Journal of Machine Learning Research 15 (2014) 3401-3423 Submitted 1/13; Revised 6/14; Published 10/14

Matrix Completion with the Trace Norm: Learning, Bounding, and Transducing

Ohad Shamir OHAD.SHAMIR@WEIZMANN.AC.IL
Department of Computer Science and Applied Mathematics
Weizmann Institute of Science
Rehovot 7610001, Israel

Shai Shalev-Shwartz SHAIS@CS.HUJI.AC.IL
School of Computer Science and Engineering
The Hebrew University
Givat Ram, Jerusalem 9190401, Israel

Editor: Tommi Jaakkola

Abstract

Trace-norm regularization is a widely-used and successful approach for collaborative filtering and matrix completion. However, previous learning guarantees require strong assumptions, such as a uniform distribution over the matrix entries. In this paper, we bridge this gap by providing such guarantees, under much milder assumptions which correspond to matrix completion as performed in practice. In fact, we claim that previous difficulties partially stemmed from a mismatch between the standard learning-theoretic modeling of matrix completion, and its practical application. Our results also shed some light on the issue of matrix completion 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.

Keywords: collaborative filtering, matrix completion, trace-norm regularization, transductive learning, sample complexity

1. Introduction

We consider the problem of matrix completion, where the goal is to predict entries of an unknown matrix based on a subset of its observed entries. A popular approach to achieve this is via trace-norm regularization, where one seeks a matrix that 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 (Srebro et al., 2004; Salakhutdinov and Mnih, 2007; Bach, 2008; Candès and Tao, 2009).

However, in terms of distribution-free guarantees, previous results on trace-norm regularization have been surprisingly weak. Most non-trivial guarantees (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 most matrix completion tasks, this is an extremely unrealistic assumption. For example, in the Netflix challenge data set, where the matrix contains the ratings of users (rows) for movies (columns), the number and distri-

bution of ratings differ drastically between users. Modeling such data as a uniform sample is not a reasonable assumption. Another paper (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 quite 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. This partially motivated work on alternative complexity measures for matrix completion, such as the max-norm and weighted variants of the trace-norm (see further discussion below).

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 matrix completion as performed *in practice*, it is possible to learn in a distribution-free manner by observing $\mathcal{O}(n^{3/2})$ entries from an $m \times n$ matrix (where $m \leq n$, and for a reasonable trace-norm regime). Moreover, this bound is tight. When $m = \Theta(n)$, this corresponds to viewing a vanishingly small portion of the entries, hence we get a non-trivial learning guarantee. In fact, we claim that the difficulties in providing such guarantees partially stemmed from a mismatch between the standard theoretical modeling of matrix completion, and its practical application. We emphasize that our bounds are weaker than previous bounds in the literature, which required observing as few as $\tilde{\mathcal{O}}(n)$ entries (up to log factors). However, these bounds hold only under restrictive distributional assumptions, whereas our bounds hold under any distribution, and are provably tight in such a distribution-free setting.

First, we show that one can obtain such guarantees, if one takes into account that the values to be predicted are bounded. 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 matrix completion. Instead, we look at the *transductive model*, where sampling of the data is done without replacement. In the context of matrix completion, 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. Finally, we discuss how recent work,

which appeared since the preliminary version of this paper was published, relate to and strengthen our observations.

The paper is structured as follows. We begin by describing the setting and the notation we use in Section 2, and introduce the sample complexity issues of matrix completion with the trace norm in Section 3. In Section 4, we show how we can non-trivially learn with the trace-norm in an inductive i.i.d. setting, under boundedness assumptions. In Section 5, we show how similar performance can be ensured if we switch from an inductive setting to a transductive setting, where each entry appears only once in the data. We provide matching lower bounds in Section 6. In Section 7, we experimentally investigate how boundedness assumptions affect practical performance. Section 8 contains a discussion of how some recent works relate to our paper, and Section 9 contains full proofs of our results. We end with a discussion and some open issues in Section 10.

