On the Convergence of Stochastic Gradient Descent with Low-Rank Projections for Convex Low-Rank Matrix Problems

Dan Garber   DANGAR@TECHNION.AC.IL
Technion - Israel Institute of Technology, Haifa, Israel

Abstract

We revisit the use of Stochastic Gradient Descent (SGD) for solving convex optimization problems that serve as highly popular convex relaxations for many important low-rank matrix recovery problems such as *matrix completion*, *phase retrieval*, and more. The computational limitation of applying SGD to solving these relaxations in large-scale is the need to compute a potentially high-rank singular value decomposition (SVD) on each iteration in order to enforce the low-rank-promoting constraint. We begin by considering a simple and natural sufficient condition so that these relaxations indeed admit low-rank solutions. This condition is also necessary for a certain notion of low-rank-robustness to hold. Our main result shows that under this condition which involves the eigenvalues of the gradient vector at optimal points, SGD with mini-batches, when initialized with a "warm-start" point, produces iterates that are low-rank with high probability, and hence only a low-rank SVD computation is required on each iteration. This suggests that SGD may indeed be practically applicable to solving large-scale convex relaxations of low-rank matrix recovery problems. Our theoretical results are accompanied with supporting preliminary empirical evidence. As a side benefit, our analysis is quite simple and short.

Keywords: low-rank matrix optimization, stochastic gradient descent, convex optimization, matrix completion

1. Introduction

This paper is concerned with convex optimization formulations and algorithms for low-rank matrix recovery. Low-rank matrix recovery problems have numerous applications in machine learning, statistics and related field and have received much attention in recent years, with some of the most well known problems / applications being *matrix completion* Candès and Recht (2009); Recht (2011); Jaggi and Sulovský (2010); Ge et al. (2016), *phase retrieval* Candes et al. (2015); Netrapalli et al. (2013); Yurtsever et al. (2017), *robust PCA* Candès et al. (2011); Wright et al. (2009); Netrapalli et al. (2014); Yi et al. (2016); Mu et al. (2016), and more. However, these optimization problems are often NP-Hard to solve due to the explicit low-rank constraint / objective. To circumvent this difficulty, a significant body of work in recent years has been devoted to study convex relaxations to these problems, which are computationally tractable, and also often well motivated in terms of their ability to recover the correct low-rank solution (usually under certain statistical assumptions), see for instance Candès and Recht (2009); Recht (2011); Candes et al. (2015); Candès et al. (2011); Wright et al. (2009). These convex relaxations replace the explicit non-convex low-rank constraint / objective with a convex surrogate such as the sum of the singular values of the matrix, often called the nuclear norm, or the trace norm. Importantly, these convex relaxations can be formulated in the following canonical form (see for instance Jaggi and Sulovský (2010)), which

is also the main optimization problem under consideration in this paper:

$$\min_{\mathbf{X} \in \mathcal{S}_n} f(\mathbf{X}). \tag{1}$$

Here $\mathcal{S}_n$ denotes the spectrahedron in $\mathbb{S}^n$ (space of $n \times n$ real symmetric matrices), i.e., $\mathcal{S}_n := \{\mathbf{X} \in \mathbb{S}^n \mid \mathbf{X} \succeq 0, \ \mathrm{Tr}(\mathbf{X}) = 1\}$. Throughout this work, f is assumed β-smooth (Lipschitz gradient) and convex.

Additionally, motivated by cases in which $f(\mathbf{X})$ admits a finite-sum structure, i.e., $f(\mathbf{X}) := \frac{1}{m} \sum_{i=1}^{m} f_i(\mathbf{X})$, where the number of functions m is large and hence the computation of exact gradients of $f(\cdot)$ is prohibitive, or when $f(\mathbf{X})$ is given by an expectation w.r.t. some unknown distribution, i.e., $f(\mathbf{X}) := \mathbb{E}_{g \sim \mathcal{D}}[g(\mathbf{X})]$, and only a finite sample drawn i.i.d. form $\mathcal{D}$ is available (e.g., in statistically-motivated scenarios), we consider stochastic optimization methods for solving Problem (1). Concretely, we assume the standard generic model for first-order stochastic optimization, in which $f(\cdot)$ is given by a stochastic first-order oracle, which when queried with some point $\mathbf{X} \in \mathbb{S}^n$ returns a random matrix $\widehat{\nabla} \in \mathbb{S}^n$ satisfying the following standard assumptions:

$$i. \ \mathbb{E}[\widehat{\nabla} \mid \mathbf{X}] = \nabla f(\mathbf{X}), \quad ii. \ \|\widehat{\nabla}\|_F \le G, \ \|\widehat{\nabla}\| \le B, \quad iii. \ \mathbb{E}[\|\widehat{\nabla} - \nabla f(\mathbf{X})\|_F^2 \mid \mathbf{X}] \le \sigma^2,$$

for some $G, B, \sigma^2 > 0$, where for any matrix $\mathbf{M} \in \mathbb{S}^n$. $\|\mathbf{M}\|_F$ denotes the Frobenius (Euclidean) norm, and $\|\mathbf{M}\|$ denotes the spectral norm (largest singular value).

While Problem (1) is convex, it is still highly challenging to solve in large-scale via traditional first-order methods, such as projected gradient methods Nesterov (2013); Bubeck et al. (2015); Hazan and Kale (2014); Rakhlin et al. (2012) or conditional gradient-based methods Jaggi (2013); Hazan and Kale (2012); Lan and Zhou (2016); Hazan and Luo (2016); Garber and Kaplan (2019), since these require a potentially high-rank singular value decomposition (SVD) computation on each iteration (which can take as much as $O(n^3)$ runtime), and / or to store potentially high-rank matrices in memory (despite the often implicit assumption that the optimal solution is low-rank).

As a starting point let us recall the structure of the Euclidean projection onto the spectrahedron $\mathcal{S}_n$, which we denote as $\Pi_{\mathcal{S}_n}[\cdot]$.

Lemma 1 (Projection onto the spectrahedron) *Let $\mathbf{M} \in \mathbb{S}^n$ and write its eigen-decomposition as $\mathbf{M} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$. Then, it holds that $\Pi_{\mathcal{S}_n}[\mathbf{M}] = \sum_{i=1}^{n} \max\{0, \ \lambda_i - \lambda\} \mathbf{v}_i \mathbf{v}_i^\top$, where $\lambda \in \mathbb{R}$ is the unique scalar satisfying $\sum_{i=1}^{n} \max\{0, \ \lambda_i - \lambda\} = 1$.*

From the lemma it is quite obvious why at worst-case computing this projection may require a high-rank SVD (note that given the SVD of $\mathbf{M}$, computing the threshold parameter λ could be done in $O(n \log n)$ time via sorting). From this lemma we also make the following simple yet important observation.

Observation 1 (Low-rank projection requires low-rank SVD) *Given a matrix $\mathbf{M} \in \mathbb{S}^n$, if $rank\,(\Pi_{\mathcal{S}_n}[\mathbf{M}]) = r$, then only the top-$r$ components in the SVD of $\mathbf{M}$ (corresponding to the rank-r matrix $\sum_{i=1}^{r} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$) are required to compute the projection. Hence, only a rank-r SVD of $\mathbf{M}$ is required.* [1]

---

1. In particular, according to Lemma 1, if $\mathbf{M}$ admits the eigen-decomposition $\sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, then its projection onto $\mathcal{S}_n$ is rank-r if and only if $\sum_{i=1}^{r} \lambda_i \ge 1 + r \cdot \lambda_{r+1}$.

