Matrix reconstruction with the local max norm

Rina Foygel
Department of Statistics
Stanford University
rinafb@stanford.edu

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

Ruslan Salakhutdinov
Dept. of Statistics and Dept. of Computer Science University of Toronto
rsalakhu@utstat.toronto.edu

Abstract

We introduce a new family of matrix norms, the "local max" norms, generalizing existing methods such as the max norm, the trace norm (nuclear norm), and the weighted or smoothed weighted trace norms, which have been extensively used in the literature as regularizers for matrix reconstruction problems. We show that this new family can be used to interpolate between the (weighted or unweighted) trace norm and the more conservative max norm. We test this interpolation on simulated data and on the large-scale Netflix and MovieLens ratings data, and find improved accuracy relative to the existing matrix norms. We also provide theoretical results showing learning guarantees for some of the new norms.

1 Introduction

In the matrix reconstruction problem, we are given a matrix $Y \in \mathbb{R}^{n\times m}$ whose entries are only partly observed, and would like to reconstruct the unobserved entries as accurately as possible. Matrix reconstruction arises in many modern applications, including the areas of collaborative filtering (e.g. the Netflix prize), image and video data, and others. This problem has often been approached using regularization with matrix norms that promote low-rank or approximately-low-rank solutions, including the trace norm (also known as the nuclear norm) and the max norm, as well as several adaptations of the trace norm described below.

In this paper, we introduce a unifying family of norms that generalizes these existing matrix norms, and that can be used to interpolate between the trace and max norms. We show that this family includes new norms, lying strictly between the trace and max norms, that give empirical and theoretical improvements over the existing norms. We give results allowing for large-scale optimization with norms from the new family. Some proofs are deferred to the Supplementary Materials.

Notation Without loss of generality we take $n \geq m$. We let $\mathbb{R}_+$ denote the nonnegative real numbers. For any $n \in \mathbb{N}$, let $[n] = \{1, \ldots, n\}$, and define the simplex on $[n]$ as $\Delta_{[n]} = \left\{\mathbf{r} \in \mathbb{R}_+^n : \sum_i \mathbf{r}_i = 1\right\}$. We analyze situations where the locations of observed entries are sampled i.i.d. according to some distribution $\mathbf{p}$ on $[n] \times [m]$. We write $\mathbf{p}_{i\bullet} = \sum_j \mathbf{p}_{ij}$ to denote the marginal probability of row i, and $\mathbf{p}_{\mathrm{row}} = (\mathbf{p}_{1\bullet}, \ldots, \mathbf{p}_{n\bullet}) \in \Delta_{[n]}$ to denote the marginal row distribution. We define $\mathbf{p}_{\bullet j}$ and $\mathbf{p}_{\mathrm{col}}$ similarly for the columns. For any matrix M, $M_{(i)}$ denotes its ith row.

1.1 Trace norm and max norm

A common regularizer used in matrix reconstruction, and other matrix problems, is the trace norm $\|X\|_{\mathrm{tr}}$, equal to the sum of the singular values of X. This norm can also be defined via a factorization

of X [1]:

$$\frac{1}{\sqrt{nm}}\|X\|_{\mathrm{tr}} = \frac{1}{2}\min_{AB^\top = X}\left(\frac{1}{n}\sum_i \left\|A_{(i)}\right\|^2 + \frac{1}{m}\sum_j \left\|B_{(j)}\right\|^2\right) , \tag{1}$$

where the minimum is taken over factorizations of X of arbitrary dimension—that is, the number of columns in A and B is unbounded. Note that we choose to scale the trace norm by $1/\sqrt{nm}$ in order to emphasize that we are averaging the squared row norms of A and B.

Regularization with the trace norm gives good theoretical and empirical results, as long as the locations of observed entries are sampled uniformly (i.e. when $\mathbf{p}$ is the uniform distribution on $[n]\times[m]$), and, under this assumption, can also be used to guarantee approximate recovery of an underlying low-rank matrix [1, 2, 3, 4].

The factorized definition of the trace norm (1) allows for an intuitive comparison with the max norm, defined as [1]:

$$\|X\|_{\max} = \frac{1}{2}\min_{AB^\top = X}\left(\sup_i \left\|A_{(i)}\right\|_2^2 + \sup_j \left\|B_{(j)}\right\|_2^2\right) . \tag{2}$$

We see that the max norm measures the largest row norms in the factorization, while the rescaled trace norm instead considers the average row norms. The max norm is therefore an upper bound on the rescaled trace norm, and can be viewed as a more conservative regularizer. For the more general setting where $\mathbf{p}$ may not be uniform, Foygel and Srebro [4] show that the max norm is still an effective regularizer (in particular, bounds on error for the max norm are not affected by $\mathbf{p}$). On the other hand, Salakhutdinov and Srebro [5] show that the trace norm is not robust to non-uniform sampling—regularizing with the trace norm may yield large error due to over-fitting on the rows and columns with high marginals. They obtain improved empirical results by placing more penalization on these over-represented rows and columns, described next.

