SVD-free Convex-Concave Approaches for Nuclear Norm Regularization

Yichi Xiao[1*], **Zhe Li**[2*], **Tianbao Yang**[2], **Lijun Zhang**[1]
[1]National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{xiaoyc, zhanglj}@lamda.nju.edu.cn
[2]Department 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 full 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., ℓ_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:

$$\min_{A \in \mathbb{R}^{m \times n}} F(A) = f(A) + \lambda \|A\|_* \tag{1}$$

where $f(\cdot)$ is a convex loss, $\lambda > 0$ is a regularization parameter, and $\|A\|_* = \mathrm{trace}(\sqrt{A^\top A})$ 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 ℓ_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):

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

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

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

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 [Dudık *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 small-size problem that can be solved easily. Although these algorithms are efficient in solving nuclear-norm regularized problem, they are restricted to the case that $f(\cdot)$ is smooth. In contrast, we only assume the function $f(\cdot)$ is convex and it could be non-smooth.

To avoid computing full SVDs, some methods rely on the following variational characterization of the nuclear norm

$$\|A\|_* = \min_{U,V:A=UV^\top} \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2)$$

where the size of the matrices U and V is not constrained. In [Srebro *et al.*, 2005], the authors formulate the problem as semi-define programming and solve it with standard SDP solvers. However, it can not scale up to large datasets. To deal with this, some authors [Rennie and Srebro, 2005; Signoretto *et al.*, 2013] propose alternating direction methods, which decrease the objective between different variables alternatively. However, these approaches break the convexity of the original problem and ensure no global convergence.

2.2 Nuclear-norm Constrained Problems

In [Jaggi *et al.*, 2010], the authors consider a constrained version of (1), i.e.,

$$\min_{A\in\mathbb{R}^{m\times n}} f(A) \text{ s.t. } \|A\|_* \le \tau.$$

They transform it to the problem of optimizing a convex function over the set of positive semi-definite matrices with unit trace, and then use the approximate SDP solver [Hazan, 2008], which in each iteration only calculates an approximate largest eigenvector of the gradient. A similar greedy algorithm has been developed for convex optimization with a low-rank constraint [Shalev-Shwartz *et al.*, 2011]

$$\min_{A\in\mathbb{R}^{m\times n}} f(A) \text{ s.t. } \operatorname{rank}(A) \le r.$$

Recently, the Frank-Wolfe Algorithm has been applied to the nuclear-norm constrained problems [Jaggi, 2013], and also avoids the full SVD operation.

Finally, we note that moving the constraint function into the objective function - a technique utilized in this work, has been leveraged for developing stochastic gradient methods with only one projection [Mahdavi *et al.*, 2012]. However, the differences include (i) we do not need to perform any projection at the end; (ii) their algorithm is to handle the constraint on the primal variable and our algorithm is to handle the constraint on the dual variable.

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.,

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

Algorithm 1 **S**VD-fre**E** **CON**vex-Concav**E** Algorithm (SECONE)

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 U_{t+1} by

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

5: **end for**
6: **Output:** $\widehat{A}_T = \sum_{t=1}^T A_t/T$

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

$$\min_{A\in\mathbb{R}^{m\times n}} \max_{U\in\mathbb{R}^{m\times n}, \|U\|_2\leq 1} f(A) + \lambda tr(U^\top A) \qquad (2)$$

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:

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

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 σ_1 the leading singular value of U, and $\mathbf{u}_1, \mathbf{v}_1$ the corresponding left and right singular vectors. Then, we have

$$\mathbf{u}_1\mathbf{v}_1^\top \mathbb{1}[\sigma_1 > 1] \in \partial[\|U\|_2 - 1]_+.$$

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 $O(m^2n)$ 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 $\|\partial f(A)\|_F \leq G$.*[1]

We have the following theorem regarding the optimization error.

[1]To ensure the gradient is bounded, we can add a norm constraint on A if necessary.

