Tractable and Scalable Schatten Quasi-Norm Approximations for Rank Minimization

Fanhua Shang **Yuanyuan Liu** **James Cheng**
Department of Computer Science and Engineering, The Chinese University of Hong Kong

Abstract

The Schatten quasi-norm was introduced to bridge the gap between the trace norm and rank function. However, existing algorithms are too slow or even impractical for large-scale problems. Motivated by the equivalence relation between the trace norm and its bilinear spectral penalty, we define two tractable Schatten norms, i.e. the bi-trace and tri-trace norms, and prove that they are in essence the Schatten-1/2 and 1/3 quasi-norms, respectively. By applying the two defined Schatten quasi-norms to various rank minimization problems such as MC and RPCA, we only need to solve much smaller factor matrices. We design two efficient linearized alternating minimization algorithms to solve our problems and establish that each bounded sequence generated by our algorithms converges to a critical point. We also provide the restricted strong convexity (RSC) based and MC error bounds for our algorithms. Our experimental results verified both the efficiency and effectiveness of our algorithms compared with the state-of-the-art methods.

1 Introduction

The rank minimization problem has a wide range of applications in matrix completion (MC) [1], robust principal component analysis (RPCA) [2], low-rank representation [3], multivariate regression [4] and multi-task learning [5]. To efficiently solve these problems, a principled way is to relax the rank function by its convex envelope [6, 7], i.e., the trace norm (also known as the nuclear norm), which also leads to a convex optimization problem. In fact, the trace norm penalty is an ℓ_1-norm regularization of the singular values, and thus it motivates a low-rank solution. However, [8] pointed out that the ℓ_1-norm over-penalizes large entries of vectors, and results in a biased solution. Similar to the ℓ_1-norm case, the trace norm penalty shrinks all singular values equally, which also leads to over-penalize large singular values. In other words, the trace norm may make the solution deviate from the original solution as the ℓ_1-norm does. Compared with the trace norm, although the Schatten-p quasi-norm for $0<p<1$ is non-convex, it gives a closer approximation to the rank function. Therefore, the Schatten-p quasi-norm minimization has attracted a great deal of attention in images recovery [9, 10], collaborative filtering [11] and MRI analysis [12].

[13] and [14] proposed iterative reweighted lease squares (IRLS) algorithms to approximate associated Schatten-p quasi-norm minimization problems. In addition, [10] proposed an iteratively reweighted nuclear norm (IRNN) algorithm to solve non-convex surrogate minimization problems. In some recent work [15, 16, 11, 9, 10], the Schatten-p quasi-norm has been shown to be empirically superior to the trace norm. Moreover, [17] theoretically proved that the Schatten-p quasi-norm minimization with small p requires significantly fewer measurements than the convex trace norm minimization. However, all existing algorithms have to be solved iteratively and involve singular value decomposition (SVD) or eigenvalue decomposition (EVD) in each iteration. Thus they suffer from high computational cost and are even not applicable for large-scale problems [18].

In contrast, the trace norm has a scalable equivalent formulation, the bilinear spectral regularization [19, 7], which has been successfully applied in many large-scale applications, such as collaborative filtering [20, 21]. Since the Schatten-p quasi-norm is equivalent to the ℓ_p quasi-norm on the singular values, it is natural to ask the following question: can we design an equivalent matrix factorization form to some cases of the Schatten-p quasi-norm, e.g., $p=1/2$ or $1/3$?

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

In this paper we first define two tractable Schatten norms, the bi-trace (Bi-tr) and tri-trace (Tri-tr) norms. We then prove that they are in essence the Schatten-1/2 and 1/3 quasi-norms, respectively, for solving whose minimization we only need to perform SVDs on much smaller factor matrices to replace the large matrices in the algorithms mentioned above. Then we design two efficient linearized alternating minimization algorithms with guaranteed convergence to solve our problems. Finally, we provide the sufficient condition for exact recovery, and the restricted strong convexity (RSC) based and MC error bounds.

2 Notations and Background

The Schatten-p norm ($0 < p < \infty$) of a matrix $X \in \mathbb{R}^{m\times n}$ ($m \geq n$) is defined as

$$\|X\|_{S_p} = \left(\sum\nolimits_{i=1}^{n} \sigma_i^p(X)\right)^{1/p},$$

where $\sigma_i(X)$ denotes the i-th singular value of X. For $p \geq 1$ it defines a natural norm, for instance, the Schatten-1 norm is the so-called trace norm, $\|X\|_{\text{tr}}$, whereas for $p < 1$ it defines a quasi-norm. As the non-convex surrogate for the rank function, the Schatten-p quasi-norm with $0<p<1$ is the better approximation of the matrix rank than the trace norm [17] (analogous to the superiority of the ℓ_p quasi-norm to the ℓ_1-norm [14, 22]).

We mainly consider the following Schatten quasi-norm minimization problem to recover a low-rank matrix from a small set of linear observations, $b \in \mathbb{R}^l$,

$$\min_{X\in\mathbb{R}^{m\times n}} \left\{\|X\|_{S_p}^p : \mathcal{A}(X) = b\right\}, \tag{1}$$

where $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^l$ is a linear measurement operator. Alternatively, the Lagrangian version of (1) is

$$\min_{X\in\mathbb{R}^{m\times n}} \left\{\|X\|_{S_p}^p + \frac{1}{\mu} f(\mathcal{A}(X) - b)\right\}, \tag{2}$$

where $\mu > 0$ is a regularization parameter, and the loss function $f(\cdot) : \mathbb{R}^l \to \mathbb{R}$ generally denotes certain measurement for characterizing the loss term $\mathcal{A}(X)-b$ (for instance, $\mathcal{A}$ is the linear projection operator $\mathcal{P}_\Omega$, and $f(\cdot)=\|\cdot\|_2^2$ in MC problems [15, 13, 23, 10]).

The Schatten-p quasi-norm minimization problems (1) and (2) are non-convex, non-smooth and even non-Lipschitz [24]. Therefore, it is crucial to develop efficient algorithms that are specialized to solve some alternative formulations of Schatten-p quasi-norm minimization (1) or (2). So far, only few algorithms, such as IRLS [14, 13] and IRNN [10], have been developed to solve such problems. In addition, since all existing Schatten-p quasi-norm minimization algorithms involve SVD or EVD in each iteration, they suffer from a high computational cost of $O(n^2m)$, which severely limits their applicability to large-scale problems.

3 Tractable Schatten Quasi-Norm Minimization

[19] and [7] pointed out that the trace norm has the following equivalent non-convex formulations.