1.2 The weighted trace norm

To reduce overfitting on the rows and columns with high marginal probabilities under the distribution $\mathbf{p}$, Salakhutdinov and Srebro propose regularizing with the $\mathbf{p}$-weighted trace norm,

$$\|X\|_{\mathrm{tr}(\mathbf{p})} := \left\|\mathrm{diag}(\mathbf{p}_{\mathrm{row}})^{1/2}\cdot X\cdot \mathrm{diag}(\mathbf{p}_{\mathrm{col}})^{1/2}\right\|_{\mathrm{tr}} .$$

If the row and the column of entries to be observed are sampled independently (i.e. $\mathbf{p} = \mathbf{p}_{\mathrm{row}}\cdot\mathbf{p}_{\mathrm{col}}$ is a product distribution), then the $\mathbf{p}$-weighted trace norm can be used to obtain good learning guarantees even when $\mathbf{p}_{\mathrm{row}}$ and $\mathbf{p}_{\mathrm{col}}$ are non-uniform [3, 6]. However, for non-uniform non-product sampling distributions, even the $\mathbf{p}$-weighted trace norm can yield poor generalization performance. To correct for this, Foygel et al. [6] suggest adding in some "smoothing" to avoid under-penalizing the rows and columns with low marginal probabilities, and obtain improved empirical and theoretical results for matrix reconstruction using the smoothed weighted trace norm:

$$\|X\|_{\mathrm{tr}(\widetilde{\mathbf{p}})} := \left\|\mathrm{diag}(\widetilde{\mathbf{p}}_{\mathrm{row}})^{1/2}\cdot X\cdot \mathrm{diag}(\widetilde{\mathbf{p}}_{\mathrm{col}})^{1/2}\right\|_{\mathrm{tr}} ,$$

where $\widetilde{\mathbf{p}}_{\mathrm{row}}$ and $\widetilde{\mathbf{p}}_{\mathrm{col}}$ denote smoothed row and column marginals, given by

$$\widetilde{\mathbf{p}}_{\mathrm{row}} = (1-\zeta)\cdot\mathbf{p}_{\mathrm{row}} + \zeta\cdot 1/n \text{ and } \widetilde{\mathbf{p}}_{\mathrm{col}} = (1-\zeta)\cdot\mathbf{p}_{\mathrm{col}} + \zeta\cdot 1/m , \tag{3}$$

for some choice of smoothing parameter ζ which may be selected with cross-validation[1]. The smoothed empirically-weighted trace norm is also studied in [6], where $\mathbf{p}_{i\bullet}$ is replaced with $\widehat{\mathbf{p}}_{i\bullet} = \frac{\#\text{ observations in row } i}{\text{total }\#\text{ observations}}$, the empirical marginal probability of row i (and same for $\widehat{\mathbf{p}}_{\bullet j}$). Using empirical rather than "true" weights yielded lower error in experiments in [6], even when the true sampling distribution was uniform.

More generally, for any weight vectors $\mathbf{r}\in\Delta_{[n]}$ and $\mathbf{c}\in\Delta_{[m]}$ and a matrix $X\in\mathbb{R}^{n\times m}$, the $(\mathbf{r},\mathbf{c})$-weighted trace norm is given by

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

[1] Our ζ parameter here is equivalent to $1-\alpha$ in [6].

Of course, we can easily obtain the existing methods of the uniform trace norm, (empirically) weighted trace norm, and smoothed (empirically) weighted trace norm as special cases of this formulation. Furthermore, the max norm is equal to a supremum over all possible weightings [7]:

$$\|X\|_{\max} = \sup_{\mathbf{r}\in\Delta_{[n]},\mathbf{c}\in\Delta_{[m]}} \|X\|_{\mathrm{tr}(\mathbf{r},\mathbf{c})} \ .$$

2 The local max norm

We consider a generalization of these norms, which lies "in between" the trace norm and max norm. For any $\mathcal{R} \subseteq \Delta_{[n]}$ and $\mathcal{C} \subseteq \Delta_{[m]}$, we define the $(\mathcal{R},\mathcal{C})$-norm of X:

$$\|X\|_{(\mathcal{R},\mathcal{C})} = \sup_{\mathbf{r}\in\mathcal{R},\mathbf{c}\in\mathcal{C}} \|X\|_{\mathrm{tr}(\mathbf{r},\mathbf{c})} \ .$$

This gives a norm on matrices, except in the trivial case where, for some i or some j, $\mathbf{r}_i = 0$ for all $\mathbf{r} \in \mathcal{R}$ or $\mathbf{c}_j = 0$ for all $\mathbf{c} \in \mathcal{C}$.

We now show some existing and novel norms that can be obtained using local max norms.

2.1 Trace norm and max norm

We can obtain the max norm by taking the largest possible $\mathcal{R}$ and $\mathcal{C}$, i.e. $\|X\|_{\max} = \|X\|_{(\Delta_{[n]},\Delta_{[m]})}$, and similarly we can obtain the $(\mathbf{r},\mathbf{c})$-weighted trace norm by taking the singleton sets $\mathcal{R} = \{\mathbf{r}\}$ and $\mathcal{C} = \{\mathbf{c}\}$. As discussed above, this includes the standard trace norm (when $\mathbf{r}$ and $\mathbf{c}$ are uniform), as well as the weighted, empirically weighted, and smoothed weighted trace norm.

