order oracle* model: for any matrix X, we can only obtain the value $f(X)$ and the gradient $\nabla f(X)$.

To achieve the desiderata, we study how gradient descent over U performs in solving (2). This leads to the *factored gradient descent* (FGD) algorithm, which applies the simple update rule

$$U^+ = U - \eta \nabla f(UU^T) \cdot U. \tag{3}$$

We provide a set of sufficient conditions to guarantee convergence. We show that given a suitable initialization point, FGD converges to a solution close to the optimal point in sublinear or linear rate, depending on the nature of f.

Our contributions in this work can be summarized as follows:

(i) *New step size rule and* FGD. Our main algorithmic contribution is a special choice of the step size η. Our analysis showcase that η needs to depend not only on the convexity parameters of f (as is the case in standard convex optimization) but also on the top singular value of the unknown optimum. Section 3 describes the precise step size rule, and also the intuition behind it. Of course, the optimum is not known a priori. As a solution in practice, we show that choosing η based on a point that is constant relative distance from the optimum also provably works.

(ii) *Convergence of* FGD *under common convex assumptions.* We consider two cases: (i) when f is just a M-smooth convex function, and (ii) when f satisfies also *restricted strong convexity (RSC)*, *i.e.*, f satisfies strong-convexity-like conditions, but only over low rank matrices; see next section

for definitions. Both cases are based on now-standard notions, common for the analysis of convex optimization algorithms. Given a good initial point, we show that, when f is M-smooth, FGD converges sublinearly to an optimal point $X^\star$. For the case where f has RSC, FGD converges linearly to the unique $X^\star$, matching analogous result for classic gradient descent schemes, under smoothness and strong convexity assumptions.

Furthermore, for the case of smooth and strongly convex f, our analysis extends to the case $r < r^\star$, where FGD converges to a point close to the best rank-r approximation of $X^\star$.[2]

Both results hold when FGD is initialized at a point with constant relative distance from optimum. Interestingly, the linear convergence rate factor depends not only on the convexity parameters of f, but also on the spectral characteristics of the optimum; a phenomenon borne out in our experiments. Section 4 formally states these results.

(iii) *Initialization:* For specific problem settings, various initialization schemes are possible (see Jain et al. (2013); Netrapalli et al. (2013); Chen and Wainwright (2015)). In this paper, we extend such results to the case where we only have access to f via the first-order oracle: specifically, we initialize based on the gradient at zero, *i.e.*, $\nabla f(0)$. We show that, for certain condition numbers of f, this yields a constant relative error initialization (Section 5). Moreover, Section 5 lists alternative procedures that lead to good initialization points and comply with our theory.

Roadmap. The rest of the paper is organized as follows. Section 2 contains basic notation and standard convex definitions. Section 3 presents the FGD algorithm and the step size η used, along with some intuition for its selection. Section 4 contains the convergence guarantees of FGD; the main supporting lemmas and proofs of the main theorems are provided in Section 6. In Section 5, we discuss some initialization procedures that guarantee a "decent" starting point for FGD. This paper concludes with discussion on related work (Section 7).

2. Preliminaries

Notation. For matrices $X, Y \in \mathbb{R}^{n\times n}$, their inner product is $\langle X, Y\rangle = \mathrm{Tr}\left(X^\top Y\right)$. Also, $X \succeq 0$ denotes X is a positive semi-definite (PSD) matrix, while the convex set of PSD matrices is denoted $\mathbb{S}^n_+$. We use $\|X\|_F$ and $\|X\|_2$ for the Frobenius and spectral norms of a matrix, respectively. Given a matrix X, we use $\sigma_{\min}(X)$ and $\sigma_{\max}(X)$ to denote the smallest and largest *strictly positive* singular values of X and define $\tau(X) = \frac{\sigma_{\max}(X)}{\sigma_{\min}(X)}$; with a slight abuse of notation, we also use $\sigma_1(X) \equiv \sigma_{\max}(X) \equiv \|X\|_2$. X_r denotes the rank-r approximation of X via its truncated singular value decomposition. Let $\tau(X^\star_r) = \frac{\sigma_1(X^\star)}{\sigma_r(X^\star)}$ denote the condition number of $X^\star_r$; again, observe $\sigma_r(X_r) \equiv \sigma_{\min}(X_r)$. Q_A denotes the basis of the column space of matrix A. $\texttt{srank}(X) := \|X\|_F^2/\|X\|_2^2$ represents the stable rank of matrix X. We use $e_i \in \mathbb{R}^n$ to denote the standard basis vector with 1 at the i-th position and zeros elsewhere.

Without loss of generality, f is a symmetric convex function, *i.e.*, $f(X) = f(X^\top)$. Let $\nabla f(X)$ denote the gradient matrix, *i.e.*, its $(i,j)^{th}$ element is $[\nabla f(X)]_{ij} = \frac{\partial f(X)}{\partial x_{ij}}$. For $X = UU^\top$, the

2. In this case, we require $\|X^\star - X^\star_r\|_F$ to be small enough, such that the rank-constrained optimum be close to the best rank-r approximation of $X^\star$. This assumption naturally applies in applications, where *e.g.*, $X^\star$ is a superposition of a low rank latent matrix, plus a small perturbation term Javanmard and Montanari (2013); Yu et al. (2014). In Section H, we show how this assumption can be dropped by using a different step size η, where spectral norm computation of two $n \times r$ matrices is required per iteration.

gradient of f with respect to U is $\left(\nabla f(UU^\top) + \nabla f(UU^\top)^\top\right) U = 2\nabla f(X) \cdot U$, due to symmetry of f. Finally, let $X^\star$ be the optimum of $f(X)$ over $\mathbb{S}^n_+$ with factorization $X^\star = U^\star (U^\star)^T$.

For any general symmetric matrix X, let the matrix $\mathcal{P}_+(X)$ be its projection onto the set of PSD matrices. This can be done by finding all the strictly positive eigenvalues and corresponding eigenvectors $(\lambda_i, v_i : \lambda_i > 0)$ and then forming $\mathcal{P}_+(X) = \sum_{i:\lambda_i>0} \lambda_i v_i v_i^\top$.

In algorithmic descriptions, U and U^+ denote the putative solution of current and next iteration, respectively. An important issue in optimizing f over the U space is the existence of non-unique possible factorizations $UU^\top$ for any feasible point X. To see this, given factorization $X = UU^\top$ where $U \in \mathbb{R}^{n\times r}$, one can define an class of *equivalent* factorizations $UR^\top RU^\top = UU^\top$, where R belongs to the set $\{R \in \mathbb{R}^{r\times r} \ : \ R^\top R = I\}$ of orthonormal matrices. So we use a distance metric that is invariant to R in the factored space that is equivalent to distance in the matrix X space, which is defined below.

Definition 1 *Let matrices $U, V \in \mathbb{R}^{n\times r}$. Define:*

$$\textsc{Dist}\,(U, V) := \min_{R:R\in\mathcal{O}} \|U - VR\|_F\,.$$

$\mathcal{O}$ is the set of $r \times r$ orthonormal matrices R, such that $R^\top R = I_{r\times r}$. The optimal R satisfies $PQ^\top$ where $P\Sigma Q^\top$ is the singular value decomposition of $V^\top U$.

Assumptions. We will investigate the performance of non-convex gradient descent for functions f that satisfy standard smoothness conditions only, as well as the case where f further is (restricted) strongly convex. We state these standard definitions below.

Definition 2 *Let $f : \mathbb{S}^n_+ \to \mathbb{R}$ be convex and differentiable. Then, f is m-strongly convex if:*

$$f(Y) \geq f(X) + \langle \nabla f(X), Y - X\rangle + \tfrac{m}{2}\|Y - X\|_F^2\,, \quad \forall X, Y \in \mathbb{S}^n_+. \tag{4}$$

Definition 3 *Let $f : \mathbb{S}^n_+ \to \mathbb{R}$ be a convex differentiable function. Then, f is M-smooth if:*

$$\|\nabla f(X) - \nabla f(Y)\|_F \leq M \cdot \|X - Y\|_F\,, \quad X, Y \in \mathbb{S}^n_+. \tag{5}$$

This further implies the following upper bound:

$$f(Y) \leq f(X) + \langle \nabla f(X), Y - X\rangle + \tfrac{M}{2}\|Y - X\|_F^2\,. \tag{6}$$

Given the above definitions, we define $\kappa = \frac{M}{m}$ as the condition number of function f.

Finally, in high dimensional settings, often loss function f does not satisfy strong convexity globally, but only on a restricted set of directions; see Negahban and Wainwright (2012); Agarwal et al. (2010) and Section F for a more detailed discussion.

Definition 4 *A convex function f is (m, r)-restricted strongly convex if:*

$$f(Y) \geq f(X) + \langle \nabla f(X), Y - X\rangle + \tfrac{m}{2}\|Y - X\|_F^2\,, \quad \textit{for any rank-}r\textit{ matrices } X, Y \in \mathbb{S}^n_+. \tag{7}$$

3. Factored gradient descent

We solve the non-convex problem (2) via *Factored Gradient Descent* (FGD) with update rule[3]:

$$U^+ = U - \eta \nabla f(UU^\top) \cdot U.$$

FGD does this, but with two key innovations: a careful initialization and a special step size η. The discussion on the initialization is deferred until Section 5.

Step size η. Even though f is a convex function over $X \succeq 0$, the fact that we operate with the non-convex $UU^\top$ parametrization means that we need to be careful about the step size η; *e.g.*, our *constant* η selection should be such that, when we are close to $X^\star$, we do not "overshoot" the optimum $X^\star$.

In this work, we pick the step size parameter, according to the following closed-form[4]:

$$\eta = \frac{1}{16(M \left\|X^0\right\|_2 + \left\|\nabla f(X^0)\right\|_2)}. \tag{8}$$

Recall that, if we were just doing standard gradient descent on f, we would choose a step size of $1/M$, where M is a uniform upper bound on the largest eigenvalue of the Hessian $\nabla^2 f(\cdot)$.

To motivate our step size selection, let us consider a simple setting where $U \in \mathbb{R}^{n\times r}$ with $r = 1$; *i.e.*, U is a vector. For clarity, denote it as u. Let f be a separable function with $f(X) = \sum_{ij} f_{ij}(X_{ij})$. Furthermore, define the function $g : \mathbb{R}^n \to \mathbb{R}$ such that $f(uu^\top) \equiv g(u)$. It is easy to compute (see Lemma 27):

Algorithm 1 Factored gradient descent (FGD)

input Function f, target rank r, # iterations K.

1: Compute X^0 as in (12).
2: Set $U \in \mathbb{R}^{n\times r}$ such that $X^0 = UU^\top$.
3: Set step size η as in (8).
4: **for** $k = 0$ to $K - 1$ **do**
5: $\quad U^+ = U - \eta \nabla f(UU^T) \cdot U.$
6: $\quad U = U^+.$
7: **end for**

output $X = UU^\top$.

$$\nabla g(u) = \nabla f(uu^\top) \cdot u \in \mathbb{R}^n \quad \text{and} \quad \nabla^2 g(u) = \texttt{mat}\left(\texttt{diag}(\nabla^2 f(uu^\top)) \cdot \texttt{vec}\left(uu^\top\right)\right) + \nabla f(uu^\top) \in \mathbb{R}^{n\times n},$$

where $\texttt{mat} : \mathbb{R}^{n^2} \to \mathbb{R}^{n\times n}$, $\texttt{vec} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n^2}$ and, $\texttt{diag} : \mathbb{R}^{n^2\times n^2} \to \mathbb{R}^{n^2\times n^2}$ are the matricization, vectorization and diagonalization operations, respectively; for the last case, $\texttt{diag}$ generates a diagonal matrix from the input, discarding its off-diagonal elements. We remind that $\nabla f(uu^\top) \in \mathbb{R}^{n\times n}$ and $\nabla^2 f(uu^\top) \in \mathbb{R}^{n^2\times n^2}$. Note also that $\nabla^2 f(X)$ is diagonal for separable f.

Standard convex optimization suggests that η should be chosen such that $\eta < 1/\|\nabla^2 g(\cdot)\|_2$. The above suggest the following step size selection rule for M-smooth f:

$$\eta < \frac{1}{\|\nabla^2 g(\cdot)\|_2} \propto \frac{1}{M\|X\|_2 + \|\nabla f(X)\|_2}.$$

In stark contrast with classic convex optimization where $\eta \propto \frac{1}{M}$, the step size selection further depends on the spectral information of the current iterate and the gradient. Since computing $\|X\|_2, \|\nabla f(X)\|_2$ per iteration could be computational inefficient, we use the spectral norm of X^0 and its gradient $\nabla f(X^0)$ as surrogate, where X^0 is the initialization point[5].

3. The true gradient of f with respect to U is $2\nabla f(UU^\top) \cdot U$. However, for simplicity and clarity of exposition, in our algorithm and its theoretical guarantees, we absorb the 2-factor in the step size η.
4. Constant 16 in the expression (8) appears due to our analysis, where we do not optimize over the constants. One can use another constant in order to be more aggressive; nevertheless, we observed that our setting works well in practice.
5. However, as we show in Section H, one could compute $\|X\|_2$ and $\|\nabla f(X)\|_2$ per iteration in order to relax some of the requirements of our approach.

To clarify η selection further, we next describe a toy example, in order to illustrate the necessity of such a scaling of the step size. Consider the following minimization problem.

$$\underset{u\in\mathbb{R}^{n\times 1}}{\text{minimize}} \quad f(uu^\top) := ||uu^\top - Y||_F^2,$$

where $u \equiv U \in \mathbb{R}^{n\times 1}$—and thus, $X = uu^\top$, *i.e.*, we are interested in rank-1 solutions—and Y is a given rank-2 matrix such that $Y = \alpha^2 v_1 v_1^\top - \beta^2 v_2 v_2^\top$, for $\alpha > \beta \in \mathbb{R}$ and v_1, v_2 orthonormal vectors. Observe that f is a strongly convex function with rank-1 minimizer $X^\star = \alpha^2 v_1 v_1^\top$; let $u^\star = \alpha v_1$. It is easy to verify that (i) $\|X^\star\|_2 = \alpha^2$, (ii) $\|\nabla f(X^\star)\|_2 = \|2\cdot(X^\star - Y)\|_2 = 2\beta^2$, and (iii) $\|\nabla f(X_1) - \nabla f(X_2)\|_F \le M\cdot\|X_1 - X_2\|_F$, where $M = 2$.

Consider the case where $u = \frac{\alpha}{2}v_1 + \frac{\beta}{10}v_2$ is the current estimate. Then, the gradient of f at u is evaluated as:

$$\nabla f(uu^\top)\cdot u = 2\left(-\tfrac{3\alpha^2}{8}v_1v_1^\top + \tfrac{101\beta^2}{10^3}v_2v_2^\top\right)\cdot\left(\tfrac{\alpha}{2}v_1 + \tfrac{\beta}{10}v_2\right) = -\tfrac{3\alpha^3}{4}v_1 + \tfrac{101\beta^3}{500}v_2.$$

Hence, according to the update rule $u^+ = u - 2\eta\nabla f(uu^\top)\cdot u$, the next iterate satisfies:

$$u^+ = u - 2\eta\left(-\tfrac{3\alpha^3}{4}v_1 + \tfrac{101\beta^3}{500}v_2\right) = \left(\tfrac{\alpha}{2} + \eta\tfrac{3\alpha^3}{2}\right)v_1 + \left(\tfrac{\beta}{10} + \eta\tfrac{202\beta^3}{500}\right)v_2.$$

Observe that coefficients of both $v_1,\ v_2$ in u^+ include $O(\alpha^3)$ and $O(\beta^3)$ quantities.

The quality of u^+ clearly depends on how η is chosen. In the case $\eta = \frac{1}{M} = \frac{1}{2}$, such step size can result in divergence/"overshooting", as $\|X^\star\|_2 = O(\alpha^2)$ and $\|\nabla f(X^\star)\|_2 = O(\beta^2)$ can be arbitrarily large (independent of M). Therefore, it could be the case that $\textsc{Dist}(u^+, u^\star) > \textsc{Dist}(u, u^\star)$.

In contrast, consider the step size[6] $\eta = \frac{1}{16(M\|X^\star\|_2 + \|\nabla f(X^\star)\|_2)} \propto \frac{1}{C(\alpha^2+\beta^2)}$. Then, with appropriate scaling C, we observe that η lessens the effect of $O(\alpha^3)$ and $O(\beta^3)$ terms in v_1 and v_2 terms, that lead to overshooting for the case $\eta = \frac{1}{2}$. This most possibly result in $\textsc{Dist}(u^+, u^\star) \le \textsc{Dist}(u, u^\star)$.

Computational complexity. The per iteration complexity of FGD is dominated by the gradient computation. This computation is required in any first order algorithm and the complexity of this operation depends on the function f. Apart from $\nabla f(X)$, the additional computation required in FGD is matrix-matrix additions and multiplications, with time complexity upper bounded by $\texttt{nnz}(\nabla f(\cdot))\cdot r$, where $\texttt{nnz}(\nabla f(\cdot))$ denotes the number of non zeros in the gradient at the current point.[7] Hence, the per iteration complexity of FGD is much lower than traditional convex methods like projected gradient descent Nesterov (2004) or classic interior point methods Nesterov and Nemirovski (1988, 1989), as they often require a full eigenvalue decomposition per step.

Note that, for $r = O(n)$, FGD and projected gradient descent have same per iteration complexity of $O(n^3)$. However, FGD performs only a single matrix-matrix multiplication operation, which is much "cheaper" than a SVD calculation. Moreover, matrix multiplication is an easier-to-parallelize operation, as opposed to eigen decomposition operation which is inherently sequential. We notice this behavior in practice; see Sections F and G for applications in matrix sensing.

6. For illustration purposes, we consider a step size that depends on the unknown $X^\star$; in practice, our step size selection is a surrogate of this choice and our results automatically carry over, with appropriate scaling.
7. It could also occur that gradient $\nabla f(X)$ is low-rank, or low-rank + sparse, depending on the problem at hand; it could also happen that the structure of $\nabla f(X)$ leads to "cheap" matrix-vector calculations, when applied to vectors. Here, we state a more generic –and maybe pessimistic– scenario where $\nabla f(X)$ is unstructured.

4. Local convergence of FGD

In this section, we present our main theoretical results on the performance of FGD. We present convergence rates for the settings where (i) f is a M-smooth convex function, and (ii) f is a M-smooth *and* (m, r)-restricted strongly convex function. These assumptions are now standard in convex optimization. Note that, since the $UU^\top$ factorization makes the problem non-convex, it is hard to guarantee convergence of gradient descent schemes in general, without any additional assumptions.

We now state the main assumptions required by FGD for convergence:

FGD ASSUMPTIONS

- *Initialization:* We assume that FGD is initialized with a "good" starting point $X^0 = U^0(U^0)^\top$ that has constant relative error to $X_r^\star = U_r^\star(U_r^\star)^\top$.[8] In particular, we assume

$$(A1) \qquad \text{DIST}(U^0, U_r^\star) \leq \rho\sigma_r(U_r^\star) \quad \text{for } \rho := \tfrac{1}{100}\tfrac{\sigma_r(X^\star)}{\sigma_1(X^\star)} \qquad \text{(Smooth } f\text{)}$$

$$(A2) \qquad \text{DIST}(U^0, U_r^\star) \leq \rho'\sigma_r(U_r^\star) \quad \text{for } \rho' := \tfrac{1}{100\kappa}\tfrac{\sigma_r(X^\star)}{\sigma_1(X^\star)} \qquad \text{(Strongly convex } f\text{)},$$

  for the smooth and restricted strongly convex setting, respectively. This assumption helps in avoiding saddle points, introduced by the U parametrization[9].

  In many applications, an initial point U^0 with this type of guarantees is easy to obtain, often with just one eigenvalue decomposition; we refer the reader to the works Jain et al. (2013); Netrapalli et al. (2013); Chen and Wainwright (2015); Zheng and Lafferty (2015); Tu et al. (2015) for specific initialization procedures for different problem settings. See also Section 5 for a more detailed discussion. Note that the problem is still non-trivial after the initialization, as this only gives a constant error approximation.

- *Approximate rank-r optimum:* In many learning applications, such as localization Javanmard and Montanari (2013) and multilabel learning Yu et al. (2014), the true $X^\star$ emerges as the superposition of a low rank latent matrix plus a small perturbation term, such that $\|X^\star - X_r^\star\|_F$ is small. While, in practice, it might be the case $\text{rank}(X^\star) = n$—due to the presence of noise—often we are more interested in revealing the latent low-rank part. As already mentioned, we might as well set $r < \text{rank}(X^\star)$ for computational or statistical reasons. In all these cases, further assumptions w.r.t. the quality of approximation have to be made. In particular, let $X^\star$ be the optimum of (1) and f is M-smooth and (m, r)-strongly convex. In our analysis, we assume:

$$(A3) \qquad \|X^\star - X_r^\star\|_F \leq \tfrac{1}{200\kappa^{1.5}}\tfrac{\sigma_r(X^\star)}{\sigma_1(X^\star)}\sigma_r(X^\star) \qquad \text{(Strongly convex } f\text{)},$$

8. If $r = r^\star$, then one can drop the subscript. For completeness and in order to accommodate the approximate rank-r case, described below, we will keep the subscript in our discussion.
9. To illustrate this consider the following example,

$$\underset{U \in \mathbb{R}^{n\times r}}{\text{minimize}} \quad f(UU^\top) := ||UU^\top - U_r^\star(U_r^\star)^\top||_F^2.$$

Now it is easy to see that $\text{DIST}(U_{r-1}^\star, U_r^\star) = \sigma_r(U_r^\star)$ and $U_{r-1}^\star$ is a stationary point of the function considered $\big(\nabla f(U_{r-1}^\star(U_{r-1}^\star)^\top) \cdot U_{r-1}^\star = 0\big)$. We need the initial error to be further smaller than $\sigma_r(U^\star)$ by a factor of condition number of $X_r^\star$.

This assumption intuitively requires the noise magnitude to be smaller than the optimum and constrains the rank constrained optimum to be closer to $X_r^\star$.[10]

We note that, in the results presented below, we have not attempted to optimize over the constants appearing in the assumptions and any intermediate steps of our analysis. Finding such tight constants could strengthen our arguments for fast convergence; however, it does not change our claims for sublinear or linear convergence rates. Moreover, we consider the case $r \leq \text{rank}(X^\star)$; we believe the analysis can be extended to the setting $r > \text{rank}(X^\star)$ and leave it for future work.[11]

4.1. $1/k$ convergence rate for smooth f

Next, we state our first main result under smoothness condition, as in Definition 3. In particular, we prove that FGD makes progress per iteration with sublinear rate. Here, we assume only the case where $r = r^\star$; for consistency reasons, we denote $X^\star = X_r^\star$. Key lemmas and their proofs for this case are provided in Section C.

Theorem 5 (Convergence performance for smooth f) *Let $X_r^\star = U_r^\star U_r^{\star\top}$ denote an optimum of M-smooth f over the PSD cone. Let $f(X^0) > f(X_r^\star)$. Then, under assumption $(A1)$, after k iterations, the* FGD *algorithm finds solution X^k such that*

$$f(X^k) - f(X_r^\star) \leq \frac{\frac{5}{\eta} \cdot \text{DIST}(U^0, U_r^\star)^2}{k + \frac{5}{\eta} \cdot \frac{\text{DIST}(U^0, U_r^\star)^2}{f(X^0) - f(X_r^\star)}}. \tag{9}$$

The theorem states that provided (i) we choose the step size η, based on a starting point that has constant relative distance to $U_r^\star$, and (ii) we start from such a point, gradient descent on U will converge sublinearly to a point $X_r^\star$. In other words, Theorem 5 shows that FGD computes a sequence of estimates in the U-factor space such that the function values decrease with $O\left(\frac{1}{k}\right)$ rate, towards a global minimum of f function. Recall that, even in the standard convex setting, classic gradient descent schemes over X achieve the same $O\left(\frac{1}{k}\right)$ convergence rate for smooth convex functions Nesterov (2004). Hence, FGD matches the rate of convex gradient descent, under the assumptions of Theorem 5.

