Rank, Trace-Norm and Max-Norm

Nathan Srebro[1] and Adi Shraibman[2]

[1] University of Toronto Department of Computer Science, Toronto ON, CANADA
[2] Hebrew University Institute of Computer Science, Jerusalem, ISRAEL
nati@cs.toronto.edu, adidan@cs.huji.ac.il

Abstract. We study the rank, trace-norm and max-norm as complexity measures of matrices, focusing on the problem of fitting a matrix with matrices having low complexity. We present generalization error bounds for predicting unobserved entries that are based on these measures. We also consider the possible relations between these measures. We show gaps between them, and bounds on the extent of such gaps.

1 Introduction

Consider the problem of approximating a noisy (or partially observed) target matrix Y with another matrix X. This problem arises often in practice, e.g. when analyzing tabulated data such as gene expressions, word counts in a corpus of documents, collections of images, or user preferences on a collection of items.

A common general scheme for solving such problems is to select a matrix X that minimizes some combination of the *complexity* of X and the *discrepancy* between X and Y. The heart of the matter is the choice of the measure of complexity for X and the measure of discrepancy between X and Y.

The most common notion of complexity of a matrix in such tasks is its rank (as in PCA, Latent Semantic Analysis, the Aspect Model and a variety of other factor models and generalizations of these approaches). Recently, the *trace-norm* and *max-norm* were suggested as alternative measures of complexity with strong connections to maximum-margin linear classification [1]. Whereas bounding the rank corresponds to constraining the *dimensionality* of each row of U and V in a factorization $X = UV'$, bounding the trace-norm and max-norm corresponds to constraining the *norms* of rows of U and V (average row-norm for the trace-norm, and maximal row-norm for the max-norm). Unlike low-rank factorizations, such constraints lead to *convex* optimization problems.

In this paper we study the rank, trace-norm and max-norm as measures of matrix complexity, concentrating on the implications to the problem mentioned above.

We begin by considering the problem of predicting unknown entries in a partially observed matrix Y (as in collaborative prediction). We assume the prediction is made by choosing a matrix X for which some combination of the discrepancy between X and Y on the one hand, and the complexity of X on the other hand, is minimized. We present generalization error bounds for general measures of discrepancy and for the cases where the complexity measure for

X is either rank (Section 3.1, repeating a previous analysis [2]), trace-norm or max-norm (Sections 3.2 and 3.3, elaborating on and proving previously quoted bounds [1]). We make no assumptions about the matrix Y, other than that the observed entries are chosen at random. The bounds, and the complexity measures used to obtain them (cardinality, pseudodimension and Rademacher complexity), are insightful in comparing the three measures of matrix complexity we are considering.

In addition to results about generic measures of discrepancy, we also specifically consider binary target matrices: For $Y \in \pm 1^{n \times m}$, we study the minimum rank, max-norm and (normalized) trace-norm of a matrix X such that $X_{ij} Y_{ij} \geq 1$ for all i, j. We refer to these as the dimensional-complexity $\mathrm{dc}(Y)$, max-complexity $\mathrm{mc}(Y)$ and trace-complexity $\mathrm{tc}(Y)$ of a binary matrix Y.

We study relations between the three matrix complexity measures. Matrices that can be approximated by a matrix of low max-norm can also be approximated by a matrix with low rank. In Section 4 we show this for general measures of discrepancy, generalizing previous results [3, 4] for binary target matrices. But this relationship is not reversible: We give examples of explicit binary matrices with low dimensional-complexity that have high max-complexity. Previously, examples in which the max-complexity is a polynomial function of the dimensional-complexity [5], or where the dimensional-complexity is constant but the max-complexity is logarithmic in the matrix size [4] have been shown. We present an explicit construction establishing that the max-complexity is not bounded by any polynomial of the dimensional-complexity and the logarithm of the matrix size.

Similarly we give examples of matrices with low trace-complexity but high dimensional-complexity and max-complexity. This gap is related to a requirement for uniform sampling of observed entries, which we show to be necessary for generalization error bounds based on the trace-norm but not on the max-norm or rank. We also show that the gap we obtain is the largest possible gap, establishing a first lower bound on the trace-complexity in terms of the max-complexity or dimensional-complexity (Section 5).