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* (sometimes also denoted as the nuclear norm or the Ky-Fan n 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 class of rank-bounded matrices (e.g., Srebro and Shraibman, 2005). Thus, we can often optimize Equation (1) in a computationally tractable manner, learning predictors which are competitive with low-rank matrices. The trace-norm of any $m \times n$ matrix W is at least $\|W\|_F$ and at most $\mathrm{Rank}(W)\|W\|_F$, where $\|W\|_F$ is the Frobenius norm (Horn and Johnson, 1985), and therefore the trace-norm of constant-rank $m \times n$ matrices with bounded entries is $\Theta(\sqrt{mn})$. Therefore, we wish to attain learning guarantees which are non-trivial when the trace norm is at least on the order of $t = \Theta(\sqrt{mn})$. However, our theorems will hold for any t.

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 Equa-

tion (1), if we can provide a non-trivial uniform sample complexity bound, namely a bound on

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

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, $\mathcal{W}=\{W:\|W\|_{tr}\leq t\}$. Although learning with respect to this class is widely used in matrix completion, understanding its generalization and sample-complexity properties has proven quite elusive. Sample complexity bounds of the form $\mathcal{O}(\sqrt{(m+n)/|S|})$ (when $t=\Theta(\sqrt{mn})$, and ignoring logarithmic factors) were obtained under the strong assumption of a uniform distribution over the matrix entries (Srebro and Shraibman, 2005). However, this assumption does not correspond to real-world matrix completion data sets, where the distribution of the revealed entries appears to be highly non-uniform. Other works, which focused on exact matrix completion (e.g., Candès and Tao, 2009; Candès and Recht, 2009), also assume a uniform sampling distribution.

The bounds in Srebro and Shraibman (2005) are based on the Rademacher complexity of the class $\mathcal{W}$, and will be utilized in our analysis as well. Formally, we define the (empirical) 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 Equation (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 $\delta\in(0,1)$, Equation (2) is bounded with probability at least $1-\delta$ by $2R_S(\ell\circ\mathcal{W})+b_\ell\sqrt{2\log(2/\delta)/|S|}$.*

In general, the dominant term in the bound above is the Rademacher complexity $R_S(\ell\circ\mathcal{W})$. Thus, if we can upper-bound the Rademacher complexity by a quantity much smaller than 1, we get a non-trivial upper bound on Equation (2). Such a bound implies that the empirical risk (or average loss over the training set) is close to the true risk uniformly for all $W\in\mathcal{W}$, and therefore that solving Equation (1) will lead to a predictor with near-optimal risk.

Unfortunately, for the class $\mathcal{W}=\{W:\|W\|_{tr}\leq t\}$ and general distributions over the matrix entries, the Rademacher complexity can be large, leading to vacuous bounds. To see why, suppose that the loss function ℓ is 1-Lipschitz in its first argument. Then the standard

way to analyze Equation (3) (see Bartlett and Mendelson, 2003) is to use the contraction principle to upper bound it by

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

and then using Hölder's inequality to upper bound it by

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\frac{1}{|S|}\|\Gamma\|_{sp}\|W\|_{tr}\right] \;=\; t\frac{1}{|S|}\mathbb{E}[\|\Gamma\|_{sp}],$$

where Γ is a matrix whose (i,j)-th entry is defined as $\sum_{\alpha:i_\alpha=i,j_\alpha=j}\sigma_\alpha$, and $\|\cdot\|_{sp}$ is the spectral norm (i.e., the largest singular value), which is well-known to be dual to the trace-norm (Fazel et al., 2001). However, if for instance all σ_α are on the same entry i,j, then $\mathbb{E}[\|\Gamma\|_{sp}]$ equals $\mathbb{E}[|\sum_\alpha\sigma_\alpha|] = \Theta(\sqrt{|S|})$, leading to a bound of the form $\mathcal{O}(t/\sqrt{|S|})$. As discussed earlier, t is typically at least on the order of $\sqrt{mn}$, in which case we get a bound on the Rademacher complexity which is $\mathcal{O}(\sqrt{mn/|S|})$ — smaller than 1 only when the sample size $|S|$ is larger than the total number mn of matrix entries. It is a trivial bound, since the entire goal of matrix completion is prediction based on observing just a small subset of the matrix entries.

