Collaborative Filtering with the Trace Norm: Learning, Bounding, and Transducing

Ohad Shamir
Microsoft Research
1 Memorial Drive, Cambridge MA 02142 USA
ohadsh@microsoft.com

Shai Shalev-Shwartz
The Hebrew University
Givat Ram, Jerusalem 91904, Israel
shais@cs.huji.ac.il

Abstract

Trace-norm regularization is a widely-used and successful approach for collaborative filtering and matrix completion. However, its theoretical understanding is surprisingly weak, and despite previous attempts, there are no distribution-free, non-trivial learning guarantees currently known. In this paper, we bridge this gap by providing such guarantees, under mild assumptions which correspond to collaborative filtering as performed in practice. In fact, we claim that previous difficulties partially stemmed from a mismatch between the standard learning-theoretic modeling of collaborative filtering, and its practical application. Our results also shed some light on the issue of collaborative filtering with bounded models, which enforce predictions to lie within a certain range. In particular, we provide experimental and theoretical evidence that such models lead to a modest yet significant improvement.

1 Introduction

We consider the problem of matrix-based collaborative filtering, where the goal is to predict entries of an unknown matrix based on a subset of its observed entries. An increasingly popular approach to achieve this is via trace-norm regularization, where one seeks a matrix which agrees well with the observed entries, while constraining its complexity in terms of the trace-norm. The trace-norm is well-known to be a convex surrogate to the matrix rank, and has repeatedly shown good performance in practice (see for instance Srebro et al. (2004), Salakhutdinov and Mnih (2007), Bach (2008), Candès and Tao (2009)).

However, in terms of distribution-free guarantees, the current learning-theoretic understanding of trace-norm regularization is surprisingly weak. Most non-trivial guarantees currently known (e.g., Srebro and Shraibman (2005), Candès and Tao (2009), Candès and Recht (2009)) assume that the observed entries are sampled uniformly at random. In collaborative filtering, this is an extremely unrealistic assumption. For example, in the Netflix challenge dataset, where the matrix contains the ratings of users (rows) for movies (columns), the number and distribution of ratings differ drastically between users. Modeling such data as a uniform sample is not a reasonable assumption. Recently, Negahban and Wainwright (2010) studied the problem of matrix completion under a non-uniform distribution. However, the analysis is still not distribution-free, and requires strong assumptions on the underlying matrix. Moreover, the results do not apply to standard trace-norm regularization, but rather to a carefully re-weighted version of trace-norm regularization.

In practice, we know that standard trace-norm regularization works well even for data which is very non-uniform. Moreover, we know that in other learning problems, one is able to derive distribution-free guarantees, and there is no a-priori reason why this should not be possible here. Nevertheless, obtaining a non-trivial guarantee for trace-norm regularization has remained elusive. As a result, some works suggested to use other complexity measures for collaborative filtering, such as the max-norm (Srebro et al. (2004), Lee et al. (2010)) and weighted trace-norm (Salakhutdinov and Srebro (2010)).

In this paper, we bridge this gap between our theoretical understanding and practical performance of trace-norm regularization. We show that by adding very mild assumptions, which correspond to collaborative filtering as performed *in practice*, it is possible to obtain non-trivial, distribution-free guarantees on learning with the trace norm. In fact, we claim that the difficulties in providing

such guarantees partially stemmed from a mismatch between the standard theoretical modeling of collaborative filtering, and its practical application.

First, we show that one can obtain such guarantees, if one takes into account that the values to be predicted are usually bounded in practice. For example, in predicting movie ratings, it is known in advance that the ratings are on a scale of (say) 1 to 5, and practitioners usually clip their predictions to be inside this range. While this seems like an almost trivial operation, we show that taking it into account has far-reaching implications in terms of the theoretical guarantees. The proof relies on a decomposition technique which might also be useful for regularizers other than the trace-norm.

Second, we argue that the standard inductive model of learning, where the training data is assumed to be sampled i.i.d. from some distribution, may not be the best way to analyze collaborative filtering. Instead, we look at the transductive model, where sampling of the data is done without replacement. In the context of collaborative filtering, we show this makes a large difference in terms of the attainable guarantees.

Our results show that a transductive model, and boundedness assumptions, play an important role in obtaining distribution-free guarantees. This relates to a line of recent works, which suggest to incorporate prior knowledge on the range of predicted values into the learning process, by explicitly bounding the predictions. We provide an empirical study, which indicates that this indeed provides a modest, yet significant, improvement in performance, and corroborates our theoretical findings.

2 Setting

Our goal is to predict entries of an unknown $m \times n$ matrix X, based on a random subset of observed entries of X. A common way to achieve this, following standard learning approaches, is to find an $m \times n$ matrix W from a constrained class of matrices $\mathcal{W}$, which minimizes the discrepancy from X on the observed entries. More precisely, if we let $S = \{i_\alpha, j_\alpha\}$ denote the set of (row,column) observed entries, and ℓ is a loss function measuring the discrepancy between the predicted and actual value, then we solve the optimization problem

$$\min_{W \in \mathcal{W}} \frac{1}{|S|} \sum_{\alpha=1}^{|S|} \ell(W_{i_\alpha,j_\alpha}, X_{i_\alpha,j_\alpha}), \tag{1}$$

An important and widely used class of matrices $\mathcal{W}$ are those with bounded *trace-norm*. Given a matrix W, its trace-norm $\|W\|_{tr}$ is defined as the sum of the singular values. The class of matrices with bounded trace-norm has several useful properties, such as it being a convex approximation of the class of rank-bounded matrices (see e.g. Srebro and Shraibman (2005)). The trace-norm of $m \times n$ matrices with bounded entries is typically on the order of $\sqrt{mn}$, and thus we will focus on classes with trace norm bounded by $t = \Theta(\sqrt{mn})$.