Lemma 1. *Given a matrix $X \in \mathbb{R}^{m\times n}$ with rank$(X) = r \leq d$, the following holds:*

$$\|X\|_{\text{tr}} = \min_{U\in\mathbb{R}^{m\times d}, V\in\mathbb{R}^{n\times d}: X=UV^T} \|U\|_F\|V\|_F = \min_{U,V:X=UV^T} \frac{\|U\|_F^2 + \|V\|_F^2}{2}.$$

3.1 Bi-Trace Quasi-Norm

Motivated by the equivalence relation between the trace norm and its bilinear spectral regularization form stated in Lemma 1, our bi-trace (Bi-tr) norm is naturally defined as follows [18].

Definition 1. *For any matrix $X \in \mathbb{R}^{m\times n}$ with rank$(X) = r \leq d$, we can factorize it into two much smaller matrices $U \in \mathbb{R}^{m\times d}$ and $V \in \mathbb{R}^{n\times d}$ such that $X = UV^T$. Then the bi-trace norm of X is defined as*

$$\|X\|_{\text{Bi-tr}} := \min_{U,V:X=UV^T} \|U\|_{\text{tr}}\|V\|_{\text{tr}}.$$

In fact, the bi-trace norm defined above is not a real norm, because it is non-convex and does not satisfy the triangle inequality of a norm. Similar to the well-known Schatten-p quasi-norm ($0<p<1$), the bi-trace norm is also a quasi-norm, and their relationship is stated in the following theorem [18].

Theorem 1. *The bi-trace norm $\|\cdot\|_{\text{Bi-tr}}$ is a quasi-norm. Surprisingly, it is also the Schatten-1/2 quasi-norm, i.e.,*

$$\|X\|_{\text{Bi-tr}} = \|X\|_{S_{1/2}},$$

where $\|X\|_{S_{1/2}}$ is the Schatten-1/2 quasi-norm of X.

The proof of Theorem 1 can be found in the Supplementary Materials. Due to such a relationship, it is easy to verify that the bi-trace quasi-norm possesses the following properties.

Property 1. *For any matrix $X \in \mathbb{R}^{m\times n}$ with rank$(X) = r \leq d$, the following holds:*

$$\|X\|_{\text{Bi-tr}} = \min_{U,V:X=UV^T} \|U\|_{\text{tr}}\|V\|_{\text{tr}} = \min_{U,V:X=UV^T} \frac{\|U\|_{\text{tr}}^2+\|V\|_{\text{tr}}^2}{2} = \min_{U,V:X=UV^T} \left(\frac{\|U\|_{\text{tr}}+\|V\|_{\text{tr}}}{2}\right)^2.$$

Property 2. *The bi-trace quasi-norm satisfies the following properties:*

1. $\|X\|_{\text{Bi-tr}} \geq 0$, *with equality iff* $X = 0$.
2. $\|X\|_{\text{Bi-tr}}$ *is unitarily invariant, i.e.,* $\|X\|_{\text{Bi-tr}} = \|PXQ^T\|_{\text{Bi-tr}}$, *where* $P \in \mathbb{R}^{m\times m}$ *and* $Q \in \mathbb{R}^{n\times n}$ *have orthonormal columns.*

3.2 Tri-Trace Quasi-Norm

Similar to the definition of the bi-trace quasi-norm, our tri-trace (Tri-tr) norm is naturally defined as follows.

Definition 2. *For any matrix $X \in \mathbb{R}^{m\times n}$ with rank$(X) = r \le d$, we can factorize it into three much smaller matrices $U \in \mathbb{R}^{m\times d}$, $V \in \mathbb{R}^{d\times d}$ and $W \in \mathbb{R}^{n\times d}$ such that $X = UVW^T$. Then the tri-trace norm of X is defined as*

$$\|X\|_{\text{Tri-tr}} := \min_{U,V,W:X=UVW^T} \|U\|_{\text{tr}}\|V\|_{\text{tr}}\|W\|_{\text{tr}}.$$

Like the bi-trace quasi-norm, the tri-trace norm is also a quasi-norm, as stated in the following theorem.

Theorem 2. *The tri-trace norm $\|\cdot\|_{\text{Tri-tr}}$ is a quasi-norm. In addition, it is also the Schatten-1/3 quasi-norm, i.e.,*

$$\|X\|_{\text{Tri-tr}} = \|X\|_{S_{1/3}}.$$

The proof of Theorem 2 is very similar to that of Theorem 1 and is thus omitted. According to Theorem 2, it is easy to verify that the tri-trace quasi-norm possesses the following properties.

Property 3. *For any matrix $X \in \mathbb{R}^{m\times n}$ with rank$(X) = r \le d$, the following holds:*

$$\|X\|_{\text{Tri-tr}} = \min_{X=UVW^T} \left(\frac{\|U\|_{\text{tr}}+\|V\|_{\text{tr}}+\|W\|_{\text{tr}}}{3}\right)^3$$
$$= \min_{X=UVW^T} \|U\|_{\text{tr}}\|V\|_{\text{tr}}\|W\|_{\text{tr}} = \min_{X=UVW^T} \frac{\|U\|_{\text{tr}}^3+\|V\|_{\text{tr}}^3+\|W\|_{\text{tr}}^3}{3}.$$

Property 4. *The tri-trace quasi-norm satisfies the following properties:*

1. $\|X\|_{\text{Tri-tr}} \ge 0$, *with equality iff* $X = 0$.
2. $\|X\|_{\text{Tri-tr}}$ *is unitarily invariant, i.e.,* $\|X\|_{\text{Tri-tr}} = \|PXQ^T\|_{\text{Tri-tr}}$, *where* $P \in \mathbb{R}^{m\times m}$ *and* $Q \in \mathbb{R}^{n\times n}$ *have orthonormal columns.*

The following relationship between the trace-norm and Frobenius norm is well known: $\|X\|_F \le \|X\|_{\text{tr}} \le \sqrt{r}\|X\|_F$. Similarly, the analogous bounds hold for the bi-trace and tri-trace quasi-norms, as stated in the following property.

Property 5. *For any matrix $X \in \mathbb{R}^{m\times n}$ with rank$(X) = r$, the following inequalities hold:*

$$\|X\|_{\text{tr}} \le \|X\|_{\text{Bi-tr}} \le r\|X\|_{\text{tr}},$$
$$\|X\|_{\text{tr}} \le \|X\|_{\text{Bi-tr}} \le \|X\|_{\text{Tri-tr}} \le r^2\|X\|_{\text{tr}}.$$

Proof. The proof of this property involves the following properties of the ℓ_p quasi-norm. For any vectors x and y in $\mathbb{R}^n$ and $0 < p_2 \le p_1 \le 1$, we have

$$\|x\|_1 \le \|x\|_{p_1}, \quad \|x\|_{p_1} \le \|x\|_{p_2} \le n^{1/p_2-1/p_1}\|x\|_{p_1}.$$

Suppose $X \in \mathbb{R}^{m\times n}$ is of rank r, and denote its skinny SVD by $X = U\Sigma V^T$. By Theorems 1 and 2, and the properties of the ℓ_p quasi-norm, we have

$$\|X\|_{\text{tr}} = \|\text{diag}(\Sigma)\|_1 \le \|\text{diag}(\Sigma)\|_{\frac{1}{2}} = \|X\|_{\text{Bi-tr}} \le r\|X\|_{\text{tr}},$$
$$\|X\|_{\text{tr}} = \|\text{diag}(\Sigma)\|_1 \le \|\text{diag}(\Sigma)\|_{\frac{1}{3}} = \|X\|_{\text{Tri-tr}} \le r^2\|X\|_{\text{tr}}.$$

In addition,

$$\|X\|_{\text{Bi-tr}} = \|\text{diag}(\Sigma)\|_{\frac{1}{2}} \le \|\text{diag}(\Sigma)\|_{\frac{1}{3}} = \|X\|_{\text{Tri-tr}}. \qquad \square$$

It is easy to see that Property 5 in turn implies that any low bi-trace or tri-trace quasi-norm approximation is also a low trace norm approximation.

