An Accelerated Gradient Method for Trace Norm Minimization

Shuiwang Ji SHUIWANG.JI@ASU.EDU
Jieping Ye JIEPING.YE@ASU.EDU
Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287, USA

Abstract

We consider the minimization of a smooth loss function regularized by the trace norm of the matrix variable. Such formulation finds applications in many machine learning tasks including multi-task learning, matrix classification, and matrix completion. The standard semidefinite programming formulation for this problem is computationally expensive. In addition, due to the non-smooth nature of the trace norm, the optimal first-order black-box method for solving such class of problems converges as $O(\frac{1}{\sqrt{k}})$, where k is the iteration counter. In this paper, we exploit the special structure of the trace norm, based on which we propose an extended gradient algorithm that converges as $O(\frac{1}{k})$. We further propose an accelerated gradient algorithm, which achieves the optimal convergence rate of $O(\frac{1}{k^2})$ for smooth problems. Experiments on multi-task learning problems demonstrate the efficiency of the proposed algorithms.

1. Introduction

The problem of minimizing the rank of a matrix variable subject to certain constraints arises in many fields including machine learning, automatic control, and image compression. For example, in collaborative filtering we are given a partially filled rating matrix and the task is to predict the missing entries. Since it is commonly believed that only a few factors contribute to an individual's tastes, it is natural to approximate the given rating matrix by a low-rank matrix. However, the matrix rank minimization problem is NP-hard in general due to the combinatorial nature of the rank function. A commonly-used convex relaxation of the rank function is the trace norm (nuclear norm) (Fazel et al., 2001), defined as the sum of the singular values of the matrix, since it is the convex envelope of the rank function over the unit ball of spectral norm. A number of recent work has shown that the low rank solution can be recovered exactly via minimizing the trace norm under certain conditions (Recht et al., 2008a; Recht et al., 2008b; Candés & Recht, 2008).

In practice, the trace norm relaxation has been shown to yield low-rank solutions and it has been used widely in many scenarios. In (Srebro et al., 2005; Rennie & Srebro, 2005; Weimer et al., 2008a; Cai et al., 2008; Ma et al., 2008) the matrix completion problem was formulated as a trace norm minimization problem. In problems where multiple related tasks are learned simultaneously, the models for different tasks can be constrained to share certain information. Recently, this constraint has been expressed as the trace norm regularization on the weight matrix in the context of multi-task learning (Abernethy et al., 2006; Argyriou et al., 2008; Abernethy et al., 2009; Obozinski et al., 2009), multi-class classification (Amit et al., 2007), and multivariate linear regression (Yuan et al., 2007; Lu et al., 2008). For two-dimensional data such as images, the matrix classification formulation (Tomioka & Aihara, 2007; Bach, 2008) applies a weight matrix, regularized by its trace norm, on the data. It was shown (Tomioka & Aihara, 2007) that such formulation leads to improved performance over conventional methods.

A practical challenge in employing the trace norm regularization is to develop efficient algorithms to solve the resulting non-smooth optimization problems. It is well-known that the trace norm minimization problem can be formulated as a semidefinite program (Fazel et al., 2001; Srebro et al., 2005). However, such formulation is computationally expensive. To overcome this limitation, a number of algorithms have been developed recently (Rennie & Srebro, 2005; Weimer et al., 2008a; Weimer et al., 2008b; Cai et al., 2008; Ma et al., 2008). In these algorithms some form of approximation is usually employed to deal with the non-smooth trace norm term. However, a fast global convergence

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

rate for these algorithms is difficult to guarantee.

Due to the non-smooth nature of the trace norm, a simple approach to solve these problems is the subgradient method (Bertsekas, 1999; Nesterov, 2003), which converges as $O(\frac{1}{\sqrt{k}})$ where k is the iteration counter. It is known from the complexity theory of convex optimization (Nemirovsky & Yudin, 1983; Nesterov, 2003) that this convergence rate is already optimal for non-smooth optimization under the first-order black-box model, where only the function values and first-order derivatives are used.

In this paper we propose efficient algorithms with fast global convergence rates to solve trace norm regularized problems. Specifically, we show that by exploiting the special structure of the trace norm, the classical gradient method for smooth problems can be adapted to solve the trace norm regularized non-smooth problems. This results in an extended gradient algorithm with the same convergence rate of $O(\frac{1}{k})$ as that for smooth problems. Following the Nesterov's method for accelerating the gradient method (Nesterov, 1983; Nesterov, 2003), we show that the extended gradient algorithm can be further accelerated to converge as $O(\frac{1}{k^2})$, which is the optimal convergence rate for smooth problems. Hence, the non-smoothness effect of the trace norm regularization is effectively removed. The proposed algorithms extend the algorithms in (Nesterov, 2007; Tseng, 2008; Beck & Teboulle, 2009) to the matrix case. Experiments on multi-task learning problems demonstrate the efficiency of the proposed algorithms in comparison with existing ones. Note that while the present paper was under review, we became aware of a recent preprint by Toh and Yun (2009) who independently developed an algorithm that is similar to ours.

