Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

A Unified Convex Surrogate for the Schatten-p Norm

Chen Xu, Zhouchen Lin,* Hongbin Zha
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, P. R. China
Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, P. R. China
xuen@pku.edu.cn, zlin@pku.edu.cn, zha@cis.pku.edu.cn

Abstract

The Schatten-p norm ($0 < p < 1$) has been widely used to replace the nuclear norm for better approximating the rank function. However, existing methods are either 1) not scalable for large scale problems due to relying on singular value decomposition (SVD) in every iteration, or 2) specific to some p values, e.g., $1/2$, and $2/3$. In this paper, we show that for any p, p_1, and $p_2 > 0$ satisfying $1/p = 1/p_1 + 1/p_2$, there is an equivalence between the Schatten-p norm of one matrix and the Schatten-p_1 and the Schatten-p_2 norms of its two factor matrices. We further extend the equivalence to multiple factor matrices and show that all the factor norms can be *convex and smooth* for any $p > 0$. In contrast, the original Schatten-p norm for $0 < p < 1$ is non-convex and non-smooth. As an example we conduct experiments on matrix completion. To utilize the convexity of the factor matrix norms, we adopt the accelerated proximal alternating linearized minimization algorithm and establish its sequence convergence. Experiments on both synthetic and real datasets exhibit its superior performance over the state-of-the-art methods. Its speed is also highly competitive.

Introduction

In recent years, low rank matrix minimization has found wide applications, e.g., matrix completion (Candès and Recht 2009), low rank representation (Liu, Lin, and Yu 2010), multi-task learning (Dudik, Harchaoui, and Malick 2012), etc. Often, we can formulate the problem as follows:

$$\min_X F(X) = \min_X f(X) + \lambda\Omega(X), \tag{1}$$

where $f(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}^+$ is the loss function, $\Omega(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}^+$ is the spectral regularization (Abernethy et al. 2009) which ensures low rankness, and $\lambda \in \mathbb{R}^+$ balances the two terms.

As the tightest convex envelop of rank function on the unit ball of matrix operator norm, the nuclear norm regularizer is often suggested for $\Omega(X)$ (Recht, Fazel, and Parrilo 2010; Candès and Tao 2010). In fact, the nuclear norm is the ℓ_1-norm on the vector of singular values. It achieves low rankness by encouraging sparseness on the singular values. As (Fan and Li 2001) pointed out, the ℓ_1-norm is a loose approximation to the ℓ_0-norm and overpenalizes large entries of vectors. By an analogy between the rank function of matrices and the ℓ_0-norm of vectors, the nuclear norm also overpenalizes large singular values. As a tighter approximation to the rank function, the Schatten-p quasi-norm ($0 < p < 1$) is suggested to replace the nuclear norm (Nie, Huang, and Ding 2012). For the task of matrix completion, the Schatten-p quasi-norm has empirically shown to be superior to the nuclear norm. Moreover, (Zhang, Huang, and Zhang 2013) theoretically prove that for the matrix completion problem the Schatten-p quasi-norm with a small p requires much fewer observed entries than the nuclear norm minimization does.

However, the Schatten-p quasi-norm is non-convex and non-smooth. So the optimization for problem (1) is much more challenging. Recently, (Lai, Xu, and Yin 2013) propose iterative reweighted least square (IRucLp) to solve a smoothed subproblem by approximating the Schatten-p quasi-norm at each iteration. They prove that any limit point of the generated sequence is a stationary point. Moreover, (Lu et al. 2014) propose the iterative reweighted nuclear norm (IRNN) algorithm. Besides the Schatten-p quasi-norm, IRNN is able to tackle a variety of regularizations, e.g., MCP (Zhang 2010) and SCAD (Fan and Li 2001), on the singular values. A similar convergence result as IRcuLq is also established. However, both of the algorithms involve computing SVD at each iteration, which is expensive for large-scale problems.

Alternative to (1), the bilinear factorization with two factor matrix norm regularizers is suggested (Srebro, Rennie, and Jaakkola 2004; Cabral et al. 2013; Shang, Liu, and Cheng 2016a):

$$\min_{U,V} F(U,V) = \min_{U,V} f(UV^T) + \lambda\left(\Omega_u(U) + \Omega_v(V)\right), \tag{2}$$

where $U \in \mathbb{R}^{m\times d}$ and $V \in \mathbb{R}^{n\times d}$ are the unknown factor matrices. Quite often, $d \ll \min\{m, n\}$ holds. When minimizing (2), one only needs to operate on two much smaller factor matrices in contrast to the full dimensional X as (1). Thus (2) is better suited for large-size applications. As (Srebro, Rennie, and Jaakkola 2004) indicated, when $\Omega_u(U) + \Omega_v(V) = \|U\|_F^2/2 + \|V\|_F^2/2$, it can be equivalently represented as the surrogate of $\Omega(X) = \|X\|_*$ when enforcing $X = UV^T$. Let r^* denote the rank of the optimal X^* in (1), (Mazumder, Hastie, and Tibshirani 2010) proved that the minimum objective function values of (1) and (2) are

*Corresponding author.

equal once $d \geq r^*$. Quite recently, Shang, Liu, and Cheng (2016a,b) extended the surrogate of the nuclear norm regularizer $\Omega(X)$ to that of specific Schatten-p norms, where $p = 1/3$, $1/2$, or $2/3$. They proposed to use the proximal alternating linearized minimization (PALM) algorithm and established its sequence convergence. Motivated by these results, we further extend the surrogate to the general Schatten-p norm. The contributions of this paper are as follows:

(a) We show that for any p, p_1, and $p_2 > 0$ satisfying $1/p = 1/p_1 + 1/p_2$, there is an equivalence between the Schatten-p norm of X and the Schatten-p_1 and the Schatten-p_2 norms of U and V when enforcing $X = UV^T$ (See Theorem 1). The existing surrogates for $p = 1$, $1/2$, and $2/3$ are only special cases of ours. We also give an *entirely different and much simpler* proof than the existing ones.

