Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Low-Rank Matrix Recovery via Efficient Schatten p-Norm Minimization

Feiping Nie, Heng Huang and Chris Ding
University of Texas, Arlington
feipingnie@gmail.com, heng@uta.edu, chqding@uta.edu

Abstract

As an emerging machine learning and information retrieval technique, the matrix completion has been successfully applied to solve many scientific applications, such as collaborative prediction in information retrieval, video completion in computer vision, *etc*. The matrix completion is to recover a low-rank matrix with a fraction of its entries arbitrarily corrupted. Instead of solving the popularly used trace norm or nuclear norm based objective, we directly minimize the original formulations of trace norm and rank norm. We propose a novel Schatten p-Norm optimization framework that unifies different norm formulations. An efficient algorithm is derived to solve the new objective and followed by the rigorous theoretical proof on the convergence. The previous main solution strategy for this problem requires computing singular value decompositions - a task that requires increasingly cost as matrix sizes and rank increase. Our algorithm has closed form solution in each iteration, hence it converges fast. As a consequence, our algorithm has the capacity of solving large-scale matrix completion problems. Empirical studies on the recommendation system data sets demonstrate the promising performance of our new optimization framework and efficient algorithm.

Introduction

In many machine learning applications measured data can be represented as a matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, for which only a relatively small number of entries are observed. The matrix completion problem is to find a matrix with low rank or low norm based on the observed entries, and has been actively studied in statistical learning, optimization, and information retrieval areas (Candes and Recht 2008; Candes and Tao 2009; Cai, Candes, and Shen 2008; Rennie and Srebro 2005). Such formulations occurred in many recent machine learning applications such as recommender system and collaborative prediction (Srebro, Rennie, and Jaakkola 2004; Rennie and Srebro 2005; Abernethy et al. 2009), multi-task learning (Abernethy et al. 2006; Pong et al. 2010; Argyriou, Evgeniou, and Pontil 2008), image/video completion (Liu et al. 2009), and classification with multiple classes (Amit et al. 2007).

The matrix completion problem of recovering a low-rank matrix from a subset of its entries is,

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times m}} \mathrm{rank}(\mathbf{X}), \quad \text{s.t.} \quad X_{ij} = T_{ij} \quad \forall\,(i,j)\in\Omega, \quad (1)$$

where rank($\mathbf{X}$) denotes the rank of matrix $\mathbf{X}$, and $T_{ij} \in \mathbb{R}$ are observed entries from entries set Ω. Directly solve the problem (1) is difficult as the rank minimization problem is known as NP-hard. Recently, (M.Fazel 2002) proved the trace norm function is the convex envelope of the rank function over the unit ball of matrices, and thus the trace norm is the best convex approximation of the rank function. More recently, it has been shown in (Candes and Recht 2008; Candes and Tao 2009; Recht, Fazel, and Parrilo 2010) that, under some conditions, the solution of problem in Eq. (1) can be found by solving the following convex optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times m}} \|\mathbf{X}\|_*, \quad \text{s.t.} \quad X_{ij} = T_{ij} \quad \forall\,(i,j)\in\Omega, \quad (2)$$

where $\|\mathbf{X}\|_*$ is the trace norm of $\mathbf{X}$. Several methods (Toh and Yun 2009; Ji and Ye 2009; Liu, Sun, and Toh 2009; Ma, Goldfarb, and Chen 2009; Mazumder, Hastie, and Tibshirani 2009) recently have been published to solve this kind of trace norm minimization problem.

In this paper, we propose a new optimization framework to discover low-rank matrix with Schatten p-norm, which can be used to solve problems in both Eq. (1) and Eq. (2). When $p = 1$, we have the trace norm formulation as Eq. (2); when $p \to 0$, the objective becomes Eq. (1). We introduce an efficient algorithm to solve the Schatten p-norm minimization problem with guaranteed convergence. We rigorously prove the algorithm monotonically decreases the objective with $0 < p \leq 2$ that covers the range we are interested in. Empirical studies demonstrate the promising performance of our optimization framework.

Recover Low-Rank Matrix with Schatten p-Norm

The Schatten p-Norm Definitions on Matrices

In this paper, all matrices are written as boldface uppercase and vectors are written as boldface lowercase. For matrix $\mathbf{M}$, the i-th column, the i-th row and the ij-th entry of $\mathbf{M}$

are denoted by $\mathbf{m}_i$, $\mathbf{m}^i$, and M_{ij}, respectively. For vector $\mathbf{v}$, the i-th entry of $\mathbf{v}$ is denoted by v_i.

The extended Schatten p-norm ($0 < p < \infty$) of a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ was defined as

$$\|\mathbf{M}\|_{S_p} = \left(\sum_{i=1}^{\min\{n,m\}} \sigma_i^p \right)^{\frac{1}{p}} = \left(Tr((\mathbf{M}^T\mathbf{M})^{\frac{p}{2}}) \right)^{\frac{1}{p}}, \quad (3)$$

where σ_i is the i-th singular value of $\mathbf{M}$. A widely used Schatten norm is the Schatten 1-norm:

$$\|\mathbf{M}\|_{S_1} = \sum_{i=1}^{\min\{n,m\}} \sigma_i = Tr((\mathbf{M}^T\mathbf{M})^{\frac{1}{2}}), \quad (4)$$

which is also called trace norm or nuclear norm, and is also denoted by $\|\mathbf{M}\|_*$ or $\|\mathbf{M}\|_\Sigma$ in literature.