2. Problem Formulation

In this paper we consider the following problem:

$$\min_W F(W) = f(W) + \lambda||W||_* \tag{1}$$

where $W \in \mathbb{R}^{m\times n}$ is the decision matrix, $f(\cdot)$ represents the loss induced by some convex smooth (differentiable) loss function $\ell(\cdot,\cdot)$, and $||\cdot||_*$ denotes the trace norm defined as the sum of the singular values. We assume that the gradient of $f(\cdot)$, denoted as $\bigtriangledown f(\cdot)$, is Lipschitz continuous with constant L, i.e.,

$$||\bigtriangledown f(X) - \bigtriangledown f(Y)||_F \leq L||X-Y||_F, \forall X, Y \in \mathbb{R}^{m\times n},$$

where $||\cdot||_F$ denotes the Frobenius norm. Such formulation arises in many machine learning tasks such as in multi-task learning, matrix classification, and matrix completion problems.

- **Multi-task learning** (Argyriou et al., 2008): $f(W) = \sum_{i=1}^{n}\sum_{j=1}^{s_i}\ell(y_i^j, w_i^T x_i^j)$, where n is the number of tasks, $(x_i^j, y_i^j) \in \mathbb{R}^m \times \mathbb{R}$ is the jth sample in the ith task, s_i is the number of samples in the ith task, and $W = [w_1, \cdots, w_n] \in \mathbb{R}^{m\times n}$.

- **Matrix classification** (Tomioka & Aihara, 2007; Bach, 2008): $f(W) = \sum_{i=1}^{s}\ell(y_i, \mathrm{Tr}(W^T X_i))$, where $(X_i, y_i) \in \mathbb{R}^{m\times n} \times \mathbb{R}$ is the ith sample.

- **Matrix completion** (Srebro et al., 2005; Candés & Recht, 2008; Recht et al., 2008a; Ma et al., 2008): $f(W) = \sum_{(i,j)\in\Omega}\ell(M_{ij}, W_{ij})$, where $M \in \mathbb{R}^{m\times n}$ is the partially observed matrix with the entries in Ω being observed.

Since the trace norm term in the objective function in Eq. (1) is non-smooth, a natural approach for solving this problem is the subgradient method in which a sequence of approximate solutions are generated as

$$W_k = W_{k-1} - \frac{1}{t_k}F'(W_{k-1}), \tag{2}$$

where W_k is the approximate solution at the kth iteration, $\frac{1}{t_k}$ is the step size, and $F'(W) \in \partial F(W)$ is the subgradient of $F(W)$ at W and $\partial F(W)$ denotes the subdifferential (Bertsekas, 1999; Nesterov, 2003) of $F(W)$ at W. It is known (Nesterov, 2003) that the subgradient method converges as $O(\frac{1}{\sqrt{k}})$, i.e.,

$$F(W_k) - F(W^*) \leq c\frac{1}{\sqrt{k}}, \tag{3}$$

for some constant c, where $W^* = \arg\min_W F(W)$.

It is known from the complexity theory of convex optimization (Nemirovsky & Yudin, 1983; Nesterov, 2003) that this convergence rate is already optimal for non-smooth problems under the first-order black-box model. Hence, the convergence rate cannot be improved if a black-box model, which does not exploit any special structure of the objective function, is employed. We show in the following that by exploiting the structure of the trace norm, its non-smoothness can be effectively overcome and the convergence rate of the algorithm for solving the trace norm regularized problem in Eq. (1) can be improved significantly.

3. An Extended Gradient Method

First, consider the minimization of the smooth loss function without the trace norm regularization:

$$\min_W f(W). \tag{4}$$

It is known (Bertsekas, 1999) that the gradient step

$$W_k = W_{k-1} - \frac{1}{t_k} \bigtriangledown f(W_{k-1}) \tag{5}$$

for solving this smooth problem can be reformulated equivalently as a proximal regularization of the linearized function $f(W)$ at W_{k-1} as

$$W_k = \arg\min_W P_{t_k}(W, W_{k-1}), \tag{6}$$

where

$$\begin{aligned} P_{t_k}(W, W_{k-1}) \;=\; & f(W_{k-1}) + \langle W - W_{k-1}, \bigtriangledown f(W_{k-1}) \rangle \\ & + \frac{t_k}{2} ||W - W_{k-1}||_F^2, \end{aligned} \tag{7}$$

and $\langle A, B \rangle = \mathrm{Tr}(A^T B)$ denotes the matrix inner product. It has been shown (Nesterov, 2003) that the convergence rate of this algorithm is $O(\frac{1}{k})$. Note that the function P_{t_k} defined in Eq. (7) can be considered as a linear approximation of the function f at point W_{k-1} regularized by a quadratic proximal term.

