Finding Global Optima in Nonconvex Stochastic Semidefinite Optimization with Variance Reduction

Jinshan Zeng[1,2] **Ke Ma**[3,4] **Yuan Yao**[2,†]
[1] School of Computer Information Engineering, Jiangxi Normal University
[2] Department of Mathematics, Hong Kong University of Science and Technology
[3] SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences
[4] School of Cyber Security, University of Chinese Academy of Sciences
{jsh.zeng@gmail.com, make@iie.ac.cn, yuany@ust.hk} († Corresponding author)

Abstract

There is a recent surge of interest in nonconvex reformulations via low-rank factorization for stochastic convex semidefinite optimization problem in the purpose of 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 global optima effectively despite their empirical success in applications. In this paper, we provide an answer that a stochastic gradient descent method with variance reduction, can be adapted to solve the nonconvex reformulation of the original convex problem, with a *global linear convergence*, i.e., converging to a global optimum exponentially fast, at a proper initial choice in the restricted strongly convex case. Experimental studies on both simulation and real-world applications on ordinal embedding are provided to show the effectiveness of the proposed algorithms.

1 Introduction

The stochastic convex semidefinite optimization problem, arising in many applications like non-metric multidimensional scaling [1, 6], matrix sensing [12, 22], community detection [14], synchronization [4], and phase retrieval [10], is of the following form:

$$\min_{X\in\mathbb{R}^{p\times p}} \; f(X)=\frac{1}{n}\sum_{i=1}^{n} f_i(X) \quad \text{s.t.} \quad X\succeq 0, \tag{1}$$

where $f_i(X)$ is some convex, smooth cost function associated with the i-th sample, $X\succeq 0$ is the positive semidefinite (PSD) constraint.

There are many algorithms for solving problem (1), mainly including the first-order methods like the well-known projected gradient descent method [17], 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 [15] and references therein). However, most of these methods are not well-scalable due to the PSD constraint, i.e., $X\succeq 0$. To circumvent this difficult constraint, the idea of low-rank factorization was adopted in literature [8, 9] and became very popular in the past few years due to its empirical success [5]. Low-rank factorization recasts the original problem (1) 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 (1) leads to,

$$\min_{U\in\mathbb{R}^{p\times r}} \; g(U) \quad \text{where} \quad r\le p. \tag{2}$$

Problems (1) and (2) will be equivalent when $r\ge r^*$ in the sense that problem (2) can find a global optimum X^* of problem (1) with $r^*=\mathrm{rank}(X^*)$. Since the PSD constraint has been eliminated, the recast problem (2) has a significant advantage over (1), but this benefit has a 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), its *global linear convergence*[1] remains unspecified until a recent work

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

[1] By *global linear convergence*, it means that the algo-

in [5]. Moreover, facing the challenge in large scale applications with a big n, stochastic algorithms [19] have been widely adopted nowadays, that is, at each iteration, we only use the gradient information of one or a small batch of the whole sample instead of the full gradient over n samples. However, due to the variance of such stochastic gradients, the stochastic gradient descent (SGD) method only has a sublinear convergence rate even in the strongly convex case. Various variance reduction techniques have been proposed in literature (see, e.g., [13, 20]), which resume the linear convergence for strongly convex problems. However, it is still open whether these methods can be adapted to the nonconvex problem (2) while enjoying the linear convergence to global optima.

The main contribution in this paper is to fill in this gap by showing that, when adapted to the nonconvex problem (2), our proposed versions of stochastic variance reduced gradient (SVRG) method can find the global optimum of the original problem (1) at a linear convergence rate when the initial choice lies in a prescribed neighbour of the global optimum and the objective function is restricted strongly convex. The initial choice condition here improves the one proposed for FGD in [5]. Moreover, our proposal includes both the fixed step sizes and the adaptive ones using a stabilized modification of Barzilai-Borwein (BB) step sizes [3] which adapts to the non-convex problems when the curvature is not guaranteed as in strongly convex cases. Finally, experiments on both matrix sensing and ordinal embedding demonstrate the effectiveness of the proposed scheme.

The reminder of this paper is organized as follows. Section 2 introduce some algorithmic background with our proposal. Section 3 presents the main convergence results. Section 4 provides some initialization schemes. Section 5 outlines the proof of our main theorem. Section 6 provides some applications to verify our theoretical findings and show the effectiveness of the proposed algorithms. We conclude this paper in Section 7.

Notations: For any two matrices $X, Y \in \mathbb{R}^{p\times p}$, their inner product is defined as $\langle X, Y\rangle = \mathrm{tr}(X^TY)$. We denote $\mathbb{S}^p_+$ as the set of positive semidefinite matrices of size $p \times p$. For any matrix $X \in \mathbb{R}^{p\times p}$, $\|X\|_F$ 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 in the context.

rithm converges to a global optimum exponentially fast when the initial choice is in a prescribed ball.

Algorithm 1 SVRG for Problem (1)

Parameters: update frequency m, step size (or learning rate) $\{\eta_k\}$, initial point $\tilde{U}^0 \in \mathbb{R}^{p\times r}$
for k=0,1,... **do**
  $\tilde{X}^k := \tilde{U}^k(\tilde{U}^k)^T$
  $g_k = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{X}^k)\tilde{U}^k$
  $U^0 = \tilde{U}^k$
  for $t = 0, \ldots, m-1$ **do**
    $X^t = U^t U^{tT}$
    Randomly pick $i_t \in \{1, \ldots, n\}$
    $U^{t+1} = U^t - \eta_k(\nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k + g_k)$
  end for
  $\tilde{U}^{k+1} = U^m$
end for

2 Algorithms

Without loss of generality, we assume that f is a symmetric function, i.e., $f(X) = f(X^T)$ throughout the paper. For $X = UU^T$, the gradient of $g(U) := f(UU^T)$ is

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

A. FGD: The FGD method proposed by [5] can be described as follows: let U^t be the iterate at the t-th iteration and $X^t := U^t U^{tT}$, then the next iterate U^{t+1} is updated according to the following

$$U^{t+1} = U^t - \eta\nabla f(X^t)U^t, \tag{3}$$

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

B. Stochastic FGD (SFGD): As a stochastic counterpart of FGD (3), SFGD here can be naturally described as follows: at the t-th iteration, randomly pick an $i_t \in \{1, \ldots, n\}$, then update the next iteration via

$$U^{t+1} = U^t - \eta_t\nabla f_{i_t}(X^t)U^t, \tag{4}$$

where $\eta_t > 0$ is a diminishing step size.