This observation implies that when the projected matrix is low-rank, the projection can be computed via fast iterative methods (such as power iterations or the faster Lanczos method) with runtime that is proportional to only $r \cdot \text{nnz}(\mathbf{M})$ (where $\text{nnz}(\cdot)$ denotes the number of non-zero entries), as opposed to n^3 required for a full-rank SVD.

Let us denote by $\mathcal{X}^*$ the set of optimal solutions to Problem (1). Our main result in this paper is that given some optimal solution $\mathbf{X}^* \in \mathcal{X}^*$ with $\text{rank}(\mathbf{X}^*) = r$, under a simple and natural condition on the eigenvalues of the gradient vector $\nabla f(\mathbf{X}^*)$, which we present next, the standard projected stochastic gradient method with mini-batches (see Algorithm 1), when initialized close enough to $\mathbf{X}^*$, will converge with constant probability to the optimal value of Problem (1) - f^*, while requiring on each iteration a single SVD computation of rank at most r to compute the projection.

Assumption 1 *We say an optimal solution $\mathbf{X}^* \in \mathcal{X}^*$ of rank r satisfies the eigen-gap assumption if $\lambda_{n-r}(\nabla f(\mathbf{X}^*)) - \lambda_n(\nabla f(\mathbf{X}^*)) > 0$.*

Importantly, the eigen-gap assumption, even without assuming explicitly that $\mathbf{X}^*$ is of rank r, is a sufficient condition for $\mathbf{X}^*$ to have rank at most r. This follows from the following lemma (see Lemma 7 in Garber (2019)). Thus, the additional requirement that $\mathbf{X}^*$ is of rank exactly r could be understood as a non-degeneracy requirement.

Lemma 2 *Let $\mathbf{X}^* \in \mathcal{X}^*$ be any optimal solution and write its eigendecomposition as $\mathbf{X}^* = \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$. Then, the gradient vector $\nabla f(\mathbf{X}^*)$ admits an eigendecomposition such that the set of vectors $\{\mathbf{v}_i\}_{i=1}^{r}$ is a set of top eigen-vectors of $(-\nabla f(\mathbf{X}^*))$ which corresponds to the eigenvalue $\lambda_1(-\nabla f(\mathbf{X}^*)) = -\lambda_n(\nabla f(\mathbf{X}^*))$.*

In order to better motivate Assumption 1 we bring the following lemma which suggests that this condition is required for the robustness of low-rank optimal solutions. The lemma shows that when the eigengap assumption does not hold, performing a standard projected gradient step from this optimal point w.r.t. to an arbitrarily small perturbation of the optimization problem, will result in a higher-rank matrix. Here we recall the first-order optimality condition $\Pi_{\mathcal{S}_n}[\mathbf{X}^* - \beta^{-1}\nabla f(\mathbf{X}^*)] = \mathbf{X}^*$.

The lemma is a simple adaptation of Lemma 3 in Garber (2019) (which considers optimization over trace-norm balls). A proof is given in the appendix for completeness.

Lemma 3 *Let $f : \mathbb{S}^n \to \mathbb{R}$ be β-smooth and convex. Let $\mathbf{X}^* \in \mathcal{S}_n$ be an optimal solution of rank r to the optimization problem $\min_{\mathbf{X} \in \mathcal{S}_n} f(\mathbf{X})$. Let $\mu_1, \ldots, \mu_n$ denote the eigenvalues of $\nabla f(\mathbf{X}^*)$ in non-increasing order. Then, $\mu_{n-r} = \mu_n$ if and only if for any arbitrarily small $\zeta > 0$ it holds that*

$$rank(\Pi_{(1+\zeta)\mathcal{S}_n}[\mathbf{X}^* - \beta^{-1}\nabla f(\mathbf{X}^*)]) > r,$$

where $(1 + \zeta)\mathcal{S}_n = \{(1 + \zeta)\mathbf{X} \mid \mathbf{X} \in \mathcal{S}_n\}$, and $\Pi_{(1+\zeta)\mathcal{S}_n}[\cdot]$ denotes the Euclidean projection onto the convex set $(1 + \zeta)\mathcal{S}_n$.

We also refer the reader to Garber (2019) (Table 2) for an empirical evidence that Assumption 1 seems to be quite practical for real-world datasets.

Formally, the main result of this paper is the proof of the following theorem.

Algorithm 1 Projected Stochastic Gradient Descent with minibathces

---

1: input: initialization point $\mathbf{X}_1 \in \mathcal{S}_n$, batch-size L, time horizon T, sequence of step-sizes $\{\eta_t\}_{t \in [T-1]}$

2: **for** $t = 1 \ldots T - 1$ **do**

3: $\quad \widehat{\nabla}_t \leftarrow \frac{1}{L} \sum_{i=1}^{L} \widehat{\nabla}_t^{(i)}$, where $\{\widehat{\nabla}_t^{(i)}\}_{i=1}^{L}$ is produced by L calls to the stochastic oracle of $f(\cdot)$ with the input point $\mathbf{X}_t$.

4: $\quad \mathbf{X}_{t+1} \leftarrow \Pi_{\mathcal{S}_n}[\mathbf{X}_t - \eta_t \widehat{\nabla}_t]$

5: **end for**

6: return solution $\bar{\mathbf{X}}$ according to one of the following options:

$$\bar{\mathbf{X}} \leftarrow \mathbf{X}_{t_0}, \ t_0 \sim Uni\{1, \ldots, T\} \text{ (\textbf{option I})} \quad \text{or} \quad \bar{\mathbf{X}} \leftarrow \frac{1}{T} \sum_{i=1}^{T} \mathbf{X}_i \text{ (\textbf{option II})}$$

Theorem 4 *Let* $\mathbf{X}^* \in \mathcal{X}^*$ *be an optimal solution of rank* r *which satisfies Assumption 1. Consider running SGD (Algorithm 1) for* T *iterations with a fixed step-size* $\eta = \frac{R_0}{10G\sqrt{T}\log(8T)}$ *and when the first iterate* $\mathbf{X}_1$ *satisfies* $rank(\mathbf{X}_1) \leq r$ *and* $\|\mathbf{X}_1 - \mathbf{X}^*\|_F \leq R_0/2$, *where*

$$R_0 := \frac{1}{8}\Big(r\beta + \frac{\sqrt{2r}B}{\lambda_r(\mathbf{X}^*)}\Big)^{-1}\delta, \qquad 0 < \delta \leq \lambda_{n-r}(\nabla f(\mathbf{X}^*)) - \lambda_n(\nabla f(\mathbf{X}^*)),$$

and with constant minibatch size L *satisfying* $L \geq L_0 = O\Big(\max\{(\big(\frac{\sigma}{GR_0}\big)^2, \frac{B^2 r^2}{\delta^2}\log(nT)\}\Big)$. *Then, for any* T *sufficiently large, it holds with probability at least* $1/2$ *that*

1. $f(\bar{\mathbf{X}}) - f^* = O\left(\frac{GR_0 \log T}{\sqrt{T}}\right)$,

2. $\forall t \in [T-1]$: $rank(\mathbf{X}_{t+1}) \leq r$. *Moreover, if* ***option I*** *is used for the returned solution, then* $rank(\bar{\mathbf{X}}) \leq r$.

Thus, Theorem 4, together with Observation 1, imply that with constant probability, all the steps of SGD can be computed via a rank-r SVD.