Based on this equivalence relationship, we propose to solve the optimization problem in Eq. (1) by the following iterative step:

$$W_k = \arg\min_W Q_{t_k}(W, W_{k-1}) \triangleq P_{t_k}(W, W_{k-1}) + \lambda ||W||_*. \tag{8}$$

A key motivation for this formulation is that if the optimization problem in Eq. (8) can be easily solved by exploiting the structure of the trace norm, the convergence rate of the resulting algorithm is expected to be the same as that of gradient method, since no approximation on the non-smooth term is employed.

By ignoring terms that do not depend on W, the objective in Eq. (8) can be expressed equivalently as

$$\frac{t_k}{2} \left\| W - \left(W_{k-1} - \frac{1}{t_k} \bigtriangledown f(W_{k-1}) \right) \right\|_F^2 + \lambda \, ||W||_* \,. \tag{9}$$

It turns out that the minimization of the objective in Eq. (9) can be solved by first computing the singular value decomposition (SVD) of $W_{k-1} - \frac{1}{t_k} \bigtriangledown f(W_{k-1})$ and then applying some soft-thresholding on the singular values. This is summarized in the following theorem (Cai et al., 2008).

Theorem 3.1. *Let $C \in \mathbb{R}^{m \times n}$ and let $C = U \Sigma V^T$ be the SVD of C where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ have orthonormal columns, $\Sigma \in \mathbb{R}^{r \times r}$ is diagonal, and $r = rank(C)$. Then*

$$\mathcal{T}_\lambda(C) \equiv \arg\min_W \left\{ \frac{1}{2} ||W - C||_F^2 + \lambda ||W||_* \right\} \tag{10}$$

is given by $\mathcal{T}_\lambda(C) = U \Sigma_\lambda V^T$, where Σ_λ is diagonal with $(\Sigma_\lambda)_{ii} = \max\{0, \Sigma_{ii} - \lambda\}$.

The proof of this theorem is in the Appendix.

The above discussion shows that the problem in Eq. (8) can be readily solved by SVD. Furthermore, we show in the following that if the step size $\frac{1}{t_k}$ of the gradient method is chosen properly, we can achieve the same convergence rate as in the smooth case, i.e., $O(\frac{1}{k})$, despite the presence of the non-smooth trace norm regularization.

3.1. Step Size Estimation

To choose an appropriate step size we impose a condition on the relationship between the function values of F and Q_{t_k} at a certain point in Lemma 3.1. We show in Theorem 3.2 below that once this condition is satisfied at each step by choosing an appropriate step size, the convergence rate of the resulting sequence can be guaranteed.

Lemma 3.1. *Let*

$$p_\mu(Y) = \arg\min_X Q_\mu(X, Y) \tag{11}$$

where Q is defined in Eq. (8). Assume the following inequality holds:

$$F(p_\mu(Y)) \leq Q_\mu(p_\mu(Y), Y). \tag{12}$$

Then for any $X \in \mathbb{R}^{m \times n}$ we have

$$\begin{aligned} F(X) - F(p_\mu(Y)) \;\geq\; & \frac{\mu}{2} ||p_\mu(Y) - Y||_F^2 \\ & + \mu \langle Y - X, p_\mu(Y) - Y \rangle. \end{aligned} \tag{13}$$

The proof of this lemma is in the Appendix.

At each step of the algorithm we need to find an appropriate value for μ such that $W_k = p_\mu(W_{k-1})$ and the condition

$$F(W_k) \leq Q_\mu(W_k, W_{k-1}) \tag{14}$$

is satisfied. Note that since the gradient of $f(\cdot)$ is Lipschitz continuous with constant L, we have (Nesterov, 2003)

$$f(X) \leq f(Y) + \langle X - Y, \bigtriangledown f(Y) \rangle + \frac{L}{2} ||X - Y||_F^2, \forall X, Y.$$

Hence, when $\mu \geq L$ we have

$$F(p_\mu(Y)) \leq P_\mu(p_\mu(Y), Y) + \lambda ||p_\mu(Y)||_* = Q_\mu(p_\mu(Y), Y).$$

This shows that the condition in Eq. (14) is always satisfied if the update rule

$$W_k = p_L(W_{k-1}) \tag{15}$$

is applied. However, L may not be known or it is expensive to compute in practice. We propose to employ the following step size estimation strategy to ensure the condition in Eq. (14): Given an initial estimate of L as L_0, we increase this estimate with a multiplicative factor $\gamma > 1$ repeatedly until the condition in Eq. (14) is satisfied. This results in the extended gradient method in Algorithm 1 for solving the problem in Eq. (1).

