Preference Completion: Large-scale Collaborative Ranking from Pairwise Comparisons

Dohyung Park
The University of Texas at Austin
dhpark@utexas.edu

Joe Neeman
The University of Texas at Austin
joeneeman@gmail.com

Jin Zhang
The University of Texas at Austin
zj@mail.utexas.edu

Sujay Sanghavi
The University of Texas at Austin
sanghavi@mail.utexas.edu

February 7, 2015

Abstract

In this paper we consider the collaborative ranking setting: a pool of users each provides a small number of pairwise preferences between d possible items; from these we need to predict each users preferences for items they have not yet seen. We do so by fitting a rank r score matrix to the pairwise data, and provide two main contributions:
(a) we show that an algorithm based on convex optimization provides good generalization guarantees once each user provides as few as $O(r \log^2 d)$ pairwise comparisons – essentially matching the sample complexity required in the related matrix completion setting (which uses actual numerical as opposed to pairwise information), and
(b) we develop a large-scale non-convex implementation, which we call AltSVM, that trains a factored form of the matrix via alternating minimization (which we show reduces to alternating SVM problems), and scales and parallelizes very well to large problem settings. It also outperforms common baselines on many moderately large popular collaborative filtering datasets in both NDCG and in other measures of ranking performance.

1 Introduction

This paper considers the following recommendation system problem: given a set of items, a set of users, and non-numerical *pairwise comparison* data, find the underlying preference ordering of the users. In particular, we are interested in the setting where data is of the form "user i preferes item j over item k", for different ordered user-item-item triples i, j, k. Pairwise preference data is wide-spread; indeed, almost any setting where a user is presented with a menu of options – and chooses one of them – can be considered to be providing a pairwise preference between the chosen item and every other item that is presented.

Crucially, we are interested in the collaborative filtering setting, where *(a)* on the one hand the number of such pairwise preferences we have for any one user is woefully insufficient to infer anything for that user in isolation; and *(b)* on the other hand, we aim for *personalization*, i.e. for every user to possibly have different inferred preferences from every other. To reconcile these two requirements, our method relates the preferences of users to each other via a low-rank matrix, which we (implicitly) assume governs the observed preferences. Essentially, we fit a low-rank users $\times$ items *score matrix* X to pairwise comparison data by trying to ensure that $X_{ij} - X_{ik}$ is positive when user i prefers item j to item k.

Our contributions: We present two algorithms to infer the score matrix X from training data; once inferred, this can be used for predicting future preferences. While there has been some recent work on fitting

low-rank score matrices to pairwise preference data (which we review and compare to below), in this paper we present the following two contributions:
(a) *A statistical analysis for the convex relaxation:* we bound the *generalization error* of the solution to our convex program. Essentially, we show that the minimizer of the empirical loss also almost minimizes the true expected loss. We also give a lower bound showing that our error rate is sharp up to logarithmic factors.
(b) *A large-scale non-convex implementation:* We provide a non-convex algorithm that we call Alternating Support Vector Machine (AltSVM). This non-convex algorithm is more practical than the convex program in a large-scale setting; it explicitly parameterizes the low-rank matrix in factored form and minimizes the hinge loss. Crucially, each step in this algorithm can be formulated as a standard SVM that updates one of the two factors; the algorithm proceeds by alternating updates to both factors. We apply a stochastic version of dual coordinate descent Hsieh et al. (2008); Shalev-Shwartz & Zhang (2013) with lock-free parallelization. This exploits the problem structure and ensures it parallelizes well. We show that our algorithm outperforms several existing collaborative ranking algorithms in both speed and prediction accuracy, and it achieves significant speedups as the number of cores increases.

1.1 Related Work

Ranking/learning preferences is a classical problem that has been considered in a large amount of work. There are many different settings for this problem, which we discuss below.

Learning to Rank The main problem in this community has been to estimate a ranking function from given feature vectors and relevance scores. Depending on its application, a feature vector may correpond to a user-item pair or a single item. While there have been algorithms that use pairwise comparisons Herbrich et al. (2000); Joachims (2002) of the training samples, our setting is different in that our data consists *only* of pairwise comparisons. We refer the reader to the survey Liu (2009).

One ranking with pairwise comparisions In a single-user model, we are asked to learn a single ranking given pairwise comparisons. Jamieson & Nowak (2011a) and Ailon (2011) consider an active query model with noiseless responses; Jamieson & Nowak (2011b) give an algorithm for exactly recovering the true ranking under a low-rank assumption similar to ours, while Ailon (2011) approximately recovers the true ranking without such an assumption. Wauthier et al. (2013) and Negahban et al. (2012) learn a ranking from noisy pairwise comparisions; Negahban et al. (2012) consider a Bradley-Terry-Luce model similar to ours and attempt to learn an underlying score vector, while Wauthier et al. (2013) get by without structure assumptions, but only attempt to learn the ranking itself. Hajek et al. (2014) considered a problem to learn a single ranking given a more generalized partial rankings from the Plackett-Luce model and provided a minimax-optimal algorithm.

Many rankings with pairwise comparisions Given multiple users with different rankings, one could of course attempt to learn their rankings by simply applying an algorithm from the previous section to each user individually. However, it is more efficient – both statistically and computationally – to postulate some global structure and use it to relate the many users' rankings. This is the same idea that has been applied so successfully in collaborative filtering. Rendle et al. (2009) and Liu et al. (2009) were the first to take this approach. They modeled the observations as coming from a BTL model with low-rank structure (i.e., very similar to our model) and gave algorithms for learning the model parameters. Yi et al. (2013) took a purely optimization-based approach. Rather than assuming a probabilistic model, they minimized a convex objective using the hinge loss on a low-rank matrix. In a slightly different model, Hu et al. (2008) and Shi et al. (2013) consider the problem of learning from latent feedback. Recently, Lu & Negahban (2014) analyzed an algorithm which is very similar to ours for the Bradley-Terry-Luce model independently from our work.

Many rankings with 1-bit ratings Instead of moving to pairwise comparisons, some work has suggested avoiding the difficulties of numerical ratings by instead asking users to give 1-bit ratings to items; that is, each user only indicates whether they like or dislike an item. In this setting, the work of Davenport et al. (2013) is most closely related to ours, in that they assume an underlying low-rank structure and give an algorithm based on convex optimization. Also, our theoretical analysis owes a lot to their work. Xu et al. (2013) consider a slightly different goal: rather than attempting to recover the preferences of each user, they try to cluster similar users and similar items together. Yun et al. (2014a) proposed an optimization problem motivated from robust binary classification and used stochastic gradient descent to solve the problem in a large-scale setting.

Many rankings with numerical ratings The goal in this setting is the same as ours, except that the data is in the form of numerical ratings instead of pairwise comparisons. Weimer et al. (2007) attempted to directly optimize Normalized Discounted Cumulative Gain (NDCG), a widely used performance measure for ranking problems. Balakrishnan & Chopra (2012), and Volkovs & Zemel (2012) converted this problem into a learning-to-rank problem and solved it using the existing algorithms. While these works considered the low-rank matrix model, different models are proposed by Weston et al. (2012) and Lee et al. (2014). Weston et al. (2012) proposed a tensor model to rank items for different queries and users, and Lee et al. (2014) proposed a weighted sum of low-rank matrix models.

2 Empirical Risk Minimization (ERM)

Let us first formulate the problem mathematically. The task is to estimate rankings of multiple users on multiple items. We denote the numbers of users by d_1, and the number of items by d_2. We are given a set of triples $\Omega \subset [d_1] \times [d_2] \times [d_2]$, where the preference of user i between items j and k is observed if $(i, j, k) \in \Omega$. The observed comparison is then given by $\{Y_{ijk} \in \{1, -1\} : (i, j, k) \in \Omega\}$ where $Y_{ijk} = 1$ if user i prefers item j over item k, and $Y_{ijk} = -1$ otherwise. Let $\Omega_i = \{(j, k) : (i, j, k) \in \Omega\}$ denote the set of item pairs that user i has compared.

We predict rankings for multiple users by estimating a score matrix $X \in \mathbb{R}^{d_1 \times d_2}$ such that $X_{ij} > X_{ik}$ means that user i prefers item j over item k. Then the sorting order for each row provides the predicted ranking for the corresponding user.

We propose (as have others) that X is low-rank or close to low-rank, the intuition being that each user bases their preferences on a small set of features that are common among all the items. Then the empirical risk minimization (ERM) framework can naturally be formulated as

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \sum_{(i,j,k)\in\Omega} \mathcal{L}(Y_{ijk}(X_{ij} - X_{ik})) \\ \text{subject to} \quad & \operatorname{rank}(X) \leq r \end{aligned} \tag{1}$$

where $\mathcal{L}(\cdot)$ is a monotonically non-increasing loss function which induces $X_{ij} > X_{ik}$ if $Y_{ijk} = 1$, and $X_{ij} < X_{ik}$ otherwise. (e.g., hinge loss, logistic regression loss, etc.)