For now, we will consider the inductive model of learning, which parallels the standard agnostic-PAC learnability framework. The model is defined as follows: We assume there exists an unknown distribution $\mathcal{D}$ over $\{1, \ldots, m\} \times \{1, \ldots, n\}$. Each instantiation (i, j) provides the value $X_{i,j}$ of an entry at a randomly picked row i and column j. An i.i.d. sample $S = \{i_\alpha, j_\alpha\}$ of indices is chosen, and the corresponding entries $\{X_{i_\alpha,j_\alpha}\}$ are revealed. Our goal is to find a matrix $W \in \mathcal{W}$ such that its risk (or generalization error), $\mathbb{E}_{(i,j)\sim\mathcal{D}}[\ell(W_{i,j}, X_{i,j})]$, is as close as possible to the smallest possible risk over all $W \in \mathcal{W}$. It is well-known that this can be achieved by solving the optimization problem in Eq. (1), if we can provide a non-trivial uniform sample complexity bound, namely a bound on

$$\sup_{W \in \mathcal{W}} \left(\mathbb{E}_{i,j}[\ell(W_{i,j}, X_{i,j})] - \frac{1}{|S|} \sum_{\alpha=1}^{|S|} \ell(W_{i_\alpha,j_\alpha}, X_{i_\alpha,j_\alpha}) \right). \tag{2}$$

A major focus of this paper is studying the difficulties and possibilities of obtaining such bounds.

3 Sample Complexity Bounds for the Trace-Norm

Consider the class of trace-norm constrained matrices, $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$, where we assume $t = \Theta(\sqrt{mn})$. Although learning with respect to this class is widely used in collaborative filtering, understanding its generalization and sample-complexity properties has proven quite elusive. In Srebro and Shraibman (2005), sample complexity bounds of the form $O(\sqrt{(m+n)/|S|})$ (ignoring logarithmic factors) were obtained under the strong assumption of a uniform distribution over the matrix entries. However, this assumption does not correspond to real-world collaborative filtering datasets, where the distribution of the revealed entries appears to be highly non-uniform. Other

works, which focused on exact matrix completion, such as Candès and Tao (2009), Candès and Recht (2009), also assume a uniform sampling distribution.

While the bounds in Srebro and Shraibman (2005) can be adapted to a non-uniform distribution, they lead to bounds which are no better than $O((m+n)/\sqrt{|S|})$. This implies a required sample size comparable or larger than the total number of entries in the matrix. It is a trivial bound, since the entire goal of collaborative filtering is prediction based on observing just a small subset of the matrix entries.

The analysis mentioned above used Rademacher complexity to quantify the richness of the hypothesis class $\mathcal{W}$, and will be utilized in our analysis as well. Formally, we define the Rademacher complexity of a hypothesis class $\mathcal{W}$ combined with a loss function ℓ, with respect to a sample S, as

$$R_S(\ell \circ \mathcal{W}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{W \in \mathcal{W}} \frac{1}{|S|} \sum_{\alpha=1}^{|S|} \sigma_\alpha \ell(W_{i_\alpha,j_\alpha}, X_{i_\alpha,j_\alpha}) \right], \tag{3}$$

where $\sigma_1, \ldots, \sigma_{|S|}$ are i.i.d. random variables taking the values -1 and $+1$ with equal probability.

Rademacher complexities play a key role in obtaining sample complexity bounds, either in expectation or in high probability. The following is a typical example (based on (Boucheron and Lugosi, 2005, Theorem 3.2)):

Theorem 1 *The expected value of Eq. (2) is at most $2R_S(\ell \circ \mathcal{W})$. Moreover, if there is a constant b_ℓ such that $\sup_{i,j,W \in \mathcal{W}} |\ell(W_{i,j}, X_{i,j})| \leq b_\ell$, then for any $\delta \in (0,1)$, Eq. (2) is bounded with probability at least $1-\delta$ by $2R_S(\ell \circ \mathcal{W}) + b_\ell \sqrt{2\log(2/\delta)/|S|}$.*

Thus, to get non-trivial learning guarantees, one can focus on effectively bounding the Rademacher complexity $R_S(\ell \circ \mathcal{W})$.

Unfortunately, for the class $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$ and general distributions and losses, the analysis of $R_S(\ell \circ \mathcal{W})$ performed in Srebro and Shraibman (2005) appears to be tight, yet still leads to vacuous results. The main drive of our paper is that by modifying the setting in some very simple ways, which often correspond to collaborative filtering as done in practice, one can obtain non-trivial learning guarantees.

4 Results for the Inductive Model

In this section, we show that by introducing *boundedness* conditions into the learning problem, one can obtain non-trivial bounds on the Rademacher complexity, and hence on the sample complexity of learning with trace-norm constraints.

We will start with the case where we actually learn with respect to the hypothesis class of trace-norm-constrained matrices, $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$, and the only boundedness is in terms of the loss function:

Theorem 2 *Consider the hypothesis class $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$. Suppose that for all i,j the loss function $\ell(\cdot, X_{i,j})$ is both b_ℓ-bounded and l_ℓ-Lipschitz in its first argument: Namely, that $\ell(W_{i,j}, X_{i,j})| \leq b_\ell$ for any W, i, j, and that $\frac{|\ell(W_{i,j},X_{i,j}) - \ell(W'_{i,j},X_{i,j})|}{|W_{i,j} - W'_{i,j}|} \leq l_\ell$ for any W, W', i, j. Then*

$$R_S(\ell \circ \mathcal{W}) \leq \sqrt{9 C l_\ell b_\ell \frac{t(\sqrt{m}+\sqrt{n})}{|S|}},$$

where C is the universal constant appearing in Thm. 7.

Assuming $t = \Theta(\sqrt{mn})$, the theorem implies that a sample of size $O(n\sqrt{m} + m\sqrt{n})$ is sufficient to obtain good generalization performance. We note that the boundedness assumption is non-trivial, since the trace-norm constraint does not imply entries of constant magnitude (the entries can be as large as t for a matrix whose trace norm is t). On the other hand, as discussed earlier, the obtainable bound on the Rademacher complexity without a boundedness assumption is no better than $O((m+n)/\sqrt{|S|})$, which leads to a trivial required sample size of $O((m+n)^2)$. Moreover, we emphasize that the result makes no assumptions on the underlying distribution from which the data was sampled. The proof is presented in Subsection 7.1. We note that it relies on a decomposition technique which might also be useful for regularizers other than the trace-norm.

An alternative way to introduce boundedness, and get a non-trivial guarantee, is by composing the entries of a matrix W with a bounded transfer function. In particular, rather than just learning a

matrix W with bounded trace-norm, we can learn a model $\phi \circ W$, where W has bounded trace-norm, and $\phi : \mathbb{R} \mapsto I$ is a fixed mapping of each entry of W into some bounded interval $I \subseteq \mathbb{R}$. This model is used in practice, and is useful in the common situation where the entries of X are known to be in a certain bounded interval. In Sec. 6, we return to this model in greater depth. In terms of the theoretical guarantee, one can provide a result substantially similar to Thm. 2, without assuming boundedness of the loss function. The proof is provided in the appendix - it uses similar techniques to the one of Thm. 2, but applies them somewhat differently.