Algorithm 1 Extended Gradient Algorithm

Initialize $L_0, \gamma, W_0 \in \mathbb{R}^{m\times n}$
Iterate:

1. Set $\bar{L} = L_{k-1}$
2. While $F(p_{\bar{L}}(W_{k-1})) > Q_{\bar{L}}(p_{\bar{L}}(W_{k-1}), W_{k-1})$, set
$$\bar{L} := \gamma \bar{L}$$
3. Set $L_k = \bar{L}$ and $W_k = p_{L_k}(W_{k-1})$

Since when $L_k \geq L$ the condition in Eq. (14) is always satisfied, we have

$$L_k \leq \gamma L, \ \forall k. \tag{16}$$

Note that the sequence of function values generated by this algorithm is non-increasing as

$$F(W_k) \leq Q_{L_k}(W_k, W_{k-1}) \leq Q_{L_k}(W_{k-1}, W_{k-1}) = F(W_{k-1}).$$

3.2. Convergence Analysis

We show in the following theorem that when the condition in Eq. (14) is satisfied at each iteration, the extended gradient algorithm converges as $O(\frac{1}{k})$.

Theorem 3.2. *Let $\{W_k\}$ be the sequence generated by Algorithm 1. Then for any $k \geq 1$ we have*

$$F(W_k) - F(W^*) \leq \frac{\gamma L ||W_0 - W^*||_F^2}{2k}, \tag{17}$$

where $W^* = \arg\min_W F(W)$.

The proof of this theorem is in the Appendix.

4. An Accelerated Gradient Method

It is known (Nesterov, 1983; Nesterov, 2003) that when the objective function is smooth, the gradient method can be accelerated to achieve the optimal convergence rate of $O(\frac{1}{k^2})$. It was shown recently (Nesterov, 2007; Tseng, 2008; Beck & Teboulle, 2009) that a similar scheme can be applied to accelerate optimization problems where the objective function consists of a smooth part and a non-smooth part provided that the non-smooth part is "simple". In particular, it was shown that the ℓ_1-norm regularized problems can be accelerated even though they are not smooth. In this section we show that the extended gradient method in Algorithm 1 can also be accelerated to achieve the optimal convergence rate of smooth problems even though the trace norm is not smooth. This results in the accelerated gradient method in Algorithm 2.

Algorithm 2 Accelerated Gradient Algorithm

Initialize $L_0, \gamma, W_0 = Z_1 \in \mathbb{R}^{m\times n}, \alpha_1 = 1$
Iterate:

1. Set $\bar{L} = L_{k-1}$
2. While $F(p_{\bar{L}}(Z_{k-1})) > Q_{\bar{L}}(p_{\bar{L}}(Z_{k-1}), Z_{k-1})$, set
$$\bar{L} := \gamma \bar{L}$$
3. Set $L_k = \bar{L}$ and update
$$W_k = p_{L_k}(Z_k)$$
$$\alpha_{k+1} = \frac{1 + \sqrt{1 + 4\alpha_k^2}}{2} \tag{18}$$
$$Z_{k+1} = W_k + \left(\frac{\alpha_k - 1}{\alpha_{k+1}}\right)(W_k - W_{k-1}) \tag{19}$$

4.1. Discussion

In the accelerated gradient method, two sequences $\{W_k\}$ and $\{Z_k\}$ are updated recursively. In particular, W_k is the approximate solution at the kth step and Z_k is called the search point (Nesterov, 1983; Nesterov, 2003), which is constructed as a linear combination of the latest two approximate solutions W_{k-1} and W_{k-2}. The key difference between the extended and the accelerated algorithms is that the gradient step is performed at the current approximate solution W_k in the extended algorithm, while it is performed at the search point Z_k in the accelerated scheme. The idea of constructing the search point is motivated by the investigation of the information-based complexity (Nemirovsky & Yudin, 1983; Nesterov, 2003), which reveals that for smooth problems the convergence rate of the gradient method is not optimal, and thus methods with a faster convergence rate should exist. The derivation of the search point is based on the concept of estimate sequence and more details can be found in (Nesterov, 2003). Note that the sequence α_k can be updated in many ways as long as certain conditions are satisfied (Nesterov, 2003). Indeed, it was shown in (Tseng, 2008) that other schemes of updating α_k

Table 1. Comparison of the three multi-task learning algorithms (EGM, AGM, and MFL) in terms of the computation time (in seconds). In each case, the computation time reported is the time used to train the model for a given parameter value obtained by cross validation, and the averaged training time over ten random trials is reported.