2.2 Arbitrary smoothing

When using the smoothed weighted max norm, we need to choose the amount of smoothing to apply to the marginals, that is, we need to choose ζ in our definition of the smoothed row and column weights, as given in (3). Alternately, we could regularize simultaneously over all possible amounts of smoothing by considering the local max norm with

$$\mathcal{R} = \{(1-\zeta)\cdot\mathbf{p}_{\mathrm{row}} + \zeta\cdot 1/n : \text{ any } \zeta\in[0,1]\} \ ,$$

and same for $\mathcal{C}$. That is, $\mathcal{R}$ and $\mathcal{C}$ are line segments in the simplex—they are larger than any single point as for the uniform or weighted trace norm (or smoothed weighted trace norm for a fixed amount of smoothing), but smaller than the entire simplex as for the max norm.

2.3 Connection to (β,τ)-decomposability

Hazan et al. [8] introduce a class of matrices defined by a property of (β,τ)-decomposability: a matrix X satisfies this property if there exists a factorization $X = AB^\top$ (where A and B may have an arbitrary number of columns) such that[2]

$$\max\left\{\max_i \left\|A_{(i)}\right\|_2^2, \max_j \left\|B_{(j)}\right\|_2^2\right\} \le 2\beta, \ \sum_i \left\|A_{(i)}\right\|_2^2 + \sum_j \left\|B_{(j)}\right\|_2^2 \le \tau \ .$$

Comparing with (1) and (2), we see that the β and τ parameters essentially correspond to the max norm and trace norm, with the max norm being the minimal $2\beta^*$ such that the matrix is (β^*,τ)-decomposable for some τ, and the trace norm being the minimal $\tau^*/2$ such that the matrix is (β,τ^*)-decomposable for some β. However, Hazan et al. go beyond these two extremes, and rely on balancing both β and τ: they establish learning guarantees (in an adversarial online model, and thus also under an arbitrary sampling distribution $\mathbf{p}$) which scale with $\sqrt{\beta\cdot\tau}$. It may therefore be useful to consider a penalty function of the form:

$$\mathrm{Penalty}_{(\beta,\tau)}(X) = \min_{X=AB^\top}\left\{\sqrt{\max_i \left\|A_{(i)}\right\|_2^2 + \max_j \left\|B_{(j)}\right\|_2^2}\cdot\sqrt{\sum_i \left\|A_{(i)}\right\|_2^2 + \sum_j \left\|B_{(j)}\right\|_2^2}\right\} \ . \tag{4}$$

[2]Hazan et al. state the property differently, but equivalently, in terms of a semidefinite matrix decomposition.

(Note that $\max\left\{\max_i \left\|A_{(i)}\right\|_2^2, \max_j \left\|B_{(j)}\right\|_2^2\right\}$ is replaced with $\max_i \left\|A_{(i)}\right\|_2^2 + \max_j \left\|B_{(j)}\right\|_2^2$, for later convenience. This affects the value of the penalty function by at most a factor of $\sqrt{2}$.)

This penalty function does not appear to be convex in X. However, the proposition below (proved in the Supplementary Materials) shows that we can use a (convex) local max norm penalty to compute a solution to any objective function with a penalty function of the form (4):

Proposition 1. *Let $\widehat{X}$ be the minimizer of a penalized loss function with this modified penalty,*

$$\widehat{X} := \arg\min_X \left\{\mathrm{Loss}(X) + \lambda \cdot \mathrm{Penalty}_{(\beta,\tau)}(X)\right\} \ ,$$

where $\lambda \geq 0$ is some penalty parameter and $\mathrm{Loss}(\cdot)$ is any convex function. Then, for some penalty parameter $\mu \geq 0$ and some $t \in [0,1]$,

$$\widehat{X} = \arg\min_X \left\{\mathrm{Loss}(X) + \mu \cdot \|X\|_{(\mathcal{R},\mathcal{C})}\right\} \ , \textit{ where}$$

$$\mathcal{R} = \left\{\mathbf{r} \in \Delta_{[n]} : \mathbf{r}_i \geq \frac{t}{1+(n-1)t}\ \forall i\right\} \textit{ and } \mathcal{C} = \left\{\mathbf{c} \in \Delta_{[m]} : \mathbf{c}_j \geq \frac{t}{1+(m-1)t}\ \forall j\right\} \ .$$

We note that μ and t cannot be determined based on λ alone—they will depend on the properties of the unknown solution $\widehat{X}$.

Here the sets $\mathcal{R}$ and $\mathcal{C}$ impose a lower bound on each of the weights, and this lower bound can be used to interpolate between the max and trace norms: when $t = 1$, each $\mathbf{r}_i$ is lower bounded by $1/n$ (and similarly for $\mathbf{c}_j$), i.e. $\mathcal{R}$ and $\mathcal{C}$ are singletons containing only the uniform weights and we obtain the trace norm. On the other hand, when $t = 0$, the weights are lower-bounded by zero, and so any weight vector is allowed, i.e. $\mathcal{R}$ and $\mathcal{C}$ are each the entire simplex and we obtain the max norm. Intermediate values of t interpolate between the trace norm and max norm and correspond to different balances between β and τ.