4.2. Linear convergence rate under strong convexity assumption

Here, we show that, with the additional assumption that f satisfies the (m, r)-restricted strong convexity over $\mathbb{S}_+^n$, FGD achieves linear convergence rate. The proof is provided in Section B.

Theorem 6 (Convergence rate for restricted strongly convex f) *Let the current iterate be U and $X = UU^\top$. Assume $\text{DIST}(U, U_r^\star) \leq \rho' \sigma_r(U_r^\star)$ and let the step size be $\eta = \frac{1}{16\,(M\|X^0\|_2 + \|\nabla f(X^0)\|_2)}$. Then under assumptions $(A2), (A3)$, the new estimate $U^+ = U - \eta \nabla f(X) \cdot U$ satisfies*

$$\text{DIST}(U^+, U_r^\star)^2 \leq \alpha \cdot \text{DIST}(U, U_r^\star)^2 + \beta \cdot \|X^\star - X_r^\star\|_F^2, \tag{10}$$

where $\alpha = 1 - \frac{m\sigma_r(X^\star)}{64(M\|X^\star\|_2 + \|\nabla f(X^\star)\|_2)}$ and $\beta = \frac{M}{28(M\|X^\star\|_2 + \|\nabla f(X^\star)\|_2)}$. Furthermore, U^+ satisfies $\text{DIST}(U^+, U_r^\star) \leq \rho' \sigma_r(U_r^\star)$.

10. Note that the assumption $(A3)$ can be dropped by using a different step size η (see Theorem 32 in Section H). However, this requires two additional spectral norm computations per iteration.
11. Experimental results on synthetic matrix sensing settings have shown that, if we overshoot r, *i.e.*, $r > \text{rank}(X^\star)$, FGD still performs well, finding an ε-accurate solution with linear rate.

The theorem states that provided (i) we choose the step size based on a point that has constant relative distance to $U_r^\star$, and (ii) we start from such a point, gradient descent on U will converge linearly to a neighborhood of $U_r^\star$. The above theorem immediately implies linear convergence rate for the setting where f is standard strongly convex, with parameter m. This follows by observing that standard strong convexity implies restricted strong convexity for all values of rank r.

Last, we present results for the special case where $r = r^\star$; in this case, FGD finds an optimal point $U_r^\star$ with linear rate, within the equivalent class of orthonormal matrices in $\mathcal{O}$.

Corollary 7 (Exact recovery of $X^\star$) *Let $X^\star$ be the optimal point of f, over the set of PSD matrices, such that $rank(X^\star) = r$. Consider X as in Theorem 6. Then, under the same assumptions and with the same convergence factor α as in Theorem 6, we have*

$$\textsc{Dist}(U^+, U^\star)^2 \leq \alpha \cdot \textsc{Dist}(U, U^\star)^2.$$

Further, for $r = n$ we recover the exact case of semi-definite optimization. In plain words, the above corollary suggests that, given an accuracy parameter ε, FGD requires $K = O\left(\log\left(1/\varepsilon\right)\right)$ iterations in order to achieve $\textsc{Dist}(U^K, U^\star)^2 \leq \varepsilon$; recall the analogous result for classic gradient schemes for M-smooth and strongly convex functions f, where similar rates can be achieved in X space Nesterov (2004). The above are abstractly illustrated in Figure 1.

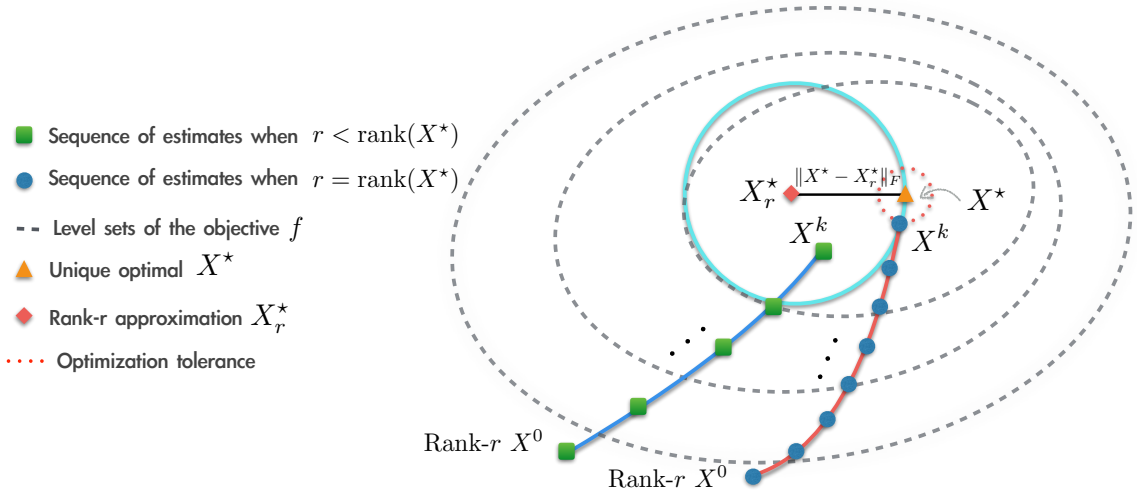

Figure 1: Abstract illustration of Theorem 6 and Corollary 7. The two curves denote the two cases: (i) $r = \text{rank}(X^\star)$ and, (ii) $r < \text{rank}(X^\star)$. (i) In the first case, the triangle marker denotes the unique optimum $X^\star$ and the dashed red circle denotes the *optimization tolerance/error*. (ii) In the case where $r < \text{rank}(X^\star)$, let the cyan circle with radius $c\|X^\star - X_r^\star\|_F$ (set $c = 1$ for simplicity) denote a *neighborhood* around $X^\star$. In this case, FGD converges to a rank-r approximation in the vicinity of $X^\star$ in sublinear rate, according to Theorem 6.

Remark 8 *By the results above, one can easily observe that the convergence rate factor α, in contrast to standard convex gradient descent results, depends both on the condition number of $X_r^\star$ and $\|\nabla f(X^\star)\|_2$, in addition to κ. This dependence is a result of the step size selection, which is different from standard step sizes, i.e., $1/M$ for standard gradient descent schemes. We also refer the reader to Section E for some discussion.*

As a ramification of the above, notice that α depends only on the condition number of $X_r^\star$ and not that of $X^\star$. This suggests that, in settings where the optimum $X^\star$ has bad condition number (and thus leads to slower convergence), it is indeed beneficial to restrict U to be a $n \times r$ matrix and only search for a rank-r approximation of the optimal solution, which leads to faster convergence rate in practice; see Figure 7 in our experimental findings at the end of Section F.3.

Remark 9 *In the setting where the optimum $X^\star$ is 0, directly applying the above theorems requires an initialization that is exactly at the optimum 0. On the contrary, this is actually an easy setting and the* FGD *converges from any initial point to the optimum.*

5. Initialization

In the previous section, we show that gradient descent over U achieves sublinear/linear convergence, once the iterates are closer to $U_r^\star$. Since the overall problem is non-convex, intuition suggests that we need to start from a "decent" initial point, in order to get provable convergence to $U_r^\star$. In the discussion that follows, we focus only on the case of M-smooth and (m, r)-restricted strongly convex case; a procedure that returns an initial point with non-trivial guarantees, for the case of just M-smooth objectives f, remains an open problem.

One way to satisfy this condition for general convex f is to use one of the standard convex algorithms and obtain U within constant error to $U^\star$ (or $U_r^\star$); then, switch to FGD to get the high precision solution. See Tu et al. (2015) for a specific implementation of this idea on matrix sensing. Such initialization procedure comes with the following guarantees; the proof can be found in Section D:

Lemma 10 *Let f be a M-smooth and (m, r)-restricted strongly convex function over PSD matrices and let $X^\star$ be the minimum of f with $rank(X^\star) = r$. Let $X^+ = \mathcal{P}_+(X - \frac{1}{M}\nabla f(X))$ be the projected gradient descent update. Then, $\|X^+ - X\|_F \leq \frac{c}{\kappa\sqrt{r}\tau(X_r)}\sigma_r(X)$ implies,*

$$\textsc{Dist}(U_r, U_r^\star) \leq \frac{c'}{\tau(X_r^\star)}\sigma_r(U_r^\star), \quad \text{for constants } c,\ c' > 0.$$

Next, we present a generic initialization scheme for general smooth and strongly convex f. We use only the *first-order oracle*: we only have access to—at most—gradient information of f. Our initialization comes with theoretical guarantees w.r.t. distance from optimum. Nevertheless, in order to show small relative distance in the form of $\textsc{Dist}(U^0, U_r^\star) \leq \rho\sigma_r(U_r^\star)$, one requires certain condition numbers of f and further assumptions on the spectrum of optimal solution $X^\star$ and rank r. However, empirical findings in Section F.3 show that our initialization performs well in practice.

Let $\nabla f(0) \in \mathbb{R}^{n\times n}$. Since the initial point should be in the PSD cone, we further consider the projection $\mathcal{P}_+(-\nabla f(0))$. By strong convexity and smoothness of f, one can observe that the point $1/M \cdot \mathcal{P}_+\left(-\nabla f(0)\right)$ is a good initialization point, within some radius from the vicinity of $X^\star$; *i.e.*,

$$\left\|\tfrac{1}{M}\mathcal{P}_+(-\nabla f(0)) - X^\star\right\|_F \ \leq\ 2\left(1 - \tfrac{m}{M}\right)\|X^\star\|_F; \tag{11}$$

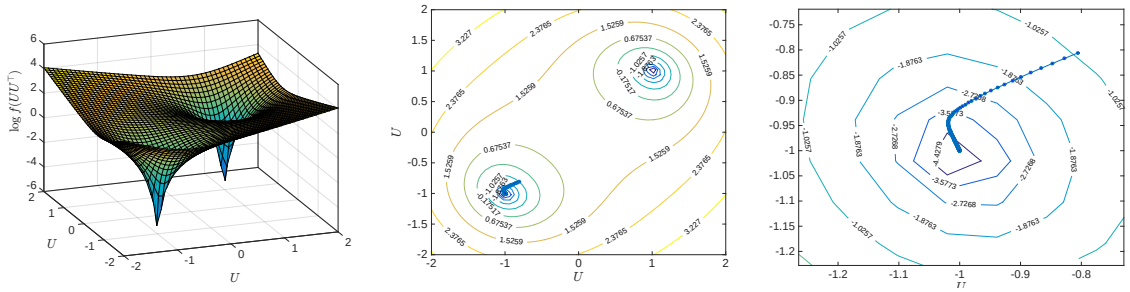

Figure 2: Abstract illustration of initialization effect on a toy example. In this experiment, we design $X^\star = U^\star U^{\star\top}$ where $U^\star = [1 \;\; 1]^\top$ (or $U^\star = -[1 \;\; 1]^\top$—these are equivalent). We observe $X^\star$ via $y = \text{vec}\,(A \cdot X^\star)$ where $A \in \mathbb{R}^{3\times 2}$ is randomly generated. We consider the loss function $f(UU^\top) = \frac{1}{2}\|y - \text{vec}\left(A \cdot UU^\top\right)\|_2^2$. Left panel: f values in logarithimic scale for various values of variable $U \in \mathbb{R}^{2\times 1}$. Center panel: Contour lines of f and the bahavior of FGD using our initialization scheme. Right panel: zoom-in plot of center plot.

see also Theorem 11. Thus, a scaling of $\mathcal{P}_+(-\nabla f(0))$ by M could serve as a decent initialization. In many recent works Jain et al. (2013); Netrapalli et al. (2013); Candes et al. (2015b); Zheng and Lafferty (2015); Chen and Wainwright (2015) this initialization has been used for specific applications.[12] Here, we note that the point $1/M \cdot \mathcal{P}_+\left(-\nabla f(0)\right)$ can be used as initialization point for generic smooth and strongly convex f.

The smoothness parameter M is not always easy to compute exactly; in such cases, one can use the surrogate $m \leq \|\nabla f(0) - \nabla f(e_1 e_1^\top)\|_F \leq M$. Finally, our initial point $U^0 \in \mathbb{R}^{n\times r}$ is a rank-r matrix such that $X_r^0 = U^0 U^{0\top}$.

We now present guarantees for the initialization discussed. The proof is provided in Section D.2.

Theorem 11 (Initialization) *Let f be a M-smooth and m-strongly convex function, with condition number $\kappa = \frac{M}{m}$, and let $X^\star$ be its minimum over PSD matrices. Let X^0 be defined as:*

$$X^0 := \frac{1}{\|\nabla f(0) - \nabla f(e_1 e_1 \top)\|_F} \mathcal{P}_+\left(-\nabla f(0)\right), \tag{12}$$

and X_r^0 is its rank-r approximation. Let $\|X^\star - X_r^\star\|_F \leq \tilde{\rho}\,\|X_r^\star\|_2$ for some $\tilde{\rho}$. Then, $\text{DIST}(U^0, U_r^\star) \leq \gamma\sigma_r(U_r^\star)$, *where* $\gamma = 4\tau(X_r^\star)\sqrt{2r}\cdot\left(\sqrt{\kappa^2 - 2/\kappa + 1}\left(\texttt{srank}^{1/2}\left(X_r^\star\right) + \tilde{\rho}\right) + \tilde{\rho}\right)$ *and* $\texttt{srank}\left(X_r^\star\right) = \frac{||X_r^\star||_F^2}{||X_r^\star||_2^2}$.

12. To see this, consider the case of least-squares objective $f(X) := \frac{1}{2}\|\mathcal{A}(X) - y\|_2^2$, where y denote the set of observations and, $\mathcal{A}$ is a properly designed sensing mechanism, depending on the problem at hand. For example, in the affine rank minimization case Zheng and Lafferty (2015); Chen and Wainwright (2015), $(\mathcal{A}(X))_i$ represents the linear system mechanism where $\text{Tr}(A_i \cdot X) = b_i$. Under this setting, computing the gradient $\nabla f(\cdot)$ at zero point, we have: $-\nabla f(0) = \mathcal{A}^*(y)$, where $\mathcal{A}^*$ is the adjoint operator of $\mathcal{A}$. Then, it is obvious that the operation $\mathcal{P}_+\left(-\nabla f(0)\right)$ is very similar to the spectral methods, proposed for initialization in the references above.

While the above result guarantees a good initialization for only small values of κ, in many applications Jain et al. (2013); Netrapalli et al. (2013); Chen and Wainwright (2015), this is indeed the case and X^0 has constant relative error to the optimum.

To understand this result, notice that in the extreme case, when f is the ℓ_2 loss function $\|X - X^\star\|_F^2$, which has condition number $\kappa = 1$ and rank$(X^\star) = r$, X^0 indeed is the optimum. More generally as the condition number κ increases, the optimum moves away from X^0 and the above theorem characterizes this error as a function of condition number of the function. See also Figure 2.

Now for the setting when the optimum is exactly rank-r we get the following result.

Corollary 12 (Initialization, exact) *Let $X^\star$ be rank-r for some $r \leq n$. Then, under the conditions of Theorem 11, we get*

$$\textsc{Dist}(U^0, U_r^\star) \leq 4\sqrt{2}r\tau(X_r^\star) \cdot \sqrt{\kappa^2 - 2/\kappa + 1} \cdot \sigma_r(U_r^\star).$$

Finally, for the setting when the function satisfies (m, r)-restricted strong convexity, the above corollary still holds as the optimum is a rank-r matrix.

Remark 13 *The above initialization strategies only attain Theorem's 6 and Corollary's 7 initialization requirements within some factor. In order to achieve such requirements, one requires further assumptions regarding the nature of the problem at hand, i.e., further restrictions on the condition number of $X^\star$ and f, as well as potential dependence on the rank parameter r. Proving global convergence, from random starting points and for a wide range of objective criteria f, remains an open problem.*

6. Convergence proofs for the FGD algorithm

In this section, we first present the key techniques required for analyzing the convergence of FGD. Later, we present proofs for both Theorems 5 and 6. Throughout the proofs we use the following notation. $X^\star$ is the optimum of problem (1) and $X_r^\star = U_r^\star R_U^\star (U_r^\star R_U^\star)^\top$ is the rank-r approximation; for the just smooth case, $X^\star = X_r^\star$, as we consider only the rank-$r^\star$ case and $r = r^\star$. Let $R_U^\star := \text{argmin}_{R:R\in\mathcal{O}} \|U - U_r^\star R\|_F$ and $\Delta = U - U_r^\star R_U^\star$.

A key property that assists classic gradient descent to converge to the optimum $X^\star$ is the fact that $\langle X^+ - X,\ X^\star - X\rangle \geq 0$ for a smooth convex function f; in the case of strongly convex f, the inner product is further lower bounded by $\frac{m}{2}\|X - X^\star\|_F^2$ (see Theorem 2.2.7 of Nesterov (2004)). Classical proofs mainly use such lower bounds to show convergence (see Theorems 2.1.13 and 2.2.8 of Nesterov (2004)).

We follow broadly similar steps in order to show convergence of FGD. In particular,

- In section 6.1, we show a lower bound for the inner product $\langle U - U^+, U - U_r^\star R_U^\star\rangle$ (Lemma 14), even though the function is not convex in U. The initialization and rank-r approximate optimum assumptions play a crucial role in proving this, along with the fact that f is convex in X.
- In sections 6.2 and 6.3, we use the above lower bound to show convergence for (i) smooth and strongly f, and (ii) just smooth f, respectively, similar to the convex setting.

6.1. Rudiments of our analysis

Next, we present the main descent lemma that is used for both sublinear and linear convergence rate guarantees of FGD.

Lemma 14 (Descent lemma) *For f being a M-smooth and (m, r)-strongly convex function and, under assumptions $(A2)$ and $(A3)$, the following inequality holds true:*

$$\tfrac{1}{\eta}\left\langle U - U^+, U - U_r^\star R_U^\star \right\rangle \geq \tfrac{2}{3}\eta\|\nabla f(X)U\|_F^2 + \tfrac{3m}{20}\cdot\sigma_r(X^\star)\textsc{Dist}(U, U_r^\star)^2 - \tfrac{M}{4}\left\|X^\star - X_r^\star\right\|_F^2.$$

Further, when f is just M-smooth convex function and, under the assumptions $f(X^+) \geq f(X_r^\star)$ and $(A1)$, we have:

$$\tfrac{1}{\eta}\left\langle U - U^+, U - U_r^\star R_U^\star \right\rangle \geq \tfrac{1}{2}\eta\|\nabla f(X)U\|_F^2.$$

Proof First, we rewrite the inner product as shown below.

$$\begin{aligned}\langle \nabla f(X)U, U - U_r^\star R_U^\star\rangle &= \left\langle \nabla f(X), X - U_r^\star R_U^\star U^\top \right\rangle \\ &= \frac{1}{2}\langle \nabla f(X), X - X_r^\star\rangle + \left\langle \nabla f(X), \frac{1}{2}(X + X_r^\star) - U_r^\star R_U^\star U^\top \right\rangle \\ &= \frac{1}{2}\langle \nabla f(X), X - X_r^\star\rangle + \frac{1}{2}\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle, \end{aligned} \tag{13}$$

which follows by adding and subtracting $\frac{1}{2}X_r^\star$.

- STRONGLY CONVEX f SETTING. For this case, the next 3 steps apply.

Step I: Bounding $\langle \nabla f(X), X - X_r^\star\rangle$. The first term in the above expression can be lower bounded using smoothness and strong convexity of f and, involves a construction of a feasible point X. We construct such a feasible point by modifying the current update to one with bigger step size $\widehat{\eta}$.

Lemma 15 *Let f be a M-smooth and (m, r)-restricted strongly convex function with optimum point $X^\star$. Moreover, let $X_r^\star$ be the best rank-r approximation of $X^\star$. Let $X = UU^\top$. Then,*

$$\langle \nabla f(X), X - X_r^\star\rangle \geq \tfrac{18\widehat{\eta}}{10}\|\nabla f(X)U\|_F^2 + \tfrac{m}{2}\left\|X - X_r^\star\right\|_F^2 - \tfrac{M}{2}\|X^\star - X_r^\star\|_F^2,$$

where $\widehat{\eta} = \frac{1}{16(M\|X\|_2 + \|\nabla f(X)Q_U Q_U^\top\|_2)} \geq \frac{5\eta}{6}$, by Lemma 21.

Proof of this lemma is provided in Section B.1.

Step II: Bounding $\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle$. The second term in equation (13) can actually be negative. Hence, we lower bound it using our initialization assumptions. Intuitively, the second term is smaller than the first one as it scales as $\textsc{Dist}(U, U_r^\star)^2$, while the first term scales as $\textsc{Dist}(U, U_r^\star)$.

Lemma 16 *Let f be M-smooth and (m, r)-restricted strongly convex. Then, under assumptions $(A2)$ and $(A4)$, the following bound holds true:*

$$\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \geq -\tfrac{2\widehat{\eta}}{25}\|\nabla f(X)U\|_F^2 - \left(\tfrac{m\sigma_r(X^\star)}{20} + M\|X^\star - X_r^\star\|_F\right)\cdot \textsc{Dist}(U, U_r^\star)^2.$$

Proof of this lemma can be found in Section B.2.

Step III: Combining the bounds in equation (13). For a detailed description, see Section B.3.

- SMOOTH f SETTING.

Similar to the strongly convex case, we can obtain a lower bound on $\langle \nabla f(X), X - X_r^\star \rangle$ (Lemma 24) and upper bound on $\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle$ (Lemma 23). Combining the bounds into equation (13) gives the result. For a detailed description, see Section C and Lemma 25.

■

6.2. Proof of linear convergence (Theorem 6)

The proof of this theorem involves showing that the potential function $\textsc{Dist}(U, U_r^\star)$ is decreasing per iteration (up to approximation error $||X^\star - X_r^\star||_F$), using the descent Lemma 14. Using the algorithm's update rule, we obtain

$$\begin{aligned}
\textsc{Dist}(U^+, U_r^\star)^2 &= \min_{R:\ R\in\mathcal{O}} \|U - U_r^\star R\|_F^2 \\
&\le \|U^+ - U_r^\star R_U^\star\|_F^2 \\
&= \|U^+ - U + U - U_r^\star R_U^\star\|_F^2 \\
&= \|U^+ - U\|_F^2 + \|U - U_r^\star R_U^\star\|_F^2 - 2\left\langle U^+ - U, U_r^\star R_U^\star - U \right\rangle, \qquad (14)
\end{aligned}$$

which follows by adding and subtracting U and then expanding the squared term.

Step I: Bounding term $\langle U - U^+, U - U^\star R \rangle$ *in* (14). By Lemma 14, we can bound the last term on the right hand side as:

$$\langle \nabla f(X)U, U - U_r^\star R_U^\star \rangle \ge \tfrac{2}{3}\eta \|\nabla f(X)U\|_F^2 + \tfrac{3m}{20}\cdot\sigma_r(X^\star)\textsc{Dist}(U, U_r^\star)^2 - \tfrac{M}{4}\|X^\star - X_r^\star\|_F^2 .$$

Furthermore, we can substitute U^+ in the first term to obtain $\|U^+ - U\|_F^2 = \eta^2 \|\nabla f(X)U\|_F^2$.

