Generalization Error Bounds for Collaborative Prediction with Low-Rank Matrices

Nathan Srebro **Tommi Jaakkola**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA
nati,tommi@mit.edu

Abstract

We prove generalization error bounds for predicting entries in a partially observed matrix by approximating the observed entries with a low-rank matrix. To do so, we bound the number of sign configurations of low-rank matrices using a result about realizable oriented matroids.

1 Introduction

"Collaborative filtering" refers to the general task of providing users with information on what items they might like, or dislike, based on their preferences so far and how they relate to the preferences of other users. This approach contrasts with a more traditional feature-based approach where predictions are made based on features of the items.

For feature-based approaches, we are accustomed to studying prediction methods in terms of probabilistic post-hoc generalization error bounds. Such results provide us a (probabilistic) bound on the performance of our predictor on future examples, in terms of its performance on the training data. These bounds hold without any assumptions on the true "model", that is the true dependence of the labels on the features, other than the central assumptions that the training examples are drawn i.i.d. from the distribution of interest.

In this paper we suggest studying the generalization ability of collaborative prediction methods. By "collaborative prediction" we indicate that the objective is to be able to predict user preferences for items, that is, entries in some unknown *target matrix* Y of user-item "ratings", based on observing a subset Y_S of the entries in this matrix[1]. We present bounds on the true average overall error $\mathcal{D}(X;Y) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{a=1}^{m}\text{loss}(X_{ia};Y_{ia})$ of the predictions X in terms of the average error over the observed entries $\mathcal{D}_S(X;Y) = \frac{1}{|S|}\sum_{ia\in S}\text{loss}(X_{ia};Y_{ia})$, without making any assumptions on the true nature of the preferences Y. What we do assume is that the subset S of entries that we observe is chosen uniformly at random. This strong assumption parallels the i.i.d. source assumption for feature-based prediction.

[1] In other collaborative filtering tasks, the objective is to be able to provide each user with a few items that overlap his top-rated items, while it is not important to be able to correctly predict the users ratings for other items. Note that it is possible to derive generalization error bounds for this objective based on bounds for the "prediction" objective.

arbitrary source distribution	⇔	target matrix Y
random training set	⇔	random set S of observed entries
hypothesis	⇔	predicted matrix X
training error	⇔	observed discrepancy $\mathcal{D}_S(X;Y)$
generalization error	⇔	true discrepancy $\mathcal{D}(X;Y)$

Figure 1: Correspondence with post-hoc bounds on the generalization error for standard feature-based prediction tasks

In particular, we present generalization error bounds on prediction using low-rank models.

Collaborative prediction using low-rank models is fairly straight forward. A low-rank matrix X is sought that minimizes the average observed error $\mathcal{D}_S(X;Y)$. Unobserved entries in Y are then predicted according to X. The premise behind such a model is that there are only a small number of factors influencing the preferences, and that a user's preference vector is determined by how each factor applies to that user. Different methods differ in how they relate real-valued entries in X to preferences in Y, and in the associated measure of discrepancy. For example, entries in X can be seen as parameters for a probabilistic models of the entries in Y, either mean parameters [1, 2] or natural parameters [3, 4], and a maximum likelihood criterion used. Or, other loss functions, such as squared error [5, 3], or zero-one loss versus the signs of entries in X, can be minimized.

Prior Work Previous results bounding the error of collaborative prediction using a low-rank matrix all assume the true target matrix Y is well-approximated by a low-rank matrix. This corresponds to a large *eigengap* between the top few singular values of Y and the remaining singular values. Azar *et al* [5] gives asymptotic results on the convergence of the predictions to the true preferences, assuming they have an eigengap. Drineas *et al* [6] analyzes the sample complexity needed to be able to predict a matrix with an eigengap, and suggests strategies for actively querying entries in the target matrix. To our knowledge, this is the first analysis of the generalization error of low-rank methods that do not make any assumptions on the true target matrix.