Corollary 5 (sample complexity) *The overall sample complexity to achieve* $f(\bar{\mathbf{X}}) - f^* \leq \epsilon$ *with probability at least* $1/2$, *when initializing from a "warm-start", is upper-bounded by* $\tilde{O}\left(\epsilon^{-2}\max\{\sigma^2, \lambda_r^2(\mathbf{X}^*)rG^2\}\right)$[2] *(note that* $\lambda_r(\mathbf{X}^*) \leq 1/r$*).*

The proof is given in the appendix. We note this sample complexity is nearly optimal (up to a logarithmic factor) in ϵ and optimal in σ, G (see for instance Bubeck et al. (2015)). Most importantly, it is independent of the eigen-gap δ.[3]

Remark 6 *It is quite important to note that while verifying the validity of Assumption 1, or the "warm start" condition, or even setting the step-size in Theorem 4 correctly, can be quite difficult*

---

2. Throughout this paper we use the notation $\tilde{O}(\cdot)$ or $\tilde{\Theta}(\cdot)$ to suppress poly-logarithmic factors.

3. Naturally, the sample complexity to obtain the required "warm-start" initialization will depend on δ, but will be independent of the overall target accuracy ϵ.

in practice, from a practical point of view, it is mainly important that the low-rank-SVD-based projection is indeed the correct Euclidean projection. This however, could be easily verified in each step t of the algorithm: if instead of computing a rank-r SVD of the point to project $\mathbf{X}_t - \eta_t \widehat{\nabla}_t$, we compute a rank-$(r+1)$ SVD, we can easily verify (using the condition on the thresholding parameter λ in Lemma 1, see Footnote 1), if the correct projection is indeed of rank at most r, and hence verify that the algorithm indeed converges correctly.

1.1. Related work

Our work is primarily motivated by the very recent work Garber (2019), which considered Problem (1) in a purely deterministic setting, i.e., when exact gradients of $f(\cdot)$ are available. In that work it is shown that, under Assumption 1, standard projected gradient methods, when initialized with a "warm-start" point, converge with their original convergence guarantees to an optimal solution using only low-rank SVD to compute the projection. However, these results are not directly extendable to the stochastic setting for two reasons. First, the "warm-start" requirement in Garber (2019) requires that the distance to an optimal solution is proportional to the step-size used. While this makes sense in the deterministic setting, since the typical step-size for projected-gradient methods is just $1/\beta$, for SGD, the step-size (e.g., when chosen to be fixed) is proportional to the target accuracy ϵ, which imposes an unrealistic initialization requirement (in particular, given that the function is Lipschitz, such a condition already implies that the initial point satisfies $f(\mathbf{X}_1) - f^* = O(\epsilon)$). Therefore, our main technical contribution is to provide an alternative analysis to the one used in Garber (2019), in which the required initial distance to an optimal solution is independent of the step-size.

Second, since the analysis of Garber (2019) (as the one in this work) only applies in a certain ball around an optimal solution, it relies on the property that the projected gradient method does not increase the distance to the optimal set from one iteration to the next. This property does not hold anymore for SGD, and here we introduce a martingale argument to show that with high probability all the iterates indeed stay within the relevant ball.

For specific low-rank matrix recovery problems, the works De Sa et al. (2015); Jin et al. (2016); Ge et al. (2016); Bhojanapalli et al. (2016b) yield global convergence guarantees for non-convex SGD which forces the low-rank constraint by explicitly factorizing the matrix variable as the product of two rank-r matrices. However, these only hold under very specific and quite strong statistical assumptions on the data. On the contrary, in this work we do not impose any statistical generative model on the data.

Finally, we note that works that analyze non-convex methods without relying on strong statistical models, such as Bhojanapalli et al. (2016a) (though they only consider the *deterministic* gradient descent method), also require "warm-start" initialization which is qualitatively similar to ours, e.g., relies on the ratio between smallest and largest singular values of the optimal solution (see Theorem 4 above).

2. Analysis

The proof of Theorem 4 follows from combining the standard convergence analysis of SGD with two main lemmas. Lemma 7, which is the main technical novelty we introduce in this paper, and believe may be of independent interest, establishes (informally) that at any step t of Algorithm 1, if $\mathbf{X}_t$ is sufficiently close to an optimal solution $\mathbf{X}^*$ which satisfies the gap assumption (Assumption 1), and the stochastic gradient is not too noisy, then $\mathbf{X}_{t+1}$ is low-rank (and hence can be computed,

given $\mathbf{X}_t, \widehat{\nabla}_t$, using only a low-rank SVD). Lemma 8 then uses a martingale concentration argument to establish that, if $\mathbf{X}_1$ is sufficiently close to some optimal solution $\mathbf{X}^*$, then with high probability, all following iterates are also sufficiently close. Combining these two lemmas ensures that with high probability, the projection onto $\mathcal{S}_n$ at each step of Algorithm 1 can be computed using only a low-rank SVD computation.

Throughout this work we let the operation $\mathbf{A} \bullet \mathbf{B}$ denote the standard inner product for any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, i.e., $\mathbf{A} \bullet \mathbf{B} = \mathrm{Tr}(\mathbf{A}\mathbf{B}^\top)$.

Lemma 7 *Let $\mathbf{X}^* \in \mathcal{X}^*$ be of rank r, and let $\mu_1 \ldots \mu_n$ denote the eigenvalues of $\nabla f(\mathbf{X}^*)$ in non-increasing order. Let $\mathbf{X} \in \mathcal{S}_n$ be a matrix such that $rank(\mathbf{X}) \leq r$, and suppose that*

$$\|\mathbf{X} - \mathbf{X}^*\|_F \leq \frac{1}{4}\left(r\beta + \frac{\sqrt{2r}B}{\lambda_r(\mathbf{X}^*)}\right)^{-1}(\delta - 4r\xi), \tag{2}$$

where $\delta := \mu_{n-r} - \mu_n$. Finally, let $\widetilde{\nabla}$ be a matrix such that $\|\widetilde{\nabla} - \nabla f(\mathbf{X})\| \leq \xi$, $\|\widetilde{\nabla}\| \leq B$. Then, for any step-size $\eta > 0$ it holds that $rank\left(\Pi_{\mathcal{S}_n}[\mathbf{X} - \eta\widetilde{\nabla}]\right) \leq r$.

Proof Let us denote $\mathbf{Y} = \mathbf{X} - \eta\widetilde{\nabla}$. From Lemma 1 it follows that a sufficient condition so that $\mathrm{rank}\left(\Pi_{\mathcal{S}_n}[\mathbf{X} - \eta\widetilde{\nabla}]\right) \leq r$, is $\sum_{i=1}^r \lambda_i(\mathbf{Y}) \geq 1 + r \cdot \lambda_{r+1}(\mathbf{Y})$ (since then the thresholding parameter λ in Lemma 1 must satisfy $\lambda \geq \lambda_{r+1}(\mathbf{Y})$).