3.3 Problem Formulations

Bounding the Schatten quasi-norm of X in (1) by the bi-trace or tri-trace quasi-norm defined above, the noiseless low-rank structured matrix factorization problem is given by

$$\min_{U,V} \left\{\mathcal{R}(U,V) = (\|U\|_{\text{tr}}+\|V\|_{\text{tr}})/2 \colon \mathcal{A}(UV^T) = b\right\}, \quad (3)$$

where $\mathcal{R}(\cdot)$ can also denote $(\|U\|_{\text{tr}}+\|V\|_{\text{tr}}+\|W\|_{\text{tr}})/3$, and $\mathcal{A}(UV^T)$ is replaced by $\mathcal{A}(UVW^T)$. In addition, (3) has the following Lagrangian forms,

$$F(U,V) := \min_{U,V} \left\{\frac{\|U\|_{\text{tr}}+\|V\|_{\text{tr}}}{2} + \frac{f(\mathcal{A}(UV^T)-b)}{\mu}\right\}, \quad (4)$$

$$\min_{U,V,W} \left\{\frac{\|U\|_{\text{tr}}+\|V\|_{\text{tr}}+\|W\|_{\text{tr}}}{3} + \frac{f(\mathcal{A}(UVW^T)-b)}{\mu}\right\}. \quad (5)$$

The formulations (3), (4) and (5) can address a wide range of problems, such as MC [13, 10], RPCA [2, 25, 26] ($\mathcal{A}$ is the identity operator, and $f(\cdot) = \|\cdot\|_1$ or $\|\cdot\|_p$ $(0 < p < 1)$), and low-rank representation [3] or multivariate regression [4] ($\mathcal{A}(X) = AX$ with A being a given matrix, and $f(\cdot) = \|\cdot\|_{2,1}$ or $\|\cdot\|_F^2$). In addition, $f(\cdot)$ may be also chosen as the Hinge loss in [19] or the structured atomic norms in [27].

4 Optimization Algorithms

In this section, we mainly propose two efficient algorithms to solve the challenging bi-trace quasi-norm regularized problem (4) with a smooth or non-smooth loss function, respectively. In other words, if $f(\cdot)$ is a smooth loss function, e.g., $f(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, we employ the proximal alternating linearized minimization (PALM) method as in [28] to solve (4). In contrast, to solve efficiently (4) with a non-smooth loss function, e.g., $f(\cdot) = \|\cdot\|_1$, we need to introduce an auxiliary variable e and obtain the following equivalent form:

$$\min_{U,V,e} \left\{\frac{\|U\|_{\text{tr}}+\|V\|_{\text{tr}}}{2} + \frac{f(e)}{\mu} \colon e = \mathcal{A}(UV^T)-b\right\}. \quad (6)$$

4.1 LADM Algorithm

To avoid introducing more auxiliary variables, inspired by [29], we propose a linearized alternating direction method (LADM) to solve (6), whose augmented Lagrangian function is given by

$$\mathcal{L}(U,V,e,\lambda,\beta)=\frac{1}{2}(\|U\|_{\rm tr}+\|V\|_{\rm tr})+\frac{f(e)}{\mu}+\langle\lambda,\mathcal{A}(UV^T)-b-e\rangle+(\beta/2)\|\mathcal{A}(UV^T)-b-e\|_2^2,$$

where $\lambda\in\mathbb{R}^l$ is the Lagrange multiplier, $\langle\cdot,\cdot\rangle$ denotes the inner product, and $\beta>0$ is a penalty parameter. By applying the classical augmented Lagrangian method to (6), we obtain the following iterative scheme:

$$U_{k+1}=\arg\min_U \frac{\|U\|_{\rm tr}}{2}+\frac{\beta_k}{2}\|\mathcal{A}(UV_k^T)-e_k-\widetilde{b}_k\|_2^2, \quad (7a)$$

$$V_{k+1}=\arg\min_V \frac{\|V\|_{\rm tr}}{2}+\frac{\beta_k}{2}\|\mathcal{A}(U_{k+1}V^T)-e_k-\widetilde{b}_k\|_2^2, \quad (7b)$$

$$e_{k+1}=\arg\min_e \frac{f(e)}{\mu}+\frac{\beta_k}{2}\|\mathcal{A}(U_{k+1}V_{k+1}^T)-e-\widetilde{b}_k\|_2^2, \quad (7c)$$

$$\lambda_{k+1}=\lambda_k+\beta_k(\mathcal{A}(U_{k+1}V_{k+1}^T)-b-e_{k+1}), \quad (7d)$$

where $\widetilde{b}_k=b-\lambda_k/\beta_k$. In many machine learning problems [15, 3, 4], $\mathcal{A}$ is not identity, e.g., the operator $\mathcal{P}_\Omega$. Due to the presence of V_k and U_{k+1}, thus we usually need to introduce some auxiliary variables to achieve closed-form solutions to (7a) and (7b). To avoid introducing additional auxiliary variables, we propose the following linearization technique for (7a) and (7b).