Solving (1) is NP-hard because of the rank constraint. As a first alternative, we propose a straightforward convex relaxation.

3 Convex Relaxation

Our first method is the convex relaxation of (1), which involves a nuclear norm constraint.

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \sum_{(i,j,k)\in\Omega} \mathcal{L}(Y_{ijk}(X_{ij} - X_{ik})) \\ \text{subject to} \quad & \|X\|_* \leq \sqrt{\lambda d_1 d_2} \end{aligned} \tag{2}$$

Here, for any matrix X, the nuclear/trace norm $\|X\|_*$ denotes the sum of its singular values; it is a well-recognized convex surrogate for low-rank structure (most famously in matrix completion).

The only parameter of this algorithm is λ, which governs the trade-off between better optimizing the likelihood of the observed data, and the strictness in imposing approximate low-rank structure. Since we motivated our algorithm with the assumption that X has low rank, we should point out how our algorithm's parameter λ compares to the rank: note that if X is a $d_1 \times d_2$ rank-r matrix whose largest absolute entry is bounded by C then $\|X\|_* \le \sqrt{r}\|X\|_F \le C\sqrt{rd_1d_2}$. In other words, λ is a parameter that takes into account both the rank of X and the size of its elements, and it is roughly proportional to the rank.

3.1 Analytic results

We analyze (2) by assuming a standard model for pairwise comparisons. Then we provide a statistical guarantee of the method under the model.

Recall the classical Bradley-Terry-Luce model Bradley & Terry (1952); Luce (1959) for pairwise preferences of a single user, which assumes that the probability of item j being preferred over k is given by a logistic of the difference of the underlying preference scores of the two items. For multiple users, we assume that there is some true score matrix $X^* \in \mathbb{R}^{d_1 \times d_2}$ and

$$\Pr(Y_{ijk} = 1) \;=\; \frac{\exp(X^*_{ij} - X^*_{ik})}{1 + \exp(X^*_{ij} - X^*_{ik})}.$$

Assume that each user-item-item triple (i, j, k) independently belongs to Ω with probability $p_{i,j,k}$, and let $m = \sum_{i,j,k} p_{i,j,k}$ be the expected size of Ω. We will assume that the $p_{i,j,k}$ are approximately balanced in the sense that no user-item pair is observed too frequently:

Assumption 3.1. *There is a constant $\kappa > 0$ such that for every i, j,*

$$\sum_k p_{i,j,k} \le \kappa \frac{m}{d_1 d_2}.$$

Note that if $\kappa = 1$ in Assumption 3.1 then the $p_{i,j,k}$ are all equal, meaning that each user-item-item triple has an equal chance to be observed.

In order to state our error bounds, we first introduce some notation: let $\mathbb{P}_X$ be the distribution of $\{Y_{i,j,k} : 1 \le i \le d_1, 1 \le j < k \le d_2\}$ (i.e. the complete distribution of all pairwise preferences, even those that are not observed).

Our main upper bound shows that if m is sufficiently large then our algorithm finds a solution with almost minimal risk. Given a loss function $\mathcal{L}$, define the expected risk of X by

$$R(X) = \frac{1}{d_1 d_2^2} \sum_{i=1}^{d_1} \sum_{j,k=1}^{d_2} \mathbb{E}_{X^*} \mathcal{L}(Y_{ijk}(\hat{X}_{ij} - \hat{X}_{ik})),$$

where the expectation is with respect to the distribution parametrized by the true parameters X^*.

Theorem 3.1. *Suppose that $\mathcal{L}$ is 1-Lipschitz, and let Y and Ω be distributed as $\mathbb{P}_{X^*}$ for some $d_1 \times d_2$ matrix X^*. Under Assumption 3.1,*

$$\mathbb{E}R(\hat{X}) \le \inf_{\{X : \|X\|_* \le \sqrt{\lambda d_1 d_2}\}} \mathbb{E}R(X) + C\kappa\sqrt{\frac{\lambda(d_1 + d_2)}{m}} \log(d_1 + d_2),$$

where C is a universal constant.

We recall that the parameter λ is related to rank in that if X is a $d_1 \times d_2$ rank-r matrix whose largest absolute entry is bounded by C then $\|X\|_* \le \sqrt{r}\|X\|_F \le C\sqrt{rd_1d_2}$. In other words, λ is a parameter that takes into account both the rank of X^* and the size of its elements, and it is roughly proportional to the rank.

In particular, Theorem 3.1 shows that once we observe $m \sim r(d_1+d_2)\log^2(d_1+d_2)$ pairwise comparisons, then we can accurately estimate the probability of any user preferring any item over any other. In other words, we need to observe about $r(1+d_2/d_1)\log^2(d_1+d_2)$ comparisons per user, which is substantially less than the $rd_2\log(d_2)$ comparisons that we would have required if each user were modelled in isolation. Moreover, our lower bound (below) shows that at least $r(1+d_2/d_1)$ comparisons per user are required, which is only a logarithmic factor from the upper bound.

Theorem 3.2. *Suppose that $\mathcal{L}'(0) < 0$. Let $\mathcal{A}$ be any algorithm that receives $\{Y_{i,j,k} : (i,j,k) \in \Omega\}$ as input and produces $\hat{X}$ as output. For any $\lambda \geq 1$ and $m \geq d_1 + d_2$, there exists X^* with $\|X^*\|_* \leq \sqrt{\lambda d_1 d_2}$ such that when Y and Ω are distributed according to $\mathbb{P}_{X^*}$ then with probability at least $\frac{1}{2}$,*

$$\mathbb{E}R(\hat{X}) \geq R(X^*) + c\min\left\{1, \sqrt{\frac{\lambda(d_1+d_2)}{m}}\right\},$$

where $c > 0$ is a constant depending only on $\mathcal{L}$.

Together, Theorems 3.1 and 3.2 show that (up to logarithmic factors) if X^* has rank r then about $r(1+d_2/d_1)$ comparisons per user are necessary and sufficient for learning the users' preferences.

3.1.1 Maximum likelihood estimation of X^*

By specializing the loss function $\mathcal{L}$, Theorem 3.1 has a simple corollary for maximum-likelihood estimation of X^*. Recall that if μ and ν are two probability distributions on a finite set S the the Kullback-Leibler divergence between them is

$$D(\mu\|\nu) = \sum_{s\in S} \mu(s)\log\frac{\mu(s)}{\nu(s)},$$

under the convention that $0\log 0 = 0$. We recall that although $D(\cdot\|\cdot)$ is not a metric it is always non-negative, and that $D(\mu\|\nu) = 0$ implies $\mu = \nu$.

Corollary 3.3. *Let Y and Ω be distributed as $\mathbb{P}_{X^*}$ for some $d_1 \times d_2$ matrix X^*. Define the loss function $\mathcal{L}$ by $\mathcal{L}(z) = \log(1+\exp(z)) - z$. Under Assumption 3.1,*

$$\frac{1}{d_1 d_2^2} \sup_{\{X:\|X\|_*\sqrt{\lambda d_1 d_2}\}} D(\mathbb{P}_{\hat{X}}\|\mathbb{P}_X) \leq C\kappa\sqrt{\frac{\lambda(d_1+d_2)}{m}}\log(d_1+d_2),$$

where C is a universal constant.

Note that the loss function in Corollary 3.3 is exactly the negative logarithm of the logistic function, and so $\hat{X}$ in Corollary 3.3 is the maximum-likelihood estimate for X^*. Thus, Corollary 3.3 shows that the distribution induced by the maximum-likelihood estimator is close to the true distribution in Kullback-Leibler divergence.

4 Large-scale Non-convex Implementation

While the convex relaxation is statistically near optimal, it is not ideal for large-scale datasets because it requires the solution of a convex program with $d_1 \times d_2$ variables. In this section we develop a non-convex variant which both scales and parallelizes very well, and has better empirical performance as compared to several existing empirical baseline methods.

Our approach is based on the following steps:

1. We represent the low-rank matrix in explicit factored form $X = UV^\top$ and replace the regularizer appropriately. This results in a non-convex optimization problem in $U \in \mathbb{R}^{d_1\times r}$ and $V \in \mathbb{R}^{d_2\times r}$, where r is the rank parameter.

2. We solve the non-convex problem by alternating between updating U while keeping V fixed, and vice versa. With the hinge loss (which we found works best in experiments), each of these becomes an SVM problem - hence we call our algorithm AltSVM.

3. The problem is of course not symmetric in U and V because users rank items but not vice versa. For the U update, each user vector naturally decouples and can be done in parallel (and in fact just reduces to the case of rankSVM Joachims (2002)).

4. For the V update, we show that this can *also* be made into an SVM problem; however it involves coupling of all item vectors, and all user ratings. We employ several tricks (detailed below) to speed up and effectively parallelize this step.