Embedding Classifiers as Linear Separators The dimensional-complexity and max-complexity have been studied in the context of embedding concept classes as low-dimensional, or large-margin, linear separators. A concept class $\mathcal{H} = \{h : \Phi \to \pm 1\}$ of binary valued functions can be represented as a $|\Phi| \times |\mathcal{H}|$ matrix Y, with $Y_{\phi,h} = h(\phi)$. The dimensional-complexity of Y is the minimum d such that each $\phi \in \Phi$ can be embedded as a point $u_\phi \in \mathbb{R}^d$ and each classifier $h \in \mathcal{H}$ can be embedded as a separating homogeneous hyperplane determined by its normal $v_h \in \mathbb{R}^d$, such that $h(\phi) = \operatorname{sign} v'_h u_\phi$. The max-complexity is the smallest M such that Φ can be embedded as points and $\mathcal{H}$ as linear separators in an infinite dimensional unit ball, where all separators separate with a margin of at least $1/M$, i.e. $\frac{|v'_h u_\phi|}{|v_h|} \geq 1/M$. Studying linear separators (in particular using kernel methods) as a generic approach to classification leads one to ask what concept classes can or cannot be embedded as low-dimensional or large-margin linear separators; that is, what matrices have high dimensional-complexity and max-complexity [4, 6].

These questions are existential questions, aimed at understanding the limits of kernel-based methods. Here, the concept class of interest is the class of matrices themselves, and we apply much of the same techniques and results in order to understand the performance of a concrete learning problem.

2 Preliminaries

Notation For vectors, $|v|_p$ is the l_p norm and $|v| = |v|_2$. For matrices, $\|X\|_{\text{Fro}} = \sqrt{\sum_{ij} X_{ij}^2}$ is the Frobenius norm; $\|X\|_2 = \max_{|u|=|v|=1} u'Xv$ is the spectral norm and is equal to the maximum singular value of X; $\|X\|_{2\to\infty} = \max_{|u|_2=1} |Xu|_\infty = \max_i |X_{i\cdot}|$ is the maximum row norm of X; $|X|_\infty = \max_{ij} |X_{ij}|$.

Discrepancy We focus on element-wise notions of discrepancy between two $n \times m$ matrices Y and X: $\mathcal{D}(X;Y) = \frac{1}{nm}\sum_{ij} g(X_{ij};Y_{ij})$, where $g(x;y)$ is some loss function. The *empirical* discrepancy for a subset $S \subset [n]\times[m]$ of the observed entries of Y is $\mathcal{D}_S(X;Y) = \frac{1}{|S|}\sum_{ij\in S} g(X_{ij};Y_{ij})$. The discrepancy relative to a distribution $\mathcal{P}$ over entries in the matrix (i.e. over $[n] \times [m]$) is $\mathcal{D}_{\mathcal{P}}(X;Y) = \mathbf{E}_{ij\sim\mathcal{P}}[g(X_{ij};Y_{ij})]$.

Since the norms are scale-sensitive measures of complexity, the scale in which the loss function changes is important. This is captured by Lipschitz continuity: A loss function $g : \mathbb{R} \times Y \to \mathbb{R}$ is L-Lipschitz if for every y, x_1, x_2, $|g(x_1;y) - g(x_2;y)| \leq L|x_1 - x_2|$.

For the special case of binary target matrices $Y \in \{\pm 1\}^{n\times m}$, the discrepancy with respect to the sign-agreement zero-one error is the (normalized) Hamming distance between $\operatorname{sign} X$ and $\operatorname{sign} Y$. It will be useful to consider the set of matrices whose sign patterns agree with the target matrix: $\mathrm{SP}(Y) = \{X | \operatorname{sign} X = \operatorname{sign} Y\}$. For scale-dependent (e.g. norm-based) complexity measures of X, considering the signs of entries in X is no longer enough, and their magnitudes must also be bounded. We consider $\mathrm{SP}^1(Y) = \{X | \forall_{ij} X_{ij}Y_{ij} \geq 1\}$, corresponding to a *margin* sign-agreement error.

