A Nonconvex Relaxation Approach for Rank Minimization Problems

Xiaowei Zhong, Linli Xu, Yitan Li, Zhiyuan Liu, Enhong Chen
School of Computer Science and Technology
University of Science and Technology of China, Hefei, China
xwzhong@mail.ustc.edu.cn, linlixu@ustc.edu.cn, {etali,lzy11}@mail.ustc.edu.cn, cheneh@ustc.edu.cn

Abstract

Recently, solving rank minimization problems by leveraging nonconvex relaxations has received significant attention. Some theoretical analyses demonstrate that it can provide a better approximation of original problems than convex relaxations. However, designing an effective algorithm to solve nonconvex optimization problems remains a big challenge. In this paper, we propose an Iterative Shrinkage-Thresholding and Reweighted Algorithm (ISTRA) to solve rank minimization problems using the nonconvex weighted nuclear norm as a low rank regularizer. We prove theoretically that under certain assumptions our method achieves a high-quality local optimal solution efficiently. Experimental results on synthetic and real data show that the proposed ISTRA algorithm outperforms state-of-the-art methods in both accuracy and efficiency.

Introduction

Rank minimization is a widely investigated problem in machine learning and computer vision where one intends to exploit low-dimensional structure in high-dimensional space. For example, in matrix completion (Candès and Recht 2009), it is common to assume that a partially observed matrix has low rank structure; in robust PCA (Candès et al. 2011), backgrounds of videos and faces under varying illumination are regarded as falling into a low rank subspace; in multi-task learning (Chen, Zhou, and Ye 2011), different tasks are supposed to share certain properties, which can be expressed as a low rank task-feature matrix; in subspace segmentation (Liu, Lin, and Yu 2010), clustering is performed on the low rank representation of the original data.

A general rank minimization problem can be formulated as

$$\min_{X} \; f(X) + \lambda \cdot \mathrm{rank}(X) \tag{1}$$

It has been proved that solving (1) is NP-hard due to the noncontinuous and nonconvex nature of the rank function. In order to tackle this NP-hard problem, general approaches usually relax the rank function to various regularizers, which can be categorized into convex and nonconvex relaxations.

The commonly used convex relaxation for the rank function is the nuclear norm $\|X\|_*$. Recht, Fazel, and Parrilo (2010) has proved that the nuclear norm is the convex envelop of the rank function over the domain $\|X\|_2 \leq 1$. In another word, the nuclear norm is a tight approximation of the rank function under simple conditions. Candès and Recht (2009) has shown that low rank solutions can be recovered perfectly via nuclear norm under incoherence assumptions. Due to the convex property of the nuclear norm, there are many sophisticated algorithms off the shelf. These algorithms can achieve a global optimal solution efficiently with theoretical guarantees, examples include SVT (Cai, Candès, and Shen 2010), ALM (Lin, Chen, and Ma 2010), APGL (Toh and Yun 2010), and FISTA (Beck and Teboulle 2009). However, the nuclear norm suffers from the major limitation that all singular values are simultaneously minimized, which implies that large singular values are penalized more heavily than small ones. In real applications, the underlying matrix may have no incoherence property, and the data may be grossly corrupted. Under these circumstances, methods based on nuclear norms usually fail to find a good solution. Even worse, the resulting global optimal solution may deviate significantly from the ground truth.

The nature of nonconvex relaxations is to overcome the imbalanced penalization of different singular values. Essentially, they will keep larger singular values large and shrink smaller ones, since the large singular values are dominant in determining the properties of a matrix, and should be penalized less to preserve the major information. A representative of nonconvex relaxations is the truncated nuclear norm (Hu et al. 2013) which is defined as the sum of the smallest r singular values. By minimizing only the smallest r singular values, one can avoid penalizing large singular values. In real applications, nonconvex relaxation methods usually perform better than convex relaxations and could be more robust to noise. On the other hand, the algorithms solving nonconvex relaxations may get trapped in bad local optimal solutions or cost too much time due to the hardness of nonconvex optimization. The approach of truncated nuclear norm could achieve more accurate solutions than nuclear norm methods empirically, however it has a two-layer loop that implies substantial computational overhead, and the number of singular values to be penalized is hard to determine.

In this paper, inspired by Candès, Wakin, and Boyd (2008), which uses the weighted ℓ_1 norm to enhance sparsity, we will introduce an intuitive and flexible weighted nuclear norm $\sum w_i \sigma_i$ defined as a weighted sum of all singular values to enhance low rank. To solve the nonconcex weighted nuclear norm problem, we propose an **Iterative Shrinkage-Thresholding and Reweighted Algorithm (ISTRA)**, which is simpler and faster to converge to a high-quality local optimal solution (i.e. the critical point) with solid theoretical guarantees compared with the state-of-the-art truncated nuclear norm algorithm.