2.4 Interpolating between trace norm and max norm

We next turn to an interpolation which relies on an upper bound, rather than a lower bound, on the weights. Consider

$$\mathcal{R}_\epsilon = \left\{\mathbf{r} \in \Delta_{[n]} : \mathbf{r}_i \leq \epsilon\ \forall i\right\} \text{ and } \mathcal{C}_\delta = \left\{\mathbf{c} \in \Delta_{[n]} : \mathbf{c}_j \leq \delta\ \forall j\right\} \ , \tag{5}$$

for some $\epsilon \in [1/n, 1]$ and $\delta \in [1/m, 1]$. The $(\mathcal{R}_\epsilon, \mathcal{C}_\delta)$-norm is then equal to the (rescaled) trace norm when we choose $\epsilon = 1/n$ and $\delta = 1/m$, and is equal to the max norm when we choose $\epsilon = \delta = 1$. Allowing ϵ and δ to take intermediate values gives a smooth interpolation between these two familiar norms, and may be useful in situations where we want more flexibility in the type of regularization.

We can generalize this to an interpolation between the max norm and a smoothed weighted trace norm, which we will use in our experimental results. We consider two generalizations—for each one, we state a definition of $\mathcal{R}$, with $\mathcal{C}$ defined analogously. The first is multiplicative:

$$\mathcal{R}^{\times}_{\zeta,\gamma} := \left\{\mathbf{r} \in \Delta_{[n]} : \mathbf{r}_i \leq \gamma \cdot ((1-\zeta) \cdot \mathbf{p}_{i\bullet} + \zeta \cdot 1/n)\ \forall i\right\} \ , \tag{6}$$

where $\gamma = 1$ corresponds to choosing the singleton set $\mathcal{R}^{\times}_{\zeta,\gamma} = \{(1-\zeta) \cdot \mathbf{p}_{\mathrm{row}} + \zeta \cdot 1/n\}$ (i.e. the smoothed weighted trace norm), while $\gamma = \infty$ corresponds to the max norm (for any choice of ζ) since we would get $\mathcal{R}^{\times}_{\zeta,\gamma} = \Delta_{[n]}$.

The second option for an interpolation is instead defined with an exponent:

$$\mathcal{R}_{\zeta,\tau} := \left\{\mathbf{r} \in \Delta_{[n]} : \mathbf{r}_i \leq ((1-\zeta) \cdot \mathbf{p}_{i\bullet} + \zeta \cdot 1/n)^{1-\tau}\ \forall i\right\} \ . \tag{7}$$

Here $\tau = 0$ will yield the singleton set corresponding to the smoothed weighted trace norm, while $\tau = 1$ will yield $\mathcal{R}_{\zeta,\tau} = \Delta_{[n]}$, i.e. the max norm, for any choice of ζ.

We find the second (exponent) option to be more natural, because each of the row marginal bounds will reach 1 simultaneously when $\tau = 1$, and hence we use this version in our experiments. On the other hand, the multiplicative version is easier to work with theoretically, and we use this in our learning guarantee in Section 4.2. If all of the row and column marginals satisfy some loose upper bound, then the two options will not be highly different.

3 Optimization with the local max norm

One appeal of both the trace norm and the max norm is that they are both SDP representable [9, 10], and thus easily optimizable, at least in small scale problems. In the Supplementary Materials we show that the local max norm is also SDP representable, as long as the sets $\mathcal{R}$ and $\mathcal{C}$ can be written in terms of linear or semi-definite constraints—this includes all the examples we mention, where in all of them the sets $\mathcal{R}$ and $\mathcal{C}$ are specified in terms of simple linear constraints.

However, for large scale problems, it is not practical to directly use SDP optimization approaches. Instead, and especially for very large scale problems, an effective optimization approach for both the trace norm and the max norm is to use the factorized versions of the norms, given in (1) and (2), and to optimize the factorization directly (typically, only factorizations of some truncated dimensionality are used) [11, 12, 7]. As we show in Theorem 1 below, a similar factorization-optimization approach is also possible for any local max norm with convex $\mathcal{R}$ and $\mathcal{C}$. We further give a simplified representation which is applicable when $\mathcal{R}$ and $\mathcal{C}$ are specified through element-wise upper bounds $R \in \mathbb{R}^n_+$ and $C \in \mathbb{R}^m_+$, respectively:

$$\mathcal{R} = \{\mathbf{r} \in \Delta_{[n]} : \mathbf{r}_i \le R_i \ \forall i\} \text{ and } \mathcal{C} = \{\mathbf{c} \in \Delta_{[m]} : \mathbf{c}_j \le C_j \ \forall j\} \ , \tag{8}$$

with $0 \le R_i \le 1$, $\sum_i R_i \ge 1$, $0 \le C_j \le 1$, $\sum_j C_j \ge 1$ to avoid triviality. This includes the interpolation norms of Section 2.4.

