On Global Linear Convergence in Stochastic Nonconvex Optimization for Semidefinite Programming

Jinshan Zeng, Ke Ma, and Yuan Yao

***Abstract*—Nonconvex reformulations via low-rank factorization for stochastic convex semidefinite optimization problem have attracted arising attention due to their empirical efficiency and scalability. Compared with the original convex formulations, the nonconvex ones typically involve much fewer variables, allowing them to scale to scenarios with millions of variables. However, it opens a new challenge that under what conditions the nonconvex stochastic algorithms may find the population minimizer within the optimal statistical precision despite their empirical success in applications. In this paper, we provide an answer that the stochastic gradient descent (SGD) method can be adapted to solve the nonconvex reformulation of the original convex problem, with a *global linear convergence* when using a fixed step size, i.e., converging exponentially fast to the population minimizer within an optimal statistical precision in the restricted strongly convex case. If a diminishing step size is adopted, the bad effect caused by the variance of gradients on the optimization error can be eliminated but the rate is dropped to be sublinear. The core of our treatment relies on a novel *second-order descent lemma*, which is more general than the existing best result in the literature and improves the analysis on both online and batch algorithms. The developed theoretical results and effectiveness of the suggested SGD are also verified by a series of experiments.**

***Index Terms*—Stochastic gradient descent, semidefinite optimization, low-rank factorization.**

Manuscript received July 25, 2018; revised March 22, 2019 and May 22, 2019; accepted June 6, 2019. Date of publication July 1, 2019; date of current version August 1, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yue Rong. The work of J. Zeng was supported in part by the National Natural Science Foundation of China under Grants 61603162 and 61876074, and in part by the Key Research Program of Jiangxi Province under Grant 20181ACE50029. The work of Y. Yao was supported in part by HKRGC under Grant 16303817, in part by 973 Program of China under Grant 2015CB85600, in part by the NNSF of China under Grants 61370004 and 11421110001, and in part by the grants from Tencent AI Lab, Si Family Foundation, Baidu BDI, and Microsoft Research-Asia. Part of this work was done while J. Zeng was visiting the Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong. *(Corresponding author: Jinshan Zeng.)*

J. Zeng is with the School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China (e-mail: jsh.zeng@gmail.com).

K. Ma is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: make@iie.ac.cn).

Y. Yao is with the Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong (e-mail: yuany@ust.hk).

Digital Object Identifier 10.1109/TSP.2019.2925609

I. Introduction

THE following semi-definite optimization problem, that is, minimizing a finite sum of convex functions with the positive semidefinite (PSD) constraint,

$$\min_{X\in\mathbb{R}^{p\times p}} f(X)=\frac{1}{n}\sum_{i=1}^{n} f_i(X) \text{ subject to } X\succeq 0, \quad (1)$$

has been found in a wide range of applications like matrix sensing [25], [40], ordinal embedding [1], [8], community detection [29], synchronization [6], 1-bit matrix completion [13], [18], phase retrieval [14], [15], subspace tracking [5], principle component analysis [4], [38], recommendation systems [22] and ptychography [24], etc.

Putting in a statistical learning background, problem (1) arises as empirical risk minimization where each f_i measures the loss on the i-th sample from the data set $\mathcal{Z}^n := \{Z_1,\ldots,Z_n\}$, which is generated independently from some identical yet unknown probability distribution $\mathbb{P}_{Z_i}$ with a low-rank parameter matrix $X\in\mathbb{S}^p_+$, where $\mathbb{S}^p_+$ denotes the set of symmetric and positive semidefinite matrices of the size $p\times p$. In this setting, $f_i(X)=\ell(X;Z_i)$ for some function $\ell:\mathbb{S}^p_+\times\mathcal{Z}\to\mathbb{R}$, often a negative logarithmic likelihood plussing a convex regularization [33], and in this case let

$$X^{tr}\in\arg\min_{X\in\mathbb{S}^p_+}\ \mathbb{E}_{\mathcal{Z}^n}[\ell(X;\mathcal{Z}^n)], \quad (2)$$

be any minimizer of the population risk $\mathbb{E}_{\mathcal{Z}^n}[\ell(X;\mathcal{Z}^n)]$ where $\mathbb{E}_{\mathcal{Z}^n}$ denotes the expectation over product measure $\prod_{i=1}^n \mathbb{P}_{Z_i}$. Since the distribution $\mathbb{P}_{Z_i}$ is generally unknown, thus, the empirical risk minimization problem (1) is commonly used as an alternative one. Henceforth, we call X^{tr} the *population minimizer*; a global optimum of problem (1), denoted by X^*, will be called as *empirical minimizer* whose rank is denoted by $r^*:=\mathrm{rank}(X^*)$. Thus, problem (1) can be reformulated as the following rank constrained problem,

$$\min_{X\in\mathbb{S}^p_+} f(X)=\frac{1}{n}\sum_{i=1}^{n} f_i(X) \text{ s.t. } \mathrm{rank}(X)\le r^*, \quad (3)$$

where X^* is also a global optimum of problem (3).

In applications, r^* is generally unknown and for the sake of reducing computational cost, one often looks for the following low rank approximate problem with $r\le r^*$ instead of solving

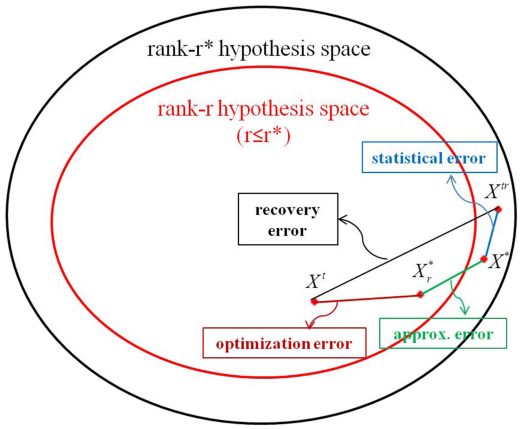


Fig. 1. Illustration of the error decomposition form (5).

the exact rank-constrained problem (3), that is,

$$\min_{X\in\mathbb{S}^p_+} \ f(X)=\frac{1}{n}\sum_{i=1}^{n} f_i(X) \text{ s.t. } \operatorname{rank}(X)\le r. \tag{4}$$

Iterative algorithms like (stochastic) gradient descent method are widely used to solve problem (4), which return a sequence $\{X^t\}_{t\in\mathbb{N}}$ of rank-r semidefinite matrices. Using the triangle inequality, the distance between the t-th iterate X^t and X^{tr}, called *recovery error* or *total error* in this paper, admits the following decomposition,

$$\underbrace{\|X^t-X^{tr}\|_F}_{\text{recovery error}} \le \underbrace{\|X^t-X_r^*\|_F}_{\text{optimization error}} + \underbrace{\|X_r^*-X^*\|_F}_{\text{approximation error}} + \underbrace{\|X^*-X^{tr}\|_F}_{\text{statistical error}}, \tag{5}$$

where X_r^* is the rank-r best approximation of X^* via truncated singular value decomposition (SVD). Such error decomposition form is illustrated in Figure 1.

There are three terms in the error decomposition (5). The first term characterizes the error between X^t and X_r^*, since the rank of X^t is at most r. Note that the first error term is mainly due to the used optimization algorithm that will be discussed below, and thus named as the *optimization error*. The second term represents the rank-r best approximation error, which can be generally bounded as follows:

$$(r^*-r)\cdot\sigma_{r^*}(X^*)\le\|X_r^*-X^*\|_F\le(r^*-r)\cdot\sigma_r(X^*),$$

where $\sigma_i(X^*)$ is the i-th largest singular value of X^*. It is obvious that $\|X_r^*-X^*\|_F=0$ when $r=r^*$ and $\|X_r^*-X^*\|_F>0$ when $r<r^*$. Such error is mainly introduced by the use of a smaller hypothesis space, i.e., the rank-r hypothesis space, to parsimoniously explore the larger model space where empirical minimizer lies in, and thus, called the *approximation error*. The last term implies the error between the empirical minimizer X^* and the population minimizer X^{tr}, and thus we name it as the *statistical error* or *sample/estimate error* throughout the rest of this paper.

There has been an extensive study in statistics on characterizing the statistical error bound $\|X^{tr}-X^*\|_F\le\mathcal{S}(n,\delta)$ in various scenarios, with probability at least $1-\delta$ ($\delta\in(0,1)$). For example, it is shown in [31], [32] that: under restricted strongly convex (RSC) condition, the minimax optimal error rates can be achieved at $\mathcal{S}(n,\delta)\approx O(\sqrt{\frac{r^*p}{n}\log p})$ for matrix sensing or completion problems, up to a logarithmic factor. In this case, the square loss with matrix nuclear norm regularization is chosen for $f_i(X)$ and X^{tr} recovers the unknown true parameter from Gaussian noise. Moreover, for 1-bit matrix completion problem using the logistic loss, it is shown in [13], [18] that the minimax optimal error rate $\mathcal{S}(n,\delta)\approx O(\sqrt{r^*p/n})$. Besides these special cases, a general treatment can be seen in [33], which enables us the following assumption in this paper.

Assumption 1 (Statistical error): Given n and a confidence $\delta\in(0,1)$, let $\mathcal{S}(n,\delta)$ be the optimal statistical error bound of problem (1), that is,

$$\|X^{tr}-X^*\|_F\le\mathcal{S}(n,\delta)\text{with probability at least } 1-\delta.$$

Moreover, it holds $\mathcal{S}(n,\delta)\to 0$ as $n\to\infty$.

Now the crucial thing is to bound the optimization error which relies on the particular algorithm in choice. There are many algorithms in optimization for solving problem (1), mainly including the traditional first-order methods like the well-known projected gradient descent method [34], interior point method [2], and more specialized path-following interior point methods which use the (preconditioned) conjugate gradient or residual scheme to compute the Newton direction (for more detail, see the survey [30] and references therein). However, most of the existing theoretical analysis of these traditional methods are based on the optimization framework, and mainly focus on the *(eventual) convergence and rate* of the iterative sequence itself, that is, under what conditions, the generated sequence converges as $t\to\infty$, and more frequently, to a stationary point, while the statistical error between the limiting point and the population minimizer X^{tr} is generally not taken into consideration. More importantly, most of these traditional optimization methods are not well-scalable in practice mainly due to the PSD constraint, i.e., $X\succeq 0$.

To circumvent this difficult constraint, the idea of low-rank factorization was adopted in optimization literature [11], [12] and became very popular in the past few years due to its empirical success (e.g., [26]). Specifically, given a small positive integer r, low-rank factorization recasts the rank-r approximate problem (4) into an unconstrained problem by introducing another rectangular matrix $U\in\mathbb{R}^{p\times r}$ with $r<p$ such that $X=UU^T$. Let $g(U):=f(UU^T)$ and problem (4) leads to,

$$\min_{U\in\mathbb{R}^{p\times r}} \ g(U). \tag{6}$$

Problems (1) and (6) will be equivalent when $r=r^*$ in the sense that problem (6) can find a global optimum X^* of problem (1) with $\operatorname{rank}(X^*)=r^*$. Since the PSD constraint has been eliminated, the recast problem (6) has a significant advantage over (1), but this benefit has corresponding cost: the objective function is no longer convex but instead *nonconvex* in general. Even for the simple first-order methods like the factored gradient descent (FGD) and gradient descent (GD) for problem (6), their *global linear convergence*[1] remains unspecified until the recent work

[1]By *global linear convergence*, it means that the algorithm converges to a global optimum exponentially fast.

in [7], [42]. The analysis in [7] only considers the optimization and approximation error terms in (5), while the analysis in [42] only considers the optimization and statistical error terms in (5) with $r = r^*$, under the statistical framework established in [31]. However, both FGD and GD suffer the scalability issue that when applied to the large scale applications with a big n, evaluation of n derivatives is expensive or prohibitive.

To meet this challenge, online or stochastic algorithms [37] have been widely adopted nowadays, that is, at each iteration, we only use the gradient information of one sample (or one mini-batch of samples), which is randomly chosen from the total samples. [19] proposed a nonconvex stochastic gradient descent (SGD) approach for the low-rank least squares problem with Gaussian ensembles, and developed its convergence based on the martingale technique. However, their analysis is only limited to the square loss and the noise models satisfying *rank-1 sampling condition* (see, [19, Condition 2]). Moreover, the exact rank r^* is imposed in their analysis. Recently, [27] proposed a SGD algorithm for online matrix completion problem and established its linear convergence based on the incoherence property. However, their analysis is limited to matrix completion with square loss and noiseless observations, and also focuses on the case $r = r^*$. It is still open how to deal with more general nonlinear loss functions such as noisy matrix sensing and nonmetric embedding. [45] proposed an accelerated SGD called stochastic variance-reduced gradient (SVRG) algorithm for matrix sensing and established the global linear convergence based on the restricted isometry property (RIP), which is generally stronger than the restricted strongly convex (RSC) condition. Later, [43] extended the algorithm in [45] to a more general low-rank matrix recovery problem and developed the global linear convergence mainly based on RSC condition and under the statistical framework established in [31]. Similar to [27], both [45] and [43] only considered the special case, i.e., $r = r^*$. So far, most of the existing works on stochastic algorithms do not consider the general case of $r < r^*$ with nonlinear losses, where the approximation error is no longer zero and should be handled via some special treatments, which is the gap to be filled in this paper. Moreover, although SVRG achieves linear convergence, it still needs to compute full gradients over all samples in its outer loop, which is impractical in some big data problems though sharing the same theoretical rate with the batch gradient descent. In this sense, SGD is still a fundamental solution for large scale applications. To our surprise, as we shall see soon, even the simple SGD equipped with a constant step size choice that is widely used in applications, can be shown to achieve global linear convergence (i.e., converging to a global optimum exponentially fast) toward an estimate with statistical precision. This is due to the existence of statistical error above, one does not need to reduce the variance of stochastic gradients for sequence convergence toward a particular minimizer; any point within a ball of optimal statistical precision will be good.