Theorem 3 *Consider the hypothesis class $\mathcal{W} = \{\phi \circ W : \|W\|_{tr} \leq t\}$. Let $\phi : \mathbb{R} \mapsto [-b_\phi, b_\phi]$ be a bounded l_ϕ-Lipschitz function, and suppose that for all i, j, $\ell(\cdot, X_{i,j})$ is l_ℓ-Lipschitz on the domain $[-b_\phi, b_\phi]$. Then*

$$R_S(\ell \circ \mathcal{W}) \leq l_\ell \sqrt{9C l_\phi b_\phi \frac{t(\sqrt{m} + \sqrt{n})}{|S|}},$$

where C is the universal constant appearing in Thm. 7.

The bound in this theorem scales similar to Thm. 2, in terms of its dependence on m, n. Another possible variant is directly learning a matrix W with both a constraint on the trace-norm, as well as an ∞-norm constraint (i.e. $\max_{i,j} |W_{i,j}|$) which forces the matrix entries to be constant. This model has some potential benefits which shall be further discussed in Sec. 8.

Theorem 4 *Consider the hypothesis class $\mathcal{W} = \{W : \|W\|_{tr} \leq t, \|W\|_\infty \leq b\}$, where $\|W\|_\infty = \max_{i,j} |W_{i,j}|$. Suppose that for all i, j, $\ell(\cdot, X_{i,j})$ is l_ℓ-Lipschitz on the domain $[-b, b]$. Then*

$$R_S(\ell \circ \mathcal{W}) \leq l_\ell \sqrt{9Cb \frac{t(\sqrt{n} + \sqrt{m})}{|S|}},$$

where C is the universal constant appearing in Thm. 7.

Assuming b is a constant (which is the reasonable assumption here), we get a similar bound as before.

So, we see that by inserting mild boundedness assumptions on the loss function or the matrix entries, it is possible to derive non-trivial guarantees for learning with trace norm constraints. These were all obtained under the standard inductive model, where we assume that our data is an i.i.d. sample from an underlying distribution. In the next section, we will discuss a different learning model, which we argue to more closely resemble collaborative filtering as done in practice, and leads to better bounds on the Rademacher complexity, without making boundedness assumptions.

5 Improved Results for the Transductive Model

In the inductive model we have considered so far, the goal is to predict well with respect to an unknown distribution over matrix entries, given an i.i.d. sample from that distribution. The *transductive* learning model (see for instance Vapnik (1998)) is different, in that our goal is to predict well with respect to a *specific* subset of entries, whose location is known in advance. More formally, we fix an arbitrary subset of S entries, and then split it uniformly at random into two subsets $S_{train} \cup S_{test}$. We are then given the values of the entries in S_{train}, and our goal is to predict the values of the entries in S_{test}. For simplicity, we will assume that $|S_{train}| = |S_{test}| = |S|/2$, but our results can be easily generalized to more general partitions.

We note that this procedure is *exactly* the one often performed in experiments reported in the literature: Given a dataset of entries, one randomly splits it into a training set and a test set, learns a matrix on the training set, and measures its performance on the test set (e.g., Toh and Yun (2009), Jaggi and Sulovský (2010)). Even for other train-test split methods, such as holding out a certain portion of entries from each row, the transductive model seems closer to reality than the inductive model. Moreover, the transductive model captures another important feature of real-world collaborative filtering: the fact that no entry is repeated in the training set. In contrast, in the inductive model the training set is collected i.i.d., so the same entry might be sampled several time over. In fact, this is virtually certain to happen whenever the sample size is at least on the order of $\sqrt{mn}$, due to the birthday paradox. This does not appear to be a mere technicality, since the proofs of our theorems in the inductive model have to rely on a careful separation of the entries according to the number of times they were sampled. However, in reality each entry appears in the dataset only once, matching the transductive learning setting.

To analyze the transductive model, we require analogues of the tools we have for the inductive model, such as the Rademacher complexity. Fortunately, such analogues were already obtained in El-Yaniv and Pechyoni (2009), and we will rely on their results. In particular, based on Theorem 1 in that paper, we can use our notion of Rademacher complexity, as defined in Eq. (3), to provide sample complexity bounds in the transductive model[1]:

Theorem 5 *Fix a hypothesis class $\mathcal{W}$, and suppose that $\sup_{i,j,W\in\mathcal{W}} |\ell(W_{i,j}, X_{i,j})| \leq b_\ell$. Let a set S of ≥ 2 distinct indices be fixed, and suppose it is uniformly and randomly split to two equal subsets S_{train}, S_{test}. Then with probability at least $1-\delta$ over the random split, it holds for any $W \in \mathcal{W}$ that*

$$\frac{\sum_{(i,j)\in S_{test}} \ell(W_{i,j}, X_{i,j})}{|S_{test}|} - \frac{\sum_{(i,j)\in S_{train}} \ell(W_{i,j}, X_{i,j})}{|S_{train}|}$$
$$\leq 4R_S(\ell \circ \mathcal{W}) + \frac{b_\ell \left(11 + 4\sqrt{\log(1/\delta)}\right)}{\sqrt{|S_{train}|}}.$$

We now present our main result for the transductive model, which implies non-trivial bounds on the Rademacher complexity of matrices with constrained trace-norm. Unlike the inductive model, here we make no additional boundedness assumptions, yet the bound is superior. The proof appears in Subsection 7.2.

Theorem 6 *Consider the hypothesis class $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$. Then in the transductive model, if we let $N_i = \max_i |\{j : (i,j) \in S\}|$ and $N_j = \max_j |\{i : (i,j) \in S\}|$, then*

$$R_S(\ell \circ \mathcal{W}) \leq C l_\ell \frac{3t\left(\sqrt{m} + \sqrt{n}\right)}{2|S|},$$

where C is the universal constant appearing in Thm. 7. Alternatively, we also have

$$R_S(\ell \circ \mathcal{W}) \leq C l_\ell \frac{t \ \max\left\{\max_i \sqrt{N_i}, \max_j \sqrt{N_j}\right\}}{|S|} \sqrt[4]{\log(\min\{m,n\})},$$

where C is the universal constant appearing in Thm. 8.