The non-convex problem can be written as

$$\min_{U,V} \sum_{(i,j,k)\in\Omega} \mathcal{L}(Y_{ijk} \cdot u_i^\top (v_j - v_k)) + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2) \tag{3}$$

where we replace the nuclear norm regularizer using the property $\|X\|_* = \min_{X=UV^\top} \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2)$ Srebro et al. (2004). $u_i^\top$ and $v_i^\top$ denote the ith rows of U and V, respectively. While this is a non-convex algorithm for which it is hard to find the global optimum, it is computationally more efficient since only $(d_1 + d_2)r$ variables are involved. We propose to use L2 hinge loss, i.e., $\mathcal{L}(x) = \max(0, 1 - x)^2$.

In the alternating minimization of (3), the subproblem for U is to solve

$$U \leftarrow \arg\min_{U\in\mathbb{R}^{d_1\times r}} \sum_{(i,j,k)\in\Omega} \mathcal{L}(Y_{ijk} \cdot u_i^\top (v_j - v_j)) + \frac{\lambda}{2}\|U\|_F^2, \tag{4}$$

while V is fixed. This can be decomposed into n independent problems for u_i's where each solves for

$$u_i \leftarrow \arg\min_{u\in\mathbb{R}^r} \frac{\lambda}{2}\|u\|_2^2 + \sum_{(j,k)\in\Omega_i} \mathcal{L}(Y_{ijk} \cdot u^\top (v_j - v_k). \tag{5}$$

This part is in general a small-scale problem as the dimension is r, and the sample size is $|\Omega_i|$ for each user i.

On the other hand, solving for V with fixed U can be written as

$$V \leftarrow \arg\min_{V\in\mathbb{R}^{d_2\times r}} \left\{ \frac{\lambda}{2}\|V\|_F^2 + \sum_{(i,j,k)\in\Omega} \mathcal{L}(\langle V, A^{(u,i,j)}\rangle) \right\} \tag{6}$$

where $A^{(i,j,k)} \in \mathbb{R}^{d_2\times r}$ is such that the lth row of $A^{(i,j,k)}$ is $Y_{ijk} \cdot u_i^\top$ if $l = j$, $-Y_{ijk} \cdot u_i^\top$ if $l = k$, and 0 otherwise. It is a much larger SVM problem than (5) as the dimension is $d_2 r$ and the sample size is $|\Omega|$.

We note that the feature matrices $\{A^{(i,j,k)} : (i, j, k) \in \Omega\}$ are highly sparse since in each feature matrix only $2r$ out of the $d_2 r$ elements are nonzero. This motivates us to apply the stochastic dual coordinate descent algorithm Hsieh et al. (2008); Shalev-Shwartz & Zhang (2013), which not only converges fast but also takes advantages of feature sparsity in linear SVMs. Each coordinate descent step takes $O(r)$ computation, and iterations over $|\Omega|$ coordinates provide linear convergence Shalev-Shwartz & Zhang (2013).

Now we describe the dual problems of our two subproblems explicitly. Let $\alpha \in \mathbb{R}^{|\Omega_i|}$ denote the dual vector for (5), in which each coordinate is denoted by α_{ijk} where $(j, k) \in \Omega_i$. Then the dual problem of (5) is to solve

$$\min_{\alpha\in\mathbb{R}^{|\Omega_i|},\alpha\geq 0} \frac{1}{2} \left\| \sum_{(j,k)\in\Omega_i} \alpha_{ijk} Y_{ijk}(v_j - v_k) \right\|_2^2 + \frac{1}{\lambda} \sum_{(j,k)\in\Omega_i} \mathcal{L}^*(-\lambda\alpha_{ijk}) \tag{7}$$

where $\mathcal{L}^*(z)$ is the convex conjugate of $\mathcal{L}$. At each coordinate descent step for α_{ijk}, we find the value of α_{ijk} minimizing (7) while all the other variables are fixed. If we maintain $u_i = \sum_{(j,k)\in\Omega_i} \alpha_{ijk} Y_{ijk}(v_j - v_k)$, then the coordinate descent step is simply to find δ^* minimizing

$$\frac{1}{2}\left\|u_i + \delta^* Y_{ijk}(v_j - v_k)\right\|_2^2 + \frac{1}{\lambda}\mathcal{L}^*(-\lambda(\alpha_{ijk} + \delta^*)) \tag{8}$$

and update $\alpha_{ijk} \leftarrow \alpha_{ijk} + \delta^*$.

The dual problem of (6) is to solve

$$\min_{\beta\in\mathbb{R}^{|\Omega|},\beta\geq 0} \frac{1}{2}\left\|\sum_{(i,j,k)\in\Omega} \beta_{ijk} A^{(i,j,k)}\right\|_F^2 + \frac{1}{\lambda}\sum_{(i,j,k)\in\Omega} \mathcal{L}^*(-\lambda\beta_{ijk}) \tag{9}$$

where β is the dual vector for the subproblem (6). Similarly to α_{ijk}, the coordinate descent step for β_{ijk} is to replace β_{ijk} by $\beta_{ijk} + \delta^*$ where δ^* minimizes

$$\frac{1}{2}\left(\left\|v_j + \delta^* Y_{ijk} u_i\right\|_2^2 + \left\|v_k - \delta^* Y_{ijk} u_i\right\|_2^2\right) + \mathcal{L}^*(-\lambda(\beta_{ijk} + \delta^*)), \tag{10}$$

and maintain $V = \sum_{(i,j,k)\in\Omega} \beta_{ijk} Y_{ijk} A^{(i,j,k)}$.

The detailed description of AltSVM is presented in Algorithm 1. In each subproblem, we run the stochastic dual coordinate descent, in which a pairwise comparison $(i, j, k) \in \Omega$ is chosen uniformly at random, and the dual coordinate descent for α_{ijk} or β_{ijk} is computed. We note that each coordinate descent step takes the same $O(r)$ computational cost in both subproblems, while the subproblem sizes are much different.

4.1 Parallelization

For each subproblem, we parallelize the stochastic dual coordinate descent algorithm asynchronously without locking. Given T processors, each processor randomly sample a triple $(i, j, k) \in \Omega$ and update the corresponding dual variable and the user or item vectors. We note that this update is for a sparse subset of the parameters. In the user part, a coordinate descent step for one sample updates only r out of the rd_1 variables. In the item part, one coordinate descent step for a sample update only $2r$ out of the rd_2 variables. This motivates us not to lock the variables when updated, so that we ignore the conflicts. This lock-free parallelism is shown to be effective in Niu et al. (2011) for stochastic gradient descent (SGD) on the sum of sparse functions. Moreover, in Anonymous (2015), it is also shown that the stochastic dual coordinate descent scales well without locking. We implemented the algorithm using the OpenMP framework. In our implementations, we also parallelized steps 3 and 13 of Algorithm 1. We show in the next section that our proposed algorithm scales up favorably.

4.2 Remark on the implementation

In Algorithm 1, the subproblem for V comes first, and then it solves for the user vectors U. We empirically observed that this order gives better convergence on practical datasets. We also note that each subproblem reuses the dual variables in the previous outer iteration. When almost converged, the features (V for solving U, and U for solving V) do not change too much. By reusing the dual variables in the previous iteration we can start with a feasible solution close to the optimum.

5 Experimental results

5.1 Pairwise data

We used the MovieLens 100k dataset, which contains 100,000 ratings given by 943 users on 1682 movies. The ratings are given as integers from one to five, but we converted them into preference data by declaring

Algorithm 1 Alternating Support Vector Machine (AltSVM)

Require: Ω, $\{Y_{ijk} : (i,j,k) \in \Omega\}$, and $\lambda \in \mathbb{R}^+$
Ensure: $U \in \mathbb{R}^{d_1 \times r}$, $V \in \mathbb{R}^{d_2 \times r}$

Initialize U, and set $\alpha, \beta \leftarrow 0 \in \mathbb{R}^{|\Omega|}$
while not converged **do**
  $v_j \leftarrow \sum_{(i,j,k)\in\Omega} \beta_{ijk} Y_{ijk} u_i - \sum_{(i,k,j)\in\Omega} \beta_{ikj} Y_{ikj} u_i$, $\forall j \in [d_2]$
  for all threads $t = 1, \ldots, T$ **in parallel do**
    for $s = 1, \ldots, S$ **do**
      Choose $(i,j,k) \in \Omega$ uniformly at random
      Find δ^* minimizing (10).
      $\beta_{ijk} \leftarrow \beta_{ijk} + \delta^*$
      $v_j \leftarrow v_j + \delta^* Y_{ijk} u_i$
      $v_k \leftarrow v_k - \delta^* Y_{ijk} u_i$
    end for
  end for
  $u_i \leftarrow \sum_{(i,j,k)\in\Omega} \alpha_{ijk} Y_{ijk} (v_j - v_k)$, $\forall i \in [d_1]$
  for all threads $t = 1, \ldots, T$ **in parallel do**
    for $s = 1, \ldots, S$ **do**
      Choose $(i,j,k) \in \Omega$ uniformly at random.
      Find δ^* minimizing (8).
      $\alpha_{ijk} \leftarrow \alpha_{ijk} + \delta^*$
      $u_i \leftarrow u_i + \delta^* Y_{ijk} (v_j - v_k)$
    end for
  end for
end while

that a user preferred one movie to another if they gave it a higher rating (if two movies received the same rating, we treated it as though the user did not provide a preference). Then we held out 20% of the data as a test set.

