Tighter Generalization Bounds for Matrix Completion Via Factorization Into Constrained Matrices

Ken-ichiro MORIDOMI†*, Nonmember, Kohei HATANO†**, and Eiji TAKIMOTO1, Members

SUMMARY We prove generalization error bounds of classes of low-rank matrices with some norm constraints for collaborative filtering tasks. Our bounds are tighter, compared to known bounds using rank or the related quantity only, by taking the additional $L_1$ and $L_\infty$ constraints into account. Also, we show that our bounds on the Rademacher complexity of the classes are optimal.

key words: matrix completion, non-negative matrix factorization, collaborative filtering, Rademacher complexity, generalization error bound

1. Introduction

Learning preferences of users over a set of items is an important task in recommendation systems. In particular, the collaborative filtering approach is known to be quite effective and popular[1]–[3]. Simply put, the collaborative filtering is an approach of inferring user’s rating for an item, which the user has not rated yet, from the existing ratings of other users. The approach is formulated as the matrix completion problem, that is, learning a user-item rating matrix from given partial entries of the matrix. More formally, we consider a (true) rating matrix $X \in \mathbb{R}^{N \times M}$ to be learned, where $N$ and $M$ are the numbers of users and items, respectively, and each component $X_{i,j}$ corresponds to user $i$’s rating for item $j$. The task is to find a hypothesis matrix $\hat{X} \in \mathbb{R}^{N \times M}$ that approximates the true matrix $X$ when only some of components of $X$ are given as a sample.

A common assumption in the previous work is that the true matrix $X \in \mathbb{R}^{N \times M}$ can be well approximated by a matrix of low rank (or low trace norm, as a convex relaxation of the rank constraint). In other words, we assume that our hypothesis matrix $\hat{X} \in \mathbb{R}^{N \times M}$ can be decomposed as $\hat{X} = UV^T$ for some $U \in \mathbb{R}^{N \times K}$ and $V \in \mathbb{R}^{M \times K}$ with a small number $K$, where $K$ gives an upper bound of the rank of $\hat{X}$. Generalization ability of algorithms (such as the empirical risk minimization) using low rank or low trace norm matrices is intensively studied in the literature (see, e.g., [4]–[7]).

Recently, further additional constraints on the class of hypothesis matrices turns out to be effective in practice. In particular, a major approach is to impose the constraints that $U$ and $V$ are non-negative with bounded norm [8]–[13]. Such a decomposition is called the non-negative matrix factorization (NMF, for short). More precisely, a typical scheme of the NMF approach is the empirical risk minimization with the norm regularization, formulated as the following optimization problem: Find non-negative matrices $U \in \mathbb{R}^{N \times K}$ and $V \in \mathbb{R}^{M \times K}$ that minimizes

$$\sum_{(i,j)} \ell((UV^T)_{i,j}, X_{i,j}) + a\|U_{i,:}\|^2 + \beta\|V_{j,:}\|^2,$$

where the sum is over all components $(i, j)$ in the sample that the learner observes, $\ell$ is a fixed loss function, $\alpha, \beta$ are regularization parameters, and $\|U_{i,:}\|$ and $\|V_{j,:}\|$ are the $L_2$-norm of the $i$-th row of $U$ and that of the $j$-th row of $V$, respectively [9], [10], [13]. Note that this formulation is essentially equivalent to the empirical risk minimization:

$$\min_{U \in \mathbb{R}^{N \times K}, V \in \mathbb{R}^{M \times K}} \sum_{(i,j)} \ell((UV^T)_{i,j}, X_{i,j})$$

sub to

$$\sum_i \|U_{i,:}\|^2 \leq a, \quad \sum_j \|V_{j,:}\|^2 \leq b$$

for appropriate choices of $a$ and $b$. Despite the empirical success of the NMF approach, no theoretical justification has been given.

In this paper, we consider different but closely related classes of hypothesis matrices, which are defined by $L_1$ and $L_\infty$ constraints as stated below.

Convex combination constraints: The first one is the class of all matrices $\hat{X} = UV^T \in \mathbb{R}^{N \times M}$ with $U \in \mathbb{R}^{N \times K}$ and $V \in \mathbb{R}^{M \times K}$ such that for every $i \in [N]$ and $j \in [M]$,

$$\|U_{i,:}\|_1 = 1, \quad \|V_{j,:}\|_\infty \leq B.$$

That is, each row $\hat{X}_{i,:}$ is a convex combination of the vectors $V_{j,:}$ with weights $U_{i,:}$. This class has a natural interpretation, just as the topic model for document classification [14]–[16], that the rating of each user $i$ (row of $\hat{X}$) is a convex combination of the ratings of $K$ latent “model” users (rows of $V^T$), where the model users are allowed to assign ratings in $[-B, B]$. There are some studies on the topic model like approach for recommendation tasks [17], [18] and showing some experimental results.

Generalization: The second one is a slightly generalized class, consisting of all matrices $\hat{X} = UV^T \in \mathbb{R}^{N \times M}$ with $U \in \mathbb{R}^{N \times K}$ and $V \in \mathbb{R}^{M \times K}$ such that for every $i \in [N]$ and $j \in [M]$,
\[ \|U_i\|_1 \leq 1, \quad \|V_{j*}\|_\infty \leq B. \]

Note that any matrix \( \hat{X} \) of rank \( R \) has a factorization \( \hat{X} = UV^T \) with \( \|U_i\|_1 \leq 1 \), because otherwise we can replace \( U \) by \( cU \) and \( V \) by \( V/c \) for a sufficiently small constant \( c \). In other words, any matrix \( \hat{X} \) of rank \( R \) is in the second class for some parameter \( B \).