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 $\widehat{A}_T = \sum_{t=1}^T A_t/T$ be the output and $U_T^* = \mathrm{argmax}_{\|U\|_2\leq 1} tr(U^\top \widehat{A}_T)$. Under Assumption 1 and $\max\{\|A_t\|_F, \|U_t\|_F\} \leq \sigma$, we have*

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

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

The above theorem implies the proposed algorithm has an $O(1/\sqrt{T})$ convergence rate. And the upper bound is minimized by choosing $c_1 = \frac{D_1}{\sqrt{2}(G+\lambda\sigma)}$ and $c_2 = \frac{D_2}{\sqrt{2}(\rho+\lambda\sigma)}$.

3.2 The Stochastic Setting

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

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

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

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

Generally speaking, it is impossible to compute the gradient of $\mathbf{E}_\xi[f(A;\xi)]$ w. r. t. A, thus the subgradient algorithm in Algorithm 1 cannot be applied here. Instead, we will first sample a random variable ξ_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 $\mathbf{E}_\xi\left[\|[\partial f(A;\xi)]\|_F^2\right] \leq G$.*

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 $\widehat{A}_T = \sum_{t=1}^T A_t/T$ be the output and $U_T^* = \arg\max_{\|U\|_2\leq 1} tr(U^\top \widehat{A}_T)$. Under Assumption 2 and*

Algorithm 2 Extension to Stochastic Setting (SECONE-S)

Initialize: $A_1 = U_1 = 0 \in \mathbb{R}^{m\times n}$
for $t = 1$ **to** T **do**
  Sample ξ_t
  Update A_{t+1} by

$$A_{t+1} = A_t - \eta_t(\partial f(A_t;\xi_t) + \lambda U_t)$$

  Update U_{t+1} by

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

end for
Output: $\widehat{A}_T = \sum_{t=1}^T A_t/T$

$\max\{\|A_t\|_F, \|U\|_F\} \leq \sigma$, *we have*

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

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

The above theorem implies the stochastic version of our 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

$$\min_{A\in\mathbb{R}^{m\times n}} f(A) + \gamma\phi(A) + \lambda\|A\|_* \tag{5}$$

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

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

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:

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

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

$$\begin{aligned}\bar{A}_{t+1} &= A_t - \eta_t(\partial f(A_t) + \lambda U_t),\\ A_{t+1} &= \text{Prox}_{\eta_t,\gamma\phi(\cdot)}(\bar{A}_{t+1}),\\ U_{t+1} &= U_t + \tau_t(\lambda A_t - \rho\partial[\|U_t\|_2 - 1]_+).\end{aligned}$$

The detailed procedure is presented in Algorithm 3.

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

Algorithm 3 Extension to Proximal Variant (SECONE-P)

Initialize: $A_1 = U_1 = 0 \in \mathbb{R}^{m\times n}$
for $t = 1$ **to** T **do**
  Update $\bar{A}_{t+1}$ by

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

  Update A_{t+1} by

$$A_{t+1} = \operatorname*{arg\,min}_{A\in\mathbb{R}^{m\times n}} \gamma\phi(A) + \frac{1}{2\eta_t}\|A-\bar{A}_{t+1}\|_F^2$$

  Update U_{t+1} by

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

end for
Output: $\widehat{A}_T = \sum_{t=1}^T A_t/T$

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

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

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

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 ℓ_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 the main theorems.

4.1 Proof of Theorem 1

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

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

For clarity, we divide the proof into two individual parts.

Part I: 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 $U \in \mathbb{R}^{m\times n}$

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

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

$$\begin{aligned} L(A_t, U) \leq & L(A, U_t) + \frac{\eta_t}{2}(G + \lambda\sigma)^2 + \frac{\tau_t}{2}(\rho + \lambda\sigma)^2 \\ & + \frac{1}{2\eta_t}(\|A - A_t\|_F^2 - \|A - A_{t+1}\|_F^2) \\ & + \frac{1}{2\tau_t}(\|U - U_t\|_F^2 - \|U - U_{t+1}\|_F^2) \end{aligned} \tag{6}$$

Let $\eta_t = c_1/\sqrt{t}$, then simple mathematics shows that

$$\begin{aligned} & \sum_{t=1}^{T} \frac{1}{2\eta_t}(\|A - A_t\|_F^2 - \|A - A_{t+1}\|_F^2) \\ \leq & \frac{1}{2\eta_1}\|A - A_1\|_F^2 + \sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|A - A_t\|_F^2 \\ \leq & \frac{1}{2\eta_1}D_1^2 + \left(\frac{1}{2\eta_T} - \frac{1}{2\eta_1}\right)D_1^2 \leq \frac{1}{2\eta_T}D_1^2 = \frac{\sqrt{T}}{2c_1}D_1^2 \end{aligned}$$

where $D_1 = \|A\| + \sigma \geq \max_t \|A - A_t\|_F$.