Theorem 1. *If $\mathcal{R}$ and $\mathcal{C}$ are convex, then the $(\mathcal{R},\mathcal{C})$-norm can be calculated with the factorization*

$$\|X\|_{(\mathcal{R},\mathcal{C})} = \frac{1}{2} \inf_{AB^\top = X} \left(\sup_{\mathbf{r}\in\mathcal{R}} \sum_i \mathbf{r}_i \left\|A_{(i)}\right\|_2^2 + \sup_{\mathbf{c}\in\mathcal{C}} \sum_j \mathbf{c}_j \left\|B_{(j)}\right\|_2^2 \right) . \tag{9}$$

In the special case when $\mathcal{R}$ and $\mathcal{C}$ are defined by (8)*, writing $(x)_+ := \max\{0, x\}$, this simplifies to*

$$\|X\|_{(\mathcal{R},\mathcal{C})} = \frac{1}{2} \inf_{AB^\top = X; a,b\in\mathbb{R}} \left\{ a + \sum_i R_i \left(\left\|A_{(i)}\right\|_2^2 - a \right)_+ + b + \sum_j C_j \left(\left\|B_{(j)}\right\|_2^2 - b \right)_+ \right\} .$$

Proof sketch for Theorem 1. For convenience we will write $\mathbf{r}^{1/2}$ to mean $\mathrm{diag}(\mathbf{r})^{1/2}$, and same for $\mathbf{c}$. Using the trace norm factorization identity (1), we have

$$2\|X\|_{(\mathcal{R},\mathcal{C})} = 2 \sup_{\mathbf{r}\in\mathcal{R},\mathbf{c}\in\mathcal{C}} \left\|\mathbf{r}^{1/2} \cdot X \cdot \mathbf{c}^{1/2}\right\|_{\mathrm{tr}} = \sup_{\mathbf{r}\in\mathcal{R},\mathbf{c}\in\mathcal{C}} \inf_{CD^\top = \mathbf{r}^{1/2}\cdot X\cdot \mathbf{c}^{1/2}} \left(\|C\|_{\mathrm{F}}^2 + \|D\|_{\mathrm{F}}^2 \right)$$

$$= \sup_{\mathbf{r}\in\mathcal{R},\mathbf{c}\in\mathcal{C}} \inf_{AB^\top = X} \left(\left\|\mathbf{r}^{1/2}\cdot A\right\|_{\mathrm{F}}^2 + \left\|\mathbf{c}^{1/2}\cdot B\right\|_{\mathrm{F}}^2 \right) \le \inf_{AB^\top = X} \left(\sup_{\mathbf{r}\in\mathcal{R}} \left\|\mathbf{r}^{1/2} A\right\|_{\mathrm{F}}^2 + \sup_{\mathbf{c}\in\mathcal{C}} \left\|\mathbf{c}^{1/2} B\right\|_{\mathrm{F}}^2 \right) ,$$

where for the next-to-last step we set $C = \mathbf{r}^{1/2}A$ and $D = \mathbf{c}^{1/2}B$, and the last step follows because $\sup\inf \le \inf\sup$ always (weak duality). The reverse inequality holds as well (strong duality), and is proved in the Supplementary Materials, where we also prove the special-case result. □

4 An approximate convex hull and a learning guarantee

In this section, we look for theoretical bounds on error for the problem of estimating unobserved entries in a matrix Y that is approximately low-rank. Our results apply for either uniform or non-uniform sampling of entries from the matrix. We begin with a result comparing the $(\mathcal{R},\mathcal{C})$-norm unit ball to a convex hull of rank-1 matrices, which will be useful for proving our learning guarantee.

4.1 Convex hull

To gain a better theoretical understanding of the $(\mathcal{R},\mathcal{C})$ norm, we first need to define corresponding vector norms on $\mathbb{R}^n$ and $\mathbb{R}^m$. For any $u \in \mathbb{R}^n$, let

$$\|u\|_{\mathcal{R}} := \sqrt{\sup_{\mathbf{r}\in\mathcal{R}} \sum_i \mathbf{r}_i u_i^2} = \sup_{\mathbf{r}\in\mathcal{R}} \left\|\mathrm{diag}(\mathbf{r})^{1/2} \cdot u\right\|_2 .$$

We can think of this norm as a way to interpolate between the ℓ_2 and ℓ_∞ vector norms. For example, if we choose $\mathcal{R} = \mathcal{R}_\epsilon$ as defined in (5), then $\|u\|_{\mathcal{R}}$ is equal to the root-mean-square of the ϵ^{-1} largest entries of u whenever ϵ^{-1} is an integer. Defining $\|v\|_{\mathcal{C}}$ analogously for $v \in \mathbb{R}^m$, we can now relate these vector norms to the $(\mathcal{R},\mathcal{C})$-norm on matrices.

Theorem 2. *For any convex $\mathcal{R} \subseteq \Delta_{[n]}$ and $\mathcal{C} \subseteq \Delta_{[m]}$, the $(\mathcal{R}, \mathcal{C})$-norm unit ball is bounded above and below by a convex hull as:*

$$\mathrm{Conv}\left\{uv^\top : \|u\|_{\mathcal{R}} = \|v\|_{\mathcal{C}} = 1\right\} \subseteq \left\{X : \|X\|_{(\mathcal{R},\mathcal{C})} \leq 1\right\} \subseteq K_G \cdot \mathrm{Conv}\left\{uv^\top : \|u\|_{\mathcal{R}} = \|v\|_{\mathcal{C}} = 1\right\},$$

where $K_G \leq 1.79$ is Grothendieck's constant, and implicitly $u \in \mathbb{R}^n$, $v \in \mathbb{R}^m$.