We note that the second bound, while containing an additional logarithmic term, depends on the distribution of the entries, and can be considerably tighter than the worst-case. To see this, suppose as usual that $t = \Theta(\sqrt{mn})$, and (for simplicity) $m = n$. Then in the worst-case, the bound becomes meaningful when $|S| = \Omega(n^{3/2})$. However, if the entries in S are (approximately) uniformly distributed throughout the matrix, then the maximal number of entries in each row and column is $O(|S|/n)$. In that case, we obtain the bound

$$R_S(\ell \circ \mathcal{W}) \leq \tilde{O}\left(\sqrt{\frac{n}{|S|}}\right)$$

(ignoring logarithmic factors), which is already meaningful when $|S| = \tilde{\Omega}(n)$. Interestingly, this bound is similar (up to logarithmic factors) to previous bounds in the inductive setting, such as in Srebro and Shraibman (2005), which relied on a uniform distribution assumption. However, our Rademacher complexity bound in Thm. 6 also applies to non-uniform distributions, and is meaningful for any distribution.

Compared to the results in Sec. 4, the result here is also superior in that the Rademacher complexity does not depend on the loss magnitude bound b_ℓ. Although this factor does appear in a different term in the overall sample complexity bound (Thm. 5), we do not know if this dependence is essential. Indeed, in the inductive setting, such terms appear only in high-probability bounds, and even then it is possible to prove oracle inequalities which depend merely on $\sqrt{b_\ell/|S|}$ rather than $b_\ell/\sqrt{|S|}$ (see for instance (Srebro et al., 2010, Theorem 1)). If such bounds can be proven in the transductive case as well, it will lead to non-trivial sample complexity results without any boundedness assumptions whatsoever[2].

[1] In El-Yaniv and Pechyoni (2009), a more general notion of transductive Rademacher complexity was defined, where the σ_α random variables could also take 0 values. However, when $|S_{train}| = |S_{test}|$, that complexity can always be upper bounded by the standard definition of Rademacher complexity - see Lemma 1 in their paper.

[2] To see why, suppose that the trace norm $t = \Theta(\sqrt{mn})$, and the loss function is Lipschitz. Then $b_\ell = \Theta(\sqrt{mn})$ as well, so a $\sqrt{b_\ell/|S|}$ term will be on the order of $\sqrt{\sqrt{mn}/|S|}$.

Another interesting feature of Thm. 6 is that the Rademacher complexity falls off at the rate of $O(1/|S|)$ rather than $O(1/\sqrt{|S|})$. While such a "fast rate" is unusual in the inductive setting, here it is a natural outcome of the different modeling of the training data. This does not lead to a $O(1/|S|)$ sample complexity bound, because the bound in Thm. 5 contains an additional low rate term $O(1/\sqrt{|S|})$. However, it still leads to a better bound because the low rate term is not explicitly multiplied by functions of m, n or t.

6 Should We Enforce Boundedness?

As mentioned previously, we often know the range of entries to be predicted (e.g. 1 to 5 for movie rating prediction). The results of Sec. 4 suggest that in the inductive model, some sort of boundedness seems essential to get non-trivial results. In the transductive model, boundedness also plays a somewhat smaller role, by appearing in the final sample-complexity bound (Thm. 5), although not in the Rademacher complexity bound (Thm. 6). These results suggest the natural idea of incorporating into the learning algorithm the prior knowledge we have on the range of entries. Indeed, several recent papers have considered the possibility of directly learning a model $\phi \circ W$, where ϕ is usually a sigmoid function (Salakhutdinov and Mnih (2007), Ma et al. (2008), Piotte and Chabbert (2009), Kozma et al. (2009)). Another common practice (not just with trace-norm regularization) is to clip the learned matrix entries to the known range. Since our theoretical results are just upper bounds, the effect of boundedness is not sufficiently clear. Thus, it is of interest to understand how clipping or enforcing boundedness in the learning model helps in practice. We note that while bounded models have been tested experimentally, we could not find in the literature a clear empirical study of their effect, in the context of trace-norm regularization.

We conducted experiments on two standard collaborative filtering datasets, movielens100K and movielens1M[3]. movielens100K contains 10^5 ratings of 943 users for 1770 movies, while movielens1M contains 10^6 ratings of 6040 users for 3706 movies. All ratings are in the range $[1, 5]$. For each dataset, we performed a random $80\% - 20\%$ of the data to obtain a training set and a test set. We considered two hypothesis classes: trace-norm constrained matrices $\{W : \|W\|_{tr} \leq t\}$, and bounded trace-norm constrained matrices $\{\phi \circ W : \|W\|_{tr} \leq t\}$, where ϕ is a sigmoid function interpolating between 1 and 5. For each hypothesis class, we trained a trace-norm regularized algorithm using the squared loss. Specifically, we used the common approach of stochastic gradient descent on a factorized representation $W = U^\top V$. This approach is based on the well-known fact, that the trace norm can also be defined as $\|W\|_{tr} = \min_{W=U^\top V} \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right)$. Thus, finding the best W by optimizing a soft trace-norm regularized problem, $\sum_{(i,j)\in S}\left(X_{i,j} - W_{i,j}\right)^2 + \lambda\|W\|_{tr}$, can be reduced to finding U, V which minimize

$$\sum_{(i,j)\in S}\left(X_{i,j} - U_i^\top V_j\right)^2 + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right). \tag{4}$$

Similarly, for learning bounded models, we can find U, V which minimize

$$\sum_{(i,j)\in S}\left(X_{i,j} - \phi(U_i^\top V_j)\right)^2 + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right). \tag{5}$$

We note that both problems are non-convex, although for the formulation in Eq. (4), it is possibly to show there are any local minimum is also a global one.

Tuning of λ was performed with a validation set. Note that in practice, for computational reasons, one often constrains U and V to have a bounded number of rows. However, this forces W to have low rank, which is an additional complexity control. Since our goal is to study the performance of trace-norm constrained matrices, and not matrices which are also low-rank, we did not constrain U, V in this manner. The downside of this is that we were unable to perform experiments on very large-scale datasets, such as Netflix, and that is why we focused on the more modest-sized movielens100K and movielens1M datasets.