Problem Formulation

Consider the general weighted nuclear norm framework for rank minimization problems

$$\min_X \ f(X) + \lambda w^T \sigma(X) \tag{2}$$

where $X \in \mathbb{R}^{m \times n}$, $q = \min(m, n)$, $w \in \mathbb{R}_+^q$ is the vector of all positive weights, $\sigma(X) = [\sigma_1(X) \cdots \sigma_q(X)]^T$, $\sigma_i(X)$ is the ith largest singular value of X, and $\lambda > 0$ is a parameter.

Note that (2) is an intuitive and unified framework rather than the true objective function[1]. To enhance low rank, we need to design a scheme to keep the weights of large singular values sufficiently small and the weights of small singular values sufficiently large, which will lead to a nearly unbiased low rank approximation. Hence, one may think of the weights in w as free parameters instead of variables in (2). Intuitively, one can set each weight w_i to be inversely proportional to the corresponding singular value $\sigma_i(X)$, which will penalize large singular values less and overcome the unfair penalization of different singular values.

Before going through the technical details, we make the following assumptions about the loss function $f(X)$ throughout this paper:

- $f : \mathbb{R}^{m \times n} \to \mathbb{R}_+$ is continuously differentiable with Lipschitz continuous gradient, i.e., for any X, Y
$$\|\nabla f(X) - \nabla f(Y)\| \leq L(f)\|X - Y\|$$
where $L(f) > 0$ is the Lipschitz constant.
- f is coercive, i.e., $f(X) \to \infty$ when $\|X\| \to \infty$.

These two assumptions are general and widely used in the design and analysis of optimization algorithms.

Methodology

In this section, we will discuss detailed techniques to solve the general problem in the unified framework (2).

Solving a Proximal Operator Problem

First, we fix w as w^k and suppose that X^k is known. Then we make a first-order approximation of $f(X)$ at X^k regularized by a quadratic proximal term:

$$P_{t^k}(X, X^k) = f(X^k) + \langle X - X^k, \nabla f(X^k) \rangle + \frac{t^k}{2}\|X - X^k\|^2 \tag{3}$$

Since optimizing f directly is hard, we minimize its first-order approximation (3) instead. Hence, our ISTRA algorithm generates the sequence $\{X^k\}$ by

$$X^{k+1} = \arg\min_X P_{t^k}(X, X^k) + \lambda (w^k)^T \sigma(X) \tag{4}$$

By ignoring constant terms and combining others, (4) can be expressed equivalently as

$$X^{k+1} = \arg\min_X \frac{t^k}{2\lambda}\left\| X - \left(X^k - \frac{1}{t^k}\nabla f(X^k) \right) \right\|_F^2 + (w^k)^T \sigma(X) \tag{5}$$

Thus, we first perform a gradient descent along the direction $-\nabla f(X^k)$ with step size $\frac{1}{t^k}$ and then solve (5), which is a nonconvex proximal operator problem (Parikh and Boyd 2013). Next we will prove (5) has a closed-form solution by exploiting the special structure of it.

Lemma 1. *(Zhang and Lu 2011) Let $\|\cdot\|$ be unitarily invariant norm on $\mathbb{R}^{m \times n}$ (i.e., $\|UXV\| = \|X\|$ for any unitary matrix U, V and any $X \in \mathbb{R}^{m \times n}$), and let $F : \mathbb{R}^{m \times n} \to \mathbb{R}$ be a unitarily invariant function (i.e. $F(UXV) = F(X)$ for any unitary matrix U, V and any $X \in \mathbb{R}^{m \times n}$). Let $A \in \mathbb{R}^{m \times n}$ be given, $q = \min(m, n)$, and let ϕ be a non-decreasing function on $[0, \infty)$. Suppose that $U \Sigma V^T$ is the singular value decomposition of A. Let operator $D : \mathbb{R}^q \to \mathbb{R}^{m \times n}$ be $D_{ij}(x) = \begin{cases} x_i & if \;\; i = j \\ 0 & otherwise \end{cases}$. Then, $X^* = UD(x^*)V^T$ is a global optimal solution of the problem*

$$\min_X \ F(X) + \phi(\|X - A\|) \tag{6}$$

where $x^* \in \mathbb{R}^q$ is a global optimal solution of the problem

$$\min_x \ F(D(x)) + \phi(\|D(x) - \Sigma\|) \tag{7}$$

Based on Lemma 1, the following conclusion is easily obtained:

