SVD-free Convex-Concave Approaches for Nuclear Norm Regularization

Yichi Xiao¹*, Zhe Li²*, Tianbao Yang², Lijun Zhang¹
¹National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
{xiaoyc, zhanglj}@lamda.nju.edu.cn
²Department of Computer Science, the University of Iowa, Iowa City, IA 52242, USA
{zhe-li-1, tianbao-yang}@uiowa.edu

Abstract
Minimizing a convex function of matrices regularized by the nuclear norm arises in many applications such as collaborative filtering and multi-task learning. In this paper, we study the general setting where the convex function could be non-smooth. When the size of the data matrix, denoted by \( m \times n \), is very large, existing optimization methods are inefficient because in each iteration, they need to perform a singular value decomposition (SVD) which takes \( O(m^2n) \) time. To reduce the computation cost, we exploit the dual characterization of the nuclear norm to introduce a convex-concave optimization problem and design a subgradient-based algorithm without performing SVD. In each iteration, the proposed algorithm only computes the largest singular vector, reducing the time complexity from \( O(m^2n) \) to \( O(mn) \). To the best of our knowledge, this is the first SVD-free convex optimization approach for nuclear-norm regularized problems that does not rely on the smoothness assumption. Theoretical analysis shows that the proposed algorithm converges at an optimal \( O(1/\sqrt{T}) \) rate where \( T \) is the number of iterations. We also extend our algorithm to the stochastic case where only stochastic subgradients of the convex function are available and a special case that contains an additional non-smooth regularizer (e.g., \( \ell_1 \) norm regularizer). We conduct experiments on robust low-rank matrix approximation and link prediction to demonstrate the efficiency of our algorithms.

1 Introduction
Low-rank matrices are preferred in real applications for different reasons. For instance, collaborative filtering uses low-rank matrices to model the fact that preferences of users are limited [Candès and Recht, 2009; Abernethy et al., 2009]. Multi-task learning uses low-rank matrices to enforce different tasks to share a common structure [Argyriou et al., 2008; Pong et al., 2010]. To yield low-rank solutions, the following nuclear-norm regularized problems have been widely adopted:

\[
\min_{A \in \mathbb{R}^{m \times n}} F(A) = f(A) + \lambda \|A\|_*
\]

where \( f(\cdot) \) is a convex loss, \( \lambda > 0 \) is a regularization parameter, and \( \|A\|_* = \text{trace}(\sqrt{A^T A}) \) denotes the nuclear norm of \( A \) (i.e., the sum of all the singular values), which is also referred to as the trace norm. Without loss of generality, we assume \( m \leq n \).

The optimization problem in (1) can be solved by first-order optimization methods such as subgradient descent [Nesterov, 2004], proximal gradient descent [Duchi and Singer, 2009; Nesterov, 2013]. Although these methods are guaranteed to converge, they are inefficient because a singular value decomposition (SVD), which takes \( O(m^2n) \) time, is required at each iteration. To reduce the computation complexity, many efficient solvers have been developed by replacing the full SVD with a partial SVD. However, those approaches either require the function \( f(\cdot) \) to be smooth [Dudık et al., 2012; Hsieh and Olsen, 2014] or are designed for nuclear-norm constrained problems instead of regularized problems [Jaggi et al., 2010; Jaggi, 2013].

In this paper, we study the general setting of (1), where the function \( f(\cdot) \) could be non-smooth, and develop a series of SVD-free optimization algorithms. First, based on the dual characterization of the nuclear norm, we reformulate (1) as a convex-concave optimization problem, and solve it by the subgradient method. In each iteration, we only need to compute the largest singular vector instead of a full SVD, thus reducing the time complexity from \( O(m^2n) \) to \( O(mn) \). As far as we know, this is the first SVD-free convex optimization algorithm for general nuclear-norm regularized problems. Theoretically the proposed algorithm converges at an \( O(1/\sqrt{T}) \) rate, which matches the optimal rate of non-smooth optimization under the first-order black-box model [Nesterov, 2004]. Second, we extend our algorithm to stochastic composite optimization, where only stochastic subgradients of \( f(\cdot) \) are available, and show that an \( O(1/\sqrt{T}) \) rate of convergence is still attainable. Finally, we study the case where an additional non-smooth regularizer such as the \( \ell_1 \)-norm is presented, and propose a proximal subgradient method that solves the problem at the same rate.

