Learning with the Weighted Trace-norm under Arbitrary Sampling Distributions

Rina Foygel
Department of Statistics
University of Chicago
rina@uchicago.edu

Ruslan Salakhutdinov
Department of Statistics
University of Toronto
rsalakhu@ustat.toronto.edu

Ohad Shamir
Microsoft Research New England
ohadsh@microsoft.com

Nathan Srebro
Toyota Technological Institute at Chicago
nati@ttic.edu

Abstract

We provide rigorous guarantees on learning with the weighted trace-norm under arbitrary sampling distributions. We show that the standard weighted-trace norm might fail when the sampling distribution is not a product distribution (i.e. when row and column indexes are not selected independently), present a corrected variant for which we establish strong learning guarantees, and demonstrate that it works better in practice. We provide guarantees when weighting by either the true or empirical sampling distribution, and suggest that even if the true distribution is known (or is uniform), weighting by the empirical distribution may be beneficial.

1 Introduction

One of the most common approaches to collaborative filtering and matrix completion is trace-norm regularization [1, 2, 3, 4, 5]. In this approach we attempt to complete an unknown matrix, based on a small subset of revealed entries, by finding a matrix with small trace-norm, which matches those entries as best as possible.

This approach has repeatedly shown good performance in practice, and is theoretically well understood for the case where revealed entries are sampled uniformly [6, 7, 8, 9, 10, 11]. Under such uniform sampling, $\Theta(n\log(n))$ entries are sufficient for good completion of an $n \times n$ matrix—i.e. a nearly constant number of entries per row. However, for arbitrary sampling distributions, the worst-case sample complexity lies between a lower bound of $\mathbf{\Omega}(n^{4/3})$ [12] and an upper bound of $\mathbf{O}(n^{3/2})$ [13], i.e. requiring between $n^{1/3}$ and $n^{1/2}$ observations per row, and indicating it is not appropriate for matrix completion in this setting.

Motivated by these issues, Salakhutdinov and Srebro [12] proposed to use a weighted variant of the trace-norm, which takes the distribution of the entries into account, and showed experimentally that this variant indeed leads to superior performance. However, although this recent paper established that the weighted trace-norm corrects a specific situation where the standard trace-norm fails, no general learning guarantees are provided, and it is not clear if indeed the weighted trace-norm always leads to the desired behavior. The only theoretical analysis of the weighted trace-norm that we are aware of is a recent report by Negahban and Wainwright [10] that provides reconstruction guarantees for a low-rank matrix with i.i.d. noise, but only when the sampling distribution is a *product distribution*, i.e. the rows index and column index of observed entries are selected independently. A product distribution assumption does not seem realistic in many cases—e.g. for the Netflix data, it would indicate that all users have the same (conditional) distribution over which movies they rate.

In this paper we rigorously study learning with a weighted trace-norm under an *arbitrary* sampling distribution, and show that this situation is indeed more complicated, requiring a correction to the weighting. We show that this correction is necessary, and present empirical results on the Netflix and MovieLens dataset indicating that it is also helpful in practice. We also rigorously consider weighting according to either the true sampling distribution (as in [10]) or the empirical frequencies, as is actually done in practice, and present evidence that weighting by the empirical frequencies might be advantageous. Our setting is also more general than that of [10]—we consider an arbitrary loss and do not rely on i.i.d. noise, instead presenting results in an agnostic learning framework.

Setup and Notation. We consider an arbitrary unknown $n \times m$ target matrix Y, where a subset of entries $\{Y_{i_t,j_t}\}_{t=1}^{s}$ indexed by $S = \{(i_1, j_1), \ldots, (i_s, j_s)\}$ is revealed to us. Without loss of generality, we assume $n \geq m$. Throughout most of the paper, we assume S is drawn i.i.d. according to some sampling distribution $p(i, j)$ (with replacement). Based on this subset on entries, we would like to fill in the missing entries and obtain a prediction matrix $\hat{X}_S \in \mathbb{R}^{n\times m}$, with low expected loss $L_p(\hat{X}_S) = \mathbf{E}_{ij\sim p}\left[\ell((\hat{X}_S)_{ij}, Y_{ij})\right]$, where $\ell(x, y)$ is some loss function. Note that we measure the loss with respect to the same distribution $p(i, j)$ from which the training set is drawn (this is also the case in [12, 10, 13]).

The trace-norm of a matrix $X \in \mathbb{R}^{n\times m}$, written $\|X\|_{\mathrm{tr}}$, is defined as the sum of its singular values. Given some distribution $p(i, j)$ on $[n] \times [m]$, the weighted trace-norm of X is given by [12]

$$\|X\|_{\mathrm{tr}(p^r,p^c)} = \left\|\mathrm{diag}\,(p^r)^{1/2} \cdot X \cdot \mathrm{diag}\,(p^c)^{1/2}\right\|_{\mathrm{tr}},$$