Step II: Combining bounds into (14). Combining the above two equations (14) becomes:

$$\begin{aligned}
\textsc{Dist}(U^+, U_r^\star)^2 &\le \eta^2\|\nabla f(X)U\|_F^2 + \|U - U_r^\star R_U^\star\|_F^2 \\
&\qquad - 2\eta\left(\tfrac{2}{3}\eta\|\nabla f(X)U\|_F^2 + \tfrac{3m}{20}\cdot\sigma_r(X^\star)\textsc{Dist}(U, U_r^\star)^2 - \tfrac{M}{4}\|X^\star - X_r^\star\|_F^2\right) \\
&= \|U - U_r^\star R_U^\star\|_F^2 + \tfrac{\eta M}{2}\|X^\star - X_r^\star\|_F^2 + \eta^2 \underbrace{\left(\|\nabla f(X)U\|_F^2 - \tfrac{4}{3}\|\nabla f(X)U\|_F^2\right)}_{\le 0} \\
&\qquad - \tfrac{3m\eta}{10}\cdot\sigma_r(X^\star)\textsc{Dist}(U, U_r^\star)^2 \\
&\overset{(i)}{\le} \|U - U_r^\star R_U^\star\|_F^2 + \tfrac{\eta M}{2}\|X^\star - X_r^\star\|_F^2 - \tfrac{3m\eta}{10}\cdot\sigma_r(X^\star)\textsc{Dist}(U, U_r^\star)^2 \\
&= \left(1 - \tfrac{3m\eta}{10}\cdot\sigma_r(X^\star)\right)\cdot\textsc{Dist}(U, U_r^\star)^2 + \tfrac{\eta M}{2}\|X^\star - X_r^\star\|_F^2 \\
&\overset{(ii)}{\le} \left(1 - \tfrac{3m}{10}\cdot\tfrac{10\eta^\star}{11}\cdot\sigma_r(X^\star)\right)\cdot\textsc{Dist}(U, U_r^\star)^2 + M\cdot\tfrac{11\eta^\star}{20}\|X^\star - X_r^\star\|_F^2 \\
&\le \left(1 - \tfrac{m\eta^\star}{4}\cdot\sigma_r(X^\star)\right)\textsc{Dist}(U, U_r^\star)^2 + \tfrac{11M\eta^\star}{20}\|X^\star - X_r^\star\|_F^2 \\
&\overset{(iii)}{\le} \left(1 - \tfrac{m\sigma_r(X^\star)}{64(M\|X^\star\|_2 + \|\nabla f(X^\star)\|_2)}\right)\textsc{Dist}(U, U_r^\star)^2 \\
&\qquad + \tfrac{M}{28(M\|X^\star\|_2 + \|\nabla f(X^\star)\|_2)}\|X^\star - X_r^\star\|_F^2 ,
\end{aligned}$$

where (i) is due to removing the negative part from the right hand side, (ii) is due to $\frac{10}{11}\eta^\star \le \eta \le \frac{11}{10}\eta^\star$ by Lemma 21, (iii) follows from substituting $\eta^\star = \frac{1}{16(M\|X^\star\|_2+\|\nabla f(X^\star)\|_2)}$. This proves the first part of the theorem.

Step III: U^+ satisfies the initial condition. Now we will prove the second part. By the above equation, we have:

$$\begin{aligned}
\textsc{Dist}(U^+, U_r^\star)^2 &\le \left(1 - \tfrac{m\eta^\star}{4}\cdot\sigma_r(X^\star)\right)\textsc{Dist}(U, U_r^\star)^2 + \tfrac{11M\eta^\star}{20}\|X^\star - X_r^\star\|_F^2 \\
&\overset{(i)}{\le} \left(1 - \tfrac{m\eta^\star}{4}\cdot\sigma_r(X^\star)\right)(\rho')^2\sigma_r(X^\star) + \tfrac{11M\eta^\star}{20}\tfrac{(\rho')^2}{4\kappa}\sigma_r^2(X^\star) \\
&= (\rho')^2\sigma_r(X^\star)\left(1 - \tfrac{m\eta^\star}{4}\cdot\sigma_r(X^\star) + \tfrac{11M\eta^\star}{80\kappa}\cdot\sigma_r(X^\star)\right) \\
&\le (\rho')^2\sigma_r(X^\star)\left(1 - \tfrac{m\eta^\star}{4}\cdot\sigma_r(X^\star) + \tfrac{m\eta^\star}{7}\sigma_r(X^\star)\right) \\
&\le (\rho')^2\sigma_r(X^\star).
\end{aligned}$$

(i) follows from substituting the assumptions on $\textsc{Dist}(U, U_r^\star)$ and $\|X^\star - X_r^\star\|_F$ and the last inequality is due to the term in the parenthesis being less than one.

6.3. Proof of sublinear convergence (Theorem 5)

Here, we show convergence of FGD when f is only a M-smooth convex function. At iterate k, we assume $f(X^k) > f(X_r^\star)$; in the opposite case, the bound follows trivially. Recall the updates of FGD over the U-space satisfy

$$U^+ = U - \eta\nabla f(X)U.$$

It is easy to verify that $X^+ = U^+(U^+)^\top = X - \eta\nabla f(X)X\Lambda - \eta\Lambda^\top X\nabla f(X)$, where $\Lambda = I - \frac{\eta}{2}Q_U Q_U^\top \nabla f(X) \in \mathbb{R}^{n\times n}$. Notice that for step size η, using Lemma A.5 we get,

$$\Lambda \succ 0, \quad \|\Lambda\|_2 \le 1 + 1/32 \quad \text{and} \quad \sigma_n(\Lambda) \ge 1 - 1/32. \tag{15}$$

Our proof proceeds using the smoothness condition on f, at point X^+. In particular,

$$\begin{aligned}
f(X^+) &\le f(X) + \langle\nabla f(X), X^+ - X\rangle + \tfrac{M}{2}\|X^+ - X\|_F^2 \\
&\overset{(i)}{\le} f(X) - 2\eta\cdot\sigma_n(\Lambda)\cdot\|\nabla f(X)U\|_F^2 + 2M\eta^2\cdot\|\nabla f(X)U\|_F^2\cdot\|X\|_2\cdot\|\Lambda\|_2^2 \\
&\overset{(ii)}{\le} f(X) - \tfrac{\eta\cdot 62}{32}\cdot\|\nabla f(X)U\|_F^2 + \tfrac{\eta}{7}\cdot\left(\tfrac{33}{32}\right)^2\cdot\|\nabla f(X)U\|_F^2 \\
&\le f(X) - \tfrac{17\eta}{10}\|\nabla f(X)U\|_F^2,
\end{aligned}$$

where (i) follows from symmetry of $\nabla f(X)$, X and

$$\begin{aligned}
\mathrm{Tr}(\nabla f(X)\nabla f(X)X\Lambda) &= \mathrm{Tr}(\nabla f(X)\nabla f(X)UU^\top) - \frac{\eta}{2}\mathrm{Tr}(\nabla f(X)\nabla f(X)UU^\top\nabla f(X)) \\
&\ge (1 - \frac{\eta}{2}\|Q_U Q_U^\top\nabla f(X)\|_2)\|\nabla f(X)U\|_F^2 \\
&\ge (1 - 1/32)\|\nabla f(X)U\|_F^2,
\end{aligned} \tag{16}$$

and (ii) is due to (15) and the fact that $\eta \leq \frac{1}{16M\|X^0\|_2} \leq \frac{1}{14M\|X\|_2}$ (see Lemma A.5). Hence,

$$f(X^+) - f(X_r^\star) \leq f(X) - f(X_r^\star) - \tfrac{18\eta}{10}\|\nabla f(X)U\|_F^2. \tag{17}$$

To bound the term $f(X) - f(X_r^\star)$ on the right hand side of (17), we use standard convexity as follows:

$$\begin{aligned}
f(X) - f(X_r^\star) &\leq \langle \nabla f(X), X - X_r^\star \rangle \\
&\overset{(i)}{=} 2\left\langle \nabla f(X), UU^\top - U_r^\star R_U^\star U^\top \right\rangle - \left\langle \nabla f(X), UU^\top + U_r^\star R_U^\star (U_r^\star R_U^\star)^\top - 2U_r^\star R_U^\star U^\top \right\rangle \\
&= 2\langle \nabla f(X)U, U - U_r^\star R_U^\star \rangle - \left\langle \nabla f(X), (U - U_r^\star R_U^\star)(U - U_r^\star R_U^\star)^\top \right\rangle \\
&\overset{(ii)}{=} 2\langle \nabla f(X)U, \Delta \rangle - \left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \\
&\leq 2\langle \nabla f(X)U, \Delta \rangle + \left| \left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \right| \\
&\overset{(iii)}{\leq} 2 \cdot \|\nabla f(X)U\|_F \cdot \textsc{Dist}(U, U_r^\star) + \tfrac{1}{40}\|\nabla f(X)U\|_2 \cdot \textsc{Dist}(U, U_r^\star) \\
&\overset{(iv)}{\leq} \tfrac{5}{2}\|\nabla f(X)U\|_F \cdot \textsc{Dist}(U, U_r^\star),
\end{aligned} \tag{18}$$

where (i) is due to $X = UU^\top$ and $X_r^\star = U_r^\star R_U^\star (U_r^\star R_U^\star)^\top$ for orthonormal matrix $R_U^\star \in \mathbb{R}^{r\times r}$, (ii) is by $\Delta := U - U_r^\star R_U^\star$, (iii) is due to Cauchy-Schwarz inequality and Lemma 22 and, (iv) is due to norm ordering $\|\cdot\|_2 \leq \|\cdot\|_F$.

From (18), we obtain to the following bound:

$$\|\nabla f(X)U\|_F \geq \tfrac{2}{5} \cdot \tfrac{f(X) - f(X_r^\star)}{\textsc{Dist}(U, U_r^\star)}. \tag{19}$$

Define $\delta = f(X) - f(X_r^\star)$ and $\delta^+ = f(X^+) - f(X_r^\star)$. Moreover, by Lemma 26, we know that $\textsc{Dist}(U, U_r^\star) \leq \textsc{Dist}(U^0, U_r^\star)$ for all iterations of FGD; thus, we have $\frac{1}{\textsc{Dist}(U, U_r^\star)} \geq \frac{1}{\textsc{Dist}(U^0, U_r^\star)}$ for every update U. Using the above definitions and substituting (19) in (17), we obtain the following recursion:

$$\delta^+ \leq \delta - \tfrac{17\eta}{10} \cdot \left(\tfrac{2}{5}\right)^2 \cdot \left(\frac{\delta}{\|\Delta\|_F}\right)^2 \leq \delta - \tfrac{\eta}{5\cdot \textsc{Dist}(U^0, U_r^\star)^2} \cdot \delta^2 \Longrightarrow \delta^+ \leq \delta\left(1 - \tfrac{\eta}{5\cdot \textsc{Dist}(U^0, U_r^\star)^2} \cdot \delta\right),$$

which can be further transformed as:

$$\frac{\left(1 - \frac{\eta}{5\cdot \textsc{Dist}(U^0, U_r^\star)^2} \cdot \delta\right)}{\delta^+} \geq \frac{1}{\delta} \Longrightarrow \frac{1}{\delta^+} \geq \frac{1}{\delta} + \tfrac{\eta}{5\cdot \textsc{Dist}(U^0, U_r^\star)^2} \cdot \frac{\delta}{\delta^+} \geq \frac{1}{\delta} + \tfrac{\eta}{5\cdot \textsc{Dist}(U^0, U_r^\star)^2}$$

since $\delta^+ \leq \delta$ from equation (17). Since each δ and δ^+ correspond to previous and new estimate in FGD per iteration, we can sum up the above inequalities over k iterations to obtain

$$\frac{1}{\delta^k} \geq \frac{1}{\delta^0} + \tfrac{\eta}{5\cdot \textsc{Dist}(U^0, U_r^\star)^2} \cdot k;$$

here, $\delta^k := f(X^k) - f(X_r^\star)$ and $\delta^0 := f(X^0) - f(X_r^\star)$. After simple transformations, we finally obtain:

$$f(X^k) - f(X_r^\star) \leq \frac{\frac{5}{\eta} \cdot \textsc{Dist}(U^0, U_r^\star)^2}{k + \frac{5}{\eta} \cdot \frac{\textsc{Dist}(U^0, U_r^\star)^2}{f(X^0) - f(X_r^\star)}}.$$

One can further obtain a bound on the right hand side that depends on $\eta^\star = \frac{1}{16\,(M\|X^\star\|_2+\|\nabla f(X^\star)\|_2)}$. By Lemma 21, we know $\eta \geq \frac{10}{11}\eta^\star$. Substituting this in the above equation gives the result.

7. Related work

Convex approaches. A significant volume of work has focused on solving the classic Semi-Definite Programming (SDP) formulation, where the *objective f (as well as any additional convex constraints) is assumed to be linear*. There, interior point methods (IPMs) constitute a popular choice for small- and moderate-sized problems; see Karmarkar (1984); Alizadeh (1995). For a comprehensive treatment of this subject, see the excellent survey in Monteiro (2003).

Large scale SDPs pointed research towards first-order approaches, which are more computationally appealing. For linear f, we note among others the work of Wen et al. (2010), a provably convergent alternating direction augmented Lagrangian algorithm, and that of Helmberg and Rendl Helmberg and Rendl (2000), where they develop an efficient first-order spectral bundle method for SDPs with the constant trace property; see also Helmberg et al. (2014) for extensions on this line of work. In both cases, no convergence rate guarantees are provided; see also Monteiro (2003). For completeness, we also mention the work of Burer (2003); Fukuda et al. (2001); Nakata et al. (2003); Toh (2004) on *second-order* methods, that take advantage of data sparsity in order to handle large SDPs in a more efficient way. However, it turns out that the amount of computations required per iteration is comparable to that of log-barrier IPMs Monteiro (2003).

Standard SDPs have also found application in the field of combinatorial optimization; there, in most cases, even a rough approximation to the discrete problem, via SDP, is sufficiently accurate and computationally affordable, than exhaustive combinatorial algorithms. Goemans and Williamson Goemans and Williamson (1995) were the first to propose the use of SDPs in approximating graph MAX CUT, where a near-optimum solution can be found in polynomial time. Klein and Lu (1996) propose an alternative approach for solving MAX CUT and GRAPH COLORING instances, where SDPs are transformed into eigenvalue problems. Then, power method iterations lead to ε-approximate solutions; however, the resulting running-time dependence on ε is worse, compared to standard IPMs. Arora, Hazan and Kale in Arora et al. (2005) derive an algorithm to approximate SDPs, as a hybrid of the Multiplicative Weights Update method and of ideas originating from an ellipsoid variant Vaidya (1989), improving upon existing algorithms for graph partitioning, computational biology and metric embedding problems.[13]

Extending to *non-linear convex f* cases, Nesterov and Nemirovski (1988, 1989) have shown how IPMs can be generalized to solve instances of (1), via the notion of self-concordance; see also Lee et al. (2012); Dinh et al. (2015) for a more recent line of work. Within the class of first-order methods, approaches for nonlinear convex f include, among others, projected and proximal gradient descent methods Nesterov (2004); Dinh et al. (2015); Jiang et al. (2012), (smoothed) dual ascent methods Nesterov (2007), as well as Frank-Wolfe algorithm variants Jaggi (2011). Note that all these schemes, often require heavy calculations, such as eigenvalue decompositions, to compute the

13. The algorithm in Arora et al. (2005) shows significant computational gains over standard IPMs per iteration, due to requiring only a power method calculation per iteration (versus a Cholesky factorization per iteration, in the latter case). However, the polynomial dependence on the accuracy parameter $\frac{1}{\varepsilon}$ is worse, compared to IPMs. Improvements upon this matter can be found in Arora and Kale (2007) where a primal-dual Multiplicative Weights Update scheme is proposed.

updates (often, to remain within the feasible set).

Burer & Monteiro factorization and related work. Burer and Monteiro Burer and Monteiro (2003, 2005) popularized the idea of solving classic SDPs by representing the solution as a product of two factor matrices. The main idea in such representation is to remove the positive semi-definite constraint by directly embedding it into the objective. While the problem becomes non-convex, Burer and Monteiro propose a method-of-multiplier type of algorithm which iteratively updates the factors in an alternating fashion. For linear objective f, they establish convergence guarantees to the optimum but do not provide convergence rates.

For generic smooth convex functions, Hazan in Hazan (2008) proposes SPARSEAPPROXSDP algorithm,[14] a generalization of the Frank-Wolfe algorithm for the vector case Clarkson (2010), where putative solutions are refined by rank-1 approximations of the gradient. At the r-th iteration, SPARSEAPPROXSDP is guaranteed to compute a $\frac{1}{r}$-approximate solution, with rank at most r, *i.e.*, achieves a sublinear $O\left(\frac{1}{\varepsilon}\right)$ convergence rate. However, depending on ε, SPARSEAPPROXSDP is not guaranteed to return a low rank solution unlike FGD. Application of these ideas in machine learning tasks can be found in Shalev-shwartz et al. (2011). Based on SPARSEAPPROXSDP algorithm, Laue (2012) further introduces "de-bias" steps in order to optimize parameters in SPARSEAPPROXSDP and do local refinements of putative solutions via L-BFGS steps. Nevertheless, the resulting convergence rate is still sublinear.[15]

Specialized algorithms – for objectives beyond the linear case – that utilize such factorization include matrix completion /sensing solvers Jain et al. (2013); Sun and Luo (2014); Zheng and Lafferty (2015); Tu et al. (2015), non-negative matrix factorization schemes Lee and Seung (2001), phase retrieval methods Netrapalli et al. (2013); Candes et al. (2015b) and sparse PCA algorithms Laue (2012). Most of these results guarantee linear convergence for various algorithms on the factored space starting from a "good" initialization. They also present a simple spectral method to compute such an initialization. For the matrix completion /sensing setting, Sa et al. (2015) have shown that stochastic gradient descent achieves global convergence at a sublinear rate. Note that these results only apply to quadratic loss objectives and not to generic convex functions f.[16] Jain et al. (2015) consider the problem of computing the matrix square-root of a PSD matrix via gradient descent on the factored space: in this case, the objective f boils down to minimizing the standard squared Euclidean norm distance between two matrices. Surprisingly, the authors show that, given an initial point that is well-conditioned, the proposed scheme is guaranteed to find an ε-accurate solution with linear convergence rate; see also Sra (2015) for a more recent discussion on this problem.

Chen and Wainwright (2015) propose a first-order optimization framework for the problem (1), where the same parametrization technique is used to efficiently accommodate the PSD constraint.[17] Moreover, the proposed algorithmic solution can accommodate extra constraints on X.[18] The set of assumptions listed in Chen and Wainwright (2015) include—apart from $X^\star$-faithfulness—local de-

14. Sparsity here corresponds to low-rankness of the solution, as in the Cholesky factorization representation. Moreover, inspired by Quantum State Tomography applications Aaronson (2007), SPARSEAPPROXSDP can also handle constant trace constraints, in addition to PSD ones.
15. For running time comparisons with FGD see Section G.
16. We recently became aware of the extension of the work Tu et al. (2015) for the non-square case $X = UV^\top$.
17. In this work, the authors further assume orthogonality of columns in U.
18. Though, additional constraints should satisfy the $X^\star$-*faithfulness* property: a constraint set on U, say $\mathcal{U}$, is faithful if for each $U \in \mathcal{U}$, that is within some bounded radius from optimal point, we are guaranteed that the closest (in the Euclidean sense)$U^\star R$ lies within $\mathcal{U}$.

scent, local Lipschitz and local smoothness conditions *in the factored space*. *E.g.*, the local descent condition can be established if $g(U) := f(UU^\top)$ is locally strongly convex and $\nabla g(\cdot)$ at an optimum point vanishes. They also require bounded gradients as their step size doesn't account for the modified curvature of $f(UU^\top)$. [19] These conditions are less standard than the global assumptions of the current work and one needs to validate that they are satisfied for each problem, separately. Chen and Wainwright (2015) presents some applications where these conditions are indeed satisfied. Their results are of the same flavor with ours: under such proper assumptions, one can prove local convergence with $O(1/\varepsilon)$ or $O(\log(1/\varepsilon))$ rate and for f instances that even fail to be locally convex.

Finally, for completeness, we also mention optimization over the Grassmannian manifold that admits tailored solvers Edelman et al. (1998); see Keshavan et al. (2010); Boumal (2014, 2015); Zhang and Balzano (2015); Uschmajew and Vandereycken (2015) for applications in matrix completion and references therein. Journée et al. (2010) presents a second-order method for (1), based on manifold optimization over the set of all equivalence class $\mathcal{O}$. The proposed algorithm can additionally accommodate constraints and enjoys monotonic decrease of the objective function (in contrast to Burer and Monteiro (2003, 2005)), featuring quadratic local convergence. In practice, the per iteration complexity is dominated by the extraction of the eigenvector, corresponding to the smallest eigenvalue, of a $n \times n$ matrix—and only when the current estimate of rank satisfies some conditions.

Table 1 summarizes the comparison of the most relevant work to ours, for the case of matrix factorization techniques.

Reference	Conv. rate	Initialization	Output rank
Hazan (2008)	$1/\varepsilon$ (Smooth f)	$X^0 = 0$	$1/\varepsilon$
Laue (2012)	$1/\varepsilon$ (Smooth f)	$X^0 = 0$	$1/\varepsilon$
Chen and Wainwright (2015)	$1/\varepsilon$ (Local Asm.)	Application dependent	r
Chen and Wainwright (2015)	$\log(1/\varepsilon)$ (Local Asm.)	Application dependent	r
This work	$1/\varepsilon$ (Smooth f)	SVD / top-r	r
This work	$\log(1/\varepsilon)$ (Smooth, RSC f)	SVD / top-r	r

Table 1: Summary of selected results on solving variants of (1) via matrix factorization. "Conv. rate" describes the number of iterations required to achieve ε accuracy. "Initialization" describes the process for starting point computation. "SVD" stands for singular value decomposition and "top-r" denotes that a rank-r decomposition is computed. For the case of Chen and Wainwright (2015), "Local Asm." refer to specific assumptions made on the U-space; we refer the reader to the footnote for a short description. "Output rank" denotes the maximum rank of solution returned for ε-accuracy.

8. Conclusion

In this paper, we focus on how to efficiently minimize a convex function f over the positive semi-definite cone. Inspired by the seminal work Burer and Monteiro (2003, 2005), we drop convexity by factorizing the optimization variable $X = UU^\top$ and show that *factored gradient descent* with

19. One can define non-trivially conditions on the original space; we defer the reader to Chen and Wainwright (2015)

a non-trivial step size selection results in linear convergence when f is smooth and (restricted) strongly convex, even though the problem is now non-convex. In the case where f is only smooth, only sublinear rate is guaranteed. In addition, we present initialization schemes that use only first order information and guarantee to find a starting point with small relative distance from optimum.

There are many possible directions for future work, extending the idea of using non-convex formulation for semi-definite optimization. Showing convergence under weaker initialization condition or without any initialization requirement is definitely of great interest. Another interesting direction is to improve the convergence rates presented in this work, by using acceleration techniques and thus, extend ideas used in the case of convex gradient descent Nesterov (2004). Finally, it would be valuable to see how the techniques presented in this paper can be generalized to other standard algorithms like stochastic gradient descent and coordinate descent.

Furthermore, we identify applications, such as sparse PCA Vu et al. (2013); Asteris et al. (2015), that require non-smooth constraints on the factors U. That being said, an extension of this work to proximal techniques for the non-convex case is a very interesting future research direction.

References

Scott Aaronson. The learnability of quantum states. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 463, pages 3089–3114. The Royal Society, 2007.

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

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

Sanjeev Arora and Satyen Kale. A combinatorial, primal-dual approach to semidefinite programs. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 227–236. ACM, 2007.

Sanjeev Arora, Elad Hazan, and Satyen Kale. Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*, pages 339–348. IEEE, 2005.

Megasthenis Asteris, Dimitris Papailiopoulos, Anastasios Kyrillidis, and Alexandros G Dimakis. Sparse PCA via bipartite matchings. *arXiv preprint arXiv:1508.00625*, 2015.

Stephen Becker, Volkan Cevher, and Anastasios Kyrillidis. Randomized low-memory singular value projection. In *10th International Conference on Sampling Theory and Applications (Sampta)*, 2013.

Rajendra Bhatia. *Perturbation bounds for matrix eigenvalues*, volume 53. SIAM, 1987.

Nicolas Boumal. *Optimization and estimation on manifolds*. PhD thesis, UC Louvain, Belgium, 2014.

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

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Sébastien Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 2014.

Samuel Burer. Semidefinite programming in the space of partial positive semidefinite matrices. *SIAM Journal on Optimization*, 14(1):139–172, 2003.

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 and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57 (4):2342–2359, 2011.

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

Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM Review*, 57(2):225–251, 2015a.

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

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

Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Coherent matrix completion. In *Proceedings of The 31st International Conference on Machine Learning*, pages 674–682, 2014.

Yuxin Chen and Sujay Sanghavi. A general framework for high-dimensional estimation in the presence of incoherence. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1570–1576. IEEE, 2010.

Kenneth L Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):63, 2010.