(b) We extend the above result to multi-factor matrices (See Corollary 1) and show that each factor matrix norm of the surrogate can be *convex and smooth* for any $p > 0$. In contrast, the Schatten-p norm $(0 < p < 1)$ is non-smooth and non-convex, and the results of Shang, Liu, and Cheng (2016a,b are only limited to two or three-factor cases which all involve the non-smooth nuclear norm.

(c) We unify the minimization of (1) and (2) for general Schatten-p norm regularizers, where the former is reformulated to the latter (See Theorem 2). We also show that the factorization formulation should be preferred when $0 < p < 1$.

(d) We conduct experiments on matrix completion as an example to test our framework. By incorporating the convexity of the factor matrix norms, our accelerated proximal alternating algorithm achieves state-of-the-art performance. We also prove its sequence convergence.

Notations and Background

Consider the SVD of a matrix $X \in \mathbb{R}^{m\times n}$: $X = U_X \mathrm{diag}\left(\sigma_i(X)\right) V_X^T$, where $\sigma_i(X)$ denotes its i-th singular value in descending order. Then the Schatten-p norm $(0 < p < \infty)$ of X is defined as

$$\|X\|_{S_p} \triangleq \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p(X) \right)^{\frac{1}{p}}. \tag{3}$$

Special cases of the Schatten-p norm include the nuclear norm ($p = 1$) and the Frobenius norm ($p = 2$). When $p \geq 1$, $\|X\|_{S_p}^p$ is convex w.r.t. X. When $p > 1$, $\|X\|_{S_p}^p$ is further differentiable everywhere with the gradient being $\nabla_X \|X\|_{S_p}^p = p U_X \mathrm{diag}\left(\sigma_i^{p-1}(X)\right) V_X^T$ (Watson 1992).

The proximal mapping of $\|X\|_{S_p}^p$ is defined as:

$$\mathbf{Prox}_{\lambda,p}(Y) = \arg\min_X \frac{1}{2}\|X - Y\|_F^2 + \frac{\lambda}{p}\|X\|_{S_p}^p \tag{4}$$

Lemma 1. *(Lu et al. 2015) Let $Y = U_Y diag\left(\sigma_i(Y)\right) V_Y^T$ be the SVD of $Y \in \mathbb{R}^{m\times n}$ with $\{\sigma_i(Y)\}$ in descending order. Then we have*

$$\mathbf{Prox}_{\lambda,p}(Y) = U_Y \mathrm{diag}(\hat{\sigma}_i) V_Y^T, \tag{5}$$

where $\hat{\sigma}_i$ is defined as the scalar proximal mapping in (4)*:*

$$\hat{\sigma}_i = \mathbf{Prox}_{\lambda,p}(\sigma_i(Y)). \tag{6}$$

When $p \geq 1$, problem (6) is strongly convex. Thus it can be easily solved by off-the-shelf algorithms. There are some special cases of p that have closed-form solutions, e.g. $\hat{\sigma}_i = \max(\sigma_i(Y) - \lambda, 0)$ if $p = 1$, which is known as the soft-thresholding (Donoho et al. 1995), and $\hat{\sigma}_i = \sigma_i(Y)/(1 + \lambda)$ when $p = 2$ by making the derivative of the objective function zero. When $p < 1$, problem (6) becomes non-smooth and non-convex. By considering its structure properly, Zuo et al. (Zuo et al. 2013) proposed the generalized iterated shrinkage algorithm (GISA), which solved (6) efficiently with high precision. Thus in this paper, we regard the proximal mapping of (4) for any $0 < p < \infty$ as an easy problem.

For the analysis on convergence of algorithms for problems with non-convex objectives, we need the definition of critical points given in (Bolte, Sabach, and Teboulle 2014):

Definition 1. ***(Critical Points)*** *Let a non-convex function $g : \mathbb{R}^n \to (\infty, +\infty]$ be a proper and lower semi-continuous function, and* $\mathrm{dom}\ g = \{x \in \mathbb{R}^n : g(x) < +\infty\}$.

- *For any $x \in \mathrm{dom}\ g$, the Fréchet sub-differential of g at x is defined as*

$$\hat{\partial} g(x) = \{u \in \mathbb{R}^n : \lim_{y \neq x,\ y \to x} \inf \frac{g(y) - g(x) - \langle u, y - x\rangle}{\|y - x\|_2} \geq 0\}, \tag{7}$$

and

$$\hat{\partial} g(x) = \emptyset \text{ if } x \notin \mathrm{dom}\ g. \tag{8}$$

- *The points whose sub-differential contains 0 are called critical points, i.e., a point x is a critical point of g if $0 \in \partial g(x)$.*

Unified Surrogate for Schatten-p Norm

Before giving our unified surrogate for the Schatten-p norm, we review three existing surrogates for specific p values, i.e., $p = 1, 2/3$, and $1/2$.