Generalization error bounds (and related online learning bounds) were previously discussed for collaborative prediction applications, but only when prediction was done for each user separately, using a feature-based method, with the other user's preferences as features [7, 8]. Although these address a collaborative prediction application, the learning setting is a standard feature-based setting. These methods are also limited, in that learning must be performed separately for each user.

Shaw-Taylor *et al* [9] discuss assumption-free post-hoc bounds on the residual errors of low-rank approximation. These results apply to a different setting, where a subset of the rows are fully observed, and bound a different quantity—the distance between rows and the learned *subspace*, rather then the distance to predicted entries.

2 Generalization Error Bound for Low Rank Matrices

Focusing on binary labels $Y_{ia} \in \pm$ and a zero-one sign agreement loss[2] , loss$(X_{ia}; Y_{ia}) = \mathbf{1}_{Y_{ia} \neq \text{sign}^{\pm} X_{ia}}$, our main result is:

Theorem 1. *For* any *matrix* $Y \in \{\pm 1\}^{n \times m}$, $n, m > 2$, $\delta > 0$ *and integer* k, *with probability at least* $1 - \delta$ *over choosing a subset* S *of entries in* Y *uniformly among all subsets*

[2] We denote sign $^{\pm}x = \begin{cases} + & \text{If } x \geq 0 \\ - & \text{otherwise} \end{cases}$, distinguishing it from the ternary variant with sign $0 = 0$.

of $|S|$ entries:

$$\forall_{X,\operatorname{rank} X<k}\mathcal{D}(X;Y) < \mathcal{D}_S(X;Y) + \sqrt{\frac{k((k+1)n+m)\log n - \log\delta}{2|S|}}$$

To prove the theorem, first fix Y as well as $X \in \mathbb{R}^{n\times m}$. When an index pair (i,a) is chosen uniformly at random, $\operatorname{loss}(X_{ia};Y_{ia})$ is a Bernoulli random variable with probability $\mathcal{D}(X;Y)$ of being one. If the entries of S are chosen independently and uniformly, $|S|\mathcal{D}_S(X;Y)$ is Binomially distributed with mean $|S|\mathcal{D}(X;Y)$ and using Chernoff's inequality:

$$\Pr_S\left(\mathcal{D}(X;Y) \geq \mathcal{D}_S(X;Y) + \epsilon\right) \leq e^{-2|S|\epsilon^2} \tag{1}$$

The distribution of S in Theorem 1 is slightly different, as S is chosen without repetitions. The mean of $\mathcal{D}_S(X;Y)$ is the same, but it is more concentrated, and (1) still holds.

Now consider all rank-k matrices. Noting that $\operatorname{loss}(X_{ia};Y_{ia})$ depends only on the *sign* of X_{ia}, it is enough to consider the equivalence classes of matrices with the same sign patterns. Let $f(n,m,k)$ be the number of such equivalence classes, i.e. the number of possible sign configurations of $n \times m$ matrices of rank at most k:

$$F(n,m,k) = \{\operatorname{sign}^{\pm} X \in \{-,+\}^{n\times m} | X \in \mathbb{R}^{n\times m}, \operatorname{rank} X \leq k\}$$
$$f(n,m,k) = \sharp F(n,m,k)$$

where $\operatorname{sign}^{\pm} X$ denotes the element-wise sign matrix $(\operatorname{sign}^{\pm} X)_{ia} = \begin{cases} 1 & \text{If } X_{ia} \geq 0 \\ -1 & \text{otherwise} \end{cases}$. For all matrices in an equivalence class, the random variable $\mathcal{D}_S(X;Y)$ is the same, and taking a union bound of the events $\mathcal{D}(X;Y) \geq \mathcal{D}_S(X;Y)+\epsilon$ for each of these $f(n,m,k)$ random variables we have:

$$\Pr_S\left(\exists_{X,\operatorname{rank} X\leq k}\mathcal{D}(X;Y) \geq \mathcal{D}_S(X;Y) + \sqrt{\frac{\log f(n,m,k) - \log\delta}{2|S|}}\right) \leq \delta \tag{2}$$

by using (1) and setting $\epsilon = \sqrt{\frac{\log f(n,m,k)-\log\delta}{2|S|}}$. The proof of Theorem 1 rests on bounding $f(n,m,k)$, which we will do in the next section.