Complexity The **rank** of a matrix X is the minimum k such that $X = UV'$, $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{m\times k}$. The *dimensional-complexity* of a sign matrix is:

$$\mathrm{dc}(Y) \doteq \min\{\operatorname{rank} X | X \in \mathrm{SP}(Y)\} = \min\{\operatorname{rank} X | X \in \mathrm{SP}^1(Y)\} \tag{1}$$

The **max-norm** (also known as the γ_2-norm [7]) of a matrix X is given by:

$$\|X\|_{\max} \doteq \min_{X=UV'} \|U\|_{2\to\infty} \|V\|_{2\to\infty} \tag{2}$$

While the rank constrains the dimensionality of rows in U and V, the max-norm constrains the norms of all rows in U and V. The **max-complexity** for a sign matrix Y is $\mathrm{mc}(Y) \doteq \min\{\|X\|_{\max} | X \in \mathrm{SP}^1(Y)\}$

The **trace-norm**[3] $\|X\|_\Sigma$ is the sum of the singular values of X (i.e. the roots of the eigenvalues of XX^t).

Lemma 1. $\|X\|_\Sigma = \min_{X=UV'} \|U\|_{Fro} \|V\|_{Fro} = \min_{X=UV'} \frac{1}{2}(\|U\|_{Fro}^2 + \|V\|_{Fro}^2)$

While the max-norm constrains the maximal norm of the rows in U and V, the trace-norm constrains the *sum* of the norms of the rows in U and V. That is, the max-norm constrains the norms uniformly, while the trace-norm constrains them on average. The **trace-complexity** of a sign matrix Y is $\mathrm{tc}(Y) \doteq \min\{\|X\|_\Sigma/\sqrt{nm} | X \in \mathrm{SP}^1(Y)\}$.

Since the maximum is greater than the average, the trace-norm is bounded by the max-norm: $\|X\|_\Sigma/\sqrt{nm} \leq \|X\|_{\max}$ and $\mathrm{tc}(Y) \leq \mathrm{mc}(Y)$. In Section 5 we see that there can be a large gap between $\|X\|_\Sigma/\sqrt{nm}$ and $\|X\|_{\max}$.

Extreme Values For any sign matrix Y, $1 \leq \mathrm{tc}(Y) \leq \mathrm{mc}(Y) \leq \|Y\|_{\max} \leq \sqrt{n}$. Rank-one sign matrices Y have $\mathrm{dc}(Y) = \mathrm{mc}(Y) = \mathrm{tc}(Y) = 1$ and are the only sign matrices for which any of the three quantities is equal to one. To obtain examples of matrices with high trace-complexity, note that:

Lemma 2. *For any* $Y \in \{\pm 1\}^{n \times m}$, $tc(Y) \geq \sqrt{nm}/\|Y\|_2$.

Proof. Let $X \in \mathrm{SP}(Y)$ s.t. $\|X\|_\Sigma = \sqrt{nm}\mathrm{tc}(Y)$, then by the duality of the spectral norm and the trace-norm, $\|X\|_\Sigma \|Y\|_2 \geq \sum_{ij} X_{ij} Y_{ij} \geq nm$. □

An example of a sign matrix with low spectral norm is the Hadamard matrix $H_p \in \{\pm 1\}^{2^p \times 2^p}$, where H_{ij} is the inner product of i and j as elements in $GF(2^p)$. Using $\|H_p\|_2 = 2^{p/2}$ we get $\mathrm{mc}(H_{\log n}) = \mathrm{tc}(H_{\log n}) = \sqrt{n}$ [5]. Although counting arguments prove that for any n, there exist $n \times n$ sign matrices for which $\mathrm{dc}(Y) > n/11$ (Lemma 3 below, following Alon *et al* [8] who give a slightly weaker bound), the Hadamard matrix, for which it is known that $\sqrt{n} \leq \mathrm{dc}(H_{\log n}) \leq n^{0.8}$ [6], is the most extreme known concrete example.

3 Generalization Error Bounds

Consider a setting in which a random subset S of the entries of Y is observed. Based on the observed entries Y_S we would like to predict unobserved entries in Y. This can be done by fitting a low-complexity matrix X to Y_S and using X to predict unobserved entries. We present generalization error bounds on the overall discrepancy in terms of the observed discrepancy. The bounds do *not* assume any structure or probabilistic assumption on Y, and hold for any (adversarial) target matrix Y. What is assumed is that the sample S is chosen at random.