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

Quoc Tran Dinh, Anastasios Kyrillidis, and Volkan Cevher. Composite self-concordant minimization. *Journal of Machine Learning Research*, 16:371–416, 2015.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Mituhiro Fukuda, Masakazu Kojima, Kazuo Murota, and Kazuhide Nakata. Exploiting sparsity in semidefinite programming via matrix completion I: General framework. *SIAM Journal on Optimization*, 11(3):647–674, 2001.

Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42 (6):1115–1145, 1995.

Elad Hazan. Sparse approximate solutions to semidefinite programs. In *LATIN 2008: Theoretical Informatics*, pages 306–316. Springer, 2008.

Christoph Helmberg and Franz Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696, 2000.

Christoph Helmberg, Michael L Overton, and Franz Rendl. The spectral bundle method with second-order information. *Optimization Methods and Software*, 29(4):855–876, 2014.

Roger A Horn and Charles R Johnson. Topics in matrix analysis. *Cambridge University Presss, Cambridge*, 37:39, 1991.

Cho-Jui Hsieh, Inderjit S Dhillon, Pradeep K Ravikumar, and Mátyás A Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*, pages 2330–2338, 2011.

Martin Jaggi. Convex optimization without projection steps. *arXiv preprint arXiv:1108.1170*, 2011.

Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*, pages 665–674. ACM, 2013.

Prateek Jain, Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Computing matrix squareroot via non convex local search. *arXiv preprint arXiv:1507.05854*, 2015.

Adel Javanmard and Andrea Montanari. Localization from incomplete noisy distance measurements. *Foundations of Computational Mathematics*, 13(3):297–345, 2013.

Kaifeng Jiang, Defeng Sun, and Kim-Chuan Toh. An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP. *SIAM Journal on Optimization*, 22(3):1042–1064, 2012.

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.

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311. ACM, 1984.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.

Philip Klein and Hsueh-I Lu. Efficient approximation algorithms for semidefinite programs arising from MAX CUT and COLORING. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 338–347. ACM, 1996.

Anastasios Kyrillidis and Volkan Cevher. Matrix recipes for hard thresholding methods. *Journal of mathematical imaging and vision*, 48(2):235–265, 2014.

Anastasios Kyrillidis, Rabeeh Karimi, Quoc Tran Dinh, and Volkan Cevher. Scalable sparse covariance estimation via self-concordance. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Soeren Laue. A hybrid algorithm for convex semidefinite optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 177–184, 2012.

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

Jason Lee, Yuekai Sun, and Michael Saunders. Proximal newton-type methods for convex optimization. In *Advances in Neural Information Processing Systems*, pages 836–844, 2012.

Yi-Kai Liu. Universal low-rank matrix recovery from Pauli measurements. In *Advances in Neural Information Processing Systems*, pages 1638–1646, 2011.

Leon Mirsky. A trace inequality of John von Neumann. *Monatshefte für Mathematik*, 79(4):303–306, 1975.

Bamdev Mishra, Gilles Meyer, and Rodolphe Sepulchre. Low-rank optimization for distance matrix completion. In *Decision and control and European control conference (CDC-ECC), 2011 50th IEEE conference on*, pages 4455–4460. IEEE, 2011.

Renato DC Monteiro. First-and second-order methods for semidefinite programming. *Mathematical Programming*, 97(1-2):209–244, 2003.

Kazuhide Nakata, Katsuki Fujisawa, Mituhiro Fukuda, Masakazu Kojima, and Kazuo Murota. Exploiting sparsity in semidefinite programming via matrix completion II: Implementation and numerical results. *Mathematical Programming*, 95(2):303–327, 2003.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

Yurii Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110(2):245–259, 2007.

Yurii Nesterov and Arkadi Nemirovski. A general approach to polynomial-time algorithms design for convex programming. *Report, Central Economical and Mathematical Institute, USSR Academy of Sciences, Moscow*, 1988.

Yurii Nesterov and Arkadi Nemirovski. *Self-concordant functions and polynomial-time methods in convex programming*. USSR Academy of Sciences, Central Economic & Mathematic Institute, 1989.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

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.

Christopher D Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2332–2341, 2015.

Shai Shalev-shwartz, Alon Gonen, and Ohad Shamir. Large-scale convex minimization with a low-rank constraint. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 329–336, 2011.

Suvrit Sra. On the matrix square root via geometric optimization. *arXiv preprint arXiv:1507.08366*, 2015.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *arXiv preprint arXiv:1411.8003*, 2014.

Kim-Chuan Toh. Solving large scale semidefinite programs via an iterative solver on the augmented systems. *SIAM Journal on Optimization*, 14(3):670–698, 2004.

Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via Procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

Andre Uschmajew and Bart Vandereycken. Greedy rank updates combined with riemannian descent methods for low-rank optimization. In *12th International Conference on Sampling Theory and Applications (Sampta)*, 2015.

Pravin M Vaidya. A new algorithm for minimizing convex functions over convex sets. In *Foundations of Computer Science, 1989., 30th Annual Symposium on*, pages 338–343. IEEE, 1989.

Vincent Q Vu, Jing Lei, et al. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

Andrew E Waters, Aswin C Sankaranarayanan, and Richard Baraniuk. Sparcs: Recovering low-rank and sparse matrices from compressive measurements. In *Advances in neural information processing systems*, pages 1089–1097, 2011.

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

Chris D White, Sujay Sanghavi, and Rachel Ward. The local convexity of solving systems of quadratic equations. *arXiv preprint arXiv:1506.07868*, 2015.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *Proceedings of The 31st International Conference on Machine Learning*, pages 593–601, 2014.

Dejiao Zhang and Laura Balzano. Global convergence of a grassmannian gradient descent algorithm for subspace estimation. *arXiv preprint arXiv:1506.07405*, 2015.

Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

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

Appendix A. Supporting lemmata

Lemma 17 (Hoffman, Wielandt Bhatia (1987)) *Let A and B be two PSD $n \times n$ matrices. Also let A be full rank. Then,*

$$\mathrm{Tr}(AB) \geq \sigma_{\min}(A)\,\mathrm{Tr}(B). \tag{20}$$

The following lemma shows that DIST, in the factor U space, upper bounds the Frobenius norm distance in the matrix X space.

Lemma 18 *Let $X = UU^\top$ and $X_r^\star = U_r^\star U_r^{\star\top}$ be two $n \times n$ rank-r PSD matrices. Let $\textsc{Dist}(U, U_r^\star) \leq \rho\sigma_r(U_r^\star)$, for some orthonormal matrix $R_U^\star$ and constant $\rho > 0$. Then,*

$$\|X - X_r^\star\|_F \leq (2+\rho)\rho \cdot \|U_r^\star\|_2 \cdot \sigma_r(U_r^\star).$$

Proof By substituting $X = UU^\top$ and $X_r^\star = U_r^\star U_r^{\star\top}$ in $\|X - X_r^\star\|_F$, we have:

$$\begin{aligned}
\|X - X_r^\star\|_F &= \|UU^\top - U_r^\star U_r^{\star\top}\|_F \\
&\overset{(i)}{=} \|UU^\top - U_r^\star R_U^\star U^\top + U_r^\star R_U^\star U^\top - U_r^\star R_U^\star (U_r^\star R_U^\star)^\top\|_F \\
&\overset{(ii)}{\leq} \textsc{Dist}(U, U_r^\star) \cdot \|U\|_2 + \textsc{Dist}(U, U_r^\star) \cdot \|U_r^\star\|_2 \\
&\overset{(iii)}{\leq} (1+\rho)\|U^\star\|_2 \cdot \textsc{Dist}(U, U_r^\star) + \textsc{Dist}(U, U_r^\star) \cdot \|U_r^\star\|_2 \\
&= (2+\rho) \cdot \textsc{Dist}(U, U_r^\star) \cdot \|U_r^\star\|_2 \\
&\overset{(iv)}{\leq} (2+\rho)\rho \cdot \|U_r^\star\|_2 \cdot \sigma_r(U_r^\star)
\end{aligned}$$

where (i) is due to the orthogonality $R_U^{\star\top} R_U^\star = I_{r\times r}$, (ii) is due to the triangle inequality, the Cauchy-Schwarz inequality and the fact that spectral norm is invariant w.r.t. orthogonal transformations and, (iii) is due to the following sequence of inequalities, based on the hypothesis of the lemma:

$$\|U\|_2 - \|U_r^\star\|_2 \leq \|U - U_r^\star R_U^\star\|_2 \leq \textsc{Dist}(U, U_r^\star) \leq \rho\sigma_r(U_r^\star)$$

and thus $\|U\|_2 \leq (1+\rho)\cdot\|U_r^\star\|_2$. The final inequality (iv) follows from the hypothesis of the lemma. ■

The following lemma connects the spectrum of U to $U_r^\star$ under the initialization assumptions.

Lemma 19 *Let U and $U_r^\star$ be $n\times r$ matrices such that $\textsc{Dist}(U, U_r^\star) \leq \rho\sigma_r(U_r^\star)$, for $\rho = \frac{1}{100}\frac{\sigma_r(X^\star)}{\sigma_1(X^\star)}$. Also, define $X_r^\star = U_r^\star U_r^{\star\top}$. Then, the following bounds hold true:*

$$(1 - 1/100)\,\sigma_1(U_r^\star) \leq \sigma_1(U) \leq (1 + 1/100)\,\sigma_1(U_r^\star),$$

$$(1 - 1/100)\,\sigma_r(U_r^\star) \leq \sigma_r(U) \leq (1 + 1/100)\,\sigma_r(U_r^\star).$$

Moreover, by definition of $\tau(V) := \frac{\sigma_r(V)}{\sigma_1(V)}$ for some V matrix, we also observe:

$$\tau(U) \leq \tfrac{101}{99}\cdot\tau(U_r^\star) \quad \textit{and} \quad \tau(X) \leq \left(\tfrac{101}{99}\right)^2\cdot\tau(X_r^\star).$$

Proof Using the norm ordering $\|\cdot\|_2 \leq \|\cdot\|_F$ and the Weyl's inequality for perturbation of singular values (Theorem 3.3.16 Horn and Johnson (1991)) we get,

$$|\sigma_i(U) - \sigma_i(U_r^\star)| \leq \tfrac{1}{100\tau(X^\star)}\sigma_r(U^\star),\ 1 \leq i \leq r.$$

Then, the first two inequalities of the lemma follow by using triangle inequality and the above bound. For the last two inequalities, it is easy to derive bounds on condition numbers by combining the first two inequalities. *Viz.*,

$$\tau\left(U\right) = \tfrac{\sigma_1(U)}{\sigma_r(U)} \leq \tfrac{1+1/100}{1-1/100}\cdot\tfrac{\sigma_1(U_r^\star)}{\sigma_r(U_r^\star)} \leq \tfrac{101}{99}\cdot\tau\left(U_r^\star\right),$$

while the last bound can be easily derived since $\tau\left(U_r^\star\right) = \sqrt{\tau\left(X_r^\star\right)}$. ■

The following lemma shows that DIST, in the factor U space, lower bounds the Frobenius norm distance in the matrix X space.

Lemma 20 *Let $X = UU^\top$ and $X_r^\star = U_r^\star U_r^{\star\top}$ be two rank-r PSD matrices. Let $\textsc{Dist}(U, U_r^\star) \leq \rho\sigma_r\left(U_r^\star\right)$, for $\rho = \frac{1}{100}\frac{\sigma_r(X^\star)}{\sigma_1(X^\star)}$. Then,*

$$\|X - X_r^\star\|_F^2 \geq \tfrac{3\sigma_r(X^\star)}{4}\textsc{Dist}(U, U_r^\star)^2.$$

Proof This proof largely follows the arguments for Lemma 5.4 in Tu et al. (2015), from which we know that

$$||X - X_r^\star||_F^2 \geq 2(\sqrt{2}-1)\sigma_r(X^\star)\textsc{Dist}(U, U_r^\star)^2. \tag{21}$$

Hence, $\|X - X_r^\star\|_F^2 \geq \frac{3\sigma_r(X^\star)}{4}\textsc{Dist}(U, U_r^\star)^2$, for the given value of ρ. ■

The following lemma shows equivalence between various step sizes used in the proofs.

Lemma 21 *Let $X^0 = U^0U^{0\top}$ and $X = UU^\top$ be two $n \times n$ rank-r PSD matrices such that $\textsc{Dist}(U, U_r^\star) \leq \textsc{Dist}(U^0, U_r^\star) \leq \rho\sigma_r(U_r^\star)$, where $\rho = \frac{1}{100} \cdot \frac{\sigma_r(X^\star)}{\sigma_1(X^\star)}$. Define the following step sizes:*

(i) $\eta = \frac{1}{16(M\|X^0\|_2 + \|\nabla f(X^0)\|_2)}$,

(ii) $\widehat{\eta} = \frac{1}{16(M\|X\|_2 + \|\nabla f(X)Q_UQ_U^\top\|_2)}$, *and*

(iii) $\eta^\star = \frac{1}{16(M\|X^\star\|_2 + \|\nabla f(X^\star)\|_2)}$.

Then, $\widehat{\eta} \geq \frac{5}{6}\eta$ holds. Moreover, assuming $\|X^\star - X_r^\star\|_F \leq \frac{\sigma_r(X^\star)}{100}\sqrt{\frac{\sigma_r(X^\star)}{\sigma_1(X^\star)}}$, the following inequalities hold:

$$\tfrac{10}{11}\eta^\star \leq \eta \leq \tfrac{11}{10}\eta^\star$$

Proof By the assumptions of this lemma and based on Lemma 19, we have, $^{98}/_{100}\,\|X^\star\|_2 \leq \|X^0\|_2 \leq {}^{103}/_{100}\,\|X^\star\|_2$; similarly $^{98}/_{100}\,\|X^\star\|_2 \leq \|X\|_2 \leq {}^{103}/_{100}\,\|X^\star\|_2$. Hence, we can combine these two set of inequalities to obtain bounds between X^0 and X, as follows:

$$\tfrac{98}{103}\left\|X^0\right\|_2 \leq \|X\|_2 \leq \tfrac{103}{98}\left\|X^0\right\|_2.$$

To prove the desiderata, we show the relationship between the gradient terms $\|\nabla f(X)Q_UQ_U^\top\|_2$, $\|\nabla f(X^0)\|_2$ and $\|\nabla f(X_r^\star)\|_2$. In particular, for the case $\widehat{\eta} \geq \frac{5}{6}\eta$, we have:

$$\begin{aligned}
\|\nabla f(X)Q_UQ_U^\top\|_2 \leq \|\nabla f(X)\|_2 &\overset{(i)}{\leq} \|\nabla f(X) - \nabla f(X^0)\|_2 + \|\nabla f(X^0)\|_2 \\
&\overset{(ii)}{\leq} M\|X - X^0\|_F + \|\nabla f(X^0)\|_2 \\
&\overset{(iii)}{\leq} M\|X - X_r^\star\|_F + M\|X^0 - X_r^\star\|_F + \|\nabla f(X^0)\|_2 \\
&\overset{(iv)}{\leq} 2M(2+\rho)\rho\|U_r^\star\|_2 \cdot \sigma_r(U_r^\star) + \|\nabla f(X^0)\|_2 \\
&\overset{(v)}{\leq} 2M \cdot (2 + \tfrac{1}{100}) \cdot \tfrac{1}{100}\|X^\star\|_2 + \|\nabla f(X^0)\|_2 \\
&\leq \tfrac{M}{20}\|X^0\|_2 + \|\nabla f(X^0)\|_2
\end{aligned}$$

where (i) follows from the triangle inequality, (ii) is due to the smoothness assumption, (iii) is due to the triangle inequality, (iv) follows by applying Lemma 18 on the first two terms on the right hand

side and, (v) is due to the fact $\|U_r^\star\|_2 \cdot \sigma_r(U_r^\star) \leq \|X^\star\|_2$ and by substituting $\rho = \frac{1}{100} \cdot \frac{\sigma_r(X^\star)}{\sigma_1(X^\star)} \leq \frac{1}{100}$. Last inequality follows from $^{98}/_{100}\,\|X^\star\|_2 \leq \left\|X^0\right\|_2$. Hence, using the above bounds in step size selection, we get

$$\widehat{\eta} = \frac{1}{16(M\|X\|_2 + \|\nabla f(X) Q_U Q_U^\top\|_2)} \overset{(i)}{\geq} \frac{1}{16\left(\frac{6M}{5}\|X^0\|_2 + \|\nabla f(X^0)\|_2\right)} \geq \frac{5}{6}\eta,$$

where (i) is based also on the bound $\|X\|_2 \leq \frac{103}{98}\|X^0\|_2$.

Similarly we show the bound $\frac{10}{11}\eta^\star \leq \eta \leq \frac{11}{10}\eta^\star$. First observe that,

$$\begin{aligned}
\|\nabla f(X^0)\|_2 &\leq \|\nabla f(X_r^\star) - \nabla f(X^0)\|_2 + \|\nabla f(X_r^\star)\|_2 \\
&\leq M\|X_r^\star - X^0\|_F + \|\nabla f(X_r^\star)\|_2 \\
&\overset{(i)}{\leq} M\|X_r^\star - X^0\|_F + M\|X^\star - X_r^\star\|_F + \|\nabla f(X^\star)\|_2 \\
&\leq M(2+\rho)\rho \cdot \|U_r^\star\|_2 \cdot \sigma_r(U_r^\star) + \frac{1}{100} M\sigma_r(X_r^\star) + \|\nabla f(X^\star)\|_2 \\
&\leq \frac{4}{100} M\|X^\star\|_2 + \|\nabla f(X^\star)\|_2.
\end{aligned}$$

Combining the above bound with $^{98}/_{100}\,\|X_r^\star\|_2 \leq \left\|X^0\right\|_2 \leq {}^{103}/_{100}\,\|X_r^\star\|_2$ gives, $\eta \geq \frac{10}{11}\eta^\star$. Similarly we can show the other bounds. ■

Appendix B. Main lemmas for the restricted strong convex case

In this section, we present proofs for the main lemmas used in the proof of Theorem 6, in Section 6.

B.1. Proof of Lemma 15

Here, we prove the existence of a non-trivial lower bound for $\langle \nabla f(X), X - X_r^\star \rangle$. Our proof differs from the standard convex gradient descent proof (see Nesterov (2004)), as we need to analyze updates without any projections. Our proof technique constructs a pseudo-iterate to obtain a bigger lower bound than the error term in Lemma 16. Here, the nature of the step size plays a key role in achieving the bound.

Let us abuse our notation and define $U^+ = U - \widehat{\eta}\nabla f(X)U$ and $X^+ = U^+U^{+\top}$. Observe that we use the surrogate step size $\widehat{\eta}$, where according to Lemma 21 satisfies $\widehat{\eta} \geq \frac{5}{6}\eta$. By smoothness of f, we get:

$$\begin{aligned}
f(X) &\geq f(X^+) - \langle \nabla f(X), X^+ - X \rangle - \tfrac{M}{2}\|X^+ - X\|_F^2 \\
&\overset{(i)}{\geq} f(X^\star) - \langle \nabla f(X), X^+ - X \rangle - \tfrac{M}{2}\|X^+ - X\|_F^2, \qquad (22)
\end{aligned}$$

where (i) follows from optimality of $X^\star$ and since X^+ is a feasible point ($X^+ \succeq 0$) for problem (1). Further, note that $X_r^\star$ is a PSD feasible point. By smoothness of f, we also get

$$\begin{aligned}
f(X_r^\star) &\leq f(X^\star) + \langle \nabla f(X^\star), X_r^\star - X^\star \rangle + \tfrac{M}{2}\|X_r^\star - X^\star\|_F^2 \\
&\overset{(i)}{=} f(X^\star) + \tfrac{M}{2}\|X_r^\star - X^\star\|_F^2, \qquad (23)
\end{aligned}$$

where (i) is due to KKT conditions Boyd and Vandenberghe (2004): since $\nabla f(X^\star)$ is orthogonal to $X^\star$, it is also orthogonal to the $n-r$ bottom eigenvectors of $X^\star$. *Viz.*, $\langle \nabla f(X^\star), X_r^\star - X^\star \rangle = 0$. Finally, since $\text{rank}(X_r^\star) = r$, by the (m, r)-restricted strong convexity of f, we get,

$$f(X_r^\star) \geq f(X) + \langle \nabla f(X), X_r^\star - X \rangle + \tfrac{m}{2} \left\| X_r^\star - X \right\|_F^2. \tag{24}$$

Combining equations (22), (23), and (24), we obtain:

$$\langle \nabla f(X), X - X_r^\star \rangle \geq \langle \nabla f(X), X - X^+ \rangle - \tfrac{M}{2} \left\| X^+ - X \right\|_F^2 + \tfrac{m}{2} \left\| X_r^\star - X \right\|_F^2 - \tfrac{M}{2} \| X_r^\star - X^\star \|_F^2. \tag{25}$$

It is easy to verify that $X^+ = X - \widehat{\eta} \nabla f(X) X \Lambda - \widehat{\eta} \Lambda^\top X \nabla f(X)$, where $\Lambda = I - \frac{\widehat{\eta}}{2} Q_U Q_U^\top \nabla f(X) \in \mathbb{R}^{n \times n}$. Notice that, for step size $\widehat{\eta}$, we have

$$\Lambda \succ 0, \quad \|\Lambda\|_2 \leq 1 + 1/32, \quad \text{and} \quad \sigma_n(\Lambda) \geq 1 - 1/32.$$

Substituting the above in (25), we obtain:

$$\begin{aligned}
&\langle \nabla f(X), X - X_r^\star \rangle - \tfrac{m}{2} \left\| X_r^\star - X \right\|_F^2 + \tfrac{M}{2} \| X_r^\star - X^\star \|_F^2 \\
&\qquad \overset{(i)}{\geq} 2\widehat{\eta} \langle \nabla f(X), \nabla f(X) X \Lambda \rangle - \tfrac{M}{2} \left\| 2\widehat{\eta} \nabla f(X) X \Lambda \right\|_F^2 \\
&\qquad = 2\widehat{\eta} \operatorname{Tr}(\nabla f(X) \nabla f(X) X \Lambda) - 2M\widehat{\eta}^2 \left\| \nabla f(X) X \Lambda \right\|_F^2 \\
&\qquad \overset{(ii)}{\geq} 2\widehat{\eta} \operatorname{Tr}(\nabla f(X) \nabla f(X) X) \cdot \sigma_n(\Lambda) - 2M\widehat{\eta}^2 \left\| \nabla f(X) U \right\|_F^2 \|U\|_2^2 \|\Lambda\|_2^2 \\
&\qquad \geq \tfrac{31 \cdot \widehat{\eta}}{16} \|\nabla f(X) U\|_F^2 - 2M\widehat{\eta}^2 \cdot \left(\tfrac{33}{32}\right)^2 \cdot \left\| \nabla f(X) U \right\|_F^2 \|U\|_2^2 \\
&\qquad = \tfrac{31 \cdot \widehat{\eta}}{16} \widehat{\eta} \|\nabla f(X) U\|_F^2 \left(1 - 2M\widehat{\eta} \left(\tfrac{33}{32}\right)^2 \cdot \tfrac{16}{31} \cdot \|X\|_2\right) \\
&\qquad \overset{(iii)}{\geq} \tfrac{18\widehat{\eta}}{10} \|\nabla f(X) U\|_F^2,
\end{aligned}$$

where (i) follows from symmetry of $\nabla f(X)$ and X, and (ii) follows from

$$\begin{aligned}
\operatorname{Tr}(\nabla f(X) \nabla f(X) X \Lambda) &= \operatorname{Tr}(\nabla f(X) \nabla f(X) U U^\top) - \frac{\eta}{2} \operatorname{Tr}(\nabla f(X) \nabla f(X) U U^\top \nabla f(X)) \\
&\geq (1 - \frac{\eta}{2} \|Q_U Q_U^\top \nabla f(X)\|_2) \|\nabla f(X) U\|_F^2 \\
&\geq (1 - 1/32) \|\nabla f(X) U\|_F^2.
\end{aligned}$$

Finally, (iii) follows by observing that $\widehat{\eta} \leq \frac{1}{16M\|X\|_2}$. Thus, we achieve the desiderata:

$$\langle \nabla f(X), X - X_r^\star \rangle \geq \tfrac{18\widehat{\eta}}{10} \|\nabla f(X) U\|_F^2 + \tfrac{m}{2} \left\| X_r^\star - X \right\|_F^2 - \tfrac{M}{2} \| X^\star - X_r^\star \|_F^2.$$

This completes the proof.