C. SVRG:[2] The SVRG method was firstly proposed by [13] for minimizing a finite sum of convex functions with a vector argument. The main idea of SVRG is adopting the variance reduction technique to accelerate SGD and achieve the linear convergence rate.

[2]Besides SVRG, there are some other accelerated stochastic methods like SAG, SDCA and their variants. We focus on SVRG mainly due to SVRG might require less storage than SAG and SDCA, and thus it may be more suitable for the applications considered in this paper.

Specifically, SVRG for solving problem (2) can be described as in Algorithm 1. There are mainly two loops including an inner loop and an outer loop in SVRG. One important implementation issue of SVRG is the tuning of the step size. There are mainly two classes of step sizes: determined or data adaptive. Here we discuss three particular choices.

(a) Fixed step size [13]:

$$\eta_k \equiv \eta, \quad \text{for some } \eta > 0. \tag{5}$$

(b) Barzilai-Borwein (BB) step size [3, 21]: given an initial $\eta_0 > 0$ and for $k \geq 1$, let $\tilde{g}_k := \nabla f(\tilde{X}^k)$,

$$\eta_k = \frac{1}{m} \cdot \frac{\|\tilde{X}^k - \tilde{X}^{k-1}\|_F^2}{|\langle \tilde{X}^k - \tilde{X}^{k-1}, \tilde{g}_k - \tilde{g}_{k-1}\rangle|}. \tag{6}$$

Note that such a BB step size is originally studied for strongly convex objective functions [21], and it may be breakout if there is no guarantee of the curvature of f like in nonconvex cases. In order to avoid such possible instability of (6) in our studies, a variant of BB step size, called the *stabilized BB* step size, is suggested as follows.

(c) Stabilized BB (SBB) step size: given an initial $\eta_0 > 0$ and an $\epsilon \geq 0$, for $k \geq 1$,

$$\eta_k = \frac{1}{m} \times \frac{\|\tilde{X}^k - \tilde{X}^{k-1}\|_F^2}{|\langle \tilde{X}^k - \tilde{X}^{k-1}, \tilde{g}_k - \tilde{g}_{k-1}\rangle| + \epsilon \|\tilde{X}^k - \tilde{X}^{k-1}\|_F^2}. \tag{7}$$

Throughout the rest of paper, with a slight abuse, we still name the original SVRG with a fixed step size as **SVRG**, and call the SVRG with stabilized BB step size (7) as **SVRG-SBB**$_\epsilon$, and particularly, we call SVRG with BB step size as **SVRG-SBB**$_0$. Besides the above three step sizes, there are some other schemes like the diminishing step size and the use of smoothing technique in BB step size as discussed in [21]. However, we mainly focus on the listed three step sizes in this paper due to they have been demonstrated to be effective in practice. Moreover, we only consider the Option-I suggested in [13] for Algorithm 1, since Option-I in SVRG is generally a more natural and better choice than Option-II as demonstrated in both [13] and [21] in the vector setting.

3 Global Linear Convergence of SVRGs

To present our main convergence results, we need the following assumptions.

Assumption 1 *Each f_i $(i = 1, \ldots, n)$ satisfies the following:*

(a) f_i is L-Lipschitz differentiable for some constant $L > 0$, i.e., f_i is smooth and ∇f_i is Lipschitz continuous satisfying

$$\|\nabla f_i(X) - \nabla f_i(Y)\|_F \leq L\|X - Y\|_F, \ \forall X, Y \in \mathbb{S}_+^p.$$

(b) f_i is (μ, r)-restricted strongly convex for some constants $\mu > 0$ and $r \leq p$, i.e., for any $X, Y \in \mathbb{S}_+^p$ with rank-r

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

Assumption 1 implies that f is also L-Lipschitz differentiable and (μ, r)-restricted strongly convex. For any L-Lipschitz differentiable and (μ, r)-restricted strongly convex function h, the following hold ([18])

$$\begin{aligned} &h(Y) \leq h(X) + \langle \nabla h(X), Y - X\rangle + \frac{L}{2}\|Y - X\|_F^2, \\ &\mu\|X - Y\|_F^2 \\ &\leq \langle \nabla h(X) - \nabla h(Y), X - Y\rangle \leq L\|X - Y\|_F^2, \end{aligned}$$

where the first inequality holds for any $X, Y \in \mathbb{S}_+^p$, and the second inequality holds for any $X, Y \in \mathbb{S}_+^p$ with rank r, the first inequality and the right-hand side of the second inequality hold for the Lipschitz continuity of ∇h, and the left-hand side of the second inequality is due to the (μ, r)-restricted strong convexity of h.

Let X^* be a global optimum of problem (1) with rank $r^* := \text{rank}(X^*)$, X_r^* be its rank-r ($r \leq r^*$) best approximation via truncated singular value decomposition (SVD), and U_r^* be a decomposition of X_r^* via $X_r^* = U_r^* U_r^{*T}$. Under Assumption 1, we define the following constants:

$$\kappa := \frac{L}{\mu}, \quad \gamma_0 := \frac{2(\sqrt{2} - 1)}{3\kappa}, \tag{8}$$

$$\bar{\eta} := \min\left\{ \frac{(1 - \sqrt{\gamma_0})^2}{\frac{\|\nabla f(X_r^*)\|_F}{L\sigma_r(X_r^*)} + (2\sqrt{\gamma_0} + \gamma_0)\tau(U_r^*)}, 1 \right\}, \tag{9}$$

$$\xi := \bar{\eta}(1 - \bar{\eta}/2), \tag{10}$$

where $\tau(X_r^*) := \frac{\sigma_1(X_r^*)}{\sigma_r(X_r^*)}$ and $\tau(U_r^*) := \frac{\sigma_1(U_r^*)}{\sigma_r(U_r^*)}$. $\kappa \geq 1$ is generally called the *condition number* of the objective function. Thus, $0 < \gamma_0 \leq \frac{2(\sqrt{2}-1)}{3}$ and $0 < \xi \leq 1/2$.