Note that since the equivalence classes we defined do not depend on the sample set, no symmetrization argument is necessary. One might suggest improving the bound using more specific equivalence classes, considering only the sign configurations of entries in S. However, not much can be gained from such refinements. Consider, for example, bounding the number of S-specific equivalence classes by $f(n,m,k,|S|) \leq |S|^V$ using VC-dimension arguments. Then we have $f(n,m,k) \leq (nm)^V$, and since for meaningful sample sizes $|S| \geq \max(n,m)$ (otherwise we cannot hope to generalize), the improvement in the bound is by at most a constant factor of two, which is lost in the symmetrization arguments. Bounding the growth function $f(n,m,k,|S|)$ directly might yield improvements for specific sample size, but since $f(n,m,k) \leq f(n,m,k,|S|)^{\log nm}$, the improvement would not be by more than a factor of $\log nm$.

3 Sign Configurations of a Low-Rank Matrix

We would like to bound the number $f(n,m,k)$ of possible sign configurations of $n \times m$ rank-k matrices over the reals. Any matrix X of rank at most k can be written as a product $X = UV$ where $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{k\times m}$. Let us first count how many sign configurations we can get for a fixed U, i.e. $F(U,m,k) = \{\text{sign } UV | V \in \mathbb{R}^{k\times m}\}$.

Lemma 2. $f(U,m,k) = \sharp F(U,m,k) < \left(5n^{k-1}\right)^m$

Proof. Interpreting each of the n rows of U as a point in $\mathbb{R}^k$, we can interpret each non-zero column v of V as specifying a homogeneous hyperplane partitioning the points in U. The corresponding column $\mathrm{sign}^{\pm}(Uv)$ of $\mathrm{sign}^{\pm}(UV)$ is a sign vector specifying the a classification of the points in U by the hyperplane v. The number of possibilities for each column is one (accounting for an all-zero v) plus the number of possible classification of n points in $\mathbb{R}^k$ using homogeneous hyperplanes, which is at most $2\sum_{i=0}^{k-1}\binom{n}{i} \le 4n^{k-1}$ [10, Theorem 3.20][3]. Multiplying the number of possibilities for each of the m columns, and bounding $1 + 4n^{k-1} \le 5n^{k-1}$ yields the desired bound. □

To bound $f(n,m,k)$ using Lemma 2 we need to bound the number of possible matrices U, or more accurately, the number of equivalence classes of matrices yielding the same set of possible column vectors $\mathcal{M}'(U) = \left\{\mathrm{sign}^{\pm}(Uv) | v \in \mathbb{R}^k\right\}$. In deriving generalization error bounds we are accustomed to bounding the number of possible classifiers for a sample set. Here, we need to bound the number of different *sample sets*: the number of possible samples of n points in $\mathbb{R}^k$ (the rows of U) that can be linearly separated in different ways. To do so, we consider slightly more relaxed equivalence classes known as *realizable oriented matroid*, for which such a bound is known.

Definition 1. *[11] The* oriented matroid realized *by the n rows of $U \in \mathbb{R}^{n\times k}$ is defined by the set of* covectors $\mathcal{M}(U) = \left\{\textit{sign}\ (Uv) \in \{-,0,+\}^n | v \in \mathbb{R}^k\right\}$.

The number of different oriented matroids realizable by n points in $\mathbb{R}^k$ is at most $n^{k(k+1)n}$[12, 13]. Since two matrices realizing the same oriented matroid also realize the same set of possible classifications (which are just projections of the covectors with 0 mapping to $+$), the same bound applies also to the number of equivalence classes of matrices U yielding identical $\mathcal{M}'(U)$. Multiplying this bound by the bound on the number of different sign configurations for each $\mathcal{M}'(U)$ (Lemma 2) we can conclude:

Theorem 3. $f(n,m,k) \le 5^m n^{k(k+1)n+m(k-1)}$

Taking the logarithm, and bounding $5\log m \le m\log n$ for $n, m > 2$, we have

$$\log f(n,m,k) \le k((k+1)n + m)\log n. \tag{3}$$

Together with (2), this proves Theorem 1.