Theorem 1. *Let $\mu > 0$, $A \in \mathbb{R}^{m \times n}$, $w \in \mathbb{R}_+^q$ be given, and let $q = \min(m, n)$, $X \in \mathbb{R}^{m \times n}$, $\sigma(X) = [\sigma_1(X) \cdots \sigma_q(X)]^T$, $\sigma(A) = [\sigma_1(A) \cdots \sigma_q(A)]^T$. Suppose that $U \Sigma V^T$ is the singular value decomposition of A. Then, $X^* = UD(x^*)V^T$ is a global optimal solution of the problem*

$$\min_X \ \frac{\mu}{2}\|X - A\|_F^2 + w^T \sigma(X) \tag{8}$$

where $x^* \in \mathbb{R}^q$ can be denoted as

$$x^* = \max\left(\sigma(A) - \frac{1}{\mu} w, 0 \right) \tag{9}$$

Proof. Let $F(X) = w^T \sigma(X)$, which is a unitarily invariant function. Let $\phi(\theta) = \frac{\mu}{2}\theta^2$, which is non-decreasing on $[0, \infty)$. It is known that Frobenius norm $\| \cdot \|_F$ is a unitarily invariant norm (Horn and Johnson 2012). Thus, we can find (8) coincides with (6). Substituting (7) with what we just defined, it is easy to obtain that x^* is a global optimal solution of the problem

$$\min_x \ \frac{\mu}{2}\|x - \sigma(A)\|_2^2 + w^T |x| \tag{10}$$

where $|x| = [|x_1| \cdots |x_q|]^T$. Using the soft-thresholding operator (Parikh and Boyd 2013) to solve (10), we can conclude that (9) is indeed the analytical solution of (10). Thus, according to Lemma 1, we can complete the proof of the theorem. □

[1] The true objective function can be derived as (12).

Let $\mu = \frac{t^k}{\lambda}, A = X^k - \frac{1}{t^k}\nabla f(X^k) = U\Sigma V^T, w = w^k$. According to Theorem 1, we can easily get that the closed-form solution of the nonconvex proximal operator problem (5) is $X^* = UD(x^*)V^T$, where $x^* = \max\Big(\sigma\Big(X^k - \frac{1}{t^k}\nabla f(X^k)\Big) - \frac{\lambda}{t^k}w^k, 0\Big)$.

Reweighting Strategy

We fix X as X^k and update w. Intuitively, we could design a scheme to make each weight w_i inversely proportional to $\sigma_i(X)$, which will penalize large singular values slightly. According to Candès, Wakin, and Boyd (2008), the weights can be updated by letting $w_i^k = \frac{1}{|x_i^k|+\epsilon}$ when minimizing the weighted ℓ_1 norm for sparsity. Here, we extend this from sparse optimization to low rank optimization. Thus, our reweighting strategy can be written as

$$w_i^k = \frac{r}{(\sigma_i(X^k) + \epsilon)^{1-r}} \tag{11}$$

where $i = 1 \cdots q$, $0 < r < 1$, and $\epsilon > 0$ is a smoothing parameter.

Next, we will explain the reason why w is reweighted by (11). Consider the problem

$$\min_X \Big\{g(X) = f(X) + \lambda \sum_{i=1}^{q} (\sigma_i(X) + \epsilon)^r\Big\} \tag{12}$$

where $0 < r < 1$. Let

$$h(v) = \sum_{i=1}^{q} (v_i + \epsilon)^r \tag{13}$$

where $v \in \mathbb{R}_+^q$. Since $0 < r < 1$, we know that $h(v)$ is concave on $\mathbb{R}_+^q$. Similar to (3), we make a first-order approximation of $h(v)$ at v^k:

$$S(v, v^k) = h(v^k) + \langle v - v^k, \nabla h(v^k)\rangle \tag{14}$$

Let $v = \sigma(X), v^k = \sigma(X^k)$, we optimize (12) by replacing the second term $h(\sigma(X))$ with its first-order approximation (14), i.e., $\min\ f(X) + \lambda S(\sigma(X), \sigma(X^k))$, which is equivalent to

$$\min_X\ f(X) + \lambda\Big[\frac{r}{(\sigma_i(X^k) + \epsilon)^{1-r}}\Big]_{i=1}^{q} \sigma(X) \tag{15}$$

According to (11), (15) can be reformulated as

$$\min_X\ f(X) + \lambda (w^k)^T \sigma(X) \tag{16}$$

We can see (16) is indeed a weighted nuclear norm problem, which falls into the unified framework (2). And the penalty

$$h(v) = h(\sigma(X)) = \sum_{i=1}^{q} (\sigma_i(X) + \epsilon)^r \to \operatorname{rank}(X)$$

when $\epsilon \to 0, r \to 0$. Thus, optimizing the nonconvex weighted nuclear norm problem (2) with reweighting strategy (11) is actually to solve the nonconvex problem (12), which is our true objective function and will augment the recovery of low rank matrices.