4.1.1 Updating U_{k+1} and V_{k+1}

Let $\varphi_k(U):=\|\mathcal{A}(UV_k^T)-b-e_k+\lambda_k/\beta_k\|_2^2/2$, then we can know that the gradient of $\varphi_k(U)$ is Lipschitz continuous with the constant t_k^φ, i.e., $\|\nabla\varphi_k(U_1)-\nabla\varphi_k(U_2)\|_F\le t_k^\varphi\|U_1-U_2\|_F$ for any $U_1,U_2\in\mathbb{R}^{m\times d}$. By linearizing $\varphi_k(U)$ at U_k and adding a proximal term, we have

$$\widehat{\varphi}_k(U,U_k)=\varphi_k(U_k)+\langle\nabla\varphi_k(U_k),U-U_k\rangle+\frac{t_k^\varphi}{2}\|U-U_k\|_F^2. \quad (8)$$

Therefore, we have

$$\begin{aligned}U_{k+1}&=\arg\min_U \frac{1}{2}\|U\|_{\rm tr}+\beta_k\widehat{\varphi}_k(U,U_k)\\&=\arg\min_U \frac{1}{2}\|U\|_{\rm tr}+\frac{\beta_k t_k^\varphi}{2}\|U-U_k+\frac{\nabla\varphi_k(U_k)}{t_k^\varphi}\|_F^2.\end{aligned} \quad (9)$$

Similarly, we have

$$V_{k+1}=\arg\min_V \frac{1}{2}\|V\|_{\rm tr}+\frac{\beta_k t_k^\psi}{2}\|V-V_k+\frac{\nabla\psi_k(V_k)}{t_k^\psi}\|_F^2, \quad (10)$$

where $\psi_k(V):=\|\mathcal{A}(U_{k+1}V^T)-b-e_k+\lambda_k/\beta_k\|_2^2/2$ with the Lipschitz constant t_k^ψ. Using the so-called matrix shrinkage operator [30], we can obtain a closed-form solution to (9) and (10), respectively. Additionally, if $f(\cdot)=\|\cdot\|_1$, the optimal solution to (7c) can be obtained by the well-known soft-thresholding operator [31].

Algorithm 1 LADM for (4) with non-smooth loss

Input: b, the given rank d and μ.
Initialize: $\beta_0=10^{-4}$, $\beta_{\max}=10^{20}$ and $\varepsilon=10^{-4}$.
1: **while** not converged **do**
2: Update t_k^φ, U_{k+1}, t_k^ψ, and V_{k+1} by (11), (9), (11), and (10), respectively.
3: Update e_{k+1} and λ_{k+1} by (7c) and (7d).
4: Update β_{k+1} by $\beta_{k+1}=\min(\rho\beta_k,\beta_{\max})$.
5: Check the convergence condition, $\|\mathcal{A}(U_{k+1}V_{k+1}^T)-b-e_{k+1}\|_2<\varepsilon$.
6: **end while**
Output: U_{k+1}, V_{k+1}, e_{k+1}.

4.1.2 Computing Step Sizes

There are two step sizes, i.e., the Lipschitz constants t_k^φ in (9) and t_k^ψ in (10), need to be set during the iteration.

$$\begin{aligned}&\|\nabla\varphi_k(U_1)-\nabla\varphi_k(U_2)\|_F=\|\mathcal{A}^*\{\mathcal{A}[(U_1-U_2)V_k^T]\}V_k\|_F\\&\le\|\mathcal{A}^*\mathcal{A}\|_2\|V_k^TV_k\|_2\|U_1-U_2\|_F,\\&\|\nabla\psi_k(V_1)-\nabla\psi_k(V_2)\|_F=\|U_{k+1}^T\mathcal{A}^*\{\mathcal{A}[U_{k+1}(V_1-V_2)^T]\}\|_F\\&\le\|\mathcal{A}^*\mathcal{A}\|_2\|U_{k+1}^TU_{k+1}\|_2\|V_1-V_2\|_F,\end{aligned}$$

where $\mathcal{A}^*$ denotes the adjoint operator of $\mathcal{A}$. Thus, both step sizes are defined in the following way:

$$\begin{cases}t_k^\varphi\ge\|\mathcal{A}^*\mathcal{A}\|_2\|V_k^TV_k\|_2,\\t_k^\psi\ge\|\mathcal{A}^*\mathcal{A}\|_2\|U_{k+1}^TU_{k+1}\|_2.\end{cases} \quad (11)$$

Based on the description above, we develop an efficient LADM algorithm to solve the Bi-tr quasi-norm regularized problem (4) with a non-smooth loss function (e.g., RPCA problems), as outlined in **Algorithm** 1. To further accelerate the convergence of the algorithm, the penalty parameter β is adaptively updated by the strategy as in [32], as well as ρ. Moreover, Algorithm 1 can be used to solve the noiseless problem (3) and also extended to solve the Tri-tr quasi-norm regularized problem (5) with a non-smooth loss function.

4.2 PALM Algorithm

By using the similar linearization technique in (9) and (10), we design an efficient PALM algorithm to solve (4) with a smooth loss function, e.g., MC problems. Specifically, by linearizing the smooth loss function $\varphi_k(U):=\|\mathcal{A}(UV_k^T)-b\|_2^2/2$ at U_k and adding a proximal term, we have the following approximation:

$$\begin{aligned}&U_{k+1}\\&=\arg\min_U \frac{\|U\|_{\rm tr}}{2}+\langle\frac{\nabla\varphi_k(U_k)}{\mu},U-U_k\rangle+\frac{t_k^\varphi}{2\mu}\|U-U_k\|_F^2\\&=\arg\min_U \frac{\|U\|_{\rm tr}}{2}+\frac{t_k^\varphi}{2\mu}\|U-U_k+\frac{\nabla\varphi_k(U_k)}{t_k^\varphi}\|_F^2,\end{aligned} \quad (12)$$

where $\nabla\varphi_k(U_k) = \mathcal{A}^*[\mathcal{A}(U_kV_k^T) - b]V_k$. Similarly,

$$V_{k+1} = \arg\min_V \frac{\|V\|_{\rm tr}}{2} + \frac{t_k^\psi}{2\mu}\|V - V_k + \frac{\nabla\psi_k(V_k)}{t_k^\psi}\|_F^2, \quad (13)$$

where $\nabla\psi_k(V_k) = \{\mathcal{A}^*[\mathcal{A}(U_{k+1}V_k^T) - b]\}^T U_{k+1}$.

4.3 Convergence Analysis

In the following, we provide the convergence analysis of our algorithms. First, we analyze the convergence of our LADM algorithm for solving (4) with a non-smooth loss function, e.g., $f(\cdot) = \|\cdot\|_1$.

Theorem 3. *Let $\{(U_k, V_k, e_k)\}$ be a sequence generated by Algorithm 1, then we have*

1. *$\{(U_k, V_k, e_k)\}$ are all Cauchy sequences;*
2. *If $\lim_{k\to\infty} \|\lambda_{k+1} - \lambda_k\|_2 = 0$, then the accumulation point of the sequence $\{(U_k, V_k, e_k)\}$ satisfies the KKT conditions for (6).*

The proof of Theorem 3 is provided in the Supplementary Materials. From Theorem 3, we can know that under mild conditions each sequence generated by our LADM algorithm converges to a critical point, similar to the LADM algorithms for solving convex problems as in [32].

Moreover, we provide the global convergence of our PALM algorithm for solving (4) with a smooth loss function, e.g., $f(\cdot) = \frac{1}{2}\|\cdot\|_2^2$.