For these hypothesis classes, we give generalization bounds in the form of the excess error bounds relative to the empirical error. Our bounds for the both classes are of the same order: \( O(B \sqrt{(MK + N \log K)/T}) \), where \( T \) is the sample size. These bounds improve the previously known bound \( \tilde{O}(\sqrt{MK/N}) \) derived for the class where only the rank \( K \) constraint is imposed [4], provided that \( B \) is reasonably small. Therefore, our results would give theoretical evidence for the empirical success of the NMF. However, our new bounds hold even when \( U \) and \( V \) have negative values in some components. This result suggests that our analysis may not yet fully capture the property of the non-negativity constraint, or the empirical success of the NMF may not rely on the non-negativity very much but mostly on the regularization.

Our technical contributions are twofold. The first one is that we develop a new technique for bounding the Rademacher complexity to derive generalization bounds. The second one is that we prove a matching lower bound of the Rademacher complexity of the first hypothesis class. This means that our generalization bounds are tightest among those derived from the Rademacher complexity argument. There are few results in the literature on deriving lower bounds of the Rademacher complexity.

On the practical viewpoint, our generalization bounds would suggest us to use the structural risk minimization (SRM) scheme, where the bound \( B \) of \( L_\infty \) constraints on \( V \) (together with the rank \( K \)) is a parameter to tune. In our experiments on real data sets, our scheme with appropriate choices of \( B \) performs better than other schemes with only the rank constraint (i.e., the case \( B = \infty \)) or the trace norm constraint. Moreover, the optimal \( B \) turns out to be roughly equal to the largest rating in \( X \), which is usually known in advance. For instance, the ratings in the MovieLens datasets are valued in the range \([1, 2, 3, 4, 5]\) and so setting \( B = 5 \) seems to be a natural choice. Our experiments show that setting \( B = 5 \) is indeed optimal. This result actually suggests us to fix \( B \) to be the largest rating, and we do not need to employ the SRM to find an appropriate value of \( B \).

### 2. Related Work

Before describing our results, we briefly review some related results on the collaborative filtering problem.

As mentioned in the previous section, Srebro et al. [4] give a generalization bound \( \tilde{O}(\sqrt{K(N + M)/T}) \) for the class of rank \( R \) matrices. There are many results for the class of matrices with the trace norm constraint, which is a convex relaxation of the rank constraint. Srebro, Shraibman and Foygel [5, 7] estimate the expected Rademacher complexity of the class of low trace norm matrices and their generalization bound is \( O(\tau \sqrt{(M + N)/(TMN)}) \) where \( \tau \) is the bound of the trace-norm. Shamir and Shalev-Shwartz [19] also give a generalization bound \( O(\sqrt{N + \sqrt{M}})/T) \). The parameter \( \tau \) is usually scaled as \( \tau = O(\sqrt{MK}) \), and in this setting the both bounds are not less than \( \tilde{O}(\sqrt{K(N + M)/T}) \).

There are some results which exploit prior knowledge of the distribution \( D \) that generates the sample [20, 21], while in our setting we only consider the worst case generalization ability. [22] derives a generalization bound under Bernstein condition, which is between \( \sqrt{(N + M)K \log(N + M)/T} \) and \( (N + M)K \log(N + M)/T \). Natarajan and Jain [23] handle non-decomposable loss function, which cannot be written as a sum of element wise loss functions.

We emphasize that all of these result have \((M + N)K \) or \((M + N)A \) factors in the generalization bound, where \( A \) is an upper bound of the trace norm scaled to be compatible to the rank. We conjecture that if we use our hypothesis classes in the distribution-dependent settings above, we would improve the factors in the generalization bounds to \( MK + N \log K \).

### 3. Preliminaries

In this section, we first define the notations used in this paper and describe the problem setting. Then we briefly review a standard method of deriving generalization bounds via the Rademacher complexity arguments.

For a positive integer \( N \), let \([N] = \{1, \ldots, N\} \). For a matrix \( Y \), we write its \( i \)-th row as \( Y_{i*} \), and \( j \)-th column as \( Y_{*,j} \). Moreover, we denote by \( \|Y\|_F \), the trace norm of \( Y \), which is defined as the sum of the singular values of \( Y \).

#### 3.1 Problem Setting

Now we fix a true rating matrix \( X \in \mathbb{R}^{N \times M} \) to be learned and a probability distribution \( D \) over \([N] \times [M] \), which is unknown to the learner. Note that the matrix \( X \) can be regarded as a function \( (i, j) \mapsto X_{i,j} \). So we consider the problem under the standard PAC learning framework. Let \( \mathcal{X} \subseteq \mathbb{R}^{N \times M} \) be a set of matrices called the hypothesis class. The input to the learner is a sample of \( X \), i.e., a sequence of triples

\[
S = \{(i_1, j_1, X_{i_1,j_1}), (i_2, j_2, X_{i_2,j_2}), \ldots, (i_r, j_r, X_{i_r,j_r})\},
\]

where each \((i_r, j_r) \in [N] \times [M] \) is independently chosen according to the distribution \( D \). When given the sample \( S \), the learner is required to produce a hypothesis matrix \( \hat{X} \) chosen from \( \mathcal{X} \), so that its generalization error is as small as possible (with high probability over the random choice of \( S \)), where the generalization error of \( \hat{X} \) is defined as

\[
\ell(\hat{X}; X) = \mathbb{E}_{(i,j) \sim D}[\ell(X_{i,j}; X_{i,j})]
\]

for a fixed loss function \( \ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \). Throughout the
paper, we assume that the function $\ell$ is $L$-Lipschitz with respect to the first argument and $|\ell|$ is bounded by a constant.