B.2. Proof of Lemma 16

We lower bound $\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle$ as follows:

$$\begin{aligned}
\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle &\overset{(i)}{=} \left\langle Q_\Delta Q_\Delta^\top \nabla f(X), \Delta\Delta^\top \right\rangle \\
&\geq -\left| \mathrm{Tr}\left(Q_\Delta Q_\Delta^\top \nabla f(X) \Delta\Delta^\top \right) \right| \\
&\overset{(ii)}{\geq} -\|Q_\Delta Q_\Delta^\top \nabla f(X)\|_2 \,\mathrm{Tr}(\Delta\Delta^\top) \\
&\overset{(iii)}{\geq} -\left(\|Q_U Q_U^\top \nabla f(X)\|_2 + \|Q_{U_r^\star} Q_{U_r^\star}^\top \nabla f(X)\|_2 \right) \textsc{Dist}(U, U_r^\star)^2. \qquad (26)
\end{aligned}$$

Note that (i) follows from the fact $\Delta = Q_\Delta Q_\Delta^\top \Delta$ and (ii) follows from $|\mathrm{Tr}(AB)| \leq \|A\|_2 \mathrm{Tr}(B)$, for PSD matrix B (Von Neumann's trace inequality Mirsky (1975)). For the transformation in (iii), we use that fact that the column space of Δ, $\textsc{Span}(\Delta)$, is a subset of $\textsc{Span}(U \cup U_r^\star)$, as Δ is a linear combination of U and $U_r^\star R_U^\star$.

To bound the first term in equation (26), we observe:

$$\begin{aligned}
&\|Q_U Q_U^\top \nabla f(X)\|_2 \cdot \textsc{Dist}(U, U_r^\star)^2 \qquad (27) \\
&\quad \overset{(i)}{=} \widehat{\eta} \cdot 16 \left(M\|X\|_2 + \|Q_U Q_U^\top \nabla f(X)\|_2 \right) \cdot \|Q_U Q_U^\top \nabla f(X)\|_2 \cdot \textsc{Dist}(U, U_r^\star)^2 \\
&\quad = \widehat{\eta} \left(16 \underbrace{M\|X\|_2 \|Q_U Q_U^\top \nabla f(X)\|_2 \cdot \textsc{Dist}(U, U_r^\star)^2}_{:=A} + 16 \|Q_U Q_U^\top \nabla f(X)\|_2^2 \cdot \textsc{Dist}(U, U_r^\star)^2 \right)
\end{aligned}$$

At this point, we desire to introduce strong convexity parameter m and condition number κ in our bound. In particular, to bound term A, we observe that $\|Q_U Q_U^\top \nabla f(X)\|_2 \leq \frac{m\sigma_r(X)}{40}$ or $\|Q_U Q_U^\top \nabla f(X)\|_2 \geq \frac{m\sigma_r(X)}{40}$. This results into bounding A as follows:

$$\begin{aligned}
&M\|X\|_2 \|Q_U Q_U^\top \nabla f(X)\|_2 \cdot \textsc{Dist}(U, U_r^\star)^2 \\
&\quad \leq \max\left\{ \tfrac{16\cdot\widehat{\eta}\cdot M\|X\|_2 \cdot m\sigma_r(X)}{40} \cdot \textsc{Dist}(U, U_r^\star)^2,\ \widehat{\eta} \cdot 16 \cdot 40\kappa\tau(X) \|Q_U Q_U^\top \nabla f(X)\|_2^2 \cdot \textsc{Dist}(U, U_r^\star)^2 \right\} \\
&\quad \leq \tfrac{16\cdot\widehat{\eta}\cdot M\|X\|_2 \cdot m\sigma_r(X)}{40} \cdot \textsc{Dist}(U, U_r^\star)^2 + \widehat{\eta} \cdot 16 \cdot 40\kappa\tau(X) \|Q_U Q_U^\top \nabla f(X)\|_2^2 \cdot \textsc{Dist}(U, U_r^\star)^2.
\end{aligned}$$

Combining the above inequalities, we obtain:

$$\begin{aligned}
&\|Q_U Q_U^\top \nabla f(X)\|_2 \cdot \textsc{Dist}(U, U_r^\star)^2 \qquad (28) \\
&\qquad \overset{(i)}{\leq} \tfrac{m\sigma_r(X)}{40} \cdot \textsc{Dist}(U, U_r^\star)^2 + (40\kappa\tau(X) + 1) \cdot 16 \cdot \widehat{\eta} \|Q_U Q_U^\top \nabla f(X)\|_2^2 \cdot \textsc{Dist}(U, U_r^\star)^2 \\
&\qquad \overset{(ii)}{\leq} \tfrac{m\sigma_r(X)}{40} \cdot \textsc{Dist}(U, U_r^\star)^2 + (41\kappa\tau(X_r^\star) + 1) \cdot 16 \cdot \widehat{\eta} \|Q_U Q_U^\top \nabla f(X)\|_2^2 \cdot (\rho')^2 \sigma_r(X_r^\star) \\
&\qquad \overset{(iii)}{\leq} \tfrac{m\sigma_r(X)}{40} \cdot \textsc{Dist}(U, U_r^\star)^2 + 16 \cdot 42 \cdot \widehat{\eta} \cdot \kappa\tau(X_r^\star) \cdot \|\nabla f(X) U\|_F^2 \cdot \tfrac{11(\rho')^2}{10} \\
&\qquad \overset{(iv)}{\leq} \tfrac{m\sigma_r(X)}{40} \cdot \textsc{Dist}(U, U_r^\star)^2 + \tfrac{2\widehat{\eta}}{25} \cdot \|\nabla f(X) U\|_F^2, \qquad (29)
\end{aligned}$$

where (i) follows from $\widehat{\eta} \leq \frac{1}{16M\|X\|_2}$, (ii) is due to Lemma 19 and bounding $\textsc{Dist}(U, U_r^\star) \leq \rho'\sigma_r(U_r^\star)$ by the hypothesis of the lemma, (iii) is due to $\sigma_r(X^\star) \leq 1.1\sigma_r(X)$ by Lemma 19 and due to the facts $\sigma_r(X)\|Q_U Q_U^\top \nabla f(X)\|_2^2 \leq \|U^\top \nabla f(X)\|_F^2$ and $(41\kappa\tau(X_r^\star) + 1) \leq 42\kappa\tau(X_r^\star)$. Finally, (iv) follows from substituting ρ' and using Lemma 19.

Next, we bound the second term in equation (26):

$$
\begin{aligned}
&\|Q_{U^\star R} Q_{U^\star R}^\top \nabla f(X)\|_2 \cdot \textsc{Dist}(U, U_r^\star)^2 \qquad (30)\\
&\overset{(i)}{\leq} \|\nabla f(X) - \nabla f(X^\star)\|_2 \cdot \textsc{Dist}(U, U_r^\star)^2\\
&\leq \|\nabla f(X) - \nabla f(X^\star)\|_F \cdot \textsc{Dist}(U, U_r^\star)^2\\
&\overset{(ii)}{\leq} M\left(\|X - X_r^\star\|_F + \|X^\star - X_r^\star\|_F\right) \cdot \textsc{Dist}(U, U_r^\star)^2\\
&\overset{(iii)}{\leq} M(2+\rho') \cdot \rho' \cdot \|U_r^\star\|_2 \cdot \sigma_r(U_r^\star) \cdot \textsc{Dist}(U, U_r^\star)^2 + M\|X^\star - X_r^\star\|_F \cdot \textsc{Dist}(U, U_r^\star)^2\\
&\overset{(iv)}{\leq} M(2+\rho')\|U_r^\star\|_2 \tfrac{1}{100\kappa\tau(U_r^\star)}\sigma_r(U_r^\star) \cdot \textsc{Dist}(U, U_r^\star)^2 + M\|X^\star - X_r^\star\|_F \cdot \textsc{Dist}(U, U_r^\star)^2\\
&\leq \tfrac{m\sigma_r(X^\star)}{40}\textsc{Dist}(U, U_r^\star)^2 + M\|X^\star - X_r^\star\|_F \cdot \textsc{Dist}(U, U_r^\star)^2, \qquad (31)
\end{aligned}
$$

where (i) follows from $\nabla f(X^\star)X^\star = 0$, (ii) is due to smoothness of f and (iii) follows from Lemma 18. Finally (iv) follows from $\textsc{Dist}(U, U_r^\star) \leq \rho'\sigma_r(U_r^\star)$ and substituting $\rho' = \frac{1}{100\kappa\tau(U_r^\star)}$.

Substituting (29), (31) in (26), we get:

$$
\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \geq -\left(\tfrac{2\widehat{\eta}}{25}\|\nabla f(X)U\|_F^2 + \tfrac{m\sigma_r(X^\star)}{20} \cdot \textsc{Dist}(U, U_r^\star)^2 + M\|X^\star - X_r^\star\|_F \cdot \textsc{Dist}(U, U_r^\star)^2\right)
$$

This completes the proof.

B.3. Proof of Lemma 14

Recall $U^+ = U - \eta\nabla f(X)U$. First we rewrite the inner product as shown below.

$$
\begin{aligned}
\frac{1}{\eta}\left\langle U - U^+, U - U_r^\star R_U^\star \right\rangle &= \left\langle \nabla f(X)U, U - U_r^\star R_U^\star \right\rangle\\
&= \left\langle \nabla f(X), X - U_r^\star R_U^\star U^\top \right\rangle\\
&= \frac{1}{2}\left\langle \nabla f(X), X - X_r^\star \right\rangle + \left\langle \nabla f(X), \frac{1}{2}(X + X_r^\star) - U_r^\star R_U^\star U^\top \right\rangle\\
&= \frac{1}{2}\left\langle \nabla f(X), X - X_r^\star \right\rangle + \frac{1}{2}\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle, \qquad (32)
\end{aligned}
$$

which follows by adding and subtracting $\frac{1}{2}X_r^\star$.

Let, $\widehat{\eta} = \frac{1}{16(M\|X\|_2 + \|\nabla f(X)Q_U Q_U^\top\|_2)}$. Using Lemmas 15 and 16, we have:

$$
\begin{aligned}
&\left\langle \nabla f(X)U, U - U_r^\star R_U^\star \right\rangle\\
&\geq \tfrac{9\widehat{\eta}}{10} \cdot \|\nabla f(X)U\|_F^2 + \tfrac{m}{4}\|X - X_r^\star\|_F^2 - \tfrac{M}{4}\|X^\star - X_r^\star\|_F^2\\
&\qquad - \tfrac{1}{2}\left(\tfrac{2\widehat{\eta}}{25} \cdot \|\nabla f(X)U\|_F^2 + \tfrac{m\sigma_r(X^\star)}{20} \cdot \textsc{Dist}(U, U_r^\star)^2 + M\|X^\star - X_r^\star\|_F \cdot \textsc{Dist}(U, U_r^\star)^2\right)\\
&= \left(\tfrac{9}{10} - \tfrac{1}{25}\right) \cdot \widehat{\eta}\|\nabla f(X)U\|_F^2 - \tfrac{M}{4}\|X^\star - X_r^\star\|_F^2\\
&\qquad + \tfrac{m}{4}\left(\|X - X_r^\star\|_F^2 - \tfrac{4\sigma_r(X^\star)}{25} \cdot \textsc{Dist}(U, U_r^\star)^2 - 2\kappa \cdot \|X^\star - X_r^\star\|_F \cdot \textsc{Dist}(U, U_r^\star)^2\right)
\end{aligned}
$$

$$\overset{(i)}{\geq} \tfrac{4\widehat{\eta}}{5} \cdot \|\nabla f(X)U\|_F^2 - \tfrac{M}{4}\|X^\star - X_r^\star\|_F^2 + \tfrac{m}{4}\left(\|X - X_r^\star\|_F^2 - \tfrac{4\sigma_r(X^\star)}{25} \cdot \textsc{Dist}(U, U_r^\star)^2 - \tfrac{\sigma_r(X^\star)}{50} \cdot \textsc{Dist}(U, U_r^\star)^2\right)$$
$$\overset{(ii)}{\geq} \tfrac{4\widehat{\eta}}{5} \cdot \|\nabla f(X)U\|_F^2 + \tfrac{3m}{20} \cdot \sigma_r(X^\star) \cdot \textsc{Dist}(U, U_r^\star)^2 - \tfrac{M}{4}\|X^\star - X_r^\star\|_F^2$$

where (i) follows from $\|X^\star - X_r^\star\| \leq \frac{\sigma_r(X^\star)}{100\kappa^{1.5}} \frac{\sigma_r(X^\star)}{\sigma_1(X^\star)} \leq \frac{\sigma_r(X^\star)}{100\kappa^{1.5}} \leq \frac{\sigma_r(X^\star)}{100\kappa}$ and (ii) follows from Lemma 20. Finally the result follows from $\widehat{\eta} \geq \frac{5}{6}\eta$ from Lemma 21.

Appendix C. Main lemmas for the smooth case

In this section, we present the main lemmas, used in the proof of Theorem 5 in Section 6. First, we present a lemma bounding the error term $\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle$, that appears in eq. (18).

Lemma 22 *Let f be M-smooth and $X = UU^\top$; also, define $\Delta := U - U_r^\star R_U^\star$. Then, for $\textsc{Dist}(U, U_r^\star) \leq \rho\sigma_r(U_r^\star)$ and $\rho = \frac{1}{100}\frac{\sigma_r(X^\star)}{\sigma_1(X^\star)}$, the following bound holds true:*

$$\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \leq \tfrac{1}{40}\|\nabla f(X)U\|_2 \cdot \textsc{Dist}(U, U_r^\star).$$

Proof By the Von Neumann's trace inequality for PSD matrices, we know that $\mathrm{Tr}(AB) \leq \mathrm{Tr}(A) \cdot \|B\|_2$, for A PSD matrix. In our context, we then have:

$$\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \leq \|\nabla f(X) Q_\Delta Q_\Delta^\top\|_2 \cdot \mathrm{Tr}(\Delta\Delta^\top)$$
$$\overset{(i)}{\leq} \left(\|\nabla f(X) Q_U Q_U^\top\|_2 + \|\nabla f(X) Q_{U_r^\star} Q_{U_r^\star}^\top\|_2\right) \cdot \textsc{Dist}(U, U_r^\star)^2, \tag{33}$$

where, (i) is because Δ can be decomposed into the column span of U and $U_r^\star$, and the orthogonality of the matrix $R_{U_r^\star}$. In sequence, we further bound the term $\|\nabla f(X) Q_{U_r^\star} Q_{U_r^\star}^\top\|_2$ as follows:

$$\|\nabla f(X) U_r^\star\|_2 \overset{(i)}{\leq} \|\nabla f(X)U\|_2 + \|\nabla f(X)\Delta\|_2$$
$$\overset{(ii)}{\leq} \|\nabla f(X)U\|_2 + \|\nabla f(X) Q_\Delta Q_\Delta^\top\|_2 \|\Delta\|_2$$
$$\overset{(iii)}{\leq} \|\nabla f(X)U\|_2 + \left(\|\nabla f(X) Q_U Q_U^\top\|_2 + \|\nabla f(X) Q_{U_r^\star} Q_{U_r^\star}^\top\|_2\right) \|\Delta\|_2$$
$$\overset{(iv)}{\leq} \|\nabla f(X)U\|_2 + \left(\|\nabla f(X) Q_U Q_U^\top\|_2 + \|\nabla f(X) Q_{U_r^\star} Q_{U_r^\star}^\top\|_2\right) \tfrac{1}{100}\sigma_r(U_r^\star)$$
$$\overset{(v)}{\leq} \|\nabla f(X)U\|_2 + \tfrac{1}{\left(1 - \frac{1}{100}\right)} \cdot \tfrac{1}{100}\|\nabla f(X)U\|_2 + \tfrac{1}{100}\|\nabla f(X)U_r^\star\|_2$$
$$\leq \tfrac{102}{100}\|\nabla f(X)U\|_2 + \tfrac{1}{100}\|\nabla f(X)U_r^\star\|_2.$$

where (i) is due to triangle inequality on $U_r^\star R_U^\star = U - \Delta$, (ii) is due to generalized Cauchy-Schwarz inequality; we denote as $Q_\Delta Q_\Delta$ the projection matrix on the column span of Δ matrix, (iii) is due to triangle inequality and the fact that the column span of Δ can be decomposed into the column span of U and $U_r^\star$, by construction of Δ, (iv) is due to

$$\|\Delta\|_2 \leq \textsc{Dist}(U, U_r^\star) \leq \tfrac{1}{100}\tfrac{\sigma_r(X^\star)}{\sigma_1(X^\star)} \cdot \sigma_r(U_r^\star) \leq \tfrac{1}{100} \cdot \sigma_r(U_r^\star).$$

Finally, (v) is due to the facts:

$$\|\nabla f(X)U_r^\star\|_2 = \|\nabla f(X)Q_{U_r^\star}Q_{U_r^\star}^\top U_r^\star\|_2 \geq \|\nabla f(X)Q_{U_r^\star}Q_{U_r^\star}^\top\|_2 \cdot \sigma_r(U_r^\star),$$

and

$$\begin{aligned}\|\nabla f(X)U\|_2 = \|\nabla f(X)Q_UQ_U^\top U\|_2 &\geq \|\nabla f(X)Q_UQ_U^\top\|_2 \cdot \sigma_r(U) \\ &\geq \|\nabla f(X)Q_UQ_U^\top\|_2 \cdot \left(1 - \tfrac{1}{100}\right) \cdot \sigma_r(U_r^\star),\end{aligned}$$

by Lemma 19. Thus:

$$\begin{aligned}\|\nabla f(X)Q_{U_r^\star}Q_{U_r^\star}^\top\|_2 &\leq \tfrac{1}{\sigma_r(U_r^\star)}\|\nabla f(X)U_r^\star\|_2 \\ &\leq \tfrac{1}{\sigma_r(U_r^\star)}\tfrac{102}{99}\|\nabla f(X)U\|_2 \\ &\leq \tfrac{101\sigma_1(U_r^\star)}{100\sigma_r(U_r^\star)}\tfrac{102}{99}\|\nabla f(X)Q_UQ_U^\top\|_2, \end{aligned} \tag{34}$$

and, combining with (33), we get

$$\begin{aligned}\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle &\leq \left(\tfrac{102\cdot 101}{100\cdot 99} + 1\right) \cdot \tfrac{\sigma_1(U_r^\star)}{\sigma_r(U_r^\star)} \cdot \|\nabla f(X)Q_UQ_U^\top\|_2 \cdot \textsc{Dist}(U, U_r^\star)^2 \\ &\leq \tfrac{1}{40}\|\nabla f(X)U\|_2 \cdot \textsc{Dist}(U, U_r^\star).\end{aligned}$$

The last inequality follows from $\textsc{Dist}(U, U_r^\star) \leq \frac{1}{100}\frac{\sigma_r(X^\star)}{\sigma_1(X^\star)} \cdot \sigma_r(U_r^\star)$. This completes the proof. ■

The following lemma lower bounds the term $\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle$; this result is used later in the proof of Lemma 25.

Lemma 23 *Let $X = UU^\top$ and define $\Delta := U - U_r^\star R_U^\star$. Let $f(X^+) \geq f(X_r^\star)$, where $X_r^\star$ is the optimum of the problem (1). Then, for $\textsc{Dist}(U, U_r^\star) \leq \rho\sigma_r(U_r^\star)$, where $\rho = \frac{1}{100}\frac{\sigma_r(X^\star)}{\sigma_1(X^\star)}$, and f being a M-smooth convex function, the following lower bound holds:*

$$\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \geq -\tfrac{\sqrt{2}}{\sqrt{2}-\frac{1}{100}} \cdot \tfrac{1}{100} \cdot \left|\langle \nabla f(X), X - X_r^\star\rangle\right|.$$

Proof Let the QR factorization of the matrix $[U \quad U_r^\star R_U^\star]_{n\times 2r}$ be $Q \cdot R$, where Q is a $n \times 2r$ orthonormal matrix and R is a $2r \times 2r$ invertible matrix (since $[U \quad U_r^\star R_U^\star]$ is assumed to be rank-$2r$). Further, let $[U \quad U_r^\star R_U^\star]^\dagger_{2r\times n}$ where $C^\dagger$ denotes the pseudo-inverse of matrix C. It is obvious that $[U \quad U_r^\star R_U^\star]^\top \cdot \left([U \quad U_r^\star R_U^\star]^\dagger\right)^\top = I_{2r\times 2r}$.

Given the above, let us re-define some quantities w.r.t. $[U \quad U_r^\star R_U^\star]$, as follows

$$\Delta = U - U_r^\star R_U^\star = [U \quad U_r^\star R_U^\star]_{n\times 2r} \cdot \begin{bmatrix} I_{r\times r} \\ -I_{r\times r}\end{bmatrix}_{2r\times r}.$$

Moreover, it is straightforward to justify that:

$$X - X_r^\star = [U \quad U_r^\star R_U^\star]_{n\times 2r} \cdot \begin{bmatrix} I_{r\times r} & 0_{r\times r} \\ 0_{r\times r} & -I_{r\times r}\end{bmatrix} \cdot [U \quad U_r^\star R_U^\star]^\top_{2r\times n}$$

Then, from the above, the two quantities $X - X_r^\star$ and Δ are connected as follows:

$$(X - X_r^\star) \cdot \left([U \quad U_r^\star R_U^\star]^\dagger \right)^\top \cdot \begin{bmatrix} I \\ I \end{bmatrix} = [U \quad U_r^\star R_U^\star] \cdot \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \cdot \underbrace{[U \quad U_r^\star R_U^\star]^\top \cdot \left([U \quad U_r^\star R_U^\star]^\dagger \right)^\top}_{=I} \cdot \begin{bmatrix} I \\ I \end{bmatrix}$$

which is equal to Δ. Then, the following sequence of (in)equalities holds true:

$$\begin{aligned}
\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle &\overset{(i)}{=} \left\langle \nabla f(X), (X - X_r^\star) \cdot \left([U \quad U_r^\star R_U^\star]^\dagger \right)^\top \cdot \begin{bmatrix} I \\ I \end{bmatrix} \cdot \Delta^\top \right\rangle \\
&\overset{(ii)}{\geq} - \left| \mathrm{Tr} \left(\underbrace{\nabla f(X) \cdot (X - X_r^\star)}_{=A} \cdot \underbrace{\left([U \quad U_r^\star R_U^\star]^\dagger \right)^\top \cdot \begin{bmatrix} I \\ I \end{bmatrix} \cdot \Delta^\top}_{=B} \right) \right| \\
&\overset{(iii)}{\geq} - \left| \mathrm{Tr} \left(\nabla f(X) \cdot (X - X_r^\star) \right) \right| \cdot \left\| \left([U \quad U_r^\star R_U^\star]^\dagger \right)^\top \cdot \begin{bmatrix} I \\ I \end{bmatrix} \cdot \Delta^\top \right\|_2 \\
&\overset{(iv)}{\geq} - \left| \langle \nabla f(X), X - X_r^\star \rangle \right| \cdot \left\| \left([U \quad U_r^\star R_U^\star]^\dagger \right)^\top \right\|_2 \cdot \left\| \begin{bmatrix} I \\ I \end{bmatrix} \right\|_2 \cdot \|\Delta\|_2 \\
&\overset{(v)}{\geq} -\sqrt{2} \cdot \left| \langle \nabla f(X), X - X_r^\star \rangle \right| \cdot \left\| \left([U \quad U_r^\star R_U^\star]^\dagger \right)^\top \right\|_2 \cdot \tfrac{1}{100} \cdot \sigma_r(U_r^\star) \\
&\overset{(vi)}{\geq} -\sqrt{2} \cdot \left| \langle \nabla f(X), X - X_r^\star \rangle \right| \cdot \tfrac{1}{\sqrt{2} - \frac{1}{100}} \cdot \tfrac{1}{\sigma_r(U_r^\star)} \cdot \tfrac{1}{100} \cdot \sigma_r(U_r^\star) \\
&= -\tfrac{\sqrt{2}}{\sqrt{2} - \frac{1}{100}} \cdot \tfrac{1}{100} \cdot \left| \langle \nabla f(X), X - X_r^\star \rangle \right|, \qquad (35)
\end{aligned}$$

where, (i) follows by substituting Δ, according to the discussion above, (ii) follows from symmetry of $\nabla f(X)$, (iii) follows from the Von Neumann trace inequality $\mathrm{Tr}(AB) \leq \mathrm{Tr}(A)||B||_2$, for a PSD matrix A; next, we show that $y^\top A y \geq 0$, $\forall y$ and $A := \nabla f(X) \cdot (X - X_r^\star)$, (iv) is due to successive application of the Cauchy-Schwarz inequality, (v) is due to $\left\| \begin{bmatrix} I & I \end{bmatrix}^\top \right\|_2 = \sqrt{2}$ and $\|\Delta\|_2 \leq \textsc{Dist}(U, U_r^\star) \leq \rho \cdot \sigma_r(U_r^\star) \leq \frac{1}{100} \cdot \sigma_r(U_r^\star)$, (vi) follows from the the following fact:

$$\begin{aligned}
\frac{1}{\left\| [U \quad U_r^\star R_U^\star]^\dagger \right\|_2} &= \sigma_r \left([U \quad U_r^\star R_U^\star] \right) \\
&= \sigma_r \left([U \quad U_r^\star R_U^\star] - [U_r^\star R_U^\star \quad U_r^\star R_U^\star] + [U_r^\star R_U^\star \quad U_r^\star R_U^\star] \right) \\
&= \sigma_r \left([U - U_r^\star R_U^\star \quad 0] + [U_r^\star R_U^\star \quad U_r^\star R_U^\star] \right) \\
&\overset{(i)}{\geq} \sigma_r \left([U_r^\star R_U^\star \quad U_r^\star R_U^\star] \right) - \|U - U_r^\star R_U^\star\|_2 \\
&\overset{(ii)}{=} \sqrt{2} \cdot \sigma_r \left(U^\star R \right) - \|U - U_r^\star R_U^\star\|_2 \\
&\overset{(iii)}{\geq} \left(\sqrt{2} - \tfrac{1}{100} \right) \cdot \sigma_r(U_r^\star),
\end{aligned}$$

where, (i) follows from a variant of Weyl's inequality, (ii) is due to $\sigma_r \left([U_r^\star R_U^\star \quad U_r^\star R_U^\star] \right) = \sqrt{2} \cdot \sigma_r \left(U_r^\star \right)$, (iii) follows from the assumption that $\|U - U_r^\star R_U^\star\|_2 \leq \textsc{Dist}(U, U_r^\star) \leq \frac{1}{100} \cdot \sigma_r \left(U_r^\star \right)$.