Theorem 4. *Let $\{(U_k, V_k)\}$ be a sequence generated by our PALM algorithm, then it is a Cauchy sequence and converges to a critical point of (4) with the squared loss, $\|\cdot\|_2^2$.*

The proof of Theorem 4 can be found in the Supplementary Materials. Theorem 4 shows the global convergence of our PALM algorithm. We emphasize that, different from the general subsequence convergence property, the global convergence property is given by $(U_k, V_k) \to (\widehat{U}, \widehat{V})$ as the number of iteration $k \to +\infty$, where $(\widehat{U}, \widehat{V})$ is a critical point of (4). On the contrary, existing algorithms for solving non-convex and non-smooth problems, such as [14] and [10], have only subsequence convergence property.

By the Kurdyka-Łojasiewicz (KL) property (for more details, see [28]) and Theorem 2 in [33], our PALM algorithm has the following convergence rate:

Theorem 5. *The sequence $\{(U_k, V_k)\}$ generated by our PALM algorithm converges to a critical point $(\widehat{U}, \widehat{V})$ of F with $f(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, which satisfies the KL property at each point of $\operatorname{dom}\partial F$ with $\phi(s) = cs^{1-\theta}$ for $c > 0$ and $\theta \in [0, 1)$. We have*

- *If $\theta = 0$, $\{(U_k, V_k)\}$ converges to $(\widehat{U}, \widehat{V})$ in finite steps;*
- *If $\theta \in (0, 1/2]$, then $\exists C > 0$ and $\gamma \in [0, 1)$ such that $\|[U_k^T, V_k^T] - [\widehat{U}^T, \widehat{V}^T]\|_F \le C\gamma^k$;*
- *If $\theta \in (1/2, 1)$, then $\exists C > 0$ such that $\|[U_k^T, V_k^T] - [\widehat{U}^T, \widehat{V}^T]\|_F \le Ck^{-\frac{1-\theta}{2\theta-1}}$.*

Theorem 5 shows us the convergence rate of our PALM algorithm for solving the non-convex and non-smooth bi-trace quasi-norm problem (4) with the squared loss $\|\cdot\|_2^2$. Moreover, we can see that the convergence rate of our PALM algorithm is at least sub-linear.

5 Recovery Guarantees

We provide theoretical guarantees for our Bi-tr quasi-norm minimization in recovering low-rank matrices from small sets of linear observations. By using the null-space property (NSP), we first provide a sufficient condition for exact recovery of low-rank matrices. We then establish the restricted strong convexity (RSC) condition based and MC error bounds.

5.1 Null Space Property

The wide use of NSP for recovering sparse vectors and low-rank matrices can be found in [22, 34]. We give the sufficient and necessary condition for exact recovery via our bi-trace quasi-norm model (3) that improves the NSP condition for the Schatten-p quasi-norm in [34]. Let $U_\star = L_{(d)}\Sigma_{(d)}^{1/2} \in \mathbb{R}^{m\times d}$, $V_\star = R_{(d)}\Sigma_{(d)}^{1/2} \in \mathbb{R}^{n\times d}$ and $\Sigma_{(d)} = \operatorname{diag}([\sigma_1(X_0), \ldots, \sigma_r(X_0), 0, \ldots, 0]) \in \mathbb{R}^{d\times d}$, where $L_{(d)}$ and $R_{(d)}$ denote the matrices consisting the top d left and right singular vectors of the true matrix X_0 (which satisfies $\mathcal{A}(X_0) = b$) with rank at most r ($r \le d$). $\mathcal{N}(\mathcal{A}) := \{X \in \mathbb{R}^{m\times n} : \mathcal{A}(X) = \mathbf{0}\}$ denotes the null space of the linear operator $\mathcal{A}$. Then we have the following theorem, the proof of which is provided in the Supplementary Materials.

Theorem 6. *X_0 can be uniquely recovered by (3), if and only if for any $Z = U_\star W_2^T + W_1 V_\star^T + W_1 W_2^T \in \mathcal{N}(\mathcal{A}) \setminus \{\mathbf{0}\}$, where $W_1 \in \mathbb{R}^{m\times d}$, $W_2 \in \mathbb{R}^{n\times d}$, we have*

$$\sum_{i=1}^{r}\sigma_i(W_1) + \sigma_i(W_2) < \sum_{i=r+1}^{d}\sigma_i(W_1) + \sigma_i(W_2). \quad (14)$$

Remark: Since $\Gamma \subset \mathcal{N}(\mathcal{A})$, where $\Gamma = \{Z | Z = U_\star W_2^T + W_1 V_\star^T + W_1 W_2^T,\ Z \in \mathcal{N}(\mathcal{A}) \setminus \{\mathbf{0}\}\}$, the sufficient condition in Theorem 6 is weaker than the corresponding sufficient condition for the Schatten-p quasi-norm in [34].

5.2 RSC based Error Bound

Unlike most of existing recovery guarantees as in [17, 34], we do not impose the restricted isometry property (RIP) on the general operator $\mathcal{A}$, rather, we require the operator $\mathcal{A}$ to satisfy a weaker and more general condition known as restricted strong convexity (RSC) [35], as shown in the following.

Assumption 1 (RSC). *We suppose that there is a positive constant $\kappa(\mathcal{A})$ such that the general operator $\mathcal{A}:\mathbb{R}^{m\times n}\to\mathbb{R}^l$ satisfies the following inequality*

$$\frac{1}{\sqrt{l}}\|\mathcal{A}(\Delta)\|_2 \geq \kappa(\mathcal{A})\|\Delta\|_F$$

for all $\Delta\in\mathbb{R}^{m\times n}$.

We mainly provide the RSC based error bound for robust recovery via our bi-trace quasi-norm algorithm with noisy measurements. To our knowledge, our recovery guarantee analysis is the first one for solutions generated by Schatten quasi-norm algorithms, not for the global optima[1] of (4) as in [36, 17, 34].

Theorem 7. *Assume $X_0\in\mathbb{R}^{m\times n}$ is a true matrix and the corrupted measurements $\mathcal{A}(X_0)+e=b$, where e is noise with $\|e\|_2\leq\epsilon$. Let $(\hat{U},\hat{V})$ be a critical point of (4) with the squared loss $\|\cdot\|_2^2$, and suppose the operator $\mathcal{A}$ satisfies the RSC condition with a constant $\kappa(\mathcal{A})$. Then*

$$\frac{\|X_0-\hat{U}\hat{V}^T\|_F}{\sqrt{mn}}\leq\frac{\epsilon}{\kappa(\mathcal{A})\sqrt{lmn}}+\frac{\mu\sqrt{d}}{2C_1\kappa(\mathcal{A})\sqrt{lmn}},$$

where $C_1=\frac{\|\mathcal{A}^*(b-\mathcal{A}(\hat{U}\hat{V}^T))\hat{V}\|_F}{\|b-\mathcal{A}(\hat{U}\hat{V}^T)\|_2}$.

The proof of Theorem 7 and the analysis of lower-boundedness of C_1 is provided in the Supplementary Materials.