For consistence, the Schatten 0-norm of a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ is defined as

$$\|\mathbf{M}\|_{S_0} = \sum_{i=1}^{\min\{n,m\}} \sigma_i^0, \quad (5)$$

where $0^0 = 0$. Under this definition, The Schatten 0-norm of a matrix $\mathbf{M}$ is exactly the rank of $\mathbf{M}$, *i.e.*, $\|\mathbf{M}\|_{S_0} = \text{rank}(\mathbf{M})$.

Low-Rank Matrix Completion Objectives via Schatten p-Norm

As mentioned before, many practical problems focus on the recovery of an unknown matrix from a sampling of its entries, which can be formulated as a matrix completion problem. It is commonly believed that only a few factors contribute to generate the matrix. That is to say, the unknown matrix is naturally of low rank. Therefore, the matrix completion problem can be cast as the following rank minimization problem:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} \|\mathbf{X}\|_{S_0} \quad \text{s.t.} \quad X_{ij} = T_{ij} \quad (i,j) \in \Omega, \quad (6)$$

where $T_{ij}((i,j) \in \Omega)$ are the known data sampled from entries set Ω.

The problem (6) is difficult to solve as the rank minimization problem is known as NP-hard. Recently, (M.Fazel 2002) proved the Schatten 1-norm (trace norm) function is the convex envelope of the Schatten 0-norm (rank) function over the unit ball of matrices, and thus the NP-hard problem (6) can be relaxed to the following convex problem:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} \|\mathbf{X}\|_{S_1} \quad \text{s.t.} \quad X_{ij} = T_{ij} \quad (i,j) \in \Omega. \quad (7)$$

In this paper, we propose to solve the general Schatten p-norm minimization problem as follows:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} \|\mathbf{X}\|_{S_p}^p \quad \text{s.t.} \quad X_{ij} = T_{ij} \quad (i,j) \in \Omega. \quad (8)$$

We will derive an efficient algorithm to solve this problem when $0 < p \leq 2$, and prove the algorithm convergence. When $0 < p < 1$, the problem (8) is a better approximation to the problem (6) than that of problem (7). More close the value p to 0, more better approximation the problem to the low rank problem.

Suppose $\mathbf{X} \in \mathbb{R}^{n \times m} (n \geq m)$. Using matrix form, problem (8) can be concisely written as:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} \|\mathbf{X}\|_{S_p}^p \quad \text{s.t.} \quad \mathbf{X} \circ \mathbf{H} = \mathbf{M}, \quad (9)$$

where $\circ$ is the Hadamard product, $\mathbf{M} \in \mathbb{R}^{n \times m}$, $M_{ij} = T_{ij}$ for $(i,j) \in \Omega$ and $M_{ij} = 0$ for other (i,j), $\mathbf{H} \in \mathbb{R}^{n \times m}$, $H_{ij} = 1$ for $(i,j) \in \Omega$ and $H_{ij} = 0$ for other (i,j).

Schatten p-Norm Minimization Algorithm

Following the work in (Nie et al. 2010), we derive the optimization algorithm from the Lagrangian function. The Lagrangian function of the problem in Eq. (9) is

$$\mathcal{L}(\mathbf{X}, \mathbf{\Lambda}) = Tr(\mathbf{X}^T\mathbf{X})^{\frac{p}{2}} - Tr\mathbf{\Lambda}^T(\mathbf{X} \circ \mathbf{H} - \mathbf{M}). \quad (10)$$

By taking the derivative of $\mathcal{L}(\mathbf{X}, \mathbf{\Lambda})$ w.r.t $\mathbf{X}$, and setting the derivative to zero, we have:

$$\frac{\partial \mathcal{L}(\mathbf{X}, \mathbf{\Lambda})}{\partial \mathbf{X}} = 2\mathbf{X}\mathbf{D} - \mathbf{H} \circ \mathbf{\Lambda} = \mathbf{0}, \quad (11)$$

where $\mathbf{D}$ is defined as $\mathbf{D} = \frac{p}{2}(\mathbf{X}^T\mathbf{X})^{\frac{p-2}{2}}$.

Using Eq. (11), we obtain that:

$$\mathbf{X} = \frac{1}{2}(\mathbf{H} \circ \mathbf{\Lambda})\mathbf{D}^{-1}. \quad (12)$$

According to Eq. (12) and the constraint $\mathbf{X} \circ \mathbf{H} = \mathbf{M}$, we have $((\mathbf{H} \circ \mathbf{\Lambda})\mathbf{D}^{-1}) \circ \mathbf{H} = 2\mathbf{M}$. Then for each i we have

$$\sum_k \Lambda_{ik} H_{ik} D_{kj}^{-1} H_{ij} = (\mathbf{\Lambda}^i(\mathbf{H}^i\mathbf{D}^{-1}\mathbf{H}^i))_j = 2M_{ij}, \quad (13)$$

where $\mathbf{H}^i$ is a diagonal matrix defined as $\mathbf{H}^i = \text{diag}(\mathbf{h}^i)$. Then Eq. (13) becomes $\mathbf{\Lambda}^i(\mathbf{H}^i\mathbf{D}^{-1}\mathbf{H}^i) = 2\mathbf{m}^i$. Thus we have

$$\mathbf{\Lambda}^i = 2\mathbf{m}^i(\mathbf{H}^i\mathbf{D}^{-1}\mathbf{H}^i)^{-1}. \quad (14)$$

Note that $\mathbf{D}$ is dependent on $\mathbf{X}$. If $\mathbf{D}$ is a known constant matrix, then each row of $\mathbf{\Lambda}$ can be calculated by Eq. (14), and then we can obtain the solution $\mathbf{X}$ by Eq. (11). Inspired by this fact, we propose an iterative algorithm to obtain the solution $\mathbf{X}$. The algorithm is described in Algorithm 1. In each iteration, $\mathbf{X}$ is calculated with the current $\mathbf{D}$, and then $\mathbf{D}$ is updated based on the current calculated $\mathbf{X}$. We will prove in the next subsection that the proposed iterative algorithm will converge when $0 < p \leq 2$.