Since $f(n,m,k)$ is symmetric in n and m, we can also state:

$$\log f(n,m,k) \le k(n + (k+1)m)\log m. \tag{4}$$

suggesting that the true bound on $\log g(n,m,k)$ might be of the form $O(k(n+m)\log(nm))$, avoiding the k^2 term. Note that the logarithmic term is unavoidable, and this is the best we can expect:

Lemma 4. *For $n > k^2$, $f(n,m,k) \ge 2^{\frac{1}{2}(k-1)m\log n}$*

Proof. Fix any matrix U with rows in general position, then for $n > k^2$:

$$\begin{aligned}\log f(n,m,k) \ge f(U,m,k) = |\mathcal{M}'(U)|^m &\ge \left(2\sum_{i=0}^{k-1}\binom{n}{i}\right)^m \\ &\ge \binom{n}{k-1}^m \ge \left(\frac{n}{k-1}\right)^{m(k-1)} \ge \sqrt{n}^{m(k-1)} = 2^{\frac{1}{2}(k-1)m\log n}\end{aligned}$$

□

[3] This is exact for points in general position. Note that this is less than the bound guaranteed by Sauer's lemma, with $k-1$ instead of k in the exponent

Overlaps between realizable oriented matroids The reasons the bound in Theorem 3 might be loose, with an excessive k^2 term, is that when $m < |\mathcal{M}'(U)| = \Theta(n^{k-1})$, we are not able to use all classification vectors in $\mathcal{M}'(U)$ (we can only use m of them). Thus, even if there are a large number of realizable oriented matroids, if many of them overlap greatly, the resulting sets $F(U, m, k)$ also overlap greatly. Summing the sizes of these sets, as we do in deriving Theorem 3, we are ignoring the overlaps between them. Bounding the extent of these overlaps might lead to a tighter bound on $f(n, m, k)$.

Ternary-sign configurations and point-hyperplane co-arrangements For classification purposes, we are only concerned with binary sign configurations, where (arbitrarily) $\operatorname{sign}^{\pm} 0 = +$. We can easily modify Theorem 3 to bound "ternary" sign configurations, distinguishing between zeros and positive entries, obtaining a bound of $(2(k+1))^m n^{k(k+1)n+m(k-1)}$. Such sign configurations can also be interpreted as follows: An oriented matroid specifies the configuration of n points with respect to all possible homogeneous hyperplanes. Instead, consider arrangements of n labeled points and m labeled homogeneous hyperplanes, where two arrangement are equivalent when each point in one arrangement lies on the same side of each hyperplane as the corresponding point in the second arrangements.

4 Low-Rank Matrices as Combined Classifiers

Rank-k matrices are those matrices which are linear combinations of k rank-1 matrices. If we view matrices as functions from pairs of indexes to the reals, we can think of rank-k matrices as "combined" classifiers, and attempt to bound their VC-dimension (and thus also $f(n, m, k)$) as such.

Rank one matrices can be written as an outer product of two vectors, and the sign-configuration of the matrix depends only on the signs and zeros of the two vectors. Therefore, the number of sign-configurations of such matrices is bounded by $f(n, m, 1) < 3^{mn}$, and their VC-dimension as indicator functions is at most $(\log 3)(n+m)$. But although we can bound the VC-dimension of combination of *indicator* functions with low VC-dimension, and thus can bound the VC-dimension of combinations of signs of rank-one matrices, this is not enough in order to bound the VC-dimension of low-rank matrices, which are combinations of real-valued functions.