We can apply the same analysis for $\tau_t = c_2/\sqrt{t}$ and obtain

$$\begin{aligned} \sum_{t=1}^{T} L(A_t, U) \leq & \sum_{t=1}^{T} L(A, U_t) + \frac{\sqrt{T}}{2c_1}D_1^2 + \frac{\sqrt{T}}{2c_2}D_2^2 \\ & + c_1\sqrt{T}(G + \lambda\sigma)^2 + c_2\sqrt{T}(\rho + \lambda\sigma)^2 \end{aligned}$$

where we use the fact $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 1 + \int_{t=1}^{T}\frac{1}{\sqrt{t}}\mathrm{d}t \leq 2\sqrt{T}$ and the notation $D_2 = \|U\|_F + \sigma \geq \max_t \|U - U_t\|_F$.

Denote that $\widehat{A}_T = \sum_{t=1}^{T} A_t/T$ and $\widehat{U}_T = \sum_{t=1}^{T} U_t/T$. By the convexity of $L(A, U)$ in terms of A and concavity in terms of U, we obtain

$$\begin{aligned} L(\widehat{A}_T, U) - L(A, \widehat{U}_T) \leq & \frac{1}{\sqrt{T}}\left(\frac{D_1^2}{2c_1} + c_1(G + \lambda\sigma)^2\right) \\ & + \frac{1}{\sqrt{T}}\left(\frac{D_2^2}{2c_2} + c_2(\rho + \lambda\sigma)^2\right) \end{aligned}$$

To summarize what we have proved, the gap $L(\widehat{A}_t, U) - L(A, \widehat{U}_t)$ decreases at an $O(1/\sqrt{T})$ rate, which indicates the solution $(\widehat{A}_T, \widehat{U}_T)$ converges to the optimal solution of the unconstrained convex-concave optimization problem in (3).

Part II: In the rest of the proof, we will show that the objective value $F(\widehat{A}_t)$ also converges to $F(A_*)$, where A_* is an optimal solution of the original problem. To see this, what we need to prove is that

$$F(\widehat{A}_T) - F(A_*) \leq L(\widehat{A}_T, U) - L(A_*, \widehat{U}_T).$$

Let U be $U_T^* = \arg\max_{\|U\|_2 \leq 1} tr(U^\top \widehat{A}_T)$ then

$$L(\widehat{A}_T, U) = f(\widehat{A}_T) + \lambda\|\widehat{A}_T\|_* = F(\widehat{A}_T).$$

Thus it remains to show that

$$\lambda\|A_*\|_* \geq \lambda tr(\widehat{U}_T^\top A_*) - \rho[\|\widehat{U}_T\|_2 - 1]_+. \tag{7}$$

Note that $tr(A^\top B) \leq \|A\|_2\|B\|_*$ for any matrices A and B. When $\|\widehat{U}_T\|_2 \leq 1$, it is easy to verify that (7) holds. In the following, we focus on the case $\|\widehat{U}_T\|_2 \geq 1$.