To estimate the performance of the learned matrix W on the test set, we used two measures which are standard in the literature: the root-mean-squared-error (RMSE), $\sqrt{\sum_{(i,j)\in S_{test}} \frac{(W_{i,j}-X_{i,j})^2}{|S_{test}|}}$, and the normalized-mean-absolute-error (NMAE), $\sum_{i,j\in S_{test}} \frac{|W_{i,j}-X_{i,j}|}{r|S_{test}|}$, where r is the range of possible values in X ($5 - 1 = 4$ for our datasets).

[3] `www.grouplens.org/node/73`

	100K (NMAE)	100K (RMSE)	1M (NMAE)	1M (RMSE)
unclipped	0.1882 ± 0.0005	0.9543 ± 0.0019	0.1709 ± 0.0003	0.8670 ± 0.0016
clipped	0.1874 ± 0.0005	0.9486 ± 0.0018	0.1706 ± 0.0002	0.8666 ± 0.0016
bounded	0.1871 ± 0.0004	0.9434 ± 0.0023	0.1698 ± 0.0002	0.8618 ± 0.0017
Δ Clipping ($*10^{-3}$)	0.77 ± 0.07	5.7 ± 0.6	0.33 ± 0.01	0.48 ± 0.04
Δ Bounding ($*10^{-3}$)	0.3 ± 0.4	5.2 ± 1.5	0.79 ± 0.02	4.8 ± 0.1

Table 1: Error on test set (mean and standard deviation over 5 repeats of the experiment). The columns refer to the dataset (movielens100K or movielens1M) and the performance measure used (NMAE or RMSE). The first two rows refer to the results using the 'unbounded' model as in Eq. (4), with the output used as-is or clipped to the range $[1-5]$. The third row refers to the results using the 'bounded' model as in Eq. (5). The fourth row is the improvement in test error by clipping the predictions after learning (i.e. the difference between the first and second row). The fifth row is the additional improvement achieved by using a bounded model (i.e., the difference between the second and third row).

The experiments were repeated 5 times over random train-test splits of the data, and the results are summarized in Table 1. From the table, we see that in almost all cases, clipping and bounding lead to a statistically significant improvement. However, note that in absolute terms, the improvement is rather modest, especially with the NMAE measure which is less sensitive to large mispredictions. This accords with our theoretical results: boundedness seems to be an important and useful property, but in the transductive model (corresponding to our experiments) it plays only a modest role.

Empirically, one would have expected the use of bounded models to help a lot (in absolute terms), if learning just trace-norm constrained matrices (without clipping/bounding) leads to many predictions being outside the interval $[1, 5]$, in which we know the ratings lie. But indeed, this does not seem to be the case. Table 2 shows the prediction with largest magnitude, over all entries in the test set, as well as the percentage of predictions which fall outside the $[1, 5]$ interval. It is clearly evident that such out-of-interval predictions are relatively rare, and this explains why the bounding and clipping only leads to modest improvements.

	100K	1M
largest value	5.95 ± 0.35	6.13 ± 0.16
% outside interval	0.69 ± 0.05	0.79 ± 0.01

Table 2: Out-of-Interval Values

We emphasize that our results should only be interpreted in the context of pure trace-norm regularization. There are many other approaches to collaborative filtering, and it is quite possible that using bounded models has more or less impact in the context of other approaches or for other application domains.

7 Proofs

Our proofs utilize the following two theorems, which bounds the expected spectral norm of random matrices.

Theorem 7 (Latała (2005)) *Let Z be a matrix composed of independent zero-mean entries. Then for some fixed constant C, $\mathbb{E}[\|Z\|_{sp}]$ is at most*

$$C\left(\max_i \sqrt{\sum_j \mathbb{E}[Z_{i,j}^2]} + \max_j \sqrt{\sum_i \mathbb{E}[Z_{i,j}^2]} + \sqrt[4]{\sum_{i,j} \mathbb{E}[Z_{i,j}^4]}\right).$$

Theorem 8 (Seginer (2000)) *Let A be an arbitrary $m \times n$ matrix, such that $m, n > 1$. Let Z denote a matrix composed of independent zero-mean entries, such that $Z_{i,j} = A_{i,j}$ with probability*

$1/2$ *and* $Z_{i,j} = -A_{i,j}$ *with probability* $1/2$. *Then for some fixed constant* C, $\mathbb{E}[\|A\|_{sp}]$ *is at most*

$$C\sqrt[4]{\log(\min\{m,n\})}\max\left\{\max_i\sqrt{\sum_j A_{i,j}^2}, \max_j\sqrt{\sum_i A_{i,j}^2}\right\}$$

7.1 Proof of Thm. 2

We write $R_S(\ell\circ\mathcal{W})$ as

$$\frac{1}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}[\sup_{W\in\mathcal{W}}\sum_{i,j}\Sigma_{i,j}\ell(W_{i,j},X_{i,j})],$$

where Σ is a matrix whose (i,j)-th entry is defined as $\sum_{\alpha:i_\alpha=i,j_\alpha=j}\sigma_\alpha$. A standard Rademacher analysis will usually proceed to reduce this to $\frac{1}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\sum_{i,j}\Sigma_{i,j}W_{i,j}\right]$. However, this will lead to a trivial bound. Instead, we will do something more refined. Given i,j, let $h_{i,j}$ be the number of times the sample S hits entry i,j, or more precisely $h_{i,j} = |\{\alpha : i_\alpha = i, j_\alpha = j\}|$. Let $p>0$ be an arbitrary parameter to be specified later, and define

$$Y_{i,j} = \begin{cases}\Sigma_{i,j} & h_{i,j} > p\\ 0 & h_{i,j}\le p\end{cases} \qquad Z_{i,j} = \begin{cases}0 & h_{i,j} > p\\ \Sigma_{i,j} & h_{i,j}\le p.\end{cases} \tag{6}$$

Clearly, we have $\Sigma = Y + Z$. Thus, we can upper bound the Rademacher complexity by

$$\frac{1}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\sum_{i,j}Y_{i,j}\ell(W_{i,j},X_{i,j})\right] + \frac{1}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\sum_{i,j}Z_{i,j}\ell(W_{i,j},X_{i,j})\right]. \tag{7}$$

Since $|\ell(W_{i,j},X_{i,j})|\le b_\ell$, the first term can be upper bounded by

$$\frac{1}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[b_\ell\sum_{i,j}|Y_{i,j}|\right] = \frac{b_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}[\|Y\|_1]. \tag{8}$$

Using the Rademacher contraction principle[4], the second term in Eq. (7) can be upper bounded by

$$\frac{l_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\sum_{i,j}Z_{i,j}W_{i,j}\right].$$