As r is used in the alternative nonconvex problem (2), the sequence $\{\tilde{X}^k\}$ generated by SVRG in Algorithm 1 is at least rank-r, and can only converge to a rank-r matrix if it is convergent. Therefore, we impose the following assumption to guarantee that the distance between X_r^* and X^* should be relatively small, otherwise, the introduced problem (2) is not a good alterative of the original problem (1).

Assumption 2 (rank-r approximation error) *Let X^* be a global optimum of problem (1), X_r^* be the rank-r approximation of X^* for a given positive integer $r \leq r^* := \mathrm{rank}(X^*)$. The following holds*

$$\|X_r^* - X^*\|_F < \frac{\sqrt{2}-1}{\sqrt{3}} \xi^{1/2} \kappa^{-1} \cdot \sigma_r(X^*),$$

where κ is specified in (8), and $\sigma_r(X^*)$ is the r-th largest singular value of X^*.

Assumption 2 is a regular assumption used in literature (say, [5]). Roughly speaking, Assumption 2 can be regarded as some noise assumption on problem (1). On the other hand, Assumption 2 is imposed to guarantee the uniqueness of the rank-r best approximation X_r^*. Otherwise, when $\|X_r^* - X^*\|_F \geq \sigma_r(X^*)$, if $X^* = \mathbf{I}_5$, i.e., an identity matrix with the size 5×5, then X_r^* with $r = 4$ has five possible candidates. Such assumption naturally holds for $r = r^*$, and when $r < r^*$, it might be satisfied if the singular values of X^* possess certain *compressible property*[3]. Under Assumption 2, we define several positive constants as follows: $\Delta := \frac{(\sqrt{2}-1)^2\xi^2\sigma_r^2(X_r^*)}{3\kappa^2} - \xi\|X_r^* - X^*\|_F^2$, $\tilde{\Delta} := \frac{4(\sqrt{2}-1)^2\xi^2\sigma_r^2(X_r^*)}{9\kappa^2} - \xi\|X_r^* - X^*\|_F^2$,

$$\gamma_l := \frac{2(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} - \sqrt{\Delta}, \tag{11}$$

$$\gamma_u := \frac{2(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} + \sqrt{\Delta}, \tag{12}$$

$$\tilde{\gamma}_l := \frac{2(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} - \sqrt{\tilde{\Delta}}, \tag{13}$$

$$\tilde{\gamma}_u := \frac{2(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} + \sqrt{\tilde{\Delta}}. \tag{14}$$

Note that the following relations hold

$$\gamma_l + \gamma_u = \frac{4(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa}, \tag{15}$$

$$\tilde{\gamma}_l < \gamma_l < \gamma_u < \tilde{\gamma}_u \leq \gamma_0\sigma_r(X_r^*), \tag{16}$$

where the last inequality of (16) holds for $0 < \xi \leq 1/2$ and $\tilde{\gamma}_u \leq 2\xi\gamma_0\sigma_r(X_r^*) \leq \gamma_0\sigma_r(X_r^*)$.

We also need the following common assumption on the stochastic direction, which has been widely used in literature on stochastic algorithms (say, [7] and reference therein).

Assumption 3 (Unbiasedness) *$\{\nabla f_{i_t}(X^t)U^t\}$ satisfies $\mathbb{E}_{i_t}[\nabla f_{i_t}(X^t)U^t] = \nabla f(X^t)U^t,\ \forall t \in \mathbb{N}$.*

If i_t is uniformly sampled, (see [16, 24] for studies on importance sampling), then the above assumption can be satisfied. Under Assumptions 1-3, let $\mathcal{N}_{\gamma_0} := \{U : \|U - U_r^*\|_F^2 \leq \gamma_0\sigma_r(X_r^*)\}$, and we define the following constants: $\mathcal{B} := \sup_{U\in\mathcal{N}_{\gamma_0}} \|UU^T\|_F$, $B_0 := \sup_{U\in\mathcal{N}_{\gamma_0}} \left\{\mathbb{E}_{i_t}[\|\nabla f_{i_t}(UU^T)\|_F^2] - \|\nabla f(UU^T)\|_F^2\right\}$, $B_1 := \sup_{U\in\mathcal{N}_{\gamma_0}} \|\nabla f(UU^T)\|_F^2$,

$$B_2 := 4\left[2L^2\mathcal{B}(\mathcal{B} + \|X_r^*\|_F) + B_0 + B_1\right], \tag{17}$$

$$\theta := \frac{2\xi B_2}{L(\sqrt{\tilde{\Delta}} - \sqrt{\Delta})} = \frac{18B_2\kappa\delta}{(\sqrt{2}-1)^2\xi\mu\sigma_r^2(X_r^*)}, \tag{18}$$

$$\eta_{\max} := \min\left\{\zeta_1, \zeta_2, \frac{1}{2\theta}\right\}, \tag{19}$$

where $\delta := \sqrt{\tilde{\Delta}} + \sqrt{\Delta}$, $\zeta_1 := \frac{1}{12\left[2L\cdot\kappa\mathcal{B} + \frac{B_0+B_1}{(\sqrt{2}-1)\mu\sigma_r(X_r^*)}\right]}$, and $\zeta_2 := \frac{(\sqrt{2}-1)\mu\xi\sigma_r(X_r^*)}{12B_2}$. It can be seen that $\mathcal{B}$ is the upper bound of $X = UU^T$, B_0 represents variance of the stochastic gradient of f, and B_1 is the upper bound of the squared Frobenius norm of gradient $\nabla f(UU^T)$, restricted to the closed ball $\mathcal{N}_{\gamma_0}$.