Algorithm Analysis

The Algorithm 1 will monotonically decrease the objective of the problem in Eq. (9) in each iteration. To prove it, we need the following lemmas:

Lemma 1 (Araki-Lieb-Thirring (Lieb and Thirring 1976; Araki 1990; Audenaert 2008)). *For any positive semi-definite matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times m}$, $q > 0$, the following inequality holds when $0 \leq r \leq 1$:*

$$Tr(\mathbf{A}^r\mathbf{B}^r\mathbf{A}^r)^q \leq Tr(\mathbf{A}\mathbf{B}\mathbf{A})^{rq}. \quad (15)$$

While for $r \geq 1$, the inequality is reversed.

Algorithm 1 An efficient iterative algorithm to solve the optimization problem in Eq. (9).

```
Input: The data $T_{ij}$ are given for $(i,j) \in \Omega$.
$0 \le p \le 2$.
Output: $\mathbf{X} \in \mathbb{R}^{n \times m}$.
1. Define $\mathbf{M} \in \mathbb{R}^{n \times m} (n \ge m)$ where $M_{ij} = T_{ij}$ for $(i,j) \in \Omega$ and $M_{ij} = 0$ for other $(i,j)$.
2. Define $\mathbf{H} \in \mathbb{R}^{n \times m}$ where $H_{ij} = 1$ for $(i,j) \in \Omega$ and $H_{ij} = 0$ for other $(i,j)$.
3. Set $t = 0$. Initialize $\mathbf{D}_t \in \mathbb{R}^{m \times m}$ as
      $\mathbf{D}_t = \frac{p}{2}(\mathbf{M}^T\mathbf{M})^{\frac{p-2}{2}}$.
repeat
   4. Calculate $\mathbf{X}_{t+1} = \frac{1}{2}(\mathbf{H} \circ \mathbf{\Lambda}_t)\mathbf{D}_t^{-1}$, where the $i$-th row of $\mathbf{\Lambda}_t$ is calculated by $\mathbf{\Lambda}_t^i = 2\mathbf{m}^i(\mathbf{H}^i\mathbf{D}_t^{-1}\mathbf{H}^i)^{-1}$.
   5. Calculate $\mathbf{D}_{t+1} = \frac{p}{2}(\mathbf{X}_{t+1}^T\mathbf{X}_{t+1})^{\frac{p-2}{2}}$.
   6. $t = t + 1$.
until Converges
```

Lemma 2. *For any positive definite matrices $\mathbf{A}, \mathbf{A}_t \in \mathbb{R}^{m \times m}$, the following inequality holds when $0 < p \le 2$:*

$$Tr(\mathbf{A}^{\frac{p}{2}}) - \frac{p}{2}Tr(\mathbf{A}\mathbf{A}_t^{\frac{p-2}{2}}) \le Tr(\mathbf{A}_t^{\frac{p}{2}}) - \frac{p}{2}Tr(\mathbf{A}_t\mathbf{A}_t^{\frac{p-2}{2}}).$$

Proof. Because $\mathbf{A}, \mathbf{A}_t$ are positive definite, $\mathbf{A}_t^{-\frac{1}{2}}\mathbf{A}\mathbf{A}_t^{-\frac{1}{2}}$ is positive definite. Suppose the i-th eigenvalues of $\mathbf{A}_t^{-\frac{1}{2}}\mathbf{A}\mathbf{A}_t^{-\frac{1}{2}}$ is $\sigma_i > 0$, then the i-th eigenvalues of $(\mathbf{A}_t^{-\frac{1}{2}}\mathbf{A}\mathbf{A}_t^{-\frac{1}{2}})^{\frac{p}{2}}$ is $\sigma_i^{\frac{p}{2}}$, and the i-th eigenvalues of $p\mathbf{A}_t^{-\frac{1}{2}}\mathbf{A}\mathbf{A}_t^{-\frac{1}{2}} - 2(\mathbf{A}_t^{-\frac{1}{2}}\mathbf{A}\mathbf{A}_t^{-\frac{1}{2}})^{\frac{p}{2}} + (2-p)\mathbf{I}$ is $p\sigma_i - 2\sigma_i^{\frac{p}{2}} + 2 - p$.

Denote $f(\sigma_i) = p\sigma_i - 2\sigma_i^{\frac{p}{2}} + 2 - p$, then we have

$$f'(\sigma_i) = p(1 - \sigma_i^{\frac{p-2}{2}}), \quad \text{and} \quad f''(\sigma_i) = \frac{p(2-p)}{2}\sigma_i^{\frac{p-4}{2}}.$$

Obviously, when $\sigma_i > 0$ and $0 < p \le 2$, then $f''(\sigma_i) \ge 0$ and $\sigma_i = 1$ is the only point that $f'(\sigma_i) = 0$. Note that $f(1) = 0$, thus when $\sigma_i > 0$ and $0 < p \le 2$, then $f(\sigma_i) \ge 0$. Therefore, $p\mathbf{A}_t^{-\frac{1}{2}}\mathbf{A}\mathbf{A}_t^{-\frac{1}{2}} - 2(\mathbf{A}_t^{-\frac{1}{2}}\mathbf{A}\mathbf{A}_t^{-\frac{1}{2}})^{\frac{p}{2}} + (2-p)\mathbf{I}$ is positive semi-definite. As a result, we have