Let $\widehat{U}_T$ have the SVD $\widehat{U}_T = P\Sigma Q^\top$, where the diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, \sigma_{r+1}, \ldots, \sigma_m)$. The sequence of singular values $\{\sigma_i\}$ is non-increasing and satisfy $\sigma_r \geq 1 > \sigma_{r+1}$. Denote $\widetilde{U}_T$ the projection of $\widehat{U}_T$ onto the unit spectral norm ball. Specifically, it has the form $\widetilde{U}_T = P\widetilde{\Sigma}Q$, where $\widetilde{\Sigma} = \mathrm{diag}(1, \ldots, 1, \sigma_{r+1}, \ldots, \sigma_m)$. It follows that

$$\begin{aligned} & tr(\widetilde{U}_T^\top A_*) \leq \|A_*\|_*, \\ & \|\widehat{U}_T - \widetilde{U}_T\|_2 \leq \sigma_1 - 1 = \|\widehat{U}_T\|_2 - 1. \end{aligned}$$

We are now in a position to prove that (7) holds. From Assumption 1, it is easy to verify $\lambda\|A_*\|_* \leq C$. We have

$$\begin{aligned} & \lambda tr(\widehat{U}_T^\top A_*) - \lambda\|A_*\|_* \leq \lambda tr((\widehat{U}_T - \widetilde{U}_T)^\top A_*) \\ \leq & \lambda\|\widehat{U}_T - \widetilde{U}_T\|_2\|A_*\|_* \leq C(\|\widehat{U}_T\|_2 - 1) \\ \leq & \rho(\|\widehat{U}_T\|_2 - 1) \end{aligned}$$

which is due to the inequality $\lambda\|A_*\|_* \leq C$ and the parameter $\rho \geq C$. This completes the proof of Theorem 1.

4.2 Proof of Theorem 2

We may abuse some notations. Denote $v_1, v_2, \cdots, v_T$ be the sequence of stochastic subgradient $\partial f(A_t, \xi_t)$. For short, let $v_{1:T}$ denote this sequence $v_1, v_2, \cdots, v_T$. Let $L(A, U) = \mathbf{E}_\xi[f(A; \xi)] + \lambda tr(U^\top A) - \rho[\|U\|_2 - 1]_+$.

Due to the linearity of expectation, we shall adopt the same procedure as in the proof of Theorem 1 and we have

$$\begin{aligned} & \mathbf{E}_{v_{1:T}}\left[\sum_{t=1}^{T} L(A_t, U) - L(A, U_t)\right] \leq \frac{\sqrt{T}}{2c_1}D_1^2 + \frac{\sqrt{T}}{2c_2}D_2^2 \\ & + c_1\sqrt{T}(G + \lambda\sigma)^2 + c_2\sqrt{T}(\rho + \lambda\sigma)^2 \end{aligned}$$

where $D_1 = \|A\|_F + \sigma$ and $D_2 = \|U\|_F + \sigma$. The remaining proof is similar to the part II of Theorem 1 and we conclude the proof by following the same analysis.

4.3 Proof of Theorem 3

We may abuse some notations from the previous section. Let $L(A, U) = f(A) + \lambda tr(U^\top A) + \gamma\phi(A) - \rho[\|U\|_2 - 1]_+$. Denote that

$$\begin{aligned} & g_t(A) = f(A) + \lambda tr(U_t^\top A),\ h(A) = \gamma\phi(A) \\ & \text{and } G_t = \partial g_t(A_t) = \partial f(A_t) + \lambda U_t. \end{aligned}$$

From the convexity of $g_t(A)$ and $h(A)$, we have

$$\begin{aligned} & \eta_t(g_t(A_t) + h(A_{t+1}) - g_t(A) - h(A)) \\ \leq & \langle A_t - A, \eta_t G_t\rangle + \langle A_{t+1} - A, \eta_t\partial h(A_{t+1})\rangle \\ = & \langle A - A_{t+1}, A_t - A_{t+1} - \eta_t G_t - \eta_t\partial h(A_{t+1})\rangle \\ & + \langle A - A_{t+1}, A_{t+1} - A_t\rangle + \eta_t\langle A_t - A_{t+1}, G_t\rangle \end{aligned}$$

From the optimality of A_{t+1} in Algorithm 3, we have

$$\langle A - A_{t+1}, A_{t+1} - A_t + \eta_t G_t + \eta_t\partial h(A_{t+1})\rangle \geq 0.$$