Unfortunately, this bound appears impossible to improve in general (see section 6.2.2 in Srebro, 2004). Srebro and Shraibman (2005) circumvent this by imposing a strong uniform distribution assumption, under which a tighter bound is attainable. The main drive of our paper is that by modifying the setting in some very simple ways, which often correspond to matrix completion as done in practice, one can obtain non-trivial learning guarantees without any distributional assumptions.

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*

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

where C is the universal constant appearing in Theorem 8.

When $t = \Theta(\sqrt{mn})$, the theorem implies that a sample of size $\mathcal{O}(n\sqrt{m} + m\sqrt{n})$ 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 $\mathcal{O}((m+n)/\sqrt{|S|})$, which leads to a trivial required sample size of $\mathcal{O}((m+n)^2)$. 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 9.1. 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 Section 7, we return to this model in greater depth. In terms of the theoretical guarantee, one can provide a result similar to Theorem 2, without assuming boundedness of the loss function.

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*

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

where C is the universal constant appearing in Theorem 8.

The bound in this theorem scales similarly to Theorem 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., $\max_{i,j} |W_{i,j}| \leq c$ for some constant c) which forces the matrix entries to be constant. This model has some potential benefits which shall be further discussed in Section 10.

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*

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

where C is the universal constant appearing in Theorem 8.

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 matrix completion 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 data set 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 matrix completion: 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 data set 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 the literature (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 Equation (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*

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.

holds for any $W \in \mathcal{W}$ *that*

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

This theorem implies that if $R_S(\ell \circ \mathcal{W})$ is effectively bounded, then the average loss over S_{train} is close to the average loss over S_{test}, uniformly for any W, and therefore minimizing the average loss over S_{train} will result in a predictor with near-optimal average loss over S_{test}.

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 9.4.

Theorem 6 *Consider the hypothesis class* $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$*. Then in the transductive model, it holds that*

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

where C *is the universal constant appearing in Theorem 8. Alternatively, letting* $N = \max_i |\{j : (i,j) \in S\}|$ *and* $M = \max_j |\{i : (i,j) \in S\}|$*, then*

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

where C *is the universal constant appearing in Theorem 9.*

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 (for simplicity) a rectangular matrix, so that $m = n$, and that $t = \Theta(\sqrt{mn}) = \Theta(n)$. 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 $\mathcal{O}(|S|/n)$. In that case, plugging $|S|/n$ instead of M and N, as well as $t = \Theta(n)$, 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 (e.g., Srebro and Shraibman, 2005)), which relied on a uniform distribution assumption. However, our Rademacher complexity bound in Theorem 6 also applies to non-uniform distributions, and is meaningful for any distribution.

Compared to the results in Section 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 cited overall sample complexity bound (Theorem 5), we conjecture that its true effect is modest at best. This is in light of recent work, which imply that using particular online matrix completion algorithms, one can learn comparatively well in a transductive setting, without explicit boundedness assumptions (see Section 8).

Another interesting feature of Theorem 6 is that the Rademacher complexity falls off at the rate of $\mathcal{O}(1/|S|)$ rather than $\mathcal{O}(1/\sqrt{|S|})$. 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 $\mathcal{O}(1/|S|)$ sample complexity bound, because the bound in Theorem 5 contains an additional low rate term $\mathcal{O}(1/\sqrt{|S|})$. 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. Lower Bounds

The previous results showed that for $m \times n$ matrices (where $m \leq n$), $\mathcal{O}(t\sqrt{n})$ samples are sufficient for learning. In this section, we show that such a sample size is also necessary, in both the inductive and transductive settings, hence establishing the tightness of our bounds. We remark that this lower bound applies in the distribution-free case (where any distribution over the matrix entries is allowed), and hence does not contradict tighter upper-bounds, which hold under distributional assumptions, such as in Negahban and Wainwright (2010); Candès and Tao (2009); Srebro and Shraibman (2005). Also, this lower bound result is not really new, and a different version of it appears in Hazan et al. (2012) for the inductive setting. However, we reproduce it here due to its relevance, and since it resolved an open problem posed in a preliminary version of our paper (Shamir and Shalev-Shwartz, 2011). For simplicity, we will consider $n \times n$ matrices.