$$\begin{aligned}
&Tr\left(\mathbf{A}_t^{\frac{p}{4}}(p\mathbf{A}_t^{-\frac{1}{2}}\mathbf{A}\mathbf{A}_t^{-\frac{1}{2}} - 2(\mathbf{A}_t^{-\frac{1}{2}}\mathbf{A}\mathbf{A}_t^{-\frac{1}{2}})^{\frac{p}{2}} + (2-p)\mathbf{I})\mathbf{A}_t^{\frac{p}{4}}\right)\\
&\ge 0\\
&\Rightarrow Tr\left(\mathbf{A}_t^{\frac{p}{4}}(p\mathbf{A}_t^{-\frac{1}{2}}\mathbf{A}\mathbf{A}_t^{-\frac{1}{2}} - 2\mathbf{A}_t^{-\frac{p}{4}}\mathbf{A}^{\frac{p}{2}}\mathbf{A}_t^{-\frac{p}{4}} + (2-p)\mathbf{I})\mathbf{A}_t^{\frac{p}{4}}\right)\\
&\ge 0\\
&\Rightarrow Tr\left(p\mathbf{A}_t^{\frac{p-2}{4}}\mathbf{A}\mathbf{A}_t^{\frac{p-2}{4}} - 2\mathbf{A}^{\frac{p}{2}} + (2-p)\mathbf{A}_t^{\frac{p}{2}}\right) \ge 0\\
&\Rightarrow Tr(\mathbf{A}^{\frac{p}{2}}) - \frac{p}{2}Tr(\mathbf{A}\mathbf{A}_t^{\frac{p-2}{2}}) \le \frac{2-p}{2}Tr(\mathbf{A}_t^{\frac{p}{2}})\\
&\Rightarrow Tr(\mathbf{A}^{\frac{p}{2}}) - \frac{p}{2}Tr(\mathbf{A}\mathbf{A}_t^{\frac{p-2}{2}}) \le Tr(\mathbf{A}_t^{\frac{p}{2}}) - \frac{p}{2}Tr(\mathbf{A}_t\mathbf{A}_t^{\frac{p-2}{2}}),
\end{aligned}$$

where the second inequality holds according to the first inequality and Lemma 1. □

Then we have the following theorem:

Theorem 1. *When $0 < p \le 2$, the Algorithm 1 will monotonically decrease the objective of the problem in Eq.(9) in each iteration till convergence.*

Proof. It can be easily verified that Eq. (12) is the solution to the following problem:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} Tr(\mathbf{X}^T\mathbf{X}\mathbf{D}) \qquad \text{s.t.} \qquad \mathbf{X} \circ \mathbf{H} = \mathbf{M} \tag{16}$$

Thus in the t iteration,

$$\mathbf{X}_{t+1} = \arg\min_{\mathbf{X}\ \mathbf{X} \circ \mathbf{H} = \mathbf{M}} Tr(\mathbf{X}^T\mathbf{X}\mathbf{D}_t), \tag{17}$$

which indicates that

$$Tr(\mathbf{X}_{t+1}^T\mathbf{X}_{t+1}\mathbf{D}_t) \le Tr(\mathbf{X}_t^T\mathbf{X}_t\mathbf{D}_t). \tag{18}$$

That is to say,

$$\begin{aligned}
&\frac{p}{2}Tr(\mathbf{X}_{t+1}^T\mathbf{X}_{t+1}(\mathbf{X}_t^T\mathbf{X}_t)^{\frac{p-2}{2}})\\
&\le \frac{p}{2}Tr(\mathbf{X}_t^T\mathbf{X}_t(\mathbf{X}_t^T\mathbf{X}_t)^{\frac{p-2}{2}}).
\end{aligned} \tag{19}$$

On the other hand, according to Lemma 2, when $0 < p \le 2$, we have

$$\begin{aligned}
&Tr((\mathbf{X}^T\mathbf{X})^{\frac{p}{2}}) - \frac{p}{2}Tr(\mathbf{X}^T\mathbf{X}(\mathbf{X}_t^T\mathbf{X}_t)^{\frac{p-2}{2}})\\
&\le Tr((\mathbf{X}_t^T\mathbf{X}_t)^{\frac{p}{2}}) - \frac{p}{2}Tr(\mathbf{X}_t^T\mathbf{X}_t(\mathbf{X}_t^T\mathbf{X}_t)^{\frac{p-2}{2}}).
\end{aligned} \tag{20}$$

Combining Ineq. (19) and Ineq. (20), we arrive at

$$Tr((\mathbf{X}^T\mathbf{X})^{\frac{p}{2}}) \le Tr((\mathbf{X}_t^T\mathbf{X}_t)^{\frac{p}{2}}). \tag{21}$$

That is to say,

$$\|\mathbf{X}_{t+1}\|_{S_p}^p \le \|\mathbf{X}_t\|_{S_p}^p. \tag{22}$$

Thus the Alg. 1 will not increase the objective of the problem in Eq. (9) in each iteration t. Note that the equalities in the above equations hold only when the algorithm converges. Therefore, the Alg. 1 monotonically decreases the objective value in each iteration till the convergence. □

In the convergence, $\mathbf{X}_t$ and $\mathbf{D}_t$ will satisfy the Eq. (12), thus the Alg. 1 will converge to the local optimum of the problem (9). When $1 \le p \le 2$, the problem in Eq. (9) is a convex problem, satisfying the Eq. (12) indicates that $\mathbf{X}$ is a global optimum solution to the problem (9).

In each iteration, the time complexity is $O(nm^2)$. Empirical results show that the convergence is fast and only several iterations are needed to converge. Therefore, the proposed method scales well in practice.