To be able to discuss combinations of real-valued functions, we must use a quantity that accounts for the complexity of the function also away from zero, and not only for the complexity of the signs of the function. One such standard measure is the *pseudodimension* of a class of functions: The *pseudodimension* $\dim_\Phi \mathcal{F}$ of a class of real-valued functions $\mathcal{F} = \{f : Q \to \mathbb{R}\}$ is the VC-dimension of the class of subsets

$$\Phi_{\mathcal{F}} = \{\, \phi_f = \{\, (q,p) \in Q \times \mathbb{R} \mid f(q) \geq p\} \mid f \in \mathcal{F}\} \,.$$

We conjecture that combinations of real-valued functions with low *pseudodimension* do exhibit a low pseudodimension:

Conjecture 5. $\dim_\Phi \sum_k \mathcal{F} \leq \tilde{O}(k \dim_\Phi \mathcal{F})$*, where* $\sum_k \mathcal{F} = \{f_1 + f_2 + \ldots + f_k | f_i \in \mathcal{F}\}$ *and the* $\tilde{O}()$ *notation indicates possible log factors.*

The conjecture is true for indicator functions. For real valued functions, the metric entropy (uniform bound on log covering number), which can be bounded in terms of the pseudodimension, does display a graceful scaling with respect to linear combinations. However, the metric entropy is a scale-sensitive measure, and thus meaningful only when we bound the entries in the matrix, while the pseudodimension is not scale-sensitive[4]. The question of

[4] Note that, under certain conditions, the metric entropy is bounded also for (unlimited) convex

the pseudodimension of combinations of low-pseudodimension real-valued functions was also raised by Dudley [14], and to our knowledge is unresolved.

We now analyze the pseudodimension of rank one matrices. We do so by bounding, for any threshold matrix $P \in \mathbb{R}^{n\times m}$ the number of relative sign matrices:

$$F_P(n,m,1) = \{\text{sign}^{\pm}(u'v - P) | u \in \mathbb{R}^n, v \in \mathbb{R}^m\}$$

Lemma 6. *For any $P \in \mathbb{R}^{n\times m}$, we have $f_P(n,m,1) = \sharp F_P(n,m,1) < (n+1)^m 2(n+1)\left(mn^2\right)^{n-1}$.*

Proof. For a fixed $u \in \mathbb{R}^n$, consider the number of possibilities for each column $\text{sign}^{\pm}(u'v_j - P_j)$ of the relative sign matrix. Varying v_j from negative infinity to positive infinity, each sign $\text{sign}^{\pm}(u_i v_j - P_{ij})$ can change at most once (or not change at all if $u_j = 0$), yielding at most $n+1$ possible sign vectors $\text{sign}^{\pm}(u'v_j - P_j)$ for a fixed u, and so at most $(n+1)^m$ sign matrices $\text{sign}^{\pm}(u'v - P)$ for a fixed u.

We must now bound the number of vectors $u \in \mathbb{R}^n$ yielding different possibilities for $\text{sign}^{\pm}(u'v - P)$. The sign pattern of u determines the sign patters of each column $\text{sign}^{\pm}(u'v_j - P_j)$ for extreme (positive and negative infinity) values of v_j. Beyond the sign pattern of u, the range of attainable sign vectors $\text{sign}^{\pm}(u'v_j - P_j)$ (for varying v_j) is determined by the order in which the signs $\text{sign}^{\pm}(u_i v_j - P_{ij})$ are flipped as v_j varies from negative infinity to positive infinity. The sign $\text{sign}^{\pm}(u_\alpha v_j - P_{\alpha j})$ will be flipped before $\text{sign}^{\pm}(u_\beta v_j - P_{\beta j})$ if $\frac{P_{\alpha j}}{u_\alpha} < \frac{P_{\beta j}}{u_\beta}$. For any column P_j, and for a fixed sign pattern of u, the order in which the signs of $\text{sign}^{\pm}(u'v_j - P_j)$ are flipped, and so the range of attainable sign columns, is determined by the signs of $(P_{\beta j}u_\alpha - P_{\alpha j}u_\beta)$. Overall then, the sign matrices $\text{sign}^{\pm}(u'v - P)$ attainable by some $u \in \mathbb{R}^n$ is determined by the location of u relative to the n hyperplanes $u_i = 0$ and the $m\binom{n}{2}$ hyperplanes $P_{\beta j}u_\alpha = P_{\alpha j}u_\beta$. That is, each u yielding a different range of sign matrices corresponds to a different covector in an oriented matroid realized by $n + m\binom{n}{2} < nm^2$ hyperplane normals in R^n. The number of covectors in such an oriented matroid is at most $2(n+1)\left(mn^2\right)^{n-1}$ (note that this is more than the number of classifications, as the distinction of being on a hyperplane or on its positive side is important).