More importantly, the proposed methodology can be generalized by replacing (11) with many other reweighting strategies. One just needs to choose alternative concave and differentiable penalty functions instead of $h(v)$ in (13). In turn, new objective functions will be obtained in (12). All these variants fall into the unified framework (2) and can be solved by the proposed ISTRA algorithm, which makes (2) a flexible framework. For example, Candès, Wakin, and Boyd (2008) defines $h(v)$ as a log-sum function.

Algorithm 1 Iterative Shrinkage-Thresholding and Reweighted Algorithm (ISTRA)

```
Input: $0 < t_{\min} < t_{\max}$, $0 < \tau < 1$, $0 < r < 1$, $\lambda > 0$,
$\delta > 0$, $\epsilon > 0$, $\rho > 1$
Output: $X^*$
Initialize: $k = -1$, $w^0 = \mathbf{1}^T$, $X^{-1}$, $X^0$
repeat
    $k = k + 1$
    update $t^k$ by (17)
    make $t^k \in [t_{\min}, t_{\max}]$
    while true do
        update $X^{k+1}$ by (5)
        if line search criterion (18) is satisfied then
            Break;
        end if
        $t^k = \rho t^k$
    end while
    update the weights $w_i^{k+1}$ by (11)    $i = 1 \cdots q$
until stop criterion $\|X^{k+1} - X^k\|^2 \le \delta$ is satisfied
```

ISTRA Algorithm

Now, we present the detailed procedure of the ISTRA algorithm which is summarized in Algorithm 1. Inspired by Gong et al. (2013), we adopt the well known Barzilai-Borwein (BB) rule (Barzilai and Borwein 1988) to initialize the step size $\frac{1}{t^k}$ in (5), which will bring us a good initial step size that can reduce the line search cost. Let

$$\Delta X^k = X^k - X^{k-1}, \quad \Delta f^k = \nabla f(X^k) - \nabla f(X^{k-1})$$

According to the BB-rule, t^k is initialized as

$$t^k = \arg\min_t \|t\Delta X^k - \Delta f^k\|^2 = \frac{\langle \Delta X^k, \Delta f^k\rangle}{\langle \Delta X^k, \Delta X^k\rangle} \tag{17}$$

In each iteration, we use the line search criterion (18) to select the step size adaptively, which will accelerate our algorithm compared with constant step size:

$$\begin{aligned} f(X^{k+1}) + \lambda (w^k)^T \sigma(X^{k+1}) \le f(X^k) + \lambda (w^k)^T \sigma(X^k) \\ - \frac{\tau}{2} t^k \|X^{k+1} - X^k\|^2 \end{aligned} \tag{18}$$

where τ is a constant in $(0, 1)$. If (18) is not satisfied, we will increase t^k (decrease the step size $\frac{1}{t^k}$) by $t^k = \rho t^k$ ($\rho > 1$). We adopt $\|X^{k+1} - X^k\|^2 \le \delta$ as the stop criterion of the algorithm, with convergence guarantees as shown in (24).

Convergence Analysis

In this section, we will give a detailed convergence analysis of the proposed ISTRA algorithm, following the insights of Beck and Teboulle (2009) and Gong et al. (2013).

Boundness of the Step Size

Theorem 2. *In each iteration of the ISTRA algorithm, the line search criterion (18) is always satisfied if* $t^k \geq \frac{L(f)}{1-\tau}$.

Proof. Since $f(X)$ is continuously differentiable with Lipschitz continuous gradient, according to Nesterov (2004), for any X^{k+1}, X^k and $t \geq L(f)$

$$\begin{aligned} f(X^{k+1}) \leq\ & f(X^k) + \langle X^{k+1} - X^k, \nabla f(X^k) \rangle \\ & + \frac{t}{2}\|X^{k+1} - X^k\|^2 \end{aligned} \tag{19}$$

Since X^{k+1} is a global minimizer of (4), we can obtain $P_{t^k}(X^{k+1}, X^k) + \lambda(w^k)^T\sigma(X^{k+1}) \leq P_{t^k}(X^k, X^k) + \lambda(w^k)^T\sigma(X^k)$, which is equivalent to

$$\begin{aligned} & \langle X^{k+1} - X^k, \nabla f(X^k) \rangle + \lambda(w^k)^T\sigma(X^{k+1}) \\ & \leq -\frac{t^k}{2}\|X^{k+1} - X^k\|^2 + \lambda(w^k)^T\sigma(X^k) \end{aligned} \tag{20}$$

Summing up (19), (20) at both sides of the inequalities and combining terms, we get

$$\begin{aligned} f(X^{k+1}) + \lambda(w^k)^T\sigma(X^{k+1}) \leq f(X^k) + \lambda(w^k)^T\sigma(X^k) \\ - \frac{t^k - t}{2}\|X^{k+1} - X^k\|^2 \end{aligned}$$