Extensions to Other Formulations

It is worth to point out that the proposed method can be easily extended to solve the other Schatten p-norm minimization problem. For example, considering a general Schatten p-norm minimization problem as follows:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times m}} f(\mathbf{X}) + \|\mathbf{X}\|_{S_p}^p \qquad \text{s.t.} \quad \mathbf{X} \in \mathcal{C} \tag{23}$$

The problem can be solved by solve the following problem iteratively:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times m}} f(\mathbf{X}) + Tr(\mathbf{X}^T\mathbf{X}\mathbf{D}) \qquad \text{s.t.} \quad \mathbf{X}\in\mathcal{C} \tag{24}$$

where $\mathbf{D}$ is the same diagonal matrix as in Eq. (11). Similar theoretical analysis can be used to prove that the iterative method will converge to a local minimum when $0 < p \leq 2$. If the problem (23) is a convex problem, i.e., $1 \leq p \leq 2$, $f(\mathbf{X})$ is a convex function and $\mathcal{C}$ is a convex set, then the iterative method will converge to the global minimum.

A more general Schatten p-norm minimization problem is as follows:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times m}} f(\mathbf{X}) + \|\mathbf{G}(\mathbf{X})\|_{S_p}^p \qquad \text{s.t.} \quad \mathbf{X}\in\mathcal{C} \tag{25}$$

where $\mathbf{G}(\mathbf{X})$ is a linear function of $\mathbf{X}$, for example, $\mathbf{G}(\mathbf{X}) = \mathbf{A}\mathbf{X} + \mathbf{B}$. This problem can be equivalently written as:

$$\min_{\mathbf{X},\mathbf{Y}} f(\mathbf{X}) + \|\mathbf{Y}\|_{S_p}^p \qquad \text{s.t.} \quad \mathbf{X}\in\mathcal{C}, \mathbf{G}(\mathbf{X}) = \mathbf{Y}, \tag{26}$$

which can be solved by solve the following problem iteratively:

$$\min_{\mathbf{X},\mathbf{Y}} f(\mathbf{X}) + Tr(\mathbf{Y}^T\mathbf{Y}\mathbf{D}) \qquad \text{s.t.} \quad \mathbf{X}\in\mathcal{C}, \mathbf{G}(\mathbf{X}) = \mathbf{Y}, \tag{27}$$

where $\mathbf{D} = \frac{p}{2}(\mathbf{Y}^T\mathbf{Y})^{\frac{p-2}{2}}$. The convergence to a local minimum is also guaranteed when $0 < p \leq 2$. When $1 \leq p \leq 2$, if $f(\mathbf{X})$ is a convex function and $\mathcal{C}$ is a convex set, then it converges to the global minimum.

A Relaxation to Handle Data Noise

In practice, the given data $T_{ij}((i,j) \in \Omega)$ might contain noise. To handle this case, a variant method is proposed to solve the following p-norm minimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times m}} \sum_{(i,j)\in\Omega} (X_{ij} - T_{ij})^2 + \lambda \|\mathbf{X}\|_{S_p}^p . \tag{28}$$

Using matrix form, problem (28) can be concisely written as:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times m}} Tr(\mathbf{X}\circ\mathbf{H} - \mathbf{M})^T(\mathbf{X}\circ\mathbf{H} - \mathbf{M}) + \lambda \|\mathbf{X}\|_{S_p}^p \tag{29}$$

where $\mathbf{M}$ and $\mathbf{H}$ are defined as before. By setting the derivative w.r.t $\mathbf{X}$ to zero, we have

$$\begin{aligned}
&(\mathbf{X}\circ\mathbf{H} - \mathbf{M})\circ\mathbf{H} + \lambda\mathbf{X}\mathbf{D} = \mathbf{0} \\
&\Rightarrow \mathbf{X}\circ\mathbf{H}\circ\mathbf{H} - \mathbf{M}\circ\mathbf{H} + \lambda\mathbf{X}\mathbf{D} = \mathbf{0} \\
&\Rightarrow \mathbf{X}\circ\mathbf{H} - \mathbf{M}\circ\mathbf{H} + \lambda\mathbf{X}\mathbf{D} = \mathbf{0} \\
&\Rightarrow \mathbf{x}^i\mathbf{H}_i + \lambda\mathbf{x}^i\mathbf{D} = \mathbf{m}^i \circ \mathbf{h}^i \\
&\Rightarrow \mathbf{x}^i = (\mathbf{m}^i \circ \mathbf{h}^i)(\mathbf{H}_i + \lambda\mathbf{D})^{-1}, \qquad (30)
\end{aligned}$$

where $\mathbf{D}$ is the same as in Eq. (11) and $\mathbf{H}_i$ is a diagonal matrix with the diagonal entries as $\mathbf{h}^i$, *i.e.* $\mathbf{H}_i = diag(\mathbf{h}^i)$.

Similarly, $\mathbf{D}$ is dependent on $\mathbf{X}$, and if $\mathbf{D}$ is known, then we can obtain the solution $\mathbf{X}$ by Eq. (30). We propose an iterative algorithm to obtain the solution $\mathbf{X}$ and the algorithm is described in Algorithm 2. In each iteration, $\mathbf{X}$ is calculated with the current $\mathbf{D}$, and then $\mathbf{D}$ is updated based on the current calculated $\mathbf{X}$. Note that Eq. (30) is the solution to the following problem