*Equal Contributions
Applications As mentioned above, nuclear norm regularization occurs in many machine learning applications. To motivate this work, we list some applications which involve a convex but non-smooth loss function:

- **Robust Low-rank Matrix Approximation:** The goal is to fit a target matrix $Y$ with a low-rank matrix $A$ in a robust way [Baccini et al., 1996; Croux and Filzmoser, 1998; Ke and Kanade, 2005]. For this problem, the objective function $f(A) = \sum_{ij} |A_{ij} - Y_{ij}| / mn$, which is non-smooth.

- **Sparse and Low-rank Link Prediction:** The goal is to discover links from a partially observed adjacency matrix $Y$ such that the link matrix is both sparse and low-rank [Richard et al., 2012]. The objective function is defined as $f(A) = \ell(A, Y) + \gamma \|A\|_1$, where $\ell(A, Y)$ is the empirical average of hinge loss over observed entries and thus non-smooth.

## 2 Related Work

In this section, we provide a brief review of existing methods for nuclear-norm regularized problems, as well as related work on nuclear-norm constrained problems.

### 2.1 Nuclear-norm Regularized Problems

Owing to the non-smoothness nature of the nuclear norm, the conventional approach for solving (1) is the subgradient descent (GD):

$$A_{t+1} = A_t - \eta_t \nabla f(A_t) + \partial \|A_t\|_*$$

where $\partial \|A_t\|_*$ denotes a subgradient of $\|\cdot\|_*$ evaluated at $A_t$, and $\eta_t > 0$ is the step size. It is well-known that GD converges to the optimum at an $O(1/\sqrt{T})$ rate, which is already optimal for first-order optimization of non-smooth functions [Nemirovski et al., 1982; Nesterov, 2004]. When the function $f(\cdot)$ is smooth, proximal gradient descent (PGD), defined as

$$A_{t+1} = \arg\min_{A \in \mathbb{R}^{m \times n}} \frac{1}{2} \|A - (A_t - \eta_t \nabla f(A_t))\|_F^2 + \eta_t \lambda \|A\|_*$$

is preferred and achieves an $O(1/T)$ rate of convergence [Nesterov, 2013]. Following the Nesterov’s method for accelerating the gradient method [Nesterov, 2004], an accelerated version of PGD that converges at an $O(1/T^2)$ rate has been developed [Ji and Ye, 2009; Toh and Yun, 2010]. Although GD and PGD are guaranteed to converge, they need to calculate the SVD of $A_t$ or $A_t - \eta_t \nabla f(A_t)$ in each iteration [Cai et al., 2010], which takes $O(m^2n)$ time. Due to the high computational cost of full SVD, GD and PGD do not scale well to large-scale problems.

To reduce the computational complexity, many efficient nuclear norm minimization solvers have been developed. In [Dudik et al., 2012], the authors lift the non-smooth convex problem into an infinitely dimensional smooth problem and apply coordinate descent to solve it. In each round, the algorithm only needs to calculate a partial SVD instead of a full SVD. In [Hsieh and Olsen, 2014], the authors propose the active subset selection (ASS) algorithm, which selects an active subspace by approximating SVD and then cast (1) to a

## 3 Main Results

In this section, we introduce the details of our SVD-free convex-concave optimization algorithm and several extensions to the basic version.

### 3.1 The Basic Algorithm

We first recall the dual characterization of the nuclear norm, i.e.,

$$\|A\|_* = \max_{U \in \mathbb{R}^{m \times n}, U \|U\|_2 \leq 1} tr(U^T A)$$

where $\|U\|_2$ represents the spectral norm of $U$. Then, we cast (1) to the following problem

$$\min_{A \in \mathbb{R}^{m \times n}} \max_{U \in \mathbb{R}^{m \times n}, \|U\|_2 \leq 1} f(A) + \lambda tr(U^T A)$$

(2)
Algorithm 1 SVD-free CONvex-ConcavE Algorithm (SÉONE)