Applying Hölder's inequality, and using the fact that the spectral norm $\|\cdot\|_{sp}$ is the dual to the trace norm $\|\cdot\|_{tr}$, we can upper bound the above by

$$\frac{l_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{W\in\mathcal{W}}[\|Z\|_{sp}\|W\|_{tr}] = \frac{l_\ell t}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\|Z\|_{sp}\right]. \tag{9}$$

Combining this with Eq. (8) and substituting into Eq. (7), we get an upper bound of the form

$$\frac{b_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\|Y\|_1\right] + \frac{l_\ell t}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\|Z\|_{sp}\right].$$

Using Lemma 9 and Lemma 10, which are given below, we can upper bound this by

$$\frac{b_\ell}{\sqrt{p}} + \frac{2.2Cl_\ell t\sqrt{p}(\sqrt{m}+\sqrt{n})}{|S|},$$

where p is the parameter used to define Y and Z in Eq. (6). Choosing $p = \frac{|S|b_\ell}{2.2Cl_\ell t(\sqrt{m}+\sqrt{n})}$, we get the bound in the theorem.

Lemma 9 *Let* Y *be a random matrix defined as in Eq. (6). Then*

$$\mathbb{E}[\|Y\|_1]\le\mathbb{E}\left[\sum_{i,j:h_{i,j}>p}\sqrt{h_{i,j}}\right]\le\frac{|S|}{\sqrt{p}}$$

[4]Strictly speaking, we use a slight generalization of it, where the loss function is allowed to differ w.r.t. every $W_{i,j}$ - see (Meir and Zhang, 2003, Lemma 5)

Proof: $\mathbb{E}[\|Y\|_1]$ equals

$$\mathbb{E}\left[\sum_{i,j:h_{i,j}>p}|\Sigma_{i,j}|\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i,j:h_{i,j}>p}\left(\left|\sum_{\alpha:(i_\alpha,j_\alpha)=(i,j)}\sigma_\alpha\right|\right)\ \middle|\ \{h_{i,j}\}\right]\right]$$

The expression inside the absolute value is the sum of $h_{i,j}$ i.i.d. random variables, and it is easily seen that its expected absolute value is at most $\sqrt{h_{i,j}}$. Therefore, we can upper bound the above by $\mathbb{E}[\sum_{i,j:h_{i,j}>p}\sqrt{h_{i,j}}]$. We can further upper bound it, in a manner which does not depend on the values of $h_{i,j}$, by

$$\max_{c\in\{1,\ldots,mn\}}\ \max_{h_1,\ldots,h_c\in\mathbb{R}:\forall i\ h_i>p,\sum_{i=1}^c h_i=|S|}\ \sum_{i=1}^c\sqrt{h_i}.$$

Note that the constraints imply that

$$|S| = \sum_{i=1}^c h_i \geq \sqrt{p}\sum_{i=1}^c\sqrt{h_i},$$

so $\sum_{i=1}^c\sqrt{h_i}$ can be at most $|S|/\sqrt{p}$ as required. ■

Lemma 10 *Let Z be a random matrix defined as in Eq. (6). Then $\mathbb{E}_{\boldsymbol{\sigma}}[\|Z\|_{sp}]$ is at most*

$$C\left(\max_i\sqrt{\sum_{j:h_{i,j}\leq p}h_{i,j}}+\max_j\sqrt{\sum_{i:h_{i,j}\leq p}h_{i,j}}+\sqrt[4]{3\sum_{i,j:h_{i,j}\leq p}h_{i,j}^2}\right),$$

where C is the universal constant which appears in the main theorem of Latała (2005). Moreover, this quantity can be upper bounded by $2.2C\sqrt{p}\,(\sqrt{m}+\sqrt{n})$

Proof: With $h_{i,j}$ held fixed, Z is a random matrix composed of independent entries. By using Thm. 7, we only need to analyze $\mathbb{E}[Z_{i,j}^2]$ and $\mathbb{E}[Z_{i,j}^4]$. For any i,j, if $h_{i,j}\leq p$ then $Z_{i,j}$ is a sum of $h_{i,j}$ i.i.d. variables taking values in $\{-1,+1\}$. Therefore, $\mathbb{E}[Z_{i,j}^2]=h_{i,j}$ and $\mathbb{E}[Z_{i,j}^4]\leq 3h_{i,j}^2$. Plugging into Thm. 7 yields the first part of the lemma. To get the second part, we can upper bound the right-hand side of the first part by

$$C\sqrt{p}\left(\sqrt{m}+\sqrt{n}+\sqrt[4]{3mn}\right)\ \leq\ C\sqrt{p}\left(\sqrt{m}+\sqrt{n}+\sqrt[4]{3/2}(\sqrt{m}+\sqrt{n})\right)\ \leq\ 2.2C\sqrt{p}\left(\sqrt{m}+\sqrt{n}\right).$$

■

7.2 Proof of Thm. 6

We write $R_S(\ell\circ\mathcal{W})$ as

$$\frac{1}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\sum_{i,j}\Sigma_{i,j}\ell(W_{i,j},X_{i,j})\right],$$

where Σ is a matrix with $\sigma_{i,j}$ in its (i,j)-th entry, if $(i,j)\in S$, and 0 otherwise. By the Rademacher contraction property[5], we can upper bound this by

$$\frac{l_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\sum_{i,j}\Sigma_{i,j}W_{i,j}\right].$$

By Hölder's inequality, this is at most

$$\frac{l_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\|\Sigma\|_{sp}\|W\|_{tr}\right] = \frac{l_\ell t}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\|\Sigma\|_{sp}\right]. \tag{10}$$

The setting so far is rather similar to the one we had in the inductive setting (see the proof of any of the theorems in Sec. 4). But now, we need to bound just the expected spectral norm of Σ, which

[5] As in the inductive case, we use in fact a slight generalization where the loss function is allowed to differ w.r.t. every $W_{i,j}$, as in (Meir and Zhang, 2003, Lemma 5).

is guaranteed to have only a single Rademacher variable in each entry. By applying Thm. 7 on Eq. (10), we get

$$R_S(\ell \circ \mathcal{W}) \leq C l_\ell \frac{t\left(\sqrt{N_i} + \sqrt{N_j} + \sqrt[4]{|S|}\right)}{|S|}.$$

Since S can contain at most m and n indices for any single row and column respectively, and $\sqrt[4]{|S|} \leq \sqrt[4]{mn} \leq \frac{1}{2}\left(\sqrt{m} + \sqrt{n}\right)$, we can upper bound the above by $3Cl_\ell t\left(\sqrt{m} + \sqrt{n}\right)/(2|S|)$.