The lower bound is based on the following theorem:

Theorem 7 *Fix a parameter $t \in \left[n, n^{3/2}\right]$, and consider the class of $n \times n$ matrices $\mathcal{W} = \{W : \|W\|_{tr} \leq t\}$. Let $\mathcal{W}'$ be the set of all matrices whose entries are $\{-1, +1\}$ on the first $\lfloor t\sqrt{n} \rfloor$ rows, and 0 everywhere else. Then $\mathcal{W}' \subset \mathcal{W}$.*

Proof We need to show that any matrix $W \in \mathcal{W}'$ has trace-norm at most t. To see why, note that W is non-zero on only $\lfloor t/\sqrt{n} \rfloor$, hence its rank is at most $\lfloor t/\sqrt{n} \rfloor$. Letting $\|\cdot\|_F$ denote the Frobenius norm and using the inequality $\|A\|_{tr} \leq \sqrt{\mathrm{rank}(A)}\|A\|_F$, we have

$$\|W\|_{tr} \;\leq\; \sqrt{\mathrm{rank}(W)}\|W\|_F \;\leq\; \sqrt{\frac{t}{\sqrt{n}}}\sqrt{n * \frac{t}{\sqrt{n}}} \;=\; t.$$

■

We now argue, based on this theorem, that learning is impossible unless the sample size $|S|$ is at least $\Omega(t\sqrt{n})$, matching our previous upper bounds (which were smaller than 1 only when $|S| > \Omega(t\sqrt{n})$). To see why, assume w.l.o.g. that S lies in the first $\left\lfloor \frac{t}{2\sqrt{n}} \right\rfloor$ rows, and let us consider first the inductive setting. Suppose we are asked to predict the values of a

matrix X, with respect to a uniform distribution over its entries in the first $\lfloor t/\sqrt{n} \rfloor$ rows, and where the value of each of these entries was independently chosen from $\{-1,+1\}$. If we are given a sample of size $|S| \leq \lfloor t\sqrt{n}/2 \rfloor$ from this matrix, it means that most of the relevant binary entries remain unobserved. Moreover, they were chosen uniformly at random, hence we have no way to predict their value. For any reasonable loss function, this would imply an expected error which is at least constant. In contrast, by the theorem above, there exists some $W \in \mathcal{W}$ which predicts perfectly all of these entries, and its expected error would be zero. In other words, for any algorithm returning a (possibly randomized) predicted matrix W,

$$\mathbb{E}_W \left[\mathbb{E}_{i,j}[\ell(W_{i,j}, X_{i,j})] - \inf_{W \in \mathcal{W}} \mathbb{E}_{i,j}[\ell(W_{i,j}, X_{i,j})] \right] \geq c,$$

for some constant $c > 0$, and hence we are unable to learn with such a sample size. A similar result holds in the transductive setting: If S is supported on those first $\lfloor t/\sqrt{n} \rfloor$ rows, and is randomly split to S_{train} and S_{test}, we have no way to predict the entries of S_{test} given S_{test}, and would achieve constant expected error. Moreover, using standard VC dimension techniques, even for larger sample sizes $|S|$ the attainable error cannot be better than $\Omega(\sqrt{t\sqrt{n}/|S|})$.

7. Should Boundedness be Enforced?

As mentioned earlier in the paper, we often know the range of entries to be predicted (e.g., 1 to 5 for movie rating prediction). The results of Section 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 smaller role, by appearing in the final sample-complexity bound (Theorem 5), although not in the Rademacher complexity bound (Theorem 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. Our theoretical results are not sufficiently refined to understand the precise effect of boundedness, so it is of interest to understand experimentally how much clipping or enforcing boundedness helps the learning process. We note that while bounded models have been tested experimentally, we could not find in prior literature a clear empirical study of their effect, in the context of trace-norm regularization.