We compared our algorithm to the following two:

- Bayesian Personalized Ranking (BPR) Rendle et al. (2009): This algorithm is based on a similar model to ours, but a different optimization procedure (essentially, a variant of stochastic gradient descent).
- Matrix completion from pairwise differences : A standard matrix completion algorithm that observes – for various triples $(i,j,k) \in \Omega$ – the difference between user i's ratings for item j and item k. Note that this algorithm has an advantage over (2) because it sees the magnitude of this difference instead of only its sign. Nevertheless, the matrix completion algorithm does not perform any better than (2). A similar phenomenon was also observed in Davenport et al. (2013).

We evaluate our performance by computing the proportion of pairwise comparisons in the test set $\mathcal{T}$ for which we correctly infer the user's preference.

$$(\text{Prediction error}) = \frac{1}{|\mathcal{T}|} \sum_{(i,j,k)\in\mathcal{T}, Y_{ijk}=1} \mathbb{I}(X_{ij} > X_{ik})$$

This is similar to the AUC statistic measured by Rendle et al. Rendle et al. (2009), and if the data were fully observed then it would measure Kendall's distance between each user's true preferences and the learned ones. However, our main reason for choosing this measure of performance is that, as an average accuracy over all pairwise comparisions, it resembles the quantity that we study in our theoretical bounds.

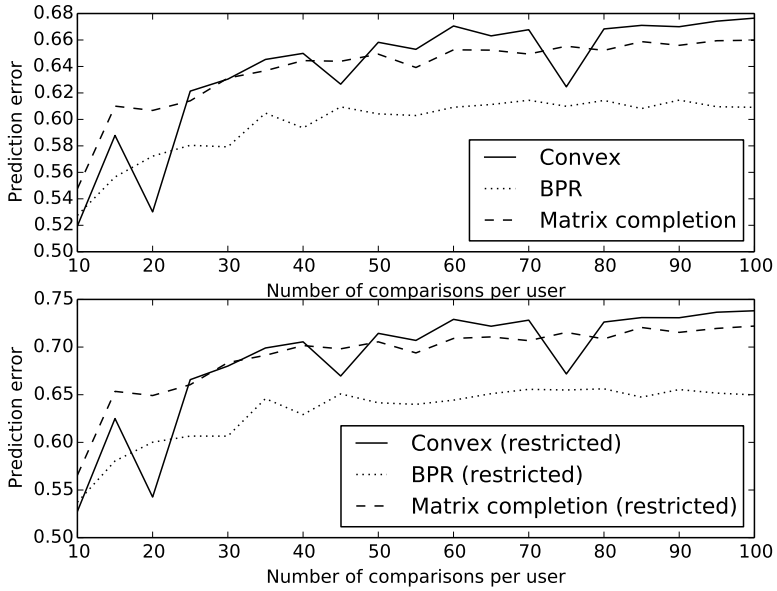


Figure 1: Prediction accuracy on the MovieLens 100k dataset, for different numbers of observed comparisions per user. For the "restricted" plots, only the pairs with a rating difference of two or more were used for evaluation.

Unsurprisingly, we were more accurate at correctly inferring strong preferences; therefore, we have also shown the accuracy obtained by only measuring performance on pairs whose rankings differ by two or more. Both the methods we considered do measurably better at predicting these orderings.

5.2 Large-scale experiments on rating data

Now we demonstrate that our algorithm performs well as a collaborative ranking method on rating data. We used the datasets specified in Table 1. Given a training set of ratings for each user, our algorithm will only use non-tying pairwise comparisons from the set, while other competing algorithms use the ratings themselves. Hence, they have more information than ours. The competing algorithms are those with publicly available codes provided by the authors.

- CofiRank Weimer et al. (2007)[1] This algorithm uses alternating minimization to directly optimize NDCG.
- Local Collaborative Ranking (LCR) Lee et al. (2014)[2] : The main idea is to predict preferences from the weighted sum of multiple low-rank matrices model.
- RobiRank Yun et al. (2014a)[3] : This algorithm uses stochastic gradient descent to optimize the loss function motivated from robust binary classification.

[1] `http://www.cofirank.org`, The dimension and the regularization parameter are set as suggested in the paper. For the rest of the parameters, we left them as provided.

[2] `http://prea.gatech.edu`, We run the code with each of the 48 sets of loss function and parameters given in the main code, and the best result is reported. We could not run this algorithm on the Netflix dataset due to time constraint.

[3] `https://bitbucket.org/d_ijk_stra/robirank`, We used the part for collaborative ranking from binary relevence score. We left the parameter settings as provide with the implementation.

	MovieLens1m	MovieLens10m	Netflix
Users	6,040	71,567	480,000
Items	3,900	10,681	17,000
Ratings	1,000,209	10,000,054	100,000,000

Table 1: Datasets to be used for simulation

- Global Ranking : To see the effect of personalized ranking, we compare the results with a global ranking of the items. We fixed U to all ones and solved for V.

The algorithms are compared in terms of two standard performance measures of ranking, which are NDCG and Precision@K. NDCG@K is the ranking measure for numerical ratings. NDCG@K for user i is defined as

$$\text{NDCG@}K(i) = \frac{\text{DCG@}K(i, \pi_i)}{\text{DCG@}K(i, \pi_i^*)}$$

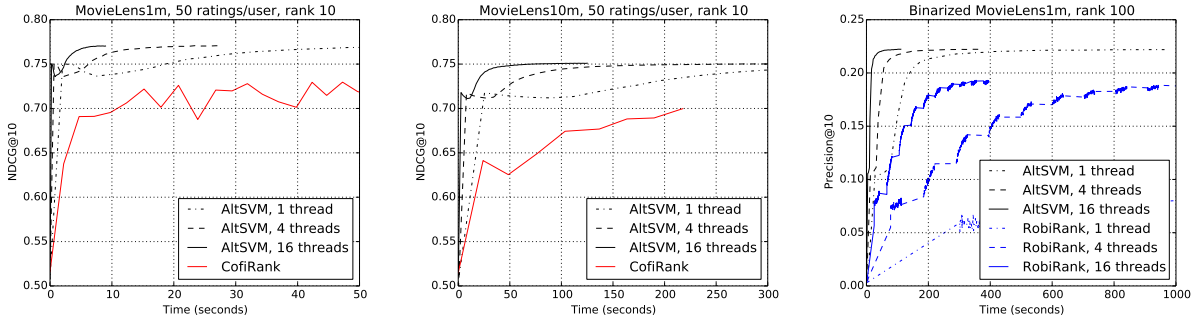
where

$$\text{DCG@}K(i, \pi_i) = \sum_{k=1}^{K} \frac{2^{M_{i\pi_i(k)}} - 1}{\log_2(k+1)},$$

and $\pi_u(k)$ is the index of the kth ranked item of $\mathcal{T}_i$ in our prediction. M_{ij} is the true rating of item j by user i in the given dataset, and π_u^* is the permutation that maximizes DCG@K. This measure counts only the top K items in our predicted ranking and put more weights on the prediction of highly ranked items. We measured NDCG@10 in our experiments. Precision@K is the ranking measure for binary ratings. Precision@K for user i is defined as

$$\text{Precision@}K(i) = \frac{1}{K} \sum_{j \in \mathcal{P}_K(i)} M_{ij}$$

where M_{ij} is the binary rating on item j by user i given in the dataset. This counts the number of relevant items in the predicted top K recommendation. These two measures are averaged over all of the users.



(a) (b) (c)

Figure 2: NDCG@10 and Precision@10 over time for different algorithms.

We first compare our algorithm with numerical rating based algorithms, CofiRank and LCR. We follow the standard setting that are used in the collaborative ranking literature Weimer et al. (2007); Balakrishnan & Chopra (2012); Volkovs & Zemel (2012); Lee et al. (2014). For each user, we subsampled N ratings, used

Datasets	N	AltSVM	AltSVM-sub	Global	CofiRank	LCR
ML1m	20	0.7308	0.6998	**0.7500**	0.7333	0.7007
	50	**0.7712**	0.7392	0.7501	0.7441	0.7081
	100	**0.7902**	0.7508	0.7482	0.7332	0.7151
ML10m	20	0.7059	0.7053	**0.7264**	0.7076	0.6977
	50	**0.7508**	0.7212	0.7176	0.6977	0.6940
	100	**0.7692**	0.7248	0.7101	0.6754	0.6899
Netflix	20	0.7132	0.6822	**0.7605**	0.6615	-
	50	**0.7642**	0.7111	**0.7640**	0.6527	-
	100	**0.8007**	0.7393	0.7656	0.6385	-

Table 2: NDCG@10 on different datasets, for different numbers of observed ratings per user.