To get the other bound in the theorem, we apply Thm. 8 instead of Thm. 8 on Eq. (10).

8 Discussion

In this paper, we analyzed the sample complexity of collaborative filtering with trace-norm regularization, obtaining the first non-trivial, distribution-free guarantees. Our results were based on either mild boundedness assumptions, or a switch from the standard inductive learning model to the transductive learning model. Moreover, we argue that such a transductive model may be a better way to model collaborative filtering as performed in practice, as it seems more natural and leads to a substantial difference in terms of obtainable results. Finally, we discussed the issue of learning with bounded models, and provided an empirical study which indicates that these lead to a modest yet significant improvement in performance, corroborating our theoretical findings.

On the theoretical side, one obvious open question is the tightness of our bounds. In a nutshell, if we assume that we learn $n \times n$ matrices with a $\Theta(n)$ trace-norm, then our bounds imply a required sample complexity of $O(n^{3/2})$. On the other hand, Salakhutdinov and Srebro (2010) presented an example requiring a sample complexity of at least $\Omega(n^{4/3})$, which applies to all the settings we have discussed earlier. Currently, we do not know how to bridge this gap. Another issue is understanding the implications of our analysis to other types of matrix regularization, such as weighted trace-norm (Salakhutdinov and Srebro (2010)).

As to the use of bounded models, we note that although they seem beneficial in our experiments, they can also lead to non-convex optimization problems. While this did not seem to hurt performance in our experiments, it might be more harmful in other datasets of applications. One possible way to enforce boundeded predictions while maintaining convexity is using ∞-norm constraints, as in Thm. 4. Minimizing the average loss with respect to such constraints is a convex optimization problem, and can be done with a generic SDP solver. However, a generic solver won't scale to large datasets. Thus, designing an *efficient* convex optimization algorithm, which combines trace-norm and ∞-norm constraints, is a potentially useful, yet non-trivial challenge.

Acknowledgements

We thank Nati Srebro and Ruslan Salakhutdinov for helpful discussions, as well as the anonymous reviewers for valuable comments.

References

F. Bach. Consistency of trace-norm minimization. *Journal of Machine Learning Research*, 9: 1019–1048, 2008.

O. Bousquet Boucheron, S. and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323 – 375, 2005.

E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9, 2009.

E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2009.

Ran El-Yaniv and Dmitry Pechyoni. Transductive rademacher complexity and its applications. *Journal of AI Research*, 35:193–234, 2009.

M. Jaggi and M. Sulovský. A simple algorithm for nuclear norm regularized problems. In *ICML*, 2010.

L. Kozma, A. Ilin, and T. Raiko. Binary principal component analysis in the netflix collaborative filtering task. In *IEEE MLSP Workshop*, 2009.

R. Latała. Some estimates of norms of random matrices. *Proceedings of the AMS*, 133(5): 1273–1282, 2005.

J. Lee, B. Recht, R. Salakhutdinov, N. Srebro, and J. Tropp. Practical large-scale optimization for max-norm regularization. In *NIPS*, 2010.

H. Ma, H. Yang, M. Lyu, and I. King. Sorec: social recommendation using probabilistic matrix factorization. In *CIKM*, 2008.

R. Meir and T. Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.
S. Negahban and M. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. arXiv:1009.2118, 2010.
M. Piotte and M. Chabbert. The pragmatic theory solution to the netflix grand prize. Available at `http://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf`, 2009.
R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, 2007.
R. Salakhutdinov and N. Srebro. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *NIPS*, 2010.
Yoav Seginer. The expected norm of random matrices. *Combinatorics, Probability & Computing*, 9(2):149–166, 2000.
N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *COLT*, 2005.
N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. In *NIPS*, 2004.
N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low-noise and fast rates. In *NIPS*, 2010.
K. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Optimization Online*, 2009.
V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

A Proof of Thm. 3

We can rewrite the definition of $R_S(\ell \circ \mathcal{W})$ (see Eq. (3)) as

$$\frac{1}{|S|} E_{\boldsymbol{\sigma}} \left[\sup_{W \in \mathcal{W}} \sum_{i,j} \Sigma_{i,j} \ell(W_{i,j}, X_{i,j}) \right],$$

where Σ is a matrix defined as $\Sigma_{i,j} = \sum_{\alpha: i_\alpha = i, j_\alpha = j} \sigma_\alpha$. Using the Rademacher contraction principle (as in (Meir and Zhang, 2003, Lemma 5)), this is at most

$$\frac{l_\ell}{|S|} E_{\boldsymbol{\sigma}} \left[\sup_{W \in \mathcal{W}} \sum_{i,j} \Sigma_{i,j} W_{i,j} \right]. \tag{11}$$

Decomposing $\Sigma = Y + Z$ as in Eq. (6) according to a parameter p, we can upper bound the Rademacher complexity by

$$\frac{l_\ell}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{W \in \mathcal{W}} \sum_{i,j} Y_{i,j} W_{i,j} \right] + \frac{l_\ell}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{W \in \mathcal{W}} \sum_{i,j} Z_{i,j} W_{i,j} \right]. \tag{12}$$

By definition of $\mathcal{W}$, $|W_{i,j}| \leq b_\phi$, so the first term can be upper bounded by

$$\frac{l_\ell}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[b_\phi \sum_{i,j} |Y_{i,j}| \right] = \frac{l_\ell b_\phi}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} [\|Y\|_1]. \tag{13}$$

The second term in Eq. (12) equals

$$\frac{l_\ell}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{W: \|W\|_{tr} \leq t} \sum_{i,j} Z_{i,j} \phi(W_{i,j}) \right] \leq \frac{l_\ell l_\phi}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{W: \|W\|_{tr} \leq t} \sum_{i,j} Z_{i,j} W_{i,j} \right],$$

again by the Rademacher contraction principle. Applying Hölder's inequality, and using the fact that the spectral norm $\|\cdot\|_{sp}$ is the dual to the trace norm $\|\cdot\|_{tr}$, we can upper bound the above by

$$\frac{l_\ell l_\phi}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{W: \|W\|_{tr} \leq t} \|Z\|_{sp} \|W\|_{tr} \right] = \frac{l_\ell l_\phi t}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\|Z\|_{sp} \right].$$

Combining this with Eq. (13) and substituting into Eq. (12), we get an upper bound of the form

$$\frac{l_\ell b_\phi}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\|Y\|_1 \right] + \frac{l_\ell l_\phi t}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\|Z\|_{sp} \right].$$

Using Lemma 9 and Lemma 10, we can upper bound this by

$$\frac{l_\ell b_\phi}{\sqrt{p}} + \frac{2.2Cl_\ell l_\phi t\sqrt{p}(\sqrt{m}+\sqrt{n})}{|S|},$$

where p is the parameter used to define Y and Z in Eq. (6). Choosing $p = \frac{|S|b_\phi}{2.2Cl_\phi t(\sqrt{m}+\sqrt{n})}$, we get the bound in the theorem.