Thus, if $\frac{t^k - t}{2} \geq \frac{\tau t^k}{2}$, i.e., $t^k \geq \frac{t}{1-\tau} \geq \frac{L(f)}{1-\tau}$, the line search criterion (18) will be always satisfied. □

In the procedure of line search, t^k is monotonically increasing due to $\rho > 1$. Hence, it is always true that $t^k \geq t_{\min}$. However t^k will not increase infinitely, because Theorem 2 guarantees that when t^k grows to exceed $\frac{L(f)}{1-\tau}$, the line search criterion (18) will be satisfied. That is, $t^k \leq \frac{\rho L(f)}{1-\tau}$. Thus, t^k is bounded with $t_{\min} \leq t^k \leq \frac{\rho L(f)}{1-\tau}$.

Remark 1. *If* $\frac{L(f)}{1-\tau} \leq t^k \leq \frac{\rho L(f)}{1-\tau}$ *is always held, the ISTRA algorithm will fall into a general family of Majorization Minimization (MM) methods (Hunter and Li 2005).*

Convergence Results

Definition 1. X^* *is called a critical point of problem (12), if* $\mathbf{0}$ *belongs to the subgradient of* $g(X)$ *at* X^**, i.e.,*

$$\mathbf{0} \in \partial g(X^*) = \nabla f(X^*) + \lambda \sum_{i=1}^{q} w_i^* \partial(\sigma_i(X^*))$$

where $\partial(\cdot)$ *is the subgradient (Nesterov 2004), and*

$$w_i^* = \frac{r}{(\sigma_i(X^*) + \epsilon)^{1-r}} \tag{21}$$

Now, we present the main convergence result.

Theorem 3. *The sequence* $\{X^k\}$ *generated by the ISTRA algorithm makes the objective function* $g(X)$ *in problem (12) monotonically decrease, and all accumulation points (i.e., the limit points of convergent subsequence in* $\{X^k\}$*) are critical points.*

Proof. Since $h(v)$ defined in (13) is concave on $\mathbb{R}_+^q$, it follows from (14) that for any $v \in \mathbb{R}_+^q$, $h(v) \leq S(v, v^k)$. Let $v = \sigma(X^{k+1})$, $v^k = \sigma(X^k)$, we have $h(\sigma(X^{k+1})) \leq S(\sigma(X^{k+1}), \sigma(X^k))$, which is equivalent to

$$h(\sigma(X^k)) - h(\sigma(X^{k+1})) \geq (w^k)^T\sigma(X^k) - (w^k)^T\sigma(X^{k+1}) \tag{22}$$

Since the sequences X^k, w^k, t^k generated by the ISTRA algorithm in each iteration are certain to satisfy the line search criterion (18), combined with (22), we have

$$\begin{aligned} & g(X^k) - g(X^{k+1}) \\ & = f(X^k) - f(X^{k+1}) + \lambda\big[h(\sigma(X^k)) - h(\sigma(X^{k+1}))\big] \\ & \geq f(X^k) - f(X^{k+1}) + \lambda(w^k)^T\sigma(X^k) - \lambda(w^k)^T\sigma(X^{k+1}) \\ & \geq \frac{\tau}{2}t^k\|X^{k+1} - X^k\|^2 \end{aligned} \tag{23}$$

which implies that the objective function $g(X)$ is monotonically decreasing. Since $g(X)$ is bounded from below, the sequence $\{g(X^k)\}$ will converge, i.e., $\lim_{k\to\infty} g(X^k) = p^*$. Since both $f(X)$ and $g(X)$ are coercive, the sequence $\{X^k\}$ is bounded. According to the Bolzano-Weierstrass theorem, there exists at least one convergent subsequence of $\{X^k\}$. Without loss of generality, assume that X^* is an arbitrary accumulation point of $\{X^k\}$. That is, there exists a subsequence $\{X^{k_j}\}$ such that $\lim_{j\to\infty} X^{k_j} = X^*$. Taking limits on both sides of (23) with $k \to \infty$, we obtain

$$\lim_{k\to\infty}(g(X^k) - g(X^{k+1})) \geq \lim_{k\to\infty} \frac{\tau}{2}t^k\|X^{k+1} - X^k\|^2$$

Since t^k is bounded, together with $\lim_{k\to\infty} g(X^k) = p^*$, we get

$$\lim_{k\to\infty} \|X^{k+1} - X^k\|^2 = 0 \tag{24}$$

Hence, we have $\lim_{k\to\infty}(X^{k+1} - X^k) = \mathbf{0}$. Substituting k with k_j, we get $\lim_{j\to\infty}(X^{k_j+1} - X^{k_j}) = \mathbf{0}$, which implies that $\lim_{j\to\infty} X^{k_j+1} = \lim_{j\to\infty} X^{k_j} = X^*$. Since $\sigma_i(X)$ is continuous, combined with (11) and (21), we have $\lim_{j\to\infty} w_i^{k_j} = \frac{r}{(\sigma_i(X^*)+\epsilon)^{1-r}} = w_i^*$. Considering the fact that X^{k_j+1} is a global optimal solution of (4), X^{k_j+1} is also a critical point of (4), we have

$$\mathbf{0} \in \nabla f(X^{k_j}) + t^{k_j}(X^{k_j+1} - X^{k_j}) + \lambda \sum_{i=1}^{q} w_i^{k_j} \partial(\sigma_i(X^{k_j+1})) \tag{25}$$