A. Contributions

Based on the error decomposition (5), we establish the *global linear convergence* (i.e., converging to a global optimum exponentially fast) and *early stopping criterion* of SGD for a more general nonlinear objective function and in the more practical algorithmic settings, that is, $r \leq r^*$ and the existence of statistical error, under the smooth and RSC assumptions, and certain an initialization scheme. Our main theorem can be stated as follows (its precise statement is shown in Theorem 2).

Main Theorem 1: Under the smooth, RSC and rank-r approximation error assumptions, SGD using a sufficiently small step size will converge to a small neighborhood of X_r^* exponentially fast with a proper initialization, which can be obtained easily by certain initialization scheme. Furthermore, suppose that the statistical error assumption holds, and if r is taken appropriately such that the rank-r approximation error is no more than the statistical precision $\mathcal{S}(n,\delta)$, then after $O(\log \frac{1}{\mathcal{S}(n,\delta)})$ iterations, with high probability, SGD can recover the population minimizer X^{tr} within a statistical precision $O(\mathcal{S}(n,\delta))$.

Besides the fixed step size above, we also establish the similar results for SGD with the diminishing step sizes, but the rate degenerates to be sublinear. It is worth noting that if there is no statistical error, the diminishing step sizes are generally necessary to eliminate the effect brought by the variance of samples.

B. Main Novelty

The main novelty of this paper is to establish the global linear convergence of SGD with a fixed step size for the low-rank stochastic semidefinite optimization problem, under some more general and weaker conditions than those in the literature. The comparisons between the existing works and this paper are presented in Table I. All these global linear convergence results are generally established by two steps, i.e., establishing the local convergence [2] to the global optimum firstly with a proper good initial guess, and then showing such proper initial guess can be obtained by some initialization schemes. For a fair comparison, we present the local convergence guarantees of all these algorithms, since their initialization schemes can be very similar, i.e., using one or several iterates of certain convex methods (say, projected gradient descent).

Compared to [27] and [45], this paper considers a general nonlinear loss, while [27] and [45] only consider the square loss. The main assumptions used in [27] and [45] respectively are the incoherence and restricted isometry property (RIP), which are generally stricter than the restricted strongly convex (RSC) condition used in this paper. Moreover, when applied to the problems considered in both [27] and [45] (in this case, $\kappa = 1$), the radius of the initialization ball required in this paper is generally larger than both of them (see, Table I).

Compared to [43], which extends [45] from the square loss to a general nonlinear loss, the radius of initialization ball of [43] is $\frac{2\sigma_r(X_r^*)}{15\kappa}$, while that of this paper is $\frac{(\sqrt{2}-1)\sigma_r(X_r^*)}{2\kappa}$, which is larger than $\frac{2\sigma_r(X_r^*)}{15\kappa}$, where κ is the "condition number" of the objective function f (to be specified in (10)), $\sigma_r(X_r^*)$ is the r-th largest singular value of the rank-r approximation X_r^* of the optimum X^* with $r \leq r^* := \mathrm{rank}(X^*)$. Moreover, [43] only considers the case of $r = r^*$, while this paper considers a general case of $r \leq r^*$. Similar claims also hold in the comparison between [42] and this paper.

[2]The *local convergence* means such kind of convergence starting from an initialization close to the global optimum.

TABLE I
COMPARISONS ON THE LOCAL CONVERGENCE GUARANTEES OBTAINED IN THE EXISTING WORKS AND THIS PAPER. THE NOTATION $\times$ MEANS SUCH KIND OF ERROR IS NOT CONSIDERED. THE TERMINOLOGIES IN THE FIRST ROW ARE THE ABBREVIATIONS OF "OPTIMIZATION ERROR", "APPROXIMATION ERROR", "STATISTICAL ERROR", "LOSS FUNCTION", "ASSUMPTION" AND "CONVERGENCE RATE", RESPECTIVELY. IT SHOULD BE POINTED OUT THAT [45] AND [43] CONSIDERED THE SVRG WITH OPTION-II (I.E., THE OUTPUT OF THE INNER LOOP IS RANDOMLY TAKEN FROM THE INNER LOOP UPDATES VIA A UNIFORM RANDOM WAY), WHILE [44] CONSIDERED THE SIMILAR CONVERGENCE OF SVRG WITH OPTION-I (I.E., THE OUTPUT OF THE INNER LOOP IS TAKEN AS THE LAST ITERATE OF THE INNER LOOP). THE LINEAR CONVERGENCE OF SVRG CONSIDERED IN [44] IS SHOWN IN THE METRIC $\tilde{\mathcal{E}}(U^t, U_r^*) := \|U^t - U_r^*\|_F^2$, UNDER THE INITIALIZATION REQUIREMENT $\tilde{\mathcal{E}}(U^0, U_r^*) = \mathcal{O}(\frac{\sigma_r(X^*)}{\kappa})$ AND RSC. IT CAN BE OBSERVED THAT THE METRIC $\tilde{\mathcal{E}}(U^0, U_r^*)$ IS GENERALLY STRICTER THAN $\mathcal{E}(U^0, U_r^*)$ USED IN THIS PAPER.

	opt err	approx err	statist err	loss	assump	$\|X^* - X_r^*\|_F$	$\mathcal{E}(U^0, U_r^*)$	rate
SGD [27]	✓	×	×	square	incoherence	0	$\frac{\sigma_r(X^*)}{10}$	linear
SVRG [45]	✓	×	✓	square	RIP	0	$\frac{\sigma_r(X^*)}{16}$	linear
SVRG [43]	✓	×	✓	general	RSC	0	$\frac{2\sigma_r(X^*)}{15\kappa}$	linear
GD [42]	✓	×	✓	general	RSC	0	$\frac{\sigma_r(X^*)}{10\kappa}$	linear
FGD [7]	✓	✓	×	general	RSC	$O(\frac{\sigma_r(X^*)}{\kappa^{1.5}\tau(X^*)})$	$O(\frac{\sigma_r(X^*)}{\kappa^2\tau^2(X^*)})$	linear
FGD (this paper)	✓	✓	✓	general	RSC	$O(\frac{\sigma_r(X^*)}{\kappa})$	$\frac{(\sqrt{2}-1)\sigma_r(X^*)}{2\kappa}$	linear
SGD (fixed, this paper)	✓	✓	✓	general	RSC	$O(\frac{\sigma_r(X^*)}{\kappa})$	$\frac{(\sqrt{2}-1)\sigma_r(X^*)}{2\kappa}$	linear

Compared to [7], this paper significantly improves the orders of both rank-r approximation error and the radius of initialization ball. More specifically, the requirement on the rank-r approximation error can be relaxed from the order $O(\frac{\sigma_r(X_r^*)}{\kappa^{1.5}\tau(X_r^*)})$ to $O(\frac{\sigma_r(X_r^*)}{\kappa})$, and the requirement on the radius of initialization can be relaxed from $O(\frac{\sigma_r(X_r^*)}{\kappa^2\tau^2(X_r^*)})$ to $O(\frac{\sigma_r(X_r^*)}{\kappa})$ (see Table I), where $\tau(X_r^*)$ is the condition number of X_r^*.

Besides more general cases considered and weaker assumptions imposed in this paper, the proof technique of this paper is also significantly different from the others. Our key technique for the proofs is the establishment of a novel *second-order descent lemma*, which generalizes the lower bound in [7, Lemma 14] with a larger domain. As a by-product, we significantly weaken the convergence conditions of FGD derived in [7] and then establish the similar results of Main Theorem 1 via applying our developed analysis framework (see, Table I). We can also show that the order $O(\sigma_r(X_r^*))$ on the radius of the initialization is tight in the sense that we can find a counter example such that FGD can not converge to the global optimum once the initialization radius is not smaller than $\sigma_r(X_r^*)$, as shown in Proposition 2. Moreover, the effectiveness as well as the developed theoretical results of the suggested SGD algorithm are verified by a series of numerical experiments.

C. Organization and Notations

The rest of this paper is organized as follows. Section II presents the main convergence results. Section III extends the developed framework to FGD. Section IV presents some related work. Section V provides a set of numerical experiments to demonstrate the effectiveness of the considered algorithm. We conclude this paper and point some future directions in Section VI. All the proofs are presented in Appendix.

For any $X, Y \in \mathbb{R}^{p\times p}$, their inner product is defined as $\langle X, Y\rangle = \mathrm{tr}(X^TY)$. For any $X \in \mathbb{R}^{p\times p}$, $\|X\|_F$ and $\|X\|_2$ denote its Frobenius and spectral norms, respectively, and $\sigma_{\min}(X)$ and $\sigma_{\max}(X)$ denote the smallest and largest *strictly positive* singular values of X, denote $\tau(X) := \frac{\sigma_{\max}(X)}{\sigma_{\min}(X)}$, with a slight abuse of notation, we also use $\sigma_1(X) \equiv \sigma_{\max}(X) \equiv \|X\|_2$, and X_r denotes the rank-r approximation of X via its truncated singular value decomposition (SVD) for any $r \le p$. $\mathbf{I}_p$ denotes the identity matrix with the size $p \times p$. We will omit the subscript p of $\mathbf{I}_p$ if there is no confusion.

II. MAIN RESULTS

In this section, we first adapt the stochastic gradient descent method to solve problem (1), then present our main results, and latter illustrate the novelty of our proof techniques.

A. Stochastic Gradient Descent Method

Throughout the paper, we assume that f is a symmetric function, i.e., $f(X) = f(X^T)$. Let $X = UU^T$, then the gradient of $g(U) := f(UU^T)$ is

$$\nabla g(U) = (\nabla f(UU^T) + \nabla f(UU^T)^T)U = 2\nabla f(X)U.$$

Factored Gradient Descent (FGD): FGD can be described as follows: let U^t be the t-th iterate and $X^t := U^t(U^t)^T$, then U^{t+1} is updated according to the following

$$U^{t+1} = U^t - \eta\nabla f(X^t)U^t, \tag{7}$$

where $\eta > 0$ is a step size.

Stochastic Gradient Descent (SGD): The SGD method for problem (1) can be described as follows: at the t-th iteration, pick an $i_t \in \{1, \ldots, n\}$ via a *uniformly random* way, and then

update the next iterate via

$$U^{t+1} = U^t - \eta_t \nabla f_{i_t}(X^t)U^t, \tag{8}$$

where $\eta_t > 0$ is a fixed (or diminishing) step size. Since i_t is uniformly sampled, it holds

$$\mathbb{E}_{i_t}[\nabla f_{i_t}(X^t)U^t] = \nabla f(X^t)U^t,\ \forall t \in \mathbb{N}. \tag{9}$$

Initialization: One of commonly used strategies is to construct the initialization directly from the observed data like in the applications of matrix sensing, matrix completion and phase retrieval (say, [7], [16], [17], [25], [35], [46]). Such strategy is generally effective for the case that the objective function has a small κ, generally the least squares case, while for the general function as considered in this paper, another common strategy is to use one of the standard convex algorithms (say, projected gradient descent (ProjGD)). Some specific implementations of this idea have been used in [7], [40], [42]. Motivated by the existing literature, this paper also suggests using the ProjGD method to generate the initialization U^0. More specifically,

a) let X^0 be the T-th iterate of ProjGD starting from zero (that is, $\tilde{X}^0 = 0$, $\tilde{X}^t = \mathrm{Proj}_{\mathbb{S}^p_+}(\tilde{X}^{t-1} - \frac{1}{L}\nabla f(\tilde{X}^{t-1}))$, $t = 1, \ldots, T$, $X^0 := \tilde{X}^T$);
b) then let the initialization U^0 be the rank-r factorization of X^0, i.e., $X^0 = U^0 U^{0T}$.

B. Assumptions

In the following, we provide some basic assumptions used in this paper, which are regular and from the existing literature (say, [7]).

Assumption 2: The objective function $f : \mathbb{S}^p_+ \to \mathbb{R}$ is L-smooth for some constant $L > 0$, i.e., its gradient ∇f is Lipschitz continuous with constant L. Moreover, f is (μ, r)-restricted strongly convex (RSC) for some constant $\mu > 0$ and a positive integer $r \leq r^*$, that is, it is convex, and for any $X, Y \in \mathbb{S}^p_+$ with at most rank r,

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

Such assumption is regular and widely used in literature (say, [33]). Let $U_r^* \in \mathbb{R}^{p\times r}$ be a rank-r factorization of X_r^* satisfying $U_r^* U_r^{*T} = X_r^*$. Based on Assumption 2, we define

$$\kappa := \frac{L}{\mu}, \gamma_0 := (\sqrt{2}-1)\kappa^{-1}, \tag{10}$$

$$\bar{\eta} := \min\left\{\frac{(1-\sqrt{\gamma_0})^2}{\frac{\|\nabla f(X_r^*)\|_F}{L\sigma_r(X_r^*)} + (2\sqrt{\gamma_0}+\gamma_0)\tau(U_r^*)}, 1\right\},$$

$$\xi := \bar{\eta}(1 - \bar{\eta}/2). \tag{11}$$

It is obvious that $0 < \xi \leq 1/2$. We also need an assumption on rank-r approximation error.