	AltSVM			RobiRank
Precision@	$C = 1000$	$C = 2000$	$C = 5000$	
1	0.2165	0.2973	**0.3635**	0.3009
2	0.1965	0.2657	**0.3297**	0.2695
5	0.1572	0.2097	**0.2697**	0.2300
10	0.1265	0.1709	**0.2223**	0.1922
100	0.0526	0.0678	**0.0819**	0.0781

Table 3: Precision@K on the binarized MovieLens1m dataset.

them for training, and took the rest of the ratings for test. The users with less than $N + 10$ ratings were dropped out. Table 2 compares AltSVM with numerical rating based algorithms. While $N = 20$ is too small so that a global ranking provides the best NDCG, our algorithm performs the best with larger N. We also ran our algorithm with subsampled pairwise comparions with the largest numerical gap (AltSVM-sub), which are as many as N for each user (the number of numerical ratings used in the other algorithms). Even with this, we could achieve better NDCG.

We have also experimented collaborative ranking on binary ratings. Our algorithm is compared with RobiRank Yun et al. (2014a), which is a recently presented algorithm for collaborative ranking with binary ratings. We ran an experiment on a *binarized* version of the Movielens1m dataset. In this case, the movies rated by a user is assumed to be relevant to the user, and the other items are not. Since it is inefficient to take all possible comparisons which are in average a half million per user, we subsampled C comparisons for each user. Both algorithms are set to estimate rank-100 matrices. Table 3 shows that our algorithm provides better performance than RobiRank.

5.3 Computational speed and Scalability

We now show the computational speed and scalability of our practical algorithm, AltSVM. The experiments were run on a single 16-core machine in the Stampede Cluster at University of Texas.

Figures 2a and 2b show NDCG@10 over time of our algorithms with 1, 4, and 16 threads, compared to CofiRank. Figure 2c shows Precision@10 over time of our algorithm with $C = 5000$. We note that our algorithm converges faster, while the sample size $|\Omega|$ for our algorithm is larger than the number of training ratings that are used in the competing algorithms. Table 4 shows the scalability of AltSVM. We measured the time to achieve 10^{-5} tolerance on the binarized MovieLens1m dataset. As can be seen in the table, we could achieve significant speedup.

References

Ailon, Nir. Active learning ranking from pairwise preferences with almost optimal query complexity. In *NIPS*, pp. 810–818, 2011.

Anonymous. Passcode: Parallel asynchronous stochastic dual co-ordinate descent. submitted to ICML, 2015.

# cores	1	2	4	8	16
Time(seconds)	963.1	691.8	365.1	188.3	111.0
Speedup	1x	1.4x	2.6x	5.1x	8.7x

Table 4: Scalability of AltSVM on the binarized MovieLens1m dataset.

Balakrishnan, Suhrid and Chopra, Sumit. Collaborative ranking. In *ACM International Conference on Web Search and Data Mining (WSDM)*, 2012.

Bradley, Ralph Allan and Terry, Milton E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, pp. 324–345, 1952.

Davenport, Mark A, Plan, Yaniv, Berg, Ewout van den, and Wootters, Mary. 1-bit matrix completion. *arXiv preprint arXiv:1209.3672*, 2013.

Hajek, Bruce, Oh, Sewoong, and Xu, Jiaming. Minimax-optimal inference from partial rankings. In *NIPS*, 2014.

Herbrich, Ralf, Graepel, Thore, and Obermayer, Klaus. *Large Margin Rank Boundaries for Ordinal Regression*, chapter 7, pp. 115–132. MIT Press, January 2000.

Hsieh, Cho-Jui, Chang, Kai-Wei, Lin, Chih-Jen, Keerthi, S. Sathiya, and Sundararajan, S. A dual coordinate descent method for large-scale linear SVM. In *ICML*, 2008.

Hu, Yifan, Koren, Yehuda, and Volinsky, Chris. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pp. 263–272. IEEE, 2008.

Jamieson, K. G. and Nowak, R. Active ranking using pairwise comparisons. In *NIPS*, 2011a.

Jamieson, Kevin G. and Nowak, Robert D. Active ranking using pairwise comparisons. In *NIPS*, 2011b.

Joachims, Thorsten. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

Lee, Joonseok, Bengio, Samy, Kim, Seungyeon, Lebanon, Guy, and Singer, Yoram. Local collaborative ranking. In *International World Wide Web Conference (WWW)*, 2014.

Liu, Nathan N, Zhao, Min, and Yang, Qiang. Probabilistic latent preference analysis for collaborative filtering. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 759–766. ACM, 2009.

Liu, Tie-Yan. *Learning to Rank for Information Retrieval*. Now Publishers Inc., 2009.

Lu, Yu and Negahban, Sahand. Individualized rank aggregation using nuclear norm regularization. *ArXiv e-prints: 1410.0860*, Oct 2014.

Luce, Duncan R. *Individual Choice Behavior*. Wiley, 1959.

Negahban, Sahand, Oh, Sewoong, and Shah, Devavrat. Iterative ranking from pair-wise comparisons. In *NIPS*, 2012.

Niu, Feng, Recht, Benjamin, Ré, Christopher, and Wright, Stephen. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *NIPS*, 2011.

Rendle, Steffen, Freudenthaler, Christoph, Gantner, Zeno, and Schmidt-Thieme, Lars. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461. AUAI Press, 2009.

Seginer, Yoav. The expected norm of random matrices. *Combinatorics Probability and Computing*, 9(2): 149–166, 2000.

Shalev-Shwartz, Shai and Zhang, Tong. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research (JMLR)*, pp. 567–599, 2013.

Shi, Yue, Karatzoglou, Alexandros, Baltrunas, Linas, Larson, Martha, Oliver, Nuria, and Hanjalic, Alan. Climf: collaborative less-is-more filtering. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 3077–3081. AAAI Press, 2013.

Srebro, Nathan, Rennie, Jason, and Jaakkola, Tommi. Maximum margin matrix factorization. In *NIPS*, 2004.

Vershynin, Roman. *Compressed sensing: theory and applications*, chapter Introduction to the non-asymptotic analysis of random matrices. Cambridge University Press, 2012.

Volkovs, Maksims N. and Zemel, Richard S. Collaborative ranking with 17 parameters. In *NIPS*, 2012.

Wauthier, Fabian L., Jordan, Michael I., and Jojic, Nebojsa. Efficient ranking from pairwise comparisons. In *ICML*, 2013.

Weimer, Markus, Karatzoglou, Alexandros, Le, Quoc V., and Smola, Alex. Cofirank: maximum margin matrix factorization for collaborative ranking. In *NIPS*, 2007.

Weston, Jason, Want, Chong, Weiss, Ron, and Berenzeig, Adam. Latent collaborative retrieval. In *ICML*, 2012.

Xu, Jiaming, Wu, Rui, Zhu, Kai, Hajek, Bruce, Srikant, R, and Ying, Lei. Jointly clustering rows and columns of binary matrices: Algorithms and trade-offs. In *ACM Sigmetrics*, 2013.

Yi, Jinfeng, Jin, Rong, Jain, Shaili, and Jain, Anil. Inferring users preferences from crowdsourced pairwise comparisons: A matrix completion approach. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

Yun, Hyokun, Raman, Parameswaran, and Vishwanathan, S. V. N. Ranking via robust binary classification and parallel parameter estimation in large-scale data. In *NIPS*, 2014a.

Yun, Hyokun, Yu, Hsiang-Fu, Hsieh, Cho-Jui, Viswanathan, S. V. N., and Dhillon, Inderjit S. NOMAD: Non-locking, stochastic multi-machine algorithm for asynchronous and descentralized matrix completion. In *VLDB*, 2014b.

A Proof of Theorem 3.1

We write $L(X)$ for the function being optimized; i.e.,

$$L(X) = \sum_{(i,j,k)\in\Omega} \mathcal{L}(Y_{i,j,k}(X_{i,j} - X_{i,k})).$$

Note that for any fixed X, $\mathbb{P}_{X^*}L(X) = mR(X)$ (where $\mathbb{P}_{X^*}$ denotes the expectation taken with respect to future samples from $\mathbb{P}_{X^*}$, as distinct from $\mathbb{E}$ which denotes the expectation over the samples used to generate $\hat{X}$). Let K be the set of $d_1 \times d_2$ matrices with nuclear norm at most 1. The proof of Theorem 3.1 proceeds in three main steps.

1. By some algebraic of manipulations L, we reduce the problem to showing a uniform law of large numbers for the family of functions $\{L(X) : X \in \sqrt{\lambda d_1 d_2}K\}$.

2. Using symmetrization and duality properties of K, we reduce the problem to bounding the norm of a matrix M whose entries are sums of random signs.

3. We bound the norm of M using various concentration inequalities and a theorem of Seginer Seginer (2000).

Since $\hat{X}$, by definition, minimizes $L(\hat{X})$, for any $\tilde{X} \in \sqrt{\lambda d_1 d_2} K$ we can bound

$$\begin{aligned}\mathbb{P}_{X^*}[L(\hat{X}) - L(\tilde{X})] &\le \mathbb{P}_{X^*}[L(\hat{X})] - L(\hat{X}) - \left(\mathbb{P}_{X^*}[L(\tilde{X})] - L(\tilde{X})\right) \\ &\le 2 \sup_{X \in \sqrt{\lambda d_1 d_2} k} |\mathbb{P}_{X^*} L(X) - L(X)|.\end{aligned}$$

In other words, it suffices to show a uniform law of large numbers for $\{L(X) : X \in \sqrt{\lambda d_1 d_2} K\}$.