3.2 Deriving Generalization Bounds

A typical strategy for achieving the goal is the empirical risk minimization, i.e., to output a matrix $\hat{X}$ in $X$ that minimizes the empirical error, defined as

$$\hat{\ell}_S(\hat{X}; X) = \frac{1}{T} \sum_{i=1}^T \ell(\hat{X}_{i,j}, X_{i,j}).$$

A generalization bound of the hypothesis class $X$ with respect to a sample size $T$ and a confidence parameter $\delta$ is a real number $M$ such that for any distribution $D$, it holds that

$$\Pr_{S \sim D^T} \left( \forall X \in X, \ell(\hat{X}; X) \leq \hat{\ell}_S(\hat{X}; X) + M \right) \geq 1 - \delta.$$

So, if we have a generalization bound $M$ that converges 0 as $T$ goes to infinity, then the inequality above justifies to use the empirical risk minimization.

Now we show a general theorem that gives a generalization bound of $X$ in terms of the Rademacher complexity. The (empirical) Rademacher complexity of $X$ with respect to the sample $S$ is defined as

$$\hat{R}_S(X) = \frac{1}{T} \mathbb{E}_S \left[ \sup_{\hat{X} \in X} \sum_{i=1}^T \sigma_i \hat{X}_{i,j} \right],$$

where $\sigma_i$'s are independent random variables taking value 1 or $-1$ with probability 1/2.

**Theorem 1** (Generalization bounds [24]). With probability at least $1 - \delta$ over the random choice of $S \in ([N] \times [M] \times \mathbb{R})^T$, it holds that for any $\hat{X} \in X$,

$$\ell(\hat{X}; X) \leq \hat{\ell}_S(\hat{X}; X) + 2L \hat{R}_S(X) + O \left( \sqrt{\frac{\ln(1/\delta)}{T}} \right).$$

Thus, to derive a generalization bound from this theorem, we need to estimate the Rademacher complexity $\hat{R}_S(X)$.

4. Generalization Bounds for Our Hypothesis Classes

In this section, we give upper bounds of the Rademacher complexity of our hypothesis classes, whereby we give generalization bounds of these classes.

4.1 Matrix Factorization with Convex Combination Constraints

Here we assume that the preferences of each user can be expressed as the convex combination of a small number $K$ of model users’ preferences. Let $P \subseteq \mathbb{R}^M$ be a finite set of vectors, so that its convex hull $\text{conv}(P)$ defines the class of model user’s preferences. In particular, if $P = \{-B, B\}^M$, then $\text{conv}(P) = [-B, B]^M$. For a set of vectors $Y \subseteq \mathbb{R}^M$, we define $Y^K$ as the set of all matrices in $\mathbb{R}^{M \times K}$ whose columns are chosen from $Y$. That is, $Y^K = \{y_1, y_2, \ldots, y_K : \forall k \in [K], y_k \in Y\}$. Let $U$ and $V$ be the sets of matrices defined as

$$U = \left\{U \in [0, 1]^{N \times K} : \forall i \in [N], \|U_{i,:}\|_1 = 1\right\}, \quad (1)$$

$$V = \left\{(\text{conv}(P))^K \right\}. \quad (2)$$

Then, first we examine the following hypothesis class:

$$X = \{\hat{X} \in \mathbb{R}^{N \times M} : \hat{X} \in U, \ V \in V\}. \quad (3)$$

In order to derive an upper bound of the Rademacher complexity of $X$, we need some technical lemmas.

**Lemma 1.** Let $F \subseteq \{f : Y \rightarrow \mathbb{Z}\}$ and $G \subseteq \{g : W \rightarrow \mathbb{Y}\}$ be finite sets of functions. Then $\text{conv}(F) \circ G \subseteq \text{conv}(F \circ G)$.

**Proof.** By the definition of the convex hull, we can write

$$\text{conv}(F) = \left\{ \sum_{f \in F} a_f f(\cdot) : a \in \Delta_F \right\},$$

where $\Delta_F$ denotes the probability simplex over $F$. Therefore, for any $h \in \text{conv}(F) \circ G$, there exist some $a \in \Delta_F$ and $g \in G$ such that $h = (\sum_{f \in F} a_f f) \circ g$. Then, for any $w \in W$, $h(w) = \sum_{f \in F} a_f f(g(w)) = \sum_{f \in F} a_f f(f \circ g)(w)$. This implies that $h \in \text{conv}(F \circ G)$. \hfill $\square$

The following lemma is crucial in our analysis.

**Lemma 2.** Let $P$ be a finite set of vectors. Then $(\text{conv}(P))^K \subseteq \text{conv}(P^K)$.

**Proof.** Let $V = (v_1, \ldots, v_K) \in (\text{conv}(P))^K$. Then we can write $v_i = \sum_{p \in P} a_{i,p} p$ for some $a_i \in \Delta_P$. Let $b_{p_1,\ldots,p_K} = \prod_{i=1}^K a_{i,p_i}$. Then observe that

$$\sum_{(p_1,\ldots,p_K) \in P^K} b_{p_1,\ldots,p_K} = \left( \sum_{p \in P} a_{1,p} \right) \cdots \left( \sum_{p \in P} a_{K,p} \right) = 1.$$

Thus $\{b_{p_1,\ldots,p_K} : (p_1,\ldots,p_K) \in P^K\}$ defines a probability distribution over $P^K$.