Algorithm 2 An efficient iterative algorithm to solve the optimization problem in Eq. (29).

```
Input: The data $T_{ij}$ are given for $(i,j) \in \Omega$. $0 \leq p \leq 2$, regularization parameter $\lambda$.
Output: $\mathbf{X} \in \mathbb{R}^{n\times m}$.
1. Define $\mathbf{M} \in \mathbb{R}^{n\times m}(n \geq m)$ where $M_{ij} = T_{ij}$ for $(i,j) \in \Omega$ and $M_{ij} = 0$ for other $(i,j)$.
2. Define $\mathbf{H} \in \mathbb{R}^{n\times m}$ where $H_{ij} = 1$ for $(i,j) \in \Omega$ and $H_{ij} = 0$ for other $(i,j)$.
3. Set $t = 0$. Initialize $\mathbf{D}_t \in \mathbb{R}^{m\times m}$ as $\mathbf{D} = \frac{p}{2}(\mathbf{M}^T\mathbf{M})^{\frac{p-2}{2}}$.
repeat
    4. Calculate $\mathbf{X}_{t+1}$, where the $i$-th row of $\mathbf{X}_{t+1}$ is calculated by $\mathbf{x}_{t+1}^i = (\mathbf{m}^i \circ \mathbf{h}^i)(\mathbf{H}_i + \lambda\mathbf{D}_t)^{-1}$.
    5. Calculate $\mathbf{D}_{t+1} = \frac{p}{2}(\mathbf{X}_{t+1}^T\mathbf{X}_{t+1})^{\frac{p-2}{2}}$.
    6. $t = t + 1$.
until Converges
```

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times m}} Tr(\mathbf{X}\circ\mathbf{H} - \mathbf{M})^T(\mathbf{X}\circ\mathbf{H} - \mathbf{M}) + \lambda Tr\mathbf{X}^T\mathbf{X}\mathbf{D}.$$

It can be similarly proved that the iterative algorithm will converge when $0 < p \leq 2$.

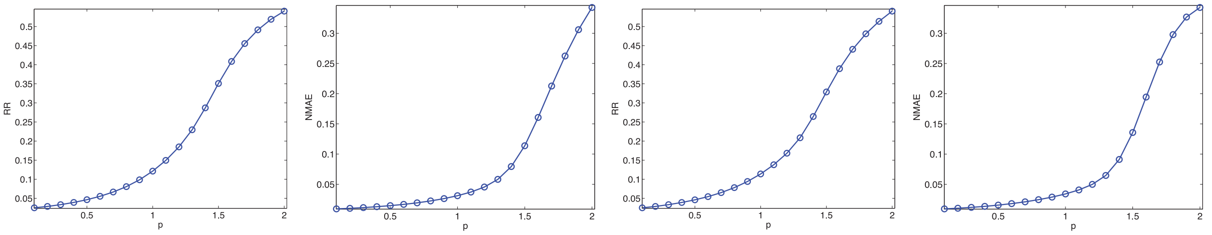
Related Work

In recent sparse learning research, several approaches have been proposed to solve the matrix completion problem (Toh and Yun 2009; Ji and Ye 2009). They tried to solve the trace norm minimization problem. In our work, we target to solve a more general Schatten p-norm problem and the trace norm formulation is a special case when $p = 1$. When $p \to 0$, our objective becomes Eq. (1) that is the exactly original problem, hence our solution is a better approximation of the true solution of low-rank matrix completion problem.

The main content of this work was finished in the beginning of 2010. Just after it was finished, we found another recent paper (Argyriou et al. 2007) also solved a Schatten p-norm problem for multi-task learning. This paper (Argyriou et al. 2007) solved the regularization problem with the regularizer as the **squared** Schatten p-norm, *i.e.*

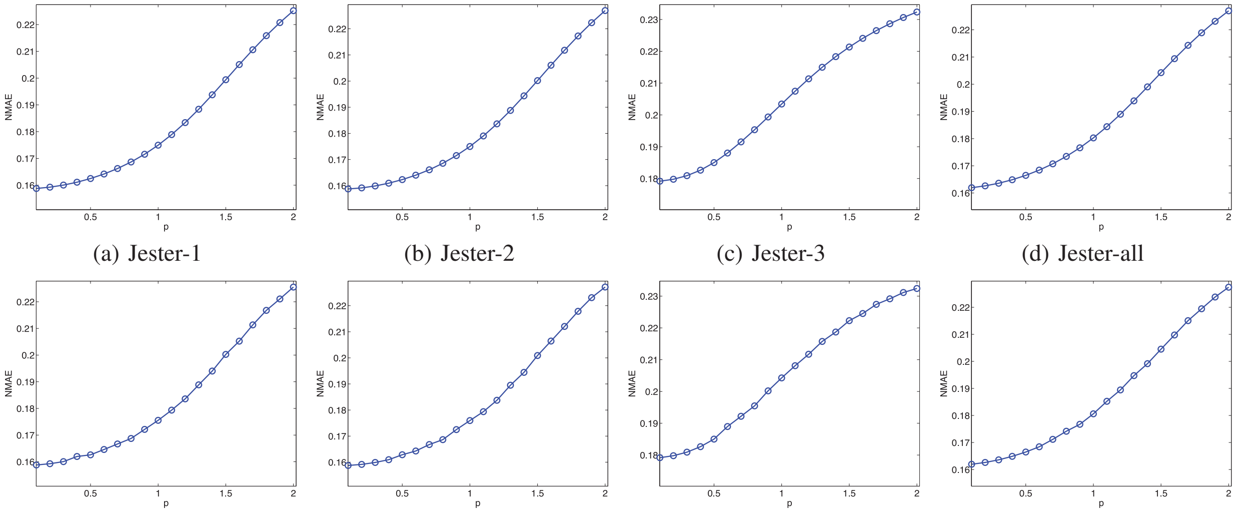
$$\min_{\mathbf{X}\in\mathbb{R}^{n\times m}} f(\mathbf{X}) + (\|\mathbf{X}\|_{S_p}^p)^2. \tag{31}$$

However, in our work, we directly use the Schatten p-norm (without square) as loss function or regularizer to solve the matrix completion problem. These two problems are totally different. Although the algorithms seem similar at first glance, they are essentially different in optimization derivations as they solve different problems. The algorithm proposed in (Argyriou et al. 2007) cannot be applied to our problem. Moreover, we further solve a constrained problem in this paper. The previous paper (Argyriou et al. 2007) didn't provide a proof of the algorithm convergence. In this



(a) RR vs. p (b) $NMAE$ vs. p (c) RR vs. p (d) $NMAE$ vs. p

Figure 1: Performance with different p between 0.1 and 1 by solving problem (9)(the first two figures) or problem (29) (the last two figures).