Let $\mathbf{X} = \mathbf{V}\Lambda\mathbf{V}^\top$ denote the eigen-decomposition of $\mathbf{X}$. In case $\mathrm{rank}(\mathbf{X}) < r$, we extend this decomposition to have rank=r by adding additional zero eigenvalues and corresponding eigenvectors, so $\mathbf{V} \in \mathbb{R}^{d \times r}$. It holds that

$$\sum_{i=1}^r \lambda_i(\mathbf{Y}) \geq \mathbf{V}\mathbf{V}^\top \bullet \mathbf{Y} = \mathbf{V}\mathbf{V}^\top \bullet (\mathbf{X} - \eta\widetilde{\nabla}) = \mathrm{Tr}(\mathbf{X}) - \eta\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla} = 1 - \eta\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla}.$$

Using the above inequality we also have

$$\begin{aligned}
\lambda_{r+1}(\mathbf{Y}) &= \sum_{i=1}^{r+1}\lambda_i(\mathbf{Y}) - \sum_{j=1}^r \lambda_j(\mathbf{Y}) \leq \sum_{i=1}^{r+1}\lambda_i(\mathbf{Y}) - \left(1 - \eta\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla}\right) \\
&\underset{(a)}{\leq} \sum_{i=1}^{r+1}\left(\lambda_i(\mathbf{X}) + \lambda_i(-\eta\widetilde{\nabla})\right) - \left(1 - \eta\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla}\right) \\
&\underset{(b)}{=} 1 + \sum_{i=1}^{r+1}\lambda_i(-\eta\widetilde{\nabla}) - \left(1 - \eta\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla}\right) = \eta\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla} - \eta\sum_{i=1}^{r+1}\lambda_{n-i+1}(\widetilde{\nabla}),
\end{aligned}$$

where (a) follows from Ky Fan's eigenvalue inequality, and (b) follows since $\sum_{i=1}^{r+1}\lambda_i(\mathbf{X}) = \sum_{i=1}^{\mathrm{rank}(\mathbf{X})}\lambda_i(\mathbf{X}) = \mathrm{Tr}(\mathbf{X}) = 1$. Thus, we arrive at the following sufficient condition so that $\mathrm{rank}\left(\Pi_{\mathcal{S}_n}[\mathbf{X} - \eta\widetilde{\nabla}]\right) \leq r$:

$$1 - \eta\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla} \geq 1 + r\left(\eta\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla} - \eta\sum_{i=1}^{r+1}\lambda_{n-i+1}(\widetilde{\nabla})\right),$$

which boils down to the sufficient condition

$$-\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla} \geq -\frac{r}{r+1}\sum_{i=1}^{r+1}\lambda_{n-i+1}(\widetilde{\nabla}). \tag{3}$$

Let $\mathbf{V}^*\Lambda^*\mathbf{V}^{*\top}$ denote the eigen-decomposition of $\mathbf{X}^*$ and recall $\mathrm{rank}(\mathbf{X}^*) = r$. Then,

$$\begin{aligned}
-\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla} &\geq -\mathbf{V}^*\mathbf{V}^{*\top} \bullet \widetilde{\nabla} - \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^*\mathbf{V}^{*\top}\|_*\|\widetilde{\nabla}\| \\
&\geq -\mathbf{V}^*\mathbf{V}^{*\top} \bullet \nabla f(\mathbf{X}^*) - \|\mathbf{V}^*\mathbf{V}^{*\top}\|_* \cdot \|\widetilde{\nabla} - \nabla f(\mathbf{X}^*)\| \\
&\quad - \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^*\mathbf{V}^{*\top}\|_* \cdot \|\widetilde{\nabla}\| \\
&\geq -\mathbf{V}^*\mathbf{V}^{*\top} \bullet \nabla f(\mathbf{X}^*) - r\|\widetilde{\nabla} - \nabla f(\mathbf{X}^*)\| - \sqrt{2r}\|\widetilde{\nabla}\| \cdot \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^*\mathbf{V}^{*\top}\|_F.
\end{aligned}$$

Since $\mathbf{X}^*$ is an optimal solution, it follows from Lemma 2 that

$$-\mathbf{V}^*\mathbf{V}^{*\top} \bullet \nabla f(\mathbf{X}^*) = -r\mu_n.$$

Also, since $f(\cdot)$ is β-smooth,

$$\|\widetilde{\nabla} - \nabla f(\mathbf{X}^*)\| \leq \|\widetilde{\nabla} - \nabla f(\mathbf{X})\| + \|\nabla f(\mathbf{X}) - \nabla f(\mathbf{X}^*)\|_F \leq \xi + \beta\|\mathbf{X} - \mathbf{X}^*\|_F. \tag{4}$$

Finally, using the Davis-Kahan $\sin\theta$ theorem (see for instance Theorem 2 in Yu et al. (2014)), we have that

$$\|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^*\mathbf{V}^{*\top}\|_F \leq \frac{2\|\mathbf{X} - \mathbf{X}^*\|_F}{\lambda_r(\mathbf{X}^*) - \lambda_{r+1}(\mathbf{X}^*)} = \frac{2\|\mathbf{X} - \mathbf{X}^*\|_F}{\lambda_r(\mathbf{X}^*)}.$$

Thus, combining these three bounds, we have that

$$-\mathbf{V}\mathbf{V}^\top \bullet \widetilde{\nabla} \geq -r\mu_n - r\xi - \sqrt{r}\|\mathbf{X} - \mathbf{X}^*\|_F\left(\sqrt{r}\beta + \frac{2\sqrt{2}\|\widetilde{\nabla}\|}{\lambda_r(\mathbf{X}^*)}\right). \tag{5}$$

On the other-hand,

$$\begin{aligned}
-\sum_{i=1}^{r+1}\lambda_{n-i+1}(\widetilde{\nabla}) = \sum_{i=1}^{r+1}\lambda_i(-\widetilde{\nabla}) &= \sum_{i=1}^{r+1}\lambda_i(-\nabla f(\mathbf{X}^*) + (\nabla f(\mathbf{X}^*) - \widetilde{\nabla})) \tag{6}\\
&\underset{(a)}{\leq} \sum_{i=1}^{r+1}\lambda_i(-\nabla f(\mathbf{X}^*)) + \sum_{i=1}^{r+1}\lambda_i(\nabla f(\mathbf{X}^*) - \widetilde{\nabla}) \\
&\leq -\sum_{i=1}^{r+1}\lambda_{n-i+1}(\nabla f(\mathbf{X}^*)) + (r+1)\|\nabla f(\mathbf{X}^*) - \widetilde{\nabla}\| \\
&\underset{(b)}{=} -((r+1)\mu_n + \delta) + (r+1)\|\nabla f(\mathbf{X}^*) - \widetilde{\nabla}\| \\
&\underset{(c)}{\leq} -((r+1)\mu_n + \delta) + (r+1)(\xi + \beta\|\mathbf{X} - \mathbf{X}^*\|_F), \tag{7}
\end{aligned}$$

where (a) follows from Ky Fan's eigenvalue inequality, and (b) follows from Lemma 2 and our assumption on the eigenvalues of $\nabla f(\mathbf{X}^*)$, and (c) follows from (4).