1: Initialize: $A_1 = U_1 = 0 \in \mathbb{R}^{m \times n}$
2: for $t = 1$ to $T$
3: \hspace{1em} Update $A_{t+1}$ by
4: \hspace{2em} $A_{t+1} = A_t - \eta_t(\partial f(A_t) + \lambda U_t)$
5: \hspace{1em} Update $U_{t+1}$ by
6: \hspace{2em} $U_{t+1} = U_t + \tau_t(\lambda A_t - \rho \partial[\|U_t\|_2 - 1]_+)$
7: \hspace{1em} end for
8: \hspace{1em} Output: $\hat{A}_T = \sum_{t=1}^{T} A_t/T$

Since the above optimization problem is convex-concave, we can apply the standard subgradient method to solve it. However, due to the presence of the spectral norm constraint of $U$, we have to project the intermediate solution onto the unit spectral norm ball, which again requires a full SVD operation.

To address this issue, we propose to remove the constraint $\|U\|_2 \leq 1$, and introduce an additional term into the objective function to control the spectral norm of $U$:

$$\min_{A \in \mathbb{R}^{m \times n}} \max_{U \in \mathbb{R}^{n \times n}} f(A) + \lambda tr(U^T A) - \rho \|U\|_2 - 1\|_+$$ (3)

where $\rho > 0$ is a parameter whose value will be specified later and $[s]_+ = \max\{s, 0\}$ is the hinge operator. To solve the above problem, we can use the standard subgradient method, which iterates as follows:

$$A_{t+1} = A_t - \eta_t(\partial f(A_t) + \lambda U_t),$$
$$U_{t+1} = U_t + \tau_t(\lambda A_t - \rho \partial[\|U_t\|_2 - 1]_+).$$

Note that the subgradient $\partial[\|U\|_2 - 1]_+$ can be computed efficiently. To show this, we denote $s_1$ the leading singular value of $U$, and $u_1, v_1$ the corresponding left and right singular vectors. Then, we have

$$u_1 v_1^T [s_1 > 1] \in \partial[\|U\|_2 - 1]_+.$$

This implies that in each iteration, we only need to compute the leading singular vectors of $U_t$ with $O(mn)$ time. By contrast, a full SVD takes $O(m^2 n)$ time. The detailed procedure is summarized in Algorithm 1.

To present the theoretical guarantee of our algorithm, we make the following assumptions of $f(\cdot)$:

**Assumption 1.** Assume that $f(A) \geq 0$ for any $A \in \mathbb{R}^{m \times n}$ and there exists $C > 0$ and $G > 0$ such that $f(0) \leq C$ and $\|\partial f(A)\|_F \leq G$.

We have the following theorem regarding the optimization error.

**Theorem 1.** Let $\rho \geq C$ and run Algorithm 1 with $\eta_t = c_1/\sqrt{t}$ and $\tau_t = c_2/\sqrt{t}$. Let $\hat{A}_T = \sum_{t=1}^{T} A_t/T$ be the output and

$$U_T^* = \arg\max_{\|\|U\|_2 \leq 1} tr(U^T \hat{A}_T).$$

Under Assumption 1 and $\max\{\|A_t\|_F, \|U_t\|_F\} \leq \sigma$, we have

$$F(\hat{A}_T) - F(A_*) \leq \frac{1}{\sqrt{T}} \left( \frac{D_1^2}{2c_1} + c_1(G + \lambda\sigma)^2 \right) + \frac{1}{\sqrt{T}} \left( \frac{D_2^2}{2c_2} + c_2(\rho + \lambda\sigma)^2 \right)$$

where $D_1 = \|A_*\|_F + \sigma$ and $D_2 = \|U_T^*\|_F + \sigma$.

3.2 The Stochastic Setting

In this subsection, we extend the basic algorithm to the stochastic setting: $f(A) = E_{\xi}[f(A; \xi)]$, where $\xi$ is a random variable. In this case, the optimization problem becomes

$$\min_{A \in \mathbb{R}^{m \times n}} E_{\xi}[f(A; \xi)] + \lambda \|A\|_*.$$

Although existing algorithms for stochastic composite optimization [Lan, 2012; Lin et al., 2014] can be applied to the above problem, they are inefficient because a full SVD operation is required in each iteration.