Assumption 3 (rank-r approximation error): The following holds

$$\|X_r^* - X^*\|_F < (\sqrt{2}-1)\xi^{1/2}\kappa^{-1}\cdot\sigma_r(X_r^*),$$

where κ and ξ are specified in (10) and (11), respectively, and $\sigma_r(X_r^*)$ is the r-th largest singular value of X_r^*.

Such assumption naturally holds for $r = r^*$, while when $r < r^*$, it might be satisfied if the singular values of X^* possess certain *compressible property* [3] or *near low-rank property* [31]. Under the above assumptions, we define several constants as follows:

$$\Delta := \frac{(\sqrt{2}-1)^2\xi^2\sigma_r^2(X_r^*)}{\kappa^2} - \xi\|X_r^* - X^*\|_F^2, \tag{12}$$

$$\gamma_l := \frac{(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{\kappa} - \sqrt{\Delta},$$

$$\gamma_u := \frac{(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{\kappa} + \sqrt{\Delta}, \tag{13}$$

$$B_0 := \sup_{\{U:\mathcal{E}(U,U_r^*)\leq\gamma_u\}} \mathbb{E}_i[\|\nabla f_i(UU^T)U - \nabla f(UU^T)U\|_F^2], \tag{14}$$

$$\eta_{\max} := \min\left\{\frac{L\Delta}{4\xi B_0}, \frac{\xi}{8L(\gamma_u + \|X_r^*\|_2)}\right\}. \tag{15}$$

For any $\eta \in (0, \eta_{\max})$, we define

$$\gamma_l^\eta := \frac{(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{\kappa} - \sqrt{\Delta - \frac{4\eta\xi B_0}{L}}, \tag{16}$$

$$\gamma_u^\eta := \frac{(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{\kappa} + \sqrt{\Delta - \frac{4\eta\xi B_0}{L}}. \tag{17}$$

It is easy to check $\gamma_l \leq \gamma_l^\eta \leq \gamma_u^\eta \leq \gamma_u \leq \gamma_0\sigma_r(X_r^*)$.

C. Linear Convergence of SGD With Fixed Step Size

In order to characterize the convergence of SGD, we need the following metric,

$$\mathcal{E}(U, V) := \min_{R\in\mathcal{O}} \|U - VR\|_F^2, \forall\, U, V \in \mathbb{R}^{p\times r},$$

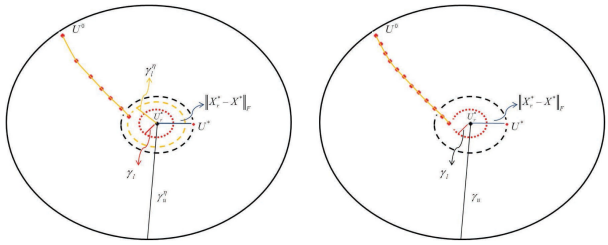
where $\mathcal{O}$ is the set of orthogonal matrices of size $r \times r$. Such metric has been used in [7] and some other references like [40]. We first give a locally linear convergence of SGD with a fixed step size starting from a proper good initialization within a prescribed ball, shown as follows.

Theorem 1 (Local linear convergence of SGD with $\eta_t \equiv \eta$): Let $\{U^t\}$ be a sequence generated by SGD (8). Let Assumptions 2 and 3 hold, and $\eta \in (0, \eta_{\max})$. The following hold:

1) Suppose that $\gamma_l^\eta < \mathcal{E}(U^0, U_r^*) < \gamma_u^\eta$. Let $\tilde{\rho} := 1 - \frac{\eta L}{2\xi}\cdot(\gamma_u^\eta - \mathcal{E}(U^0, U_r^*)) \in (0, 1)$, then
   a1) $\{\mathbb{E}[\mathcal{E}(U^t, U_r^*)]\}$ is decreasing,
   a2) $\mathbb{E}[\|X^t\|_2] \leq 2(\gamma_u + \|X_r^*\|_2)$, where $X^t = U^t(U^t)^T$, and
   a3) $\mathbb{E}[\mathcal{E}(U^t, U_r^*)] \leq (\mathcal{E}(U^0, U_r^*) - \gamma_l^\eta)\cdot\tilde{\rho}^t + \gamma_l^\eta$;
2) If $\mathcal{E}(U^0, U_r^*) \leq \gamma_l^\eta$, then $\mathbb{E}[\mathcal{E}(U^t, U_r^*)] \leq \gamma_l^\eta$ for any $t \in \mathbb{N}$.

The proof of Theorem 1 is provided in Appendix C. Parts (1) and (2) of Theorem 1 show certain *locally linear convergence* (i.e., converging to a global optimum starting from some proper initial point) of SGD, as depicted in Figure 2(a). From Figure 2(a), starting from an initialization lying in a γ_u^η-neighborhood of U_r^*, SGD converges exponentially fast until

[3] $\sigma_i(X^*)$ decays in a power law, i.e., $\sigma_i(X^*) \leq Ci^{-q}, i = 1, 2, \ldots, p$ for some constants $C, q > 0$.



(a) SGD (η, $O(\tilde{\rho}^t)$) (b) SGD (η_t, $O(\frac{1}{t})$)

Fig. 2. Local convergence paths of SGD with different step sizes.

achieving a much smaller γ_l^η-neighborhood of U_r^* in expectation, then stagnates and never jumps out from this smaller neighborhood. From (16)-(17) and by assumptions of Theorem 1, γ_l^η and γ_u^η satisfy

$$\frac{\left(\sqrt{2}+1\right)\kappa\|X^*-X_r^*\|_F^2}{2\sigma_r(X_r^*)}+\frac{2\left(\sqrt{2}+1\right)\eta B_0}{\mu\sigma_r(X_r^*)}\le\gamma_l^\eta \tag{18}$$

$$\le\frac{\left(\sqrt{2}+1\right)\kappa\|X^*-X_r^*\|_F^2}{\sigma_r(X_r^*)}+\frac{4\left(\sqrt{2}+1\right)\eta B_0}{\mu\sigma_r(X_r^*)}$$

$$<\frac{(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{\kappa}<\gamma_u^\eta$$

$$\le\frac{(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{\kappa}+\sqrt{\frac{(\sqrt{2}-1)^2\xi^2\sigma_r^2(X_r^*)}{\kappa^2}-\frac{4\eta\xi B_0}{L}}, \tag{19}$$

where the third inequality holds for $0<\eta<\eta_{\max}\le\frac{L\Delta}{4\xi B_0}$ and the definition of Δ (12). This implies that the radius of the limiting ball (i.e., γ_l^η) is intrinsically related to $\|X^*-X_r^*\|_F$ and ηB_0, while the radius of the initialization ball (i.e., γ_u^η) is intrinsically related to $\sigma_r(X_r^*)$. Particularly, when $r=r^*$,

$$\gamma_l^\eta\le\frac{4\left(\sqrt{2}+1\right)\eta B_0}{\mu\sigma_r(X_r^*)},\gamma_u^\eta>\frac{(\sqrt{2}-1)\xi\sigma_r(X^*)}{\kappa}. \tag{20}$$

Since η can be sufficiently small, (20) implies that the radius of the initialization ball is generally much bigger than the radius of the limiting ball.

Theorem 1 shows the locally linear convergence of SGD with a fixed step size starting from a proper good initial point. Thus, it is crucial to show whether such proper initial point can be easily achieved in practice. In the following, we provide a proposition to show that such a desired initial point can be easily obtained by the suggested initialization scheme, i.e., only few iterations of ProjGD are needed.

Proposition 1 (Feasibility of initialization scheme): Let Assumption 2 hold with $1<\kappa\le 64(\sqrt{2}-1)$ (in this case, $\xi=1/2$), and $r=r^*$. Let U^0 be generated by the initialization scheme described in Section II-A. If

$$T\ge\log_{1-\kappa^{-1}}\frac{(\sqrt{2}-1)^2\sigma_r^2(X^*)}{\kappa\|X^*\|_F^2}, \tag{21}$$

then $\mathcal{E}(U^0,U^*)\le\frac{(\sqrt{2}-1)\sigma_r(X^*)}{2\kappa}<\gamma_u^\eta$.

Similar result of Proposition 1 has been established in [42] for the gradient descent method. According to [42, Theorem 5.7], the condition on κ is $\kappa\in(1,4/3)$, while the requirement in Proposition 1 is $\kappa\in(1,64(\sqrt{2}-1)]$, which significantly relaxes the condition in [42]. The proof of Proposition 1 is provided in Appendix E.

With the help of Proposition 1, we can boost the locally linear convergence of SGD with a fixed step size shown in Theorem 1 to the following globally linear convergence.

Theorem 2 (Global linear convergence of SGD with $\eta_t\equiv\eta$): Let Assumption 2 hold with $1<\kappa\le 64(\sqrt{2}-1)$, and $r=r^*$. Let $\{U^t\}$ be a sequence generated by SGD with U^0 via the initialization scheme in Section II-A (where T satisfies (21)), and $\eta\in(0,\eta_{\max})$ (where $\eta_{\max}$ is defined in (15) with $\xi=1/2$). Then for any $t\in\mathbb{N}$,

$$\mathbb{E}[\mathcal{E}(U^t,U_r^*)]\le(\mathcal{E}(U^0,U_r^*)-\gamma_l^\eta)\cdot\tilde{\rho}^t+\gamma_l^\eta,$$

where γ_l^η is defined in (16), and $\tilde{\rho}$ is defined in Theorem 1.

Furthermore, suppose that Assumption 1 holds, and if $0<\eta<\min\{\eta_{\max},\frac{\mu\sigma_r(X_r^*)\mathcal{S}(n,\delta)}{4(\sqrt{2}+1)B_0}\}$, then after $\mathcal{T}^*=O(\log(\frac{1}{\mathcal{S}(n,\delta)}))$ iterations, with probability at least $1-\delta$, the following holds

$$\mathbb{E}[\|X^t-X^{tr}\|_F]=O(\mathcal{S}(n,\delta)),\ \forall t\ge\mathcal{T}^*.$$

As demonstrated by the final part of Theorem 2, if there is a statistical error (possibly introduced by the use of noisy data or finite samples), then it is not necessary for SGD to run amount of iterations to achieve a high convergence precision, but fewer iterations (in the order of $O(\log(1/\mathcal{S}(n,\delta)))$) are sufficient to achieve the statistical precision $\mathcal{S}(n,\delta)$ with high probability, as long as r is taken appropriately such that the rank-r approximation error is smaller than the statistical error. In this sense, the final part of Theorem 2 gives certain *early stopping criterion* of SGD. Moreover, Theorem 2 shows that the practical performance of SGD may be improved via gradually shrinking the step size (but not necessary diminishing to 0) during the update procedure. We leave this in our future work.

D. $O(1/t)$-Convergence of SGD With Diminishing Step Size

By Lemma 4 in Appendix A,

$$\|X^t-X_r^*\|_F\le(2+\sqrt{\gamma_0})\|U_r^*\|_2\sqrt{\mathcal{E}(U^t,U_r^*)}$$
$$\le 3\|U_r^*\|_2\sqrt{\mathcal{E}(U^t,U_r^*)}$$

due to $\gamma_0\le\sqrt{2}-1$. Based on this inequality and (18), even if $\|X_r^*-X^*\|_F=0$, the optimization error does not approach to 0 but to the order $O(\sqrt{\eta B_0\tau(X_r^*)/\mu})$ mainly due to the variance term B_0. In order to circumvent this issue, the diminishing step size is generally adopted for the update of SGD, that is, η_t is assumed to satisfy: $\sum_{t=0}^\infty\eta_t=\infty$ and $\sum_{t=0}^\infty\eta_t^2<\infty$. In this paper, without loss of generality, we choose

$$\eta_t=\frac{\eta}{t+1}, \tag{22}$$

for some $\eta\in(0,\eta_{\max})$, where $\eta_{\max}$ is specified in (15). Similarly, we can define two positive sequences $\{\gamma_l^t\}$ and $\{\gamma_u^t\}$ as follows

$$\gamma_l^t:=\frac{(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{\kappa}-\sqrt{\Delta-\frac{4\eta_t\xi B_0}{L}},$$
$$\gamma_u^t:=\frac{(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{\kappa}+\sqrt{\Delta-\frac{4\eta_t\xi B_0}{L}}. \tag{23}$$

Then γ_l^t is decreasing and γ_u^t is increasing with $\lim_{t\to\infty}\gamma_l^t = \gamma_l$ and $\lim_{t\to\infty}\gamma_u^t = \gamma_u$, and for any $t \in \mathbb{N}$, $\gamma_l \le \gamma_l^t \le \gamma_l^\eta \le \gamma_u^\eta \le \gamma_u^t \le \gamma_u$, where γ_l and γ_u are specified in (13).

Theorem 3 (Local convergence of SGD with diminishing η_t): Let $\{U^t\}$ be a sequence generated by SGD with η_t as in (22). Let Assumptions 2 and 3 hold. The following hold:

1) Assume that U^0 satisfies $\gamma_l < \mathcal{E}(U^0, U_r^*) < \gamma_u^\eta$. Then there hold
   a1) $\{\mathbb{E}[\mathcal{E}(U^t, U_r^*)]\}$ is decreasing,
   a2) $\mathbb{E}[\|X^t\|_2] \le 2(\gamma_u + \|X_r^*\|_2)$, and
   a3) $\mathbb{E}[\mathcal{E}(U^t, U_r^*)] - \gamma_l = O(\frac{1}{t})$.
2) If $\mathcal{E}(U^0, U_r^*) \le \gamma_l$, then $\mathbb{E}[\mathcal{E}(U^t, U_r^*)] \le \gamma_l, \forall t \in \mathbb{N}$.