Taking limits on both sides of (25) with $j \to \infty$, by considering the boundness of t^{k_j}, the continuity of $\nabla f(X)$ and $\sigma_i(X)$, and the semi-continuity of subgradient $\partial(\cdot)$ (Rockafellar 1970), we obtain

$$\mathbf{0} \in \nabla f(X^*) + \lambda \sum_{i=1}^{q} w_i^* \partial(\sigma_i(X^*))$$

Therefore, any accumulation point X^* is a critical point of the objective function $g(X)$ in problem (12). □

Convergence Rate

According to (24), we use $\|X^{k+1} - X^k\|^2 \le \delta$ as a stop criterion in the ISTRA algorithm. Thus, $\|X^{k+1} - X^k\|^2$ can be a quantity to measure the rate of the subsequence of $\{X^k\}$ converging to a critical point.

Theorem 4. *Suppose that $\{X^k\}$ is the sequence generated by the ISTRA algorithm, and X^* is an accumulation point of $\{X^k\}$, then*

$$\min_{0 \le k \le n} \|X^{k+1} - X^k\|^2 \le 2(g(X^0) - g(X^*)) \Big/ n\tau t_{\min}$$

which indicates that the ISTRA algorithm can converge with sublinear rate $O(\frac{1}{n})$.

Proof. Since $t^k \ge t_{\min}$, considering (23), we obtain

$$\frac{\tau}{2} t_{\min} \|X^{k+1} - X^k\|^2 \le g(X^k) - g(X^{k+1})$$

Summing this inequality over $k = 0 \cdots n$, we have

$$\begin{aligned} \frac{\tau}{2} t_{\min} \sum_{k=0}^{n} \|X^{k+1} - X^k\|^2 &\le g(X^0) - g(X^{n+1}) \\ &\le g(X^0) - g(X^*) \end{aligned}$$

which implies that

$$\min_{0 \le k \le n} \|X^{k+1} - X^k\|^2 \le 2(g(X^0) - g(X^*)) \Big/ n\tau t_{\min}$$

□

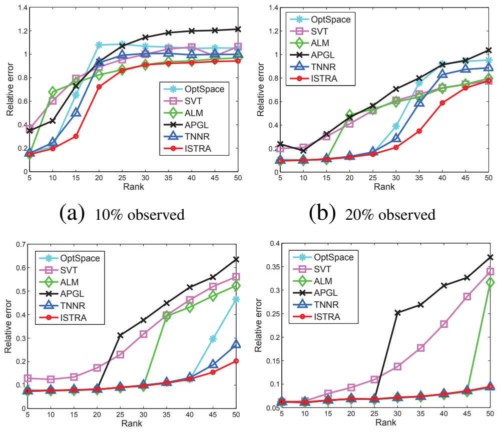
Since (2) and (12) are both nonconvex problems, it is unrealistic to solve them globally. However from the analysis above, we prove the proposed ISTRA algorithm can efficiently find a critical point that is a high-quality local optimal solution with sublinear convergence rate $O(\frac{1}{n})$, which will enhance the recovery of low rank solutions.

Experiments

In this section, we conduct experiments on the matrix completion task with both synthetic and real data. We compare the ISTRA algorithm with five commonly used matrix completion methods, among which SVT (Cai, Candès, and Shen 2010), ALM (Lin, Chen, and Ma 2010), APGL (Toh and Yun 2010) are based on the nuclear norm, OptSpace (Keshavan, Montanari, and Oh 2010) adopts matrix factorization, and TNNR (Hu et al. 2013) is the state-of-the-art nonconvex algorithm using the truncated nuclear norm.

Synthetic Data

We generate synthetic $m \times n$ matrix by $M + aZ$, where M is the ground truth matrix of rank b, Z is Gaussian white noise, and a controls the noise level. M is generated by $M = AB$, where $A \in \mathbb{R}^{m \times b}$ and $B \in \mathbb{R}^{b \times n}$ both have i.i.d. Gaussian entries. The set of observed entries Ω is uniformly sampled. We adopt the widely used measure called relative error ($RE = \|X^* - M\|_F / \|M\|_F$) to evaluate the accuracy of the recovered matrix X^*.