DATA SET	YEAST		LETTERS		DIGITS		DMOZ	
PERCENTAGE	5%	10%	5%	10%	5%	10%	5%	10%
EGM	2.24	3.37	4.74	5.67	62.51	29.59	133.21	146.58
AGM	0.34	0.49	0.62	0.91	2.41	2.39	1.59	1.42
MFL	2.33	17.27	2.49	9.66	15.50	42.64	74.24	31.49

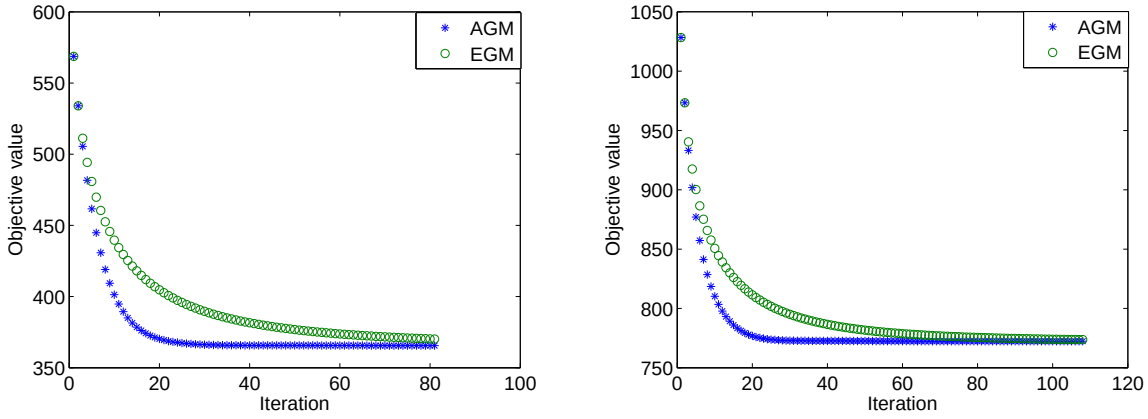

Figure 1. The convergence of EGM and AGM on the *yeast* data set when 5% (left figure) and 10% (right figure) of the data are used for training. On the first data set EGM and AGM take 81 and 1122 iterations, respectively, to converge, while on the second data set they take 108 and 773 iterations, respectively.

can lead to better practical performance, though the theoretical convergence rate remains the same. Note that the sequence of objective values generated by the accelerated scheme may increase. It, however, can be made non-increasing by a simple modification of the algorithm as in (Nesterov, 2005).

4.2. Convergence Analysis

We show in the following that by performing the gradient step at the search point Z_k instead of at the approximate solution W_k, the convergence rate of the gradient method can be accelerated to $O(\frac{1}{k^2})$. This result is summarized in the following theorem.

Theorem 4.1. *Let $\{W_k\}$ and $\{Z_k\}$ be the sequences generated by Algorithm 2. Then for any $k \geq 1$ we have*

$$F(W_k) - F(W^*) \leq \frac{2\gamma L||W_0 - W^*||_F^2}{(k+1)^2}. \qquad (20)$$

The proof of this theorem follows the same strategy as in (Beck & Teboulle, 2009) and it is in the Appendix.

5. Experiments

We evaluate the proposed extended gradient method (EGM) and the accelerated gradient method (AGM) on four multi-task data sets. The *yeast* data set was derived from a yeast gene classification problem consisting of 14 tasks. The *letters* and *digits* are handwritten words and digits data sets (Obozinski et al., 2009), which consist of 8 and 10 tasks, respectively. The *dmoz* is a text categorization data set obtained from DMOZ (http://www.dmoz.org/) in which each of the 10 tasks corresponds to one of the subcategories of the Arts category. For each data set we randomly sample 5% and 10% of the data from each task for training.

To evaluate the efficiency of the proposed formulations, we report the computation time of the multi-task feature learning (MFL) algorithm (Argyriou et al., 2008), as MFL involves a formulation that is equivalent to EGM and AGM. For all methods, we terminate the algorithms when the relative changes in the objective is below 10^{-8}, since the objective values of MFL and EGM/AGM are not directly comparable.

The averaged computation time over ten random trials for each method is reported in Table 1. We can observe that AGM is by far the most efficient method in all cases. The relative efficiency of EGM and AGM differs significantly across data sets, demonstrating that the performance of AGM is very stable for different problems. In order to investigate the convergence behaviors of EGM and AGM, we plot the objective values of these two methods on the *yeast* data set in Figure 1. We can observe that in both cases AGM converges much faster than EGM, especially at early iterations. This is consistent with our theoretical results and confirms that the proposed accelerated scheme can reach the optimal objective value rapidly.

6. Conclusion and Discussion

In this paper we propose efficient algorithms to solve trace norm regularized problems. We show that by exploiting the special structure of the trace norm, the optimal convergence rate of $O(\frac{1}{\sqrt{k}})$ for general non-smooth problems can be improved to $O(\frac{1}{k})$. We further show that this convergence rate can be accelerated to $O(\frac{1}{k^2})$ by employing the Nesterov's method. Experiments on multi-task learning problems demonstrate the efficiency of the proposed algorithms.

As pointed out in the paper, another important application of the trace norm regularization is in matrix completion problems. We plan to apply the proposed formulations to this problem in the future. The proposed algorithms require the computation of SVD, which may be computationally expensive for large-scale problems. We will investigate approximate SVD techniques in the future to further reduce the computational cost.

