Primal-Dual methods for sparse constrained matrix completion

Yu Xin
MIT CSAIL

Tommi Jaakkola
MIT CSAIL

Abstract

We develop scalable algorithms for regular and non-negative matrix completion. In particular, we base the methods on trace-norm regularization that induces a low rank predicted matrix. The regularization problem is solved via a constraint generation method that explicitly maintains a sparse dual and the corresponding low rank primal solution. We provide a new dual block coordinate descent algorithm for solving the dual problem with a few spectral constraints. Empirical results illustrate the effectiveness of our method in comparison to recently proposed alternatives.

1 Introduction

Matrix completion lies at the heart of collaborative filtering. Given a sparsely observed rating matrix of users and items, the goal is to reconstruct the remaining entries so as to be able to recommend additional items to users. By placing constraints on the underlying rating matrix, and assuming favorable conditions for the selection of observed entries (cf. [12]), we may be able to recover the matrix from sparse [6] and potentially also noisy observations. The most commonly used constraint on the underlying matrix is that user preferences vary only across a few prominent underlying dimensions. The assumption can be made explicit so that the matrix has low rank or introduced as a regularizer (convex relaxations of rank) [7, 14]. An explicit low rank assumption about the underlying matrix results in a non-convex optimization problem. For this reason, recent work has focused on optimization algorithms (e.g., [9, 16, 15, 13]) and statistical recovery questions (e.g., [6, 2]) pertaining to convex relaxations of rank or alternative convex formulations (e.g., [1]).

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22.

Our focus in this paper is on optimization algorithms for trace-norm regularized matrix completion. Trace-norm (a.k.a nuclear norm) is a 1-norm penalty on the singular values of the matrix and thus leads to a low rank solution with sufficient regularization. One key difficulty with this approach is that while the resulting optimization problem is convex, it is not differentiable. A number of approaches have been suggested to deal with this problem (e.g., [4]). In particular, many variants of proximal gradient methods (e.g., [9]) are effective as they fold the non-smooth regularization penalty into a simple proximal update that makes use of singular value decomposition. An alternative strategy would be to cast the trace-norm itself as a minimization problem over weighted Frobenius norms that are both convex and smooth.

Another key difficulty arises from the sheer size of the full rating matrix even if the observations are sparse. This is a problem with all convex optimization approaches (e.g., proximal gradient methods) that explicitly maintain the predicted rating matrix (rank constraint would destroy convexity). The scaling problem can be remedied by switching to the dual regularization problem where dual variables are associated with the few observed entries [14]. The standard dual approach would, however, require us to solve an additional reconstruction problem (using complementary slackness) to realize the actual rating matrix.

We introduce here a new primal-dual approach that avoids both of these problems, leading to a sparse dual optimization problem while also iteratively generating a low rank estimated matrix. The approach scales well to large and especially sparse matrices that are unsuitable for methods that maintain the full matrix or require singular value decomposition. We also extend the approach to non-negative matrix completion.

2 Trace norm regularization for matrix completion

We consider the well-known sparse matrix completion problem [6]. The goal is to predict the missing entries of a $n \times m$ ($n \geq m$) real valued target matrix Y

based on a small subset of observed entries. We index the observed entries as $Y_{u,i}$, $(u,i) \in D$. A typical approach to this problem would constrain the predicted matrix W to have low rank: $W = UV^T$, where the smaller dimension of U and V is (substantially) less than m. A convex relaxation of the corresponding estimation problem is obtained via trace norm regularization (e.g., [7, 14])

$$J(W) \quad = \quad \sum_{(u,i)\in D} \text{Loss}(Y_{u,i}, W_{u,i}) + \lambda \|W\|_* \quad (1)$$

where $\text{Loss}(Y_{u,i}, W_{u,i})$ is a convex loss function of $W_{u,i}$ such as the squared loss. To simplify the ensuing notation, we assume that the loss function depends only on the difference between W and Y so that $\text{Loss}(Y_{u,i}, W_{u,i}) = \text{Loss}(Y_{u,i} - W_{u,i})$. The trace-norm $\|W\|_*$ is a 1-norm penalty on the singular values of the matrix, i.e., $\|W\|_* = \sum_{j=1}^{m} \sigma_j(W)$ where $\sigma_j(W) \geq 0$ is the j^{th} singular value of W. For large enough λ, some of the singular values are set exactly to zero resulting in a low rank predicted matrix W.

One key optimization challenge comes from the regularization penalty $\|W\|_*$ which is convex but not differentiable. Effective algorithms based on (accelerated) proximal gradient updates have been developed to overcome this issue (e.g., [9, 2]). However, such methods update the full matrix W in each iteration and are thus unsuitable for large problems.

3 A primal-dual algorithm

We consider here primal-dual algorithms that operate only over the observed entries, thus leading to a sparse estimation problem. In our formulation, the dual variables arise from Legendre conjugate transformations of the loss functions:

$$\text{Loss}(z) = \max_q \{qz - \text{Loss}^*(q)\} \quad (2)$$

where the conjugate function $\text{Loss}^*(q)$ is also convex. For example, for the squared loss, $\text{Loss}^*(q) = q^2/2$. The resulting dual variables, $Q_{u,i}$, $(u,i) \in D$, one variable for each observed entry, can be viewed as a sparse $n \times m$ matrix Q with remaining entries set to zero.