We are interested in predicting unobserved entries not only as an application of matrix learning (e.g. when predicting a user's preferences based on preferences of the user and other users, or completing missing experimental data), but also as a conceptual learning task where the different measures of complexity can be compared and related. Even when learning is done for some other purpose

[3] Also known as the *nuclear norm* and the *Ky-Fan n-norm*.

arbitrary source distribution	⇔	target matrix Y
random training set	⇔	random set S of observed entries
hypothesis	⇔	concept matrix X
training error	⇔	observed discrepancy $\mathcal{D}_S(X;Y)$
generalization error	⇔	true discrepancy $\mathcal{D}(X;Y)$

Fig. 1. Correspondence with post-hoc bounds on the generalization error for standard feature-based prediction tasks

(e.g. understanding structure or reconstructing a latent signal), the ability of the model to predict held-out entries is frequently used as an ad-hoc indicator of its fit to the true underlying structure. Bounds on the generalization ability for unobserved entries can be used as a theoretical substitute to such measures (with the usual caveats of using generalization error bounds).

The Pseudodimension and the Rademacher Complexity To obtain generalization error bounds, we consider matrices as functions from index pairs to entry values, and calculate the pseudodimension of the class of low-rank matrices and the Rademacher complexity of the classes of low max-norm and low trace-norm matrices. Recall that:

Definition 1. *A class $\mathcal{F}$ of real-valued functions* pseudo-shatters *the points $x_1, \ldots, x_n$ with thresholds $t_1, \ldots, t_n$ if for every binary labeling of the points $(s_1, \ldots, s_n) \in \{+,-\}^n$ there exists $f \in \mathcal{F}$ s.t. $f(x_i) \leq t_i$ iff $s_i = -$. The* pseudodimension *of a class $\mathcal{F}$ is the supremum over n for which there exist n points and thresholds that can be shattered.*

Definition 2. *The empirical Rademacher complexity of a class $\mathcal{F}$ over a specific sample $S = (x_1, x_2, \ldots)$ is given by: $\hat{R}_S(\mathcal{F}) = \frac{2}{|S|}\mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}} \left|\sum_i \sigma_i f(x_i)\right|\right]$, where the expectation is over the uniformly distributed random signs σ_i.*

The Rademacher complexity with respect to a distribution D is the expectation, over a sample of $|S|$ points drawn i.i.d. from D: $R^D_{|S|}(\mathcal{F}) = \mathop{\mathbf{E}}_S\left[\hat{R}_S(\mathcal{F})\right]$.

It is well known how to obtain Generalization error bounds in terms of the pseudodimension and Rademacher complexity. Our emphasis is on calculating the pseudodimension and the Rademacher complexity. We do not present the tightest possible bounds in terms of these measures.

3.1 Low-Rank Matrices

Generalization error bounds for prediction with low-rank matrices can be obtained by considering the number of sign configurations of low-rank matrices [2] (following techniques introduced in [8]):

Lemma 3 ([9]). $|\{Y \in \{\pm 1\}^{n\times m} | dc(Y) \leq k\}| \leq (8em/k)^{k(n+m)}$

This bound is tight up to a multiplicative factor in the exponent: for $m > k^2$, $|\{Y \in \{\pm 1\}^{n\times m} | \mathrm{dc}(Y) \leq k\}| \geq m^{\frac{1}{2}(k-1)n}$.

Using the bound of Lemma 3, a union bound of Chernoff bounds yields a generalization error bound for the zero-one sign agreement error (since only signs of entries in X are relevant). Generalization error bounds for other loss functions can be obtained by using a similar counting argument to bound the pseudodimension of the class $\mathcal{X}^k = \{X | \operatorname{rank} X \leq k\}$. To do so, we need to bound not only the number of sign configurations of such matrices, but the number of sign configurations relative to any threshold matrix T:

Lemma 4 ([2]). $\forall_{T\in\mathbb{R}^{n\times m}} |\{\operatorname{sign}(X-T) | \operatorname{rank} X \leq k\}| \leq \left(\frac{8em}{k}\right)^{k(n+m)}$