This result is a nontrivial extension of Srebro and Shraibman [1]'s analysis for the max norm and the trace norm. They show that the statement holds for the max norm, i.e. when $\mathcal{R} = \Delta_{[n]}$ and $\mathcal{C} = \Delta_{[m]}$, and that the trace norm unit ball is exactly equal to the corresponding convex hull (see Corollary 2 and Section 3.2 in their paper, respectively).

Proof sketch for Theorem 2. To prove the first inclusion, given any $X = uv^\top$ with $\|u\|_{\mathcal{R}} = \|v\|_{\mathcal{C}} = 1$, we apply the factorization result Theorem 1 to see that $\|X\|_{(\mathcal{R},\mathcal{C})} \leq 1$. Since the $(\mathcal{R}, \mathcal{C})$-norm unit ball is convex, this is sufficient. For the second inclusion, we state a weighted version of Grothendieck's Inequality (proof in the Supplementary Materials):

$$\begin{aligned}\sup\left\{\langle Y, UV^\top\rangle : U \in \mathbb{R}^{n\times k}, V \in \mathbb{R}^{m\times k}, \left\|U_{(i)}\right\|_2 \leq a_i\ \forall i,\ \left\|V_{(j)}\right\|_2 \leq b_j\ \forall j\right\} \\ = K_G \cdot \sup\left\{\langle Y, uv^\top\rangle : u \in \mathbb{R}^n, v \in \mathbb{R}^m, |u_i| \leq a_i\ \forall i,\ |v_j| \leq b_j\ \forall j\right\}\ .\end{aligned}$$

We then apply this weighted inequality to the dual norm of the $(\mathcal{R}, \mathcal{C})$-norm to prove the desired inclusion, as in Srebro and Shraibman [1]'s work for the max norm case (see Corollary 2 in their paper). Details are given in the Supplementary Materials. □

4.2 Learning guarantee

We now give our main matrix reconstruction result, which provides error bounds for a family of norms interpolating between the max norm and the smoothed weighted trace norm.

Theorem 3. *Let $\mathbf{p}$ be any distribution on $[n] \times [m]$. Suppose that, for some $\gamma \geq 1$, $\mathcal{R} \supseteq \mathcal{R}^{\times}_{1/2,\gamma}$ and $\mathcal{C} \supseteq \mathcal{C}^{\times}_{1/2,\gamma}$, where these two sets are defined in* (6)*. Let $S = \{(i_t, j_t) : t = 1, \ldots, s\}$ be a random sample of locations in the matrix drawn i.i.d. from $\mathbf{p}$, where $s \geq n$. Then, in expectation over the sample S,*

$$\sum_{ij} \mathbf{p}_{ij} \left|Y_{ij} - \widehat{X}_{ij}\right| \leq \underbrace{\inf_{\|X\|_{(\mathcal{R},\mathcal{C})} \leq \sqrt{k}} \sum_{ij} \mathbf{p}_{ij} \left|Y_{ij} - X_{ij}\right|}_{\textit{Approximation error}} + \underbrace{\mathcal{O}\left(\sqrt{\frac{kn}{s}}\right) \cdot \left(1 + \frac{\log(n)}{\sqrt{\gamma}}\right)}_{\textit{Excess error}},$$

where $\widehat{X} = \arg\min_{\|X\|_{(\mathcal{R},\mathcal{C})} \leq \sqrt{k}} \sum_{t=1}^{s} |Y_{i_t j_t} - X_{i_t j_t}|$. Additionally, if we assume that $s \geq n\log(n)$, then in the excess risk bound, we can reduce the term $\log(n)$ to $\sqrt{\log(n)}$.

Proof sketch for Theorem 3. The main idea is to use the convex hull formulation from Theorem 2 to show that, for any X with $\|X\|_{(\mathcal{R},\mathcal{C})} \leq \sqrt{k}$, there exists a decomposition $X = X' + X''$ with $\|X'\|_{\max} \leq \mathcal{O}(\sqrt{k})$ and $\|X''\|_{\mathrm{tr}(\widetilde{\mathbf{p}})} \leq \mathcal{O}(\sqrt{k/\gamma})$, where $\widetilde{\mathbf{p}}$ represents the smoothed marginals with smoothing parameter $\zeta = 1/2$ as in (3). We then apply known bounds on the Rademacher complexity of the max norm unit ball [1] and the smoothed weighted trace norm unit ball [6], to bound the Rademacher complexity of $\left\{X : \|X\|_{(\mathcal{R},\mathcal{C})} \leq \sqrt{k}\right\}$. This then yields a learning guarantee by Theorem 8 of Bartlett and Mendelson [13]. Details are given in the Supplementary Materials. □

As special cases of this theorem, we can re-derive the existing results for the max norm and smoothed weighted trace norm. Specifically, choosing $\gamma = \infty$ gives us an excess error term of order $\sqrt{kn/s}$ for the max norm, previously shown by [1], while setting $\gamma = 1$ yields an excess error term of order $\sqrt{kn\log(n)/s}$ for the smoothed weighted trace norm as long as $s \geq n\log(n)$, as shown in [6].