The trace norm regularization in the primal formulation translates into a semi-definite constraint in the dual (cf. [7, 14]):

$$\begin{array}{ll} \text{maximize} & tr(Q^T Y) - \sum_{(u,i)\in D} \text{Loss}^*(Q_{u,i}) \quad (3) \\ \text{subject to} & Q^T Q \leq \lambda^2 I \end{array}$$

We derive the dual problem in the next section as the derivation will be reused with slight modification in the context of non-negative matrix factorization. Note that the constraint $Q^T Q \leq \lambda^2 I$ can be equivalently written as $\|Qb\|^2 \leq \lambda^2$ for all b such that $\|b\| = 1$. In other words, it is a constraint on the spectral norm of Q. We will solve the dual by iteratively adding spectral constraints that are violated.

3.1 Derivation of the dual

The trace norm of matrix W with factorization $W = UV^T$ is given by (e.g., [14])

$$\|W\|_* = \min_{W=UV^T} \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2) \quad (4)$$

where U and V need not be low rank so the factorization always exists. Consider then an extended symmetric matrix

$$X = \begin{pmatrix} UU^T & UV^T \\ VU^T & VV^T \end{pmatrix} \quad (5)$$

whose upper right and lower left components equal W and W^T, respectively. As a result, $\|W\|_* = \min_{W=X^{UR}} \text{tr}(X)/2$ where X^{UR} is the upper right part of X. By construction, X is symmetric and positive semi-definite.

We can also expand the observations into a symmetric matrix

$$Z = \begin{pmatrix} 0 & Y \\ Y^T & 0 \end{pmatrix} \quad (6)$$

and use Ω as the index set to identify observed entries in the upper right and lower left components of Z. Formally, $\Omega = \{(i,j) | 1 \leq i,j \leq m+n, \quad (i, j-n) \ \ or \ \ (i-n, j) \in D\}$.

With these definitions, the primal trace norm regularization problem is equivalent to the extended problem

$$\min_{X\in S} \sum_{(u,i)\in\Omega} \text{Loss}(Z_{u,i} - X_{u,i}) + \lambda \text{tr}(X) \quad (7)$$

where S is the cone of positive semi-definite matrices in $\mathbb{R}^{(m+n)\times(m+n)}$.

We introduce dual variables for each observed entry in Z via Legendre conjugate transformations of the loss functions. The Lagrangian involving both primal and dual variables is then given by

$$\begin{array}{rcl} \mathcal{L}(A, X, Z) & = & \displaystyle\sum_{(u,i)\in\Omega} \Big[A_{u,i}(Z_{u,i} - X_{u,i}) - \text{Loss}^*(A_{u,i}) \Big] \\ & + & \lambda tr(X) \end{array} \quad (8)$$

where A can be written as

$$A = \begin{pmatrix} 0 & Q \\ Q^T & 0 \end{pmatrix} \quad (9)$$

and Q is sparse such that $Q_{u,i} = 0$ is $(u,i) \notin D$. To introduce the dual variables for the positive semi-definite constraint, we consider the dual cone of S which is defined as

$$S^* = \{E \in \mathbb{R}^{(m+n)\times(m+n)}, \mathrm{tr}(E^T M) \geq 0, \forall M \in S\} \quad (10)$$

S is self-dual so that $S^* = S$. The Lagrangian is then

$$\mathcal{L}(A,E,X,Z) = \sum_{(u,i)\in\Omega} \Big[A_{u,i}(Z_{u,i} - X_{u,i}) - \mathrm{Loss}^*(A_{u,i})\Big]$$

$$+ \lambda tr(X) - \mathrm{tr}(E^T X) \quad (11)$$

To solve for X, we set $d/X\, \mathcal{L}(A,E,X,Z) = 0$, and get

$$\lambda I - A = E \in S \quad (12)$$

Inserting the solution back into the Lagrangian, we obtain

$$L(A) = \sum_{(u,i)\in\Omega} A_{u,i} Z_{u,i} - \mathrm{Loss}^*(A_{u,i}) \quad (13)$$

Since E does not show up in the Lagrangian, we can replace the equation (12) with a constraint $\lambda I - A \in S$. The formulation can be simplified by considering the original Q and Y which correspond to the upper right components of A and Z. The dual problem in these variables is then

$$\begin{array}{ll} \text{maximize} & tr(Q^T Y) - \sum_{(u,i)\in D} \mathrm{Loss}^*(Q_{u,i}) \\ \text{subject to} & \lambda^2 I - Q^T Q \in S \end{array} \quad (14)$$

3.2 Solving the dual

There are three challenges with the dual. The first one is the separation problem, i.e., finding vectors b, $\|b\| = 1$, such that $\|Qb\|^2 > \lambda^2$, where Q refers to the current solution. Each such b can be found efficiently precisely because Q is sparse. The second challenge is effectively solving the dual under a few spectral constraints. For this, we derive a new block coordinate descent approach (cf. [16]). The third challenge concerns the problem of reconstructing the primal matrix W from the dual solution. By including only a few spectral constraints in the dual, we obtain a low-rank primal solution. We can thus explicitly maintain a primal-dual pair of the relaxed problem (fewer constraints) throughout the optimization.

The separation problem. We will iteratively add constraints represented by b. The separation problem we must solve is then: given the current solution Q, find b for which $\|Qb\|^2 > \lambda^2$. This is easily solved by finding the eigenvector of $Q^T Q$ with the largest eigenvalue. For example, the power method

$$b = \mathrm{randn}(m,1). \text{ Iterate } b \leftarrow Q^T Q b, \quad b \leftarrow b/\|b\| \quad (15)$$

is particularly effective with sparse matrices. If $\|Qb\|^2 > \lambda^2$ for the resulting b, then we add a single constraint $\|Qb\|^2 \leq \lambda^2$ into the dual. Note that b does not have to be solved exactly; any b provides a valid albeit not necessarily the tightest constraint. An existing constraint can also be easily tightened later on with a few iterations of the power method, starting with the current b. We can fold this tightening together with the block coordinate optimization discussed below.