Plugging (5) and (6) into (3), we arrive at the following sufficient condition so that $\text{rank}\left(\Pi_{\mathcal{S}_n}[\mathbf{X} - \eta\widetilde{\nabla}]\right) \leq r$,

$$-r\mu_n - r\xi - \sqrt{r}\|\mathbf{X} - \mathbf{X}^*\|\left(\sqrt{r}\beta + \frac{2\sqrt{2}\|\widetilde{\nabla}\|}{\lambda_r(\mathbf{X}^*)}\right) \geq -r\left(\mu_n + \frac{\delta}{r+1}\right) + r\left(\xi + \beta\|\mathbf{X} - \mathbf{X}^*\|_F\right),$$

which is equivalent to the condition $\|\mathbf{X} - \mathbf{X}^*\|_F \leq \left(2r\beta + \frac{2\sqrt{2r}\|\widetilde{\nabla}\|}{\lambda_r(\mathbf{X}^*)}\right)^{-1}\left(\frac{r}{r+1}\delta - 2r\xi\right)$. Simplifying the above expression gives the result. ∎

Lemma 8 *Fix $p \in (0,1)$. Let $\mathbf{X}_1, \ldots, \mathbf{X}_T$ be a sequence generated by Algorithm 1 such that for all $t \in [T-1]$, $\eta_t = \eta$ for some $\eta > 0$ satisfying $1/\eta = \tilde{\Theta}(\sqrt{T})$, and with mini-batch size $L = O(poly(\log T))$. Then, for any $\mathbf{X}^* \in \mathcal{X}^*$ and any T large enough, it holds with probability at least $1 - p$ that for all $t \in [T]$:*

$$\|\mathbf{X}_t - \mathbf{X}^*\|_F^2 \leq \|\mathbf{X}_1 - \mathbf{X}^*\|_F^2 + G^2 T\eta^2 + \sqrt{40T\eta^2\sigma^2/L}\sqrt{\log(T/p)}.$$

Proof Define the auxiliary sequence $\{\mathbf{Y}_t\}_{t=1}^T$ as follows: $\mathbf{Y}_1 = \mathbf{X}_1$ and for all $t \in [T-1]$, $\mathbf{Y}_{t+1} := \mathbf{X}_t - \eta_t\widehat{\nabla}_t$. Recall that with these definitions we have that for all $t \in [T]$, $\mathbf{X}_t = \Pi_{\mathcal{S}_n}[\mathbf{Y}_t]$.

Throughout the proof let us fix some optimal solution $\mathbf{X}^* \in \mathcal{X}^*$. We begin with the observation that for all $t \in [T-1]$ it holds that

$$\begin{aligned}
\mathbb{E}[\|\mathbf{Y}_{t+1} - \mathbf{X}^*\|_F^2 \mid \mathbf{X}_t] &= \mathbb{E}[\|\mathbf{X}_t - \eta_t\widehat{\nabla}_t - \mathbf{X}^*\|_F^2 \mid \mathbf{X}_t] \\
&\leq \|\mathbf{X}_t - \mathbf{X}^*\|_F^2 - 2\eta_t\mathbb{E}[(\mathbf{X}_t - \mathbf{X}^*) \bullet \widehat{\nabla}_t \mid \mathbf{X}_t] + \eta_t^2 G^2 \\
&= \|\mathbf{X}_t - \mathbf{X}^*\|_F^2 - 2\eta_t(\mathbf{X}_t - \mathbf{X}^*) \bullet \nabla f(\mathbf{X}_t) + \eta_t^2 G^2 \\
&\underset{(a)}{\leq} \|\mathbf{X}_t - \mathbf{X}^*\|_F^2 - 2\eta_t(f(\mathbf{X}_t) - f(\mathbf{X}^*)) + \eta_t^2 G^2 \underset{(b)}{\leq} \|\mathbf{Y}_t - \mathbf{X}^*\|_F^2 + \eta_t^2 G^2,
\end{aligned}$$

where (a) follows from the convexity of $f(\cdot)$, and (b) follows since $\mathbf{X}_t$ is the projection of $\mathbf{Y}_t$ onto $\mathcal{S}_n$.

For all $t \in [T]$ define the random variable $Z_t := \|\mathbf{Y}_t - \mathbf{X}^*\|_F^2 - G^2\sum_{i=1}^{t-1}\eta_i^2$. Note that $Z_1, \ldots, Z_T$ forms a submartingale sequence w.r.t. the filtration $\mathbf{F} := \{\mathcal{F}_t := \{\mathbf{X}_1, \ldots, \mathbf{X}_t\}\}_{t=1}^{T-1}$. This holds since using the above inequality, we have that for all $t \in [T-1]$:

$$\mathbb{E}[Z_{t+1}|\mathcal{F}_t] = \mathbb{E}[\|\mathbf{Y}_{t+1} - \mathbf{X}^*\|_F^2|\mathbf{X}_t] - G^2\sum_{i=1}^{t}\eta_i^2 \leq \|\mathbf{Y}_t - \mathbf{X}^*\|_F^2 + G^2\eta_t^2 - G^2\sum_{i=1}^{t}\eta_i^2 = Z_t.$$

We continue to show that this submartingale has bounded-differences and to upper-bound its variance. It holds for all $2 \leq t \leq T$ that

$$\begin{aligned}
Z_t - \mathbb{E}[Z_t|\mathcal{F}_{t-1}] &= \|\mathbf{Y}_t - \mathbf{X}^*\|_F^2 - \mathbb{E}[\|\mathbf{Y}_t - \mathbf{X}^*\|_F^2|\mathbf{X}_{t-1}] \\
&= \|\mathbf{X}_{t-1} - \mathbf{X}^* - \eta_{t-1}\widehat{\nabla}_{t-1}\|_F^2 - \mathbb{E}[\|\mathbf{X}_{t-1} - \mathbf{X}^* - \eta_{t-1}\widehat{\nabla}_{t-1}\|_F^2|\mathbf{X}_{t-1}] \\
&\leq -2\eta_{t-1}(\mathbf{X}_{t-1} - \mathbf{X}^*) \bullet \widehat{\nabla}_{t-1} + \eta_{t-1}^2\|\widehat{\nabla}_{t-1}\|_F^2 + 2\eta_{t-1}(\mathbf{X}_{t-1} - \mathbf{X}^*) \bullet \mathbb{E}[\widehat{\nabla}_{t-1}|\mathbf{X}_{t-1}] \\
&\underset{(a)}{\leq} 4\sqrt{2}\eta_{t-1}G + \eta_{t-1}^2 G^2 \underset{(b)}{\leq} 6\eta G,
\end{aligned}$$

8

where (a) follows the Cauchy-Schwarz inequality and plugging the Euclidean diameter of $\mathcal{S}_n$ and the bound G on the norm of the stochastic gradients, and (b) holds for any T sufficiently large. We continue to upper-bound the conditional variance. For any $2 \leq t \leq T$ we have that