The $K$-th column vector $v_K$ can be represented as follows:

$$v_K = \sum_{p \in P} a_K p = \left( \sum_{p \in P} a_{1,p} \right) \times \left( \sum_{p \in P} a_{2,p} \right) \times \cdots \times \left( \sum_{p \in P} a_{K-1,p} \right) \sum_{p \in P} a_{K,p} p_K$$

$$= \left( \sum_{p \in P} a_{1,p} \right) \cdots \left( \sum_{p \in P} a_{K-1,p} \right) \sum_{p \in P} a_{K,p} p_K.$$
\[
\begin{align*}
\hat{\mathcal{R}}_S(\mathcal{P} \circ \mathcal{K}) &\leq B \sqrt{\frac{8 \ln |\mathcal{P} \circ \mathcal{K}|}{T} + \frac{N \ln K}{T}} \\
&\leq B \sqrt{\frac{8(K \ln |\mathcal{P}| + \ln |\mathcal{K}|)}{T}}. 
\end{align*}
\]

Using the fact $|\mathcal{K}| = K^K$, we obtain the theorem. \hfill \Box

Now we choose $\mathcal{P} = [-B, B]^M$ so that $\text{conv}(\mathcal{P}) = [-B, B]^M$, i.e., the $M$-dimensional $L_\infty$-norm ball. Clearly, $|\mathcal{P}| = 2^M$. Combining Theorem 1 and Theorem 2, we get the following generalization bound for the class $\mathcal{X}$ induced by $\mathcal{P}$.

**Corollary 3.** Let $\mathcal{P} = [-B, B]^M$ and $\mathcal{X}$ be defined as in (1), (2) and (3). Note here that $V = [-B, B]^{MK}$. Then

\[
\hat{\mathcal{R}}_S(\mathcal{X}) \leq B \sqrt{\frac{MK \ln 2 + N \ln K}{T}}, 
\]

and an generalization bound of $\mathcal{X}$ is

\[
O\left(\frac{LB \sqrt{MK + N \ln K}}{T} + \sqrt{\frac{\ln(1/\delta)}{T}}\right). 
\]

Another choice of $\mathcal{P}$ may be useful in some applications and lead an even smaller generalization bound. For example, if we take $\mathcal{P} = \{-e_1, e_1, -e_2, e_2, \ldots, -e_M, e_M\}$ where each $e_i \in \{0, 1\}^M$ is the unit vector whose $i$-th element is 1, then we have $\text{conv}(\mathcal{P}) = \{v \in \mathbb{R}^M : \|v\|_1 \leq 1\}$, which is the $M$-dimensional $L_1$-norm ball. In this case $|\mathcal{P}| = 2M$. Therefore, we have the following corollary.

**Corollary 4.** Let $\mathcal{P} = \{-e_1, e_1, -e_2, e_2, \ldots, -e_M, e_M\}$ and $\mathcal{X}$ be defined as in (1), (2) and (3). Then, a generalization bound of $\mathcal{X}$ is

\[
O\left(\frac{LB \sqrt{MK + N \ln K}}{T} + \sqrt{\frac{\ln(1/\delta)}{T}}\right). 
\]

4.2 Matrix Factorization with $L_1$ and $L_\infty$ Norm Constraints

Now we consider a slightly generalized hypothesis class. For a finite set $\mathcal{P} \in \mathbb{R}^M$, let

\[
\mathcal{U} = \{U \in \mathbb{R}^{N \times K} : \forall i \in [N], \|U_{i,\cdot}\|_1 \leq 1\},
\]

\[
\tilde{\mathcal{V}} = \left(\text{conv}(\mathcal{P})\right)^K,
\]

and

\[
\tilde{\mathcal{X}} = \{\tilde{X} = UV^T \in \mathbb{R}^{N \times M} : U \in \mathcal{U}, \ V \in \tilde{\mathcal{V}}\}. 
\]

Only the difference from the class $\mathcal{X}$ discussed in the previous subsection (i.e., (1), (2) and (3)) is that $U \in \mathcal{U}$ can contain negative components and the $L_1$-norm of each row $U_{i,\cdot}$ is not necessarily equals to 1.

The next theorem says that the problem of learning with the hypothesis class $\tilde{\mathcal{X}}$ can be reduced to that with the hypothesis class $\mathcal{X}$.
Theorem 5. \( \tilde{X} \subset X \), where \( X \) is defined as in (1), (2) and (3) with \( K = 2K + 1 \) and \( \mathcal{P} = \tilde{U} \cup \{ -a : a \in \mathcal{P} \} \).

Proof. It suffices to show that any \( \tilde{X} = \tilde{U}V^{T} \in \tilde{X} \) can be represented as \( U^{T}V \) for some \( U \in \mathcal{U} \) and \( V \in (\text{conv}(\mathcal{P}))^{K} \). To see this, let \( U^{+} = [0,1]^{N \times K} \) and \( U^{-} = [0,1]^{N \times K} \) be non-negative matrices such that \( U^{+}_{i,j} = \tilde{U}_{i,j}+ \) and \( U^{-}_{i,j} = -\tilde{U}_{i,j} \), respectively, where \( [x]_{+} = x \) if \( x \geq 0 \) and \( [x]_{-} = 0 \) if \( x < 0 \). Moreover, let \( u \in [0,1]^{K} \) be a vector whose \( i \)-th component is \( u_{i} = 1 - \|\tilde{U}_{i,\cdot}\| \). Then, it is straightforward to show that

\[
\tilde{U}V^{T} = [U^{+}, U^{-}, u][\mathcal{V}, -\mathcal{V}, 0]^{T}.
\]

Moreover, it is easy to show that \( U = [U^{+}, U^{-}, u] \in \mathcal{U} \) and \( \mathcal{V} = [\mathcal{V}, -\mathcal{V}, 0] \in (\text{conv}(\mathcal{P}))^{K} \).