Following the derivation of (3), we convert the above problem to the unconstrained convex-concave optimization problem

$$\min_{A \in \mathbb{R}^{m \times n}} \max_{U \in \mathbb{R}^{n \times n}} E_{\xi}[f(A; \xi)] + \lambda tr(U^T A) - \rho \|U\|_2 - 1\|_+$$ (4)

Generally speaking, it is impossible to compute the gradient of $E_{\xi}[f(A; \xi)]$ w. r. t. $A$, thus the subgradient algorithm in Algorithm 1 cannot be applied here. Instead, we will first sample a random variable $\xi_t$ and use the stochastic gradient $\partial f(A_t; \xi_t)$ to update the intermediate solution. Specifically, the updating rules are as follows:

$$A_{t+1} = A_t - \eta_t(\partial f(A_t; \xi_t) + \lambda U_t),$$
$$U_{t+1} = U_t + \tau_t(\lambda A_t - \rho \partial[\|U_t\|_2 - 1]_+).$$

The complete procedure is summarized in Algorithm 2.

Before presenting the convergence rate, we make the following assumption.

**Assumption 2.** Assume that $f(A) \geq 0$ for any $A \in \mathbb{R}^{m \times n}$ and there exists $C > 0$ and $G > 0$ such that $f(0) \leq C$ and $E_{\xi} \|\partial f(A; \xi)\|_F^2 \leq G^2$.

The above assumption requires that the stochastic gradient is bounded in expectation, which is different from Assumption 1 in the deterministic setting. Then, we have the following theorem.

**Theorem 2.** Let $\rho \geq C$ and run Algorithm 2 with $\eta_t = c_1/\sqrt{t}$ and $\tau_t = c_2/\sqrt{t}$. Let $\hat{A}_T = \sum_{t=1}^{T} A_t/T$ be the output and $U_T^* = \arg\max_{\|\|U\|_2 \leq 1} tr(U^T \hat{A}_T)$. Under Assumption 2 and $\max\{\|A_t\|_F, \|U_t\|_F\} \leq \sigma$, we have

$$E[F(\hat{A}_T)] - F(A_*) \leq \frac{1}{\sqrt{T}} \left( \frac{D_1^2}{2c_1} + c_1(G + \lambda\sigma)^2 \right) + \frac{1}{\sqrt{T}} \left( \frac{D_2^2}{2c_2} + c_2(\rho + \lambda\sigma)^2 \right)$$
The algorithm shares the same $O(1/\sqrt{T})$ rate of convergence with the deterministic version. Given the non-smoothness of the objective function, this rate cannot be improved in general.

3.3 Problems with an Additional Regularizer

In this subsection, we consider the case that besides the nuclear norm regularizer, there is an additional non-smooth regularizer. The optimization problem is given by

$$\min_{A \in \mathbb{R}^{m \times n}} f(A) + \gamma \phi(A) + \lambda \|A\|_*$$

where $\phi(\cdot)$ is a non-smooth regularizer such as $\|A\|_1$ and $\gamma > 0$ is a regularizer parameter. Note that the proximal gradient descent [Nesterov, 2013] can not be directly applied to (5), because there are two regularizers and there is no closed-form solution to the proximal mapping.

To address this limitation, we propose to solve the following problem

$$\min_{A \in \mathbb{R}^{m \times n}} \max_{U \in \mathbb{R}^{m \times n}} f(A) + \gamma \phi(A) + \lambda tr(U^T A) - \rho \|\|U\|_2 - 1\|_+.$$ 

In this way, we only have one regularizer $\gamma \phi(\cdot)$, which can be handled by proximal mapping. To be specific, we define the proximal mapping of the convex function $h(\cdot)$ as:

$$\text{Prox}_{\eta,h(\cdot)}(\hat{A}) = \arg \min_{A \in \mathbb{R}^{m \times n}} h(A) + \frac{1}{2\eta} \|A - \hat{A}\|_F^2.$$ 

We let $h(A) = \gamma \phi(A)$, and introduce the updating rules as:

$$\hat{A}_{t+1} = A_t - \eta_t \partial f(A_t) + \lambda U_t,$$

$$A_{t+1} = \text{Prox}_{\eta_t,\gamma \phi(\cdot)}(\hat{A}_{t+1}),$$

$$U_{t+1} = U_t + \tau_t (\lambda A_t - \rho \partial \|U_t\|_2 - 1\|_+).$$

The detailed procedure is presented in Algorithm 3.