Primal-dual block coordinate descent. The second problem is to solve the dual subject to $\|Qb^l\|^2 \leq \lambda^2$, $l = 1,\ldots,k$, instead of the full set of constraints $Q^T Q \leq \lambda^2 I$. This partially constrained dual problem can be written as

$$tr(Q^T Y) - \sum_{(u,i)\in D} \mathrm{Loss}^*(Q_{u,i}) - \sum_{l=1}^{k} h(\|Qb^l\|^2 - \lambda^2) \quad (16)$$

where $h(z) = \infty$ if $z > 0$ and $h(z) = 0$ otherwise. The advantage of this form is that each $h(\|Qb\|^2 - \lambda^2)$ term is a convex function of Q (a convex non-decreasing function of a convex quadratic function of Q, therefore itself convex). We can thus obtain its conjugate dual as

$$\begin{aligned} h(\|Qb\|^2 - \lambda^2) &= \sup_{\xi\geq 0}\{\xi(\|Qb\|^2 - \lambda^2)/2\} \\ &= \sup_{\xi\geq 0, v} \{v^T Q b - \|v\|^2/(2\xi) - \xi\lambda^2/2\} \end{aligned}$$

where the latter form is jointly concave in (ξ, v) where b is assumed fixed. This step lies at the core of our algorithm. By relaxing the supremum over (ξ, v), we obtain a *linear*, not quadratic, function of Q. The new Lagrangian is given by

$$\begin{aligned} \mathcal{L}(Q,V,\xi) &= tr(Q^T Y) - \sum_{(u,i)\in D} \mathrm{Loss}^*(Q_{u,i}) \\ &- \sum_{l=1}^{k} \Big[(v^l)^T Q b^l - \|v^l\|^2/(2\xi^l) - \xi^l\lambda^2/2\Big] \\ &= tr(Q^T(Y - \sum_l v^l (b^l)^T)) \\ &- \sum_{(u,i)\in D} \mathrm{Loss}^*(Q_{u,i}) + \sum_{l=1}^{k}\Big[\frac{\|v^l\|^2}{2\xi^l} + \frac{\xi^l\lambda^2}{2}\Big] \end{aligned}$$

which can be maximized with respect to Q for fixed (ξ^l, v^l), $l = 1,\ldots,k$. Indeed, our primal-dual algorithm seeks to iteratively minimize $\mathcal{L}(V,\xi) = \max_Q \mathcal{L}(Q,V,\xi)$ while explicitly maintaining $Q = Q(V,\xi)$. Note also that by maximizing over Q, we reconstitute the loss terms $\mathrm{tr}(Q^T(Y - W)) - \sum_{(u,i)\in D} \mathrm{Loss}^*(Q_{u,i})$. The predicted rank k matrix W is therefore obtained explicitly $W = \sum_l v^l (b^l)^T$.

By allowing k constraints in the dual, we search over rank k predictions.

The iterative algorithm proceeds by selecting one l, fixing (ξ^j, v^j), $j \neq l$, and optimizing $\mathcal{L}(V,\xi)$ with respect to (ξ^l, v^l). As a result, $\mathcal{L}(V,\xi)$ is monotonically decreasing. Let $\tilde{W} = \sum_{j\neq l} v^j (b^j)^T$, where only the observed entries need to be evaluated. We consider two variants for minimizing $\mathcal{L}(V,\xi)$ with respect to (ξ^l, v^l):

Method 1: When the loss function is not strictly convex, we first solve ξ^l as function of v^l resulting in $\xi^l = \|v^l\|/\lambda$. Recall that $\mathcal{L}(V,\xi)$ involves a maximization over Q that reconstitutes the loss terms with a predicted matrix $W = \tilde{W} + v^l (b^l)^T$. By dropping terms not depending on v^l, the relevant part of $\min_{\xi^l} \mathcal{L}(V,\xi)$ is given by

$$\sum_{(u,i)\in D} \text{Loss}(Y_{u,i} - \tilde{W}_{u,i} - v^l_u b^l_i) + \lambda \|v^l\| \qquad (17)$$

which is a simple primal group Lasso minimization problem over v^l. Since b^l is fixed, the problem is convex and can be solved by standard methods.

Method 2: When the loss function is strictly convex, we can first solve $v^l = \xi^l Q b^l$ and explicitly maintain the maximizing $Q = Q(\xi^l)$. The minimization problem over $\xi^l \geq 0$ is

$$\max_Q \quad \Big\{ tr(Q^T(Y - \tilde{W})) - \sum_{(u,i)\in D} \text{Loss}^*(Q_{u,i}) \\ -\xi^l(\|Qb^l\|^2 - \lambda^2)/2 \Big\} \qquad (18)$$

where we have dropped all the terms that remain constant during the iterative step. For the squared loss, for example, $Q(\xi^l)$ is obtained in closed form:

$$Q_{u,\mathcal{I}_u}(\xi^l) \;=\; (Y_{u,\mathcal{I}_u} - \tilde{W}_{u,\mathcal{I}_u}) \left(1 - \frac{\xi^l b^l_{\mathcal{I}_u} (b^l_{\mathcal{I}_u})^T}{1 + \xi^l \|b^l_{\mathcal{I}_u}\|^2}\right),$$

where $\mathcal{I}_u = \{i : (u,i) \in D\}$ is the index set of observed elements for a row (user) u. In general, an iterative solution is required. The optimal value $\xi^l \geq 0$ is subsequently set as follows. If $\|Q(0)b^l\|^2 \leq \lambda^2$, then $\xi^l = 0$. Otherwise, since $\|Q(\xi^l)b^l\|^2$ is monotonically decreasing as a function of ξ^l, we find (e.g., via bracketing) $\xi^l > 0$ such that $\|Q(\xi^l)b^l\|^2 = \lambda^2$.