Let $\{\eta_k\}$ be a sequence satisfying $\eta_k \in (0, \eta_{\max})$ for any $k \in \mathbb{N}$. Given a positive integer m, define

$$\rho_k := 1 - \frac{\eta_k(\sqrt{2}-1)^2\xi\mu\sigma_r^2(X_r^*)}{18\kappa\delta}, \tag{20}$$

$$\tilde{\rho}_k := \rho_k^m + (1-\rho_k^m)\eta_k\theta. \tag{21}$$

It is easy to check that $0 < \rho_k < 1$ and $0 < \tilde{\rho}_k < 1$. Based on the above defined constants, we present our main theorem as follows.

Theorem 1 (Linear convergence of SVRG) *Let $\{\tilde{U}^k\}$ be a sequence generated by Algorithm 1. Suppose that Assumptions 1-3 hold, and that $\eta_k \in (0, \eta_{\max})$. The following hold: (a) if $\gamma_l < \|\tilde{U}^0 - U_r^*\|_F^2 < \gamma_u$, there hold*

(a1) $\{\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2]\}$ is monotonically decreasing,

*(a2) (**Linear convergence**) for any $k \geq 1$,*

$$\begin{aligned}&\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] \\ &\leq \left(\prod_{i=0}^{k-1}\tilde{\rho}_i\right)\cdot\|\tilde{U}^0 - U_r^*\|_F^2 + \tilde{\gamma}_l\times \\ &\left[\sum_{t=0}^{k-2}\left(\prod_{i=t+1}^{k-1}\tilde{\rho}_i\cdot(1-(\rho_t)^m)\right) + (1-(\rho_{k-1})^m)\right].\end{aligned} \tag{22}$$

(b) In addition, if $\|\tilde{U}^0 - U_r^*\|_F^2 \leq \gamma_l$, then $\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] \leq \gamma_l$ for any $k \in \mathbb{N}$.

The above theorem holds for a generic step size satisfying $\eta_k \in (0, \eta_{\max})$. Actually, if $\{\eta_k\}$ is lower bounded by a positive constant $\eta_{\min}$ and obviously, $\eta_{\min} < \eta_{\max}$,

[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 $C, q > 0$

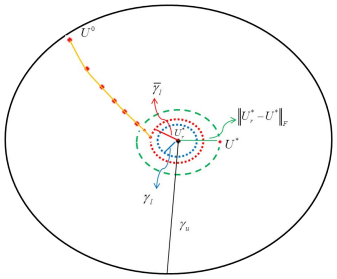

Figure 1: Convergence path of SVRG.

then by (20) and (21), $\rho_k \in (\rho_{\min}, \rho_{\max})$ and $\tilde{\rho}_k \in (\theta\eta_{\min}, \tilde{\rho}_{\max})$, where $\rho_{\min} := 1 - \frac{\eta_{\max}(\sqrt{2}-1)^2\xi\mu\sigma_r^2(X_r^*)}{18\kappa\delta}$, $\rho_{\max} := 1 - \frac{\eta_{\min}(\sqrt{2}-1)^2\xi\mu\sigma_r^2(X_r^*)}{18\kappa\delta}$, and $\tilde{\rho}_{\max} := \rho_{\max}^m + (1-\rho_{\max}^m)\theta\eta_{\max} < 1$. Thus, $\prod_{i=0}^{k-1}\tilde{\rho}_i \le (\tilde{\rho}_{\max})^k$, and

$$\begin{aligned}
&\sum_{t=0}^{k-2}\left(\prod_{i=t+1}^{k-1}\tilde{\rho}_i\cdot(1-(\rho_t)^m)\right)+(1-(\rho_{k-1})^m)\\
&\le (1-\rho_{\min}^m)\cdot\left[1+\sum_{t=0}^{k-2}(\tilde{\rho}_{\max})^{k-t-1}\right]\\
&= (1-\rho_{\min}^m)\cdot\frac{1-(\tilde{\rho}_{\max})^k}{1-\tilde{\rho}_{\max}} = \frac{1-\rho_{\min}^m}{1-\rho_{\max}^m}\cdot\frac{1-(\tilde{\rho}_{\max})^k}{1-\theta\eta_{\max}}.
\end{aligned}$$

Let $\bar{\gamma}_l := \frac{1-\rho_{\min}^m}{1-\rho_{\max}^m}\cdot\frac{\tilde{\gamma}_l}{1-\theta\eta_{\max}}$. According to the above two inequalities, (22) implies that

$$\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] - \bar{\gamma}_l \le (\tilde{\rho}_{\max})^k\,(\|\tilde{U}^0 - U_r^*\|_F^2 - \bar{\gamma}_l),$$

which shows the linear convergence of SVRG. Thus, Theorem 1 shows certain *global linear* convergence of SVRG, that is, the convergence to a global optimum starting from some good initial point, as depicted in Figure 1. From Figure 1, starting from an initialization lying in a γ_u-neighborhood of U_r^*, SVRG converges exponentially fast until achieving a small $\bar{\gamma}_l$-neighborhood of U_r^*; while if the initialization lies in the γ_l-ball of U_r^*, then SVRG will never escape from this small ball in expectation.

The comparisons on convergence results between FGD [5] and SVRG in the restricted strongly convex case are shown in Table 1. The convergence result of SVRG is presented in expectation. From Table 1, the requirement on the rank-r approximation error can be relaxed from the order $\mathcal{O}(\frac{\sigma_r(X_r^*)}{\kappa^{1.5}\tau(X_r^*)})$ to $\mathcal{O}(\frac{\sigma_r(X_r^*)}{\kappa})$, and the requirement on the radius of initialization can be relaxed from $\mathcal{O}(\frac{\sigma_r(X_r^*)}{\kappa^2\tau^2(X_r^*)})$ to $\mathcal{O}(\frac{\sigma_r(X_r^*)}{\kappa})$, where κ is the "condition number" of the objective function f (specified in (8)), $\sigma_r(X_r^*)$ and $\tau(X_r^*)$ are respectively the r-th largest singular value and the condition number of the rank-r approximation X_r^* of the optimum X^* with $r \le r^* := \mathrm{rank}(X^*)$.