We establish the convergence rate of Algorithm 3 in the following theorem.

**Theorem 3.** Under the same condition as Theorem 1 and assume $\phi(0) = 0$, we have

$$F(\hat{A}_T) - F(A_*) \leq \frac{1}{\sqrt{T}} \left( \frac{D_1^2}{2c_1} + c_1 (G + \lambda \sigma)^2 \right) + \frac{1}{\sqrt{T}} \left( \frac{D_2^2}{2c_2} + c_2 (\rho + \lambda \sigma)^2 \right)$$

where $D_1 = \|A_*\|_F + \sigma$ and $D_2 = \|U_*\|_F + \sigma$.

**Algorithm 2** Extension to Stochastic Setting (SECONE-S)

1: Initialize: $A_1 = U_1 = 0 \in \mathbb{R}^{m \times n}$
2: for $t = 1$ to $T$ do
3: Sample $\xi_t$
4: Update $A_{t+1}$ by
   $$A_{t+1} = A_t - \eta_t \partial f(A_t; \xi_t) + \lambda U_t$$
5: Update $U_{t+1}$ by
   $$U_{t+1} = U_t + \tau_t (\lambda A_t - \rho \partial \|U_t\|_2 - 1\|_+)$$
6: end for
7: Output: $\hat{A}_T = \sum_{t=1}^T A_t / T$

The theorem proves that the proximal variant of our algorithm also converges at an $O(1/\sqrt{T})$ rate. Though the convergence rate is as same as the one of SECONE, the proximal mapping for $\ell_1$ norm usually gives us sparse solutions. Also, it is trivial to extend this method to the stochastic optimization algorithm by following the derivation in Section 3.2.

4 Theoretical analysis

In this section, we provide proofs of Theorems 1 and 3. The proof of Theorem 2 is included in the supplementary.

4.1 Proof of Theorem 1

We start with the unconstrained convex-concave optimization problem (3) by denoting the objective function as

$$L(A, U) = f(A) + \lambda tr(U^T A) - \rho \|\|U\|_2 - 1\|_+.$$ 

For clarity, we divide the proof into two individual parts. Part 1: Recall that $A_{t+1}$ is the update of subgradient descent applied to $L(A_t, U_t)$, according to the standard analysis of subgradient descent update, we have for any $A \in \mathbb{R}^{m \times n}$

$$L(A_t, U_t) \leq L(A, U_t) + \frac{\eta_t}{2} \|\partial f(A_t) + \lambda U_t\|_F^2 + \frac{1}{2\eta_t} (\|A - A_t\|_F^2 - \|A - A_{t+1}\|_F^2)$$

Similarly, $U_{t+1}$ is the update of subgradient descent applied to $L(A_t, U_t)$, hence for any $U \in \mathbb{R}^{m \times n}$

$$L(A_t, U_t) \leq L(A_t, U_t) + \frac{\tau_t}{2} \|\lambda A_t - \rho \partial \|U_t\|_2 - 1\|_+^2 + \frac{1}{2\tau_t} (\|U - U_t\|_F^2 - \|U - U_{t+1}\|_F^2)$$

With the assumption max$\|\|A_t\|_F, \|U_t\|_F\| \leq \sigma$ and $\|\partial f(A_t)\|_F \leq G$, we combine the above inequalities

$$L(A, U) \leq L(A, U_t) + \frac{\eta_t}{2} (G + \lambda \sigma)^2 + \frac{\tau_t}{2} (\rho + \lambda \sigma)^2 + \frac{1}{2\eta_t} (\|A - A_t\|_F^2 - \|A - A_{t+1}\|_F^2) + \frac{1}{2\tau_t} (\|U - U_t\|_F^2 - \|U - U_{t+1}\|_F^2)$$