Proposition 1. ***(Bi-Frobenius Norm Surrogate*** *(Srebro, Rennie, and Jaakkola 2004; Mazumder, Hastie, and Tibshirani 2010)) Given matrices $U \in \mathbb{R}^{m\times d}$, $V \in \mathbb{R}^{n\times d}$, and $X \in \mathbb{R}^{m\times n}$ with $rank(X) = r \leq d$, the following holds:*

$$\|X\|_* = \min_{U,V:X=UV^T} \frac{\|U\|_F^2}{2} + \frac{\|V\|_F^2}{2}. \tag{9}$$

Proposition 2. ***(Frobenius/Nuclear and Bi-Nuclear Norm Surrogate*** *Shang, Liu, and Cheng (2016a,b)) Given matrices $U \in \mathbb{R}^{m\times d}$, $V \in \mathbb{R}^{n\times d}$, and $X \in \mathbb{R}^{m\times n}$ with $rank(X) = r \leq d$, the following holds:*

$$\begin{aligned} \frac{3}{2}\|X\|_{S_{2/3}}^{2/3} &= \min_{U,V:X=UV^T} \|U\|_* + \frac{1}{2}\|V\|_F^2 \quad \textit{and} \\ 2\|X\|_{S_{1/2}}^{1/2} &= \min_{U,V:X=UV^T} \|U\|_* + \|V\|_*. \end{aligned} \tag{10}$$

Note that we have rewritten Proposition 2 in a more consistent way than the original one in Shang, Liu, and Cheng (2016a,b). Combining the above two propositions, one may induce that there are some unified surrogates for Schatten-p norm. In fact, we have:

Theorem 1. ***(Bi-Schatten-p Norm Surrogate)*** *Given matrices $U \in \mathbb{R}^{m\times d}$, $V \in \mathbb{R}^{n\times d}$, and $X \in \mathbb{R}^{m\times n}$ with $rank(X) = r \leq d$, for any p, p_1 and $p_2 > 0$ satisfying $1/p = 1/p_1 + 1/p_2$, we have*

$$\frac{1}{p}\|X\|_{S_p}^p = \min_{U,V:X=UV^T} \frac{1}{p_1}\|U\|_{S_{p_1}}^{p_1} + \frac{1}{p_2}\|V\|_{S_{p_2}}^{p_2}. \quad (11)$$

We provide an *entirely different and much simpler* proof than those in Shang, Liu, and Cheng (2016a,b), (Mazumder, Hastie, and Tibshirani 2010), which use the property of the specific p, p_1, and p_2 values shown in Propositions 1 and 2 [1]. The core idea is to utilize the property of general Schatten-p norms derived from the determinant of matrices:

Lemma 2. *(Horn and Johnson 1991)[Theorem 3.3.14 (c)] For any matrices $A \in \mathbb{R}^{m\times l}$ and $B \in \mathbb{R}^{n\times l}$, denoting $\{\sigma_i(\cdot)\}$ as the singular values in descending order, we have*

$$\sum_{i=1}^{\min\{m,n,l\}} \sigma_i^p(AB^T) \leq \sum_{i=1}^{\min\{m,n,l\}} \sigma_i^p(A)\sigma_i^p(B), \ \ \forall\, p > 0. \quad (12)$$

For the completed proof of Theorem 1, please refer to the Supplementary Material[2]. By extending Theorem 1 to multiple factors, we have [3]:

Corollary 1. ***(Multi-Schatten-p Norm Surrogate)*** *Given I ($I \geq 2$) matrices X_i, $i = 1, \ldots, I$, where $X_1 \in \mathbb{R}^{m\times d_1}$, $X_i \in \mathbb{R}^{d_{i-1}\times d_i}, i = 2, \ldots, I-1$, $X_I \in \mathbb{R}^{d_I\times n}$, and $X \in \mathbb{R}^{m\times n}$ with $rank(X) = r \leq \min\{d_i, i = 1, \ldots, I\}$, for any $p, p_1, \ldots, p_I > 0$ satisfying $1/p = \sum_1^I 1/p_i$, we have*

$$\frac{1}{p}\|X\|_{S_p}^p = \min_{X_i:X=\prod_{i=1}^I X_i} \sum_{i=1}^{I} \frac{1}{p_i}\|X_i\|_{S_{p_i}}^{p_i}. \quad (13)$$

For the ease of computation, we assume that $d_i = d$, $i = 1, \ldots, I$ in the rest of this paper. Special cases of Corollary 1 include ($p = 1/I, p_i = 1$, $i = 1, \ldots, I$), e.g., Tri-Nuclear norm surrogate in (Shang, Liu, and Cheng 2016b), and ($p = 2/I, p_i = 2$, $i = 1, \ldots, I$). When $I = 2$, Corollary 1 reduces to Theorem 1. Corollary 1 can be proved by induction, using Theorem 1. In fact, the two do not just differ in the number of factors. Corollary 1 enables us to choose some particular p_i values which endow the operator $\|\cdot\|_{S_{p_i}}^{p_i}$ with nice properties, especially when $0 < p < 1$, the case that we are mainly interested in.

Proposition 3. *For any $0 < p < 1$, there always exist $I \in N$ and p_i such that $1/p = \sum_i^I 1/p_i$, where all p_i satisfy one of the cases: (a) $p_i \geq 1$ or (b) $p_i > 1$.*

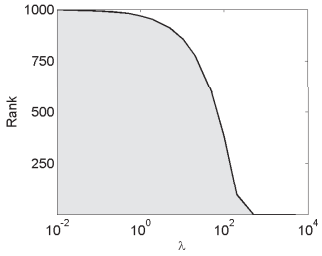


Figure 1: Region of equivalence between (14) and (15) on proximal mapping (4) with Y being a 1000×1000 random matrix and $p = 1/2$. When (15) is initialized in the white area, it is equivalent to the result obtained with (14) (black line). When the rank is known a priori, better reconstruction results can be found in the grey area, by using the factorization formulation (Cabral et al. 2013).