Appendix

Proof of Theorem 3.1

Proof. Since the objective function in Eq. (10) is strongly convex, a unique solution exists for this problem and hence it remains to show that the solution is $\mathcal{T}_\lambda(C)$. Recall that $Z \in \mathbb{R}^{m\times n}$ is the subgradient of a convex function $h : \mathbb{R}^{m\times n} \to \mathbb{R}$ at X_0 if

$$h(X) \geq h(X_0) + \langle Z, X - X_0\rangle \tag{21}$$

for any X. The set of subgradients of h at X_0 is called the subdifferential of h at X_0 and it is denoted as $\partial h(X_0)$. It is well-known (Nesterov, 2003) that W^* is the optimal solution to the problem in Eq. (10) if and only if $\mathbf{0} \in \mathbb{R}^{m\times n}$ is a subgradient of the objective function at W^*, i.e.,

$$\mathbf{0} \in W^* - C + \lambda\partial||W^*||_*. \tag{22}$$

Let $W = P_1\Lambda P_2^T$ be the SVD of W where $P_1 \in \mathbb{R}^{m\times s}$ and $P_2 \in \mathbb{R}^{n\times s}$ have orthonormal columns, $\Sigma \in \mathbb{R}^{s\times s}$ is diagonal, and $s = \text{rank}(W)$. It can be verified that (Bach, 2008; Recht et al., 2008a)

$$\begin{aligned}\partial||W||_* &= \{P_1P_2^T + S : S \in \mathbb{R}^{m\times n}, P_1^TS = 0,\\ &\quad SP_2 = 0, ||S||_2 \leq 1\},\end{aligned} \tag{23}$$

where $||\cdot||_2$ denotes the spectral norm of a matrix. Decomposing the SVD of C as

$$C = U_0\Sigma_0V_0^T + U_1\Sigma_1V_1^T,$$

where $U_0\Sigma_0V_0^T$ corresponds to the part of SVD with singular values greater than λ. Then we have the SVD of $\mathcal{T}_\lambda(C)$ as

$$\mathcal{T}_\lambda(C) = U_0(\Sigma_0 - \lambda I)V_0^T$$

and thus

$$C - \mathcal{T}_\lambda(C) = \lambda(U_0V_0^T + S)$$

where $S = \frac{1}{\lambda}U_1\Sigma_1V_1^T$. It follows from the facts that $U_0^TS = 0$, $SV_0 = 0$, and $||S||_2 \leq 1$ that

$$C - \mathcal{T}_\lambda(C) \in \lambda\partial||\mathcal{T}_\lambda(C)||_*,$$

which shows that $\mathcal{T}_\lambda(C)$ is an optimal solution. □

Proof of Lemma 3.1

Proof. Since both the loss function f and the trace norm are convex, we have

$$\begin{aligned}f(X) &\geq f(Y) + \langle X - Y, \bigtriangledown f(Y)\rangle,\\ \lambda||X||_* &\geq \lambda||p_\mu(Y)||_* + \lambda\langle X - p_\mu(Y), g(p_\mu(Y))\rangle,\end{aligned}$$

where $g(p_\mu(Y)) \in \partial||p_\mu(Y)||_*$ is the subgradient of the trace norm at $p_\mu(Y)$. Summing up the above two inequalities we obtain that

$$\begin{aligned}F(X) \geq\ & f(Y) + \langle X - Y, \bigtriangledown f(Y)\rangle\\ &+\lambda||p_\mu(Y)||_* + \lambda\langle X - p_\mu(Y), g(p_\mu(Y))\rangle.\end{aligned} \tag{24}$$

By combining the condition in Eq. (14), the result in Eq. (24), and the relation

$$Q_\mu(p_\mu(Y), Y) = P_\mu(p_\mu(Y), Y) + \lambda||p_\mu(Y)||_*,$$

we obtain that

$$\begin{aligned}&F(X) - F(p_\mu(Y)) \geq F(X) - Q_\mu(p_\mu(Y), Y)\\ &\geq \langle X - p_\mu(Y), \bigtriangledown f(Y) + \lambda g(p_\mu(Y))\rangle - \frac{\mu}{2}||p_\mu(Y) - Y||_F^2\\ &= \mu\langle X - p_\mu(Y), Y - p_\mu(Y)\rangle - \frac{\mu}{2}||p_\mu(Y) - Y||_F^2\\ &= \mu\langle Y - X, p_\mu(Y) - Y\rangle + \frac{\mu}{2}||p_\mu(Y) - Y||_F^2,\end{aligned}$$

where the first equality follows from that $p_\mu(Y)$ is a minimizer of $Q_\mu(X,Y)$ as in Eq. (11), and thus

$$\bigtriangledown f(Y) + \mu(p_\mu(Y) - Y) + \lambda g(p_\mu(Y)) = 0.$$

This completes the proof of the lemma. □

Proof of Theorem 3.2

Proof. Applying Lemma 3.1 with $(X = W^*, Y = W_n, \mu = L_{n+1})$ and making use of the fact that for any three matrices A, B, and C of the same size

$$||B-A||_F^2 + 2\langle B-A, A-C\rangle = ||B-C||_F^2 - ||A-C||_F^2, \tag{25}$$

we obtain that

$$\frac{2}{L_{n+1}}(F(W^*)-F(W_{n+1})) \geq ||W_{n+1}-W^*||_F^2 - ||W_n-W^*||_F^2.$$