(a) Jester-1 (b) Jester-2 (c) Jester-3 (d) Jester-all

(e) Jester-1 (f) Jester-2 (g) Jester-3 (h) Jester-all

Figure 2: NMAE vs. different p between 0.1 and 2 by solving problem in Eq. (9) (the first row) or problem Eq. (29) (the second row) on the Jester data.

paper, we provide a rigorous proof on the algorithm convergency and our approaches can be easily extended to the much more general Schatten p-norm minimization problem.

Experimental Results

In this section, we present the numerical results for solving problem (9) and problem (29) with $0 < p \leq 2$, and compare the performance between different values of p. In the experiments, p is selected between 0.1 and 2 with an interval 0.1. For the problem (29), the regularization parameter λ is simply set to 1 in the experiments.

Numerical Experiments on Low-Rank Matrix Recovery Problem

In this experiment, we consider the low-rank matrix recovery problem. A low-rank matrix $\mathbf{T} \in \mathbb{R}^{100\times 100}$ is randomly generated as follows: First, a matrix $\mathbf{B} \in \{0,1\}^{100\times 20}$ is randomly generated. Then let $\mathbf{T} = \mathbf{B}\mathbf{B}^T$. Thus the rank of $\mathbf{T}$ is 20. We randomly sample 5000 entries in $\mathbf{T}$ as the entries set Ω, and recover the other 5000 entries in $\mathbf{T}$ by solving problem (9) or problem (29).

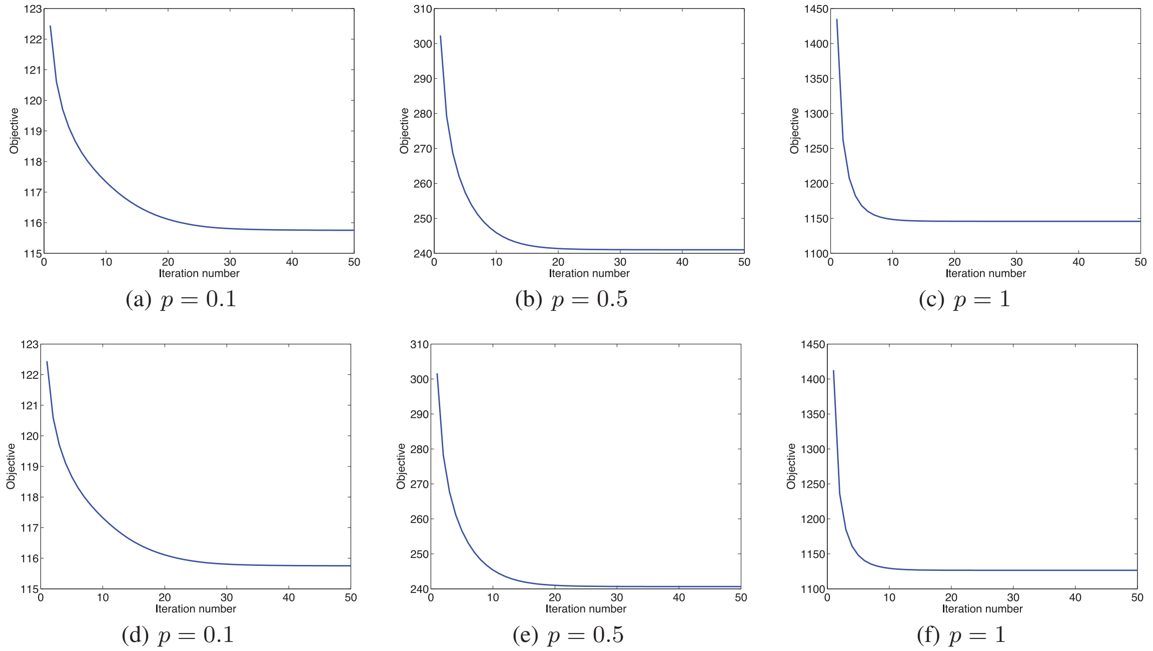
We use two metrics to measure the performance of the recovery. Suppose the i-th singular value of $\mathbf{X}$ is σ_i. One metric is to measure the rank of recovered matrix $\mathbf{X}$. Due to the numerical issue, we use the following metric to measure it instead of calculate the rank of $\mathbf{X}$ directly: $RR = \sum_{i=21}^{100} \sigma_i / \sum_{i=1}^{100} \sigma_i$. Smaller the value RR indicates lower the rank of $\mathbf{X}$.

The other metric is to measure the error of recovery. We use the Normalized Mean Absolute Error (NMAE) as in (Goldberg et al. 2001). The $NMAE$ is defined as:

$$\frac{1}{T_{\max} - T_{\min}} \cdot \frac{1}{|\Gamma \setminus \Omega|} \cdot \sum_{(i,j)\in\Gamma\setminus\Omega} |X_{ij} - T_{ij}| , \qquad (32)$$

where $T_{\max} = \max_{ij} T_{ij}$, $T_{\min} = \min_{ij} T_{ij}$, Γ is the entries set that the ground truth of T_{ij} are known. Smaller the value $NMAE$ indicates better the recovery of $\mathbf{X}$.