Multiplying the number of different sign configurations possible with any one u by the number of vectors u yielding different possibilities yields the desired bound, $(n+1)^m 2(n+1)\left(mn^2\right)^{n-1}$. □

Theorem 7. *The pseudodimension of rank one, $n \times m$ matrices over the reals is less than $2(n+m)\log(n+m)$.*

Proof. Consider the largest subset $S \subset ([n] \times [m]) \times \mathbb{R}$ that is shattered by $\Phi_{\mathcal{R}_1}$. For S to be shattered, there cannot be to different points in S corresponding to the same index pair, and so S can be seen as defining a partial matrix $P \in \mathbb{R}^{n\times m}$ with $P_{ij} = p$ for each $(ij,p) \in S$. Shattering S corresponds to attaining all possible $2^{|S|}$ sign patterns relative to these entries of P. Filling in the other entries of P arbitrarily, we get at least $2^{|S|}$ possible sign patterns relative to this complete P. But we know from Lemma 6 that for any matrix, we cannot get more then $(n+1)^m 2(n+1)\left(mn^2\right)^{n-1}$ sign patterns relative to it, and so $\dim_\Phi \mathcal{R}_1 = |S| < \log\left((n+1)^m 2(n+1)\left(mn^2\right)^{n-1}\right) < 2(m+n)\log mn$. □

combinations, whereas the pseudodimension of a convex combination of a small class of classifiers can be infinite. However, although bounds for convex combinations are inherently scale sensitive, bounds for finite linear combinations can be combinatorial, as in the case, for example for a linear combination of indicator functions of low VC-dimension.

5 Other Loss Functions

In Section 2 we considered generalization error bounds for a zero-one loss function. More commonly, though, other loss functions are used, and it is desirable to obtain generalization error bounds for general loss functions. If Conjecture 5 is true, we could bound the pseudodimension of low-rank matrices, and obtain generalization error bounds for any bounded monotone loss function, over all low-rank matrices. Here, we present a more limited approach, which requires bounding the entries of the low-rank matrix. We do so by calculating the covering number of bounded-entry low-rank matrices.

Lemma 8. *For any $B, \epsilon > 0$, there exists an ϵ-net $\mathcal{N}$ of $\left(\frac{2Bk}{\epsilon}\right)^{k(n+m)}$ matrices, such that for any rank-k matrix $X \in [-B,B]^{n\times m}$, there exists a matrix $\tilde{X} \in \mathcal{N}$ with $|X_{ia} - \tilde{X}_{ia}| < \epsilon$ for all i, a.*

Proof. Construct the set $\mathcal{N}$ by taking products $\tilde{U}\tilde{V}'$ of all $n \times k$ matrices $\tilde{U}$ and $m \times k$ matrices $\tilde{V}$ with values in $[-\sqrt{B}, \sqrt{B}]$ discratized to $2\epsilon' = \frac{\epsilon}{k\sqrt{B}}$. A rank-$k$ matrix $X \in [-B,B]^{n\times m}$ can be factored to $X = UV'$, with $U \in [-\sqrt{B}, \sqrt{B}]^{n\times k}, V \in [-\sqrt{B}, \sqrt{B}]^{n\times k}$. For $\tilde{X} = \tilde{U}\tilde{V}' \in \mathcal{N}$ where $|U_{i\alpha} - \tilde{U}_{i\alpha}| < \epsilon'$ and $|V_{a\alpha} - \tilde{U}_{a\alpha}| < \epsilon'$ for all i, a, α, we have:

$$\begin{aligned}\left|X_{ia} - \tilde{X}_{ia}\right| &= \left|\sum_\alpha U_{i\alpha}V_{a\alpha} - \sum_\alpha \tilde{U}_{i\alpha}\tilde{V}_{a\alpha}\right| \\ &= \left|\sum_\alpha \tilde{U}_{i\alpha}(\tilde{V}_{a\alpha} - V_{a\alpha}) + \sum_\alpha V_{a\alpha}(\tilde{U}_{i\alpha} - U_{i\alpha})\right| \\ &\leq \sum_\alpha \sqrt{B}\left|\tilde{V}_{a\alpha} - V_{a\alpha}\right| + \sum_\alpha \sqrt{B}\left|\tilde{U}_{a\alpha} - U_{a\alpha}\right| \leq 2k\sqrt{B}\epsilon' = \epsilon\end{aligned}$$

□

We can now use arguments similar to those of Theorem 1 to provide a generalization error bound for any Lipschitz bounded loss function, by approximating each matrix with a matrix in $\mathcal{N}$ and taking a union bound over them.

Theorem 9. *For any L-Lipschitz continuous loss function taking values in the interval $[0,1]$, any matrix Y, $\delta > 0$ integer k, and $B > 0$, with probability at least $1 - \delta$ over choosing a subset S of entries in Y uniformly among all subsets of $|S|$ entries:*

$$\forall_{X \in [-B,B]^{n\times m}, \operatorname{rank} X < k} \mathcal{D}(X;Y) < \mathcal{D}_S(X;Y) + \sqrt{\frac{k(n+m)\log\frac{14LB\sqrt{k|S|}}{\sqrt{n+m}} - \log\delta}{|S|}}$$

Proof. Fix X and consider $\text{loss}(X_{ia}; Y_{ia})$ as a random variable where (i,a) is chosen uniformly at random. Then $\mathcal{D}_S(X;Y)$ is the average of $|S|$ such random variables, each with mean $\mathcal{D}(X;Y)$ and bounded between zero and one, and using Hoeffding's inequality (and arguments similar to those in Theorem 1) we have:

$$\Pr_S\left(\mathcal{D}(X;Y) \geq \mathcal{D}_S(X;Y) + \epsilon_1\right) \leq e^{-2|S|\epsilon_1^2} \qquad (5)$$

To get a bound which is uniform over all rank-k $X \in [-B,B]^{n\times m}$, consider an ϵ_2-net $\mathcal{N}$ for these matrices guaranteed by Lemma 8. The Lipschitz continuity ensures this is an $L\epsilon_2$-net for the elementwise loss matrices, i.e. for any rank-k $X \in [-B,B]^{n\times m}$, there exists

$\tilde{X} \in \mathcal{N}$ with $|\text{loss}(X_{ia}; Y_{ia}) - \text{loss}(\tilde{X}_{ia}; Y_{ia})| \leq L\epsilon_2$ for all i, a and so also $|\mathcal{D}(X;Y) - \mathcal{D}(\tilde{X};Y)| \leq L\epsilon_2$ and $|\mathcal{D}_S(X;Y) - \mathcal{D}_S(\tilde{X};Y)| \leq L\epsilon_2$ yielding $\mathcal{D}(X;Y) - \mathcal{D}_S(X;Y) \leq \mathcal{D}(\tilde{X};Y) - \mathcal{D}_S(\tilde{X};Y) + 2L\epsilon_2$. Taking a union bound over the "bad" events $\mathcal{D}(\tilde{X};Y) \geq \mathcal{D}_S(\tilde{X};Y) + \epsilon_1$ for all matrices $\tilde{X} \in \mathcal{N}$ we have (for rank-k B-bounded matrices):

$$\Pr_S \left(\exists_X \mathcal{D}(X;Y) - \mathcal{D}_S(X;Y) > \epsilon_1 + 2L\epsilon_2\right) \leq e^{-2|S|\epsilon_1^2} \left(\frac{2kB}{\epsilon_2}\right)^{k(n+m)}$$

Setting $\epsilon_1 = \sqrt{\frac{k(n+m)\log\frac{2LBk}{\epsilon_2} - \log\delta}{2|S|}}$ and $\epsilon_2 = \sqrt{\frac{k(n+m)}{48L^2|S|}}$ we get the desired bound. □