Since \( \hat{R}_{S}(\tilde{X}) \leq \hat{R}_{S}(X) \) by the theorem above, we have the same order of generalization bound of \( \tilde{X} \) as shown in Theorem 2.

5. A Matching Lower Bound on the Rademacher Complexity

We show that the bound stated in Corollary 3 is almost tight and cannot be significantly improved in the worst case.

Theorem 6. Let \( X \) be the matrix class defined in Corollary 3. For any sufficiently large \( T \), there exists a sample \( S \) such that

\[
\hat{R}_{S}(X) = \Omega \left( B \sqrt{\frac{MK + N \log K}{T}} \right).
\]

Proof. Let \( L = \lfloor \log_{2} K \rfloor \) and \( J_{1} \cup J_{2} \cup \cdots \cup J_{L} = \{ 1, 2, \ldots, M \} \) be a partition of the column set. Consider a sample \( S \) that satisfies \( \mathcal{T}(i,u) = T/(NL) \) for each \( i \in [N] \) and \( u \in [L] \), and \( |\{ t : j_{t} = j \}| = T/M \) for each \( j \in [M] \), where \( \mathcal{T}(i,u) = \{ t : i_{t} = i, j_{t} \in J_{u} \} \).

From the proof of Theorem 2, we have

\[
\frac{T}{B} \hat{R}_{S}(X) = \frac{1}{B} \mathbb{E} \left[ \sup_{x \in K} \max_{V} \sum_{t=1}^{T} \sigma_{t} V_{j_{t}, x_{(i)}} \right] \\
= \frac{1}{B} \mathbb{E} \left[ \max_{x \in K} \sum_{j=1}^{M} \sum_{t,j_{t}=j} \sigma_{t} V_{j_{t}, x_{(i)}} \right] \\
= \frac{1}{B} \mathbb{E} \left[ \max_{x \in K} \sum_{j=1}^{M} \sum_{t,j_{t}=j} \sigma_{t} V_{j_{t}, x_{(i)}} \right] \\
= \mathbb{E} \left[ \max_{x \in K} \sum_{j=1}^{M} \sum_{t,j_{t}=j} \sigma_{t} \right] = \mathbb{E} \left[ \max_{x \in K} \sum_{j=1}^{M} \sum_{t,j_{t}=j} \sigma_{t} \right].
\]

So, it suffices to show that (8) is lower bounded by \( \Omega(\sqrt{MKT}) \) and \( \Omega(\sqrt{NLT}) \).

First we give a proof of the \( \Omega(\sqrt{MKT}) \) part. To this end, we replace the max operator in (8) by the expectation under the uniform distribution over \( \mathcal{K} \). Note then that \( T_{j,\ell} = |\{ t : j_{t} = j, x_{(i_{t})} = \ell \}| \) is a binomial distributed random variable with parameters \( (T/M, 1/K) \), i.e.,

\[
T_{j,\ell} \sim B(T/M, 1/K)
\]

and thus \( \mu \equiv \mathbb{E}_{x}[T_{j,\ell}] = T/(MK) \) for each \( j \in [M] \) and \( \ell \in [K] \). So, by Chernoff bound we get

\[
\text{Pr}[T_{j,\ell} \leq T/(2MK)] = \text{Pr}[T_{j,\ell} \leq \mu/2] \leq e^{-\mu T/(8MK)} = e^{-T/(8MK)}.
\]

Thus we can say \( \text{Pr}[T_{j,\ell} \geq T/(2MK)] \geq 1/2 \) if \( T \geq 8MK \ln 2 \). Moreover, \( \mathbb{E}_{x}[\sum_{j,\ell} \mathbb{1}[x_{(j,\ell)}=\ell]] \geq c \sqrt{T} \) for some constant \( c > 0 \) by the random-walk argument. Therefore,

\[
\mathbb{E} \left[ \max_{x \in K} \sum_{j=1}^{M} \sum_{\ell=1}^{K} \sigma_{j} \right] \\
\geq \mathbb{E} \left[ \mathbb{E}_{x} \left[ \sum_{j=1}^{M} \sum_{\ell=1}^{K} \mathbb{1}[x_{(j,\ell)}=\ell] \sigma_{j} \right] \right] \\
\geq \frac{c}{2} \sum_{j=1}^{M} \sum_{\ell=1}^{K} \sqrt{\frac{T}{2MK}} = \Omega \left( \sqrt{MKT} \right).
\]