where $p^r \in \mathbb{R}^n$ and $p^c \in \mathbb{R}^m$ denote vectors of the row- and column-marginals respectively. Note that the weighted trace-norm only depends on these marginals (but not their joint distribution) and that if p^r and p^c are uniform, then $\|X\|_{\mathrm{tr}(p^r,p^c)} = \frac{1}{\sqrt{nm}}\|X\|_{\mathrm{tr}}$. The weighted trace-norm does not generally scale with n and m, and in particular, if X has rank r and entries bounded in $[-1, 1]$, then $\|X\|_{\mathrm{tr}(p^r,p^c)} \leq \sqrt{r}$ regardless of which $p\,(i, j)$ is used. This motivates us to define the class

$$\mathcal{W}_r\,[p] = \left\{X \in \mathbb{R}^{n\times m} : \|X\|_{\mathrm{tr}(p^r,p^c)} \leq \sqrt{r}\right\},$$

although we emphasize that our results do not directly depend on the rank, and $\mathcal{W}_r\,[p]$ certainly includes full-rank matrices. We analyze here estimators of the form $\hat{X}_S = \arg\min\{\hat{L}_S(X) \,:\, X \in \mathcal{W}_r\,[p]\}$ where $\hat{L}_S(X) = \frac{1}{s}\sum_{t=1}^{s}\ell(X_{i_t,j_t}, Y_{i_t,j_t})$ is the empirical error on the observed entries.

Although we focus mostly on the standard inductive setting, where the samples are drawn i.i.d. and the guarantee is on generalization for future samples drawn by the same distribution, our results can also be stated in a transductive model, where a training set and a test set are created by splitting a fixed subset of entries uniformly at random (as in [13]). The transductive setting is discussed, and transductive variants of our Theorems are given, in Section 4.2 and in the Supplementary Materials.

2 Learning with the Standard Weighting

In this Section, we consider learning using the weighted trace-norm as suggested by Salakhutdinov and Srebro [12], i.e. when the weighting is according to the sampling distribution $p(i, j)$. Following the approach of [6] and [11], we base our results on bounding the Rademacher complexity of $\mathcal{W}_r\,[p]$, as a class of functions mapping index pairs to entry values. However, we modify the analysis for the weighted trace-norm with non-uniform sampling.

For a class of matrices $\mathcal{X}$ and a sample $S = \{(i_1, j_1), \ldots, (i_s, j_s)\}$ of indexes in $[n] \times [m]$, the empirical Rademacher complexity of the class (with respect to S) is given by

$$\hat{\mathcal{R}}_S(\mathcal{X}) = \mathbf{E}_{\sigma\sim\{\pm1\}^s}\left[\sup_{X\in\mathcal{X}}\frac{1}{s}\sum_{t=1}^{s}\sigma_t X_{i_t j_t}\right],$$

where σ is a vector of signs drawn uniformly at random. Intuitively, $\hat{\mathcal{R}}_S(\mathcal{X})$ measures the extent to which the class $\mathcal{X}$ can "overfit" data, by finding a matrix X which correlates as strongly as possible to a sample from a matrix of random noise. For a loss $\ell(x, y)$ that is Lipschitz in x, the Rademacher complexity can be used to uniformly bound the deviations $|L_p(X) - \hat{L}_S(X)|$ for all $X \in \mathcal{X}$, yielding a learning guarantee on the empirical risk minimizer [14].

2.1 Guarantees for Special Sampling Distributions

We begin by providing guarantees for an arbitrary, possibly unbounded, Lipschitz loss $\ell(x,y)$, but only under sampling distributions which are *either* product distributions (i.e. $p(i,j) = p^r(i)p^c(j)$) *or* have uniform marginals (i.e. p^r and p^c are uniform, but perhaps the rows and columns are not independent). In Section 2.3 below, we will see why this severe restriction on p is needed.

Theorem 1. *For an l-Lipschitz loss ℓ, fix any matrix Y, sample size s, and distribution p, such that p is either a product distribution or has uniform marginals. Let $\hat{X}_S = \arg\min\left\{\hat{L}_S(X) \,:\, X \in \mathcal{W}_r[p]\right\}$. Then, in expectation over the training sample S drawn i.i.d. from the distribution p,*

$$L_p(\hat{X}_S) \leq \inf_{X \in \mathcal{W}_r[p]} L_p(X) + \mathbf{O}\left(l \cdot \sqrt{\frac{rn\log(n)}{s}}\right) . \tag{1}$$

Here and elsewhere we state learning guarantees in expectation for simplicity. Since the guarantees are obtained by bounding the Rademacher complexity, one can also immediately obtain high-probability guarantees, with logarithmic dependence on the confidence parameter, via standard techniques (e.g. [14]).