Let $\epsilon_{i,j,k}$ be i.i.d. ± 1-valued variables and let $\xi_{i,j,k}$ be the indicator that $(i,j,k) \in \Omega$. By Giné-Zinn's symmetrization (as in Davenport et al. (2013)),

$$\begin{aligned}&\mathbb{E} \sup_{X \in \sqrt{\lambda d_1 d_2} K} |\mathbb{P}_{X^*} L(X) - L(X)| \\ &\le 2\mathbb{E} \sup_{X \in \sqrt{\lambda d_1 d_2} K} \left| \sum_{i,j,k \in \Omega} \epsilon_{i,j,k} \mathcal{L}(Y_{i,j,k}(X_{i,j} - X_{i,k})) \right|.\end{aligned}$$

Since $\mathcal{L}$ is 1-Lipschitz, we obtain

$$\begin{aligned}\mathbb{E} \sup_{X \in \sqrt{\lambda d_1 d_2} K} |\mathbb{P}_{X^*}[L(X)] - L(X)| &\le 2\mathbb{E} \sup_{X \in \sqrt{\lambda d_1 d_2} K} \left| \sum_{i,j,k \in \Omega} \epsilon_{i,j,k} Y_{i,j,k}(X_{i,j} - X_{i,k}) \right| \\ &= 2\mathbb{E} \sup_{X \in \sqrt{\lambda d_1 d_2} K} \left| \sum_{i,j,k} \xi_{i,j,k} \epsilon_{i,j,k}(X_{i,j} - X_{i,k}) \right|,\end{aligned}$$

where in the last line, we recognized that $\epsilon_{i,j,k} Y_{i,j,k}$ has the same distribution as $\epsilon_{i,j,k}$. Now, let M denote the matrix where $M_{ij} = \sum_k (\xi_{i,j,k}\epsilon_{i,j,k} - \xi_{i,k,j}\epsilon_{i,k,j})$. Then

$$\sum_{i,j,k} \xi_{i,j,k}\epsilon_{i,j,k}(X_{i,j} - X_{i,k}) = \operatorname{tr}(M^T X)$$

and so

$$\sup_{X \in \sqrt{\lambda d_1 d_2} K} \sum_{i,j,k} \xi_{i,j,k}\epsilon_{i,j,k}(X_{i,j} - X_{i,k}) = \sup_{X \in \sqrt{\lambda d_1 d_2} K} \operatorname{tr}(M^T X) = \sqrt{\lambda d_1 d_2} \|M\|.$$

Putting everything together, we have (for any $\tilde{X} \in \sqrt{\lambda d_1 d_2} K$)

$$\mathbb{E}\left[\mathbb{P}_{X^*}[L(\hat{X})] - \mathbb{P}_{X^*}[L(\tilde{X})]\right] \le 4\sqrt{\lambda d_1 d_2}\mathbb{E}\|M\|.$$

Together with the following lemma (which we prove in Appendix B), this completes the proof of Theorem 3.1

Lemma A.1. *With* $p = \frac{m}{d_1 d_2}$,

$$\mathbb{E}\|M\| \le C\kappa\sqrt{p(d_1 + d_2)}\log(d_1 d_2).$$

B Proof of Lemma A.1

We will decompose M into two parts, $M = M^{(1)} - M^{(2)}$, with

$$M^{(1)}_{ij} = \sum_{k \neq j} \xi_{i,j,k} \epsilon_{i,j,k}$$
$$M^{(2)}_{ij} = \sum_{k \neq j} \xi_{i,k,j} \epsilon_{i,k,j}.$$

Then $\|M\| \leq \|M^{(1)}\| + \|M^{(2)}\|$. Since $M^{(1)}$ and $M^{(2)}$ have the same distribution,

$$\mathbb{E}\|M\| \leq 2\mathbb{E}\|M^{(1)}\|,$$

and so we are reduced to studying $M^{(1)}$, which has i.i.d. entries. Now, we apply Seginer's theorem Seginer (2000):

$$\mathbb{E}\|M^{(1)}\| \leq C\left(\mathbb{E}\max_i \|M^{(1)}_{i*}\|_2 + \mathbb{E}\max_j \|M^{(1)}_{*j}\|_2\right), \tag{11}$$

where $M^{(1)}_{i*}$ denotes the ith row of $M^{(1)}$ and $M^{(1)}_{*j}$ denotes the jth column, and $\|\cdot\|_2$ denotes the Euclidean norm.

We will separate the task of bounding $\mathbb{E}\max_i \|M^{(1)}_{i*}\|_2$ into two parts: if $\|x\|_0$ denotes the number of non-zero coordinates in x and $\|x\|_\infty$ denotes $\max_j |x_j|$ then $\|x\|_2 \leq \sqrt{\|x\|_0}\|x\|_\infty$; with the Cauchy-Schwarz inequality, this implies that

$$\left(\mathbb{E}\left[\max_i \|M^{(1)}_{i*}\|_2\right]\right)^2 \leq \mathbb{E}\left[\max_i \|M^{(1)}_{i*}\|_0\right]\mathbb{E}\left[\max_i \|M^{(1)}_{i*}\|^2_\infty\right] \tag{12}$$

First, we will show that every row of $M^{(1)}$ is sparse. Let $Z_{ij} = \sum_{k\neq j}\xi_{i,j,k}$ and let Y_{ij} be the indicator that $Z_{ij} > 0$. Recalling that $\mathbb{E}\xi_{i,j,k} = p_{i,j,k}$, we have (by Assumption 3.1) $\mathbb{E}Z_{ij} \leq \kappa p$. Since Z_{ij} takes non-negative integer values, we have $\Pr(Y_{ij} = 1) = \Pr(Z_{ij} > 0) \leq \kappa p$. By Bernstein's inequality, for any fixed i

$$\Pr(\|M^{(1)}_{i*}\|_0 \geq \kappa d_2 p + t) \leq \Pr(\sum_{j=1}^{d_2} Y_{ij} \geq \kappa d_2 p + t) \leq \exp\left(-\frac{t^2/2}{\kappa p d_2 + t/3}\right).$$

Integrating by parts, we have

$$\mathbb{E}\left[\|M^{(1)}_{i*}\|_0\right] \leq \kappa d_2 p + \int_{\kappa d_2 p}^\infty \Pr(\|M^{(1)}_{i*}\|_0 \geq t)\, dt \leq \kappa d_2 p + \frac{3}{8}.$$

Next, we will consider the size of the elements in $M^{(1)}$. First of all, $M^{(1)}_{ij} \leq Z_{ij}$ (this fairly crude bound will lose us a factor of $\sqrt{\log(d_1 d_2)}$). Now, Bernstein's inequality applied to Z_{ij} gives

$$\Pr(M^{(1)}_{ij} \geq \kappa p + t) \leq \Pr(Z_{ij} \geq \kappa p + t) \leq \exp\left(-\frac{t^2/2}{\kappa p + t/3}\right).$$

Taking a union bound over i and j, if $t \geq C\kappa\log(d_1 d_2)$ then

$$\Pr(\max_{ij} M^{(1)}_{ij} \geq t) \leq d_1 d_2 \exp(-ct) \leq \exp(-c't).$$

Integrating by parts,

$$\mathbb{E}\left[\max_{ij} M^{(1)}_{ij}\right] \leq \kappa\log^2(d_1 d_2) + \int_{\kappa\log^2(d_1 d_2)}^\infty \Pr(\max_{ij} M^{(1)}_{ij} \geq \sqrt{t})\, dt \leq \kappa\log^2(d_1 d_2) + C.$$

Going back to (12), we have shown that

$$\mathbb{E}\max_i \|M_{i*}^{(1)}\| \le C\kappa\sqrt{pd_2}\log(d_1d_2).$$

The same argument applies to $M_{*j}^{(1)}$ (but with $\sqrt{pd_1}$ instead of $\sqrt{pd_2}$), and so we conclude from (11) that

$$\mathbb{E}\|M^{(1)}\| \le C\kappa\sqrt{p(d_1+d_2)}\log(d_1d_2).$$

C Proof of Theorem 3.2

C.1 A sketch of the proof

The proof of Theorem 3.2 uses Fano's inequality.

1. We construct matrices $X^1, \dots, X^\ell$. These matrices all have small nuclear norm, and for every pair i, j the KL-divergence between the induced observation distributions is $\Theta(\log \ell)$. We construct these matrices randomly, using concentration inequalities and a union bound to show that we can take ℓ of the order $\sqrt{\lambda m(d_1+d_2)}$.
2. We apply Fano's inequality to show that if we generate data according to a randomly chosen X^i, then any algorithm has a reasonable chance to choose a different X^j (using the fact that the KL-divergence is $O(\log \ell)$). Since the KL-divergence is $\Omega(\log \ell)$, this implies that the algorithm incurs a substantial penalty whenever it makes a wrong choice.