When ensuring condition (a), the operator $\|\cdot\|_{S_{p_i}}^{p_i}$ becomes convex as mentioned before. In the later experiments, we will show that such convexity can be employed for acceleration. When ensuring condition (b), the operator becomes differentiable. In other words, we transform the original *non-smooth* function into a *smooth* one. Thus it is possible to utilize some gradient-based methods, which give much freedom to the optimization.

By substituting the spectral regularization in (1) with the Schatten-p norm, we have

$$\min_X F(X) = \min_X f(X) + \frac{\lambda}{p}\|X\|_{S_p}^p. \quad (14)$$

According to Corollary 1, by rewriting X in the multi-linear form as (13), the above problem becomes

$$\min_{\mathcal{X}} F(\mathcal{X}) = \min_{X_i,\ i=1,\ldots,I} f\left(\prod_{i=1}^I X_i\right) + \sum_{i=1}^{I} \frac{\lambda}{p_i}\|X_i\|_{S_{p_i}}^{p_i}, \quad (15)$$

where $1/p = \sum_1^I 1/p_i$ and X_i is defined as Corollary 1. $\mathcal{X} = (X_1, X_2, \ldots, X_I)$ denotes the set of all unknown X_i's. By mild modification on (Mazumder, Hastie, and Tibshirani 2010)[Theorem 3] for the Bi-Frobenius norm surrogate, we have the following connections between (14) and (15):

Theorem 2. *Suppose $\hat{X}^*$ is a solution to* (14)*, and let r^* be its rank. If $d \geq r^*$, the solutions of 14 and 15 are equivalent. For any solution $\hat{\mathcal{X}}$ to* (15)*, $\prod_{i=1}^I \hat{X}_i$ is a solution to* (14)*. On the other hand, the SVD of $\hat{X}^* = \hat{U}_X^*\hat{\Sigma}_X^*\hat{V}_X^{*T}$ provides one such solution to* (15) *with $\hat{X}_1 = \hat{U}_X^*\hat{\Sigma}_X^{*p/p_1}$, $\hat{X}_i = \hat{\Sigma}_X^{*p/p_i}, i = 2, \ldots, I-1$, and $\hat{X}_I = \hat{\Sigma}_X^{*p/p_I}\hat{V}_X^{*T}$.*

[1] In fact, we have tried to extend the proof in Shang, Liu, and Cheng (2016a,b) to the general case and found that it needs to ensure one of p_1 or p_2 to be greater than 1, which is less general compared with Theorem 1.

[2] All proofs in this paper are in Supplementary Material, which is available at https://arxiv.org/abs/1611.08372

[3] When we were preparing the camera ready version of this paper, Shang et al. told us that they also got the same result independently (Shang, Liu, and Cheng 2016c). But their proof still followed that in Shang, Liu, and Cheng (2016a,b).

The matrix factorization formulation (15) defines a bi-parameterized family of models indexed by (d, λ), while the spectral penalty formulation (14) defines a uni-parameterized family. As Theorem 2 indicates, this family is a special path in the two-dimensional grid of solutions $\hat{\mathcal{X}}_{d,\lambda}$. Figure 1 shows the relationship. In real applications, it often occurs that the intrinsic rank $r \ll \min\{m, n\}$. When r is unknown, we may overestimate $d \geq r$, while $d \ll \min\{m, n\}$ still holds. When r is known a priori, e.g., Structure from Motion in computer vision, better reconstruction results can be found by setting $d = r$ in (15) than using (14) (Cabral et al. 2013). In addition, the factorization formulation requires far less memory on the unknowns ($\mathcal{O}(d(m+n)) \ll \mathcal{O}(mn)$). It also avoids the SVD computation on the full matrix, whose cost is as large as $\mathcal{O}(\min\{m, n\}mn)$. For $0 < p < 1$, (14) is a non-smooth and non-convex problem. In contrast, by choosing some appropriate p_i's as discussed earlier, (15) can be smooth[4] (although still non-convex) and hold some good properties for acceleration. In conclusion, it is preferable to model the Schatten-p norm based problem as (15) instead of (14).

Optimization on Matrix Completion

In this section, as a concrete example we consider solving the matrix completion problem. Then problem (15) can be written as follows:

$$\min_{\mathcal{X}} F(\mathcal{X}) = \min_{X_i,\, i=1,\ldots,I} \frac{1}{2} \left\| W \odot \left(M - \prod_{i=1}^{I} X_i \right) \right\|_F^2 + \sum_{i=1}^{I} \frac{\lambda}{p_i} \|X_i\|_{S_{p_i}}^{p_i}, \tag{16}$$

which is a non-convex problem, where $M \in \mathbb{R}^{m \times n}$ is the low rank measurement matrix. W is a 0-1 binary mask with the same size as M. The entry value of W being 0 means that the component at the same position in M is missing, and 1 otherwise. The operator $\odot$ is the Hadamard element-wise product. By utilizing the smoothness of the first part in (16), we use the PALM proposed in (Bolte, Sabach, and Teboulle 2014), which can also be regarded as block coordinate descent (BCD) of Gauss-Seidel type. At each iteration, PALM minimizes F cyclically over each of $X_1, \ldots, X_I$ while fixing the remaining blocks at their last updated values. Let X_i^k denote the value of X_i at the k-th update, $A_{-i}^k = X_1^k \cdots X_{i-1}^k$ and $A_{+i}^{k-1} = X_{i+1}^{k-1} \cdots X_I^{k-1}$. Then we can represent the first part of each subproblem minimizing F as follows:

$$f_i^k(X_i) = \frac{1}{2} \|W \odot (M - A_{-i}^k X_i A_{+i}^{k-1})\|_F^2. \tag{17}$$