Proof. We will show how to bound the expected Rademacher complexity $\mathbf{E}_S\left[\hat{\mathcal{R}}_S(\mathcal{W}_r[p])\right]$, from which the desired results follows using standard arguments (Theorem 8 of [14][1]). Following [11] by including the weights, using the duality between spectral norm $\|\cdot\|_{\mathrm{sp}}$ and trace-norm, we compute:

$$\mathbf{E}_S\left[\hat{\mathcal{R}}_S(\mathcal{W}_r[p])\right] = \frac{\sqrt{r}}{s}\mathbf{E}_{S,\sigma}\left[\left\|\sum_{t=1}^{s}\sigma_t \frac{e_{i_t,j_t}}{\sqrt{p^r(i_t)\,p^c(j_t)}}\right\|_{\mathrm{sp}}\right] = \frac{\sqrt{r}}{s}\mathbf{E}_{S,\sigma}\left[\left\|\sum_{t=1}^{s} Q_t\right\|_{\mathrm{sp}}\right] ,$$

where $e_{i,j} = e_i e_j^T$ and $Q_t = \sigma_t \frac{e_{i_t,j_t}}{\sqrt{p^r(i_t)p^c(j_t)}} \in \mathbb{R}^{n\times m}$. Since the Q_t's are i.i.d. zero-mean matrices, Theorem 6.1 of [15], combined with Remarks 6.4 and 6.5 there, establishes that $\mathbf{E}_{S,\sigma}\left[\|\sum_{t=1}^{s} Q_t\|_{\mathrm{sp}}\right] = \mathbf{O}\big(\rho\sqrt{\log(n)} + R\log(n)\big)$, where R and ρ are defined to satisfy $\|Q_t\|_{\mathrm{sp}} \leq R$ (almost surely) and $\rho^2 = \max\left\{\left\|\sum \mathbf{E}\left[Q_t^T Q_t\right]\right\|_{\mathrm{sp}}, \left\|\sum \mathbf{E}\left[Q_t Q_t^T\right]\right\|_{\mathrm{sp}}\right\}$. Calculating these bounds (see Supplementary Material), we get $R \leq \sqrt{\frac{nm}{\min_{i,j}\{np^r(i)\cdot mp^c(j)\}}}$, and

$$\rho \leq \sqrt{s\max\left\{\max_i \sum_j \frac{p(i,j)}{p^r(i)\,p^c(j)}, \max_j \sum_i \frac{p(i,j)}{p^r(i)\,p^c(j)}\right\}} \leq \sqrt{\frac{sn}{\min_{i,j}\{np^r(i)\cdot mp^c(j)\}}} .$$

If p has uniform row- and column-marginals, then for all i,j, $np^r(i) = mp^c(j) = 1$. This yields $\mathbf{E}_S\left[\hat{\mathcal{R}}(\mathcal{W}_r[p])\right] \leq \mathbf{O}\left(\sqrt{\frac{rn\log(n)}{s}}\right)$, as desired. (Here we assume $s > n\log(n)$, since otherwise we need only establish that excess error is $\mathbf{O}(l\sqrt{r})$, which holds trivially for any matrix in $\mathcal{W}_r[p]$.)

If p does not have uniform marginals, but instead is a product distribution, then the quantity R defined above is potentially unbounded, so we cannot apply the same simple argument. However, we can consider the "p-truncated" class of matrices

$$\mathcal{Z} = \left\{ Z(X) = \left(X_{ij}\mathbb{I}\left\{p(i,j) \geq \frac{\log(n)}{s\sqrt{nm}}\right\}\right)_{ij} : X \in \mathcal{W}_r[p] \right\} .$$

By a similar calculation of the expected spectral norms, we can now bound $\mathbf{E}_S\left[\hat{\mathcal{R}}_S(\mathcal{Z})\right] \leq \mathbf{O}\left(\sqrt{\frac{rn\log(n)}{s}}\right)$. Applying Theorem 8 of [14], this bounds $\left(L_p(Z(\hat{X}_S)) - \hat{L}_S(Z(\hat{X}_S))\right)$ (in expectation). Since $Z(\hat{X}_S)_{ij} \neq (\hat{X}_S)_{ij}$ only on the extremely low-probability entries, we can

[1] Theorem 8 of [14] gives a learning guarantee holding with high probability, but their proof of this theorem (in particular, the last series of displayed equations) contains a guarantee in expectation, which we use here.

also bound $\left(L_p(\hat{X}_S) - L_p(Z(\hat{X}_S))\right)$ and $\left(\hat{L}_S(Z(\hat{X}_S)) - \hat{L}_S(\hat{X}_S)\right)$. Combining these steps, we can bound $\left(L_p(\hat{X}_S) - \hat{L}_S(\hat{X}_S)\right)$. We similarly bound $\hat{L}_S(X^*) - L_p(X^*)$, where $X^* = \arg\min_{X \in \mathcal{W}_r[p]} L_p(X)$. Since $\hat{L}_S(\hat{X}_S) \leq \hat{L}_S(X^*)$, this yields the desired bound on excess error. The details are given in the Supplementary Materials. □

Examining the proof of Theorem 1, we see that we can generalize the result by including distributions p with row- and column-marginals that are lower-bounded. More precisely, if p satisfies $p^r(i) \geq \frac{1}{Cn}$, $p^c(j) \geq \frac{1}{Cm}$ for all i, j, then the bound (1) holds, up to a factor of C. Note that this result does not require an upper bound on the row- and column-marginals, only a lower bound, i.e. it only requires that no marginals are too low. This is important to note since the examples where the unweighted trace-norm fails under a non-uniform distribution are situations where some marginals are very *high* (but none are too low) [12]. This suggests that the low-probability marginals could perhaps be "smoothed" to satisfy a lower bound, without removing the advantages of the weighted trace-norm. We will exploit this in Section 3 to give a guarantee that holds more generally for arbitrary p, when smoothing is applied.