Next we give a proof of the \( \Omega(\sqrt{NLT}) \) part. For each \( \sigma \in \{-1,1\}^{L} \), we define \( \kappa \in \mathcal{K} \) as described below, and replace the max operator in (8) by this particular \( \kappa \), which gives a lower bound. For each \( i \in [N] \) and \( u \in [L] \), let

\[
\ell_{i,u} = \begin{cases} 
1 & \text{if } \sum_{t \in \mathcal{T}(i,u)} \sigma_{t} \geq 0, \\
-1 & \text{otherwise}.
\end{cases}
\]

and let the \( L \)-bit sequence \( \ell_{i} = (\ell_{i,1}, \ell_{i,2}, \ldots, \ell_{i,L}) \) be identified with an integer in \( [K] \) whose binary representation is \( \ell_{i} \). For example, \((-1,-1,-1) = 1, (-1,-1,1) = 2, \ldots, (1,1,1) = 8 \) when \( L = 3 \). Now, our \( \kappa \) is given by \( \kappa(i) = \ell_{i} \) for each \( i \). Then, (8) is lower bounded by

\[
\mathbb{E} \left[ \sum_{j=1}^{M} \sum_{\ell=1}^{K} \mathbb{1}[x_{(j,\ell)}=\ell] \sigma_{j} \right] \\
\geq \sum_{j=1}^{M} \sum_{\ell=1}^{K} \sum_{t,j_{t}=j} \mathbb{1}[x_{(j_{t},\ell)}=\ell] \sigma_{t} \\
= \sum_{u=1}^{L} \sum_{j=1}^{M} \sum_{\ell=1}^{K} \sum_{t,j_{t}=j} \mathbb{1}[x_{(j_{t},\ell)}=\ell] \sigma_{t} \\
= \sum_{u=1}^{L} \sum_{j=1}^{M} \sum_{\ell=1}^{K} \sum_{t,j_{t}=\ell} \mathbb{1}[t \in \mathcal{T}(i,u)] \sigma_{t} \\
\text{Now by the definition of } \ell_{i,u}, \text{ we have}
\]

\[
\sum_{t \in \mathcal{T}(i,u)} \sigma_{t} = \ell_{i,u} \sum_{t \in \mathcal{T}(i,u)} \sigma_{t}.
\]

Therefore, we get
where the last inequality is derived from the random-walk argument.

6. Experimental Results

The $L_{\infty}$ norm bound $B$ on $V$ has a role of controlling the complexity of the hypothesis class. Roughly, as the value of $B$ decreases, the excess error gets smaller while the empirical error grows larger. So we may somehow need to seek for an appropriate value of $B$ to minimize the generalization error. In the first experiment, we examine the training and test errors for various choices of $B$ and see when the test error is minimized. In the second experiments, we compare the performance of our method with standard ones.

For both experiments, we use the following three datasets. First dataset is MovieLens-100k which contains 100000 ratings of 1682 movies by 943 users. All ratings are valued in $\{1, 2, 3, 4, 5\}$. Second one is the sushi preference dataset [25] which contains 50000 ratings of 100 kind of sushi items by 5000 users. All ratings are valued in $\{0, 1, 2, 3, 4\}$. The last data is a synthetic data. We made this data as follows: (i) Generate $U_{i,k}$ and $V_{j,k}$ according to the uniform distribution over $[0, 1]$, (ii) Generate a random noise matrix $R_{i,j}$ according to the standard normal distribution, and (iii) Set $X = \alpha UV^T + (1 - \alpha)R$ and use randomly chosen 50% entries of $X$ as the sample and the rest as the test set. The parameters we use are $N = 200$, $M = 600$, $K = 10$, and $\alpha = 0.75$. This data contains 59736 entries with range $[-0.347, 4.569]$. The mean value of $X_{i,j}$ is $\mu = 1.886$ and its variance is $\sigma = 0.349$.

We examine the square loss $\ell(x, y) = \frac{1}{2}(x - y)^2$ in all experiments. We conduct 5-fold cross validation and measure the mean squared error

$$\text{MSE}(\hat{X}, T^*) = \frac{1}{|T^*|} \sum_{(i,j) \in T^*} \ell(\hat{X}_{i,j}, X_{i,j})$$

where $T^*$ is the test set.

6.1 Changing $L_{\infty}$ Norm Constraint

In the first experiment, we use the hypothesis classes defined in (1), (2) and (3) with $K = 10$ and $\mathcal{P} = [-B, B]$ for various values of $B$. Since the problem of empirical risk minimization is NP-hard, we employ the alternating least square optimization scheme to get a hypothesis $\hat{X}$ that attains a local minimum of the training error $\text{MSE}(\hat{X}, S)$ where $S$ is the sample. (See Appendix A for implementation details).

The results are shown in Figs. 1, 2 and 3. We see that in the first two datasets, the training error remains small enough when $B$ is as small as the largest rating values (i.e., 5 for MovieLens-100k and 4 for sushi3b), and thus the test error is minimized around those values. For the last dataset, we can see a similar phenomenon since most data falls into the range $[\mu - 2\sigma, \mu + 2\sigma] = [1.188, 2.584]$ and thus $B = 3$ would be an appropriate choice.
The results suggest that we do not need to employ the SRM to seek for an appropriate value of $B$ but fixing $B$ to be the largest (typical) value of the ratings is good enough.

6.2 Comparison with Trace Norm Regularization

In the second experiments, we compare our method with the conventional trace norm regularized empirical risk minimization.

As in the first experiment, we also use the alternating least square optimization scheme to get a hypothesis from the hypothesis classes defined in (1), (2) and (3) with appropriate choices of $B$ for each dataset. But now we choose the best choice of rank $K$ by grid search from $\{1, 2, 4, 8, 10\}$ for MovieLens and sushi dataset, and $\{4, 8, 11, 12, 14\}$ for synthetic data set.