5.3 Error Bound on Matrix Completion

Although the MC problem is a practically important application of (4), the projection operator $\mathcal{P}_\Omega$ in (15) does not satisfy the standard RIP and RSC conditions in general [1, 37, 38]. Therefore, we also need to provide the recovery guarantee for performance of our Bi-tr quasi-norm minimization for solving the following MC problem.

$$\min_{U,V}\left\{\frac{\|U\|_{\rm tr}+\|V\|_{\rm tr}}{2}+\frac{1}{2\mu}\|\mathcal{P}_\Omega(UV^T)-\mathcal{P}_\Omega(D)\|_F^2\right\}. \quad (15)$$

Without loss of generality, assume that the observed matrix $D\in\mathbb{R}^{m\times n}$ can be decomposed as a true matrix X_0 of rank $r\leq d$ and a random Gaussian noise E, i.e., $D=X_0+E$. We give the following recovery guarantee for our Bi-tr quasi-norm minimization (15).

Theorem 8. *Let $(\widehat{U},\widehat{V})$ be a critical point of the problem (15) with given rank d, and $m\geq n$. Then there exists an absolute constant C_2, such that with probability at least $1-2\exp(-m)$,*

$$\frac{\|X_0-\widehat{U}\widehat{V}^T\|_F}{\sqrt{mn}}\leq\frac{\|E\|_F}{\sqrt{mn}}+C_2\delta\left(\frac{md\log(m)}{|\Omega|}\right)^{1/4}+\frac{\mu\sqrt{d}}{2C_3\sqrt{|\Omega|}},$$

where $\delta=\max_{i,j}|D_{i,j}|$ and $C_3=\frac{\|\mathcal{P}_\Omega(D-\hat{U}\hat{V}^T)\hat{V}\|_F}{\|\mathcal{P}_\Omega(D-\hat{U}\hat{V}^T)\|_F}$.

The proof of Theorem 8 and the analysis of lower-boundedness of C_3 can be found in the Supplementary Materials. When the samples size $|\Omega|\gg md\log(m)$, the second and third terms diminish, and the recovery error is essentially bounded by the "average" magnitude of entries of noise E. In other words, only $O(md\log(m))$ observed entries are needed, significantly lower than $O(mr\log^2(m))$ in standard matrix completion theories [37, 39, 7], which will be confirmed by the following experimental results.

6 Experimental Results

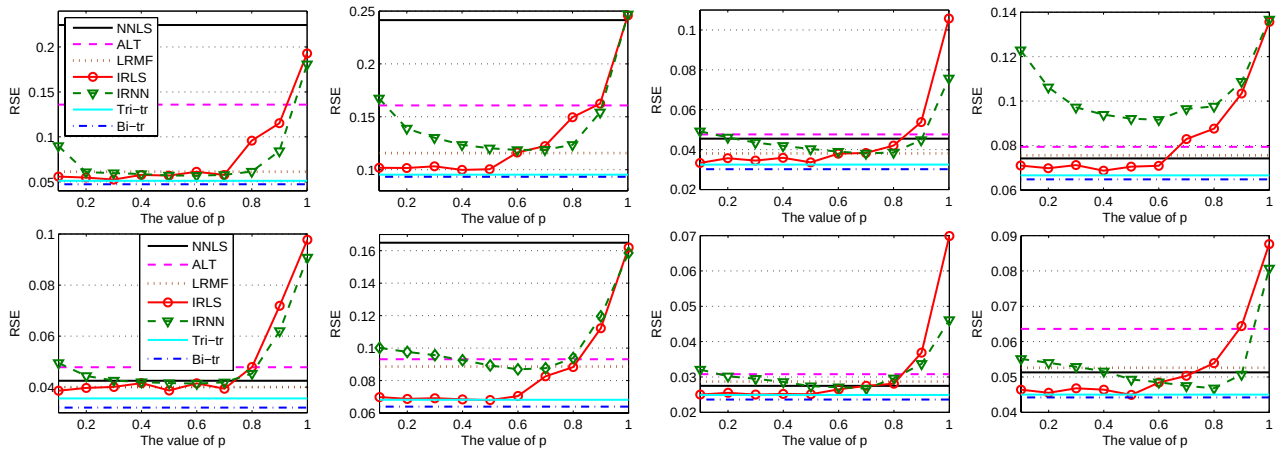
We evaluate both the effectiveness and efficiency of our methods (i.e., the Bi-tr and Tri-tr methods) for solving MC and RPCA problems, such as collaborative filtering and text separation. All experiments were conducted on an Intel Xeon E7-4830V2 2.20GHz CPU with 64G RAM.

6.1 Synthetic Matrix Completion

The synthetic matrices $X_0\in\mathbb{R}^{m\times n}$ with rank r are generated randomly by the following procedure: the entries of both random matrices $P\in\mathbb{R}^{m\times r}$ and $Q\in\mathbb{R}^{n\times r}$ are first generated as independent and identically distributed (i.i.d.) numbers, and then $X_0=PQ^T$ is assembled. The experiments are conducted on random matrices with different noise factors, $nf=0.1$ or 0.2, where the observed subset is corrupted by i.i.d. standard Gaussian random variables as in [18]. In both cases, the sampling ratio (SR) is set to 20% or 30%. We use the relative standard error (RSE := $\|X-X_0\|_F/\|X_0\|_F$) as the evaluation measure, where X denotes the recovered matrix.

We compare our methods with two trace norm solvers: NNLS [40] and ALT [4], one bilinear spectral regularization method, LRMF [20], and two Schatten-p norm methods, IRLS [14] and IRNN [10]. The recovery results of IRLS and IRNN ($p\in\{0.1,0.2,\ldots,1\}$) on noisy random matrices are shown in Figure 1, from which we can observe that as a scalable alternative to trace norm regularization, LRMF with relatively small ranks often obtains more accurate solutions than its trace norm counterparts, i.e., NNLS and ALT. If p is chosen from the range of $\{0.3,0.4,0.5,0.6\}$, IRLS and IRNN have similar performance, and usually outperform NNLS, ALT and LRMF in terms of RSE, otherwise they sometimes perform much worse than the latter three methods, especially $p=1$. This means that

[1] It is well known that the Schatten-p quasi-norm ($0<p<1$) problems in [15, 11, 14, 10, 9] are non-convex, non-smooth and non-Lipschitz [24]. The recovery guarantees in [36, 17, 34] are naturally based on the global optimal solution of associated models.