$$
\begin{aligned}
\mathbf{Var}\left(Z_t|\mathcal{F}_{t-1}\right) &= \mathbf{Var}\left(\|\mathbf{Y}_t - \mathbf{X}^*\|_F^2 | \mathcal{F}_{t-1}\right) \\
&= \mathbf{Var}\left(\|\mathbf{X}_{t-1} - \mathbf{X}^* - \eta_{t-1}\widehat{\nabla}_{t-1}\|_F^2 | \mathbf{X}_{t-1}\right) \\
&= \mathbf{Var}\left(\eta_{t-1}^2 \|\widehat{\nabla}_{t-1}\|_F^2 - 2\eta_{t-1}(\mathbf{X}_{t-1} - \mathbf{X}^*) \bullet \widehat{\nabla}_{t-1} | \mathbf{X}_{t-1}\right) \\
&\underset{(a)}{\leq} 2\eta_{t-1}^4 \mathbf{Var}\left(\|\widehat{\nabla}_{t-1}\|_F^2 | \mathbf{X}_{t-1}\right) + 8\eta_{t-1}^2 \mathbf{Var}\left((\mathbf{X}_{t-1} - \mathbf{X}^*) \bullet \widehat{\nabla}_t | \mathbf{X}_{t-1}\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathbf{Var}\left(Z_t|\mathcal{F}_{t-1}\right) &\leq 2\eta_{t-1}^4 \mathbf{Var}\left(\|\widehat{\nabla}_{t-1}\|_F^2 | \mathbf{X}_{t-1}\right) + 8\eta_{t-1}^2 \mathbf{Var}\left((\mathbf{X}_{t-1} - \mathbf{X}^*) \bullet (\widehat{\nabla}_t - \nabla f(\mathbf{X}_{t-1})) | \mathbf{X}_{t-1}\right) \\
&\leq 2\eta_{t-1}^4 \mathbb{E}\left[\|\widehat{\nabla}_{t-1}\|_F^4 | \mathbf{X}_{t-1}\right] + 8\eta_{t-1}^2 \mathbb{E}\left[\left((\mathbf{X}_{t-1} - \mathbf{X}^*) \bullet (\widehat{\nabla}_{t-1} - \nabla f(\mathbf{X}_{t-1}))\right)^2 | \mathbf{X}_{t-1}\right] \\
&\underset{(b)}{\leq} 8\eta_{t-1}^2 \|\mathbf{X}_{t-1} - \mathbf{X}^*\|_F^2 \mathbb{E}\left[\|\widehat{\nabla}_{t-1} - \nabla f(\mathbf{X}_{t-1})\|_F^2 | \mathbf{X}_{t-1}\right] + 2\eta_{t-1}^4 G^4 \\
&\underset{(c)}{\leq} 2\eta_{t-1}^4 G^4 + 16\eta_{t-1}^2 \sigma^2 / L,
\end{aligned}
$$

where (a) follows since $\mathbf{Var}(X + Y) \leq 2(\mathbf{Var}(X) + \mathbf{Var}(Y))$, (b) follows from the Cauchy-Schwarz inequality, and (c) follows from plugging the Euclidean diameter of $\mathcal{S}_n$ and the variance of the mini-batch stochastic gradient.

Now, using a standard concentration argument for submartingales (see Theorem 7.3 in Chung and Lu (2006), which we apply with parameters $a_i = 0, \phi_i = 0$), we have that for any $\Delta = O(\text{poly}(\log T))$, $L = O(\text{poly}(\log T))$, $1/\eta = \tilde{\Theta}(\sqrt{T})$, and T large enough,

$$
\begin{aligned}
\Pr\left(Z_t \geq Z_1 + \Delta\right) &\leq \exp\left(\frac{-\Delta^2}{\sum_{i=1}^{t-1}(4\eta_i^4 G^4 + 32\eta_i^2 \sigma^2 / L) + 4\eta G\Delta}\right) \\
&\leq \exp\left(\frac{-\Delta^2}{4T\eta^4 G^4 + 32T\eta^2 \sigma^2 / L + 4\eta G\Delta}\right) \leq \exp\left(\frac{-\Delta^2}{40T\eta^2 \sigma^2 / L}\right).
\end{aligned}
$$

Thus, for $\Delta := \sqrt{40T\eta^2 \sigma^2 / L}\sqrt{\log\frac{1}{p'}}$ it holds that with probability at least $1 - p'$ that $\|\mathbf{Y}_t - \mathbf{X}^*\|_F^2 \leq \|\mathbf{Y}_1 - \mathbf{X}^*\|_F^2 + G^2 \sum_{i=1}^{t-1} \eta_i^2 + \Delta \leq \|\mathbf{Y}_1 - \mathbf{X}^*\|_F^2 + G^2 T\eta^2 + \Delta$.

Now, recalling that $\mathbf{Y}_1 = \mathbf{X}_1$, and since $\mathbf{X}_t$ is the projection of $\mathbf{Y}_t$ onto $\mathcal{S}_n$, we have that with probability at least $1 - p'$ it holds that $\|\mathbf{X}_t - \mathbf{X}^*\|_F^2 \leq \|\mathbf{X}_1 - \mathbf{X}^*\|_F^2 + G^2 T\eta^2 + \Delta$. The Lemma now follows from setting $p' = p/T$ and using the union-bound for all $t \in [T]$. ∎

We can now prove Theorem 4.

Proof [Proof of Theorem 4]

Suppose for now that the iterates $\mathbf{X}_1, \ldots, \mathbf{X}_T$ are computed using exact Euclidean projection. Then, standard results (see for instance proof of Theorem 6.1 in Bubeck et al. (2015)) give that for any step-size $\eta > 0$ it holds that

$$
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{X}_t)\right] - f^* = O\left(\frac{\|\mathbf{X}_1 - \mathbf{X}^*\|_F^2}{\eta T} + \eta G^2\right).
$$

Thus, for both options of the returned solution $\bar{\mathbf{X}}$ in Algorithm 1 (using the convexity of $f(\cdot)$ for **option II**), we have that

$$\mathbb{E}\left[f(\bar{\mathbf{X}})\right] - f^* = O\left(\frac{\|\mathbf{X}_1 - \mathbf{X}^*\|_F^2}{\eta T} + \eta G^2\right).$$

In particular, using Markov's inequality, we have that with probability at least $3/4$ it holds that

$$f(\bar{\mathbf{X}}) - f^* = O\left(\frac{\|\mathbf{X}_1 - \mathbf{X}^*\|_F^2}{\eta T} + \eta G^2\right).$$

Using a standard Matrix Hoeffding concentration argument (see for instance Tropp (2012)), we have that with probability at least $9/10$, under the batch-size listed in the theorem, it holds that $\forall t \in [T-1]: \|\widehat{\nabla}_t - \nabla f(\mathbf{X}_t)\| \le \delta/(2r)$.

Also, using Lemma 8, we have for any T sufficiently large that with probability at least $7/8$ it holds that for all $1 \le t \le T-1$ $\|\mathbf{X}_t - \mathbf{X}^*\|_F^2 \le \|\mathbf{X}_1 - \mathbf{X}^*\|_F^2 + G^2 T \eta^2 + \sqrt{40\eta^2\sigma^2 T/L}\sqrt{\log(8T)}$.

Thus, we have that for $\eta = \frac{R_0}{10G\log(8T)\sqrt{T}}$ and $L \ge (\sigma/G)^2 R_0^{-2}$, combining all of the above guarantees, we have that with probability at least $1/2$, all following three guarantees hold:

$$i)\ f(\bar{\mathbf{X}}) - f^* = O\left(\frac{GR_0\log T}{\sqrt{T}}\right), \quad ii)\ \forall t \in [T-1]: \quad \|\mathbf{X}_t - \mathbf{X}^*\|_F \le R_0,$$

$$iii)\ \forall t \in [T-1]: \quad \|\widehat{\nabla}_t - \nabla f(\mathbf{X}_t)\| \le \frac{\delta}{2r}.$$

Thus, by invoking Lemma 7, with the above probability, for all $t \in [T]$ it holds that $\text{rank}(\mathbf{X}_t) \le r$. In particular, using **option I** in Algorithm 1, the returned solution $\bar{\mathbf{X}}$ is also of rank at most r. ∎

3. Preliminary Empirical Evidence

Table 1: Information on experiments. Column $\text{rank}(\mathbf{X}^*)$ is taken from Garber (2019). Column "SVD rank" records the SVD rank used to compute the projection on each iteration, and the column "max rank" records the maximum rank of any of the iterates produced by the algorithm.