Corollary 1. *pseudodimension*$(\mathcal{X}^k) \leq k(n+m)\log\frac{8em}{k}$

Theorem 1 ([2]). *For any monotone loss function bounded by M,* any $n \times m$ *matrix Y,* any *distribution $\mathcal{P}$ of index pairs (i,j), $n, m > 2$, $\delta > 0$ and integer k, with probability at least $1-\delta$ over choosing a set S of $|S|$ index pairs according to $\mathcal{P}$, for all matrices X with* $\operatorname{rank} X \leq k$*:*

$$\mathcal{D}_{\mathcal{P}}(X;Y) < \mathcal{D}_S(X;Y) + 6\sqrt{\frac{k(n+m)\log\frac{8em}{k}\log\frac{M|S|}{k(n+m)} - \log\delta}{|S|}}$$

3.2 Low Trace-Norm Matrices

In order to calculate the Rademacher complexity of the class $\mathcal{X}[M] = \{X | \, \|X\|_{\Sigma} \leq M\}$, we observe that this class is convex and that any unit-trace-norm matrix is a convex combination of unit-norm rank-one matrices $X = \sum D_{aa}(U_{\cdot a}V'_{\cdot a})$, where $X = UDV'$ is the SVD and $U_{\cdot a}, V_{\cdot a}$ are columns of U, V. Therefore, $\mathcal{X}[1] = \operatorname{conv}\mathcal{X}_1[1]$, where $\mathcal{X}_1[1] \doteq \{uv' \mid u \in \mathbb{R}^n, v \in \mathbb{R}^m, |u| = |v| = 1\}$ is the class of unit-norm rank-one matrices. We use the fact that the Rademacher complexity does not change when taking convex combinations, and calculate the Rademacher complexity of $\mathcal{X}_1[1]$. We first analyze the empirical Rademacher complexity for any fixed sample S, possibly with repeating index pairs. We then bound the average Rademacher complexity for a sample of $|S|$ index pairs drawn uniformly at random from $[n] \times [m]$ (with repetitions).

The Empirical Rademacher Complexity For an empirical sample $S = \{(i_1, j_1), (i_2, j_2), \ldots\}$ of $|S|$ index pairs, the empirical Rademacher complexity of rank-one unit-norm matrices is the expectation:

$$\hat{R}_S(\mathcal{X}_1[1]) = \mathbf{E}_\sigma\left[\sup_{|u|=|v|=1}\left|\frac{2}{|S|}\sum_{\alpha=1}^{|S|}\sigma_\alpha u_{i_\alpha} v_{j_\alpha}\right|\right] \tag{3}$$

where σ_α are uniform ± 1 random variables. For each index pair (i,j) we will denote by s_{ij} the number of times it appears in the empirical sample S, and consider the random variables $\sigma_{ij} = \sum_{\alpha \text{ s.t. } (i_\alpha, j_\alpha)=(i,j)} \sigma_\alpha$.

Since the variables σ_α are independent, $\mathbf{E}\left[\sigma_{ij}^2\right] = s_{ij}$, and we can calculate:

$$\hat{R}_S(\mathcal{X}_1[1]) = \mathbf{E}_\sigma\left[\sup_{|u|,|v|=1}\left|\tfrac{2}{|S|}\sum_{i,j}\sigma_{ij}u_iv_j\right|\right] = \tfrac{2}{|S|}\mathbf{E}_\sigma\left[\sup_{|u|,|v|=1}|u'\sigma v|\right] = \tfrac{2\mathbf{E}_\sigma\left[\|\sigma\|_2\right]}{|S|}$$

where σ is an $n \times m$ matrix of σ_{ij}.

The Rademacher complexity is equal to the expectation of the spectral norm of the random matrix σ (with a factor of $\frac{2}{|S|}$). Using the Frobenius norm to bound the spectral norm, we have:

$$\hat{R}_S(\mathcal{X}_1[1]) \leq \frac{2}{|S|}\mathbf{E}_\sigma[\|\sigma\|_{\text{Fro}}] \leq \frac{2}{|S|}\sqrt{|S|} = \frac{2}{\sqrt{|S|}} \tag{4}$$

As a supremum over all sample sets S, this bound is tight: consider a sample of $|S|$ index pairs, all in the same column. The rank-one unit-norm matrix attaining the supremum would match the signs of the matrix with $\pm 1/\sqrt{|S|}$ yielding an empirical Rademacher complexity of $2/\sqrt{|S|}$. The form of (4) is very disappointing, and does not lead to meaningful generalization error bounds.