4 Constraints and convergence

The algorithm described earlier monotonically decreases $\mathcal{L}(V,\xi)$ for any fixed set of constraints corresponding to b^l, $l = 1, \ldots, k$. Any additional constraint further decreases this function. We consider here how quickly the solution approaches the dual optimum as new constraints are added. To this end, we write our algorithm more generally as minimizing

$$F(B) \;=\; \max_Q \big\{ \text{tr}(Q^T Y) - \text{Loss}^*(Q_{u,i}) \\ +\; 1/2\text{tr}((\lambda^2 I - Q^T Q)B) \big\} \qquad (19)$$

where B is a positive semi-definite $m \times m$ matrix. For any fixed set of constraints, our algorithm minimizes $F(B)$ over the cone $B = \sum_{l=1}^k \xi^l b^l (b^l)^T$ that corresponds to constraints $\text{tr}((\lambda^2 I - Q^T Q) b^l (b^l)^T) > 0, \forall l$. $F(B)$ is clearly convex as a point-wise supremum of linear functions of B. For simplicity, we assume that the loss function $\text{Loss}(z)$ is strictly convex with a Lipschitz continuous derivative (e.g., the squared loss). In this case, $F(B)$ is also differentiable with gradient $dF(B) = 1/2(\lambda^2 I - Q^T Q)$ where $Q = Q(B)$ (unique). Moreover, B^* that minimizes $F(B)$ remains bounded, as does $Q(B)$. Under these assumptions $F(B)$ has a Lipschitz continuous derivative but need not itself be strictly convex.

Theorem 4.1. *Under the assumptions discussed above,* $F(B^r) - F(B^*) = O(1/r)$.

Proof. Consider $(B^0, Q^0), (B^1, Q^1), \ldots$ the sequence generated by the primal-dual algorithm. Since $dF(B)$ is Lipschitz continuous with some constant L, $F(B)$ has a quadratic upper bound:

$$F(B) \;\leq\; F(B^{r-1}) + \langle dF(B^{r-1}), B - B^{r-1} \rangle \\ +L/2 \|B - B^{r-1}\|_F^2 \qquad (20)$$

In each iteration, we add a constraint $(b^r)^T Q^T Q b^r \leq \lambda^2$ and fully optimize Q and ξ^i to satisfy all the constraints that have been added. Prior to including the r^{th} constraint, from complementary slackness, we have for each $i < r$, $\xi^i (b^i)^T (\lambda^2 I - Q^T Q) b^i = 0$ where Q is optimized based on $r-1$ constraints. This means that $tr((\lambda^2 I - Q^T Q) \sum_{i=0}^{r-1} \xi^i b^i (b^i)^T) = tr((\lambda^2 I - Q^T Q) B^{r-1}) = 0$, or $\langle dF(B^{r-1}), B^{r-1} \rangle = 0$. Let $B^r = B^{r-1} + \xi^r b^r (b^r)^T$, where $b^r = \text{argmax}_{\|b\|=1} b^T Q^T Q b$, i.e., the largest eigenvalue λ_r^2 ($\lambda_r > \lambda$). Here, λ_r is the largest eigenvalue prior to including b^r. As a result, the upper bound evaluated at B^r yields

$$F(B^r) \leq F(B^{r-1}) + \xi^r(\lambda^2 - \lambda_r^2) + 2L(\xi^r)^2$$

By setting $\xi^r = (\lambda_r^2 - \lambda^2)/4L$ (non-negative), we get

$$F(B^r) \leq F(B^{r-1}) - (\lambda_r^2 - \lambda^2)^2/8L$$

The optimal value of ξ^r may be different. Moreover, at each iteration, B^r, including all of its previous constraints, are optimized. Thus the actual decrease may be somewhat larger.

From the convexity of $F(B)$ and complementary slackness, we have

$$F(B^{r-1}) - F(B^*) \;\leq\; \langle dF(B^{r-1}), B^{r-1} - B^* \rangle \\ = -\langle dF(B^{r-1}), B^* \rangle$$

where B^* is the optimum. B^* is a positive semi-definite matrix. Therefore

$$\begin{aligned} -\langle dF(B^{r-1}), B^*\rangle &\leq -\langle Proj_{neg}(dF(B^{r-1})), B^*\rangle \\ &\leq (\lambda_r^2 - \lambda^2)\mathrm{tr}(B^*) = (\lambda_r^2 - \lambda^2)C \end{aligned}$$

where $Proj_{neg}(\cdot)$ is a projection to negative semi-definite matrices. The minimum eigenvalue of $dF(B^{r-1}) = 1/2(\lambda^2 - Q^TQ)$ is $\lambda^2 - \lambda_r^2$. Combining this with the sufficient decrease, we get

$$F(B^{r-1}) - F(B^r) \geq \frac{(F(B^{r-1}) - F(B^*))^2}{8LC^2}$$

The above inequality implies

$$\begin{aligned} &\frac{1}{F(B^r) - F(B^*)} - \frac{1}{F(B^{r-1}) - F(B^*)} \\ &\geq \frac{F(B^{r-1}) - F(B^*)}{8LC^2(F(B^r) - F(B^*))} \geq \frac{1}{8LC^2} \end{aligned}$$