(a) 20% SR and $nf=0.1$ (b) 20% SR and $nf=0.2$ (c) 30% SR and $nf=0.1$ (d) 30% SR and $nf=0.2$

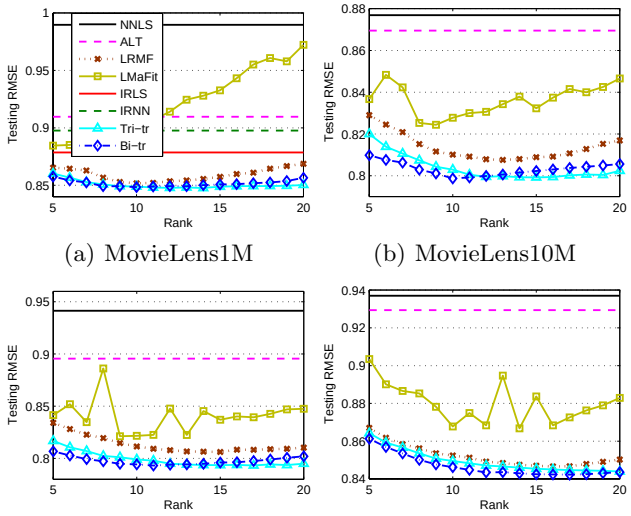
Figure 1: The recovery accuracy of NNLS, ALT, LRMF, IRLS, IRNN, and our Tri-tr and Bi-tr methods on noisy random matrices of size 100×100 (the first row) or 200×200 (the second row).

both our methods (which are in essence the Schatten-1/2 and 1/3 quasi-norm algorithms) should perform better than them. As expected, the RSE results of both our methods under all of these settings are consistently much better than those of the other approaches. This clearly justifies the usefulness of our Bi-tr and Tri-tr quasi-norm penalties. Moreover, the running time of all these methods on random matrices with different sizes is provided in the Supplementary Materials, which shows that our methods are much faster than the other methods. This confirms that both our methods have very good scalability and can address large-scale problems.

6.2 Collaborative Filtering

We test our methods on the real-world recommendation system datasets: MovieLens1M, MovieLens10M and MovieLens20M[2], and Netflix [41]. We randomly choose 90% as the training set and the remaining as the testing set, and the experimental results are reported over 10 independent runs. Besides those methods used above, we also compare our methods to one of the fastest methods, LMaFit [42], and use the root mean squared error (RMSE) as evaluation measure.

The testing RMSE of all those methods on the four datasets is reported in Figure 2, where the rank varies from 5 to 20 (the running time of all methods are provided in Supplementary Materials). From all these results, we can observe that for these fixed ranks, the matrix factorization methods including LMaFit, LRMF and our methods significantly perform better than the trace norm solvers including NNLS and ALT in terms of RMSE, especially on the three larger datasets, as



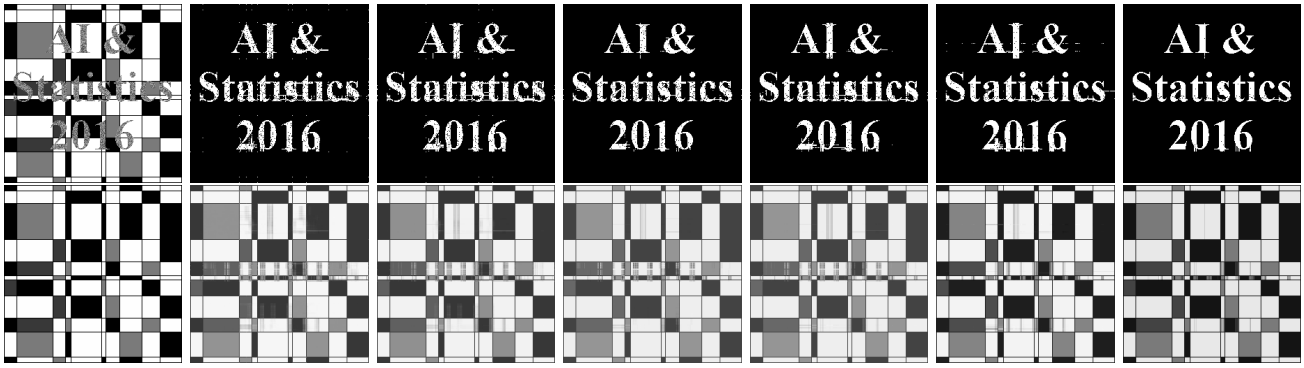
(a) MovieLens1M (b) MovieLens10M

(c) MovieLens20M (d) Netflix

Figure 2: Evolution of the testing RMSE of different methods with ranks varying from 5 to 20.

shown in Figures 2(b)-(d). In most cases, the sophisticated matrix factorization based approaches outperform LMaFit as a baseline method without any regularization term. This suggests that those regularized models can alleviate the over-fitting problem of matrix factorization. The testing RMSE of both our methods varies only slightly when the number of the given rank increases, while that of the other matrix factorization methods changes dramatically. This further means that our methods perform much more robust than them in terms of the given ranks. More importantly, both our methods under all of the rank settings consistently outperform the other methods in terms of prediction accuracy. This confirms that our Bi-tr or Tri-tr quasi-norm regularized models can provide a

[2] http://www.grouplens.org/node/73



(a) Images (b) PCP (c) LRMF+ℓ_1 (d) S_p+ℓ_p (e) Tri-tr+ℓ_1 (f) Bi-tr+ℓ_1 (g) Bi-tr+$\ell_{1/2}$

Figure 3: Text separation results. The first and second rows mainly show the detected texts and the recovered background images: (a) Input image (upper) and original image (bottom); (b) AUC: 0.8939, RSE: 0.1494; (c) AUC: 0.9058, RSE: 0.1406; (d) AUC: 0.9425, RSE: 0.1342; (e) AUC: 0.9356, RSE: 0.1320; (f) AUC: 0.9389, RSE: 0.1173; (g) AUC: **0.9731**, RSE: **0.0853**.

good estimation of a low-rank matrix. Note that IRLS and IRNN could not run on the three larger datasets due to runtime exceptions. Moreover, our methods are much faster than LRMF, NNLS, ALT, IRLS and IRNN on all these datasets, and are comparable in speed with LMaFit. This shows that our methods have very good scalability and can solve large-scale problems.

6.3 Text Separation

We conducted an experiment on artificially generated data to separate some text from an image. The ground-truth image is of size 256×256 with rank equal to 10. Figure 3(a) shows the input image together with the original image. The input data are generated by setting 10% of the randomly selected pixels as missing entries. We compare our Bi-tr+ℓ_1, Tri-tr+ℓ_1 and Bi-tr+$\ell_{1/2}$ methods (see Supplementary Materials for the details) to three state-of-the-art methods, including PCP [2], LRMF+ℓ_1 [43] and S_p+ℓ_p [11] with $0<p\leq 1$. For fairness, we set the rank of all methods to 15, and $\varepsilon=10^{-4}$ for all these algorithms.