By further linearizing $f_i^k(X_i)$ at some point $\hat{X}_i^{k-1}$, the subproblem becomes

$$\min_{X_i} \left\langle \nabla f_i^k(\hat{X}_i^{k-1}), X_i - \hat{X}_i^{k-1} \right\rangle + \frac{L_i^{k-1}}{2} \|X_i - \hat{X}_i^{k-1}\|_F^2 + \frac{\lambda}{p_i} \|X_i\|_{S_{p_i}}^{p_i}, \tag{18}$$

[4] If $f(\cdot)$ is also smooth.

Algorithm 1 Minimizing $F(\mathcal{X})$ in (16) with accelerated PALM.

Input: $k = 1$ and $X_i^{-1} = X_i^0, i = 1, \ldots, I$.

```
while not converged do
    for i = 1, 2, ..., I do
        Compute L_i^{k-1} as (20) and w_i^{k-1} as (22).
        Update X̂_i^{k-1} as (21).
        Update X_i^k by solving (18).
    end for
    if F(X^k) ≥ F(X^{k-1}) then
        Reupdate X_i^k by solving (18) with X̂_i^{k-1} = X_i^{k-1}, i = 1, ..., I.
    end if
    k = k + 1
end while
```

Output: The factors $(X_1, \ldots, X_I)$

which can be formulated as the proximal mapping (4) and solved efficiently or even in closed-form solution for specific p_i values. $\nabla f_i^k(\hat{X}_i^{k-1})$ is the gradient of $f_i^k(X_i)$ at $\hat{X}_i^{k-1}$:

$$\nabla f_i^k(\hat{X}_i^{k-1}) = (A_{-i}^k)^T \left(W \odot (M - A_{-i}^k \hat{X}_i^{k-1} A_{+i}^{k-1}) \right) (A_{+i}^{k-1})^T. \tag{19}$$

L_i^{k-1} is the Lipschitz constant of $\nabla f_i^k(X_i)$:

$$L_i^{k-1} = \max\{\|A_{-i}^k\|_2^2 \|A_{+i}^{k-1}\|_2^2, \epsilon\}, \tag{20}$$

where $\|A\|_2$ denotes the largest singular value of A and $\epsilon > 0$ is some small constant ensuring that L_i^{k-1} is bounded away from 0 for convergence. For the basic version of PALM, we let $\hat{X}_i^{k-1} = X_i^{k-1}$. When $p_i \geq 1$, the linearized subproblem (18) becomes convex. The acceleration technique proposed by (Xu and Yin 2013) can then be adopted, where $\hat{X}_i^{k-1}$ is extrapolated as

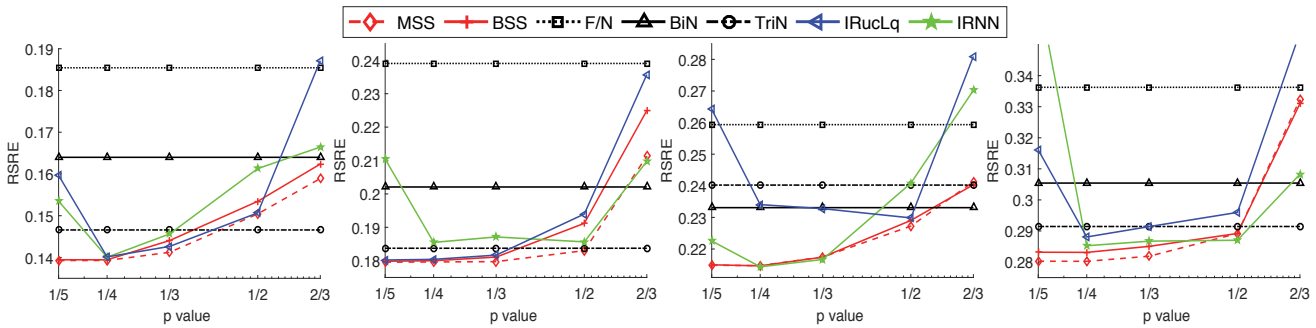
$$\hat{X}_i^{k-1} = X_i^{k-1} + w_i^{k-1}(X_i^{k-1} - X_i^{k-2}), \tag{21}$$

where w_i^{k-1} is defined as

$$w_i^{k-1} = \min\left\{ \frac{t_{k-1} - 1}{t_k}, 0.9999 \sqrt{\frac{L_i^{k-2}}{L_i^{k-1}}} \right\} \tag{22}$$

with $t_0 = 1$ and $t_k = \left(1 + \sqrt{1 + 4t_{k-1}^2}\right)/2$.