Algorithm	$\|X^* - X_r^*\|_F$	Initialization
FGD ([5])	$\mathcal{O}\left(\frac{\sigma_r(X_r^*)}{\kappa^{1.5}\tau(X_r^*)}\right)$	$\mathcal{O}\left(\frac{\sigma_r(X_r^*)}{\kappa^2\tau^2(X_r^*)}\right)$
SVRG (our)	$\mathcal{O}\left(\frac{\sigma_r(X_r^*)}{\kappa\tau^{0.5}(X_r^*)}\right)$	$\mathcal{O}\left(\frac{\sigma_r(X_r^*)}{\kappa\tau(X_r^*)}\right)$

Table 1: Comparisons on convergence results (in order) between FGD [5] and SVRG (this paper) in the restricted strongly convex case.

In the following, we give a corollary to show the convergence of SVRG when adopting the considered three step-size strategies (5)-(7).

Corollary 1 (Convergence for step size choices) *Under conditions of Theorem 1, all claims in Theorem 1 hold, if one of the following conditions holds:*

(1) $\eta \in (0, \eta_{\max})$ *when a fixed step size is adopted;*

(2) $m > \frac{1}{(\mu+\epsilon)\eta_{\max}}$ *for any* $\epsilon \ge 0$ *when SBB step size is adopted.*

By the definition of SBB step size (7), and Assumption 1, we have

$$\frac{1}{m(L+\epsilon)} \le \eta_k \le \frac{1}{m(\mu+\epsilon)}.$$

Thus, if $m > \frac{1}{(\mu+\epsilon)\eta_{\max}}$, then $\eta_k < \eta_{\max}$ for any $k \in \mathbb{N}$.

From (11)-(13), if $r = r^*$ then $\|X_r^* - X^*\|_F = 0$, and thus $\tilde{\gamma}_l = 0$ and $\gamma_u = \frac{(2+\sqrt{3})\cdot(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa}$. However, in this case, $\gamma_l = \frac{(2-\sqrt{3})\cdot(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} > 0$. Thus, we cannot claim the exact recovery of a global optimum directly from Theorem 1 even if $\|\tilde{U}^0 - U_r^*\|_F^2 \le \gamma_l$. To circumvent this problem, we use a more consecutive step size, and get the following corollary showing the exact recovery of SVRG. Let

$$\bar{\eta}_{\max} := \min\left\{\frac{L\gamma_u}{2B_2\xi}, \eta_{\max}\right\}. \tag{23}$$

Corollary 2 (Exact recovery when $r = r^*$) *Let $\{\tilde{U}^k\}$ be a sequence generated by Algorithm 1. Let Assumptions 1 and 3 hold. If the following conditions hold: (a) $r = r^*$, (b) $\eta_k \in (0, \bar{\eta}_{\max})$, and (c) $\|\tilde{U}^0 - U_r^*\|_F^2 < \frac{(2+\sqrt{3})\cdot(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa}$, then SVRG exactly recover the global optimum X^* in expectation at a linear rate.*

Corollary 2 shows that if fortunately, we can take r as the exact rank r^* of the global optimum, then SVRG can exactly find the global optimum in expectation exponentially fast, as long as the initialization lies in a neighborhood of the global optimum. The proof of this corollary is presented in (Supplementary Material: Section 2.2).

4 On Initialization Schemes

According to our main theorem (see, Theorem 1), the initialization should be close to U_r^* to get the provable convergence. In the following, we discuss some potential initialization schemes.

Scheme I: One common way is to use one of the standard convex algorithms (say, projected gradient descent method) and obtain a good initialization U^0, then switch to SVRG to get a higher precision solution. A specific implementation of this idea has been used in [22] to deal with the matrix sensing problem, and some theoretical guarantees of this scheme have been developed in [5].

Scheme II: Another way is firstly to get $X^0 := \frac{1}{\|\nabla f(0)-\nabla f(\mathbf{e}_1\mathbf{e}_1^T)\|_F}\cdot \mathrm{Proj}_{\mathbb{S}_+^p}(-\nabla f(0))$, then take $U^0 \in \mathbb{R}^{p\times r}$ such that $U^0 U^{0^T} = X_r^0$, where X_r^0 is the rank-r best approximation of X^0 via SVD, and $\mathbf{e}_1 \in \mathbb{R}^p$ is the vector with 1 as the first component and 0 as the other components. The effectiveness of such scheme is guaranteed by [5, Corollary 12] when the objective function is *well-conditioned*, i.e., has a small κ.

Scheme III: Note that the previous two schemes need at least one SVD, which might be prohibitive in large scale applications. To avoid such an issue, random initialization can be exploited which actually works well in many applications.

5 Outline of Proofs

To prove Theorem 1, we need the following key lemma, which gives an error estimate of the inner loop.

Lemma 1 (A key lemma) *Let $\{U^t\}_{t=0}^m$ be the sequence at the k-th inner loop. Let Assumptions 1, 2 and 3 hold. Let $\eta_k \in (0, \eta_{\max})$. If $\gamma_l < \mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] < \gamma_u$, then the sequence $\{\mathbb{E}[\|U^t - U_r^*\|_F^2]\}$ is monotonically decreasing for $t = 0, \ldots, m$, and*

$$\begin{aligned}\mathbb{E}_{i_t}[\|U^{t+1} - U_r^*\|_F^2] &\le \frac{\eta_k L}{2}\|X^* - X_r^*\|_F^2 \\ &+ \|U^t - U_r^*\|_F^2 - \frac{2(\sqrt{2}-1)}{3}\eta_k\mu\sigma_r(X_r^*)\|U^t - U_r^*\|_F^2 \\ &+ \frac{\eta_k L}{2\xi}\|U^t - U_r^*\|_F^4 + \eta_k^2 B_2\cdot\|\tilde{U}^k - U_r^*\|_F^2\end{aligned} \tag{24}$$

where B_2 is specified in (17); while if $\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] \le \gamma_l$, then $\mathbb{E}[\|U^t - U_r^*\|_F^2] \le \gamma_l$ for any $t = 0, \ldots, m$.