The above lead to the inequality:

$$\left\| \left([U \quad U_r^\star R_U^\star]^\dagger \right)^\top \right\|_2 = \left\| [U \quad U_r^\star R_U^\star]^\dagger \right\|_2 \leq \frac{1}{\sqrt{2}-\frac{1}{100}} \cdot \frac{1}{\sigma_r(U_r^\star)}.$$

In the above inequalities (35), we used the fact that symmetric version of A is a PSD matrix, where $A := \nabla f(X)\Delta(U + U_r^\star R_U^\star)^\top = \nabla f(X) \cdot (X - X_r^\star)$ is a PSD matrix, *i.e.*, given a vector y, $y^\top \nabla f(X) \cdot (X - X_r^\star)y \geq 0$. To show this, let $g(t) = f(X + tyy^\top)$ be a function from $\mathbb{R} \to \mathbb{R}$. Hence, $\nabla g(t) = \left\langle \nabla f(X + tyy^\top), yy^\top \right\rangle$. Now, consider g restricted to the level set $\{t : f(X + tyy^\top) \leq f(X)\}$. Note that, since f is convex, this set is convex and further X belongs to this set from the hypothesis of the lemma. Also $f(X_r^\star) \leq f(X + tyy^\top)$, for t in this set from the optimality of $X_r^\star$. Let t^* be the minimizer of $g(t)$ over this set. Then, by convexity of g,

$$\left\langle \nabla f(X), yy^\top \right\rangle \cdot -t^* = \nabla g(0) \cdot (0 - t^*) \geq g(0) - g(t^*) \geq 0.$$

Further, since $g(t^*) = f(X + t^* yy^\top) \geq f(X_r^\star)$, $X + t^* yy^\top - X_r^\star$ is orthogonal to y. Hence, $(X + t^* yy^\top - X_r^\star)y = 0$. Combining this with the above inequality gives, $\left\langle \nabla f(X), (X - X_r^\star)yy^\top \right\rangle \geq 0$. This completes the proof. ■

We next present a lemma for lower bounding the term $\langle \nabla f(X), X - X_r^\star \rangle$. This result is used in the following Lemma 25, where we bound the term $\langle \nabla f(X)U, U - U_r^\star R_U^\star \rangle$.

Lemma 24 *Let f be a M-smooth convex function with optimum point $X_r^\star$. Then, under the assumption that $f(X^+) \geq f(X_r^\star)$, the following holds:*

$$\langle \nabla f(X), X - X_r^\star \rangle \geq \frac{18\widehat{\eta}}{10} \|\nabla f(X)U\|_F^2.$$

Proof The proof follows much like the proof of the Lemma for strong convex case (Lemma 15), except for the arguments used to bound equation (22). For completeness, we here highlight the differences; in particular, we again have by smoothness of f:

$$f(X) \geq f(X^+) - \langle \nabla f(X), X^+ - X \rangle - \frac{M}{2} \|X^+ - X\|_F^2 ,$$

where we consider the same notation with Lemma 15. By the assumptions of the Lemma, we have $f(X^+) \geq f(X_r^\star)$ and, thus, the above translates into:

$$f(X) \geq f(X_r^\star) - \langle \nabla f(X), X^+ - X \rangle - \frac{M}{2} \|X^+ - X\|_F^2 ,$$

hence eliminating the need for equation (23). Combining the above and assuming just smoothness (*i.e.*, the restricted strong convexity parameter is $m = 0$), we obtain a simpler version of eq. (25):

$$\langle \nabla f(X), X - X_r^\star \rangle \geq \langle \nabla f(X), X - X^+ \rangle - \frac{M}{2} \|X^+ - X\|_F^2 . \tag{36}$$

Then, the result easily follows by the same steps in Lemma 15. ■

Next, we state an important result, relating the gradient step in the factored space $U^+ - U$ to the direction to the optimum $U - U^\star$. The result borrows the outcome of Lemmas 22-24.

Lemma 25 *Let $X = UU^\top$ and define $\Delta := U - U_r^\star R_U^\star$. Assume $f(X^+) \geq f(X_r^\star)$ and $\textsc{Dist}(U, U_r^\star) \leq \rho\sigma_r(U_r^\star)$, where $\rho = \frac{1}{100}\frac{\sigma_r(X^\star)}{\sigma_1(X^\star)}$. For f being a M-smooth convex function, the following descent condition holds for the U-space:*

$$\langle \nabla f(X)U, U - U_r^\star R_U^\star \rangle \geq \tfrac{\eta}{2} \cdot \|\nabla f(X)U\|_F^2.$$

Proof Expanding the term $\langle \nabla f(X)U, U - U_r^\star R_U^\star \rangle$, we obtain the equivalent characterization:

$$\begin{aligned}
\langle \nabla f(X)U, U - U_r^\star R_U^\star \rangle &= \left\langle \nabla f(X), X - U_r^\star R_U^\star U^\top \right\rangle \\
&= \tfrac{1}{2}\langle \nabla f(X), X - X_r^\star \rangle + \left\langle \nabla f(X), \tfrac{1}{2}(X + X_r^\star) - U_r^\star R_U^\star U^\top \right\rangle \\
&= \tfrac{1}{2}\langle \nabla f(X), X - X_r^\star \rangle + \tfrac{1}{2}\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \qquad (37)
\end{aligned}$$

which follows by the definition of X and adding and subtracting $\frac{1}{2}X_r^\star$ term. By Lemma 24, we can bound the first term on the right hand side as:

$$\tfrac{1}{2}\langle \nabla f(X), X - X_r^\star \rangle \geq \tfrac{18\widehat{\eta}}{20} \cdot \|\nabla f(X)U\|_F^2. \qquad (38)$$

Observe that $\langle \nabla f(X), X - X_r^\star \rangle \geq 0$. By Lemma 23, we can lower bound the last term on the right hand side of (37) as:

$$\tfrac{1}{2}\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \geq -\tfrac{\sqrt{2}}{\sqrt{2}-\frac{1}{100}} \cdot \tfrac{1}{200}\left|\langle \nabla f(X), X - X_r^\star \rangle\right| = -\tfrac{\sqrt{2}}{\sqrt{2}-\frac{1}{100}} \cdot \tfrac{1}{200}\langle \nabla f(X), X - X_r^\star \rangle. \qquad (39)$$

Combining (38) and (39) in (37), we get:

$$\begin{aligned}
\langle \nabla f(X)U, U - U_r^\star R_U^\star \rangle &\geq \tfrac{1}{2}\langle \nabla f(X), X - X_r^\star \rangle - \tfrac{\sqrt{2}}{\sqrt{2}-\frac{1}{100}} \cdot \tfrac{1}{200}\langle \nabla f(X), X - X_r^\star \rangle \\
&\geq \left(1 - \tfrac{\sqrt{2}}{\sqrt{2}-\frac{1}{100}} \cdot \tfrac{1}{100}\right) \cdot \tfrac{1}{2}\langle \nabla f(X), X - X_r^\star \rangle \\
&\geq \tfrac{98}{100} \cdot \tfrac{18\widehat{\eta}}{20} \cdot \|\nabla f(X)U\|_F^2 \\
&\overset{(i)}{\geq} \tfrac{98}{100} \cdot \tfrac{18}{20} \cdot \tfrac{5\eta}{6} \cdot \|\nabla f(X)U\|_F^2 \\
&\geq \tfrac{7\eta}{10} \cdot \|\nabla f(X)U\|_F^2 \geq \tfrac{\eta}{2} \cdot \|\nabla f(X)U\|_F^2,
\end{aligned}$$

where (i) follows from $\widehat{\eta} \geq \frac{5}{6}\eta$ in Lemma 21. This completes the proof. ■

We conclude this section with a lemma that proves that the distance $\textsc{Dist}(U, U_r^\star)$ is non-increasing per iteration of FGD. This lemma is used in the proof of sublinear convergence of FGD (Theorem 5), in Section 6.

Lemma 26 *Let $X = UU^\top$ and $X^+ = U^+(U^+)^\top$ be the current and next estimate of* FGD. *Assume f is a M-smooth convex function such that $f(X^+) \geq f(X_r^\star)$. Moreover, define $\Delta := U - U_r^\star R_U^\star$ and $\textsc{Dist}(U, U_r^\star) \leq \rho\sigma_r(U_r^\star)$, where $\rho = \frac{1}{100}\frac{\sigma_r(X^\star)}{\sigma_1(X^\star)}$. Then, the following inequality holds:*

$$\textsc{Dist}(U^+, U_r^\star) \leq \textsc{Dist}(U, U_r^\star).$$

This further implies $\textsc{Dist}(U, U_r^\star) \leq \textsc{Dist}(U^0, U_r^\star)$ for any estimate U of FGD.

Proof Let $R_U^\star = \arg\min_{R\in\mathcal{O}} \|U - U_r^\star R\|_F^2$. Expanding $\textsc{Dist}(U^+, U_r^\star)^2$, we obtain:

$$\begin{aligned}
\textsc{Dist}(U^+, U_r^\star)^2 &= \min_{R\in\mathcal{O}} \|U^+ - U_r^\star R\|_F^2 \qquad (40)\\
&\le \|U^+ - U_r^\star R_U^\star\|_F^2\\
&= \|U^+ - U + U - U_r^\star R_U^\star\|_F^2\\
&= \|U^+ - U\|_F^2 + \|U - U_r^\star R_U^\star\|_F^2 - 2\left\langle U^+ - U, U_r^\star R_U^\star - U\right\rangle\\
&= \eta^2\|\nabla f(X)U\|_F^2 + \textsc{Dist}(U, U_r^\star)^2 - 2\eta\left\langle \nabla f(X)U, U - U_r^\star R_U^\star\right\rangle\\
&\le \textsc{Dist}(U, U_r^\star)^2, \qquad (41)
\end{aligned}$$

where last inequality is due to Lemma 25. ■

Appendix D. Initialization proofs

D.1. Proof of Lemma 10

The proof borrows results from standard projected gradient descent. In particular, we know from Theorem 3.6 in Bubeck (2014) that, for consecutive estimates X^+, X and optimal point $X^\star$, projected gradient descent satisfies:

$$\|X^+ - X^\star\|_F^2 \le \left(1 - \tfrac{1}{\kappa}\right)\cdot\|X - X^\star\|_F^2.$$

By taking square root of the above inequality, we further have:

$$\|X^+ - X^\star\|_F \le \sqrt{1 - \tfrac{1}{\kappa}}\cdot\|X - X^\star\|_F \le \left(1 - \tfrac{1}{2\kappa}\right)\cdot\|X - X^\star\|_F, \qquad (42)$$

since $\sqrt{1 - \frac{1}{\kappa}} \le 1 - \frac{1}{2\kappa}$, for all values of $\kappa > 1$.

Given the above, the following (in)equalities hold true:

$$\begin{aligned}
\|X - X^+\|_F &= \|X - X^\star + X^\star - X^+\|_F\\
&\overset{(i)}{\ge} \|X - X^\star\|_F - \|X^+ - X^\star\|_F\\
&\overset{(ii)}{\ge} \frac{1}{2\kappa}\|X - X^\star\|_F \Rightarrow\\
\|X - X^\star\|_F &\le 2\kappa\cdot\|X - X^+\|_F,
\end{aligned}$$

where (i) is due to the lower bound on triangle inequality and (ii) is due to (42). Under the assumptions of the lemma, if $\|X - X^+\|_F \le \frac{c}{\kappa\sqrt{r}\tau(X_r)}\sigma_r(X)$, the above inequality translates into:

$$\|X - X^\star\|_F \le \tfrac{2c}{\sqrt{r}\tau(X_r)}\sigma_r(X).$$

By construction, both X and $X^\star$ are PSD matrices; moreover, X can be a matrix with $\text{rank}(X) > r$. Hence,

$$||X_r - X^\star||_F \le \sqrt{r}||X_r - X^\star||_2 \le 2\sqrt{r}\|X - X^\star\|_2 \le 2\sqrt{r}\|X - X^\star\|_F,$$

using Weyl's inequalities. Thus, $\|X_r - X^\star\|_F \leq \frac{4c}{\tau(X_r)}\sigma_r(X)$. Define $X_r = U_r U_r^\top$ and $X^\star = U_r^\star (U_r^\star)^\top$. Further, by Lemma 5.4 of Tu et al. (2015), we have:

$$\|X_r - X^\star\|_F \geq \sqrt{2\left(\sqrt{2}-1\right)}\sigma_r(U_r^\star) \cdot \textsc{Dist}(U_r, U_r^\star).$$

The above lead to:

$$\sqrt{2\left(\sqrt{2}-1\right)}\sigma_r(U_r^\star) \cdot \textsc{Dist}(U_r, U_r^\star) \leq \frac{4c}{\tau(X_r)}\sigma_r(X).$$

Recall that $\sigma_r(X) = \sigma_r^2(U_r)$; then, by Lemma 19, there is constant $c'' > 0$ such that $\frac{4c}{\tau(X_r)}\sigma_r^2(U) \leq \frac{c''}{\tau(X_r^\star)}\sigma_r^2(U_r^\star)$. Combining all the above, we conclude that there is constant $c' > 0$ such that:

$$\textsc{Dist}(U_r, U_r^\star) \leq \frac{c'}{\tau(X_r^\star)}\sigma_r(U_r^\star).$$

D.2. Proof of Theorem 11

Recall $X^0 = \mathcal{P}_+\left(\frac{-\nabla f(0)}{\|\nabla f(0) - \nabla f(e_1 e_1')\|_F}\right)$. Here, we remind that $\mathcal{P}_+(\cdot)$ is the projection operator onto the PSD cone and $\mathcal{P}_-(\cdot)$ is the projection operator onto the negative semi-definite cone.

To bound $\|X^0 - X^\star\|_F$, we will bound each individual term in its squared expansion

$$\|X^0 - X^\star\|_F^2 = \|X^0\|_F^2 + \|X^\star\|_F^2 - 2\left\langle X^0, X^\star\right\rangle.$$

From the smoothness of f, we get the following:

$$M\|X^\star\|_F \geq \|\nabla f(0) - \nabla f(X^\star)\|_F \overset{(i)}{\geq} \|\mathcal{P}_-(\nabla f(0)) - \mathcal{P}_-(\nabla f(X^\star))\|_F \overset{(ii)}{=} \|\mathcal{P}_-(\nabla f(0))\|_F.$$

where (i) follows from non-expansiveness of projection operator and (ii) follows from the fact that $\nabla f(X^\star)$ is PSD and hence $\mathcal{P}_-(\nabla f(X^\star)) = 0$. Finally, observe that $\mathcal{P}_-(\nabla f(0)) = \mathcal{P}_+(-\nabla f(0))$. The above combined imply:

$$\|\mathcal{P}_+(-\nabla f(0))\|_F \leq M\|X^\star\|_F \quad \Longrightarrow \quad \left\|X^0\right\|_F \leq \frac{M}{\|\nabla f(0) - \nabla f(e_1 e_1')\|_F} \cdot \|X^\star\|_F \leq \kappa\|X^\star\|_F$$

where we used the fact that $m \leq \left\|\nabla f(0) - \nabla f(e_1 e_1^\top)\right\|_F \leq M$ and $\kappa = M/m$. Hence $\|X^0\|_F^2 \leq \kappa^2\|X^\star\|_F^2$.

Using the strong convexity of f around $X^\star$, we observe

$$f(0) \geq f(X^\star) + \langle \nabla f(X^\star), 0 - X^\star\rangle + \frac{m}{2}\|X^\star\|_F^2 \geq f(X^\star) + \frac{m}{2}\|X^\star\|_F^2,$$

where the last inequality follows from first order optimality of $X^\star$, $\langle \nabla f(X^\star), 0 - X^\star\rangle \geq 0$ and 0 is a feasible point for problem (1). Similarly, using strong convexity of f around 0, we have

$$f(X^\star) \geq f(0) + \langle \nabla f(0), X^\star\rangle + \frac{m}{2}\|X^\star\|_F^2$$

Combining the above two inequalities we get, $\langle -\nabla f(0), X^\star \rangle \geq m \left\| X^\star \right\|_F^2$. Moreover:

$$\begin{aligned}\langle -\nabla f(0), X^\star \rangle &= \langle \mathcal{P}_+ \left(-\nabla f(0)\right) + \mathcal{P}_- \left(-\nabla f(0)\right), X^\star \rangle \\ &= \langle \mathcal{P}_+ \left(-\nabla f(0)\right), X^\star \rangle + \underbrace{\langle \mathcal{P}_- \left(-\nabla f(0)\right), X^\star \rangle}_{\leq 0}\end{aligned}$$

since $X^\star$ is PSD. Thus, $\langle \mathcal{P}_+(-\nabla f(0)), X^\star \rangle \geq \langle -\nabla f(0), X^\star \rangle$ and

$$\left\langle X^0, X^\star \right\rangle \geq \frac{m}{\left\| \nabla f(0) - \nabla f(e_1 e_1') \right\|_F} \left\| X^\star \right\|_F^2 \geq \frac{1}{\kappa} \left\| X^\star \right\|_F^2, \tag{43}$$

where we used the fact that $m \leq \left\| \nabla f(0) - \nabla f(e_1 e_1^\top) \right\|_F \leq M$. Given the above inequalities, we can now prove the following:

$$\begin{aligned}\left\| X^0 - X^\star \right\|_F^2 &= \left\| X^0 \right\|_F^2 + \left\| X^\star \right\|_F^2 - 2 \left\langle X^0, X^\star \right\rangle \\ &\leq \kappa^2 \left\| X^\star \right\|_F^2 + \left\| X^\star \right\|_F^2 - \frac{2}{\kappa} \left\| X^\star \right\|_F^2 \\ &= \left(\kappa^2 - \frac{2}{\kappa} + 1 \right) \left\| X^\star \right\|_F^2.\end{aligned}$$

Now we know that $\left\| X^0 - X^\star \right\|_F \leq \sqrt{\kappa^2 - 2/\kappa + 1} \left\| X^\star \right\|_F$. Now, by triangle inequality $\left\| X^0 - X_r^\star \right\|_F \leq \sqrt{\kappa^2 - 2/\kappa + 1} \left\| X^\star \right\|_F + \left\| X^\star - X_r^\star \right\|_F$. By $||.||_2 \leq ||.||_F$ and Weyl's inequality for perturbation of singular values (Theorem 3.3.16 Horn and Johnson (1991)) we get,

$$\left\| X_r^0 - X_r^\star \right\|_2 \leq 2\sqrt{\kappa^2 - 2/\kappa + 1} \left\| X^\star \right\|_F + 2 \left\| X^\star - X_r^\star \right\|_F.$$

By the assumptions of the theorem, we have $\left\| X^\star - X_r^\star \right\|_F \leq \tilde{\rho} \left\| X^\star \right\|_2$. Therefore,

$$\left\| X_r^0 - X_r^\star \right\|_F \leq 2\sqrt{2r} \left(\sqrt{\kappa^2 - 2/\kappa + 1} \left\| X^\star \right\|_F + \tilde{\rho} \left\| X^\star \right\|_2 \right).$$

Now again using triangle inequality and substituting we get $\left\| X^\star \right\|_F \leq \mathtt{srank}^{1/2} \left\| X^\star \right\|_2 + \tilde{\rho} \left\| X^\star \right\|_2$. Finally combining this with Lemma 20 gives the result.

Appendix E. Dependence on condition number in linear convergence rate

It is known that the convergence rate of classic gradient descent schemes depends only on the condition number $\kappa = \frac{M}{m}$ of the function f. However, in the case of FGD, we notice that convergence rate also depends on condition number $\tau(X_r^\star) = \frac{\sigma_1(X^\star)}{\sigma_r(X^\star)}$, as well as $\|\nabla f(X^\star)\|_2$.

To elaborate more on this dependence, let us recall the update rule of FGD, as presented in Section 3. In particular, one can observe that the gradient direction has an extra factor U, multiplying $\nabla f(UU^\top)$, as compared to the standard gradient descent on X. One way to reveal how this extra factor affects the condition number of the Hessian of f, we consider the special case of separable functions; see the definition of separable functions in the next lemma. Next, we show that the condition number of the Hessian – for this special case – has indeed a dependence on both $\tau(X_r^\star)$ and $\|\nabla f(X^\star)\|_2$, a scaling similar to the one appearing in the convergence rate α of FGD.

Lemma 27 (Dependence of Hessian on $\tau(X_r^\star)$ and $\|\nabla f(X^\star)\|_2$) *Let f be a smooth, twice differentiable function over the PSD cone. Further, assume f is a separable function over the matrix entries, such that $f(X) = \sum_{(i,j)} \varphi_{ij}(X_{ij})$, where $(i,j) \in [n] \times [n]$, and let φ_{ij}'s be M-smooth and m-strongly convex functions, $\forall i, j$. Finally, let $X^\star = U^\star (U^\star)^\top$ be rank-r and let $\nabla^2_{UR} f(X)$ denote the Hessian of f w.r.t. the U factor, for some orthonormal matrix $R \in \mathcal{O}$. Then,*

$$\sigma_1\left(\nabla^2_{U^\star} f(X^\star)\right) \leq C \cdot (M\|X^\star\|_2 + \|\nabla f(X^\star)\|_2),$$

for constant C. Further, for any unit vector $y \in \mathbb{R}^{nr\times 1}$ such that columns of $\mathtt{mat}(y) \in \mathbb{R}^{n\times r}$ are orthogonal to $U^\star$, i.e., $\mathtt{mat}(y)^\top U^\star = 0$, we further have:

$$y^\top \nabla^2_{U^\star R} f(X^\star) y \geq c \cdot m\sigma_r(X^\star),$$

for some constant c.