Summing over all r, we get

$$F(B^r) - F(B^*) \leq \frac{8LC^2(F(B^0) - F(B^*))}{r(F(B^0) - F(B^*)) + 8LC^2} = O(\frac{1}{r})$$

5 Convex formulation of nonnegative matrix factorization

The non-negative matrix factorization(NMF) problem is typically written as

$$\min_{U,V} \sum_{(u,i)\in D} (Y_{u,i} - (UV^T)_{u,i})^2 \tag{21}$$

where $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{m\times k}$ are both matrices with nonnegative entries. The optimization problem defined in this way is non-convex and NP-hard in general [20]. As before, we look for convex relaxations of this problem via trace norm regularization. This step requires some care, however. The set of matrices UV^T, where U and V are non-negative and of rank k, is not the same as the set of rank k matrices with non-negative elements. The difference is well-known even if not fully characterized.

Let us first introduce some basic concepts such as completely positive and co-positive matrices. A matrix W is completely positive if there exists a nonnegative matrix B such that $W = BB^T$. It can be seen from the definition that each completely positive matrix is also positive semi-definite. A matrix C is co-positive if for any $v \geq 0$, i.e., a vector with non-negative entries, $v^TCv \geq 0$. Unlike completely positive matrices, co-positive matrices may be indefinite. We will denote the set of completely positive symmetric matrices and co-positive matrices as $\mathcal{S}^+$ and $\mathcal{C}^+$, respectively. $\mathcal{C}^+$ is the dual cone of $\mathcal{S}^+$, i.e., $(\mathcal{S}^+)^* = \mathcal{C}^+$. (the dimensions of these cones are clear from context).

Following the derivation of the dual in section 3, we consider expanded symmetric matrices X and Z in $\mathbb{R}^{(m+n)\times(m+n)}$, defined as before. Instead of requiring X to be positive semi-definite, however, for NMF we constrain X to be a completely positive matrix, i.e., in $\mathcal{S}^+$. The primal optimization problem can be given as

$$\min_{X\in\mathcal{S}^+} \sum_{(u,i)\in\Omega} \mathrm{Loss}(X_{u,i} - Z_{u,i}) + \lambda\mathrm{tr}(X) \tag{22}$$

Notice that the primal problem involves infinite constraints corresponding to $X \in \mathcal{S}^+$ and is difficult to solve directly.

We start with the Lagrangian involving primal and dual variables:

$$\begin{aligned} \mathcal{L}(A, C, X, Z) &= \sum_{(u,i)\in\Omega} \left[A_{u,i}(Z_{u,i} - X_{u,i}) - \mathrm{Loss}^*(A_{u,i})\right] \\ &+ \lambda tr(X) - \mathrm{tr}(C^TX) \end{aligned} \tag{23}$$

where $C \in \mathcal{C}^+$ and $A = [0, Q; Q^T, 0]$ as before. By setting $d/X\ \mathcal{L}(A, E, X, Z) = 0$, we get

$$\lambda I - A = C \in \mathcal{C}^+ \tag{24}$$

Substituting the result back into the Lagrangian, the dual problem becomes

$$\begin{aligned} \text{maximize} \quad & tr(A^TZ) - \textstyle\sum_{(u,i)\in\Omega} \mathrm{Loss}^*(A) \\ \text{subject to} \quad & \lambda I - A \in \mathcal{C}^+ \end{aligned} \tag{25}$$

where A is a sparse matrix in the sense that $A_{u,i} = 0$ if $(u, i) \notin \Omega$. The co-positivity constraint is equivalent to $v^T(\lambda I - A)v \geq 0$, $\forall v \geq 0$.

Similarly to the primal dual algorithm, in each iteration $v^l = [a^l; b^l]$ is selected as the vector that violates the constraint the most. This vector can be found by solving

$$\max_{v\geq 0, \|v\|\leq 1} v^TAv = \max_{a,b\geq 0, \|a\|^2+\|b\|^2\leq 1} a^TQb \tag{26}$$

where Q is the matrix of dual variables (before expansion), and $a \in \mathbb{R}^m$, $b \in \mathbb{R}^n$. While the sub-problem of finding the most violating constraint is NP-hard in general, it is unnecessary to find the global optimum. Any non-negative v that violates the constraint $v^T(\lambda I - A)v \geq 0$ can be used to improve the current solution. At the optimum, $\|a\|^2 = \|b\|^2 = 1/2$, and thus it is equivalent to solve

$$\max_{a,b\geq 0, \|a\|^2=\|b\|^2=1/2} a^TQb \tag{27}$$

A local optimum can be found by alternatingly optimizing a and b according to

$$a = h(Qb)/(\sqrt{2}\|h(Qb)\|)$$
$$b = h(Q^T a)/(\sqrt{2}\|h(Q^T a)\|) \quad (28)$$

where $h(x) = \max(0, x)$ is the element-wise non-negative projection. The running time across the two operations is directly proportional to the number of observed entries. The stationary point $v_s = [a_s; b_s]$ satisfies $\|v_s\| = 1$ and $Av_s + \max(0, -Av_s) + \|h(Av_s)\|v_s = 0$ which are necessary (but not sufficient) optimality conditions for (26). As a result, many elements in a and b are exactly zero so that the decomposition is not only non-negative but sparse.