We compare our method with the standard trace norm regularization approach, which minimizes the empirical loss regularized by the trace norm:

$$\tilde{X}_{\text{Tr}} = \arg\min_{X \in \mathbb{R}^{N \times M}} \frac{1}{2} \sum_{(i,j) \in S} (x_{i,j} - \tilde{x}_{i,j})^2 + \lambda \|X\|_{\text{Tr}}$$

where $\lambda$ is a parameter. We also choose the best $\lambda$ from $\{0.1, 0.2, 0.4, 0.8, 1, 2, 4, 6, 8, 10, 20, 40\}$ by grid search. It is well known that this is the tightest convex relaxation of the rank constraint. To solve this minimization problem, we use SoftImpute implemented at fancyimpute 0.0.19 ¥. This is an implementation of SoftImpute algorithm proposed by Mazumder et al. [26].

The results are summarized in Table 1.

| Data set    | Our approach $\tilde{X}_{\text{Tr}}$ | SoftImpute $\tilde{X}_{\text{Tr}}$ |
|-------------|--------------------------------------|----------------------------------|
| MovieLens   | $0.881 (K = 2, B = 5)$               | $1.011 (\lambda = 8.0)$          |
| sushi       | $1.363 (K = 1, B = 4)$               | $1.388 (\lambda = 8.0)$          |
| synthetic   | $0.079 (K = 11, B = 4.569)$          | $0.399 (\lambda = 4.0)$          |

The results are summarized in Table 1.

Clearly, our method outperforms the conventional method for all datasets. Note that in the real datasets, the best rank $K$ is very small. Especially, on the sushi data set, it is 1. This would imply that all users have almost the same preference.

7. Conclusion

In this paper, we consider the collaborative filtering problem and derive tighter generalization error bounds for matrix factorization with $L_1$ and $L_{\infty}$ norm constraints.

One of our future work is to derive generalization bounds with other norm constraints such as the Frobenius norm. Also, one could improve our results for strongly convex loss functions or under some easy data setting, say, under the Bernstein conditions [22].

Acknowledgments

This work is supported in part by JSPS KAKENHI Grant Number 16K00305 and JSPS KAKENHI Grant Number 15H02667, respectively.

References

[1] N. Rao, H.F. Yu, P.K. Ravikumar, and I.S. Dhillon, “Collaborative filtering with graph information: Consistency and scalable methods,” Advances in Neural Information Processing Systems 28 (NIPS ’15), pp.2107–2115, 2015.
[2] B. Li, Y. Wang, A. Singh, and Y. Vorobeychik, “Data poisoning attacks on factorization-based collaborative filtering,” Advances in Neural Information Processing Systems 29 (NIPS ’16), pp.1885–1893, 2016.
[3] Y. Zheng, B. Tang, W. Ding, and H. Zhou, “A neural autoregressive approach to collaborative filtering,” Proc. 33rd International Conference on Machine Learning (ICML’16), pp.764–773, 2016.
[4] N. Srebro, N. Alon, and T.S. Jaakkola, “Generalization error bounds for collaborative prediction with low-rank matrices,” Advances in Neural Information Processing Systems 17 (NIPS’04), pp.1321–1328, 2004.
[5] N. Srebro, J.D.M. Rennie, and T.S. Jaakkola, “Maximum-margin matrix factorization,” Advances in Neural Information Processing Systems 17 (NIPS’04), pp.1329–1336, 2005.
[6] N. Srebro and A. Shraibman, “Rank, trace-norm and max-norm,” 18th Annual Conference on Learning Theory (COLT’05), pp.545–560, 2005.
[7] R. Foygel and N. Srebro, “Concentration-based guarantees for low-rank matrix reconstruction,” The 24th Annual Conference on Learning Theory (COLT’11), pp.315–340, 2011.
[8] D.D. Lee and H.S. Seung, “Learning the parts of objects by nonnegative matrix factorization,” Nature, vol.401, pp.788–791, 1999.
[9] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” Computer, vol.42, no.8, pp.30–37, Aug. 2009.
[10] Q. Gu, J. Zhou, and C. Ding, “Collaborative filtering: Weighted non-negative matrix factorization incorporating user and item graphs,” Proc. 2010 SIAM International Conference on Data Mining (SDM’10), pp.199–210, 2010.
[11] S. Zhang, W. Wang, J. Ford, and F. Makedon, “Learning from incomplete ratings using non-negative matrix factorization,” Proc. 2006 SIAM International Conference on Data Mining (SDM’06), pp.549–553, 2006.
[12] X. Luo, M. Zhou, Y. Xia, and Q. Zhu, “An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems,” IEEE Trans. Industrial Informatics, vol.10, no.2, pp.1273–1284, 2014.
[13] S.S. Du, Y. Liu, B. Chen, and L. Li, “Maxios: Large scale nonnegative matrix factorization for collaborative filtering,” Neural Information Processing Systems 27 (NIPS’14), workshop on Distributed Machine Learning and Matrix Computations, 2014.
[14] D.M. Blei, “Introduction to probabilistic topic models,” Communications of the ACM, 2011.
[15] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol.3, pp.993–1022, March 2003.
[16] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, “A practical algorithm for topic modeling with provable guarantees,” Proc. 30th International Conference on Machine Learning (ICML’13), pp.280–288, 2013.
[17] C. Wang and D.M. Blei, “Collaborative topic modeling for recommending scientific articles,” Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’11), pp.448–456, 2011.
[18] Y. Bao, H. Fang, and J. Zhang, “Topicmcf: Simultaneously exploiting ratings and reviews for recommendation,” Proc. 28th AAAI Conference on Artificial Intelligence (AAAI’14), pp.2–8, 2014.
[19] O. Shamir and S. Shalev-Shwartz, “Collaborative filtering with the