B Proof of Thm. 4

Before we begin, we will need the following technical result:

Lemma 11 *The dual of the norm* $\|W\| = \max\{\|W\|_{tr}/t, \|W\|_\infty/b\}$ *equals*

$$\|\Sigma\|_* = \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp},$$

Proof: By definition of a dual norm, we have

$$\|\Sigma\|_* = \sup_{W:\|W\|\leq 1} \langle \Sigma, W\rangle,$$

and our goal is to show that

$$\sup_{W:\|W\|\leq 1} \langle \Sigma, W\rangle = \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}.$$

First, we note that the dual norm of $\|W\|_{tr}$ is well-known to be the spectral norm $\|W\|_{sp}$ (namely, its largest singular value), and the dual of $\|W\|_\infty$ is the 1-norm $\|W\|_1 = \sum_{i,j}|W_{i,j}|$. Now, for any Y, Z such that $Y + Z = \Sigma$, we have by Hölder's inequality that

$$\sup_{W:\|W\|\leq 1} \langle \Sigma, W\rangle = \sup_{W:\|W\|\leq 1} \langle Y, W\rangle + \langle Z, W\rangle \leq \sup_{W:\|W\|\leq 1} \|Y\|_1\|W\|_\infty + \|Z\|_{sp}\|W\|_{tr} \leq b\|Y\|_1 + t\|Z\|_{sp}.$$

This holds for any Y, Z, and in particular

$$\sup_{W:\|W\|\leq 1} \langle \Sigma, W\rangle \leq \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}. \tag{14}$$

It remains to show the opposite direction, namely

$$\sup_{W:\|W\|\leq 1} \langle \Sigma, W\rangle \geq \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}.$$

To show this, let W^* be the matrix which maximizes the inner product above. We know that $\|W^*\| \leq 1$, which means that either $\|W^*\|_\infty \leq b$, or $\|W^*\|_{tr} \leq t$. If $\|W^*\|_\infty \leq b$, it follows that

$$\sup_{W:\|W\|\leq 1} \langle \Sigma, W\rangle = \sup_{W:\|W\|_\infty \leq b} \langle \Sigma, W\rangle = b\|\Sigma\|_1 \geq \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}.$$

In the other case, if $\|W^*\|_{tr} \leq t$, it follows that

$$\sup_{W:\|W\|\leq 1} \langle \Sigma, W\rangle = \sup_{W:\|W\|_{tr} \leq t} \langle \Sigma, W\rangle = t\|\Sigma\|_{sp} \geq \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}.$$

So in either case,

$$\sup_{W:\|W\|\leq 1} \langle \Sigma, W\rangle \geq \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}.$$

Combining this with Eq. (14), the result follows. ■

We now turn to the proof of Thm. 4 itself. Since $\ell(W_{i,j}, X_{i,j})$ is assumed to be l_ℓ-Lipschitz, we can use the Rademacher contraction principle to upper bound $R_S(\ell \circ \mathcal{W})$ by

$$l_\ell \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}} \frac{1}{|S|}\sum_{\alpha=1}^{|S|} \sigma_\alpha W_{i_\alpha,j_\alpha}\right] = \frac{l_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}} \sum_{i,j} \Sigma_{i,j}W_{i,j}\right],$$

where Σ is a matrix defined as $\Sigma_{i,j} = \sum_{\alpha:i_\alpha=i,j_\alpha=j} \sigma_\alpha$.

Thinking of Σ, W as vectors, the equation above is the expected supremum of an inner product between Σ and W. By Hölder's inequality, we can upper bound this by

$$\frac{l_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\|\Sigma\|_*\|W\|\right] \tag{15}$$

for any norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$. In particular, we will choose the norm $\|W\| = \max\{\|W\|_{tr}/t, \|W\|_\infty/b\}$. Note that by definition of W, $\sup_{W\in\mathcal{W}}\|W\| \le 1$. Also, by Lemma 11,

$$\|\Sigma\|_* = \min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp},$$

where $\|Y\|_1 = \sum_{i,j}|Y_{i,j}|$, and $\|Z\|_{sp}$ is the spectral norm of Z . Thus, we can upper bound Eq. (15) by

$$\frac{l_\ell}{|S|}\mathbb{E}_\Sigma\left[\min_{Y+Z=\Sigma} b\|Y\|_1 + t\|Z\|_{sp}\right]. \tag{16}$$

Recall that Σ is random matrix, where each entry is the sum of Rademacher variables. Let $h_{i,j}$ denote the number of variables 'hitting' entry (i,j) - formally, $h_{i,j} = |\{\alpha : (i_\alpha = i, j_\alpha = j\}|$. We can upper bound Eq. (16) by replacing the optimal decomposition of Σ into Y, Z by any fixed decomposition rule. In particular, for an arbitrary parameter p, we can decompose Σ into Y, Z as in Eq. (6), and get an upper bound on Eq. (16) of the form

$$\frac{l_\ell}{|S|}\left(b\mathbb{E}_\Sigma[\|Y\|_1] + t\mathbb{E}_\Sigma[\|Z\|_{sp}]\right). \tag{17}$$

Bounds for the two expectations are provided in Lemma 9 and Lemma 10. Plugging them in, we get

$$\frac{bl_\ell}{\sqrt{p}} + \frac{2.2l_\ell Ct\sqrt{p}\,(\sqrt{m}+\sqrt{n})}{|S|}.$$

Choosing $p = \frac{b|S|}{2.2Ct(\sqrt{m}+\sqrt{n})}$ and simplifying, we get the bound in the theorem.