The results of different methods are shown in Figure 3, where the text detection accuracy (the score Area Under the receiver operating characteristic Curve, AUC) and the RSE of low-rank component recovery are reported. Note that we present the best performance results of S_p+ℓ_p with all choices of p in $\{0.1, 0.2, \ldots, 0.9\}$. For both low-rank component recovery and text separation, our Bi-tr+$\ell_{1/2}$ method is significantly better than the other methods, not only visually but also quantitatively. In addition, our Bi-tr+ℓ_1 and Tri-tr+ℓ_1 methods have very similar performance to the S_p+ℓ_p method, and all these three methods outperform PCP and LRMF+ℓ_1 in terms of AUC and RSE. Moreover, the running time of PCP, LRMF+ℓ_1, S_p+ℓ_p, Tri-tr+ℓ_1, Bi-tr+ℓ_1 and Bi-tr+$\ell_{1/2}$ is 31.57sec, 6.91sec, 163.65sec, 0.96sec, 0.57sec and 1.62sec, respectively. In other words, our three methods are at least 7, 12 and 4 times faster than the other methods, respectively. This is a very impressive result as our three methods are nearly 170, 290 or 100 times faster than the most related S_p+ℓ_p method, which further confirms that our methods have good scalability.

7 Conclusions

In this paper, we defined two tractable Schatten quasi-norm formulations, and then proved that they are in essence the Schatten-1/2 and 1/3 quasi-norms, respectively. By applying the two defined quasi-norms to various rank minimization problems, such as MC and RPCA, we achieved some challenging non-smooth and non-convex problems. Then we designed two classes of efficient PALM and LADM algorithms to solve our problems with smooth and non-smooth loss functions, respectively. Finally, we established that each bounded sequence generated by our algorithms converges to a critical point, and also provided the recovery performance guarantees for our algorithms. Experiments on real-world data sets showed that our methods outperform the state-of-the-art methods in terms of both efficiency and effectiveness. For future work, we are interested in analyzing the recovery bound for our algorithms to solve the Bi-tr or Tri-tr quasi-norm regularized problems with non-smooth loss functions.

Acknowledgements

We thank the reviewers for their valuable comments. The authors are supported by the Hong Kong GRF 2150851. The project is funded by Research Committee of CUHK.

References

[1] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[2] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):1–37, 2011.

[3] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[4] C. Hsieh and P. A. Olsen. Nuclear norm minimization via active subspace selection. In *ICML*, pages 575–583, 2014.

[5] A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. In *NIPS*, pages 25–32, 2007.

[6] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *ACC*, pages 4734–4739, 2001.

[7] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52:471–501, 2010.

[8] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its Oracle properties. *J. Am. Statist. Assoc.*, 96:1348–1361, 2001.

[9] Z. Lu and Y. Zhang. Schatten-p quasi-norm regularized matrix optimization via iterative reweighted singular value minimization. *arXiv:1401.0869v2*, 2015.

[10] C. Lu, J. Tang, S. Yan, and Z. Lin. Generalized nonconvex nonsmooth low-rank minimization. In *CVPR*, pages 4130–4137, 2014.

[11] F. Nie, H. Wang, X. Cai, H. Huang, and C. Ding. Robust matrix completion via joint Schatten p-norm and L_p-norm minimization. In *ICDM*, pages 566–574, 2012.

[12] A. Majumdar and R. K. Ward. An algorithm for sparse MRI reconstruction by Schatten p-norm minimization. *Magn. Reson. Imaging*, 29:408–417, 2011.

[13] K. Mohan and M. Fazel. Iterative reweighted algorithms for matrix rank minimization. *J. Mach. Learn. Res.*, 13:3441–3473, 2012.

[14] M. Lai, Y. Xu, and W. Yin. Improved iteratively rewighted least squares for unconstrained smoothed ℓ_p minimization. *SIAM J. Numer. Anal.*, 51(2):927–957, 2013.

[15] G. Marjanovic and V. Solo. On ℓ_p optimization and matrix completion. *IEEE Trans. Signal Process.*, 60(11):5714–5724, 2012.

[16] F. Nie, H. Huang, and C. Ding. Low-rank matrix recovery via efficient Schatten p-norm minimization. In *AAAI*, pages 655–661, 2012.

[17] M. Zhang, Z. Huang, and Y. Zhang. Restricted p-isometry properties of nonconvex matrix recovery. *IEEE Trans. Inform. Theory*, 59(7):4316–4323, 2013.

[18] F. Shang, Y. Liu, and J. Cheng. Scalable algorithms for tractable Schatten quasi-norm minimization. In *AAAI*, pages 2016–2022, 2016.

[19] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. In *NIPS*, pages 1329–1336, 2004.

[20] K. Mitra, S. Sheorey, and R. Chellappa. Large-scale matrix factorization with missing data under additional constraints. In *NIPS*, pages 1642–1650, 2010.

[21] A. Aravkin, R. Kumar, H. Mansour, B. Recht, and F. J. Herrmann. Fast methods for denoising matrix completion formulations, with applications to robust seismic data interpolation. *SIAM J. Sci. Comput.*, 36(5):S237–S266, 2014.

[22] S. Foucart and M. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Appl. Comput. Harmon. Anal.*, 26:397–407, 2009.

[23] Y. Liu, F. Shang, H. Cheng, and J. Cheng. A Grassmannian manifold algorithm for nuclear norm regularized least squares problems. In *UAI*, pages 515–524, 2014.

[24] W. Bian, X. Chen, and Y. Ye. Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Math. Program.*, 149:301–327, 2015.

[25] F. Shang, Y. Liu, J. Cheng, and H. Cheng. Robust principal component analysis with missing data. In *CIKM*, pages 1149–1158, 2014.

[26] F. Shang, Y. Liu, J. Cheng, and H. Cheng. Recovering low-rank and sparse matrices via robust bilateral factorization. In *ICDM*, pages 965–970, 2014.

[27] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, pages 427–435, 2013.

[28] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146:459–494, 2014.

[29] J. Yang and X. Yuan. Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Math. Comp.*, 82:301–329, 2013.

[30] J. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, 20(4):1956–1982, 2010.

[31] I. Daubechies, M. Defrise, and C. DeMol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pur. Appl. Math.*, 57(11):1413–1457, 2004.

[32] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *NIPS*, pages 612–620, 2011.

[33] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116:5–16, 2009.

[34] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi. A simplified approach to recovery conditions for low rank matrices. In *ISIT*, pages 2318–2322, 2011.

[35] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for highdimensional analysis of M-estimators with decomposable regularizers. In *NIPS*, pages 1348–1356, 2009.

[36] A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930, 2011.

[37] E. Candès and Y. Plan. Matrix completion with noise. *Proc. IEEE*, 98(6):925–936, 2010.

[38] P. Jain, R. Meka, and I. Dhillon. Guaranteed rank minimization via singular value projection. In *NIPS*, pages 937–945, 2010.

[39] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, 2010.

[40] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pac. J. Optim.*, 6:615–640, 2010.

[41] KDDCup. ACM SIGKDD and Netflix. In *Proc. KDD Cup and Workshop*, 2007.

[42] Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Math. Prog. Comp.*, 4(4):333–361, 2012.

[43] R. Cabral, F. Torre, J. Costeira, and A. Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *ICCV*, pages 2488–2495, 2013.