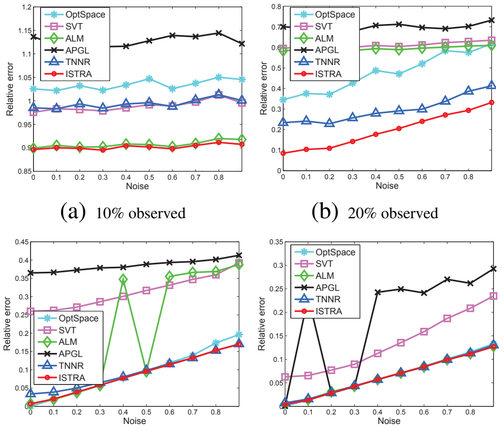


(a) 10% observed (b) 20% observed

(c) 30% observed (d) 40% observed

Figure 1: Relative error versus rank with different observations

First, we fix the matrix size and noise level to be 400×300, $a = 0.5$ respectively, and change the rank with different observed ratios. The results are shown in Figure 1. Next, we fix the matrix size and rank to be 400×300, $b = 30$ respectively, and change the noise level with different observed ratios. The results are shown in Figure 2. As can be observed from Figure 1-2, the proposed ISTRA algorithm is more robust to noise and more reliable as the underlying rank increases. Particularly, our algorithm has notable advantages when less entries are observed, and therefore is able to survive more corrupted data, which will significantly enhance the low rank recovery in real applications.



(a) 10% observed (b) 20% observed

(c) 30% observed (d) 40% observed

Figure 2: Relative error versus noise with different observations

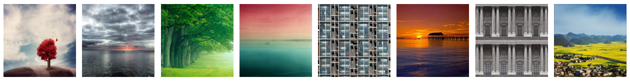


Figure 3: Images used in experiments (number 1-8)

Real Image Data

Here we consider the task of image inpainting which can also be treated as a matrix completion problem. Regarding a noisy image as three separate incomplete matrices (3 channels), we aim to recover missing pixels by exploiting the low

Table 1: PSNR values of recovered images with text mask and iteration numbers of SVD computation

Image	OptSpace	SVT	ALM	APGL	TNNR	ISTRA	#SVD (TNNR)	#SVD (ISTRA)
1	20.23	18.62	21.32	22.70	26.63	**26.68**	520	**451**
2	19.79	18.97	20.40	22.94	**30.66**	30.56	572	**547**
3	19.81	19.21	21.37	22.13	24.74	**24.81**	795	**352**
4	32.64	24.37	27.75	36.57	36.75	**37.87**	**346**	421
5	18.21	18.13	23.83	18.58	**24.87**	24.82	625	**611**
6	23.91	22.24	25.44	27.83	30.72	**31.02**	1072	**930**
7	21.95	17.95	22.24	25.83	**28.93**	28.88	611	**565**
8	16.27	12.35	19.48	18.06	20.76	**20.82**	538	**423**

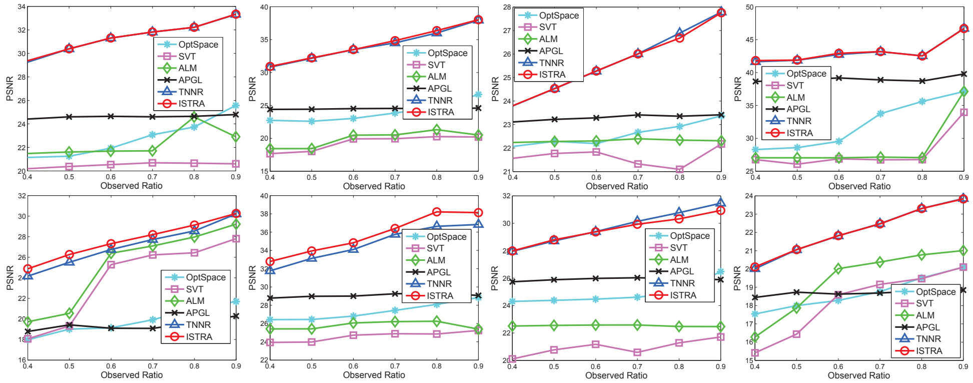


Figure 4: PSNR values of recovered images with different observed ratios for random mask (top row: images 1-4, bottom row: images 5-8)

rank structure. The quality of recovered image is evaluated by the well known PSNR (Peak Signal-to-Noise Ratio) measure. Higher PSNR values indicate better performance.