Parts (1) and (2) of Theorem 3 establish the global convergence and sublinear rate of SGD with diminishing step sizes in expected error $\mathcal{E}(U^t, U_r^*)$. Its convergence path is illustrated in Figure 2(b). It is well known that the rate $O(1/t)$ is optimal for SGD in the vector setting (see, [9] and reference therein). Particularly, if $r = r^*$, then $\gamma_l = 0$. In this case, Theorem 3 shows that SGD with diminishing step sizes can exactly recover the global optimum X^* in expectation. However, when there is a statistical error, exact recovery of X^* might be not desired, instead, we only need to run $O\left(1/\mathcal{S}(n,\delta)\right)$ iterations such that the algorithm can find the population minimizer X^{tr} within an optimal statistical precision. It is worth noting that Theorem 3 also holds for generic diminishing step sizes. The proof of Theorem 3 is presented in Appendix D.

Similarly, by Theorem 3 and Proposition 1, we can establish the globally sublinear convergence of SGD with diminishing step size, shown as follows.

Theorem 4 (Global convergence of SGD with diminish. η_t): Let Assumption 2 hold with $1 < \kappa \le 64(\sqrt{2}-1)$, and $r = r^*$. Let $\{U^t\}$ be a sequence generated by SGD with U^0 via the initialization scheme in Section II-A (where T satisfies (21)), and η_t is defined as in (22). Then for $t = 1, 2, \ldots$,

$$\mathbb{E}[\mathcal{E}(U^t, U_r^*)] - \gamma_l = O\left(\frac{1}{t}\right).$$

Furthermore, suppose that Assumption 1 holds, then after $\mathcal{T}^* = O(\frac{1}{\mathcal{S}(n,\delta)})$ iterations, with probability at least $1-\delta$, the following holds

$$\mathbb{E}[\|X^t - X^{tr}\|_F] = O(\mathcal{S}(n,\delta)), \forall t \ge \mathcal{T}^*.$$

E. Main Technique Novelty: Second-Order Descent Lemma

Our key development for the proofs is a *second-order descent* lemma, which generalizes the lower bound in [7, Lemma 14] with a larger domain.

Lemma 1 (Second-order descent lemma): Let Assumption 2 hold. For any $U \in \mathbb{R}^{p\times r}$, let $X = UU^T$, Q_U be a basis of the column space of U, and $\mathcal{P}_U := Q_U Q_U^T$. If $\mathcal{E}(U, U_r^*) < \gamma_0\sigma_r(X_r^*)$ for some γ_0 specified in (10), then the following holds

$$2\langle\nabla f(X)U, U - V_U^*\rangle \ge (\sqrt{2}-1)\mu\sigma_r(X_r^*)\mathcal{E}(U, U_r^*) - \frac{L}{2\xi}\cdot\mathcal{E}^2(U, U_r^*) + \frac{\xi}{2L}\|\mathcal{P}_U\nabla f(X)\|_F^2 - \frac{L}{2}\|X^* - X_r^*\|_F^2, \tag{24}$$

where $V_U^* := U_r^* R_U^*$ and $R_U^* := \arg\min_{R\in\mathcal{O}}\|U - U_r^* R\|_F^2$.

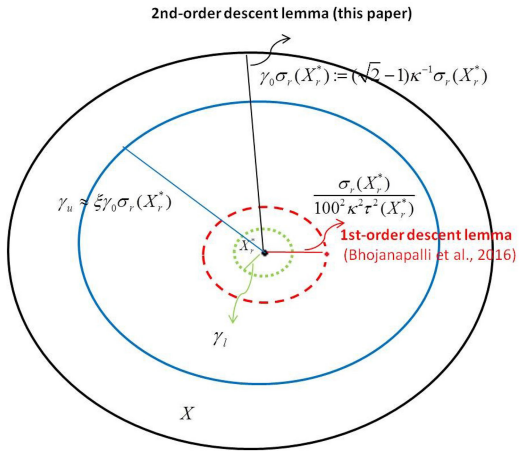


Fig. 3. Illustration on the second-order descent lemma (Lemma 1) with a larger domain of radius $O(\sigma_r(X_r^*)/\kappa)$ (black), than the previous (first-order) descent lemma with smaller domain of radius $O(\sigma_r(X_r^*)/\kappa^2\tau^2(X_r^*))$ (red).

Lemma 1 is similar to [7, Lemma 14], in which the bound is

$$2\langle\nabla f(X)U, U - V_U^*\rangle \ge 0.3\mu\sigma_r(X_r^*)\cdot\mathcal{E}(U, U_r^*) + \frac{4\eta}{3}\|\nabla f(X)U\|_F^2 - \frac{L}{2}\|X^* - X_r^*\|_F^2, \tag{25}$$

under the assumptions: $\mathcal{E}(U, U_r^*) \le \frac{\sigma_r(X_r^*)}{100^2\kappa^2\tau^2(X_r^*)} = O(\frac{\sigma_r(X_r^*)}{\kappa^2\tau^2(X_r^*)})$ which generally admits a smaller domain than ours: $\mathcal{E}(U, U_r^*) < \frac{(\sqrt{2}-1)\sigma_r(X_r^*)}{\kappa} = O(\frac{\sigma_r(X_r^*)}{\kappa})$, as depicted in Figure 3. To achieve such a more general lemma, in addition to the first order term of $\mathcal{E}(U, U_r^*)$ which has been used in literature, the second order term $\mathcal{E}^2(U, U_r^*)$ is introduced to characterize the lower bound of $2\langle\nabla f(X)U, U - V_U^*\rangle$, where the name *second-order descent lemma* comes. It leads to our relaxed conditions for both SGD and FGD (shown latter). Proof details are provided in Appendix B.

III. Extensions and Discussions

In this section, we apply the established proof technique to FGD and improve the existing convergence results in [7], then give some discussions on the radius of the initialization ball.

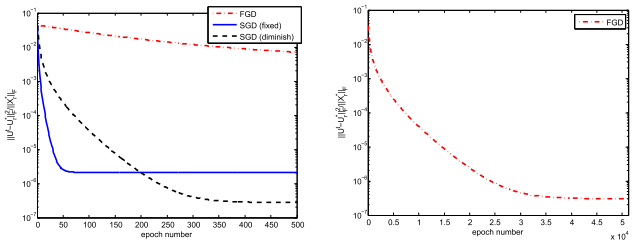
According the similar proof of Theorem 1 with $B_0 = 0$, we can easily derive the following improved convergence results of FGD.

Corollary 1 (Global linear convergence of FGD): Let $\{U^t\}$ be a sequence generated by FGD (7). Let Assumptions 2 and 3 hold, and $0 < \eta < \frac{\xi}{8(\gamma_u + \|X_r^*\|_2)L}$. Then all claims in Theorem 1 (1) and (2) hold without expectation, where γ_l^η, γ_u^η and $\tilde{\rho}$ in Theorem 1 should be replaced by γ_l, γ_u and $\rho := 1 - \frac{\eta L}{2\xi}\cdot(\gamma_u - \mathcal{E}(U^0, U_r^*)) \in (0,1)$, respectively.

Moreover, if $1 < \kappa \le 64(\sqrt{2}-1)$, $r = r^*$, and T satisfies (21), then for any $t \in \mathbb{N}$,

$$\mathcal{E}(U^t, U_r^*) \le (\mathcal{E}(U^0, U_r^*) - \gamma_l)\cdot\rho^t + \gamma_l.$$

Furthermore, suppose that Assumption 1 holds, then after $\mathcal{T}^* = O(\log(\frac{1}{\mathcal{S}(n,\delta)}))$ iterations, with probability at least $1-\delta$,



(a) Iterative error (b) Iterative error of FGD

Fig. 4. Experiments for matrix sensing problem. It requires about 350 and 4×10^4 epochs for SGD (diminish) and FGD, respectively, to achieve the precision 3.2×10^{-7}.

the following holds

$$\|X^t - X^{tr}\|_F = O(\mathcal{S}(n, \delta)), \forall t \geq \mathcal{T}^*.$$

From Corollary 1, we significantly weaken the convergence conditions of FGD derived in [7]. More specifically, the requirement on the rank-r approximation error can be relaxed from the order $O(\frac{\sigma_r(X_r^*)}{\kappa^{1.5}\tau(X_r^*)})$ to $O(\frac{\sigma_r(X_r^*)}{\kappa})$, and the requirement on the radius of initialization can be relaxed from $O(\frac{\sigma_r(X_r^*)}{\kappa^2\tau^2(X_r^*)})$ to $O(\frac{\sigma_r(X_r^*)}{\kappa})$, where κ is the "condition number" of the objective function f (specified in (10)), $\sigma_r(X_r^*)$ and $\tau(X_r^*)$ are respectively the r-th largest singular value and the condition number of X_r^*. We can also show that the order $O(\sigma_r(X_r^*))$ on the radius of the initialization is tight in the sense that we can find a counter example such that FGD can not converge to the global optimum once the initialization radius is not smaller than $\sigma_r(X_r^*)$.

Proposition 2 (Necessity of order $O(\sigma_r(X_r^*))$): Suppose that $r = r^*$. There exists a counter example (shown in Appendix F) such that $\mathcal{E}(U^0, U_r^*) = \sigma_r(X_r^*)$ and the other conditions in Corollary 1 hold, but FGD does not converge to X^*.

IV. Related Work

It is well-known that the most intuitive benefit of stochastic gradient descent (SGD) method is that SGD employs information (a cheaper gradient at each iteration) more efficiently than the batch gradient descent method, while its main disadvantage is the slower convergence due to the variance of the stochastic direction [9]. To address this limitation, methods endowed with *variance reduction* capabilities have been developed. One of the typical methods is the stochastic variance reduced gradient (SVRG) [28]. Two loops including an inner loop and outer loop, are employed in the iterate of SVRG. Two options are introduced to update the outer loop from the inner loop, one is to take the last iterate of the inner loop (Option I) and the other one is to pick randomly from the inner loop (Option II). In the vector case (i.e., the argument variable is vector), the linear convergence of outer loop with Option II is proved in [28], while the linear convergence of SVRG with Option I (as well as adopting Barzilai-Borwein step size strategy) is provided in [39]. More specifically, SVRG performs linear convergence (in expectation) of the outer loop via mainly exploiting an inner loop to reduce the variance of the stochastic direction. As we can see from Theorem 1, the *negative effect* of the variance introduced by the stochastic direction can not be eliminated if we use a fixed step size, while it can be eliminated via using the diminishing step size as shown in Theorem 3, however, in the diminishing step-size case, the convergence speed degrades to be a sublinear rate (i.e., $O(\frac{1}{t})$).

Besides such type of convergence results developed in this paper, there is another type of the convergence results of SGD and SVRG for nonconvex optimization, that is, the convergence to a stationary point, which is generally characterized by the decay of the term $\mathbb{E}[\|\nabla f(x^t)\|_2^2]$. Such kind of results can be found in many recent works (say, [3], [21], [36]). In [21], the convergence rate of SGD is proved to be $O(\frac{1}{\sqrt{T}})$ in the sense that $\mathbb{E}[\|\nabla f(x^t)\|_2^2] \leq O(\frac{1}{\sqrt{T}}),\ t = 1, \ldots, T$. Such rate can be improved to $O(\frac{1}{T})$ for SVRG as shown in [3], [36]. It can be observed that this type of convergence results is very different with the type established in this paper, in the sense that the convergence results established in [3], [21], [36] focus on the convergence to a stationary point and the convergence rates are non-asymptotic, while the convergence results obtained in this paper are about the convergence to a global optimum and the convergence rates are asymptotic.

V. Numerical Experiments

In this section, we implement two applications to show the effectiveness of the proposed algorithm and also verify our developed theoretical results. The code is available in https://github.com/alphaprime/Stochastic_Factored_Gradient_Descent.

A. Matrix Sensing

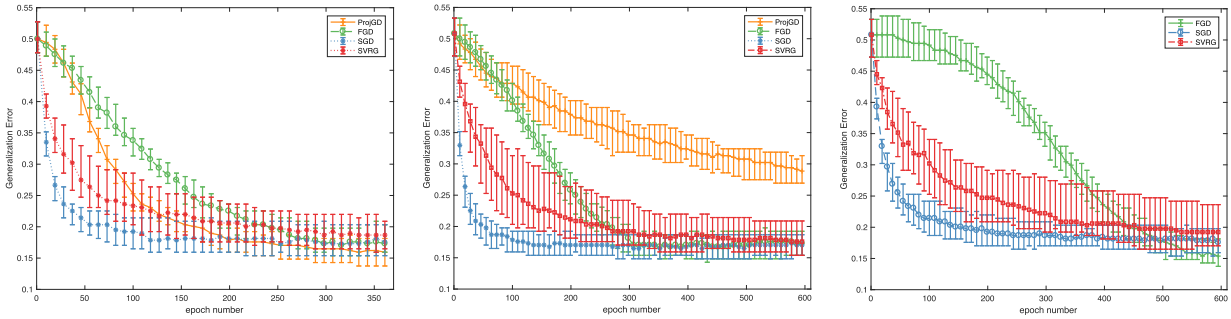
We consider the following matrix sensing problem

$$\min_{X \succeq 0} f(X) = \frac{1}{2n} \sum_{i=1}^{n} (b_i - \langle A_i, X \rangle)^2,$$

where $X \in \mathbb{R}^{p \times p}$ is a low-rank matrix, $A_i \in \mathbb{R}^{p \times p}$ is a sub-Gaussian independent measurement matrix of the i-th sample, $b_i \in \mathbb{R}$, and $n \in \mathbb{N}$ is the sample size.