Summing the above inequality over $n = 0, \cdots, k-1$ and making use of the inequality in Eq. (16), we get

$$\sum_{n=0}^{k-1}(F(W_{n+1})-F(W^*)) \leq \frac{\gamma L}{2}(||W_0-W^*||_F^2 - ||W_k-W^*||_F^2).$$

It follows from $F(W_{n+1}) \leq F(W_n)$ and $F(W_n) \geq F(W^*)$ that

$$\begin{aligned} k(F(W_k) - F(W^*)) &\leq \sum_{n=0}^{k-1}(F(W_{n+1}) - F(W^*)) \\ &\leq \frac{\gamma L}{2}||W_0 - W^*||_F^2, \end{aligned}$$

which leads to Eq. (17). □

Proof of Theorem 4.1

Proof. Let us denote

$$\begin{aligned} v_k &= F(W_k) - F(W^*), \\ U_k &= \alpha_k W_k - (\alpha_k - 1)W_{k-1} - W^*. \end{aligned}$$

Applying Lemma 3.1 with $(X = W_k, Y = Z_{k+1}, L = L_{k+1})$ and $(X = W^*, Y = Z_{k+1}, L = L_{k+1})$, respectively, we obtain the following two inequalities:

$$\begin{aligned} \frac{2}{L_{k+1}}(v_k - v_{k+1}) &\geq ||W_{k+1} - Z_{k+1}||_F^2 \\ &\quad +2\langle W_{k+1} - Z_{k+1}, Z_{k+1} - W_k\rangle, \end{aligned} \tag{26}$$

$$\begin{aligned} -\frac{2}{L_{k+1}}v_{k+1} &\geq ||W_{k+1} - Z_{k+1}||_F^2 \\ &\quad +2\langle W_{k+1} - Z_{k+1}, Z_{k+1} - W^*\rangle. \end{aligned} \tag{27}$$

Multiplying both sides of Eq. (26) by $(\alpha_{k+1} - 1)$ and adding it to Eq. (27), we get

$$\begin{aligned} \frac{2}{L_{k+1}}((\alpha_{k+1}-1)v_k - \alpha_{k+1}v_{k+1}) &\geq \alpha_{k+1}||W_{k+1}-Z_{k+1}||_F^2 \\ +2\langle W_{k+1} - Z_{k+1}, \alpha_{k+1}Z_{k+1} - (\alpha_{k+1}-1)W_k - W^*\rangle. \end{aligned}$$

Multiplying the last inequality by α_{k+1} and making use of the equality $\alpha_k^2 = \alpha_{k+1}^2 - \alpha_{k+1}$ derived from Eq. (18), we get

$$\begin{aligned} \frac{2}{L_{k+1}}(\alpha_k^2 v_k - \alpha_{k+1}^2 v_{k+1}) &\geq ||\alpha_{k+1}(W_{k+1} - Z_{k+1})||_F^2 \\ +2\alpha_{k+1}\langle W_{k+1} - Z_{k+1}, \alpha_{k+1}Z_{k+1} - (\alpha_{k+1}-1)W_k - W^*\rangle. \end{aligned}$$

Applying the equality in Eq. (25) to the right-hand side of the above inequality, we get

$$\begin{aligned} \frac{2}{L_{k+1}}(\alpha_k^2 v_k - \alpha_{k+1}^2 v_{k+1}) \geq &||\alpha_{k+1}W_{k+1} - (\alpha_{k+1}-1)W_k \\ &- W^*||_F^2 - ||\alpha_{k+1}Z_{k+1} - (\alpha_{k+1}-1)W_k - W^*||_F^2. \end{aligned}$$

It follows from Eq. (19) and the definition of U_k that

$$\frac{2}{L_{k+1}}(\alpha_k^2 v_k - \alpha_{k+1}^2 v_{k+1}) \geq ||U_{k+1}||_F^2 - ||U_k||_F^2,$$

which combined with $L_{k+1} \geq L_k$ leads to

$$\frac{2}{L_k}\alpha_k^2 v_k - \frac{2}{L_{k+1}}\alpha_{k+1}^2 v_{k+1} \geq ||U_{k+1}||_F^2 - ||U_k||_F^2. \tag{28}$$