What advantage does this new result offer over the existing results for the max norm and for the smoothed weighted trace norm? To simplify the comparison, suppose we choose $\gamma = \log^2(n)$, and define $\mathcal{R} = \mathcal{R}^{\times}_{1/2,\gamma}$ and $\mathcal{C} = \mathcal{C}^{\times}_{1/2,\gamma}$. Then, comparing to the max norm result (when $\gamma = \infty$), we see

Table 1: Matrix fitting for the five methods used in experiments.

Norm	Fixed parameters	Free parameters
Max norm	ζ arbitrary; $\tau = 1$	λ
(Uniform) trace norm	$\zeta = 1$; $\tau = 0$	λ
Empirically-weighted trace norm	$\zeta = 0$; $\tau = 0$	λ
Arbitrarily-smoothed emp.-wtd. trace norm	$\tau = 0$	ζ; λ
Local max norm	—	ζ; τ; λ

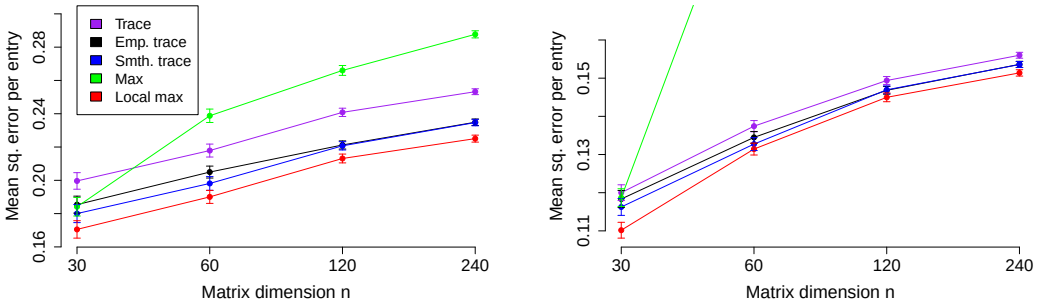

Figure 1: Simulation results for matrix reconstruction with a rank-2 (left) or rank-4 (right) signal, corrupted by noise. The plot shows per-entry squared error averaged over 50 trials, with standard error bars. For the rank-4 experiment, max norm error exceeded 0.20 for each $n = 60, 120, 240$ and is not displayed in the plot.

that the excess error term is the same in both cases (up to a constant), but the approximation error term may in general be much lower for the local max norm than for the max norm. Comparing next to the weighted trace norm (when $\gamma = 1$), we see that the excess error term is lower by a factor of $\log(n)$ for the local max norm. This may come at a cost of increasing the approximation error, but in general this increase will be very small. In particular, the local max norm result allows us to give a meaningful guarantee for a sample size $s = \Theta(kn)$, rather than requiring $s \geq \Theta(kn \log(n))$ as for any trace norm result, but with a hypothesis class significantly richer than the max norm constrained class (though not as rich as the trace norm constrained class).

5 Experiments

We test the local max norm on simulated and real matrix reconstruction tasks, and compare its performance to the max norm, the uniform and empirically-weighted trace norms, and the smoothed empirically-weighted trace norm.

5.1 Simulations

We simulate $n \times n$ noisy matrices for $n = 30, 60, 120, 240$, where the underlying signal has rank $k = 2$ or $k = 4$, and we observe $s = 3kn$ entries (chosen uniformly without replacement). We performed 50 trials for each of the 8 combinations of (n, k).

Data For each trial, we randomly draw a matrix $U \in \mathbb{R}^{n \times k}$ by drawing each row uniformly at random from the unit sphere in $\mathbb{R}^n$. We generate $V \in \mathbb{R}^{m \times k}$ similarly. We set $Y = UV^\top + \sigma \cdot Z$, where the noise matrix Z has i.i.d. standard normal entries and $\sigma = 0.3$ is a moderate noise level. We also divide the n^2 entries of the matrix into sets $S_0 \sqcup S_1 \sqcup S_2$ which consist of $s = 3kn$ training entries, s validation entries, and $n^2 - 2s$ test entries, respectively, chosen uniformly at random.

Methods We use the two-parameter family of norms defined in (7), but replacing the true marginals $\mathbf{p}_{i\bullet}$ and $\mathbf{p}_{\bullet j}$ with the empirical marginals $\widehat{\mathbf{p}}_{i\bullet}$ and $\widehat{\mathbf{p}}_{\bullet j}$. For each $\zeta, \tau \in \{0, 0.1, \ldots, 0.9, 1\}$ and each penalty parameter value $\lambda \in \{2^1, 2^2, \ldots, 2^{10}\}$, we compute the fitted matrix

$$\widehat{X} = \arg\min \left\{ \sum_{(i,j) \in S_0} (Y_{ij} - X_{ij})^2 + \lambda \cdot \|X\|_{(\mathcal{R}_{\zeta,\tau}, \mathcal{C}_{\zeta,\tau})} \right\}. \tag{10}$$

(In fact, we use a rank-8 approximation to this optimization problem, as described in Section 3.) For each method, we use S_1 to select the best ζ, τ, and λ, with restrictions on ζ and/or τ as specified by the definition of the method (see Table 1), then report the error of the resulting fitted matrix on S_2.