2.2 Guarantees for bounded loss

In Theorem 1, we showed a strong bound on excess error, but only for a restricted class of distributions p. We now show that if the loss function ℓ is bounded, then we can give a non-trivial, but weaker, learning guarantee that holds uniformly over all distributions p. Since we are in any case discussing Lipschitz loss functions, requiring that the loss function be bounded essentially amounts to requiring that the entries of the matrices involved be bounded. That is, we can view this as a guarantee on learning matrices with bounded entries. In Section 2.3 below, we will show that this boundedness assumption is unavoidable if we want to give a guarantee that holds for arbitrary p.

Theorem 2. *For an l-Lipschitz loss ℓ bounded by b, fix any matrix Y, sample size s, and any distribution p. Let $\hat{X}_S = \arg\min\left\{\hat{L}_S(X) \,:\, X \in \mathcal{W}_r[p]\right\}$ for $r \geq 1$. Then, in expectation over the training sample S drawn i.i.d. from the distribution p,*

$$L_p(\hat{X}_S) \leq \inf_{X \in \mathcal{W}_r[p]} L_p(X) + \mathbf{O}\left((l+b) \cdot \sqrt[3]{\frac{rn\log(n)}{s}}\right) . \tag{2}$$

The proof is provided in the Supplementary Materials, and is again based on analyzing the expected Rademacher complexity, $\mathbf{E}_S\left[\hat{\mathcal{R}}(\ell \circ \mathcal{W}_r[p])\right] \leq \mathbf{O}\left((l+b) \cdot \sqrt[3]{\frac{rn\log(n)}{s}}\right)$.

2.3 Problems with the standard weighting

In the previous Sections, we showed that for distributions p that are either product distributions or have uniform marginals, we can prove a square-root bound on excess error, as shown in (1). For arbitrary p, the only learning guarantee we obtain is a cube-root bound given in (2), for the special case of bounded loss. We would like to know whether the square-root bound might hold uniformly over all distributions p, and if not, whether the cube-root bound is the strongest result that we can give for the bounded-loss setting, and whether any bound will hold uniformly over all p in the unbounded-loss setting.

The examples below demonstrate that we cannot improve the results of Theorems 1 and 2 (up to log factors), by constructing degenerate examples using non-product distributions p with non-uniform marginals. Specifically, in Example 1, we show that in the special case of bounded loss, the cube-root bound in (2) is the best possible bound (up to the log factor) that will hold for all p, by giving a construction for arbitrary $n = m$ and arbitrary $s \leq nm$, such that with 1-bounded loss, excess error is $\mathbf{\Omega}\left(\sqrt[3]{\frac{n}{s}}\right)$. In Example 2, we show that with unbounded (Lipschitz) loss, we cannot bound excess error better than a constant bound, by giving a construction for arbitrary $n = m$ and arbitrary $s \leq nm$ in the unbounded-loss regime, where excess error is $\mathbf{\Omega}(1)$. For both examples we fix $r = 1$. We note that both examples can be modified to fit the transductive setting, demonstrating that smoothing is necessary in the transductive setting as well.

Example 1. Let $\ell(x,y) = \min\{1, |x-y|\} \leq 1$, let $a = (2s/n)^{2/3} < n$, and let matrix Y and block-wise constant distribution p be given by

$$Y = \begin{pmatrix} A & \mathbf{0}_{a\times\frac{n}{2}} \\ \mathbf{0}_{(n-a)\times\frac{n}{2}} & \mathbf{0}_{(n-a)\times\frac{n}{2}} \end{pmatrix}, \quad (p(i,j)) = \begin{pmatrix} \frac{1}{2s}\cdot\mathbf{1}_{a\times\frac{n}{2}} & \mathbf{0}_{a\times\frac{n}{2}} \\ \mathbf{0}_{(n-a)\times\frac{n}{2}} & \frac{1-\frac{an}{4s}}{(n-a)\frac{n}{2}}\cdot\mathbf{1}_{(n-a)\times\frac{n}{2}} \end{pmatrix},$$

where $A \in \{\pm 1\}^{a\times\frac{n}{2}}$ is any sign matrix. Clearly, $\|Y\|_{\mathrm{tr}(p^r,p^c)} \leq 1$, and so $\inf_{X\in\mathcal{W}_r[p]} L_p(X) = 0$. Now suppose we draw a sample S of size s from the matrix Y, according to the distribution p. We will show an ERM $\hat{Y}$ such that in expectation over S, $L_p(\hat{Y}) \geq \frac{1}{8}\sqrt[3]{\frac{n}{s}}$.

Consider Y^S where $Y^S_{ij} = Y_{ij}\mathbb{I}\{ij \in S\}$, and note that $\left\|Y^S\right\|_{\mathrm{tr}(p^r,p^c)} \leq 1$. Since $\hat{L}_S(Y^S) = 0$, it is clearly an ERM. We also have $L_p(Y^S) = \frac{N}{2s}$, where N is the number of ± 1's in Y which are not observed in the sample. Since $\mathbf{E}[N] \geq \frac{an}{4}$, we see that $\mathbf{E}\left[L_p(Y^S)\right] \geq \frac{1}{2s}\cdot\frac{an}{4} \geq \frac{1}{8}\sqrt[3]{\frac{n}{s}}$.