We conducted experiments on two standard matrix completion data sets,[2] movielens100K and movielens1M. 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 data set, 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

2. These data sets are taken from `www.grouplens.org/node/73`

trained a trace-norm regularized algorithm using the squared loss. Specifically, we used the common approach of stochastic gradient descent on a factorized representation $W = U^\top V$: First, we note that for any t, minimizing $\sum_{(i,j)\in S}(X_{i,j} - W_{i,j})^2$ over all $W : \|W\|_{tr} \leq t$ is equivalent to minimizing

$$\sum_{(i,j)\in S}(X_{i,j} - W_{i,j})^2 + \lambda\|W\|_{tr} \tag{4}$$

over all matrices W, where λ is some suitable soft-regularization parameter. Second, we use the fact that the trace norm can also be written as $\|W\|_{tr} = \min_{W=U^\top V}\frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right)$, so minimizing *Equation* (4) over W is equivalent to minimizing

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

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

$$\sum_{(i,j)\in S}\left(X_{i,j} - \phi(U_i^\top V_j)\right)^2 + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right). \tag{6}$$

We note that both problems are non-convex, although for the formulation in Equation (5), 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 data sets, such as Netflix, and that is why we focused on the more modest-sized movielens100K and movielens1M data sets.

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 data sets).

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.

	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 data set (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 Equation (5), 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 Equation (6). 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).

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 matrix completion, 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.

8. Follow-Up Work

Since the preliminary version of this paper appeared (Shamir and Shalev-Shwartz, 2011), several related works have been published. In this section, we survey these results, and discuss how they relate to the current work and the insights it provides.

While this work focuses on a stochastic setting, a closely related problem has been matrix completion in an *online* setting. In online learning (Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz, 2012), rather than having examples sampled from a stochastic process, the examples arrive in an online fashion and are arbitrary, possibly provided by an all-powerful adversary. The goal in this setting is to minimize *regret*, namely the difference between the learner's loss and that of the best single hypothesis from some hypothesis class. In the context of matrix completion, this can be modeled as a sequential game where at each round a matrix entry is arbitrarily chosen, and the learner needs to predict its value. The actual entry value is then revealed, and the learner suffers some loss (such as the absolute difference between the prediction and actual value). In our case, the regret can be measured with respect to the class of matrices with bounded trace-norm. Note that this setting is generally harder than our stochastic setting, since the entry are chosen arbitrarily rather than in a stochastic manner, and it is known that in general, any online learning algorithm can be converted to a learning algorithm in a stochastic setting, with similar guarantees. Despite the difference between the settings, regret guarantees in the online learning setting are often strikingly similar to sample complexity guarantees in the stochastic learning setting.

The problem of online matrix completion with trace-norm bounded matrix has been dealt with in several recent works. Interestingly, the same insights provided in our work — the importance of entry boundedness or a transductive model — were crucial for attaining online learning algorithms. Considering $n \times m$ matrices with bounded trace-norm t as well as bounded entries, Hazan et al. (2012) showed that one can efficiently obtain vanishing regret after $\mathcal{O}(t\sqrt{n})$ rounds (assuming $m \leq n$). Note that this parallels our sample complexity guarantees in a stochastic setting (assuming bounded entries), which imply learnability for sample size $\mathcal{O}(t\sqrt{n})$. Alternatively, if one considers a transductive online setting (where each entry can be chosen only once), Cesa-Bianchi and Shamir (2011) showed that one can also efficiently obtain vanishing regret after $\mathcal{O}(t\sqrt{n})$ rounds. In Rakhlin et al. (2012) this was shown to be possible for Lipschitz-continuous losses, even if the entries are not explicitly bounded — the transductive setting alone suffices to achieve results of this order.

Another recent related work is Shalev-Shwartz et al. (2011), which deals with a supposedly different problem: Approximately solving convex optimization problems over the (non-convex) domain of low-rank matrices. However, one of their results provides an alternative justification of our $\mathcal{O}(t\sqrt{n})$ sample complexity guarantee, for the case of bounded trace-norm matrices whose entries are clipped to a bounded range (Theorem 3). To sketch the argument, Shalev-Shwartz et al. (2011, Section 4) show that if we have $|S|$ observed entries in our matrix, then for every matrix W with bounded trace-norm $\|W\|_{tr}$, there exists a low-rank matrix $\bar{W}$, with rank $\mathcal{O}(\|W\|_{tr}^2/|S|)$, which approximates W arbitrarily well in terms of average loss over the observed entries. Since a matrix of rank r is parameterized by $\mathcal{O}(rn)$ parameters, it follows that the generalization error of clipped r-rank matrices is arbitrarily small when $|S| \geq \tilde{\Omega}(rn)$. This indirectly provides a generalization error bound for our original matrix W. Plugging in $r = \|W\|_{tr}^2/|S|$, and noting that in our case $\|W\|_{tr} = n$, we get that learnability is possible for a sample of size $|S| \geq \tilde{\Omega}(n^{3/2}) = \tilde{\Omega}(t\sqrt{n})$.