Results The results for these simulations are displayed in Figure 1. We see that the local max norm results in lower error than any of the tested existing norms, across all the settings used.

Table 2: Root mean squared error (RMSE) results for estimating movie ratings on Netflix and MovieLens data using a rank 30 model. Setting $\tau = 0$ corresponds to the uniform or weighted or smoothed weighted trace norm (depending on ζ), while $\tau = 1$ corresponds to the max norm for any ζ value.

MovieLens				
$\zeta \setminus \tau$	0.00	0.05	0.10	1.00
0.00	0.7852	0.7827	0.7838	0.7918
0.05	0.7836	**0.7822**	0.7842	—
0.10	0.7831	0.7837	0.7846	—
0.15	0.7833	0.7842	0.7854	—
0.20	0.7842	0.7853	0.7866	—
1.00	0.7997			

Netflix				
$\zeta \setminus \tau$	0.00	0.05	0.10	1.00
0.00	0.9107	0.9092	0.9094	0.9131
0.05	0.9095	**0.9090**	0.9107	—
0.10	0.9096	0.9098	0.9122	—
0.15	0.9102	0.9111	0.9131	—
0.20	0.9126	0.9344	0.9153	—
1.00	0.9235			

5.2 Movie ratings data

We next compare several different matrix norms on two collaborative filtering movie ratings datasets, the Netflix [14] and MovieLens [15] datasets. The sizes of the data sets, and the split of the ratings into training, validation and test sets[3], are:

Dataset	# users	# movies	Training set	Validation set	Test set
Netflix	480,189	17,770	100,380,507	100,000	1,408,395
MovieLens	71,567	10,681	8,900,054	100,000	1,000,000

We test the local max norm given in (7) with $\zeta \in \{0, 0.05, 0.1, 0.15, 0.2\}$ and $\tau \in \{0, 0.05, 0.1\}$. We also test $\tau = 1$ (the max norm—here ζ is arbitrary) and $\zeta = 1, \tau = 0$ (the uniform trace norm). We follow the test protocol of [6], with a rank-30 approximation to the optimization problem (10).

Table 2 shows root mean squared error (RMSE) for the experiments. For both the MovieLens and Netflix data, the local max norm with $\tau = 0.05$ and $\zeta = 0.05$ gives strictly better accuracy than any previously-known norm studied in this setting. (In practice, we can use a validation set to reliably select good values for the τ and ζ parameters[4].) For the MovieLens data, the local max norm achieves RMSE of 0.7822, compared to 0.7831 achieved by the smoothed empirically-weighted trace norm with $\zeta = 0.10$, which gives the best result among the previously-known norms. For the Netflix dataset the local max norm achieves RMSE of 0.9090, improving upon the previous best result of 0.9095 achieved by the smoothed empirically-weighted trace norm [6].

6 Summary

In this paper, we introduce a unifying family of matrix norms, called the "local max" norms, that generalizes existing methods for matrix reconstruction, such as the max norm and trace norm. We examine some interesting sub-families of local max norms, and consider several different options for interpolating between the trace (or smoothed weighted trace) and max norms. We find norms lying strictly between the trace norm and the max norm that give improved accuracy in matrix reconstruction for both simulated data and real movie ratings data. We show that regularizing with any local max norm is fairly simple to optimize, and give a theoretical result suggesting improved matrix reconstruction using new norms in this family.

Acknowledgements

R.F. is supported by NSF grant DMS-1203762. R.S. is supported by NSERC and Early Researcher Award.

[3] For Netflix, the test set we use is their "qualification set", designed for a more uniform distribution of ratings across users relative to the training set. For MovieLens, we choose our test set at random from the available data.

[4] To check this, we subsample half of the test data at random, and use it as a validation set to choose (ζ, τ) for each method (as specified in Table 1). We then evaluate error on the remaining half of the test data. For MovieLens, the local max norm gives an RMSE of 0.7820 with selected parameter values $\zeta = \tau = 0.05$, as compared to an RMSE of 0.7829 with selected smoothing parameter $\zeta = 0.10$ for the smoothed weighted trace norm. For Netflix, the local max norm gives an RMSE of 0.9093 with $\zeta = \tau = 0.05$, while the smoothed weighted trace norm gives an RMSE of 0.9098 with $\zeta = 0.05$. The other tested methods give higher error on both datasets.

References

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

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

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

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

[5] 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.

[6] R. Foygel, R. Salakhutdinov, O. Shamir, and N. Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. *Advances in Neural Information Processing Systems*, 24, 2011.

[7] 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.

[8] E. Hazan, S. Kale, and S. Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. *25th Annual Conference on Learning Theory (COLT)*, 2012.

[9] M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference*, volume 6, pages 4734–4739, 2002.

[10] N. Srebro, J.D.M. Rennie, and T.S. Jaakkola. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems*, 18, 2005.

[11] J.D.M. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

[12] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20:1257–1264, 2008.

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

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

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