Example 2. Let $\ell(x,y) = |x-y|$. Let $Y = \mathbf{0}_{n\times n}$; trivially, $Y \in \mathcal{W}_r[p]$. Let $p(1,1) = \frac{1}{s}$, and $p(i,1) = p(1,j) = 0$ for all $i,j > 1$, yielding $p^r(1) = p^c(1) = \frac{1}{s}$. (The other entries of p may be defined arbitrarily.) We will show an ERM $\hat{Y}$ such that, in expectation over S, $L_p(\hat{Y}) \geq 0.25$. Let A be the matrix with $X_{11} = s$ and zeros elsewhere, and note that $\|A\|_{\mathrm{tr}(p^r,p^c)} = 1$. With probability ≥ 0.25, entry $(1,1)$ will not appear in S, in which case $\hat{Y} = A$ is an ERM, with $L_p(\hat{Y}) = 1$.

The following table summarizes the learning guarantees that can be established for the (standard) weighted trace-norm. As we saw, these guarantees are tight up to log-factors.

	1-Lipschitz, 1-bounded loss	1-Lipschitz, unbounded loss
p = product	$\sqrt{\frac{rn\log(n)}{s}}$	$\sqrt{\frac{rn\log(n)}{s}}$
p^r, p^c = uniform	$\sqrt{\frac{rn\log(n)}{s}}$	$\sqrt{\frac{rn\log(n)}{s}}$
p arbitrary	$\sqrt[3]{\frac{rn\log(n)}{s}}$	1

3 Smoothing the weighted trace norm

Considering Theorem 1 and the degenerate examples in Section 2.3, it seems that in order to be able to generalize for non-product distributions, we need to enforce some sort of uniformity on the weights. The Rademacher complexity computations in the proof of Theorem 1 show that the problem lies not with large entries in the vectors p^r and p^c (i.e. if p^r and/or p^c are "spiky"), but with the small entries in these vectors. This suggests the possibility of "smoothing" any overly low row- or column-marginals, in order to improve learning guarantees.

In Section 3.1, we present such a smoothing, and provide guarantees for learning with a smoothed weighted trace-norm. The result suggests that there is no strong negative consequence to smoothing, but there might be a large advantage, if confronted with situations as in Examples 1 and 2. In Section 3.2 we check the smoothing correction to the weighted trace-norm on real data, and observe that indeed it can also be beneficial in practice.

3.1 Learning guarantee for arbitrary distributions

Fix a distribution p and a constant $\alpha \in (0,1)$, and let $\tilde{p}$ denote the smoothed marginals:

$$\tilde{p}^r(i) = \alpha\cdot p^r(i) + (1-\alpha)\cdot\tfrac{1}{n}, \quad \tilde{p}^c(j) = \alpha\cdot p^c(j) + (1-\alpha)\cdot\tfrac{1}{m}\,. \tag{3}$$

In the theoretical results below, we use $\alpha = \frac{1}{2}$, but up to a constant factor, the same results hold for any fixed choice of $\alpha \in (0,1)$.

Theorem 3. *For an l-Lipschitz loss ℓ, fix any matrix Y, sample size s, and any distribution p. Let $\hat{X}_S = \arg\min\left\{\hat{L}_S(X) \,:\, X \in \mathcal{W}_r[\tilde{p}]\right\}$. Then, in expectation over the training sample S drawn i.i.d. from the distribution p,*

$$L_p(\hat{X}_S) \leq \inf_{X\in\mathcal{W}_r[\tilde{p}]} L_p(X) + \mathbf{O}\left(l\cdot\sqrt{\frac{rn\log(n)}{s}}\right)\,. \tag{4}$$

Proof. We bound $\mathbf{E}_{S\sim p}\left[\hat{\mathcal{R}}_S(\mathcal{W}_r\left[\tilde{p}\right])\right] \leq \mathbf{O}\left(\sqrt{\frac{rn\log(n)}{s}}\right)$, and then apply Theorem 8 of [14]. The proof of this Rademacher bound is essentially identical to the proof in Theorem 1, with the modified definition of $Q_t = \sigma_t \frac{e_{i_t,j_t}}{\sqrt{\tilde{p}^r(i)\tilde{p}^c(j)}}$. Then $\|Q_t\|_{\mathrm{sp}} \leq \max_{ij} \frac{1}{\sqrt{\tilde{p}^r(i)\tilde{p}^c(j)}} \leq 2\sqrt{nm} \doteq R$, and $\mathbf{E}\left[\left\|\sum_{t=1}^s Q_t Q_t^T\right\|_{\mathrm{sp}}\right] = s\cdot\max_i \sum_j \frac{p(i,j)}{\tilde{p}^r(i)\tilde{p}^c(j)} \leq s\cdot\max_i \sum_j \frac{p(i,j)}{\frac{1}{2}p^r(i)\cdot\frac{1}{2m}} \leq 4sm$. Similarly, $\mathbf{E}\left[\left\|\sum_{t=1}^s Q_t^T Q_t\right\|_{\mathrm{sp}}\right] \leq 4sn$. Setting $\rho \doteq \sqrt{4sn}$ and applying [15], we obtain the result. □