setting		low rank SGD			high rank SGD		
trace (τ)	rank($\mathbf{X}^*$)	step-size	SVD rank	max rank	step-size	SVD rank	max rank
3000	10	0.02	10	10	$1/\sqrt{t}$	250	250
3500	41	0.007	41	41	$1/\sqrt{t}$	250	250
4000	70	0.005	70	70	$1/\sqrt{t}$	250	250

The goal of this section is to motivate our theoretical investigation from an empirical point of view. Our main result, Theorem 4, relies on an eigen-gap assumption (Assumption 1), a "warm-start" initialization, and certain choice of step-size which depends on several parameters. In Garber (2019) it was already demonstrated that Assumption 1 holds empirically for the highly popular matrix completion task. Here, we demonstrate empirically, that SGD with low-rank projections converges correctly (i.e., the projection with low-rank SVD is always the accurate projection) for matrix completion with a very simple initialization scheme, and is competitive with a standard implementation of SGD, which uses high-rank SVD.
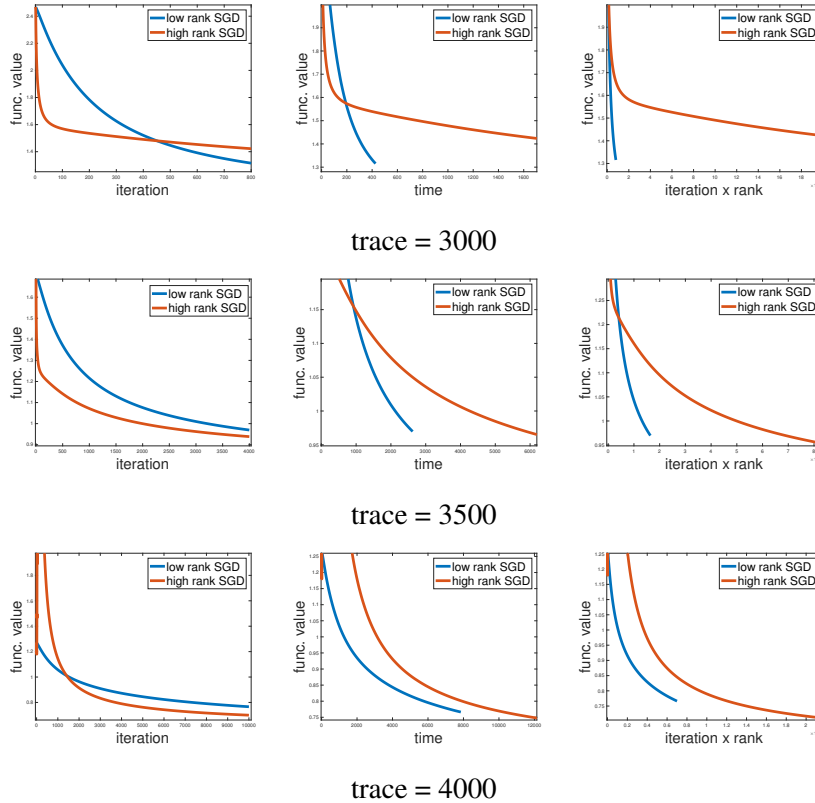
Figure 1: Performance of low rank SGD vs. (standard) high rank SGD. Each graph is the average of 5 i.i.d runs.

We use the standard MovieLens100K dataset (943x1682 matrix with 100,000 observed entries) Harper and Konstan (2016). [4] Formally, our objective is the following:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}:\|\mathbf{X}\|_*\leq\tau} \frac{1}{|S|} \sum_{(i,j,r)\in S} (\mathbf{X}_{i,j} - r)^2, \tag{8}$$

where S is the set of observed entries (each entry is a triplet consisting of a matrix entry (i,j) and a scalar ranking (r)), and $\|\cdot\|_*$ is the trace norm of a matrix (sum of singular values). Problem (8) could be directly formulated in the form of the canonical problem (1) using standard manipulations (see for instance Jaggi and Sulovský (2010)).

Following the experiments in Garber (2019), we use different values for the trace norm bound τ, which in turn affects the rank of the optimal solution $\mathbf{X}^*$. For both variants of SGD and for all experiments we use a batch-size of $L = 5000$ (5% of the data).

For low rank SGD we always compute the projection using thin SVD with rank equal that of the optimal solution (see Table 1). Also, we use a fixed step-size on all iterations which is tuned manually for every value of τ, so that indeed throughout all iterations, the rank of the true projection is at most the rank of the SVD used (which we verify by examining the condition on the threshold

---

parameter λ in Lemma 1). Thus, to be clear, with this tuned step-size, the low rank projection is **always** (up to negligible numerical error) the correct projection, which matches our theoretical investigation.

For the standard (high rank) implementation of SGD, in order to allow for more realistic implementation, we set the SVD rank used to compute the projection to 250 (instead of $\min\{m, n\}$, see Table 1). In all experiments we use a diminishing step-size of $\eta_t = 1/\sqrt{t}$ which follows the standard theoretical convergence results on SGD (up to constants, see Bubeck et al. (2015) for instance), without additional tuning.

We initialize both variants with the same point (based on assigning each unobserved entry the mean value of the observed ones and taking a low rank SVD with rank that matches that of the optimal solution). Each experiment is the average of 5 i.i.d runs (due to the randomness in the mini-batch). The experiments were implemented in MATLAB with the svds command used to compute thin SVD. We record the objective value (8) as a function of the number of iterations (for both variants we calculate the objective at the average of iterates obtained so far), and the runtime (in seconds). Additionally, to give an approximate measure of time that is implementation-independent, we also plot the function value vs. the number of iterations scaled by the SVD rank used by each algorithm. This is because in theory (and also often in practice) the time to compute a thin SVD scales linearly with the rank of the SVD required.

It can be seen in Figure 1 that standard SGD (with step size $1/\sqrt{t}$) seems to exhibit faster converge rates in terms of #iterations (perhaps with $\tau = 3000$ being the exception), due to the smaller step-size required by the low rank variant to guarantee low rank projections. However, when examining either the runtime or the convergence rate scaled by SVD rank, we see that as expected, low rank SGD is significantly faster. Also, as recorded in Table 1, while all iterates of low rank SGD indeed remain low rank, the iterates of high rank SGD always reach at some point the maximal rank used of 250, indicating that using a larger step-size indeed comes with a price.

4. Discussion

The main message we hope to convey in this work is that, perhaps in contrast to current popular belief, convex optimization methods can indeed be efficient for large-scale low-rank matrix problems, from the point of view of both theory and practice. We thus believe that it is worthwhile to continue studying their efficient implementations, perhaps under suitable assumptions.

There are two avenues for further research which could be of interest. First, Theorem 4 holds only with constant probability and not with high probability. Second, our analysis requires taking mini-batches. Since our objective is smooth, we may expect that these mini-batches will improve the convergence rate (see for instance Theorem 6.3 in Bubeck et al. (2015) which, roughly speaking, shows the rate improves by a factor of $\sqrt{L}$, where L is the mini-batch size). Unfortunately, our current analysis requires taking too small step-sizes (in order for the iterates to stay close enough to the optimal solution, see Lemma 8) to leverage the variance reduction due to the mini-batch.

Acknowledgments

References

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

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016b.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

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

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

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

Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127, 2006.

Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *International Conference on Machine Learning*, pages 2332–2341, 2015.