In any application of Fano's inequality, the key is to construct a large number of admissible models that are close to one another in KL-divergence. Specifically, if we can construct distributions $\mathbb{P}_1, \dots, \mathbb{P}_\ell$ with $D(\mathbb{P}_i\|\mathbb{P}_j) + 1 \le \frac{1}{2}\log \ell$ for all i, j, then given a single sample from some $\mathbb{P}_i$, no algorithm can accurately identify which $\mathbb{P}_i$ it came from. In order to apply this denote by $\mathbb{P}_{X,m}$ the distribution of the data when the true parameters are X. We will construct $X^1 \dots, X^\ell \in \sqrt{\lambda d_1 d_2}K$ such that for all $i \ne j$,

$$D(\mathbb{P}_{X^i,m}\|\mathbb{P}_{X^j,m}) + 1 \le \frac{1}{2}\log \ell, \tag{13}$$

$$R_j(X^i) \ge R_j(X^j) + c\frac{\log \ell}{m} \tag{14}$$

for some constant $c > 0$, where R_j denotes the expected risk when the true parameters are given by X^j. Given a single observation from some $\mathbb{P}_{X^j,m}$, (13) will imply (by Fano's inequality) that no algorithm can correctly identify which X^j was the true parameter. On the other hand, (14) will imply that if the algorithm makes a mistake – say it chooses X^i for $i \ne j$ – then its risk will be $c\frac{\log \ell}{m}$ larger than the best in the class. In particular, if we can prove (13) and (14) with $\log \ell \sim \sqrt{\lambda m(d_1+d_2)}$ then it will imply Theorem 3.2.

We construct a set of matrices satisfying (13) and (14) using a probabilistic method. Supposing that $d_2 \ge d_1$, we choose a parameter $\gamma > 0$ and set B to be an integer that is approximately $\lambda\gamma^{-2}$. We define X^1 by filling its top $B \times d_2$ block with independent, uniform $\pm\gamma$ entries, and then copying that top block B/d_1 times to fill the matrix. Then let $X^2, \dots, X^\ell$ be independent copies of X^1. First of all, each $X^i \in \sqrt{\lambda d_1 d_2}K$ because $\|X^i\|_* \le \sqrt{\mathrm{rank}(X^i)}\|X^i\|_F \le \sqrt{\lambda d_1 d_2}$.

Now, let us consider $D(\mathbb{P}_{X^1,m}\|\mathbb{P}_{X^2,m})$. For a single i, j, k triple, there is probability $1/4$ of having $X^1_{i,j} - X^1_{i,k}$ different from $X^2_{i,j} - X^2_{i,k}$, in which case they differ by 4γ. If γ is bounded above, each different entry contributes $\Theta(\alpha^2\gamma^2)$ to the KL-divergence between $\mathbb{P}_{X^1,m}$ and $\mathbb{P}_{X^2,m}$. Since about m entries are observed in $\mathbb{P}_{X^1,m}$, we see that

$$D(\mathbb{P}_{X^1,m}\|\mathbb{P}_{X^2,m}) \asymp m\gamma^2. \tag{15}$$

On the other hand, $R_1(X^1)$ and $R_1(X^2)$ differ by $\Theta(\gamma^2)$, because for a constant fraction of triples i, j, k, the chance that $Y_{i,j,k}$ is 1 differs by $O(\gamma)$ in X^1 and X^2, and on the event that $Y_{i,j,k}$ differs in these two models the loss differs by another $O(\gamma)$ factor.

Applying standard concentration inequalities, we show that one can apply the union bound to $\ell = \exp(cBd_2)$ of these matrices. In view of (13) and (15), we need to take $Bd_2 = \frac{\lambda^2}{\gamma^2 d_1} \asymp m\gamma^2$. Eliminating γ, we end up with $\log \ell \asymp \sqrt{\lambda m/d_1}$ (which is within a constant factor of $\sqrt{\lambda m(d_1+d_2)}$ under our assumption that $d_2 \geq d_1$).

C.2 Some concentration lemmas

We begin by quoting some standard concentration results (see, e.g. Vershynin (2012)).

Definition C.1. *A random variable X is σ^2-subgaussian if $\mathbb{E}e^{\theta X} \leq e^{\theta^2\sigma^2/2}$ for all $\theta > 0$. A random variable X is L-subexponential if $\mathbb{E}e^{\theta X} \leq (1-\theta^2 L^2)$ for $\theta < 1/L$.*

One can easily show that the product of two subgaussian variables is subexponential:

Lemma C.2. *If X is σ^2-subgaussian and Y is τ^2-subgaussian then XY is $C\sigma\tau$-subexponential for a universal constant C.*

Moreover, one has a Bernstein-type inequality for sums of independent subexponential variables.

Lemma C.3. *If $X_1, \ldots, X_k$ are i.i.d. L-subexponential then*

$$\Pr(\sum_i X_i \geq t) \leq \exp\left(-\frac{ct^2}{L^2 k + Lt}\right).$$

C.3 Construction of a packing set

Let $0 < \gamma < 1$ be some parameter to be determined such that $B := \lambda\gamma^{-2}$ is an integer.

Proposition C.4. *Suppose that $\mathcal{L}'(0) < 0$. For every sufficiently small γ (depending on $\mathcal{L}$), there exists a set $\mathcal{X} \subset \sqrt{\lambda d_1 d_2} K$ of $\exp(cBd_2)$ $d_1 \times d_2$ matrices such that for any two $X^1, X^2 \in \mathcal{X}$,*

$$\frac{1}{d_1 d_2^2} \sum_{i=1}^{d_1} \sum_{j,k=1}^{d_2} \mathbb{E}_{X^1}[\mathcal{L}(Y(X^2_{ij} - X^2_{ik})) - \mathcal{L}(Y(X^1_{ij} - X^1_{ik}))] \geq c\gamma^2$$

and for any m,

$$\frac{1}{m} D(\mathbb{P}_{X^1,m} \| \mathbb{P}_{X^2,m}) \leq C\gamma^2,$$

where $0 < c < C$ are universal constants.

Following Davenport et al., we construct this set $\mathcal{X}$ randomly: let X be a random $B \times d_2$ matrix, where each element is chosen independently to be either γ or $-\gamma$.

Lemma C.5. *Let X^1 and X^2 be independent copies of X. Then with probability at least $1 - \exp(-cBd_2)$,*

$$\sum_{i=1}^{B} \sum_{j,k=1}^{d_2} (X^1_{ij} - X^1_{ik} - X^2_{ij} + X^2_{ik})^2 \geq 2\gamma^2 B d_2^2,$$

where $c > 0$ is a universal constant.

Before proving Lemma C.5, let us see how it implies Proposition C.4. First of all, for X a random $B \times d_2$ matrix as above, let $\tilde{X}$ be the $d_1 \times d_2$ matrix obtained by stacking $\lceil d_1/B \rceil$ copies of X, and filling out any remaining entries by zeros. Then, for random X and Y, with high probability

$$\begin{aligned}\sum_{i=1}^{d_1} \sum_{j,k=1}^{d_2} (\tilde{X}^1_{ij} - \tilde{X}^1_{ik} - \tilde{X}^2_{ij} + \tilde{X}^2_{ik})^2 &= \lceil d_1/B \rceil \sum_{i=1}^{B} \sum_{j,k=1}^{d_2} (X^1_{ij} - X^1_{ik} - X^2_{ij} + X^2_{ik})^2 \\ &\asymp \gamma^2 d_1 d_2^2, \end{aligned} \tag{16}$$

where the lower bound for the last line came from Lemma C.5, and the upper bound just came from the observation that each term in the sum is bounded by $16\gamma^2$. Let $\mathcal{X}$ be the set obtained by choosing $\exp(cBd_2/4)$ random copies of $\tilde{X}$ in this way. The high-probability estimate in Lemma C.5 implies that with high probability, *every* pair $\tilde{X}^1, \tilde{X}^2$ in $\mathcal{X}$ satisfies (16). Now,

$$\begin{aligned} D(\mathbb{P}_{X^1,m} \| \mathbb{P}_{X^2,m}) &= \mathbb{E}_\Omega \left[\sum_{(i,j,k)\in\Omega} D(f(X^1_{ij} - X^1_{ik}) \| f(X^2_{ij} - X^2_{ik})) \right] \\ &\asymp \frac{m}{d_1 d_2^2} \sum_{i,j,k} (X^1_{ij} - X^1_{ik} - X^2_{ij} + X^2_{ik})^2, \end{aligned}$$

where $f(x) = e^x/(1+e^x)$ is the logistic function, and the last line follows from a Taylor expansion of $D(f(x)\|f(y))$ around $x = y$, because all the X^1_{ij} and X^2_{ij} are bounded by $\gamma < 1$. Together with (16), this proves the first inequality in Proposition C.4; the second inequality follows because each term of the form $D(f(X_{ij} - X_{ik})\|f(Y_{ij} - Y_{ik}))$ is bounded by a constant times γ^2. This proves the second inequality of Proposition C.4.