Given v^l, $l = 1, \ldots, k$, the Lagrangian is then

$$\begin{aligned} L(A, \xi) &= tr(A^T Z) - \sum_{(u,i)\in\Omega} \text{Loss}^*(A) \\ &- \sum_l \xi^l((v^l)^T A v^l - \lambda) \end{aligned}$$

where $\xi^l \geq 0$ and $\sum_{l=1}^k \xi^l v^l (v^l)^T \in \mathcal{S}^+$.

To update ξ^l, we fix ξ^j $(j \neq l)$ and optimize the objective with respect to A and ξ^l,

$$\begin{aligned} \max_{A, \xi^l \geq 0} \quad & tr(A^T(Z - \textstyle\sum_{j\neq l} \xi^j v^j (v^j)^T)) \\ & - \textstyle\sum_{(u,i)\in\Omega} \text{Loss}^*(A) - \xi^l((v^l)^T A v^l - \lambda) \end{aligned}$$

It can be seen from above formulation that our primal solution without the l^{th} constraint can be reconstructed from the dual as $\tilde{X} = \sum_{j\neq l} \xi^j v^j (v^j)^T \in \mathcal{S}^+$. Only a small fraction of entries of this matrix (those corresponding to observations) need to be evaluated. If the loss function is the squared loss, then the variables A and ξ^l have closed form expressions. Specifically, for a fixed ξ^l,

$$A(\xi^l)_{u,i} = Z_{u,i} - \tilde{X}_{u,i} - \xi^l v_u^l v_i^l \quad (29)$$

If $(v^l)^T A(0) v^l \leq \lambda$, then the optimum $\hat{\xi}^l = 0$. Otherwise

$$\hat{\xi}^l = \frac{\sum_{(u,i)\in\Omega}(Z_{u,i} - \tilde{X}_{u,i}) v_u^l v_i^l - \lambda}{\sum_{(u,i)\in\Omega}(v_u^l v_i^l)^2} \quad (30)$$

The update of A and ξ^l takes time linear in the number of observed elements. If we update all ξ^l in each iteration, then the running time is k times longer and is the same as the update times in multiplicative update methods ([22]) that aim to directly solve the optimization problem (21). The multiplicative update method revises the decomposition into U and V iteratively while keeping U and V nonnegative. Because the objective is non-convex, the algorithm gets frequently trapped in locally optimal solutions. It is also likely to convergence slowly, and often requires a proper initialization scheme [23]. In contrast, the primal-dual method discussed here solves a convex optimization problem. With a large enough λ, only a few constraints are expected to be necessary, resulting in a low rank reconstruction. The key difficulty in our approach is identifying a violating constraint. The simple iterative method may fail even though a violating constraint exists. This is where randomization is helpful in our approach. We initialize a and b randomly and run the separation algorithm several times so as to get a reasonable guarantee of finding a violated constraint when they exist.

6 Experiments

There are many variations of our primal-dual (PD) method pertaining to how the constraints are enforced and when they are updated. We introduced the method as one that fully optimizes all ξ^l corresponding to the available constraints prior to searching for additional constraints. Another variant, analogous to gradient based methods, is to update only the last ξ^k associated with the new constraint without ever going back and iteratively optimizing any of the previous $\xi^l (l < k)$ (or the constraints themselves). The method adds constraints fasters as ξ^l are updated only once when introduced. Another variant is to update all the previous ξ^l together with their constraint vectors b^l (through the power method) before introducing a new constraint. By updating b^l as well, this approach replaces the previous constraints with tighter ones. This protocol can reduce the number of constraints that are added is thus useful in cases where a very low rank solution is desirable.

We compare here the two variants of our method to recent approaches proposed in the literature. We begin with the state of art method proposed in [18] which we refer to as JS. Similar to PD, JS formulates a trace norm regularization problem using extended matrices $X, Z \in \mathbb{R}^{(m+n)\times(m+n)}$ which are derived from the predictions and observed data. It solves the following optimization problem,

$$\min_{X\geq 0, \text{tr}(X)=t} \text{Loss}(X, Z) \quad (31)$$

Instead of penalizing the trace of X, JS fixes $\text{tr}(X)$ to a constant t. In each iteration, JS finds the maximum eigenvector of the gradient, v^k, and updates X according to $X = (1 - t_k)X + t_k v^k (v^k)^T$ where t_k is an optimized step size. During the optimization process, the trace norm constraint is always satisfied. Though JS and PD attempt to optimize different objective functions, if we set $t = \text{tr}(X^*)$ where X^* is the optimum

matrix from our regularized objective (1), then JS will converge to the same solution in the limit.

The comparison is on Movielens 10M dataset which contain 10^7 ratings of 69878 users on 10677 movies. Following the setup in [18], we use partition r_b provided with the dataset which has 10 test ratings per user. The regularization parameter λ in the PD method is set to 50. For the JS method, the corresponding value of the constant $t = tr(X^*) = 30786$ where X^* is the optimum from the PD method that fully optimizes all ξ^l before searching for additional constraints.