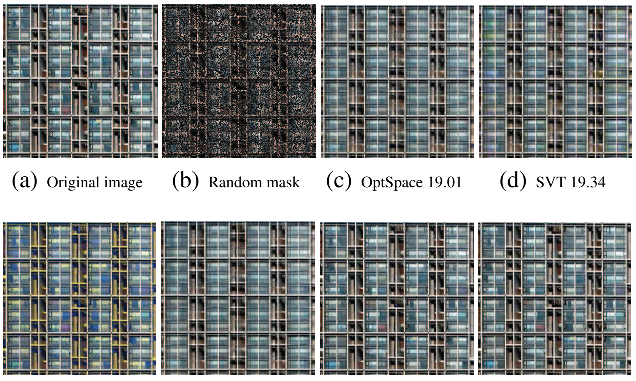
We test all methods using 8 images in Figure 3. First, we solve the matrix completion tasks with random mask, where the missing pixels are randomly sampled. The results are shown in Figure 4 and 5. Then we conduct experiments on text mask, which is more complicated since the missing pixels covered by text are not randomly distributed and the text may cover some dominant image information. The results are shown in Figure 6 and Table 1. We can see that the proposed ISTRA algorithm achieves higher or comparable PSNR values but requires less SVD iterations than the sate-of-the-art truncated nuclear norm method TNNR, which demonstrates the accuracy and efficiency of our method.

Conclusion

In this paper, we propose the ISTRA algorithm to solve rank minimization problems using the nonconvex weighted nuclear norm. We prove theoretically that the ISTRA algorithm can efficiently find a critical point that is a high-quality local optimal solution with sublinear convergence rate. The experiments further verify the accuracy and efficiency of our method.

Acknowledgement

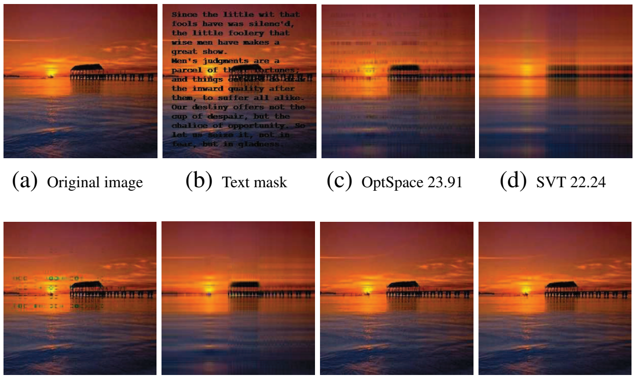
Research supported by the National Natural Science Foundation of China (No. 61375060), the Fundamental Research Funds for the Central Universities (WK0110000036), and the National Science Foundation for Distinguished Young Scholars of China (No. 61325010).



(a) Original image (b) Random mask (c) OptSpace 19.01 (d) SVT 19.34

(e) ALM 20.55 (f) APGL 19.42 (g) TNNR 25.55 (h) ISTRA 26.27

Figure 5: Recovered images and PSNR values by different methods (50% pixels are randomly masked)



(a) Original image (b) Text mask (c) OptSpace 23.91 (d) SVT 22.24

(e) ALM 25.44 (f) APGL 27.83 (g) TNNR 30.72 (h) ISTRA 31.02

Figure 6: Recovered images and PSNR values by different methods (with text on the image)

References

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

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* 2(1):183–202.

Cai, J.-F.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

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

Candès, E. J.; Li, X.; Ma, Y.; and Wright, J. 2011. Robust principal component analysis? *Journal of the ACM (JACM)* 58(3):11.

Candès, E. J.; Wakin, M. B.; and Boyd, S. P. 2008. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications* 14(5-6):877–905.

Chen, J.; Zhou, J.; and Ye, J. 2011. Integrating low-rank and group-sparse structures for robust multi-task learning. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*.

Gong, P.; Zhang, C.; Lu, Z.; Huang, J.; and Ye, J. 2013. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning (ICML)*.

Horn, R. A., and Johnson, C. R. 2012. *Matrix analysis*. Cambridge university press.

Hu, Y.; Zhang, D.; Ye, J.; Li, X.; and He, X. 2013. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(9):2117–2130.

Hunter, D. R., and Li, R. 2005. Variable selection using mm algorithms. *Annals of statistics* 33(4):1617.

Keshavan, R. H.; Montanari, A.; and Oh, S. 2010. Matrix completion from a few entries. *Information Theory, IEEE Transactions on* 56(6):2980–2998.

Lin, Z.; Chen, M.; and Ma, Y. 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *International Conference on Machine Learning (ICML)*.

Nesterov, Y. 2004. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer.

Parikh, N., and Boyd, S. 2013. Proximal algorithms. *Foundations and Trends in optimization* 1(3):123–231.

Recht, B.; Fazel, M.; and Parrilo, P. A. 2010. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* 52(3):471–501.

Rockafellar, R. T. 1970. *Convex analysis*. Princeton university press.

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

Zhang, Y., and Lu, Z. 2011. Penalty decomposition methods for rank minimization. In *Advances in Neural Information Processing Systems (NIPS)*.