Specifically, we let $p = 128$, the optimal matrix $X^* := U^* U^{*T}$ be a low-rank matrix with $\mathrm{rank}(X^*) = 4$ and the sample size $n = 8p$. For both FGD and SGD, we take $r = r^*$, and construct the initialization U^0 empirically via 10 iterations of projected gradient descent. For FGD, we use $\eta = \frac{1}{4L\|X^*\|_F}$, where L is the Lipschitz constant of ∇f. For SGD, we use a more consecutive fixed step size $\hat{\eta} = \frac{1}{8L\|X^*\|_F}$ and a diminishing step size $\eta_t = \frac{\hat{\eta}}{t+1}$. Henceforth, we denote SGD (fixed) and SGD (diminish) as abbreviations of SGD with a fixed step size and SGD with diminishing step sizes, respectively. The experiment results are shown in Figure 4. An epoch of SGD includes n iterations of SGD, and an epoch of FGD is exactly an iteration of FGD. The iterative error curves of SGD and FGD are shown along epochs since both of them exploit a full scan of gradients over sample per epoch and their computational complexities per epoch are thus comparable. From Figure 4, we can observe the following phenomena:

- SGD can significantly speed up FGD in the sense that much fewer epochs are required for SGD to achieve the same precision. Specifically, about 350 and 4×10^4 epochs are required for SGD (diminish) and FGD, respectively, to



(a) GNMDS (b) STE (c) TSTE

Fig. 5. Experiments for ordinal embedding problem using different models. From (a)-(c), SGD is much faster than both ProjGD and FGD, and also slightly faster than SVRG in terms of epoch.

achieve the same limiting precisions. The speedup of SGD in the perspective of epoch number is more than 100 times of FGD. This exhibits the main advantage of SGD over FGD.

- In the beginning, SGD (fixed) will be much faster than SGD (diminish) until SGD (fixed) attains to its limitation, i.e., γ_l^η. This is because the rate of SGD (fixed) is $O((\tilde{\rho}^n)^t)$, while that of SGD (diminish) is $O(\frac{1}{nt})$, where t is the epoch number. Moreover, we can also observe that SGD (diminish) is much faster than FGD at the initial several epochs. This is mainly due to that $n = 1024$ and thus, $\frac{1}{nt}$ will be much smaller than $O(\rho^t)$ when t is relatively small.

These experiment results demonstrate the effectiveness of SGD and also verify our developed theoretical results.

B. Ordinal Embedding

In this subsection, we apply SGD to the ordinal embedding problem, which aims to learn representation of data objects as points in a low-dimensional space. There are mainly three existing models to deal with this problem, i.e., the generalized nonmetric multidimensional scaling (GNMDS) [1], the stochastic triplet embedding (STE) and its variant replacing the Gaussian kernel in STE with Student-t kernel (called TSTE for short) [41]. Their objective functions are shown as follows

$$f(X) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \ell_c(X) + \lambda \cdot \mathrm{tr}(X),$$

where $\mathcal{C}$ is a set of ordinal constraints, $|\mathcal{C}|$ is its cardinality, and ℓ_c is some loss function (say, the hinge loss for GNMDS, the scale-invariant loss with Gaussian kernel for STE, and the scale-invariant loss with Student-t kernel for TSTE). In this application, we compare the performance of SGD with the state-of-the-art methods including the projected gradient descent (ProjGD) method, FGD and an accelerated SGD method, i.e., SVRG suggested in [43],[4] to show its effectiveness.

Music artist dataset: We implement SGD on a real world dataset called *Music artist dataset*, collected by [20] via a web-based survey. In this dataset, there are 1032 users and 412 music artists. The number of triplets on the similarity between music artists is 213472. A *triplet* (i, j, k) indicates an ordinal constraint like $d_{ij}^2(X) \le d_{ik}^2(X)$, which means that "*music artist i is more similar to artist j than artist k*", where $d_{ij}^2(X)$ is the Euclidean distance between artists i and j, $i, j, k \in \{1, \ldots, p\}$, and p is the number of total kinds of music artists. Specifically, we use the data pre-processed by [41] via removing the inconsistent triplets from the original dataset. In this dataset, there are 9107 triplets for $p = 400$ artists. The genre labels for all artists are gathered using Wikipedia, to distinguish nine music genres (rock, metal, pop, dance, hip hop, jazz, country, gospel, and reggae). Thus, the desired dimension $r^* = 9$.

For each method, we implement independently 50 trials, and then record: (a) the generalization error and (b) the running time. For each trail, 80% triplets are randomly picked as the training set and the rest as the test set. All three methods start with the same initial point, which is chosen according to the suggested initialization scheme with $T = 10$. Each curve in Figure 5 shows the trend of generalization error of one method with respect to the epoch number.

From Figure 5, SGD can significantly speed up the batch methods in terms of epoch number for all three models. Particularly, the generalization error curve of SGD decays much faster than those of batch methods at the first 50 epochs, which means that SGD can quickly get an admissible result (say, for GNMDS model, its generalization error after 50 epochs is about 0.2). However, if a better generalization error is desired, then the advantage of SGD in the computational cost will gradually lose. Similar phenomenon can be also observed from Table II in terms of running time. From Table II, when the training error desired is 0.15, there are about 30 times and 6 times speedup for ProjGD and FGD, respectively. Moreover, by Figure 5 and Table II, the performance of SGD is also slightly better than SVRG in terms of both generalization error and training time. The outperformance of SGD over SVRG in terms of generalization error might be mainly due to the existence of statistical error (in this case, one may not need to reduce the variance of stochastic gradients for sequence convergence toward a particular minimizer).

[4]As considered in [43], the output of the inner loop of SVRG is randomly chosen from all the iterates in the inner loop via a uniformly random way, i.e., *Option II* is adopted.

TABLE II
COMPARISONS ON CPU TIME (SECOND.) OF DIFFERENT ALGORITHMS FOR THE MUSIC ARTIST DATASET. THE VALUES IN THE FIRST COLUMN MEAN THE TRAINING ERROR (TRAINERR) ACHIEVED. NAN IN THIS TABLE MEANS THAT THE CONSIDERED ALGORITHM CAN NOT ACHIEVE THE GIVEN PRECISION WITHIN AN ACCEPTABLE CPU TIME.

GNMDS						
TrainErr	method	min	max	mean	median	std
0.15	ProjGD	19.1460	37.7370	33.0843	36.4140	5.9570
	FGD	3.1730	3.7340	3.5511	3.5800	0.1261
	SVRG	6.5610	35.2910	14.9224	11.7270	8.9589
	SGD	**0.3750**	**0.5720**	**0.4921**	**0.4935**	**0.0414**
0.10	ProjGD	33.2510	37.9230	36.6465	36.8795	1.0384
	FGD	3.9260	4.4100	4.2234	4.2405	0.0952
	SVRG	12.8870	35.4390	24.1533	23.2580	8.4652
	SGD	**0.9950**	**1.3780**	**1.2292**	**1.2360**	**0.0878**
0.05	ProjGD	35.9540	38.2200	37.0647	37.0735	0.4511
	FGD	4.8840	5.3810	5.1745	5.1815	0.1065
	SVRG	33.9010	35.4390	34.6991	34.7435	0.4319
	SGD	**3.7970**	**3.9900**	**3.8739**	**3.8785**	**0.0390**
STE						
TrainErr	method	min	max	mean	median	std
0.15	ProjGD	NaN	NaN	NaN	NaN	NaN
	FGD	2.6080	3.0540	2.7410	2.7320	0.0883
	SVRG	7.5170	59.1890	23.3683	13.5385	19.1085
	SGD	**0.3060**	**0.4550**	**0.3784**	**0.3780**	**0.0094**
0.10	ProjGD	NaN	NaN	NaN	NaN	NaN
	FGD	3.2740	3.6270	3.4012	3.3985	0.0844
	SVRG	16.3070	59.3330	35.7929	30.1635	17.1202
	SGD	**0.8550**	**1.2100**	**1.0327**	**1.0370**	**0.0732**
0.05	ProjGD	NaN	NaN	NaN	NaN	NaN
	FGD	**4.2310**	**4.7280**	**4.4184**	**4.4090**	**0.1008**
	SVRG	50.5800	60.8430	57.6998	57.9515	1.6374
	SGD	6.7680	7.7830	7.4048	7.4020	0.1346
TSTE						
TrainErr	method	min	max	mean	median	std
0.15	FGD	9.4100	10.5050	9.7593	9.7470	0.2095
	SVRG	6.3840	NaN	23.0009	15.7890	17.8931
	SGD	**0.5310**	**0.7240**	**0.6065**	**0.5990**	**0.0481**
0.10	FGD	10.7330	11.8990	11.1459	11.1325	0.2223
	SVRG	13.9850	NaN	36.9182	31.8430	18.8322
	SGD	**1.2880**	**1.9390**	**1.6284**	**1.6345**	**0.1560**
0.05	FGD	12.6700	14.0660	13.1275	13.1020	0.2716
	SVRG	37.8560	NaN	56.5836	60.3290	6.7433
	SGD	**8.2410**	**8.7180**	**8.4136**	**8.3985**	**0.0895**

VI. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we consider a nonlinear stochastic semidefinite optimization problem in the scenario of statistical learning. An error decomposition consisting of approximation error, optimization error, and statistical error, is proposed for the analysis of an algorithm designed for this kind of problem. Stochastic gradient descent method is particularly considered in this paper. Specifically, under assumptions of RSC, smoothness and rank-r approximation error, we can show that SGD with a fixed (diminishing) step size can converge to a smaller neighborhood of a global optimum at a linear (sublinear) rate as long as the initialization lies in a neighborhood of the optimum, which can be easily achieved by the suggested initialization scheme. Furthermore, if the rank r is taken appropriately such that the rank-r approximation error is smaller than the optimal statistical error $\mathcal{S}(n,\delta)$, then after about $O(-\log \mathcal{S}(n,\delta))$ (or $O(1/\mathcal{S}(n,\delta))$) iterations, SGD with a fixed (or diminishing) step size may approach the population minimizer within the statistical precision. To establish the convergence of SGD, we develop a novel *second-order descent lemma*, which improves and generalizes the best existing ones in literature. It is left open whether the prescribed initialization ball has tightest rates.

APPENDIX

In the appendix, we present the main proofs of this paper.

A. Preliminary Lemmas

We firstly provide several preliminary lemmas, which will be frequently used in our proofs.

Lemma 2 (Lemma 5.4 in [40]): For any $U, V \in \mathbb{R}^{p\times r}$, then we have

$$\|UU^T - VV^T\|_F^2 \geq 2(\sqrt{2}-1)\sigma_r^2(V)\cdot \mathcal{E}(U,V).$$

Lemma 3: For any $U \in \mathbb{R}^{p\times r}$, let $X = UU^T$. If $\mathcal{E}(U, U_r^*) \leq \gamma\sigma_r(X_r^*)$ for some constant $0 < \gamma < 1$, then

$$\sigma_r(X) \geq (1-\sqrt{\gamma})^2\sigma_r(X_r^*).$$

Proof: Using the norm ordering $\|\cdot\|_2 \leq \|\cdot\|_F$ and the Weyl's inequality for perturbation of singular values (see, [23, Theorem 3.3.16]), we get

$$|\sigma_i(U) - \sigma_i(U_r^*)| \leq \sqrt{\gamma}\sigma_r(U_r^*), 1 \leq i \leq r, \tag{26}$$

which implies that

$$\sigma_r(U) \geq (1-\sqrt{\gamma})\sigma_r(U_r^*).$$

Thus, $\sigma_r(X) \geq (1-\sqrt{\gamma})^2\sigma_r(X_r^*)$. ■

Lemma 4: For any $U \in \mathbb{R}^{p\times r}$, let $X = UU^T$. If $\mathcal{E}(U, U_r^*) \leq \gamma\sigma_r(X_r^*)$ for some constant $\gamma > 0$, then

$$\begin{aligned}\|X - X_r^*\|_F &\leq (2+\sqrt{\gamma})\|U_r^*\|_2\mathcal{E}^{1/2}(U, U_r^*)\\ &\leq (2\sqrt{\gamma}+\gamma)\cdot\tau(U_r^*)\cdot\sigma_r(X_r^*),\end{aligned}$$

where $\tau(U_r^*) := \frac{\sigma_1(U_r^*)}{\sigma_r(U_r^*)}$.

Proof: Let $R_U^* := \arg\min_{R\in\mathcal{O}}\|U - U_r^*R\|_F^2$ and $V_U^* := U_r^*R_U^*$. Note that

$$\begin{aligned}\|X - X_r^*\|_F &= \|UU^T - U_r^*R_U^*{R_U^*}^T{U_r^*}^T\|_F\\ &= \|U(U-V_U^*)^T + (U-V_U^*){V_U^*}^T\|_F\\ &\leq (\|U\|_2 + \|V_U^*\|_2)\|U - V_U^*\|_F\\ &\leq (2+\sqrt{\gamma})\|U_r^*\|_2\mathcal{E}^{1/2}(U, U_r^*),\end{aligned}$$

where the first inequality holds for the triangle inequality and the inequality: $\|AB^T\|_F \leq \|A\|_2\cdot\|B\|_F$ for any two matrices with the same sizes, where A is full-column rank, and the second inequality is due to (26). Substituting the hypothesis of this lemma yields the final inequality of the lemma. ■

For any matrix $U \in \mathbb{R}^{p\times r}$, let Q_U be a basis of the column space of U. Denote $\mathcal{P}_U := Q_UQ_U^T$. Then $\mathcal{P}_U \cdot U = U$. For any

matrix $Y \in \mathbb{R}^{p\times p}$, $\mathcal{P}_U Y$ is a projection of Y onto the subspace spanned by $X := UU^T$.