Proof By the definition of gradient, we know that $\nabla_U f(UU^\top) = (\nabla f(UU^\top) + \nabla f(UU^\top)^\top)U$; for simplicity, we assume $\nabla f(UU^T)$ be symmetric. Since X is symmetric, with $\nabla f(UU^T)_{ij} = \varphi'_{ij}(X_{ij})$ and $\varphi'_{ij}(X_{ij}) = \varphi'_{ji}(X_{ji})$. By the definition of Hessian, the entries of $\nabla^2_U f(UU^T)$ are given by:

$$\left(\nabla^2_U f(UU^T)\right)_{ij,kl} = \frac{\partial}{\partial U_{kl}} \sum_{p=1}^{n} \varphi'_{ip}(X_{ip}) U_{pj} = \underbrace{\sum_{p=1}^{n} \frac{\partial \varphi'_{ip}(X_{ip})}{\partial U_{kl}} U_{pj}}_{:=T_1} + \underbrace{\sum_{p=1}^{n} \varphi'_{ip}(X_{ip}) \frac{\partial U_{pj}}{\partial U_{kl}}}_{:=T_2}.$$

In particular, for T_1 we observe the following cases:

$$T_1 = \begin{cases} \varphi''_{ik}(X_{ik}) U_{il} U_{kj} & \text{if } i \neq k, \\ \sum_p \varphi''_{ip}(X_{ip}) U_{pl} U_{pj} + \varphi''_{ii}(X_{ii}) U_{il} U_{kj} & \text{if } i = k. \end{cases}$$

while, for T_2 we further have:

$$T_2 = \begin{cases} 0 & \text{if } j \neq l, \\ \varphi'_{ik}(X_{ik}) & \text{if } j = l. \end{cases}$$

Consider now the case where gradient and Hessian information is calculated at the optimal point $X^\star$. Based on the above, the Hessian of f w.r.t $U^\star$ turns out to be a sum of three PSD $nr \times nr$ matrices, as follows:

$$\nabla^2_{U^\star} f(X^\star) = A + B + C,$$

where

(*i*) $A = (\widehat{U^\star})^T G \widehat{U^\star}$, where G is a $n^2 \times n^2$ diagonal matrix with diagonal elements $\varphi''_{ij}(X^\star_{ij})$ and $\widehat{U^\star}$ is a $n^2 \times nr$ matrix with $U^\star$ repeated n times on the diagonal. It is easy to see that

$$\|A\|_2 \leq \|\varphi''_{ij}\|_\infty \sigma_{\max}(U^\star)^2 = M\|X^\star\|_2.$$

Similarly, we have $\sigma_{nr}(A) \geq \min \varphi''_{ij} \cdot \sigma_{\min}(U^\star)^2 = m\sigma_{\min}(X^\star)$.

(*ii*) B is a $nr \times nr$ matrix, with $B_{ij,kl} = \varphi''_{ik}(X^\star_{ik})U^\star_{il}U^\star_{kj}$. Again, it is easy to verify that $\|B\|_2 \leq M\|X^\star\|_2$. Now for y perpendicular to $U^\star$, notice that $y^\top By = 0$, since the columns of B are concatenation of scaled columns of $U^\star$.

(*iii*) C is a $nr \times nr$ diagonal block-matrix, with $n \times n$ blocks $\nabla f(X^\star)$ repeated r times. It is again easy to see that $\|C\|_2 \leq \|\nabla f(X^\star)\|_2$, since C is a block diagonal matrix. Moreover, by KKT optimality condition $\nabla f(X^\star)X^\star = 0$, $\mathrm{rank}(\nabla f(X^\star)) \leq n - r$ and thus, $\sigma_{nr}(C) = 0$.

Combining the above results and observing that all the three matrices are PSD, we conclude that $\sigma_1\left(\nabla^2_{U^\star} f(X^\star)\right) \leq C \cdot (M\|X^\star\|_2 + \|\nabla f X^\star\|_2)$. Regarding the lower bound on $\sigma_{nr}\left(\nabla^2_{U^\star} f(X^\star)\right)$, we observe the following: due to $UU^\top$ factorization and for $U^\star$ optimum, we know that also $U^\star R$ is optimum, where gradient $\nabla f(U^\star(U^\star)^\top) = \nabla f(U^\star RR^\top(U^\star)^\top) = 0$. This further indicates that the hessian of f is zero along directions corresponding to columns of $U^\star$, and thus $\sigma_{nr}\left(\nabla^2_{U^\star} f(X^\star)\right) = 0$ along these directions; see figure 3 (right panel) for an example. However, for any other directions orthogonal to $U^\star$, we have $y^\top\left(\nabla^2_{U^\star R} f(X^\star)\right) y \geq c \cdot m\sigma_{\min}(X^\star)$, for some constant c. This completes the proof.

■

To show this dependence in practice, we present some simulation results in Figure 3. We observe that the convergence rate does indeed depend on $\tau(X^\star_r)$.

Appendix F. Test case I: Matrix sensing problem

In this section, we briefly describe and compare algorithms designed specifically for the *matrix sensing* problem, using the variable parametrization $X = UU^\top$. To accommodate the PSD constraint, we consider a variation of the matrix sensing problem where one desires to find $X^\star$ that minimizes[20]:

$$\underset{X \in \mathbb{R}^{n\times n}}{\text{minimize}} \quad \tfrac{1}{2}\|b - \mathcal{A}(X)\|_F^2 \quad \text{subject to} \quad \mathrm{rank}(X) \leq r,\ X \succeq 0. \tag{44}$$

W.l.o.g., we assume $b = \mathcal{A}(X^\star)$ for some rank-r $X^\star$. Here, $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^p$ is the linear sensing mechanism, such that the i-th entry of $\mathcal{A}(X)$ is given by $\langle A_i,\ X\rangle$, for $A_i \in \mathbb{R}^{n\times n}$ sub-Gaussian independent measurement matrices.

Jain et al. (2013) is one of the first works to propose a provable and efficient algorithm for (44), operating in the U-factor space, while Sa et al. (2015) solves (44) in the stochastic setting; see also Zheng and Lafferty (2015); Tu et al. (2015); Chen and Wainwright (2015). To guarantee convergence, most of these algorithms rely on *restricted isometry assumptions*; see Definition 28 below.

To compare the above algorithms with FGD, Subsection F.1 further describes the notion of restricted strong convexity and its connection with the RIP. Then, Subsection F.2 provides explicit comparison results of the aforementioned algorithms, with respect to the convergence rate factor α, as well as initialization conditions assumed, for each case.

20. This problem is a special case of affine rank minimization problem Recht et al. (2010), where no PSD constraints are present.

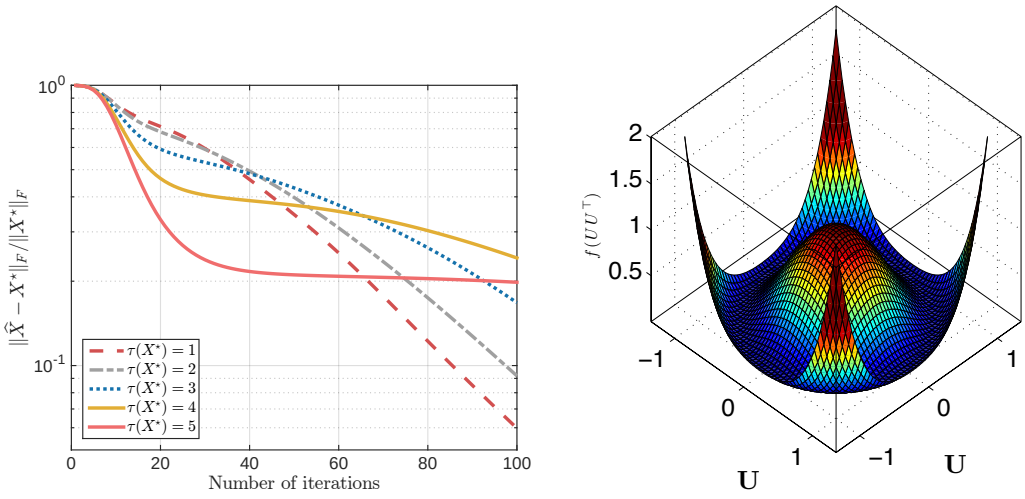

Figure 3: Left panel: Assume dimension $n = 50$. We consider the matrix sensing setup Recht et al. (2010) and generate $m = \lceil 2n \log n \rceil$ Gaussian linear measurements of $n \times n$ matrices $X^\star$ of rank $r = 2$, with varying condition number $\tau(X^\star)$. We compute matrix $X = UU^\top$, U is $n \times r$ tall matrix, by minimizing the standard least squares lost function, using our scheme. In the plot, we show the log error versus total number of iterations. Observe that, varying the condition number of $X^\star$, higher $\tau(X^\star)$ leads to slower convergence. Right panel: Contour of function $(u_1^2 + u_2^2 - 1)^2$. Observe the "ring" of points (u_1, u_2) where f is minimized. This illustrates the existence of multiple points with zero gradient and, thus, directions where the hessian of the objective is zero.

F.1. Restricted isometry property and restricted strong convexity

To shed some light on the notion of restricted strong convexity and how it relates to the RIP, consider the matrix sensing problem, as described above. According to (44), we consider the quadratic loss function:

$$f(X) = \tfrac{1}{2}\|b - \mathcal{A}(X)\|_F^2.$$

Since the Hessian of f is given by $\mathcal{A}^*\mathcal{A}$, restricted strong convexity suggests that Negahban and Wainwright (2012):

$$\|\mathcal{A}(Z)\|_2^2 \geq C \cdot \|Z\|_F^2, \quad Z \in \mathbb{R}^{n \times n},$$

for a restricted set of directions Z, where $C > 0$ is a small constant. This bound implies that the quadratic loss function, as defined above, is strongly convex in such a restricted set of directions Z.[21]

A similar but stricter notion is that of *restricted isometry property* for low rank matrices Candes and Plan (2011); Liu (2011):

21. One can similarly define the notion of restricted smoothness condition, where $\|\mathcal{A}(Z)\|_2^2$ is upper bounded by $\|Z\|_F^2$.

Definition 28 (Restricted Isometry Property (RIP)) *A linear map $\mathcal{A}$ satisfies the r-RIP with constant δ_r, if*

$$(1-\delta_r)\|X\|_F^2 \leq \|\mathcal{A}(X)\|_2^2 \leq (1+\delta_r)\|X\|_F^2,$$

is satisfied for all matrices $X \in \mathbb{R}^{n\times n}$ such that rank$(X) \leq r$.

The correspondence of restricted strong convexity with the RIP is obvious: both lower bound the quantity $\|\mathcal{A}(X)\|_2^2$, where X is drawn from a restricted set. It turns out that linear maps that satisfy the RIP for low rank matrices, also satisfy the restricted strong convexity; see Theorem 2 in Chen and Sanghavi (2010).

By assuming RIP in (44), the condition number of f depends on the RIP constants of the linear map $\mathcal{A}$; in particular, one can show that $\kappa = \frac{M}{m} \propto \frac{1+\delta}{1-\delta}$, since the eigenvalues of $\mathcal{A}^*\mathcal{A}$ lie between $1-\delta$ and $1+\delta$, when restricted to low-rank matrices. For δ sufficiently small and dimension n sufficiently large, $\kappa \approx 1$, which, with high probability, is the case for $\mathcal{A}$ drawn from a sub-Gaussian distribution.

F.2. Comparison

Given the above discussion, the following hold true for FGD, under RIP settings:

(i) In the noiseless case, $b = \mathcal{A}(X^\star)$ and thus, $\|\nabla f(X^\star)\|_2 = \| -2\mathcal{A}^*\left(b - \mathcal{A}(X^\star)\right)\|_2 = 0$. Combined with the above discussion, this leads to convergence rate factor

$$\alpha \lesssim 1 - \tfrac{c_4}{\tau(U_r^\star)^2},$$

in FGD.

(ii) In the noisy case, $b = \mathcal{A}(X^\star) + e$ where e is an additive noise term; for this case, we further assume that $\|\mathcal{A}\left(e\right)\|_2$ is bounded. Then,

$$\alpha \lesssim 1 - \frac{c_4}{\tau(U_r^\star)^2 + \frac{\|\mathcal{A}(e)\|_2}{(1-\delta)\sigma_r(X^\star)}}.$$

Table 2 summarizes convergence rate factors α and initialization conditions of state-of-the-art approaches for the noiseless case.

F.3. Empirical results

We start our discussion on empirical findings with respect to the convergence rate of the algorithm, how the step size and initialization affects its efficiency and some comparison plots with an efficient first-order projected gradient solver. We note that the experiments presented below are performed as a proof of concept and are not complete in the set of algorithms we could compare with.

Linear convergence rate and step size selection: To show the convergence rate of the factored gradient descent in practice, we solve affine rank minimization problems instances with synthetic data. In particular, the ground truth $X^\star \in \mathbb{R}^{n\times n}$ is synthesized as a rank-r matrix as $X^\star = U^\star\left(U^\star\right)^\top$, where $U^\star \in \mathbb{R}^{n\times r}$. In sequence, we sub-sample $X^\star$ by observing $m = C_{\text{sam}} \cdot p \cdot r$ entries, according to:

$$y = \mathcal{A}(X^\star) \in \mathbb{R}^m. \tag{45}$$

Reference	$\text{DIST}(U^+, U^\star)^2 \leq \alpha \cdot \text{DIST}(U, U^\star)^2$	$\text{DIST}(U^0, U^\star) \leq \cdots$
Jain et al. (2013)	$\alpha = \frac{1}{16}$ †	$\sqrt{6\delta} \cdot \frac{\sigma_r(X^\star)}{\sigma_1(X^\star)} \sigma_r(U_r^\star)$
Tu et al. (2015)	$\alpha = 1 - \frac{c_1}{\tau(U_r^\star)^4}$	$\frac{1}{4}\sigma_r(U_r^\star)$
Zheng and Lafferty (2015)	$\alpha = 1 - \frac{c_2}{\tau(U_r^\star)\cdot r}$	$\sqrt{\frac{3}{16}} \cdot \sigma_r(U_r^\star)$
Chen and Wainwright (2015)	$\alpha = 1 - \frac{c_3}{\tau(U_r^\star)^{10}}$	$(1-\tau) \cdot \sigma_r(U_r^\star)$
This work	$\alpha = 1 - \frac{c_4}{\tau(U_r^\star)^2}$	$\frac{1}{100} \cdot \frac{\sigma_r(X^\star)}{\sigma_1(X^\star)} \sigma_r(U_r^\star)$

Table 2: Comparison of related work for the matrix sensing problem. All methods use $UU^\top$ parametrization of the variable X and admit linear convergence. $\tau = \sqrt{12\delta}$ according to Chen and Wainwright (2015). $c_i > 0$, $\forall i$ denote absolute constants. In Jain et al. (2013), the proposed algorithm is designed to solve the rectangular case where $X = UV^\top$; the reported factor α and initial conditions could be improved for the case of (2). † Note that this convergence is in terms of subspace distance.

We use permuted and sub-sampled noiselets for the linear operator $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$; for more information, see Waters et al. (2011). $\mathbf{y} \in \mathbb{R}^m$ contains the linear measurements of $X^\star$ through $\mathcal{A}$ in vectorized form. We consider the noiseless case, for ease of exposition. Under this setting, we solve (2) with $f(UU^\top) := 1/2 \cdot \|y - \mathcal{A}\left(UU^\top\right)\|_2^2$. We use as a stopping criterion the condition $\|U^+ (U^+)^\top - UU^\top\|_F < \texttt{tol} \cdot \|U^+ (U^+)^\top\|_F$ where $\texttt{tol} := 5 \cdot 10^{-6}$.

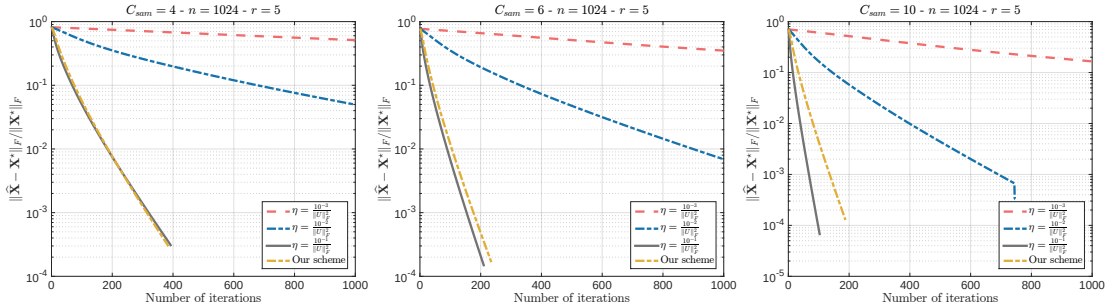

Figure 4: Median error per iteration of factored gradient descent algorithm for different step sizes, over 20 Monte Carlo iterations. The number of measurements is fixed to $C_{\text{sam}} \cdot n \cdot r$ for varying $C_{\text{sam}} \in \{4, 6, 10\}$. Here, $n = 1204$ and rank $r = 5$. Curves show convergence behavior of factored gradient descent as a function of the step size selection. One can observe that arbitrary step size selections can lead to slow convergence. Moreover, good constant step size selections – for a specific problem configuration, do not necessarily translate into good performance for a different setting; e.g., observe how the constant step size convergence rates worsen *faster*, as we decrease the number of observations.

Figure 4 show the linear convergence of our approach as well as the efficiency of our step selection, as compared to other arbitrary constant step size selections. All instances use our initialization

point. It is worth mentioning that the performance of our step size can be inferior to specific constant step size selections; however, finding such a good constant step size usually requires trial-and-error rounds and do not come with convergence guarantees. Moreover, we note that one can perform line search procedures to find the "best" step size per iteration; although, for more complicated f instances, such step size selection might not be computationally desirable, even infeasible.

Impact of avoiding low-rank projections on the PSD cone: In this experiment, we compare factored gradient descent with a variant of the Singular Value Projection (SVP) algorithm Jain et al. (2010); Becker et al. (2013)[22]. For the purpose of this experiment, the SVP variant further projects on the PSD cone, along with the low rank projection. Its main difference is that it does not operate on the factor U space but requires projection over the (low-rank) positive semi-definite cone per iteration. In the discussion below, we refer to this variant as SVP (SDP).

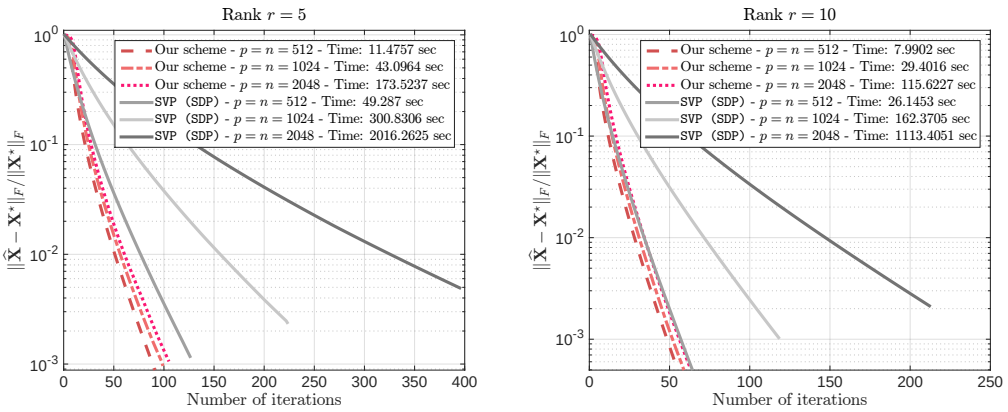

Figure 5: Median error per iteration for factored gradient descent and SVP (SDP) algorithms, over 20 Monte Carlo iterations. For all cases, the number of measurements is fixed to $C_{\text{sam}} \cdot n \cdot r$ for $C_{\text{sam}} = 6$. From left to right, we consider different rank configurations: (i) $r = 5$ and (ii) $r = 10$. Both schemes use the same initialization point. Both plots show better convergence rate performance in terms of iterations due to our step size selection. In addition, factored gradient descent avoids performing SVD operations per iteration, a fact that leads also to lower per iteration complexity; see also Table 3.

We perform two experiments. In the first experiment, we compare factored gradient descent with SVP (SDP), as designed in Jain et al. (2010); *i.e.*, while we use our initialization point for both schemes, step size selections are different. Figure 5 shows some convergence rate results: clearly our step size selection performs better in practice, in terms of the total number of iterations required for convergence.

In the second experiment, we would like to highlight the time bottleneck introduced by the projection operations: for this aim, *we use the same initialization points and step sizes* for both the algorithms under comparison. Thus, the only difference lies in the SVD computations of SVP (SDP) to retain a PSD low rank estimate per iteration. Table 3 presents reconstruction error and

22. SVP is a non-convex, first-order, projected gradient descent scheme for low rank recovery from linear measurements.

Model		$\|\widehat{X} - X^\star\|_F / \|X^\star\|_F$		Time (sec)	
n	r	SVP (SDP)	Our scheme	SVP (SDP)	Our scheme
512	5	1.1339e-03	**8.4793e-04**	36.9652	**11.4757**
	10	4.6552e-04	**4.4954e-04**	19.6089	**7.9902**
	20	**1.6541e-04**	2.0571e-04	10.6052	**6.4149**
1024	5	2.4224e-03	**9.9180e-04**	225.6230	**43.0964**
	10	1.0203e-03	**4.5103e-04**	121.7779	**29.4016**
	20	4.1149e-04	**2.3442e-04**	67.6272	**22.9616**
2048	5	4.8500e-03	**1.0093e-03**	1512.1969	**173.5237**
	10	2.0836e-03	**4.6735e-04**	835.0538	**115.6227**
	20	9.4893e-04	**2.6417e-04**	458.8766	**88.1960**

Table 3: Summary of comparison results for reconstruction and efficiency. Observe that both our scheme and SVP (SDP) require more iterations to converge as r radically decreases. This justifies the higher time-complexity observed; see also Figure 5 for comparison.

execution time results. It is obvious that projecting on the low-rank PSD code per iteration constitutes a computational bottleneck per iteration, which slows down (w.r.t. total time required) the convergence of SVP (SDP).

Initialization. Here, we evaluate the importance of our initialization point selection:

$$X^0 \;:=\; \mathcal{P}_+\left(\frac{-\nabla f(0)}{\|\nabla f(0) - \nabla f(e_1 e_1')\|_F}\right) \tag{46}$$

To do so, we consider the following settings: we compare random initializations against the rule (46), both for constant step size selections and our step size selection. In all cases, we work with the factored parametrization.

Figure 6 shows the results. Left panel presents results for constant step size selections where $\eta = {}^{0.1}/_{\|U\|_F^2}$ and right panel uses our step size selection; again, note that the selection of the constant step size is after many trial-and-errors for best step size selection, based on the specific configuration. Both figures compare the performance of factored gradient descent when (i) a random initialization point is selected and, (ii) our initialization is performed, according to (46). All curves depict median reconstruction errors over 20 Monte Carlo iterations. For all cases, the number of measurements is fixed to $C_{\text{sam}} \cdot n \cdot r$ for $C_{\text{sam}} = 10$, $n = 1024$ and rank $r = 20$.