For better reference, we summarize the algorithm for minimizing $F(\mathcal{X})$ in (16) in Algorithm 1 . The running time is dominated by performing matrix multiplications. The total time complexity is $\mathcal{O}(mnd)$, where $d \ll \min\{m, n\}$. We terminate the algorithm when all the magnitudes of gradients (19) over the Lipschitz constants in (20) are below a threshold. For further acceleration, we adopt the backtracking continuation technique (Toh and Yun 2010) to find a proper local Lipschitz constant instead of the global ones as shown in (20). Namely, we initially underestimate L_i^{k-1} by multiplying a factor $\rho < 1$. We then increase ρ gradually along the iteration until it approaches the upper bound, i.e., 1. As pointed out



(a) $\sigma = 0.3, o\% = 20\%$ (b) $\sigma = 0.4, o\% = 20\%$ (c) $\sigma = 0.3, o\% = 15\%$ (d) $\sigma = 0.4, o\% = 15\%$

Figure 2: Synthetic experiments on the Schatten-p norm regularized algorithms, i.e., our MSS and BSS, F/N (Shang, Liu, and Cheng 2016a), BiN (Shang, Liu, and Cheng 2016a), TriN (Shang, Liu, and Cheng 2016b), IRucLq (Lai, Xu, and Yin 2013), and IRNN (Lu et al. 2014), with varying p values, noise magnitude σ, and observed data percentage $o\%$. As F/N, BiN, and TriN are only for $p = 2/3$, $p = 1/2$, and $p = 1/3$, respectively, we plot them across different p values for comparison with others.

in (Xu et al. 2016), such a technique can further improve the quality of the solution for non-convex optimization. Note that the inner factors are of much smaller size while taking almost the same updating cost as the side factors, i.e., X_1 and X_I. When the number of factors I becomes big, the redundancy on the inner factors may weaken the effectiveness on decreasing the objective. Thus we modify Algorithm 1 by updating only one or two inner factors in a shuffling order in each cycle (lines $2-6$). The convergence result remains unchanged, which is as follows:

Theorem 3. *(Sequence Convergence) Let* $\{(X_1^k, \ldots, X_I^k)\}$ *be a sequence generated by Algorithm 1 with all* $p_i > 0$ *being rational, then it is a Cauchy sequence and converges to a critical point of* (16).

Note that the sequence convergence is stronger than those in the existing general Schatten-p solvers, e.g., IRucLq (Lai, Xu, and Yin 2013) and IRNN (Lu et al. 2014), where they only prove that any limit point is a stationary point, which is *subsequence* convergent.

Experiments

We test our framework in two variants: Bi-Schatten-p norm Surrogate (BSS) with $p_1 = p_2 = 2p$ and Multi-Schatten-p norm Surrogate (MSS) with $p_i \geq 1$ and extrapolation as (21). We do not use extrapolation for BSS when the p_1 and $p_2 \geq 1$ in order to test the effectiveness of this technique by comparing with MSS. We compare them with several state-of-the-art algorithms for Schatten-p norm regularized problems, i.e., F/N ($p = 2/3$) (Shang, Liu, and Cheng 2016a), BiN ($p = 1/2$) (Shang, Liu, and Cheng 2016a), TriN ($p = 1/3$) (Shang, Liu, and Cheng 2016b), IRNN[5] (Lu et al. 2014) and IRucLq [6] (Lai, Xu, and Yin 2013). As IRNN and IRucLq are not suitable for large-scale problems due to high computing costs and memory requirements, we also include the state-of-the-art solvers for matrix completion, i.e., NNLS[7] (Toh and Yun 2010), LMaFit[8] (Wen, Yin, and Zhang 2012), and Soft-ALS[9] (Hastie et al. 2015) on real datasets. We implement F/N, BiN, and TriN by ourselves, which also use PALM. Their main differences from ours are that they do not employ either the continuation technique or the extrapolated-based acceleration. We initialize all algorithms with the same random matrices. All the codes are run in Matlab on a desktop PC with a 3.4 GHz CPU and 20 GB RAM.

Synthetic Data

We first generate synthetic data matrices $M = U_0 V_0^T$, where $U_0 \in \mathbb{R}^{100\times 5}$ and $V_0 \in \mathbb{R}^{100\times 5}$. The entries of U_0 and V_0 are sampled i.i.d. from the standard Gaussian distribution $\mathcal{N}(0,1)$. Then $\mathcal{N}(0,\sigma)$ Gaussian noise is added independently to every entry of M and a portion ($\%$) of them are picked out uniformly as the observed data. We conduct experiments by choosing different p values, i.e., $p = 1/5, 1/4, 1/3, 1/2$, and $2/3$, varying the noise magnitude $\sigma = 0.3, 0.4$ and the observed data percentage $o\% = 15\%, 20\%$. At each combination of these hyper parameters, we repeat the experiments 30 times. All the compared algorithms use the same data in each trial. For MSS with multiple factors, we set all $p_i = 1$ when $p = 1/5, 1/4$, or $1/3$ and all $p_i = 2$ when $p = 1/2$ or $2/3$[10]. To cope with different Schatten-p norm regularizers, the regularization parameter λ of all compared algorithms is tuned in the range $[1, 20]$. As done in (Lai, Xu, and Yin 2013), d is overestimated as $3 \times 5 = 15$.

We use the relative square root error (RSRE), i.e., $\|X - U_0 V_0^T\|_F / \|U_0 V_0^T\|_F$, to evaluate the performance of recovery. The average values over all trails are shown in Fig. 2. Among all the compared algorithms, our MSS achieves the most plausible performance, which reports the least or the second least RSREs in most cases. Besides, MSS shows less sensitivity to the change of p values than other general

[5] https://sites.google.com/site/canyilu/

[6] http://www.math.ucla.edu/~wotaoyin/

[7] http://www.math.nus.edu.sg/~mattohkc/NNLS.html

[8] http://lmafit.blogs.rice.edu/

[9] http://cran.r-project.org/web/packages/softImpute/

[10] In fact, we have tested different numbers of factors (corresponding to different p_i's) on the same p while find no difference in the measured performance.