Lemma 5: Let Assumption 2 hold. For any $U \in \mathbb{R}^{p\times r}$, let $X = UU^T$. If $\mathcal{E}(U, U_r^*) < \gamma_0 \sigma_r(X_r^*)$, where γ_0 is specified in (10), then the following hold:

a) $\|\nabla f(X)\|_F \le \|\nabla f(X_r^*)\|_F + (2\sqrt{\gamma_0} + \gamma_0) L\tau(U_r^*)\sigma_r(X_r^*)$,

b) $\bar{X} := X - \frac{\bar{\eta}}{L}\mathcal{P}_U \nabla f(X)\mathcal{P}_U$ is symmetric and positive semidefinite with rank r, where $\bar{\eta}$ is specified in (11),

c) $(\mathbf{I} - \mathcal{P}_U)X_r^* = 0$.

Proof:

a) Note that

$$\begin{aligned}\|\nabla f(X)\|_F &\le \|\nabla f(X_r^*)\|_F + L\|X - X_r^*\|_F \\ &\le \|\nabla f(X_r^*)\|_F + (2\sqrt{\gamma_0} + \gamma_0) \\ &\quad \times L\tau(U_r^*)\sigma_r(X_r^*),\end{aligned}$$

where the first inequality holds for the L-smoothness of f, and the second inequality holds for Lemma 4.

b) Since $X - \frac{\bar{\eta}}{L}\cdot \mathcal{P}_U\nabla f(X)\mathcal{P}_U = \mathcal{P}_U(X - \frac{\bar{\eta}}{L}\cdot \nabla f(X))\mathcal{P}_U$, thus, the i-th eigenvalue $\lambda_i(X - \frac{\bar{\eta}}{L}\cdot \mathcal{P}_U\nabla f(X)\mathcal{P}_U) = 0$ for $i = r+1, \ldots, p$. While for any $i = 1, \ldots, r$,

$$\begin{aligned}&\lambda_i\left(X - \frac{\bar{\eta}}{L}\cdot \mathcal{P}_U\nabla f(X)\mathcal{P}_U\right) \\ &\ge \lambda_i(X) - \frac{\bar{\eta}}{L}\cdot \lambda_{\max}(\mathcal{P}_U\nabla f(X)\mathcal{P}_U) \\ &\ge \sigma_r(X) - \frac{\bar{\eta}}{L}\cdot \sigma_{\max}(\nabla f(X)) \ge (1 - \sqrt{\gamma_0})^2\sigma_r(X_r^*) \\ &- \frac{\bar{\eta}}{L}\cdot (\|\nabla f(X_r^*)\|_F + (2\sqrt{\gamma_0} + \gamma_0)L\tau(U_r^*) \\ &\sigma_r(X_r^*)) \ge 0,\end{aligned}$$

where the third inequality holds for Lemma 3 and **(a)** of this lemma, and the final inequality holds for the definition of $\bar{\eta}$ (11). Therefore, $\bar{X}$ is positive semidefinite.

c) By $\mathcal{E}(U, U_r^*) < \gamma_0\sigma_r(X_r^*)$ and $0 < \gamma_0 \le \sqrt{2} - 1$, we have $\sigma_i(X)\cdot\sigma_i(X_r^*) > 0,\ i \in \{1, \ldots, r\}$ and

$$\sigma_i(X_r^*) = 0,\ \sigma_i(X) = 0, i \in \{r+1, \ldots, p\},$$

which implies that X_r^* lies in the subspace spanned by X. In other words, X_r^* does not lie in the orthogonal subspace of the subspace spanned by X, that is, the following holds

$$(\mathbf{I} - \mathcal{P}_U)X_r^* = 0.$$

Thus, we end the proof of this lemma. ■

B. Proof of Lemma 1

Note that

$$\begin{aligned}&2\langle \nabla f(X)U, U - V_U^*\rangle = 2\langle \nabla f(X), X - V_U^* U^T\rangle \\ &= \langle \nabla f(X), X - X_r^*\rangle + \langle \nabla f(X), X + X_r^* - 2V_U^* U^T\rangle,\end{aligned} \tag{27}$$

where the lower bounds of these two terms in the right-hand side of the above equality are provided by the following two lemmas, respectively. Thus, substituting (28) and (30) into (27) yields the claim of this lemma.

Lemma 6: Under conditions of Lemma 1, there holds

$$\begin{aligned}&\langle \nabla f(X), X + X_r^* - 2V_U^* U^T\rangle \\ &\ge -\frac{\xi}{2L}\|\mathcal{P}_U\nabla f(X)\|_F^2 - \frac{L}{2\xi}\cdot \mathcal{E}^2(U, U_r^*),\end{aligned} \tag{28}$$

where ξ is specified in (11).

Proof: Note that

$$\begin{aligned}&\langle \nabla f(X), X + X_r^* - 2V_U^* U^T\rangle \\ &= \langle \mathcal{P}_U\nabla f(X) + (\mathbf{I} - \mathcal{P}_U)\nabla f(X), X + X_r^* - 2V_U^* U^T\rangle \\ &= \langle \mathcal{P}_U\nabla f(X), X + X_r^* - 2V_U^* U^T\rangle \\ &= \langle \mathcal{P}_U\nabla f(X), (U - V_U^*)(U - V_U^*)^T\rangle \\ &\ge -\frac{(1 - \bar{\eta}/2)\bar{\eta}}{2L}\|\mathcal{P}_U\nabla f(X)\|_F^2 - \frac{L}{2\bar{\eta}(1 - \bar{\eta}/2)}\cdot \mathcal{E}^2(U, U_r^*),\end{aligned} \tag{29}$$

where the second equality is due to $\langle(\mathbf{I} - \mathcal{P}_U)\nabla f(X), X\rangle = 0$, $\langle(\mathbf{I} - \mathcal{P}_U)\nabla f(X), V_U^* U^T\rangle = 0$ and $\langle(\mathbf{I} - \mathcal{P}_U)\nabla f(X), X_r^*\rangle = 0$ by $(\mathbf{I} - \mathcal{P}_U)U = 0$ and Lemma 5(c), and the last equality holds for $X_r^* = U_r^* U_r^{*T} = (U_r^* R_U^*)(U_r^* R_U^*)^T = V_U^* V_U^{*T}$, and the inequality holds for the basic inequality: $\langle Y, Z\rangle \ge -\frac{\delta}{2}\|Y\|_F^2 - \frac{1}{2\delta}\|Z\|_F^2$ for any $Y, Z \in \mathbb{R}^{p\times p}$ and $\delta > 0$. ■

Lemma 7: Under conditions of Lemma 1, there holds

$$\begin{aligned}&\langle \nabla f(X), X - X_r^*\rangle \ge \frac{\xi}{L}\cdot \|\mathcal{P}_U\nabla f(X)\|_F^2 \\ &+ (\sqrt{2} - 1)\mu\sigma_r(X_r^*)\mathcal{E}(U, U_r^*) - \frac{L}{2}\|X^* - X_r^*\|_F^2,\end{aligned} \tag{30}$$

where ξ is specified in (11).

Proof: To bound $\langle \nabla f(X), X - X_r^*\rangle$, we utilize the following three inequalities mainly by the L-smoothness and (μ, r)-restricted strong convexity of f, that is,

$$\text{(i) } f(X_r^*) \ge f(X) + \langle \nabla f(X), X_r^* - X\rangle + \frac{\mu}{2}\|X_r^* - X\|_F^2,$$

$$\text{(ii) } f(X) \ge f(X^*) + (1 - \bar{\eta}/2)\bar{\eta}L^{-1}\cdot \|\mathcal{P}_U\nabla f(X)\|_F^2,$$

$$\text{(iii) } f(X^*) \ge f(X_r^*) - \frac{L}{2}\|X^* - X_r^*\|_F^2,$$

where (i) holds for the (μ, r)-restricted strong convexity of f, (ii) holds for the following inequality induced by the L-smoothness of f, i.e.,

$$\begin{aligned}&f(X) \ge f(\bar{X}) + \langle \nabla f(X), X - \bar{X}\rangle - \frac{L}{2}\|X - \bar{X}\|_F^2 \\ &(\text{where } \bar{X} := X - \frac{\bar{\eta}}{L}\mathcal{P}_U\nabla f(X)\mathcal{P}_U) \\ &= f(\bar{X}) + (1 - \bar{\eta}/2)\bar{\eta}L^{-1}\cdot \|\mathcal{P}_U\nabla f(X)\|_F^2 \\ &(\because \langle \nabla f(X), \mathcal{P}_U\nabla f(X)\mathcal{P}_U\rangle = \|\mathcal{P}_U\nabla f(X)\|_F^2),\end{aligned}$$

and $f(\bar{X}) \ge f(X^*)$ since X^* is an optimum and $\bar{X}$ is a feasible point by Lemma 5(b), and (iii) holds for the L-smoothness of f

and the optimality condition $\nabla f(X^*)U^* = 0$, i.e.,

$$f(X_r^*) \le f(X^*) + \langle \nabla f(X^*), X^* - X_r^* \rangle + \frac{L}{2}\|X^* - X_r^*\|_F^2 = f(X^*) + \frac{L}{2}\|X^* - X_r^*\|_F^2,$$

where the equality holds for $\nabla f(X^*)U^* = 0$, which directly implies the following facts: $\nabla f(X^*)U_r^* = 0$, $\nabla f(X^*)X^* = 0$ and $\nabla f(X^*)X_r^* = 0$ due to $X^* = U^*U^{*T}$ and $X_r^* = U_r^*U_r^{*T}$. Summing the inequalities (i)–(iii) yields

$$\begin{aligned}&\langle \nabla f(X), X - X_r^* \rangle \\ &\ge \frac{\mu}{2}\|X - X_r^*\|_F^2 + \xi L^{-1} \cdot \|\mathcal{P}_U \nabla f(X)\|_F^2 - \frac{L}{2}\|X^* - X_r^*\|_F^2 \\ &\ge (\sqrt{2}-1)\mu\sigma_r(X_r^*)\mathcal{E}(U, U_r^*) - \frac{L}{2}\|X^* - X_r^*\|_F^2 \\ &\quad + \xi L^{-1} \cdot \|\mathcal{P}_U \nabla f(X)\|_F^2,\end{aligned}$$

where the second inequality is due to Lemma 2, i.e., $\|X - X_r^*\|_F^2 \ge 2(\sqrt{2}-1)\sigma_r(X_r^*)\mathcal{E}(U, U_r^*)$. ■

C. Proof of Theorem 1

To prove Theorem 1, we first justify the following lemma.

Lemma 8: Let Assumptions 2 and 3 hold. Given the current iterate U, let $X = UU^T$, i_t be randomly chosen from $\{1, \ldots, n\}$, η be a fixed step size, and U^+ be the next iterate updated via (8). If $\|X\|_2 \le \mathcal{B}$ for some constant $\mathcal{B} > 0$, $0 < \eta < \frac{\xi}{4L\mathcal{B}}$, and $\mathcal{E}(U, U_r^*) < \gamma_u^\eta$, then there holds

$$\begin{aligned}&\mathbb{E}_{i_t}[\mathcal{E}(U^+, U_r^*)] \\ &\le \mathcal{E}(U, U_r^*) - \frac{\eta L}{2\xi} \times \left(\frac{2(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{\kappa} \cdot \mathcal{E}(U, U_r^*) \right. \\ &\left. -\mathcal{E}^2(U, U_r^*) - \xi\|X_r^* - X^*\|_F^2 - \frac{4\eta\xi B_0}{L} \right). \end{aligned} \tag{31}$$

Proof: Given the current iterate U, let

$$R_U^* := \arg\min_{R \in \mathcal{O}} \|U - U_r^* R\|_F^2, V_U^* := U_r^* R_U^*.$$

Then $\mathcal{E}(U, U_r^*) = \|U - V_U^*\|_F^2$. Note that

$$\begin{aligned}&\mathcal{E}(U^+, U_r^*) = \min_{R \in \mathcal{O}} \|U^+ - U_r^* R\|_F^2 \\ &\le \|U^+ - V_U^*\|_F^2 = \|(U^+ - U) + (U - V_U^*)\|_F^2 \\ &= \|U - V_U^*\|_F^2 + 2\langle U^+ - U, U - V_U^* \rangle + \|U^+ - U\|_F^2 \\ &= \mathcal{E}(U, U_r^*) - 2\eta\langle \nabla f_{i_t}(X)U, U - V_U^* \rangle + \eta^2 \|\nabla f_{i_t}(X)U\|_F^2,\end{aligned}$$

and

$$\begin{aligned}&\|\nabla f_{i_t}(X)U\|_F^2 = \|\nabla f_{i_t}(X)U - \nabla f(X)U + \nabla f(X)U\|_F^2 \\ &\le 2(\|\nabla f_{i_t}(X)U - \nabla f(X)U\|_F^2 + \|\nabla f(X)U\|_F^2).\end{aligned}$$