By Taylor expansion again, if γ is sufficiently small (depending on $\mathcal{L}$) then

$$\mathcal{L}(Y_{i,j,k}(X^2_{i,j} - X^2_{i,k})) - \mathcal{L}(Y_{i,j,k}(X^1_{i,j} - X^1_{i,k})) \asymp Y_{i,j,k}(X^1_{i,j} - X^1_{i,k} - X^2_{i,j} + X^2_{i,k}).$$

Now, if i, j, k is a triple for which $2\gamma = X^1_{i,j} - X^1_{i,k} > X^2_{i,j} - X^2_{i,k}$ (and under the event of Lemma C.5, there are at least cBd_2^2 such triples) then $\mathbb{E}_{X^1}[Y_{i,j,k}] \asymp \gamma$ and so

$$\mathbb{E}_{X^1}[\mathcal{L}(Y_{i,j,k}(X^2_{i,j} - X^2_{i,k})) - \mathcal{L}(Y_{i,j,k}(X^1_{i,j} - X^1_{i,k}))] \asymp \gamma^2.$$

The same holds when i, j, k is a triple for which $-2\gamma = X^1_{i,j} - X^1_{i,k} < X^2_{i,j} - X^2_{i,k}$. Finally, if i, j, k is a triple such that $X^1_{i,j} - X^1_{i,k} = X^2_{i,j} - X^2_{i,k}$ then the expectation is zero. Summing over all triples, we see that on the event that Lemma C.5 holds,

$$\frac{1}{Bd_2^2} \sum_{i,j,k} \mathbb{E}_{X^1}[\mathcal{L}(Y_{i,j,k}(X^2_{i,j} - X^2_{i,k})) - \mathcal{L}(Y_{i,j,k}(X^1_{i,j} - X^1_{i,k}))] \geq c\gamma^2.$$

After summing over all $\lceil d_1/B \rceil$ blocks, this proves the first inequality of Proposition C.4.

Proof of Lemma C.5. We expand the square:

$$\begin{aligned} \sum_{ijk} (X_{ij} - X_{ik} - Y_{ij} + Y_{ik})^2 &= 2\sum_{ijk} X_{ij}^2 + Y_{ij}^2 + 2X_{ij}Y_{ik} - X_{ij}X_{ik} - Y_{ij}Y_{ik} - 2X_{ij}Y_{ij} \\ &= 4\gamma^2 Bd_2^2 + 2\sum_{ijk} 2X_{ij}Y_{ik} - X_{ij}X_{ik} - Y_{ij}Y_{ik} - 2X_{ij}Y_{ij}. \end{aligned} \tag{17}$$

We may study each of the cross-terms separately: for the $X_{ij}Y_{ik}$ term, note that $\sum_j X_{ij}$ and $\sum_k Y_{ik}$ are both $\gamma^2 d_2$-subgaussian (by Hoeffding's inequality). Hence, $\sum_{jk} X_{ij}Y_{ik}$ is $C\gamma^2 d_2$-subexponential (by Lemma C.2) and so by Lemma C.3,

$$\Pr\left(\left| \sum_{ijk} X_{ij}Y_{ik} \right| \geq \frac{1}{8}\gamma^2 Bd_2^2 \right) \leq 2\exp(-cBd_2).$$

The similar argument applies to the $X_{ij}X_{ik}$ term: $\sum_j X_{ij}$ is $\gamma^2 d_2$-subgaussian and so $\sum_{ijk} X_{ij}X_{ik} = \sum_i (\sum_j X_{ij})^2$ is $C\gamma^2 d_2$-subexponential; hence

$$\Pr\left(\left| \sum_{ijk} X_{ij}X_{ik} \right| \geq \frac{1}{8}\gamma^2 Bd_2^2 \right) \leq 2\exp(-cBd_2).$$

Of course, the $Y_{ij}Y_{ik}$ term is identical. Finally, note that $\sum_{ijk} X_{ij}Y_{ij} = d_2 \sum_{ij} X_{ij}Y_{ij}$. Since the terms in this sum are i.i.d., we may apply Hoeffding's inequality to obtain

$$\Pr\left(\left|\sum_{ijk} X_{ij}Y_{ij}\right| \geq \frac{1}{8}\gamma^2 B d_2^2\right) = \Pr\left(\left|\sum_{ij} X_{ij}Y_{ij}\right| \geq \frac{1}{8}\gamma^2 B d_2\right) \leq 2\exp(-cB^2 d_2^2).$$

Putting everything together, we see that with high probability, the total of all the cross-terms in (17) is at most half of the first term. □

C.4 Completing the proof

Let C denote the constant from Proposition C.4. Assume that $d_1 \leq d_2$ and that m is large enough so

$$\sqrt{\frac{d_2}{m}} \leq 8C\sqrt{\lambda} \leq \sqrt{\frac{m}{d_2}}. \tag{18}$$

Note that under the assumptions $\lambda \geq 1$ and $m \geq d_1 + d_2$ from Theorem 3.2, the lower bound of (18) is satisfied. Moreover, if the upper bound of (18) is not satisfied then we may decrease λ until it is; the conclusion of Theorem 3.2 will not be affected because as long as (18) fails, the minimum in Theorem 3.2 will be 1.

By the lower bound in (18), there is an integer B such that

$$B \leq \sqrt{\frac{\lambda m}{d_2}} \leq 2B;$$

fix this B and define γ by

$$\gamma^2 = \lambda/B \asymp \sqrt{\frac{\lambda d_2}{m}}.$$

By the upper bound in (18), $\gamma \leq 1$.

Now, Fano's inequality states that if we first select a random $X \in \mathcal{X}$ and then draw a sample from $\mathbb{P}_{X,m}$, then any algorithm trying to identify X can succeed with probability at most

$$\frac{\min\{D(\mathbb{P}_{X,m} \| \mathbb{P}(Y,m)) : X, Y \in \mathcal{X}\} + 1}{\log|\mathcal{X}|} \leq \frac{2Cm\gamma^2}{Bd_2} \leq \frac{1}{2}.$$

Finally, note that by the first inequality in Proposition C.4, the error incurred by choosing the wrong $X \in \mathcal{X}$ is at least $c\gamma^2 \asymp \sqrt{\frac{\lambda d_2}{m}}$.

Now, we have so far only discussed the case $d_2 \geq d_1$. The case $d_1 \leq d_2$ is not exactly equivalent because our model is not symmetric in its treatment of users and items. However, the proof of Theorem 3.2 does not change very much. We take horizontally stacked blocks of size $d_1 \times B$ instead of $B \times d_2$. The main difference is in the calculation leading to (16): there are extra cross-terms appearing due to the fact that items in different blocks need to be compared with one another. However, all of these additional terms may be controlled with Lemmas C.2 and C.3 in much the same way as the existing terms are controlled.

D Comparison to Stochastic Gradient Descent

Another practical algorithm to optimize (3) is Stochastic Gradient Descent (SGD). We have experimented SGD on the same datasets in Table 1. We ran the algorithm with the same regularization parameters and different step sizes. The statistical results for SGD were observed to be no better than AltSVM, and hence we did not present them in the main paper.

Datasets	N	NDCG@10
ML1m	20	0.6852
	50	0.7666
	100	0.7728
ML10m	20	0.6977
	50	0.7452
	100	0.7659

Table 5: NDCG@10 of SGD on different datasets, for different numbers of observed ratings per user.

Precision@	SGD with $C = 5000$
1	0.1556
2	0.1498
5	0.1236
10	0.1031
100	0.0441

Table 6: Precision@K for SGD of (3) on the binarized MovieLens1m dataset.

Let us first describe the SGD procedure. At each step, ones chooses a triple $(i, j, k) \in \Omega$ uniformly at random and run a SGD step, which can be written as

$$u_i^+ \leftarrow u_i - \eta \cdot \left\{ g \cdot (v_j - v_k) + \frac{\lambda}{|\Omega_i|} u_i \right\}$$
$$v_j^+ \leftarrow v_j - \eta \cdot \left\{ g \cdot u_i + \frac{\lambda}{|\Omega^j|} v_j \right\}$$
$$v_j^+ \leftarrow v_j - \eta \cdot \left\{ -g \cdot u_i + \frac{\lambda}{|\Omega^k|} v_k \right\}$$

where $\Omega^{(j)}$ denotes the number of comparisons in Ω which involve item j. η is a step size and $g \in \partial\mathcal{L}(u_i^\top (v_j - v_k))$.

The following tables show the statistical result of SGD. The step size is chosen by $\eta = \frac{\alpha}{1+\beta t}$ as suggested in Yun et al. (2014b). α and β were the powers of 10^{-1}, and the best result is reported. The results are comparable to AltSVM, but it did not achieve better results. We note that this is the best result from several different step sizes, while AltSVM does not have any other parameter to choose except for the regularization parameter.