Schatten-p norm solvers, i.e., IRNN, IRucLq, and BSS. At $p = 2/3$ ($1/2$ or $1/3$), F/N (BiN or TriN) is inferior to our BSS and MSS, which confirms the effectiveness of the continuation (and extrapolation for MSS) technique. Across all the four subfigures, all the general Schatten-p norm solvers achieve reasonably good performance at $p = 1/4$. So in the rest of this paper, we only report the results with $p = 1/4$.

Real Data

We conduct experiments on three real-world recommendation system datasets: MovieLens 1M, MovieLens 10M[11], and Netflix (SIGKDD 2007). The corresponding observed matrices are of size 6040×3449 with $o\% = 4.80\%$, 69878×10677 with $o\% = 1.34\%$, and 480189×17770 with $o\% = 1.18\%$, respectively. Here we fix the regularization $\lambda = 200$ and tune it for other algorithms in the range $[1, 200]$. Following the experimental setup in (Shang, Liu, and Cheng 2016a), we randomly pick out 80% of the observed entries as the training data and use the remaining 20% for testing. The root mean squared errors (RMSEs) on the test set T, i.e., $\sqrt{\sum_{(i,j)\in T}(X_{ij} - M_{ij})^2/|T|}$, are measured during the computation. Due to the non-convexity of the Schatten-p norm solvers, we repeat them with different random initializations on MovieLens 1M and 10M, while find no significant difference. Results ($d = 10$ for the factorization formulation) of all compared algorithms are shown in the first row of Fig. 3. As IRNN and IRucLq are slow and require large memory, we do not apply them to MovieLen 10M and Netflix. From the figure, we can see some obvious gaps in the time vs. testing RMSE curves between our BSS and MSS. This is mainly caused by the extrapolation technique employed by MSS. The testing RMSE of LMaFit increases after several iterations on MovieLens 1M, which may be caused by the intrinsic unregularized model. Among all compared algorithms, our MSS achieves the best performance across all the three data sets. It reaches the smallest testing RMSEs with the least time in less than 500 iterations. We also conduct experiments to test the sensitivity with respect to different estimated ranks d. The results are shown in the second row, where our MSS shows an apparent superiority over others, especially on large-scale datasets, i.e., MovienLen 10M and Netflix.

Conclusions

In this paper, we propose a unified surrogate for the Schatten-p norm with two factor matrix norms. We further extend it to multiple factor matrices and show that all the factor norms can be *convex and smooth* for any $p > 0$. In contrast, the original Schatten-p norm for $0 < p < 1$ is non-convex and non-smooth. We establish equivalence between the surrogate formulation and the original problem and show that the former should be preferred in practice. As an example we conduct experiments on matrix completion. By utilizing the convexity of the factor matrix norms, our accelerated PALM achieves the state-of-the-art performance. Its sequence convergence is also established.

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

Acknowledgements

Zhouchen Lin is supported by National Basic Research Program of China (973 Program) (grant no. 2015CB352502), National Natural Science Foundation (NSF) of China (grant nos. 61625301 and 61231002), and Qualcomm.

References

Abernethy, J.; Bach, F.; Evgeniou, T.; and Vert, J.-P. 2009. A new approach to collaborative filtering: Operator estimation with spectral regularization. *The Journal of Machine Learning Research* 10:803–826.

Bolte, J.; Sabach, S.; and Teboulle, M. 2014. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming* 146(1-2):459–494.

Cabral, R.; Torre, F. D. L.; Costeira, J. P.; and Bernardino, A. 2013. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *ICCV*, 2488–2495. IEEE.

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

Candès, E. J., and Tao, T. 2010. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on* 56(5):2053–2080.

Donoho, D. L.; Johnstone, I. M.; Kerkyacharian, G.; and Picard, D. 1995. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)* 301–369.

Dudik, M.; Harchaoui, Z.; and Malick, J. 2012. Lifted coordinate descent for learning with trace-norm regularization. In *AISTATS*, volume 22, 327–336.

Fan, J., and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456):1348–1360.

Hastie, T.; Mazumder, R.; Lee, J. D.; and Zadeh, R. 2015. Matrix completion and low-rank SVD via fast alternating least squares. *Journal of Machine Learning Research* 16:3367–3402.

Horn, R. A., and Johnson, C. R. 1991. Topics in matrix analysis. *Cambridge UP, New York*.

Lai, M.-J.; Xu, Y.; and Yin, W. 2013. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM Journal on Numerical Analysis* 51(2):927–957.

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

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

Lu, C.; Zhu, C.; Xu, C.; Yan, S.; and Lin, Z. 2015. Generalized singular value thresholding. In *AAAI*, 1805–1811.

Mazumder, R.; Hastie, T.; and Tibshirani, R. 2010. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research* 11:2287–2322.

Nie, F.; Huang, H.; and Ding, C. H. 2012. Low-rank matrix recovery via efficient Schatten p-norm minimization. In *AAAI*.

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.