Thus,

$$\begin{aligned}&\mathcal{E}(U^+, U_r^*) \le \mathcal{E}(U, U_r^*) - 2\eta\langle \nabla f_{i_t}(X)U, U - V_U^* \rangle \\ &+ 2\eta^2\|\nabla f_{i_t}(X)U - \nabla f(X)U\|_F^2 + 2\eta^2\|\nabla f(X)U\|_F^2.\end{aligned}$$

Taking expectation of both sides of the above inequality over i_t, by (9) and the definition of B_0 (14), we have

$$\begin{aligned}&\mathbb{E}_{i_t}[\mathcal{E}(U^+, U_r^*)] \le \mathcal{E}(U, U_r^*) \\ &- 2\eta\langle \nabla f(X)U, U - V_U^* \rangle + 2\eta^2 B_0 + 2\eta^2\|\nabla f(X)U\|_F^2 \\ &\le \mathcal{E}(U, U_r^*) - 2\eta\langle \nabla f(X)U, U - V_U^* \rangle + 2\eta^2 B_0 \\ &+ 2\eta^2\|X\|_2 \cdot \|\mathcal{P}_U \nabla f(X)\|_F^2,\end{aligned}$$

where the final inequality holds for

$$\|\nabla f(X)U\|_F^2 = \|\mathcal{P}_U \nabla f(X)U\|_F^2 + \|(\mathbf{I} - \mathcal{P}_U)\nabla f(X)U\|_F^2,$$

$\|(\mathbf{I} - \mathcal{P}_U)\nabla f(X)U\|_F^2 = 0$ and $\|\mathcal{P}_U \nabla f(X)U\|_F^2 \le \|X\|_2 \|\mathcal{P}_U \nabla f(X)\|_F^2$. By Lemma 1, it follows

$$\begin{aligned}&\mathbb{E}_{i_t}[\mathcal{E}(U^+, U_r^*)] \le \mathcal{E}(U, U_r^*) + \frac{\eta L}{2\xi}\mathcal{E}^2(U, U_r^*) \\ &- \left(\sqrt{2}-1\right)\eta\mu\sigma_r(X_r^*)\mathcal{E}(U, U_r^*) + \frac{\eta L}{2}\|X^* - X_r^*\|_F^2 + 2\eta^2 B_0 \\ &- \eta\|\mathcal{P}_U \nabla f(X)\|_F^2 \left(\frac{\xi}{2L} - 2\eta\|X\|_2\right) \\ &\le \mathcal{E}(U, U_r^*) + \frac{\eta L}{2\xi} \cdot \mathcal{E}^2(U, U_r^*) \\ &- \left(\sqrt{2}-1\right)\eta\mu\sigma_r(X_r^*)\mathcal{E}(U, U_r^*) \\ &+ \frac{\eta L}{2}\|X^* - X_r^*\|_F^2 + 2\eta^2 B_0,\end{aligned} \tag{32}$$

where the second inequality is due to the assumptions of this lemma, i.e., $\eta \le \frac{\xi}{4L\mathcal{B}}$ and $\|X\|_2 \le \mathcal{B}$. Therefore, we prove this lemma. ■

Based on Lemma 8, we prove Theorem 1.

Proof of Theorem 1: From Lemma 8 and by the definitions of γ_l^η and γ_u^η (17), we can easily verify that if $\gamma_l^\eta < \mathbb{E}[\mathcal{E}(U^t, U_r^*)] < \gamma_u^\eta$ for any $t \in \mathbb{N}$, there hold

i) $\mathbb{E}[\mathcal{E}(U^{t+1}, U_r^*)] < \mathbb{E}[\mathcal{E}(U^t, U_r^*)]$,

ii) $\|X^t\|_2 = \|U^t - U_r^* R_{U^t}^* + U_r^* R_{U^t}^*\|_2^2 \le 2(\|U^t - U_r^* R_{U^t}^*\|_2^2 + \|X_r^*\|_2) \le 2(\mathcal{E}(U^t, U_r^*) + \|X_r^*\|_2)$,

where $R_{U^t}^* := \arg\min_{R \in \mathcal{O}} \|U^t - U_r^* R\|_F^2$. Since $\gamma_l^\eta < \mathcal{E}(U^0, U_r^*) < \gamma_u^\eta$, then inductively by (i), $\{\mathbb{E}[\mathcal{E}(U^t, U_r^*)]\}$ is monotonically decreasing, and $\mathbb{E}[\mathcal{E}(U^t, U_r^*)] \le \mathcal{E}(U^0, U_r^*)$, which together with (ii) shows that $\mathbb{E}[\|X^t\|_2] \le 2(\mathcal{E}(U^0, U_r^*) + \|X_r^*\|_2) \le 2(\gamma_u + \|X_r^*\|_2) =: \mathcal{B}$. Thus, (a1) and (a2) hold.

Moreover, by (31) and the definitions of γ_l^η and γ_u^η (17), we have

$$\begin{aligned}&\mathbb{E}[\mathcal{E}(U^t, U_r^*)] \le \mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] \\ &- \frac{\eta L}{2\xi} \cdot (\gamma_u^\eta - \mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)]) \cdot (\mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] - \gamma_l^\eta) \\ &\le \mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] \\ &- \frac{\eta L}{2\xi} \cdot (\gamma_u^\eta - \mathcal{E}(U^0, U_r^*)) \cdot (\mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] - \gamma_l^\eta),\end{aligned} \tag{33}$$

where the second inequality holds for $\gamma_l^\eta < \mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] \leq \mathcal{E}(U^0, U_r^*) < \gamma_u^\eta$ for any $t \in \mathbb{N}_+$. The above inequality implies

$$\begin{aligned}&\mathbb{E}[\mathcal{E}(U^t, U_r^*)] - \gamma_l^\eta \\ &\leq \left[1 - \frac{\eta L}{2\xi} \cdot (\gamma_u^\eta - \mathcal{E}(U^0, U_r^*))\right] \cdot (\mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] - \gamma_l^\eta),\end{aligned} \tag{34}$$

which shows the linear convergence (a3).

For any $z \in \mathbb{R}_+$, we define a univariate function $h(z) = z - \frac{\eta L}{2\xi} \cdot \left((\gamma_l^\eta + \gamma_u^\eta)z - z^2 - \gamma_l^\eta \gamma_u^\eta\right)$. Then its derivative is

$$\begin{aligned}h'(z) &= 1 - \frac{\eta L}{2\xi} \cdot (\gamma_l^\eta + \gamma_u^\eta) + \frac{\eta L}{\xi} \cdot z \\ &> 1 - \frac{\gamma_l^\eta + \gamma_u^\eta}{16(\gamma_u + \|X_r^*\|_2)} + \frac{\eta L}{\xi} \cdot z > 0, \forall z \leq \gamma_l^\eta,\end{aligned}$$

where the first inequality holds for $\eta < \eta_{\max} \leq \frac{\xi}{8L(\gamma_u + \|X_r^*\|_2)}$. Thus, for any $0 < z \leq \gamma_l^\eta$,

$$h(z) \leq h(\gamma_l^\eta) = \gamma_l^\eta,$$

which shows that **(2)** holds.

In the following, we prove the final part of this theorem. From (17),

$$\begin{aligned}&\frac{\left(\sqrt{2}+1\right)\kappa\|X^* - X_r^*\|_F^2}{2\sigma_r(X_r^*)} + \frac{2\left(\sqrt{2}+1\right)\eta B_0}{\mu\sigma_r(X_r^*)} \leq \gamma_l^\eta \\ &\leq \frac{\left(\sqrt{2}+1\right)\kappa\|X^* - X_r^*\|_F^2}{\sigma_r(X_r^*)} + \frac{4\left(\sqrt{2}+1\right)\eta B_0}{\mu\sigma_r(X_r^*)}.\end{aligned}$$

Let $C := \sqrt{\mathcal{E}(U^0, U_r^*) - \gamma_l^\eta}$. Applying the inequality (i.e., $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$) twice to the inequality in **(a3)** yields

$$\begin{aligned}\sqrt{\mathbb{E}[\mathcal{E}(U^t, U_r^*)]} &\leq C(\sqrt{\tilde{\rho}})^t + \sqrt{\gamma_l^\eta} \\ &\leq C(\sqrt{\tilde{\rho}})^t + \frac{\sqrt{3\kappa}\|X^* - X_r^*\|_F}{\sigma_r(U_r^*)} + \sqrt{\frac{10\eta B_0}{\mu\sigma_r(X_r^*)}}.\end{aligned}$$

By Lemma 4,

$$\begin{aligned}\|X^t - X_r^*\|_F &\leq (2 + \sqrt{\gamma_0})\|U_r^*\|_2\sqrt{\mathcal{E}(U^t, U_r^*)} \\ &\leq 3\|U_r^*\|_2\sqrt{\mathcal{E}(U^t, U_r^*)},\end{aligned}$$

where the second inequality is due to $\gamma_0 \leq \sqrt{2} - 1$. Thus, for any $t \in \mathbb{N}$,

$$\begin{aligned}&\mathbb{E}[\|X^t - X_r^*\|_F] \leq 3C\|U_r^*\|_2(\sqrt{\tilde{\rho}})^t \\ &+ 3\sqrt{3\kappa}\tau(U_r^*)\|X^* - X_r^*\|_F + 3\tau(U_r^*)\sqrt{10\eta B_0/\mu}.\end{aligned} \tag{35}$$

Substituting (35) into (5) yields

$$\begin{aligned}&\mathbb{E}[\|X^t - X^{tr}\|_F] \leq 3C\|U_r^*\|_2(\sqrt{\tilde{\rho}})^t + \mathcal{S}(n, \delta) \\ &+ (3\sqrt{3\kappa}\tau(U_r^*) + 1)\|X^* - X_r^*\|_F + 3\tau(U_r^*)\sqrt{10\eta B_0/\mu}.\end{aligned} \tag{36}$$

Based on (36), we can claim the final part of this theorem. Thus, we finish the proof. ■

D. Proof of Theorem 3

Proof: Note that (31) still holds for the diminishing step size (22) for each $t \in \mathbb{N}$, that is,

$$\begin{aligned}&\mathbb{E}_{i_t}[\mathcal{E}(U^{t+1}, U_r^*)] \leq \mathcal{E}(U^t, U_r^*) \\ &- \frac{\eta_t L}{2\xi} \cdot \left(\frac{2(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{\kappa} \cdot \mathcal{E}(U^t, U_r^*) - \mathcal{E}^2(U^t, U_r^*)\right. \\ &\left. - \xi\|X_r^* - X^*\|_F^2 - \eta_t \cdot \frac{4\xi B_0}{L}\right).\end{aligned} \tag{37}$$

Based on (37) and following the similar derivations of the proof of Theorem 1, we can claim that Theorem 3 (a1) and (a2) hold. In the following, we mainly prove Theorem 3 (a3). By (37) and according to the similar derivations of (33), for any $t \in \mathbb{N}_+$, we have

$$\begin{aligned}&\mathbb{E}[\mathcal{E}(U^t, U_r^*)] \leq \mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] - \frac{\eta_{t-1}L}{2\xi} \\ &\times (\gamma_u^{t-1} - \mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)]) \cdot (\mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] - \gamma_l^{t-1}) \\ &\leq \mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] - \frac{\eta_{t-1}L}{2\xi} \cdot (\gamma_u^\eta - \mathcal{E}(U^0, U_r^*)) \\ &\cdot (\mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] - \gamma_l^{t-1}),\end{aligned} \tag{38}$$

where the second inequality holds for $\gamma_u^t \geq \gamma_u^0 = \gamma_u^\eta$ and by Theorem 3 (a1), $\mathbb{E}[\mathcal{E}(U^{t-1}, U_r^*)] \leq \mathcal{E}(U^0, U_r^*)$. Let $a_t := \mathbb{E}[\mathcal{E}(U^t, U_r^*)] - \gamma_l$ and $b_t := \gamma_l^t - \gamma_l, \forall t \in \mathbb{N}$, and $c := \frac{\eta L}{2\xi}(\gamma_u^\eta - \mathcal{E}(U^0, U_r^*))$. (38) implies

$$a_t \leq \left(1 - \frac{c}{t}\right)a_{t-1} + \frac{c \cdot b_{t-1}}{t}. \tag{39}$$

Note that

$$\begin{aligned}b_t &= \sqrt{\Delta} - \sqrt{\Delta - \frac{4\eta_t\xi B_0}{L}} \\ &= \frac{4\eta\xi B_0}{L\left(\sqrt{\Delta} + \sqrt{\Delta - \frac{4\eta\xi B_0}{(t+1)L}}\right)} \cdot \frac{1}{t+1} \\ &\leq \frac{4\eta\xi B_0}{2L\sqrt{\Delta - \frac{4\eta\xi B_0}{L}}} \cdot \frac{1}{t+1} := \frac{c_1}{t+1},\end{aligned} \tag{40}$$

where $c_1 := \frac{2\eta\xi B_0}{L\sqrt{\Delta - \frac{4\eta\xi B_0}{L}}}$. Substituting (40) into (39) yields

$$a_t \leq \left(1 - \frac{c}{t}\right)a_{t-1} + \frac{c \cdot c_1}{t^2}. \tag{41}$$

Based on (41), it is easy to check that

$$a_t = O\left(\frac{c_1}{t+1}\right).$$

Thus, we have proved the first two parts of this theorem. While the final part of this theorem can be easily proved via the similar proof of Theorem 1. ■

E. Proof of Proposition 1

Proof: We first show $\xi = \frac{1}{2}$ when $1 < \kappa \leq 64(\sqrt{2}-1)$ and $r = r^*$. By the definition of ξ (11), it suffices to show

$$\frac{(1-\sqrt{\gamma_0})^2}{(2\sqrt{\gamma_0}+\gamma_0)\tau(U_r^*)} \geq 1.$$

By the above inequality, the definition of γ_0 (10) and some basic derivations, the above inequality holds if

$$\frac{4(\sqrt{2}-1)(\tau(U_r^*)-1)^2}{(\sqrt{\tau^2(U_r^*)+3\tau(U_r^*)}+\tau(U_r^*)+1)^2} \leq \kappa \leq 4(\sqrt{2}-1)(\sqrt{\tau^2(U_r^*)+3\tau(U_r^*)}+\tau(U_r^*)+1)^2.$$

Since $\tau(U_r^*) \geq 1$ and $\kappa > 1$ by their definitions, it is easy to check that

$$1 < \kappa \leq 64(\sqrt{2}-1)$$

implies the above inequality.