**Algorithm 3** Extension to Proximal Variant (SECONE-P)

1: Initialize: $A_1 = U_1 = 0 \in \mathbb{R}^{m \times n}$
2: for $t = 1$ to $T$ do
3: Update $A_{t+1}$ by
   $$A_{t+1} = A_t - \eta_t \partial f(A_t) + \lambda U_t$$
4: Update $A_{t+1}$ by
   $$A_{t+1} = \arg \min_{A \in \mathbb{R}^{m \times n}} \gamma \phi(A) + \frac{1}{2\eta_t} \|A - A_{t+1}\|_F^2$$
5: Update $U_{t+1}$ by
   $$U_{t+1} = U_t + \tau_t (\lambda A_t - \rho \partial \|U_t\|_2 - 1\|_+)$$
6: end for
7: Output: $\hat{A}_T = \sum_{t=1}^T A_t / T$
Let $\eta_t = c_1/\sqrt{t}$, then simple mathematics shows that
\[
\sum_{t=1}^{T} \frac{1}{2\eta_t} (\|A - A_t\|_F^2 - \|A - A_{t+1}\|_F^2)
\leq \frac{1}{2\eta_1} \|A - A_1\|_F^2 + \sum_{t=1}^{T-1} \left( \frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t} \right) \|A - A_t\|_F^2
\leq \frac{1}{2\eta_1} D_1^2 + \left( \frac{1}{2\eta_T} - \frac{1}{2\eta_1} \right) D_1^2 \leq \frac{1}{2\eta_1} D_1^2 + \sqrt{T} \frac{1}{2\eta_2} D_2^2
\]
where $D_1 = \|A\| + \sigma \geq \max_t \|A - A_t\|_F$.

We can apply the same analysis for $c_2/\sqrt{t}$ and obtain
\[
\sum_{t=1}^{T} L(A_t, U_t) \leq \sum_{t=1}^{T} L(A, U_t) + \left( \frac{D_1^2}{2\eta_1} + c_1(G + \lambda\sigma)^2 \right) + \frac{1}{\sqrt{T}} \left( \frac{D_2^2}{2\eta_2} + c_2(\rho + \lambda\sigma)^2 \right)
\]
where we use the fact $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 1 + \int_{1}^{T} \frac{1}{\sqrt{t}} dt \leq 2\sqrt{T}$.

From the convexity of $g_t(A)$ and $h_t(A)$, we have
\[
\eta_t (g_t(A_t) + h_t(A_{t+1}) - g_t(A_t) - h_t(A))
\leq \langle A_t - A, \eta_t G_t \rangle + \langle A_{t+1} - A, \eta_t \partial h(A_{t+1}) \rangle
= \langle A - A_t + 1, A_t - A_{t+1} - \eta_t G_t + \eta_t \partial h(A_{t+1}) \rangle
+ \langle A - A_{t+1}, A_{t+1} - A_t \rangle + \eta_t \langle A - A_{t+1}, G_t \rangle
\]
from the optimality of $A_{t+1}$ in Algorithm 3, we have
\[
\langle A - A_{t+1}, A_t - A_{t+1} - \eta_t G_t + \eta_t \partial h(A_{t+1}) \rangle \geq 0.
\]

By combining the above two inequality and choosing $A = A_*$, which is the optimal solution, we obtain
\[
\eta_t (g_t(A_t) + h_t(A_{t+1}) - g_t(A_t) - h_t(A))
\leq \langle A_* - A_t + 1, A_t - A_{t+1} \rangle + \eta_t \langle A - A_{t+1}, G_t \rangle
\leq \frac{1}{2} (\|A_* - A\|_F^2 + \|A_* - A_{t+1}\|_F^2 - \|A_{t+1} - A_t\|_F^2)
+ \frac{1}{2} (\|A_t - A_{t+1}\|_F^2 + \eta_t^2 \|G_t\|_F^2)
\leq \frac{1}{2} (\|A_* - A\|_F^2 + \|A_* - A_{t+1}\|_F^2) + \frac{\eta_t^2}{2} \|G_t\|_F^2.
\]

The second inequality follows from Cauchy-Schwarz inequality. Let’s consider $L(A_t, U_t) - L(A_*, U_t)$, which is
\[
L(A_t, U_t) - L(A_*, U_t) = g_t(A_t) + h_t(\hat{A}_t) - g_t(A_t) - h_t(A) + \eta_t \langle \hat{A}_t - A_{t+1}, G_t \rangle
\leq \frac{1}{2\eta_t} (\|A_* - A_t\|_F^2 + \|A_* - A_{t+1}\|_F^2) + \frac{\eta_t}{2} \|G_t\|_F^2.
\]

Note that the trailing term $\gamma(\phi(A_t) - \partial h(A_{t+1}))$ has little impact on the convergence as we assume $\phi(A_1) = \phi(0) = 0$.