In order to ensure that we perform a fair comparison with JS, rather than updating ξ^l for all $l = 1, 2, ..., k$ at iteration k, only ξ^k is updated (the first variant introduced above). In other words, all ξ^l are updated only once when the corresponding constraint is incorporated. Figure 1-a compares the test RMSE of JS and PD as a function of time. The test error of PD decreases much faster than JS. Figure 1-b compares the test error as a function of rank, i.e., as a function of iteration rather than time. PD performs significantly better than JS when rank is small at the beginning owing to the better optimization step. In Figure 2-a, we compare the regularized primal objective function values $J(W) = L(W) + \lambda\|W\|_*$ as a function of time. Note that the objectives used by the methods are different and thus JS does not directly optimize this function. However, it will converge to the same limiting value at the optimum given how the parameters t and λ are set. From the figure, $J(W)$ does not converge to the optimum over the training set as the limiting values should agree. This is also evident in Figure 2-b that shows the training error as a function of time. PD is consistently lower than JS.

We also compare our PD method to a recently proposed GECO algorithm (see [19]). GECO seeks to solve the rank constrained optimization problem

$$\min_{rank(W)\leq r} \text{Loss}(W, Y) \tag{32}$$

It maintains a decomposition in the form $W = UV^T$ and, at each iteration, increases the rank of U and V by concatenating vectors u and v that correspond to the largest singular values of the gradient of $\text{Loss}(W, Y)$. Moreover, it optimizes the expanded U and V by rotating and scaling the corresponding vectors. This solidifying step is computationally expensive but nevertheless improves the fixed-rank result. In order to compare fairly with their method, we use the second variant of our method. In other words, before adding a new constraint in iteration k, we update all ξ^l including b^l for $l = 1, 2, ...k$ for a fixed number of times related to their algorithm (q times). Note that our

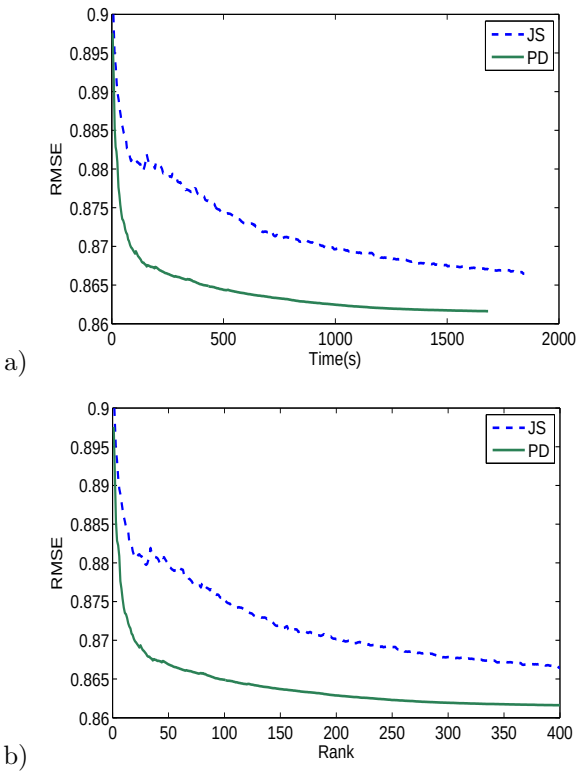

Figure 1: a) test RMSE comparison of JS and PD as a function of training time. b) test RMSE comparison of JS and PD as a function of rank.

algorithm here is different from that used in comparison to JS. Due to the complexity of GECO, we used a smaller dataset, Movielens 1M, which contain 10^6 ratings of 6040 users about 3706 movies. Following the setup in [19], 50% of the ratings are used for training, and the rest for testing.

Figure 3 compares the test RMSE of GECO and PD. GECO becomes very slow when the rank increases (the complexity can be $O(r^6)$ as a function of rank r). It already took 17 hours to get rank 9 solutions from GECO, so we didn't attempt to use it on bigger datasets. Figure 3-b compares the test RMSE as a function of rank. PD is substantially faster and needs fewer constraints to perform well.

7 Discussion

Our primal-dual algorithm is monotone and heavily exploits sparsity of observations for efficient computations. Constraints are added to the dual problem in a sequential cutting plane fashion. The method also maintains a low rank primal solution corresponding to the partially constrained dual. We believe our primal-dual method (PD) is particularly suitable for sparse large scale problems, and empirical results con-

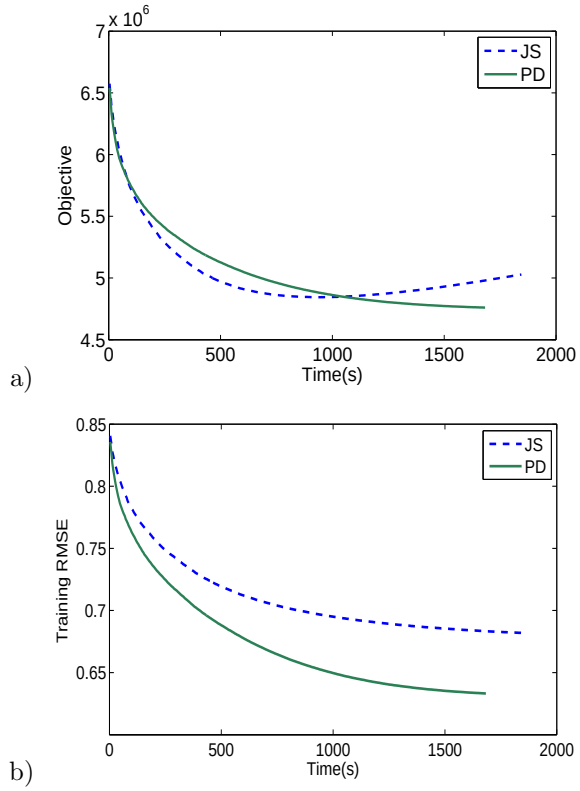

Figure 2: a) objective $J(W)$ as a function of time. b) training RMSE as a function of time.