Moving from Theorem 1 to Theorem 3, we are competing with a different class of matrices: $\inf_{X\in\mathcal{W}_r[p]} L_p(X) \rightsquigarrow \inf_{X\in\mathcal{W}_r[\tilde{p}]} L_p(X)$. In most applications we can think of, this change is not significant. For example, we consider the low-rank matrix reconstruction problem, where the trace-norm bound is used as a surrogate for rank. In order for the (squared) weighted trace-norm to be a lower bound on the rank, we would need to assume $\left\|\mathrm{diag}\left(p^r\right)^{1/2} X \mathrm{diag}\left(p^c\right)^{1/2}\right\|_F^2 \leq 1$ [11]. If we also assume that $\left\|(X^*)_{(i)}\right\|_2^2 \leq m$ and $\left\|(X^*)^{(j)}\right\|_2^2 \leq n$ for all rows i and columns j — i.e. the row and column magnitudes are not "spiky" — then $X^* \in \mathcal{W}_r\left[\tilde{p}\right]$. Note that this condition is much weaker than placing a spikiness condition on X^* itself, e.g. requiring $|X^*|_\infty \leq 1$.

3.2 Results on Netflix and MovieLens Datasets

We evaluated different models on two publicly-available collaborative filtering datasets: Netflix [16] and MovieLens [17]. The Netflix dataset consists of 100,480,507 ratings from 480,189 users on 17,770 movies. Netflix also provides a qualification set containing 1,408,395 ratings, but due to the sampling scheme, ratings from users with few ratings are overrepresented relative to the training set. To avoid dealing with different training and test distributions, we also created our own validation and test sets, each containing 100,000 ratings set aside from the training set. The MovieLens dataset contains 10,000,054 ratings from 71,567 users on 10,681 movies. We again set aside test and validation sets of 100,000 ratings. Ratings were normalized to be zero-mean.

When dealing with large datasets the most practical way to fit trace-norm regularized models is via stochastic gradient descent [18, 3, 12]. For computational reasons, however, we consider rank-truncated trace-norm minimization, by optimizing within the restricted class $\{X : X \in \mathcal{W}_r\left[p\right], \mathrm{rank}\left(X\right) \leq k\}$ for $k = 30$ and $k = 100$, and for various values of smoothing parameters α (as in (3)). For each value of α and k, the regularization parameter was chosen by cross-validation.

The following table shows root mean squared error (RMSE) for the experiments. For both k=30 and k=100 the weighted trace-norm with smoothing ($\alpha = 0.9$) significantly outperforms the weighted trace-norm without smoothing ($\alpha = 1$), even on the differently-sampled Netflix qualification set. The proposed weighted trace-norm with smoothing outperforms max-norm regularization [19], and performs comparably to "geometric" smoothing [12]. On the Netfix qualification set, using k=30, max-norm regularization and geometric smoothing achieve RMSE 0.9138 [19] and 0.9091 [12], compared to 0.9096 achieved by the weighted trace-norm with smoothing. We note that geometric smoothing was proposed by [12] as a heuristic without any theoretical or conceptual justification.

	Netflix						MovieLens			
α	k	Test	Qual	k	Test	Qual	k	Test	k	Test
1	30	0.7604	0.9107	100	0.7404	0.9078	30	0.7852	100	0.7821
0.9	30	0.7589	0.9096	100	0.7391	0.9068	30	0.7831	100	0.7798
0.5	30	0.7601	0.9173	100	0.7419	0.9161	30	0.7836	100	0.7815
0.3	30	0.7712	0.9198	100	0.7528	0.9207	30	0.7864	100	0.7871
0	30	0.7887	0.9249	100	0.7659	0.9236	30	0.7997	100	0.7987

4 The empirically-weighted trace norm

In practice, the sampling distribution p is not known exactly — it can only be estimated via the locations of the entries which are observed in the sample. Defining the empirical marginals

$$\hat{p}^r\left(i\right) = \frac{\#\{t : i_t = i\}}{s},\ \hat{p}^c\left(j\right) = \frac{\#\{t : j_t = j\}}{s},$$

we would like to give a learning guarantee when $\hat{X}_S$ is estimated via regularization on the $\hat{p}$-weighted trace-norm, rather than the p-weighted trace-norm.

In Section 4.1, we give bounds on excess error when learning with smoothed empirical marginals, which show that there is no theoretical disadvantage as compared to learning with the smoothed true marginals. In fact, we provide evidence that suggests there might even be an *advantage* to using the empirical marginals. To this end, in Section 4.2, we introduce the transductive learning setting, and give a result based on the empirical marginals which implies a sample complexity bound that is better by a factor of $\log^{1/2}(n)$. In Section 4.3, we show that in low-rank matrix reconstruction simulations, using empirical marginals indeed yields better reconstructions.