The sketch proof of Lemma 1: We prove this lemma by induction. Specifically, we first show that if $\gamma_l < \mathbb{E}[\|U^t - U_r^*\|_F^2] < \gamma_u$, then $\mathbb{E}[\|U^{t+1} - U_r^*\|_F^2] \le \mathbb{E}[\|U^t - U_r^*\|_F^2] < \gamma_u$ for $t = 0, \ldots, m-1$. Furthermore, $\mathbb{E}[\|U^{t+1} - U_r^*\|_F^2]$ can be estimated via noting that

$$\begin{aligned}&\mathbb{E}_{i_t}[\|U^{t+1} - U_r^*\|_F^2] \\ &= \|U^t - U_r^*\|_F^2 + \eta_k^2\mathbb{E}_{i_t}[\|v_k^t\|_F^2] \\ &- 2\eta_k\mathbb{E}_{i_t}[\langle v_k^t, U^t - U_r^*\rangle],\end{aligned}$$

where $v_k^t = \nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k + \nabla f(\tilde{X}^k)\tilde{U}^k$, and then establish the bounds of both $\mathbb{E}_{i_t}[\|v_k^t\|_F^2]$ and $\mathbb{E}_{i_t}[\langle v_k^t, U^t - U_r^*\rangle]$ via two lemmas shown in (Supplementary Material: Lemma 2 and Lemma 3), respectively. The specific proof of this lemma is presented in (Supplementary Material: Section A).

Based on Lemma 1, we show the proof of Theorem 1 as follows.

Proof of Theorem 1: By Lemma 1, if $\gamma_l < \|\tilde{U}^0 - U_r^*\|_F^2 < \gamma_u$, then for any $k \in \mathbb{N}$ and $t = 0, \ldots, m-1$,

$$\mathbb{E}[\|\tilde{U}^{k+1} - U_r^*\|_F^2] \le \mathbb{E}[\|U^t - U_r^*\|_F^2] \le \mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2].$$

From (24) and by the definitions of $\tilde{\gamma}_l$ and $\tilde{\gamma}_u$, at the k-th inner loop, there holds

$$\begin{aligned}&\mathbb{E}[\|U^{t+1} - U_r^*\|_F^2] - \tilde{\gamma}_l \\ &\le \left[1 - \frac{\eta_k L}{2\xi}(\tilde{\gamma}_u - \mathbb{E}[\|U^t - U_r^*\|_F^2])\right]\times \\ &\big(\mathbb{E}[\|U^t - U_r^*\|_F^2] - \tilde{\gamma}_l\big) + \eta_k^2 B_2\cdot\|\tilde{U}^k - U_r^*\|_F^2 \\ &\le \left[1 - \frac{\eta_k L}{2\xi}\left(\sqrt{\tilde{\Delta}} - \sqrt{\Delta}\right)\right]\cdot(\mathbb{E}[\|U^t - U_r^*\|_F^2] - \tilde{\gamma}_l) \\ &+ \eta_k^2 B_2\cdot\|\tilde{U}^k - U_r^*\|_F^2 \\ &:= \rho_k(\mathbb{E}[\|U^t - U_r^*\|_F^2] - \tilde{\gamma}_l) + \eta_k^2 B_2\cdot\|\tilde{U}^k - U_r^*\|_F^2,\end{aligned}$$

where the second inequality holds for $\mathbb{E}[\|U^t - U_r^*\|_F^2] < \gamma_u$ and $\eta_k < \eta_{\max} \le \frac{(\sqrt{2}-1)\mu\xi\sigma_r(x_r^*)}{12B_2} \le \frac{6}{(\sqrt{2}-1)\mu\sigma_r(X_r^*)} \le \frac{2\xi}{L(\sqrt{\tilde{\Delta}}-\sqrt{\Delta})}$. By the above inequality, we have

$$\begin{aligned}&\mathbb{E}[\|\tilde{U}^{k+1} - U_r^*\|_F^2] - \tilde{\gamma}_l = \mathbb{E}[\|U^m - U_r^*\|_F^2] - \tilde{\gamma}_l \\ &\le (\rho_k)^m(\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] - \tilde{\gamma}_l) \\ &+ \eta_k^2 B_2\cdot\frac{1-(\rho_k)^m}{1-\rho_k}\cdot\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] \\ &\le (\rho_k)^m(\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] - \tilde{\gamma}_l) \\ &+ \eta_k\theta(1-(\rho_k)^m)\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2],\end{aligned}$$

where the final inequality is due to the definition of ρ_k (20), i.e., $\rho_k = 1 - \frac{\eta_k B_2}{\theta}$ and θ is specified in (18). Therefore,

$$\begin{aligned}&\mathbb{E}[\|\tilde{U}^{k+1} - U_r^*\|_F^2] \le (1-(\rho_k)^m)\tilde{\gamma}_l \\ &+ [(\rho_k)^m + \eta_k\theta(1-(\rho_k)^m)]\cdot\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] \\ &:= \tilde{\rho}_k\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] + (1-(\rho_k)^m)\tilde{\gamma}_l.\end{aligned}$$

Based on the above inequality, we get (22) via a recursive way, and thus complete the proof of this theorem. □

6 Experiments

In this section, we present two application examples to show the effectiveness of the proposed algorithm and also verify our developed theoretical results.