6 Discussion

We suggest a framework for studying the generalization ability of collaborative prediction methods, and present a first generalization error bound for collaborative prediction with low-rank models. The core component of this bound is a combinatorial result on the number of sign configurations of low-rank matrices, the proof of which is based on classic results by Goodman, Pollack and Alon on the number of realizable oriented matroids. The connection between machine learning and oriented matroids is a natural one, as the covectors of an oriented matroid can be viewed as possible classification vectors. In statistical machine learning, or empirical process theory, one is usually satisfied with a bound on the *size* of an oriented matroid (the number of covectors, or number of different classifications of a sample set), and uses Sauer's lemma to obtain it. Here, we require also a bound on the *number* of oriented matroids, or number of different sample sets. The theory of oriented matroids provides us with such a bound. A better understanding the overlaps between realizable oriented matroids, and not only their number, can help improve the bound we present.

Beyond borrowing results about oriented matroids, the object of study itself, i.e. sign configurations of low rank matrices, has a natural interpretation in terms of point and hyperplane arrangements.

Here, we present a bound on the logarithm of the number of sign configurations which is of the form $O((kn + k^2 m)\log(nm))$. Although this bound already yields useful bounds on the generalization error, we conjecture that the dependence on k^2 is extraneous, and that the true bound is of the form $\Theta(k(n+m)\log(nm))$. This type of dependence more closely matches the parametric dimension of rank-k matrices and the lower bound (Lemma 4).

We also suggest an alternative approach for bounding the number of sign configurations of low rank matrices, by considering such matrices as convex combinations of rank-one matrices, and bounding their pseudodimension. This approach is based on a conjecture about the pseudodimension of combined classifiers. We note that proving this conjecture will not only tighten the bounds we present here, and yield generalization error bounds for low-rank collaborative prediction with general loss functions, but would also allow relaxing boundedness assumptions in general results about combined classifiers.

Acknowledgments We would like to thank Erik Demaine for directing us to oriented matroids and John Barnett for showing that a bounded-entry low-rank matrix has a bounded-entry factorization.

References

[1] Thomas Hofmann. Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst.*, 22(1):89–115, 2004.

[2] Benjamin Marlin. Modeling user rating profiles for collaborative filtering. In *NIPS*17*, 2004.

[3] Nathan Srebro and Tommi Jaakkola. Weighted low rank approximation. In *20th International Conference on Machine Learning*, 2003.

[4] Benjamin Marlin and Richard S. Zemel. The multiple multiplicative factor model for collaborative filtering. In *To appear in ICML*, 2004.

[5] Yossi Azar, Amos Fiat, Anna R. Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *ACM Symposium on Theory of Computing*, pages 619–626, 2001.

[6] Petros Drineas, Iordanis Kerenidis, and Prabhakar Raghavan. Competitive recommendation systems. In *ACM Symposium on Theory of Computing*, 2002.

[7] K. Crammer and Y. Singer. Pranking with ranking. In *NIPS*14*, 2002.

[8] Sanjoy Dasgupta, Wee Sun Lee, and Philip M. Long. A theoretical analysis of query selection for collaborative filtering. *Machine Learning*, 51(3):283–298, 2003.

[9] John Shawe-Taylor, Nello Cristianini, and Jaz Kandola. On the concentration of spectral properties. In *NIPS*15*, 2002.

[10] Gerhard Wesp. *Counting certain covectors in oriented matroids*. PhD thesis, Slazburg, 1999.

[11] A. Bj orner, M. Las Vergnas, B. Strumfels, N. White, and G. Ziegler, editors. *Oriented Matroids*. Cambridge University Press, 2nd edition edition, 1999.

[12] Jacob Goodman and Richard Pollack. Upper bounds for configurations and polytopes in $\mathbb{R}^d$. *Discrete and Computational Geometry*, 1:219–227, 1986.

[13] Noga Alon. The number of polytopes, configurations and real matroids. *Mathematika*, 33:62–71, 1986.

[14] R. M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15(4), 1987.