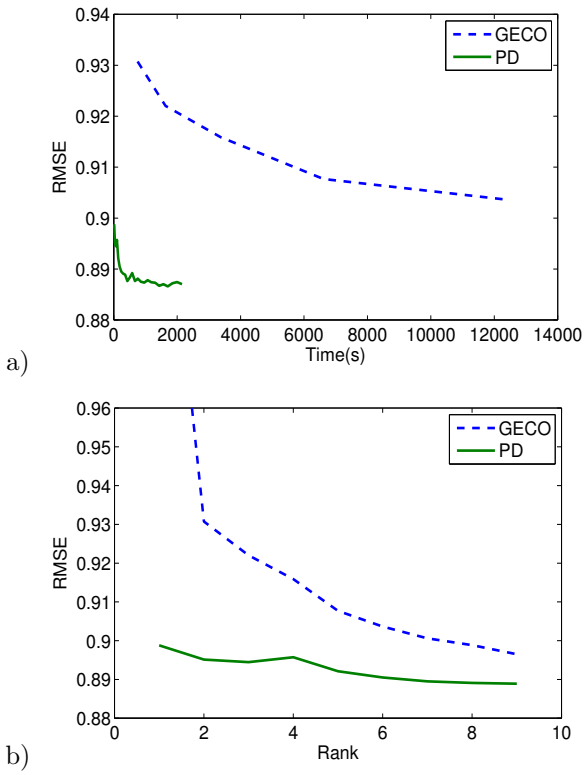

Figure 3: a) test RMSE comparison of GECO and PD as a function of training time. b) test RMSE comparison of GECO and PD as a function of rank.

firm that the stronger systematic optimization of each constraint is beneficial in comparison to other recent approaches.

There are a number of possible extensions of the basic approach. We already showed how non-negative matrix factorization problems can be solved within a similar framework with the caveat that the separation problem is substantially harder in this case. Other variations such as non-negative factorization with sparse factors are also possible.

References

[1] J. Abernethy, F. Bach, T. Evgeniou and J. Vert, *A New Approach to Collaborative Filtering: Operator Estimation with Spectral Regularization*, Journal of Machine Learning Research, Volume 10, 2009.

[2] A. Agarwal, S. N. Negahban, and M. J. Wainwright, *Fast global convergence of gradient methods for high-dimensional statistical recovery*, NIPS, 2010.

[3] A. Argyriou, C. A. Micchelli, and M. Pontil, *On Spectral Learning*, Journal of Machine Learning Research 11(2010), 905-923.

[4] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2009.

[5] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear System Theory*, Princeton University Press, 2005.

[6] E. J. Candes and B. Recht, *Exact matrix completion via convex optimization*, Found. of Comput. Math., 9 717-772, 2008.

[7] M. Fazel, H. Hindi, and S. Boyd, *A rank minimization heuristic with application to minimum order system approximation*, Proceedings American Control Conference, volume 6, pages 4734-4739, 2001.

[8] A. Goldberg, X. Zhu, B. Recht, J. Sui, and R. Nowak, *Transduction with matrix completion: Three birds with one stone*, NIPS, 2010.

[9] S. Ji and J. Ye, *An accelerated gradient method for trace norm minimization*, ICML, 2009.

[10] A. S. Lewis, *The Convex Analysis of Unitarily Invariant Matrix Functions*, Journal of Convex Analysis Volume 2 (1995), No.1/2, 173183.

[11] S. Ma, D. Goldfarb, and L. Chen, *Fixed point and Bregman iterative methods for matrix rank minimization*, Technical Report, Department of IEOR, Columbia University, October, 2008.

[12] B. M. Marlin, R. S. Zemel, S. Roweis, and M. Slaney, *Collaborative Filtering and the Missing at Random Assumption*, UAI, 2007.

[13] T. K. Pong, P. Tseng, S. Ji, and J. Ye, *Trace Norm Regularization: Reformulations, Algorithms, and Multi-task Learning*, SIAM Journal on Optimization, 2009.

[14] N. Srebro, J. Rennie and T. Jaakkola, *Maximum Margin Matrix Factorizations*, In Advances in Neural Information Processing Systems 17, 2005.

[15] K-C. Toh, S. Yun, *An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems*, preprint, Department of Mathematics, National University of Singapore, March 2009.

[16] P. Tseng, *Dual Coordinate Ascent Methods for Non-strictly Convex Minimization*, Mathematical Programming, Vol. 59, pp. 231247, 1993

[17] G. A. Watson, *Characterization of the subdifferential of some matrix norms*, Linear Algebra and its Applications, Volume 170, June 1992, Pages 33-45.

[18] M. Jaggi and with Marek Sulovsk, *A Simple Algorithm for Nuclear Norm Regularized Problems*, ICML 2010.

[19] S. Shalev-Shwartz, A. Gonenand and O. Shamir *Large-Scale Convex Minimization with a Low-Rank Constraint*, ICML 2011.

[20] Vavasis, S. *On the complexity of nonnegative matrix factorization.* arxiv.org, 0708.4149.

[21] J. B. Hiriart-Urruty and A. Seeger, *A variational approach to co-positive matrices*, SIAM Review 52 (2010), no. 4, 593629.

[22] D. D. Lee and H. S. Seung. *Algorithms for nonnegative matrix factorization.* NIPS 2001.

[23] C. Boutsidis and E. Gallopoulos, *SVD based initialization: A head start for nonnegative matrix factorization*, Pattern Recognition 2008.