6.1 Matrix Sensing

We consider the following matrix sensing problem

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

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 $p=5000$, the optimal matrix $X^*:=U^*U^{*T}$ be a low-rank matrix with $\mathrm{rank}(X^*)=5$ and the sample size $n=10p$. In such high-dimensional regime, the generic semidefinite optimization methods generally do not work. Therefore, we only compare the performance of the low-rank factorization based methods, i.e., FGD [5], SFGD, and SVRG with three different step sizes studied in this paper. In this experiment, r is set as r^*, and the initialization is constructed via the optimum U^* with a random perturbation, and the step sizes for all algorithms are tuned in the hand-optimal way (shown in the figure). For three SVRG algorithms, the update frequency of the inner loop m is set as the sample size n. The experiment results are shown in Figure 2. An epoch of SFGD includes n iterations of SFGD, an epoch of FGD is exactly an iteration of FGD, and an epoch of SVRG is an iteration of outer loop. The iterative error curves of SVRG, SFGD and FGD are shown along epochs since all of them exploit a full scan of gradients over sample per epoch and their computational complexities per epoch are thus comparable.

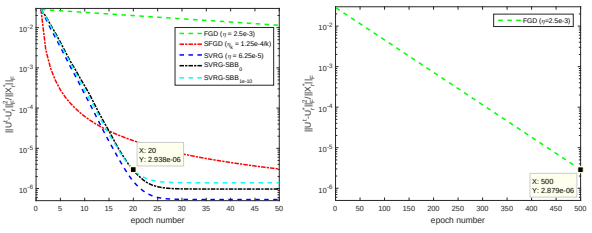

Figure 2: Experiments for matrix sensing problem. Left: trends of iterative errors of five algorithms. Right: trend of iterative error of FGD. It requires about 20, 50 and 500 epochs for SVRG, SFGD and FGD respectively, to achieve the precision 3×10^{-6}.

From Figure 2, we can observe that all three SVRG algorithms converge exponentially fast to the global optimum with high precisions. To achieve the precision 3×10^{-6}, it requires about 50 and 500 epochs for SFGD and FGD, respectively, while about 20 epochs are generally sufficient for three SVRG algorithms. In terms of epoch number, the considered SVRG methods are much faster than both FGD and SFGD. These experiment results demonstrate the effectiveness of SVRG and also verify our developed theoretical results.

6.2 Ordinal Embedding

In this subsection, we apply SVRG to the ordinal embedding problem, of which the Stochastic Triplet Embedding (STE) [23] is one of the typical models. The objective function is 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 the logistic loss. To show the effectiveness of the considered SVRG methods, we compare the performance of SVRG (using fixed, SBB_0 and SBB_ϵ step sizes, where $\epsilon=0.02$) with SFGD, FGD and the projected gradient descent (ProjGD) method, where the last two are batch methods.

A. Music artist dataset: We implement SVRG on the first real world dataset called *Music artist dataset*, collected by [11] 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, $i,j,k\in\{1,\dots,p\}$, and p is the number of total kinds of music artists. Specifically, we use the data pre-processed by [23] 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).

For each method, we implement independently 50 trials, and then record their test errors. For each trail, 80% triplets are randomly picked as the training set and the rest as the test set. All methods start with the same initial point, which is chosen randomly. Each curve in Figure 3 shows the trend of test error of one method with respect to the epoch number.

From Figure 3, SVRG with SBB step sizes can significantly speed up SFGD and the batch methods in terms of epoch number. Particularly, the test error curves of two SVRG-SBB methods decay much faster than those of SFGD, FGD and ProjGD at the initial 50 epochs.

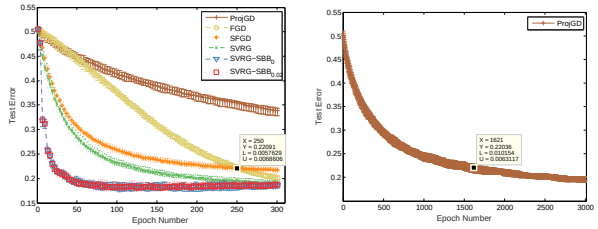

Figure 3: Experiments for Music artist data. To achieve the test error 0.22, about 40 epochs for SVRG-SBB$_0$ and SVRG-SBB$_{0.02}$, and 130 epochs for SVRG (fixed step size), and 260 epochs for both SFGD and FGD, and 1600 epochs for ProjGD are required.

B. eurodist dataset: We implement SVRG on another real world dataset called *eurodist dataset*, which describes the "driving" distances between 21 cities in Europe, and is available in the stats library of **R**. In this dataset, there are 21945 comparisons in total. A *quadruplet* (i, j, k, l) indicates an ordinal constraint like $d_{ij}^2(X) \leq d_{kl}^2(X)$, which means that "*the distance between cities i and j is shorter than the distance between cities l and k*", where $d_{ij}^2(X)$ is the "driving" distance between cities i and j, $i, j, k, l \in \{1, \ldots, 21\}$. One of the main task of this dataset is to embed these 21 cities in 2-dimensional space.

In this experiment, we first abstract all 3990 triplet ordinal comparisons from the total data set, and then use these triplets for learning. A *triplet* (i, j, k)[4] indicates an ordinal constraint like $d_{ik}^2(X) \leq d_{jk}^2(X)$, which means that "*the distance between cities i and k is less than the distance between cities j and k*". For each method, we implement independently 50 trials, and then record their test errors. For each trail, 80% triplets are randomly picked as the training set and the rest as the test set. All methods start with the same initial point, which is chosen randomly. Each curve in Figure 4 shows the trend of test error of one method with respect to the epoch number. The embedding results of all algorithms via comparing with the classical MDS (using the Matlab commond: mds.m) which is allowed to use the actual distance scores between all cities, are shown in (Supplementary Material: Appendix B).

From Figure 4, SVRG with SBB step sizes can speed up SFGD and both batch methods in terms of epoch number. Particularly, the test error curves of two SVRG-SBB methods decay much faster than those of SFGD, FGD and ProjGD at the initial 50 epochs.