The results are shown in Fig. (1). Interestingly, we can see from the figures that the rank of the recovered matrix $\mathbf{X}$ monotonically decreases, and the recovery quality simultaneously becomes better when p decrease. The results clearly demonstrate the effectiveness of the recovery algorithm with Schatten p-norm.



(a) $p = 0.1$ (b) $p = 0.5$ (c) $p = 1$

(d) $p = 0.1$ (e) $p = 0.5$ (f) $p = 1$

Figure 3: The algorithm convergency results to solve the optimization problem in Eq. (9) (the first three figures) or problem in Eq. (29) (the last three figures) when $p = 0.1, 0.5, 1$.

Numerical Experiments on Real Matrix Completion Problem

In this experiment, we verify the performance on real world matrix completion problem with Jester joke data set. The Jester joke data set contains 4.1 million ratings for 100 jokes from 73421 users The data set consists of three files: Jester-1, Jester-2 and Jester-3. Jester-1 contains 24983 users who have rated 36 or more jokes, Jester-2 contains 23500 users who have rated 36 or more jokes, and Jester-3 contains 24938 users who have rated between 15 and 35 jokes. We combines all the three data sets and produce a new data set Jester-all, which contains 73421 users.

For each data set, we only have an incomplete data matrix.Let Γ be the set of entries that the ratings have been provided by users. In the experiment, we randomly choose half of entries in Γ to construct the entries set Ω, the $NMAE$ is used to measure the performance. The results are shown in Fig. (2). We have consistent conclusion that the recovery algorithm with Schatten p-norm would further improve the recovery performance over recent developed recovery algorithm with Schatten 1-norm (*i.e.*, trace norm) when $p < 1$.

Experiments on Convergency Analysis

When we run the experiments we also record the objective values after each iteration. Fig. 3 reports the algorithm convergency results to solve the optimization problem in Eq. (9) or problem in Eq. (29) when $p = 0.1, 0.5, 1$. In both figures, we can see that when the value of p decreases, the algorithm convergency speed is also reduced. When $p = 1$, *i.e.* the trace norm problem, our algorithms converge very fast within ten iterations. Although our algorithm need more iterations when $p = 0.1$, the computational speed is still fast, because the algorithms have the closed form solution in each iteration.

Conclusions

In this paper, we studied the approaches of solving a low-rank factorization model for the matrix completion problem that recovers a low-rank matrix from a subset of its entries. A new Schatten p-Norm optimization framework has been proposed to solve rank norm and trace norm objectives. We derived an efficient algorithm to minimize Schatten p-Norm objective and proved that our algorithm monotonically decreases the objective till convergence. The time complexity analysis reveals our method efficiently works for large-scale matrix completion problems. Experiments on real data sets validated the performance of our new algorithm.

Acknowledgement

This research was partially supported by NSF-CCF 0830780, NSF-DMS 0915228, NSF-CCF 0917274, NSF-IIS 1117965.

References

Abernethy, J.; Bach, F.; Evgeniou, T.; and Vert, J. P. 2006. Low-rank matrix factorization with attributes. *(Technical Report N24/06/MM). Ecole des Mines de Paris*.

Abernethy, J.; Bach, F.; Evgeniou, T.; and Vert, J. P. 2009. A new approach to collaborative filtering: Operator estimation with spectral regularization. *JMLR* 10:803–826.

Amit, Y.; Fink, M.; Srebro, N.; and Ullman, S. 2007. Uncovering shared structures in multiclass classification. *ICML*.

Araki, H. 1990. On an inequality of lieb and thirring. *Letters in Mathematical Physics* 19(2):167–170.

Argyriou, A.; Micchelli, C. A.; Pontil, M.; and Ying, Y. 2007. A spectral regularization framework for multi-task structure learning. In *NIPS*.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Machine Learning* 73(3):243–272.

Audenaert, K. 2008. On the araki-lieb-thirring inequality. *International Journal of Information and Systems Sciences* 4(1):78–83.

Cai, J.-F.; Candes, E. J.; and Shen, Z. 2008. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization* 20(4):1956–1982.

Candes, E., and Recht, B. 2008. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*.

Candes, E. J., and Tao, T. 2009. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* 56(5):2053–2080.

Goldberg, K. Y.; Roeder, T.; Gupta, D.; and Perkins, C. 2001. Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retr.* 4(2):133–151.

Ji, S., and Ye, Y. 2009. An accelerated gradient method for trace norm minimization. *ICML*.

Lieb, E., and Thirring, W. 1976. Inequalities for the moments of the eigenvalues of the schrodinger hamiltonian and their relation to sobolev inequalities.g. *Essays in Honor of Valentine Borgmann* 269–303.

Liu, J.; Musialski, P.; Wonka, P.; and Ye, J. 2009. Tensor completion for estimating missing values in visual data. *ICCV*.

Liu, Y.-J.; Sun, D.; and Toh, K.-C. 2009. An implementable proximal point algorithmic framework for nuclear norm minimization. *Optimization Online*.

Ma, S.; Goldfarb, D.; and Chen, L. 2009. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*.

Mazumder, R.; Hastie, T.; and Tibshirani, R. 2009. Spectral regularization algorithms for learning large incomplete matrices. *submitted to JMLR*.

M.Fazel. 2002. Matrix rank minimization with applications. *Ph.D. dissertation, Stanford University*.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*.

Pong, T. K.; Tseng, P.; Ji, S.; and Ye, J. 2010. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*.

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.

Rennie, J., and Srebro, N. 2005. Fast maximum margin matrix factorization for collaborative prediction. *ICML*.

Srebro, N.; Rennie, J.; and Jaakkola, T. 2004. Maximum-margin matrix factorization. *NIPS* 17:1329–1336.

Toh, K., and Yun, S. 2009. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Optimization Online*.