Dependence of α on $\frac{\sigma_1(X^\star)}{\sigma_r(X^\star)}$. Here, we highlight the dependence of $\frac{\sigma_1(X^\star)}{\sigma_r(X^\star)}$ on the convergence rate of factored gradient descent. Consider the following matrix sensing toy example: let $X^\star := U^\star (U^\star)^\top \in \mathbb{R}^{n\times n}$ for $n = 50$ and assume $\text{rank}(X^\star) > r$. We desire to compute a (at most) rank-r

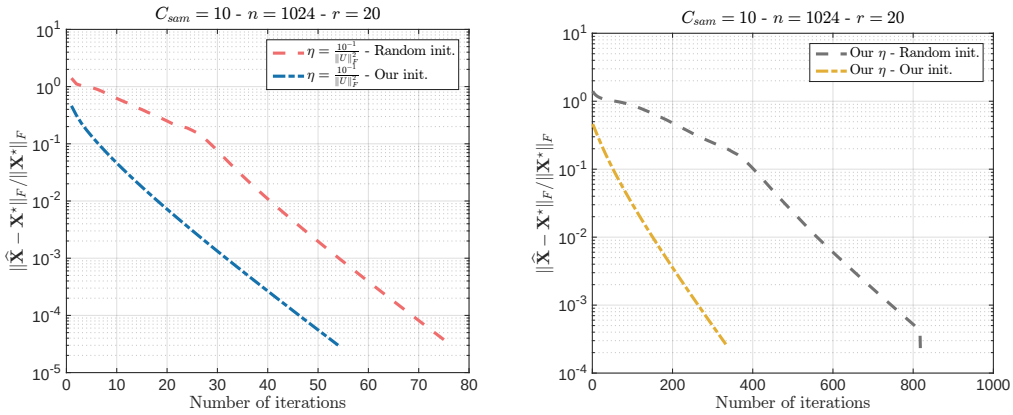

Figure 6: Median error per iteration for different initialization set ups. Left panel presents results for constant step size selections where $\eta = {}^{0.1}/\|U\|_F^2$ and right panel uses our step size selection. Both figures compare the performance of factored gradient descent when (i) a random initialization point is selected and, (ii) our initialization is performed, according to (46). All curves depict median reconstruction errors over 20 Monte Carlo iterations. For all cases, the number of measurements is fixed to $C_{\text{sam}} \cdot n \cdot r$ for $C_{\text{sam}} = 10$, $n = 1024$ and rank $r = 20$.

approximation of $X^\star$ by minimizing the simple least squares loss function:

$$\begin{aligned} \underset{X \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad & \frac{1}{2}\|X - X^\star\|_F^2 \\ \text{subject to} \quad & X \succeq 0, \quad \text{rank}(X) \leq r \end{aligned} \tag{47}$$

For this example, let us consider $r = 3$ and design $X^\star$ according to the following three scenarios: we fix $\sigma_1(X^\star) = \sigma_2(X^\star) = 100$ and vary $\sigma_3(X^\star) \in \{1, 10, 20\}$. This leads to condition numbers for these three cases as: (i) $\frac{\sigma_1(X^\star)}{\sigma_3(X^\star)} = 100$, (ii) $\frac{\sigma_1(X^\star)}{\sigma_3(X^\star)} = 10$ and, (iii) $\frac{\sigma_1(X^\star)}{\sigma_3(X^\star)} = 5$. The convergence behavior is shown in Figure 7(Left panel). It is obvious that factored gradient descent suffers – w.r.t. convergence rate – as the condition number $\frac{\sigma_1(X^\star)}{\sigma_3(X^\star)}$ get worse; especially, for the case where $\frac{\sigma_1(X^\star)}{\sigma_3(X^\star)} = 100$, factored gradient descent reaches a plateau after the $\sim$80-th iteration, where the steps towards solution become smaller. As the condition number improves, factored gradient descent enjoys faster convergence to the optimum, which shows the dependence of the algorithm on $\frac{\sigma_1(X^\star)}{\sigma_3(X^\star)}$ also in practice.

As a second setting, we fix $r = 2$, thereby computing a rank-2 approximation. As Figure 7(Right panel) illustrates, for all values of $\sigma_3(X^\star)$, factored gradient descent performs similarly, enjoying fast convergence towards the optimum $X^\star$. Thus, while the condition number of original $X^\star$ varies to a large degree for $r = 3$, the convergence rate factor α only depends on $\frac{\sigma_1(X^\star)}{\sigma_2(X^\star)} = 1$, for $r = 2$. This leads to similar convergence behavior for all three scenarios described above.

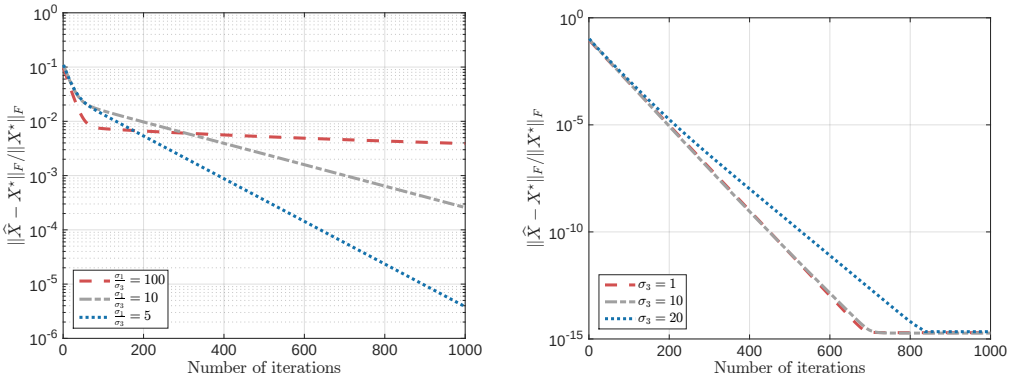

Figure 7: Toy example on the dependence of α on the term $\frac{\sigma_1(X^\star)}{\sigma_r(X^\star)}$. Here, $X^\star := U^\star (U^\star)^\top \in \mathbb{R}^{n\times n}$ for $n = 50$. We use factored gradient descent to solve (47) for $r = 3$. Left panel: As condition number $\frac{\sigma_1(X^\star)}{\sigma_3(X^\star)}$ improves, factored gradient descent enjoys faster convergence in practice, as dictated by our theory. Right panel: convergence rate behavior of factored gradient descent when $r = 2$ in (47).

Appendix G. Test case II: PSD problems with high-rank solutions

As a final example, we consider problems of the form:

$$\underset{X\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad f(X) \quad \text{subject to} \quad X \succeq 0,$$

where $X^\star$ is the minimizer of the above problem and $\text{rank}(X^\star) = O(n)$. In this particular case and assuming we are interested in finding high-ranked $X^\star$, we can reparameterized the above problem as follows:

$$\underset{U\in\mathbb{R}^{n\times O(n)}}{\text{minimize}} \quad f(UU^\top).$$

Observe that U is a square $n \times O(n)$ matrix. Under this setting, FGD performs the recursion:

$$\underbrace{U^+}_{n\times O(n)} = \underbrace{U}_{n\times O(n)} - \eta \underbrace{\nabla f(UU^\top)}_{n\times n} \cdot \underbrace{U}_{n\times O(n)} .$$

Due to the matrix-matrix multiplication, the per-iteration time complexity of FGD is $O(n^3)$, which is comparable to a SVD calculation of a $n \times n$ matrix. In this experiment, we study the performance of FGD in such high-rank cases and compare it with state-of-the-art approaches for PSD constrained problems.

For the purpose of this experiment, we only consider first-order solvers; *i.e.*, second order methods such as interior point methods are excluded as, in high dimensions, it is prohibitively expensive the hessian of f. To this end, the algorithms to compare include: (i) standard projected gradient descent approach Kyrillidis and Cevher (2014) and (ii) Frank-Wolfe type of algorithms, such as the one in Hazan (2008). We note that this experiment can be seen as a proof of concept on how avoiding SVD calculations help in practice.[23]

23. Here, we assume a standard Lanczos implementation of SVD, as the one provided in Matlab enviroment.

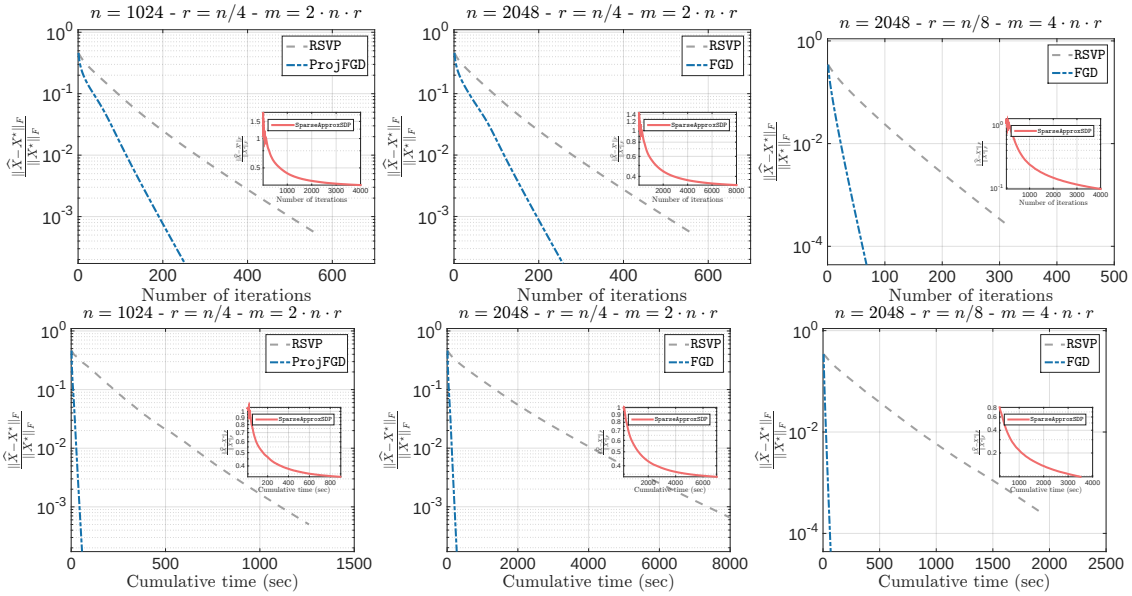

Figure 8: Convergence performance of algorithms under comparison w.r.t. $\frac{\|\widehat{X}-X^\star\|_F}{\|X^\star\|_F}$ vs. (i) the total number of iterations (top) and (ii) the total execution time (bottom).

Experiments. We consider the simple example of matrix sensing Kyrillidis and Cevher (2014): we obtain a set of measurements $y \in \mathbb{R}^m$ according to the linear model:

$$y = \mathcal{A}(X^\star).$$

Here, $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ is a sensing mechanism such that $(\mathcal{A}(X))_i = \mathrm{Tr}(A_i X)$ for some Gaussian random matrices A_i, $i = 1, \ldots, m$. The ground truth matrix $X^\star$ is design such that $\mathrm{rank}(X^\star) = n/4$ and $\mathrm{Tr}(X^\star) = 1$.[24]

Figure 8 and Table 4 show some results for the following settings: (i) $n = 1024$, $r = n/4$ and $m = 2nr$, (ii) $n = 2048$, $r = n/4$ and $m = 2nr$, (iii) $n = 2048$, $r = n/8$ and $m = 4nr$. From our finding, we observe that, even for high rank cases—where $r = O(n)$—performing matrix factorization and optimizing over the factors results into a much faster convergence, as compared to low-rank projection algorithms, such as `RSVP` in Becker et al. (2013). Furthermore, FGD performs better than `SparseApproxSDP` Hazan (2008) in practice: while `SparseApproxSDP` is a Frank-Wolfe type-of algorithm (and thus, the per iteration complexity is low), it admits *sublinear* convergence which leads to suboptimal performance, in terms of total execution time. However, `RSVP` and `SparseApproxSDP` algorithms do not assume specific initialization procedures to work in theory.

24. The reason we design $X^\star$ such that $\mathrm{Tr}(X^\star) = 1$ is such that the algorithm `SparseApproxSDP` Hazan (2008) applies; this is due to the fact that `SparseApproxSDP` is designed for QST problems, where trace constraint is present in the optimization criterion.

Appendix H. Convergence without tail bound assumptions

In this section, we show how assumptions (A3) and (A4) can be dropped by using a different step size η, where spectral norm calculation of two $n \times r$ matrices is required per iteration. Here, we succinctly describe the main theorems and how they differ from the case where η as in (8) is used. We also focus only on the case of restricted strongly convex functions. Similar extension is possible without restricted strong convexity.

Our discussion is organized as follows: we first re-define key lemmas (*e.g.*, descent lemmas, etc.) for a different step size; then, we state the main theorems and a sketch of their proof. In the analysis below we use as step size:

$$\widehat{\eta} = \frac{1}{16\left(M\left\|X\right\|_2 + \left\|\nabla f(X) Q_U Q_U^\top\right\|_2\right)}$$

H.1. Key lemmas

Next, we present the main descent lemma that is used for both sublinear and linear convergence rate guarantees of FGD.

Lemma 29 (Descent lemma) *For f being a M-smooth and (m, r)-strongly convex function and under assumptions $(A2)$ and $f(X^+) \geq f(X_r^\star)$, the following inequality holds true:*

$$\tfrac{1}{\widehat{\eta}} \left\langle U - U^+, U - U_r^\star R_U^\star \right\rangle \geq \tfrac{3}{5}\widehat{\eta}\|\nabla f(X) U\|_F^2 + \tfrac{3m}{20} \cdot \sigma_r(X^\star)\|\Delta\|_F^2.$$

Proof [Proof of Lemma 29] By (13), we have:

$$\left\langle \nabla f(X) U, U - U_r^\star R_U^\star \right\rangle = \frac{1}{2}\left\langle \nabla f(X), X - X_r^\star \right\rangle + \frac{1}{2}\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle, \tag{48}$$

Step I: Bounding $\langle \nabla f(X), X - X_r^\star \rangle$. For this term, we have a variant of Lemma 15, as follows:

Algorithm	$\|\widehat{X} - X^\star\|_F / \|X^\star\|_F$	Total time (sec)	Time per iter. (sec - median)
	Setting: $n = 1024$, $r = n/4$, $m = 2nr$.		
RSVP	4.9579e-04	1262.3719	2.1644
SparseApproxSDP	3.3329e-01	895.9605	2.1380e-01
FGD	1.6763e-04	57.8495	2.1961e-01
	Setting: $n = 2048$, $r = n/4$, $m = 2nr$.		
RSVP	4.9537e-04	8412.6981	14.6811
SparseApproxSDP	3.3526e-01	26962.0379	8.7761e-01
FGD	1.6673e-04	272.8102	1.0040e+00
	Setting: $n = 2048$, $r = n/8$, $m = 4nr$.		
RSVP	2.4254e-04	1945.6714	5.9763
SparseApproxSDP	9.6725e-02	3506.8147	8.6440e-01
FGD	3.8917e-05	68.5689	9.2567e-01

Table 4: Comparison of related work in high-rank matrix sensing problems. We construct $X^\star$ with $\mathrm{Tr}(X^\star) = 1$ such that Hazan (2008) applies. It is apparent that avoiding SVDs helps in practice.

Lemma 30 *Let f be a M-smooth and (m,r)-restricted strongly convex function with optimum point $X^\star$. Assume $f(X^+) \geq f(X_r^\star)$. Let $X = UU^\top$. Then,*

$$\langle \nabla f(X), X - X_r^\star \rangle \geq \tfrac{18\widehat{\eta}}{10} \|\nabla f(X)U\|_F^2 + \tfrac{m}{2} \|X - X_r^\star\|_F^2 ,$$

where $\widehat{\eta} = \frac{1}{16(M\|X\|_2 + \|\nabla f(X) Q_U Q_U^\top\|_2)}$.

The proof of this lemma is provided in Section I.1.

Step II: Bounding $\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle$. For the second term, we have the following variant of Lemma 16.

Lemma 31 *Let f be M-smooth and (m,r)-restricted strongly convex. Then, under assumptions $(A2)$ and $f(X^+) \geq f(X_r^\star)$, the following bound holds true:*

$$\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \geq -\tfrac{6\widehat{\eta}}{25} \|\nabla f(X)U\|_F^2 - \tfrac{3m\sigma_r(X^\star)}{40} \cdot \|\Delta\|_F^2 .$$

Proof of this lemma can be found in Section I.2.

Step III: Combining the bounds in equation (48). The rest of the proof is similar to that of Lemma 14.

■

H.2. Proof of linear convergence

For the case of (restricted) strongly convex functions f, we have the following revised theorem:

Theorem 32 (Convergence rate for restricted strongly convex f) *Let current iterate be U and $X = UU^\top$. Assume $\textsc{Dist}(U, U_r^\star) \leq \rho' \sigma_r(U_r^\star)$ and let the step size be $\widehat{\eta} = \frac{1}{16\,(M\|X\|_2 + \left\|\nabla f(X) Q_U Q_U^\top\right\|_2)}$. Then under assumptions $(A2)$ and $f(X^+) \geq f(X_r^\star)$, the new estimate $U^+ = U - \eta \nabla f(X) \cdot U$ satisfies*

$$\textsc{Dist}(U^+, U_r^\star)^2 \leq \alpha \cdot \textsc{Dist}(U, U_r^\star)^2, \tag{49}$$

where $\alpha = 1 - \frac{m\sigma_r(X^\star)}{64(M\|X^\star\|_2 + \|\nabla f(X_r^\star)\|_2)}$. Furthermore, U^+ satisfies $\textsc{Dist}(U^+, U_r^\star) \leq \rho' \sigma_r(U_r^\star)$.

The proof follows the same steps as that of theorem 6, except from the fact Lemmas 30 and 31 are used.

Appendix I. Main lemmas for convergence proof without tail bound assumptions

I.1. Proof of Lemma 30

Let $U^+ = U - \widehat{\eta} \nabla f(X) U$ and $X^+ = U^+ (U^+)^\top$. By smoothness of f, we get:

$$\begin{aligned} f(X) &\geq f(X^+) - \langle \nabla f(X), X^+ - X \rangle - \tfrac{M}{2} \|X^+ - X\|_F^2 \\ &\overset{(i)}{\geq} f(X_r^\star) - \langle \nabla f(X), X^+ - X \rangle - \tfrac{M}{2} \|X^+ - X\|_F^2 , \end{aligned} \tag{50}$$

where (i) follows from hypothesis of the lemma and since X^+ is a feasible point ($X^+ \succeq 0$) for problem (1). Finally, since $\text{rank}(X_r^\star) = r$, by the (m, r)-restricted strong convexity of f, we get,

$$f(X_r^\star) \geq f(X) + \langle \nabla f(X), X_r^\star - X \rangle + \tfrac{m}{2} \|X_r^\star - X\|_F^2 . \tag{51}$$

Combining equations (50) and (51), we obtain:

$$\langle \nabla f(X), X - X_r^\star \rangle \geq \langle \nabla f(X), X - X^+ \rangle - \tfrac{M}{2} \|X^+ - X\|_F^2 + \tfrac{m}{2} \|X_r^\star - X\|_F^2 , \tag{52}$$

instead of (25) in the proof where η is used. The rest of the proof follows the same steps as that of Lemma 15 and we get:

$$\langle \nabla f(X), X - X_r^\star \rangle \geq \tfrac{18\widehat{\eta}}{10} \|\nabla f(X) U\|_F^2 + \tfrac{m}{2} \|X_r^\star - X\|_F^2 .$$

Moreover, for the case where f is just M-smooth and $X^\star \equiv X_r^\star$, the above bound becomes:

$$\langle \nabla f(X), X - X_r^\star \rangle \geq \tfrac{18\widehat{\eta}}{10} \|\nabla f(X) U\|_F^2 .$$

This completes the proof.

I.2. Proof of Lemma 31

Similar to Lemma 16, we have:

$$\|Q_U Q_U^\top \nabla f(X)\|_2 \cdot \|\Delta\|_F^2 = \widehat{\eta} \left(16 \underbrace{M \|X\|_2 \|Q_U Q_U^\top \nabla f(X)\|_2 \|\Delta\|_F^2}_{:=A} + 16 \|Q_U Q_U^\top \nabla f(X)\|_2^2 \cdot \|\Delta\|_F^2 \right)$$

At this point, we desire to introduce strong convexity parameter m and condition number κ in our bound. In particular, to bound term A, we observe that $\|Q_U Q_U^\top \nabla f(X)\|_2 \leq \frac{m\sigma_r(X)}{40\tau(U_r^\star)}$ or $\|Q_U Q_U^\top \nabla f(X)\|_2 \geq \frac{m\sigma_r(X)}{40\tau(U_r^\star)}$. This results into bounding A as follows:

$$\begin{aligned}
&M\|X\|_2 \|Q_U Q_U^\top \nabla f(X)\|_2 \|\Delta\|_F^2 \\
&\quad \leq \max \left\{ \tfrac{16 \cdot \widehat{\eta} \cdot M \|X\|_2 \cdot m\sigma_r(X)}{40\tau(U_r^\star)} \cdot \|\Delta\|_F^2,\ \widehat{\eta} \cdot 16 \cdot 40\tau(U_r^\star)\kappa\tau(X) \|Q_U Q_U^\top \nabla f(X)\|_2^2 \cdot \|\Delta\|_F^2 \right\} \\
&\quad \leq \tfrac{16 \cdot \widehat{\eta} \cdot M \|X\|_2 \cdot m\sigma_r(X)}{40\tau(U_r^\star)} \cdot \|\Delta\|_F^2 + \widehat{\eta} \cdot 16 \cdot 40\kappa\tau(X)\tau(U_r^\star) \|Q_U Q_U^\top \nabla f(X)\|_2^2 \cdot \|\Delta\|_F^2 .
\end{aligned}$$

Combining the above inequalities, we obtain:

$$\begin{aligned}
&\|Q_U Q_U^\top \nabla f(X)\|_2 \|\Delta\|_F^2 \qquad (53) \\
&\overset{(i)}{\leq} \tfrac{m\sigma_r(X)}{40\tau(U_r^\star)} \cdot \|\Delta\|_F^2 + (40\kappa\tau(X)\tau(U_r^\star) + 1) \cdot 16 \cdot \widehat{\eta} \|Q_U Q_U^\top \nabla f(X)\|_2^2 \cdot \|\Delta\|_F^2 \\
&\overset{(ii)}{\leq} \tfrac{m\sigma_r(X)}{40\tau(U_r^\star)} \cdot \|\Delta\|_F^2 + (41\kappa\tau(X_r^\star)\tau(U_r^\star) + 1) \cdot 16 \cdot \widehat{\eta} \|Q_U Q_U^\top \nabla f(X)\|_2^2 \cdot (\rho')^2 \sigma_r(X_r^\star) \\
&\overset{(iii)}{\leq} \tfrac{m\sigma_r(X)}{40\tau(U_r^\star)} \cdot \|\Delta\|_F^2 + 16 \cdot 42 \cdot \widehat{\eta} \cdot \kappa\tau(X_r^\star)\tau(U_r^\star) \cdot \|\nabla f(X) U\|_F^2 \cdot \tfrac{11(\rho')^2}{10} \\
&\overset{(iv)}{\leq} \tfrac{m\sigma_r(X)}{40\tau(U_r^\star)} \cdot \|\Delta\|_F^2 + \tfrac{2\widehat{\eta}}{25\tau(U_r^\star)} \cdot \|\nabla f(X) U\|_F^2 , \qquad (54)
\end{aligned}$$

where (i) follows from $\widehat{\eta} \leq \frac{1}{16M\|X\|_2}$, (ii) is due to Lemma 19 and bounding $\|\Delta\|_F \leq \rho' \sigma_r(U_r^\star)$ by the hypothesis of the lemma, (iii) is due to $\sigma_r(X^\star) \leq 1.1\sigma_r(X)$ by Lemma 19, $\sigma_r(X)\|Q_U Q_U^\top \nabla f(X)\|_2^2 \leq \|U^\top \nabla f(X)\|_F^2$ and $(41\kappa\tau(X_r^\star) + 1) \leq 42\kappa\tau(X_r^\star)$. Finally, (iv) follows from substituting ρ' and using Lemma 19.

From Lemma 16, we also have the following bound:

$$\|Q_{U^\star R} Q_{U^\star R}^\top \nabla f(X)\|_2 \leq \frac{102 \cdot 101\tau(U_r^\star)}{99 \cdot 100} \|Q_U Q_U^\top \nabla f(X)\|_2. \tag{55}$$

This follows from equation (34). Then, the proof completes when we combine the above two inequalities to obtain:

$$\left\langle \nabla f(X), \Delta\Delta^\top \right\rangle \geq -\left(\tfrac{6\widehat{\eta}}{25} \|\nabla f(X) U\|_F^2 + \tfrac{3m\sigma_r(X^\star)}{40} \cdot \|\Delta\|_F^2 \right)$$