[4]A triplet (i, j, k) is a special quadruplet (i, k, j, k). We only use all triplets in this experiment because the existing and our codes are only suitable for dealing with triplet ordinal constraints. We will further prepare the codes for dealing with the quadruplet ordinal constraints.

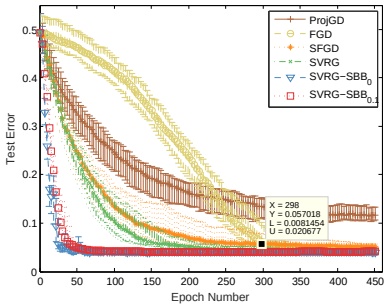

Figure 4: Experiments for *eurodist* dataset. To achieve the test error 0.057, about 40 epochs for SVRG-SBB$_0$ and SVRG-SBB$_{0.1}$, and 150 epochs for SVRG with a fixed step size, and 300 epochs for SFGD and FGD, and more than 450 epochs for ProjGD are required.

7 Conclusion

In this paper, we consider a nonconvex stochastic semidefinite optimization problem, which emerges in many fields of science and engineering. For the first time up to our knowledge, provable global linear convergence is established for stochastic variance reduced gradient (SVRG) algorithms to solve this nonconvex problem. Specifically, under common assumptions of restricted strong convexity of the objective function and small rank-r approximation error, we can show that SVRG can converge to a global optimum at a linear rate as long as the initialization lies in a neighborhood of the optimum. The initial choice condition significantly improves the existing results for deterministic gradient descent. Moreover, our choice of step sizes includes both fixed and adaptive ones using Barzilai-Borwein (BB) step size with stabilization in nonconvex settings. Application examples show that the proposed scheme is promising in fast solving some large scale problems.

Acknowledgment

The work of Jinshan Zeng is supported in part by the National Natural Science Foundation (NNSF) of China (No.61603162, 11501440), and the Doctoral start-up foundation of Jiangxi Normal University. Yuan Yao's work is supported in part by HKRGC grant 16303817, 973 Program of China (No. 2015CB85600, 2012CB825501), NNSF of China (No. 61370004, 11421110001), as well as grants from Tencent AI Lab, Si Family Foundation, Baidu BDI, and Microsoft Research-Asia.

References

[1] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, and S. Belongie (2007). Generalized non-metric multidimensional scaling, In *AISTATS*, pp. 11-18.

[2] F. Alizadeh (1995). Interior point methods for semidefinite programming with applications to combinatorial optimization, *SIAM Journal on Optimization*, 5(1): 13-51.

[3] J. Barzilai, and J. M. Borwein (1988). Two-point step size gradient methods, *IMA Journal of Numerical Analysis*, 8(1): 141-148.

[4] A.S. Bandeira, N. Boumal, and V. Voroninski (2016). On the low-rank approach for semidefinite programs arising in synchronization and community detection, In *COLT*, 49: 1-22.

[5] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi (2016). Dropping convexity for faster semidefinite optimization, In *COLT*, vol 49: 1-53.

[6] I. Borg, and P.J. Groenen (2005) *Modern multidimensional scaling: Theory and applications*, Springer.

[7] L. Bottou, F.E. Curtis, and J. Nocedal (2016). Optimization methods for large-scale machine learning, *arXiv:1606.04838*.

[8] S. Burer, and R. D. Monteiro (2003). A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization, *Mathematical Programming*, 95(2): 329-357.

[9] S. Burer, and R. D. Monteiro (2005), Local minima and convergence in low-rank semidefinite programming, *Mathematical Programming*, 103(3): 427-444.

[10] E.J. Candes, X. Li, and M. Soltanolkotabi (2015). Phase retrieval from coded diffraction patterns, *Applied and Computational Harmonic Analysis*, 39: 277-299.

[11] D. P. W. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence (2002). The quest for ground truth in musical artist similarity, In *Third International Conference on Music Information Retrieval*.

[12] P. Jain, P. Netrapalli, and S. Sanghavi (2013). Low-rank matrix completion using alternating minimization, In *Proceedings of 45th annual ACM symposium on Symposium on theory of computing*, pp. 665-674.

[13] R. Johnson, and T. Zhang (2013). Accelerating stochastic gradient descent using predictive variance reduction, In *Adcances in Neural Information Processing Systems*, pp. 315-323.

[14] A. Montanari, and S. Sen (2016). Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 814-827.

[15] R. D. Monteiro (2003). First- and second-order methods for semidefinite programming. *Mathematical Programming*, 97: 209-244.

[16] D. Needell, N. Srebro, and R. Ward (2014). Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm, In *Advances in Neural Information Processing Systems*, pp. 1017-1025.

[17] Y. Nestrov, and A. Nemirovski (1989). *Self-concordant functions and polynomial-time methods in convex programming*, USSR Academy of Sciences, Central Economic & Mathematic Institute.

[18] Y. Nestrov (2004). *Introductory lectures on convex optimization*, volumn 87. Springer Science & Business Media.

[19] H. Robbins, and S. Monro (1951). A stochastic approximation method, *The Annals of Mathematical Statistics*, 22(3):400-407.

[20] M. Schmidt, N. L. Roux, and F. Bach (2017). Minimizing finite sums with the stochastic average gradient, *Mathematical Programming*, Ser. A, 162(1-2): 83-112.

[21] C. Tan, S. Ma, Y.H. Dai, and Y. Qian (2016). Barzilai-Borwein step size for stochastic gradient descent, In *Advances in Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.

[22] S. Tu, R. Boczar, M. Simchowitz, M. Soltanokotabi, and B. Recht (2016). Low-rank solutions of linear matrix equations via procrustes flow, In *Proceedings of the 33rd International Conference on Machine Learning*, New York, NY, USA.

[23] L. van der Maaten, and K. Weinberger (2012). Stochastic triplet embedding, In *IEEE International workshop on machine learning for signal processing (MLSP)*, pp. 1-6.

[24] P. Zhao, and T. Zhang (2015). Stochastic optimization with importance sampling for regularized loss minimization, In *Proceedings of the International Conference on Machine Learning*.