Finally, we note that several recent works explored the possibility of replacing the standard trace-norm constraint by other matrix norms. These include the max-norm (Srebro et al., 2004; Lee et al., 2010); weighted trace-norm (Salakhutdinov and Srebro, 2010) and smoothed/empirical variants (Foygel et al., 2011); and 'local' max-norms (Foygel et al.,

2012). An important motivation of these works is that they allow us to learn non-trivial classes of matrices, with a sample complexity of $\mathcal{O}(n)$ — smaller than earlier trivial results for the trace-norm and the $\mathcal{O}(n^{3/2})$ results we obtain here (when the trace-norm is $\Theta(n)$). Essentially, this is achieved by using classes of matrices which are less rich than trace-norm-bounded one, hence are statistically easier to learn.

9. Proofs of Upper Bounds

In our proofs, we use $\|\cdot\|_{sp}$ to denote the spectral norm of matrices, which is well-known to be the dual of the trace-norm (see for instance Fazel et al., 2001).

Our proofs utilize the following two theorems, which bounds the expected spectral norm $\|\cdot\|_{sp}$ of random matrices.

Theorem 8 (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 9 (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*

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

9.1 Proof of Theorem 2

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

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

where Γ is a matrix whose (i,j)-th entry is defined as $\sum_{\alpha: i_\alpha = i, j_\alpha = j} \sigma_\alpha$. As discussed in Section 3, a standard analysis will proceed to reduce this to

$$\frac{1}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}} \sum_{i,j} \Gamma_{i,j} W_{i,j}\right], \tag{8}$$

but this leads to a trivial bound. However, examining the analysis in Section 3, we see that the problem is when a single entry is "hit" many times in the sample. This will cause the magnitude of that entry to be very large (as much as $\Theta(\sqrt{|S|})$), and as a result make Equation (8) as large as $\Theta(t/\sqrt{|S|})$. However, recall that our original goal is to bound Equation (7), not Equation (8), and in Equation (7) we have the loss operator, which is

bounded by a constant b_ℓ. Therefore, even if some $\Gamma_{i,j}$ has a large value, it can only be multiplied by a factor as large as b_ℓ, and *not* the trace-norm bound t. This observation is the key for our analysis.

Intuitively, instead of going directly from Equation (7) to Equation (8), we first decompose Γ into two matrices Y and Z, where Y contains the "heavily-hit" entries, and Z the "lightly-hit" entries, where the two types of entries are differentiated according to some threshold p. We perform a different type analysis for each matrix, and then tune p appropriately to get the desired result.

More formally, 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

$$Y_{i,j} = \begin{cases} \Gamma_{i,j} & h_{i,j} > p \\ 0 & h_{i,j} \leq p \end{cases} \qquad Z_{i,j} = \begin{cases} 0 & h_{i,j} > p \\ \Gamma_{i,j} & h_{i,j} \leq p. \end{cases} \tag{9}$$

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

$$\frac{1}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\sum_{i,j} Y_{i,j}\ell(W_{i,j}, X_{i,j})\right] + \frac{1}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\sum_{i,j} Z_{i,j}\ell(W_{i,j}, X_{i,j})\right]. \tag{10}$$

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

$$\frac{1}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[b_\ell \sum_{i,j} |Y_{i,j}|\right] = \frac{b_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}[\|Y\|_1]. \tag{11}$$

Using the Rademacher contraction principle,[3] the second term in Equation (10) 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

$$\frac{l_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}} \sup_{W\in\mathcal{W}} [\|Z\|_{sp}\|W\|_{tr}] = \frac{l_\ell t}{|S|}\mathbb{E}_{\boldsymbol{\sigma}} [\|Z\|_{sp}]. \tag{12}$$

Combining this with Equation (11) and substituting into Equation (10), we get an upper bound of the form

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

Using Lemma 10 and Lemma 11, which are given below, we can upper bound this by

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

3. 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).

where p is the parameter used to define Y and Z in Equation (9). Choosing $p = \frac{|S| b_\ell}{2.2 C l_\ell t(\sqrt{m}+\sqrt{n})}$, we get the bound in the theorem.