By combining the above two inequality and choosing $A = A_*$ which is the optimal solution, we obtain

$$\begin{aligned}
&\eta_t(g_t(A_t) + h(A_{t+1}) - g_t(A_*) - h(A_*)) \\
\leq &\langle A_* - A_{t+1}, A_{t+1} - A_t\rangle + \eta_t\langle A_t - A_{t+1}, G_t\rangle \\
\leq &\frac{1}{2}\left(\|A_* - A_t\|_F^2 - \|A_* - A_{t+1}\|_F^2 - \|A_{t+1} - A_t\|_F^2\right) \\
&+ \frac{1}{2}\left(\|A_t - A_{t+1}\|_F^2 + \eta_t^2\|G_t\|_F^2\right) \\
\leq &\frac{1}{2}\left(\|A_* - A_t\|_F^2 - \|A_* - A_{t+1}\|_F^2\right) + \frac{\eta_t^2}{2}\|G_t\|_F^2
\end{aligned}$$

The second inequality follows from Cauchy-Schwartz inequality. Let's consider $L(A_t, U_t) - L(A_*, U_t)$, which is

$$\begin{aligned}
&L(A_t, U_t) - L(A_*, U_t) \\
=&g_t(A_t) + h(A_t) - g_t(A_*) - h(A_*) \\
\leq &\frac{1}{2\eta_t}(\|A_* - A_t\|_F^2 - \|A_* - A_{t+1}\|_F^2) + \frac{\eta_t}{2}\|G_t\|_F^2 \\
&+ \gamma(\phi(A_t) - \phi(A_{t+1}))
\end{aligned}$$

Note that the trailing term $\gamma(\phi(A_t) - \phi(A_{t+1}))$ 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.

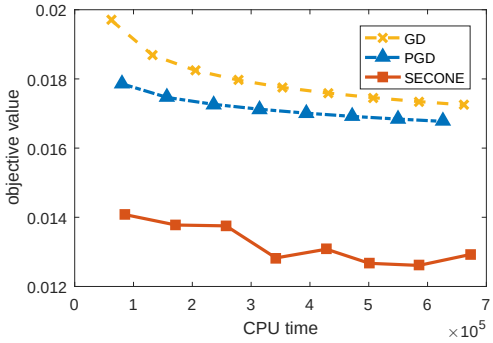

Figure 1: Results of robust low-rank matrix approximation

Table 1: Statistics for matrix approximation

Method	c_1	c_2	T	Total CPU time
SECONE	1e8	1e4	8000	6.73e5
PGD	1e6		80	6.26e5
GD	1e6		90	6.62e5

5 Experiments

We present numerical experiments on real datasets to demonstrate the efficiency of the proposed algorithms.

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]:

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

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 $n = 11269$ instances, each of which has $n = 20302$ features (we filter the features which appear less than 7 times). According to Theorem 1, we set step sizes in Algorithm 1 as $\eta_t = c_1/\sqrt{t}$ and $\tau_t = c_2/\sqrt{t}$, where c_1, c_2 are some constants. The same step size $\eta_t = c_1/\sqrt{t}$ is also used for GD and PGD. We tune the value of c_1 and c_2 in a range of $\{10^{-5}, 10^{-4}, \ldots, 10^{10}\}$ and report the best results based on the objective value.

In Fig. 1, we plot the objective value versus the running time for $\lambda = 1 \times 10^{-6}$. We choose this value of λ because it can produce a low-rank output, and the convergence behavior is insensitive to λ. As can be seen, SECONE decreases much faster than GD and PGD. This is as expected as 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.

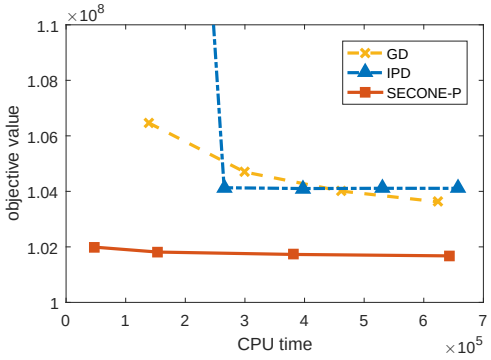

Figure 2: Results of sparse and low-rank link prediction

Table 2: Statistics for link prediction

Method	c_1 or θ	c_2	T	Total CPU time
SECONE-P	1	1e-5	2000	6.43e5
IPD	0.1		50	6.57e5
GD	1		40	6.25e5

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 $m = n = 15000$ users only with the highest degree. We flip 15% of

[2]http://qwone.com/~jason/20Newsgroups/

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 θ of IPD is searched in the range of $\{10^{-2}, 1, \ldots, 100\}$.