Applying Lemma 3.1 with $(X = W^*, Y = Z_1, L = L_1)$, we obtain

$$\begin{aligned} &F(W^*) - F(W_1) = F(W^*) - F(p_{L_1}(Z_1)) \\ &\geq \frac{L_1}{2}||p_{L_1}(Z_1) - Z_1||^2 + L_1\langle Z_1 - W^*, p_{L_1}(Z_1) - Z_1\rangle \\ &= \frac{L_1}{2}||W_1 - Z_1||^2 + L_1\langle Z_1 - W^*, W_1 - Z_1\rangle \\ &= \frac{L_1}{2}||W_1 - W^*||^2 - \frac{L_1}{2}||Z_1 - W^*||^2. \end{aligned}$$

Hence, we have

$$\frac{2}{L_1}v_1 \leq ||Z_1 - W^*||^2 - ||W_1 - W^*||^2. \tag{29}$$

It follows from Eqs. (28) and (29) that $\frac{2}{L_k}\alpha_k^2 v_k \leq ||W_0 - W^*||^2$, which combined with $\alpha_k \geq (k+1)/2$ yields

$$F(W_k) - F(W^*) \leq \frac{2L_k||W_0 - W^*||^2}{(k+1)^2} \leq \frac{2\gamma L||W_0 - W^*||^2}{(k+1)^2}.$$

This completes the proof of the theorem. □

Acknowledgments

This work was supported by NSF IIS-0612069, IIS-0812551, CCF-0811790, NIH R01-HG002516, and NGA HM1582-08-1-0016.

References

Abernethy, J., Bach, F., Evgeniou, T., & 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., & Vert, J.-P. (2009). A new approach to collaborative filtering: Operator estimation with spectral regularization. *J. Mach. Learn. Res.*, *10*, 803–826.

Amit, Y., Fink, M., Srebro, N., & Ullman, S. (2007). Uncovering shared structures in multiclass classification. In *Proceedings of the International Conference on Machine Learning*, 17–24.

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

Bach, F. R. (2008). Consistency of trace norm minimization. *J. Mach. Learn. Res.*, *9*, 1019–1048.

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

Bertsekas, D. P. (1999). *Nonlinear programming*. Athena Scientific. 2nd edition.

Cai, J.-F., Candés, E. J., & Shen, Z. (2008). *A singular value thresholding algorithm for matrix completion* (Technical Report 08-77). UCLA Computational and Applied Math.

Candés, E. J., & Recht, B. (2008). *Exact matrix completion via convex optimization* (Technical Report 08-76). UCLA Computational and Applied Math.

Fazel, M., Hindi, H., & Boyd, S. P. (2001). A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American Control Conference*, 4734–4739.

Lu, Z., Monteiro, R. D. C., & Yuan, M. (2008). Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Submitted to Mathematical Programming*.

Ma, S., Goldfarb, D., & Chen, L. (2008). *Fixed point and Bregman iterative methods for matrix rank minimization* (Technical Report 08-78). UCLA Computational and Applied Math.

Nemirovsky, A. S., & Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. John Wiley & Sons Ltd.

Nesterov, Y. (1983). A method for solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Dokl.*, *27*, 372–376.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, *103*, 127–152.

Nesterov, Y. (2007). *Gradient methods for minimizing composite objective function* (Technical Report 2007/76). CORE, Université catholique de Louvain.

Obozinski, G., Taskar, B., & Jordan, M. I. (2009). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*. In press.

Recht, B., Fazel, M., & Parrilo, P. (2008a). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *Submitted to SIAM Review*.

Recht, B., Xu, W., & Hassibi, B. (2008b). Necessary and sufficient condtions for success of the nuclear norm heuristic for rank minimization. In *Proceedings of the 47th IEEE Conference on Decision and Control*, 3065–3070.

Rennie, J. D. M., & Srebro, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the International Conference on Machine Learning*, 713–719.

Srebro, N., Rennie, J. D. M., & Jaakkola, T. S. (2005). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, 1329–1336.

Toh, K.-C., & Yun, S. (2009). An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Preprint, Department of Mathematics, National University of Singapore, March 2009.

Tomioka, R., & Aihara, K. (2007). Classifying matrices with a spectral regularization. In *Proceedings of the International Conference on Machine Learning*, 895–902.

Tseng, P. (2008). On accelerated proximal gradient methods for convex-concave optimization. *Submitted to SIAM Journal on Optimization*.

Weimer, M., Karatzoglou, A., Le, Q., & Smola, A. (2008a). COFI$^{\text{rank}}$ - maximum margin matrix factorization for collaborative ranking. In *Advances in Neural Information Processing Systems*, 1593–1600.

Weimer, M., Karatzoglou, A., & Smola, A. (2008b). Improving maximum margin matrix factorization. *Machine Learning*, *72*, 263–276.

Yuan, M., Ekici, A., Lu, Z., & Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B*, *69*, 329–346.