Lemma 10 *Let Y be a random matrix defined as in Equation (9). Then*

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

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

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

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 11 *Let Z be a random matrix defined as in Equation (9). Then the expected spectral norm $\mathbb{E}_{\boldsymbol{\sigma}}[\|Z\|_{sp}]$ 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 Theorem 8, 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 Theorem 8 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

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

■

9.2 Proof of Theorem 3

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

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

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

$$\frac{l_\ell}{|S|} E_{\boldsymbol{\sigma}} \left[\sup_{W \in \mathcal{W}} \sum_{i,j} \Gamma_{i,j} W_{i,j} \right]. \tag{13}$$

Decomposing $\Gamma = Y + Z$ as in Equation (9) according to a parameter p, we can upper bound the Rademacher complexity by

$$\frac{l_\ell}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{W \in \mathcal{W}} \sum_{i,j} Y_{i,j} W_{i,j} \right] + \frac{l_\ell}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{W \in \mathcal{W}} \sum_{i,j} Z_{i,j} W_{i,j} \right]. \tag{14}$$

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

$$\frac{l_\ell}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[b_\phi \sum_{i,j} |Y_{i,j}| \right] = \frac{l_\ell b_\phi}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} [\|Y\|_1]. \tag{15}$$

The second term in Equation (14) 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 Equation (15) and substituting into Equation (14), 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 10 and Lemma 11, we can upper bound this by

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

where p is the parameter used to define Y and Z in Equation (9). Choosing $p = \frac{|S| b_\phi}{2.2 C l_\phi t (\sqrt{m} + \sqrt{n})}$, we get the bound in the theorem.

9.3 Proof of Theorem 4

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

Lemma 12 *The dual of the norm* $\|W\| = \max\{\|W\|_{tr}/t, \|W\|_\infty/b\}$ *equals*

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

where $\|Y\|_1 = \sum_{i,j} |Y_{i,j}|$ *and* $\|Z\|_{sp}$ *is the spectral norm of* Z*.*

It is possible to prove the lemma directly using duality of infimal convolution. However, for the sake of completeness we give below a self-contained proof.

Proof By definition of a dual norm, we have

$$\|\Gamma\|_* = \sup_{W:\|W\|\le 1} \langle \Gamma, W\rangle,$$

and our goal is to show that

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

First, we recall that the dual norm of $\|W\|_{tr}$ is the spectral norm $\|W\|_{sp}$, 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 $Y + Z = \Gamma$, we have by Hölder's inequality that

$$\begin{aligned}
\sup_{W:\|W\|\le 1} \langle \Gamma, W\rangle &= \sup_{W:\|W\|\le 1} \langle Y, W\rangle + \langle Z, W\rangle \\
&\le \sup_{W:\|W\|\le 1} \|Y\|_1\|W\|_\infty + \|Z\|_{sp}\|W\|_{tr} \\
&\le b\|Y\|_1 + t\|Z\|_{sp}.
\end{aligned}$$

This holds for any Y, Z, and in particular

$$\sup_{W:\|W\|\le 1} \langle \Gamma, W\rangle \le \min_{Y+Z=\Gamma} b\|Y\|_1 + t\|Z\|_{sp}. \tag{16}$$

It remains to show the opposite direction, namely

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

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

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

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

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

So in either case,

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

Combining this with Equation (16), the result follows. ■

We now turn to the proof of Theorem 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

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

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

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

$$\frac{l_\ell}{|S|} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{W \in \mathcal{W}} \|\Gamma\|_* \|W\| \right] \tag{17}$$

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, $\sup_{W \in \mathcal{W}} \|W\| \leq 1$. Also, by Lemma 12,

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

where $\|Y\|_1 = \sum_{i,j} |Y_{i,j}|$, and $\|Z\|_{sp}$ is the spectral norm of Z . Thus, we can upper bound Equation (17) by

$$\frac{l_\ell}{|S|} \mathbb{E}_\Gamma \left[\min_{Y+Z=\Gamma} b\|Y\|_1 + t\|Z\|_{sp} \right]. \tag{18}$$