Using the same argument as in the proof of Theorem 1, we can then easily carry out the rest proof of this theorem.

5 Experiments

We present numerical experiments on real datasets to demonstrate the efficiency of the proposed algorithms, and more results can be found in the supplementary.
5.1 Robust Low-rank Matrix Approximation

We consider the robust low-rank approximation problem [Baccini et al., 1996; Croux and Filzmoser, 1998; Ke and Kanade, 2005]:

$$\min_{A \in \mathbb{R}^{m \times n}} \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} |Y_{ij} - A_{ij}| + \lambda \|A\|_*,$$

where $Y$ is a given data matrix. Due to the non-smoothness of the objective function, we compare our method with two classical methods: subgradient descent (GD) and proximal subgradient descent (PGD) [Duchi and Singer, 2009]. We use the News20$^2$ dataset, which contains $m = 11, 269$ instances, each of which has $n = 20,302$ features (we filter the features which appear less than 7 times). According to Theorem 1, we set step sizes in Algorithm 1 as $\eta = c_1 / \sqrt{t}$ and $\tau = c_2 / \sqrt{t}$, where $c_1, c_2$ are some constants. The same step size $\eta = c_1 / \sqrt{t}$ is also used for GD and PGD. We tune the value of $c_1$ and $c_2$ in a range of $\{1e-5, 1e-4, \ldots, 1e10\}$ and report the best results based on the objective value.

In Fig. 1, we plot the objective value versus the running time for $\lambda = 1e-6$. We choose this value of $\lambda$ because it can produce a low-rank output, and the convergence behavior is insensitive to $\lambda$. As can be seen, SECONE decreases much faster than GD and PGD. This is as expected because SECONE is SVD-free and time-efficient, which is also convinced by the statistics shown in Table 1. As can be seen, each iteration of SECONE takes much less time than other two methods.

5.2 Sparse and Low-rank Link Prediction

Given the adjacency matrix $Y$ of a graph with 0/1 filled entries, we consider the sparse and low-rank link prediction problem:

$$\min_{A \in \mathbb{R}^{m \times n}} \sum_{ij} \max(1 - (2Y_{ij} - 1) \cdot A_{ij}, 0) + \gamma \|A\|_1 + \lambda \|A\|_*.$$

Following the setting in [Richard et al., 2012], we perform experiments on the Facebook100 dataset which contains the friendship relations between students. We select a single university with 41,554 students and keep the first $m = n = 15,000$ users only with the highest degree. We flip 15% of randomly chosen entries and the goal is to learn a sparse and low-rank matrix from the noisy adjacency matrix $Y$.

We compare Algorithm 3 (SECONE-P) with subgradient descent (GD) and Incremental Proximal Decent (IPD), which is an iterative algorithm designed for the above problem but with no theoretical guarantees [Richard et al., 2012]. The step sizes in SECONE-P and GD are set in the same way as in Section 5.1. The parameter $\theta$ of IPD is searched in the range of $\{1e-2, 1e-3, \ldots, 1e2\}$.

In Fig. 2, we plot objective value versus the running time when $\lambda = 10$ and $\gamma = 0.4$. As can be seen, SECONE-P converges much faster than other methods. The tuning of IPD is somewhat tricky since it does not converge to the optimum. The statistics of different methods are shown in Table 2. Again, the running time per iteration of SECONE-P is much smaller than other methods.

6 Acknowledgements

This work was partially supported by the NSFC (61603177), JiangsuSF (BK20160658), and the Collaborative Innovation Center of Novel Software Technology and Industrialization of Nanjing University. Z. Li and T. Yang are partially supported by National Science Foundation (IIS-1463988, IIS-1545995).

References