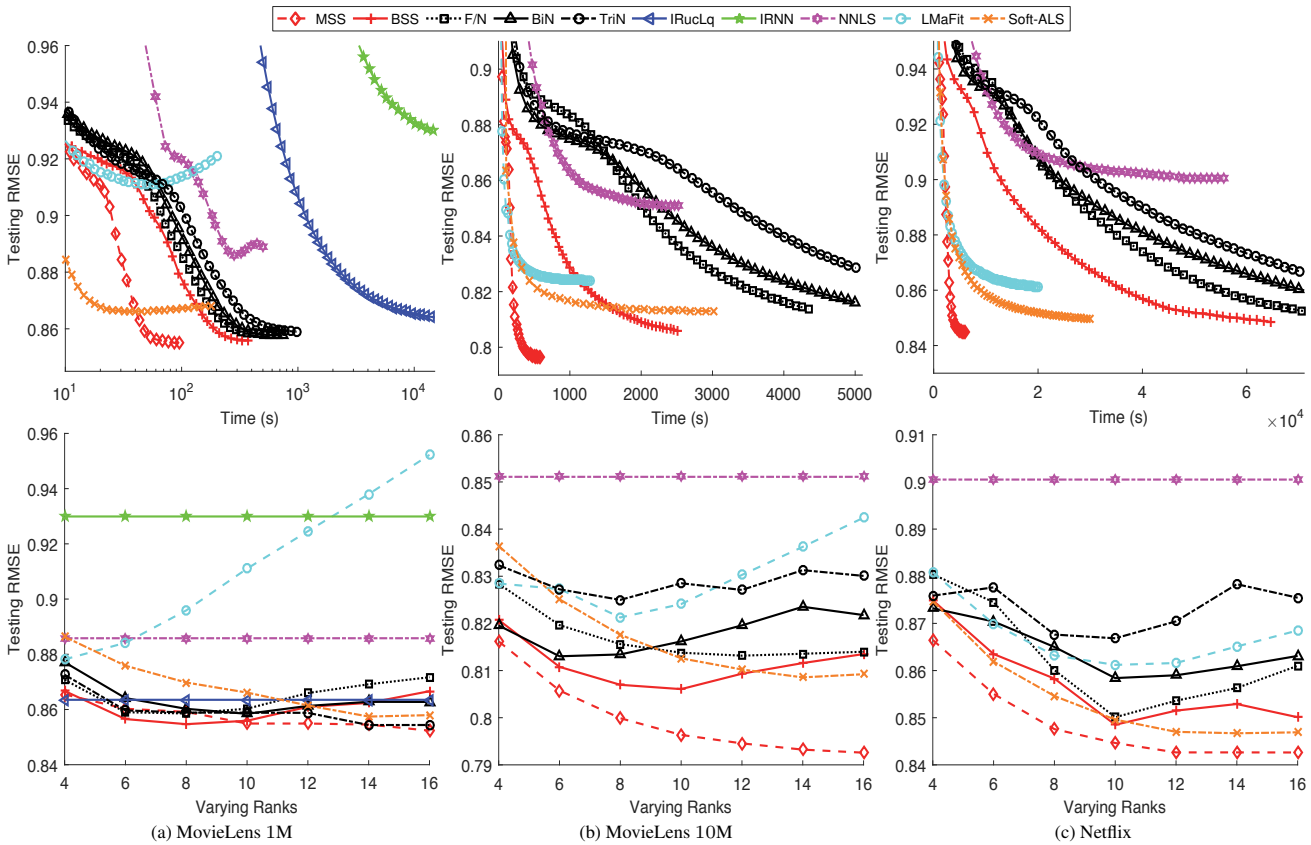


Figure 3: Matrix completion on recommendation system data sets using our MSS and BSS, F/N (Shang, Liu, and Cheng 2016a), BiN (Shang, Liu, and Cheng 2016a), TriN (Shang, Liu, and Cheng 2016b), IRucLq (Lai, Xu, and Yin 2013), IRNN (Lu et al. 2014), NNLS (Toh and Yun 2010), LMaFit (Wen, Yin, and Zhang 2012), Soft-ALS (Hastie et al. 2015). In the first row, we depict the testing RMSE along the executing time ($d = 10$ for the factorization formulation). Note that the subfigure (a) is in log-10 scale in the time axis. IRucLq and IRNN are not included in the last two datasets due to their unaffordable computing time and high memory requirements. The second row tests the sensitivity w.r.t. varying ranks d.

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

Shang, F.; Liu, Y.; and Cheng, J. 2016b. Tractable and scalable Schatten quasi-norm approximations for rank minimization. In *AISTATS*, 620–629.

Shang, F.; Liu, Y.; and Cheng, J. 2016c. Unified scalable equivalent formulations for Schatten quasi-norms. *arXiv preprint arXiv:1606.00668*.

SIGKDD, A. 2007. Netflix. In *Proceedings of KDD cup and workshop*.

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

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.

Watson, G. A. 1992. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications* 170:33–45.

Wen, Z.; Yin, W.; and Zhang, Y. 2012. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation* 4(4):333–361.

Xu, Y., and Yin, W. 2013. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences* 6(3):1758–1789.

Xu, C.; Lin, Z.; Zhao, Z.; and Zha, H. 2016. Relaxed majorization-minimization for non-smooth and non-convex optimization. In *AAAI*, 812–818.

Zhang, M.; Huang, Z.-H.; and Zhang, Y. 2013. Restricted-isometry properties of nonconvex matrix recovery. *Information Theory, IEEE Transactions on* 59(7):4316–4323.

Zhang, C.-H. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 894–942.

Zuo, W.; Meng, D.; Zhang, L.; Feng, X.; and Zhang, D. 2013. A generalized iterated shrinkage algorithm for non-convex sparse coding. In *ICCV*, 217–224.