Dan Garber. On the convergence of projected-gradient methods with low-rank projections for smooth convex minimization over trace-norm balls and related problems. *CoRR*, abs/1902.01644, 2019. URL http://arxiv.org/abs/1902.01644.

Dan Garber and Atara Kaplan. Fast stochastic algorithms for low-rank and nonsmooth matrix problems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 286–294, 2019.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):19, 2016.

Elad Hazan and Satyen Kale. Projection-free online learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML*, 2012.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1):2489–2512, 2014. URL http://dl.acm.org/citation.cfm?id=2670328.

Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. *CoRR*, abs/1602.02101, 2016.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning, ICML*, 2013.

Martin Jaggi and Marek Sulovský. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th International Conference on Machine Learning, ICML*, 2010.

Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 4520–4528, 2016.

Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

Cun Mu, Yuqian Zhang, John Wright, and Donald Goldfarb. Scalable robust matrix recovery: Frank–wolfe meets proximal methods. *SIAM Journal on Scientific Computing*, 38(5):A3291–A3317, 2016.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

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

Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578. Omnipress, 2012.

Benjamin Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, Aug 2012.

John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in neural information processing systems*, pages 2080–2088, 2009.

Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust pca via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2014.

Alp Yurtsever, Madeleine Udell, Joel A. Tropp, and Volkan Cevher. Sketchy decisions: Convex low-rank matrix optimization with optimal storage. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pages 1188–1196, 2017. URL http://proceedings.mlr.press/v54/yurtsever17a.html.

Appendix A. Proof of Lemma 3

We first restate the lemma and then prove it.

Lemma 9 *Let $f : \mathbb{S}^n \to \mathbb{R}$ be β-smooth and convex. Let $\mathbf{X}^* \in \mathcal{S}_n$ be an optimal solution of rank r to the optimization problem $\min_{\mathbf{X} \in \mathcal{S}_n} f(\mathbf{X})$. Let $\mu_1, \ldots, \mu_n$ denote the eigenvalues of $\nabla f(\mathbf{X}^*)$ in non-increasing order. Then, $\mu_{n-r} = \mu_n$ if and only if for any arbitrarily small $\zeta > 0$ it holds that*

$$rank(\Pi_{(1+\zeta)\mathcal{S}_n}[\mathbf{X}^* - \beta^{-1}\nabla f(\mathbf{X}^*)]) > r,$$

where $(1+\zeta)\mathcal{S}_n = \{(1+\zeta)\mathbf{X} \mid \mathbf{X} \in \mathcal{S}_n\}$, and $\Pi_{(1+\zeta)\mathcal{S}_n}[\cdot]$ denotes the Euclidean projection onto the convex set $(1+\zeta)\mathcal{S}_n$.

Proof Let us write the eigen-decomposition of $\mathbf{X}^*$ as $\mathbf{X}^* = \sum_{i=1}^r \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$. It follows from the optimality of $\mathbf{X}^*$ that for all $i \in [r]$, $\mathbf{v}_i$ is also an eigenvector of $\nabla f(\mathbf{X}^*)$ which corresponds to the smallest eigenvalue μ_n (see Lemma 7 in Garber (2019)). Thus, if we let $\rho_1, \ldots, \rho_n$ denote the eigenvalues (in non-increasing order) of $\mathbf{Y} := \mathbf{X}^* - \beta^{-1}\nabla f(\mathbf{X}^*)$, it holds that

$$\begin{aligned} \forall i \in [r]: \quad \rho_i &= \lambda_i - \beta^{-1}\mu_n; \\ \forall i > r: \quad \rho_i &= \lambda_i - \beta^{-1}\mu_{n-i+1}. \end{aligned}$$

Recall that $\sum_{i=1}^r \lambda_i = 1$ and $\lambda_{r+1} = 0$.

It is well known that for any matrix $\mathbf{M} \in \mathbb{S}^n$ with eigen-decomposition $\mathbf{M} = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{u}_i^\top$, the projection of $\mathbf{M}$ onto the set $(1+\zeta)\mathcal{S}_n$, for any $\zeta \geq 0$ is given by

$$\Pi_{(1+\zeta)\mathcal{S}_n}[\mathbf{M}] = \sum_{i=1}^n \max\{0, \ \sigma_i - \sigma\}\mathbf{u}_i \mathbf{u}_i^\top,$$

where $\sigma \in \mathbb{R}$ is the unique scalar such that $\sum_{i=1}^n \max\{0, \ \sigma_i - \sigma\} = 1 + \zeta$.

Now, we can see that $rank(\Pi_{(1+\zeta)\mathcal{S}_n}[\mathbf{Y}]) \leq r$ if and only if $\sigma \geq \rho_{r+1} = -\beta^{-1}\mu_{n-r}$. However, in this case, we have

$$\begin{aligned} 1 + \zeta = \sum_{i=1}^n \max\{0, \ \rho_i - \sigma\} &= \sum_{i=1}^r \max\{0, \ \rho_i - \sigma\} \leq \sum_{i=1}^r \max\{0, \ \rho_i - (-\beta^{-1}\mu_{n-r})\} \\ &= \sum_{i=1}^r (\rho_i - (-\beta^{-1}\mu_{n-r})) = \sum_{i=1}^r (\lambda_i + \beta(\mu_{n-r} - \mu_n)) \\ &= 1 + \beta r(\mu_{n-r} - \mu_n) < 1 + \zeta \quad \forall \zeta > \beta r(\mu_{n-r} - \mu_n). \end{aligned}$$

Thus, for any fixed $\zeta > 0$, it follows that $rank(\Pi_{(1+\zeta)\mathcal{S}_n}[\mathbf{Y}]) \leq r$ if and only if $\beta r(\mu_{n-r} - \mu_n) \geq \zeta$. This proves the lemma. ∎

Appendix B. Proof of Corollary 5

We first restate the corollary.

Corollary 10 *The overall sample complexity to achieve $f(\bar{\mathbf{X}}) - f^* \leq \epsilon$ with probability at least $1/2$, when initializing from a "warm-start", is upper-bounded by $\tilde{O}\left(\epsilon^{-2} \max\{\sigma^2, \lambda_r^2(\mathbf{X}^*) r G^2\}\right)$ (note that $\lambda_r(\mathbf{X}^*) \leq 1/r$).*

Proof The overall sample complexity is given simply by the number of iterations to reach ϵ error times the size of the minibatch and is thus upper-bounded by:

$$\tilde{O}\left(\frac{G^2 R_0^2}{\epsilon^2} \cdot \max\{(\sigma/G)^2 R_0^{-2}, \frac{B^2 r^2}{\delta^2}\}\right) = \tilde{O}\left(\frac{1}{\epsilon^2} \max\{\sigma^2, \frac{R_0^2 G^2 B^2 r^2}{\delta^2}\}\right) =$$

$$\tilde{O}\left(\frac{1}{\epsilon^2} \max\{\sigma^2, \left(\frac{\lambda_r(\mathbf{X}^*)}{\lambda_r(\mathbf{X}^*) r \beta + \sqrt{r}B}\right)^2 G^2 B^2 r^2\}\right) =$$

$$\tilde{O}\left(\frac{1}{\epsilon^2} \max\{\sigma^2, \left(\frac{\lambda_r(\mathbf{X}^*)}{\sqrt{r}B}\right)^2 G^2 B^2 r^2\}\right) = \tilde{O}\left(\frac{1}{\epsilon^2} \max\{\sigma^2, \lambda_r^2(\mathbf{X}^*) r G^2\}\right).$$

∎