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

[Abernethy *et al.*, 2009] Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10:803–826, 2009.

[Argyriou *et al.*, 2008] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[Baccini *et al.*, 1996] A. Baccini, Ph. Besse, and A. de Falguerolles. A l_1-norm PCA and a heuristic approach. In *Proceedings of the International Conference on Ordinal and Symbolic Data Analysis*, pages 359–368, 1996.

[Cai *et al.*, 2010] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

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

[Croux and Filzmoser, 1998] Christophe Croux and Peter Filzmoser. Robust factorization of a data matrix. In *Proceedings in Computational Statistics*, pages 245–250, 1998.

[Duchi and Singer, 2009] John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

[Dudık *et al.*, 2012] Miroslav Dudık, Zaıd Harchaoui, and Jérôme Malick. Lifted coordinate descent for learning with trace-norm regularization. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22, pages 327–336, 2012.

[Hazan, 2008] Elad Hazan. Sparse approximate solutions to semidefinite programs. In *Proceedings of the 8th Latin American Conference on Theoretical Informatics*, pages 306–316, 2008.

[Hsieh and Olsen, 2014] Cho-Jui Hsieh and Peder A Olsen. Nuclear norm minimization via active subspace selection. In *Proceedings of The 31st International Conference on Machine Learning*, pages 575–583, 2014.

[Jaggi *et al.*, 2010] Martin Jaggi, Marek Sulovsk, et al. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th International Conference on Machine Learning*, pages 471–478, 2010.

[Jaggi, 2013] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435, 2013.

[Ji and Ye, 2009] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 457–464, 2009.

[Ke and Kanade, 2005] Qifa Ke and Takeo Kanade. Robust l_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 739–746, 2005.

[Lan, 2012] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133:365–397, 2012.

[Lin *et al.*, 2014] Qihang Lin, Xi Chen, and Javier Peña. A sparsity preserving stochastic gradient methods for sparse regression. *Computational Optimization and Applications*, 58(2):455–482, 2014.

[Mahdavi *et al.*, 2012] Mehrdad Mahdavi, Tianbao Yang, Rong Jin, Shenghuo Zhu, and Jinfeng Yi. Stochastic gradient descent with only one projection. In *Advances in Neural Information Processing Systems (NIPS)*, pages 503–511, 2012.

[Nemirovski *et al.*, 1982] Arkadi Nemirovski, David Borisovich Yudin, and E-R Dawson. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, 1982.

[Nesterov, 2004] Yurii Nesterov. *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied optimization*. Kluwer Academic Publishers, 2004.

[Nesterov, 2013] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[Pong *et al.*, 2010] Ting Kei Pong, Paul Tseng, Shuiwang Ji, and Jieping Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20(6):3465–3489, 2010.

[Rennie and Srebro, 2005] Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 713–719, 2005.

[Richard *et al.*, 2012] Emile Richard, Pierre-Andre Savalle, and Nicolas Vayatis. Estimation of simultaneously sparse and low rank matrices. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1351–1358, 2012.

[Shalev-Shwartz *et al.*, 2011] Shai Shalev-Shwartz, Alon Gonen, and Ohad Shamir. Large-scale convex minimization with a low-rank constraint. In *Proceedings of the 28th International Conference on Machine Learning*, pages 329–336, 2011.

[Signoretto *et al.*, 2013] Marco Signoretto, Volkan Cevher, and JA Suykens. An svd-free approach to a class of structured low rank matrix optimization problems with application to system identification. In *Proceedings of the IEEE Conference on Decision and Control*, 2013.

[Srebro *et al.*, 2005] Nathan Srebro, Jason Rennie, and Tommi S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, pages 1329–1336, 2005.

[Toh and Yun, 2010] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.