4.1 Guarantee for the standard (inductive) setting

We first show that when learning with the smoothed empirical marginals, defined as

$$\check{p}^r(i) = \tfrac{1}{2}\left(\hat{p}^r(i) + \tfrac{1}{n}\right),\ \check{p}^c(j) = \tfrac{1}{2}\left(\hat{p}^c(j) + \tfrac{1}{m}\right)\ ,$$

we can obtain the same guarantee as for learning with the smoothed (true) marginals, given by $\tilde{p}$.

Theorem 4. *For an l-Lipschitz loss ℓ, fix any matrix Y, sample size s, and any distribution p. Let $\hat{X}_S = \arg\min\left\{\hat{L}_S(X) \,:\, X \in \mathcal{W}_r[\check{p}]\right\}$. Then, in expectation over the training sample S drawn i.i.d. from the distribution p,*

$$L_p(\hat{X}_S) \leq \inf_{X \in \mathcal{W}_r[\tilde{p}]} L_p(X) + \mathbf{O}\left(l \cdot \sqrt{\frac{r\max\{n,m\}\log(n+m)}{s}}\right). \tag{5}$$

Note that although we regularize using the (smoothed) empirically-weighted trace-norm, we still compare ourselves to the best possible matrix in the class defined by the (smoothed) true marginals.

The proof of this Theorem (in the Supplementary Material) uses Theorem 3 and involves showing that when $s = \mathbf{\Omega}(n\log(n))$, which is required for all Theorems so far to be meaningful, the true and empirical marginals are the same up to a constant factor. For this to be the case, such a sample size is even necessary. In fact, the $\log(n)$ factor in our analysis (e.g. in the proof of Theorem 1) arises from the bound on the expected spectral norm of a matrix, which, for a diagonal matrix, is just a bound on the deviation of empirical frequencies. Might it be possible, then, to avoid this logarithmic factor by using the empirical marginals? Although we could not establish such a result in the inductive setting, we now turn to the transductive setting, where we could indeed obtain a better guarantee.

4.2 Guarantee for the transductive setting

In the transductive model, we fix a set $\overline{S} \subset [n] \times [m]$ of size $2s$, and then randomly split $\overline{S}$ into a training set S and a test set T of equal size s. The goal is to obtain a good estimator for the entries in T based on the values of the entries in S, as well as the locations (indexes) of all elements on $\overline{S}$. We will use the smoothed empirical marginals of $\overline{S}$, for the weighted trace-norm.

We now show that, for bounded loss, there may be a benefit to weighting with the smoothed empirical marginals—the sample size requirement can be lowered to $s = \mathbf{O}(rn\log^{1/2}(n))$.

Theorem 5. *For an l-Lipschitz loss ℓ bounded by b, fix any matrix Y and sample size s. Let $\overline{S} \subset [n] \times [m]$ be a fixed subset of size $2s$, split uniformly at random into training and test sets S and T, each of size s. Let $\overline{p}$ denote the smoothed empirical marginals of $\overline{S}$. Let $\hat{X}_S = \arg\min\left\{\hat{L}_S(X) \,:\, X \in \mathcal{W}_r[\overline{p}]\right\}$. Then in expectation over the splitting of $\overline{S}$ into S and T,*

$$\hat{L}_T(\hat{X}_S) \leq \inf_{X \in \mathcal{W}_r[\overline{p}]} \hat{L}_T(X) + \mathbf{O}\left(l \cdot \sqrt{\frac{rn\log^{1/2}(n)}{s}} + \frac{b}{\sqrt{s}}\right). \tag{6}$$

This result (proved in the Supplementary Materials) is stated in the transductive setting, with a somewhat different sampling procedure and evaluation criteria, but we believe the main difference is in the use of the empirical weights. Although it is usually straightforward to convert a transductive guarantee to an inductive one, the situation here is more complicated, since the hypothesis class depends on the weighting, and hence on the sample $\overline{S}$. Nevertheless, we believe such a conversion might be possible, establishing a similar guarantee for learning with the (smoothed) empirically

weighted trace-norm also in the inductive setting. Furthermore, since the empirical marginals are close to the true marginals when $s = \Theta(n \log(n))$, it might be possible to obtain a learning guarantee for the true (non-empirical) weighting with a sample of size $s = \mathbf{O}\big(n(r \log^{1/2}(n) + \log(n))\big)$.

Theorem 5 can be viewed as a transductive analog to Theorem 3 (with weights based on the combined sample $\overline{S}$). In the Supplementary Materials we give transductive analogs to Theorems 1 and 2. As mentioned in Section 2.3, our lower bound examples can also be stated in the transductive setting, and thus all our guarantees and lower bounds can also be obtained in this setting.