Recall that Γ 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 Equation (18) by replacing the optimal decomposition of Γ into Y, Z by any fixed decomposition rule. In particular, for an arbitrary parameter p, we can decompose Γ into Y, Z as in Equation (9), and get an upper bound on Equation (18) of the form

$$\frac{l_\ell}{|S|} \left(b\mathbb{E}_\Gamma[\|Y\|_1] + t\mathbb{E}_\Gamma[\|Z\|_{sp}] \right). \tag{19}$$

Bounds for the two expectations are provided in Lemma 10 and Lemma 11. Plugging them in, we get

$$\frac{bl_\ell}{\sqrt{p}} + \frac{2.2 l_\ell C t \sqrt{p} \left(\sqrt{m} + \sqrt{n}\right)}{|S|}.$$

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

9.4 Proof of Theorem 6

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

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

where Γ 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,[4] we can upper bound this by

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

By Hölder's inequality, this is at most

$$\frac{l_\ell}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{W\in\mathcal{W}}\|\Gamma\|_{sp}\|W\|_{tr}\right]=\frac{l_\ell t}{|S|}\mathbb{E}_{\boldsymbol{\sigma}}\left[\|\Gamma\|_{sp}\right]. \tag{20}$$

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 Section 4). But now, we need to bound just the expected spectral norm of Γ, which is guaranteed to have only a single Rademacher variable in each entry. By applying Theorem 8 on Equation (20), we get

$$R_S(\ell\circ\mathcal{W})\le Cl_\ell\frac{t\left(\sqrt{M}+\sqrt{N}+\sqrt[4]{|S|}\right)}{|S|}.$$

Since S can contain at most m and n indices for any single row and column respectively, and $\sqrt[4]{|S|}\le\sqrt[4]{mn}\le\frac{1}{2}\left(\sqrt{m}+\sqrt{n}\right)$, 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 Theorem 9 instead of Theorem 8 on Equation (20).

10. Discussion

In this paper, we analyzed the sample complexity of matrix completion 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 matrix completion as performed in practice, as it seems more natural and leads to a substantial difference in terms of obtainable results. We also discussed the issue of learning with bounded models, and provided an empirical study which indicates that these lead to a modest improvement in performance, in line with our theoretical findings. We also show that our results are essentially tight, and discuss some recent work which relates to the results and insights provided here.

One interesting open question arises from our experiments in Section 7. In all our experiments, minimizing the squared loss over the training data (with trace-norm regularization)

4. 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).

resulted in matrices whose entries have reasonably small values, even when boundedness was not enforced. This is probably an important factor in explaining why explicitly enforcing boundedness resulted in only a modest performance improvement. However, if boundedness is not enforced, there is no a-priori reason why the resulting matrix shouldn't have some very large values (up to the trace-norm constraint) in some of the test set entries. Thus, we may raise the following conjecture: If we minimize training loss over data, consisting of bounded entries, over trace-norm constrained matrices, then the resulting matrix will have bounded entries as well. If this conjecture holds, it means that enforcing boundedness will always lead to only a modest performance improvement.

Acknowledgements

We thank Nati Srebro and Ruslan Salakhutdinov for helpful discussions.

References

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

P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

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.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi and O. Shamir. Efficient online learning via randomized rounding. In *NIPS*, 2011.

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

M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE, 2001.

R. Foygel, R. Salakhutdinov, O. Shamir, and N. Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. In *NIPS*, 2011.

R. Foygel, N. Srebro, and R. Salakhutdinov. Matrix reconstruction with the local max norm. In *NIPS*, 2012.

E. Hazan, S. Kale, and S. Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. In *COLT*, 2012.

R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

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.

A. Rakhlin, O. Shamir, and K. Sridharan. Relax and randomize : From value to algorithms. In *NIPS*, 2012.

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.

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

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

S. Shalev-Shwartz, A. Gonen, and O. Shamir. Large-scale convex minimization with a low-rank constraint. In *ICML*, 2011.

O. Shamir and S. Shalev-Shwartz. Collaborative filtering with the trace norm: Learning, bounding, and transducing. In *COLT*, 2011.

N. Srebro. *Learning with Matrix Factorizations*. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, August 2004.

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.

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.