Then we show the computational complexity of such initialization scheme. From [10, Theorem 3.6], we have that for consecutive updates $\tilde{X}^{t-1}$, $\tilde{X}^t$, and the optimum X^*, projected gradient descent satisfies:

$$\|\tilde{X}^t - X^*\|_F^2 \leq (1-\kappa^{-1}) \cdot \|\tilde{X}^{t-1} - X^*\|_F^2,$$

which implies for any $t \in \mathbb{N}$

$$\|\tilde{X}^t - X^*\|_F^2 \leq (1-\kappa^{-1})^t \cdot \|\tilde{X}^0 - X^*\|_F^2 = (1-\kappa^{-1})^t \cdot \|X^*\|_F^2.$$

Thus, by the hypothesis of T,

$$\|X^0 - X^*\|_F^2 = \|\tilde{X}^T - X^*\|_F^2 \leq \frac{2(\sqrt{2}-1)^2\xi\sigma_r^2(X^*)}{\kappa}.$$

By Lemma 2, there holds

$$\mathcal{E}(U^0, U^*) \leq \frac{\|X^0 - X^*\|_F^2}{2(\sqrt{2}-1)\sigma_r(X^*)} \leq \frac{(\sqrt{2}-1)\xi\sigma_r(X^*)}{\kappa} < \gamma_u^\eta.$$

Therefore, we end the proof. ■

F. Proof of Proposition 2

Proof: Given a positive integer $r^* \geq 2$, consider the following optimization problem

$$\min_{X \in \mathbb{R}^{(r^*+2)\times(r^*+2)}} f(X) = \frac{1}{2}\|X - A\|_F^2 \text{ s.t. } X \succeq 0, \quad (42)$$

where

$$A = \begin{pmatrix} \mathbf{I}_{\mathbf{r^*-1}} & \mathbf{0}_{r^*-1} & \mathbf{0}_{r^*-1} & \mathbf{0}_{r^*-1} \\ \mathbf{0}_{r^*-1}^T & 1 & 0 & 0 \\ \mathbf{0}_{r^*-1}^T & 0 & 0 & 0 \\ \mathbf{0}_{r^*-1}^T & 0 & 0 & -1 \end{pmatrix},$$

$\mathbf{I}_{r^*-1}$ is an identity matrix of the size $(r^*-1)\times(r^*-1)$, $\mathbf{0}_{r^*-1} \in \mathbb{R}^{(r^*-1)\times 1}$ denotes an all 0's vector, and $\mathbf{0}_{r^*-1}^T$ denotes the transpose of $\mathbf{0}_{r^*-1}$. It is easy to check that the global optimum of the problem (42) is

$$X^* = \begin{pmatrix} \mathbf{I}_{\mathbf{r^*-1}} & \mathbf{0}_{r^*-1} & \mathbf{0}_{r^*-1} & \mathbf{0}_{r^*-1} \\ \mathbf{0}_{r^*-1}^T & 1 & 0 & 0 \\ \mathbf{0}_{r^*-1}^T & 0 & 0 & 0 \\ \mathbf{0}_{r^*-1}^T & 0 & 0 & 0 \end{pmatrix},$$

and $\mathrm{rank}(X^*) = r^*$, and

$$U^* = \begin{pmatrix} \mathbf{I}_{\mathbf{r^*-1}} & \mathbf{0}_{r^*-1} \\ \mathbf{0}_{r^*-1}^T & 1 \\ \mathbf{0}_{r^*-1}^T & 0 \\ \mathbf{0}_{r^*-1}^T & 0 \end{pmatrix}.$$

Moreover, it is obvious that the constant $\kappa = 1$ in this case. In the following, we apply FGD to solve this problem. Particularly, we take $r = r^*$, then $U_r^* = U^*, \sigma_r(X_r^*) = 1, \tau(X_r^*) = 1$ and $\gamma_l = 0$ in this case. We consider the following initialization

$$U^0 = \begin{pmatrix} \mathbf{I}_{\mathbf{r^*-1}} & \mathbf{0}_{r^*-1} \\ \mathbf{0}_{r^*-1}^T & 0 \\ \mathbf{0}_{r^*-1}^T & 0 \\ \mathbf{0}_{r^*-1}^T & 0 \end{pmatrix}.$$

Then $\mathcal{E}(U^0, U^*) = 1 = \sigma_r(X_r^*)$, and it is easy to check that

$$U^0 {U^0}^T U^0 = U^0, AU^0 = U^0.$$

Thus, $(U^0{U^0}^T - A)U^0 = 0$, and

$$U^1 = U^0 - \eta(U^0{U^0}^T - A)U^0 = U^0, \forall \eta > 0,$$

which implies that FGD converges from the first step, that is, $U^t = U^0, \forall t \in \mathbb{N}$. However, $\mathcal{E}(U^0, U^*) = 1$ but not 0. ■

ACKNOWLEDGMENT

The authors would like to thank three anonymous reviewers and the Associate Editor for their constructive and helpful comments.

REFERENCES

[1] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, and S. Belongie, "Generalized non-metric multidimensional scaling," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2007, pp. 11–18.

[2] F. Alizadeh, "Interior point methods for semidefinite programming with applications to combinatorial optimization," *SIAM J. Optim.*, vol. 5, no. 1, pp. 13–51, 1995.

[3] Z. Allen-Zhu and E. Hazan, "Variance reduction for faster non-convex optimization," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 699–707.

[4] R. Arora, A. Cotter, and K. Livescu, "Stochastic optimization for PCA and PLS," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput*, 2012, pp. 861–868.

[5] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput.*, 2010, pp. 704–711.

[6] A. S. Bandeira, N. Boumal, and V. Voroninski, "On the low-rank approach for semidefinite programs arising in synchronization and community detection," in *Proc. Int. Conf. Learn. Theory*, 2016, vol. 49, pp. 1–22.

[7] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, "Dropping convexity for faster semi-definite optimization," in *Proc. Int. Conf. Learn. Theory*, 2016, pp. 530–582.

[8] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. New York, NY, USA: Springer, 2005.

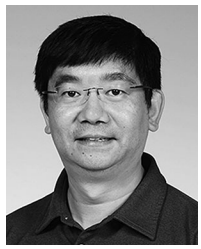
[9] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, 2018.

[10] S. Bubeck, "Theory of convex optimization for machine learning," 2014, arXiv:1405.4980.
[11] S. Burer and R. D. C. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Math. Program.*, vol. 95, no. 2, pp. 329–357, 2003.
[12] S. Burer and R. D. C. Monteiro, "Local minima and convergence in low-rank semidefinite programming," *Math. Program.*, vol. 103, no. 3, pp. 427–444, 2005.
[13] T. Cai and W.-X. Zhou, "A max-norm constrained minimization approach to 1-bit matrix completion," *J. Mach. Learn. Res.*, vol. 14, pp. 3619–3647, 2013.
[14] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imag. Sci.*, vol. 6, no. 1, pp. 199–225, 2013.
[15] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval from coded diffraction patterns," *Appl. Comput. Harmon. Anal.*, vol. 39, vol. 2, pp. 277–299, 2015.
[16] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.
[17] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," 2015, arXiv:1509.03025.
[18] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Inf. Inference: J. IMA*, vol. 3, pp. 189–223, 2014.
[19] C. De Sa, K. Olukotun, and C. Ré, "Global convergence of stochastic gradient descent for some non-convex matrix problems," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, vol. 37, pp. 2332–2341.
[20] D. P. W. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence, "The quest for ground truth in musical artist similarity," in *Proc. 3rd Int. Conf. Music Inf. Retrieval*, 2002.
[21] S. Ghadimi and G. Lan, "Stochastic first- and zeroth-order methods for nonconvex stochastic programming," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2341–2368, 2013.
[22] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, "WTF: The who to follow service at Twitter," in *Proc. Int. Conf. World Wide Web*, 2013, pp. 505–514.
[23] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1991.
[24] R. Horstmeyer, R. Y. Chen, X. Ou, B. Ames, J. A. Tropp, and C. Yang, "Solving ptychography with a convex relaxation," *New J. Phys.*, vol. 17, 2015, Art. no. 053044.
[25] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Symp. Theory Comput.*, 2013, pp. 665–674.
[26] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi, "Phase transitions in semidefinite relaxations," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 6, pp. 2218–2223, 2016.
[27] C. Jin, S. M. Kakade, and P. Netrapalli, "Provable efficient online matrix completion via nonconvex stochastic gradient descent," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4527–4535.
[28] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 315–323.
[29] A. Montanari and S. Sen, "Semidefinite programs on sparse random graphs and their application to community detection," in *Proc. 48th Annu. ACM Symp. Theory Comput.*, 2016, pp. 814–827.
[30] R. D. C. Monteiro, "First- and second-order methods for semidefinite programming," *Math. Program.*, vol. 97, pp. 209–244, 2003.
[31] S. N. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *Ann. Statist.*, vol. 39, no. 2, pp. 1069–1097, 2011.
[32] S. N. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, pp. 1665–1697, 2012.
[33] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Stat. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.
[34] Y. Nestrov and A. Nemirovski, *Self-Concordant Functions and Polynomial-Time Methods in Convex Programming*. Moscow, Russia: USSR Acad. Sci. Central Econ., Math. Inst., 1989.
[35] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2796–2804.
[36] S. J. Reddi, A. Hefny, S. Sra, B. Póczós, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, vol. 48, pp. 314–323.
[37] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.
[38] O. Shamir, "A stochastic PCA and SVD algorithm with an exponential convergence rate," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, vol. 37, pp. 144–152.
[39] C. Tan, S. Ma, Y.-H. Dai, and Y. Qian, "Barzilai-Borwein step size for stochastic gradient descent," in *Proc. 30th Annu. Conf. Neural Inf. Process. Syst.*, 2016, pp. 685–693.
[40] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 864–973.
[41] L. van der Maaten and K. Weinberger, "Stochastic triplet embedding," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2012, pp. 1–6.
[42] L. Wang, X. Zhang, and Q. Gu, "A unified computational and statistical framework for nonconvex low-rank matrix estimation," in *Proc. 20th Int. Conf. Artif. Intell. Stat.*, PMLR, 2017, vol. 54, pp. 981–990.
[43] L. Wang, X. Zhang, and Q. Gu, "A universal variance reduction-based catalyst for nonconvex low-rank matrix recovery," 2017, arXiv:1701.02301.
[44] J. Zeng, K. Ma, and Y. Yao, "Finding global optima in nonconvex stochastic semidefinite optimization with variance reduction," in *Proc. 21st Int. Conf. Artif. Intell. Stat.*, 2018.
[45] X. Zhang, L. Wang, and Q. Gu, "Stochastic variance-reduced gradient descent for low-rank matrix recovery from linear measurments," 2017, arXiv:1701.00481.
[46] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 109–117.

Jinshan Zeng received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2015. He is currently an Associate Professor with the School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China. His research interest includes nonconvex optimization, distributed optimization, and machine learning.

He received the Best Paper Award of the 7th International Congress of Chinese Mathematicians in 2018. He served as a Reviewer Editor for the journal *Frontiers in Applied Mathematics and Statistics*, as well as a Reviewer for *Mathematical Programming*, *SIAM Journal of Optimization*, IEEE TRANSACTIONS ON SIGNAL PROCESSING, etc.

Ke Ma received the Ph.D. degree from the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, in 2019. He is currently an Assistant Researcher with the Institute of Information Engineering, Chinese Academy of Sciences. His research interest includes statistical learning, data science, and stochastic optimization for large-scale data analysis.



Yuan Yao received the B.S.E. and M.S.E. degrees in control engineering from the Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, the M.Phil. degree in mathematics from the City University of Hong Kong, Hong Kong, in 2002, and the Ph.D. degree in mathematics from the University of California, Berkeley, Berkeley, CA, USA, in 2006. Since then, he has been with Stanford University, Stanford, CA, USA, and in 2009, he joined the Department of Probability and Statistics, School of Mathematical Sciences, Peking University, Beijing, China. He is currently an Associate Professor of Mathematics, Chemical, and Biological Engineering, and by courtesy, Computer Science and Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong. His current research interests include topological and geometric methods for high-dimensional data analysis and statistical machine learning, with applications in computational biology, computer vision, and information retrieval. He is a member of American Mathematical Society, Association for Computing Machinery, Institute of Mathematical Statistics, and Society for Industrial and Applied Mathematics. He was the area or session Chair in the Conference and Workshop on Neural Information Processing Systems and International Council for Industrial and Applied Mathematics, as well as a Reviewer for *Foundation of Computational Mathematics*, IEEE TRANSACTIONS ON INFORMATION THEORY, *Journal of Machine Learning Research*, and *Neural Computation*, etc.