4.3 Simulations with empirical weights

In order to numerically investigate the possible advantage of empirical weighting, we performed simulations on low-rank matrix reconstruction under uniform sampling with the unweighted, and the smoothed empirically weighted, trace-norms. We choose to work with uniform sampling in order to emphasize the benefit of empirical weights, even in situations where one might not consider to use any weights at all. In all the experiments, we attempt to reconstruct a possibly noisy, random rank-2 "signal" matrix M with singular values $\frac{1}{\sqrt{2}}(n, n, 0, \ldots, 0)$, ensuring $\|M\|_F = n$. We measure error using the squared loss[2]. Simulations were performed using MATLAB, with code adapted from the SOFTIMPUTE code developed by [20]. We performed two types of simulations:

Sample complexity comparison in the noiseless setting: We define $Y = M$, and compute $\hat{X}_S = \arg\min \big\{ \|X\| : \hat{L}_S(X) = 0 \big\}$, where $\|X\| = \|X\|_{\mathrm{tr}}$ or $= \|X\|_{\mathrm{tr}(\hat{p}^r, \hat{p}^c)}$, as appropriate. In Figure 1(a), we plot the average number of samples per row needed to get average squared error (over 100 repetitions) of at most 0.1, with both uniform weighting and empirical weighting.

Excess error comparison in the noiseless and noisy settings: We define $Y = M + \nu N$, where noise N has i.i.d. standard normal entries. We compute $\hat{X}_S = \arg\min \big\{ \|X\| : \hat{L}_S(X) \leq \nu^2 \big\}$. In Figure 1(b), we plot the resulting average squared error (over 100 repetitions) over a range of sample sizes s and noise levels ν, with both uniform weighting and empirical weighting. A larger plot including standard error bars is shown in the Supplementary Materials.

The results from both experiments show a significant benefit to using the empirical marginals.

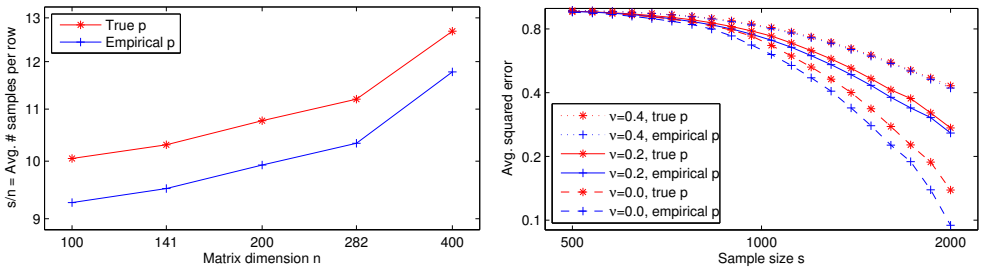

Figure 1: *(a) Left: Sample size needed to obtain avg. error* 0.1*, with respect to* n*. (b) Right: Excess error level over a range of sample sizes, for fixed* $n = 200$*. (Axes are on a logarithmic scale.)*

5 Discussion

In this paper, we prove learning guarantees for the weighted trace-norm by analyzing expected Rademacher complexities. We show that weighting with smoothed marginals eliminates degenerate scenarios that can arise in the case of a non-product sampling distribution, and demonstrate in experiments on the Netflix and MovieLens datasets that this correction can be useful in applied settings. We also give results for empirically-weighted trace-norm regularization, and see indications that using the empirical distribution may be better than using the true distribution, even if it is available.

[2] Although Lipschitz in a bounded domain, it is probably possible to improve all our results (removing the square root) for the special case of the squared-loss, possibly with an i.i.d. noise assumption, as in [10].

References

[1] M. Fazel. *Matrix rank minimization with applications*. PhD Thesis, Stanford University, 2002.

[2] N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems*, 17, 2004.

[3] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. *Advances in Neural Information Processing Systems*, 20, 2007.

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

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

[6] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. *18th Annual Conference on Learning Theory (COLT)*, pages 545–560, 2005.

[7] B. Recht. A simpler approach to matrix completion. *arXiv:0910.0651*, 2009.

[8] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.

[9] V. Koltchinskii, A. Tsybakov, and K. Lounici. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *arXiv:1011.6256*, 2010.

[10] S. Negahban and M. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *arXiv:1009.2118*, 2010.

[11] R. Foygel and N. Srebro. Concentration-based guarantees for low-rank matrix reconstruction. *24th Annual Conference on Learning Theory (COLT)*, 2011.

[12] R. Salakhutdinov and N. Srebro. Collaborative Filtering in a Non-Uniform World: Learning with the Weighted Trace Norm. *Advances in Neural Information Processing Systems*, 23, 2010.

[13] O. Shamir and S. Shalev-Shwartz. Collaborative filtering with the trace norm: Learning, bounding, and transducing. *24th Annual Conference on Learning Theory (COLT)*, 2011.

[14] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[15] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *arXiv:1004.4389*, 2010.

[16] J. Bennett and S. Lanning. The netflix prize. In *Proceedings of KDD Cup and Workshop*, volume 2007, page 35. Citeseer, 2007.

[17] MovieLens Dataset. Available at http://www.grouplens.org/node/73. 2006.

[18] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. *ACM Int. Conference on Knowledge Discovery and Data Mining (KDD'08)*, pages 426–434, 2008.

[19] J. Lee, B. Recht, R. Salakhutdinov, N. Srebro, and J. Tropp. Practical Large-Scale Optimization for Max-Norm Regularization. *Advances in Neural Information Processing Systems*, 23, 2010.

[20] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.