¥https://pypi.python.org/pypi/fancyimpute
trace norm: Learning, bounding, and transducing,” The 24th Annual Conference on Learning Theory (COLT ’11), pp.661–678, 2011.
[20] V. Kolchinskii, K. Lounici, and A.B. Tsybakov, “Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion,” The Annals of Statistics, vol.39, no.5, pp.2302–2329, 2011.
[21] O. Klopp, “Noisy low-rank matrix completion with general sampling distribution,” Bernoulli, vol.20, no.1, pp.282–303, 2014.
[22] P. Alquier, V. Cottet, and G. Lecué, “Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions,” ArXiv e-prints, Feb. 2017.
[23] N. Natarajan and P. Jain, “Regret bounds for non-decomposable metrics with missing labels,” Advances in Neural Information Processing Systems 29 (NIPS 2016), pp.2874–2882, 2016.
[24] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of Machine Learning, Adaptive computation and machine learning series, MIT Press, 2012.
[25] T. Kamishima, “Nantonac collaborative filtering: Recommendation based on order responses,” Proc. Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’03), pp.583–588, 2003.
[26] R. Mazumder, T. Hastie, and R. Tibshirani, “Spectral regularization algorithms for learning large incomplete matrices,” J. Mach. Learn. Res., vol.11, pp.2287–2322, Aug. 2010.

Appendix A: Implementation Details

In this section we describe an algorithm to find a hypothesis from the hypothesis class. We adopt the empirical risk minimization which selects the hypothesis \( \hat{X} \) as

\[
\hat{X} = \arg\min_{X \in \mathcal{X}} \hat{\ell}_S(\hat{X}, X) = \arg\min_{U \in \mathcal{U}, V \in \mathcal{V}} \hat{\ell}_S(UV^T; X)
\]

where \( \mathcal{X} \) is the hypothesis class.

In the implementation of experiments, we use the conventional alternating minimization which iterates

\[
\hat{U} = \arg\min_{U \in \mathcal{U}} \hat{\ell}_S(UV^T; X) \quad \text{and} \quad \hat{V} = \arg\min_{V \in \mathcal{V}} \hat{\ell}_S(UV^T; X)
\]

until it converges. Each optimization step is just a convex optimization. Especially, if we choose \( \ell \) as squared loss, it is a quadratic programming. So we can solve optimization of \( U \) and \( V \) efficiently using common QP solvers.

Of course our optimization of \( \hat{X} \) is not convex due to the rank constraint, we get a local minimizer \( \hat{X} = \hat{U}\hat{V}^T \) in the experiments and we use it as the hypothesis instead of the global minimizer.

Because of our bounds based on Theorem 1 holds for any \( \hat{X} \) in the hypothesis class \( \mathcal{X} \), we can use the generalization error bounds under this optimization scheme.

Appendix B: Yet Another Proof of Lemma 2

Proof. Given \( V = (v_1, \ldots, v_K) \in (\text{conv}(\mathcal{P}))^K \) with \( v_i = \sum_{p \in \mathcal{P}} \alpha_p^{(i)} p \) \( (i \in [K]) \), we can construct a convex combination \( X = \sum_{p_1, \ldots, p_K \in \mathcal{P}} \alpha_{p_1, \ldots, p_K} (p_1, \ldots, p_K) \) by the following procedure.

1. Let \( Y \leftarrow V \) and \( k \leftarrow 1 \).
2. While \( Y \geq 0 \), repeat:
   a. Let \( \mathcal{P}^{(i)} = \{ p \in \mathcal{P} \mid \alpha_p^{(i)} > 0 \} \).
   b. Let \( (i_k, p_{i_k}) = \arg\min_{i \in [K]} \min_{p \in \mathcal{P}^{(i)}} \alpha_p^{(i)} \).
   c. Pick up any \( p_j \in \mathcal{P}^{(i_k)} \) for \( j \neq i_k \) and construct a vector \( q_k = (p_1, \ldots, p_{i_k}, \ldots, p_K) \).
   d. Update \( Y \leftarrow Y - \alpha_{p_{i_k}} q_k \).
   e. Update all \( \alpha_p^{(i)} \) and \( k \leftarrow k + 1 \).
3. Output \( \sum_{k \geq 1} \alpha_{p_{i_k}} q_k \).

Note that the while loop terminates in \( K|\mathcal{P}| \) trials since the procedure increases a zero entry by at least one at each loop. Also, one can observe that each \( \alpha_{p_{i_k}}^{(i_k)} \geq 0 \) and \( \sum_{k \geq 1} \alpha_{p_{i_k}}^{(i_k)} = 1 \). Therefore the procedure outputs a convex decomposition of \( V \).

Ken-ichiro Moridomi received the B.E., Master of Informatics degrees and Ph.D. from Kyushu University in 2013, 2015 and 2018 respectively. He now works for So-net Media Networks Corporation.

Kohei Hatano received Ph.D. from Tokyo Institute of Technology in 2005. Currently, he is an associate professor at Research and Development Division in Kyushu University Library. He is also the leader of the Computational Learning Theory team at RIKEN AIP. His research interests include machine learning, computational learning theory, online learning and their applications.

Eiji Takimoto received Dr.Eng. degree from Tohoku University in 1991. Currently, he is a professor at Department of Informatics in Kyushu University. His research interests include computational complexity, computational learning theory, and online learning.