Even though the empirical Rademacher complexity for a specific sample might be very high, in what follows we show that the *expected* Rademacher complexity, for a uniformly chosen sample, is low. Using the Frobenius norm to bound the Spectral norm of σ will no longer be enough, and in order to get a meaningful bound we must analyze the expected spectral norm more carefully.

Bounding $\mathbf{E}_{\boldsymbol{\sigma}}[\|\boldsymbol{\sigma}\|_{\mathbf{2}}]$ In order to bound the expected spectral norm of σ, we apply Theorem 3.1 of [10], which bounds the expected spectral norm of matrices with entries of fixed magnitudes but random signs in terms of the maximum row and column magnitude norms. If S contains no repeated index pairs ($s_{ij} =$ 0 or 1), we are already in this situation, as the magnitudes of σ are equal to s. When some index pairs are repeated, we consider a different random matrix, $\tilde{\sigma}_{ij} = \epsilon_{ij}s_{ij}$, where ϵ_{ij} are i.i.d. unbiased signs. Using $\tilde{\sigma}$ instead of σ gives us an upper bound on the empirical Rademacher complexity (Lemma 12 from the Appendix). Applying Theorem 3.1 of [10] to $\tilde{\sigma}_{ij}$, we obtain:

$$\hat{R}_S(\mathcal{X}_1[1]) \leq \frac{2}{|S|}\mathbf{E}_\epsilon[\|\tilde{\sigma}\|_2]\frac{2}{|S|} \leq K(\ln m)^{\frac{1}{4}}\left(\max_i |s_{i\cdot}| + \max_j |s_{\cdot j}|\right) \tag{5}$$

where $|s_{i\cdot}|$ and $|s_{\cdot j}|$ are norms of row and column vectors of the matrix s, and K is the absolute constant guaranteed by Theorem 3.1 of [10].

Bounding the Row and Column Norms For the worst samples, the norm of a single row or column vector of s might be as high as $|S|$, but for random uniformly drawn samples, we would expect the row and column norms to be roughly $\sqrt{|S|/n}$ and $\sqrt{|S|/m}$. To make this estimate precise we proceed in two steps[4]. We first use Bernstein's inequality to bound the maximum value of s_{ij},

[4] We assume here $nm > |S| > n \geq m > 3$. See [9] for more details.

uniformly over all index pairs: $\Pr_S(\max_{ij} s_{ij} > 9 \ln n) \leq \frac{1}{|S|}$. When the maximum entry in s is bounded, the norm of a row can be bounded by the square root of the number of observations in the row. In the second step we use Bernstein's inequality again to bound the expected maximum number of observations in a row (similarly column) by $6(\frac{|S|}{n} + \ln |S|)$. Combining these results we can bound the Rademacher complexity, for a random sample set where each index pair is chosen uniformly and independently at random:

$$\begin{aligned}
&R^{\text{uniform}}_{|S|}(\mathcal{X}_1[1]) = \mathbf{E}_S\left[\hat{R}_S(\mathcal{X}_1[1])\right] \\
&\leq \Pr\left(\max_{ij} s_{ij} > 9\ln n\right) \sup_S \hat{R}_S(\mathcal{X}_1[1]) + \mathbf{E}_S\left[\hat{R}_S(\mathcal{X}_1[1]) \,\middle|\, \max_{ij} s_{ij} \leq 9\ln n\right] \\
&\leq \frac{1}{|S|} \cdot \frac{2}{\sqrt{|S|}} + \frac{2}{|S|} K(\ln m)^{\frac{1}{4}} \mathbf{E}_S\left[\max_i |s_{i\cdot}| + \max_j |s_{\cdot j}| \,\middle|\, \max_{ij} s_{ij} \leq 9\ln n\right] \\
&\leq \frac{2}{|S|^{3/2}} + \frac{2K(\ln m)^{\frac{1}{4}}}{|S|}\sqrt{9\ln n}\left(\sqrt{6(\frac{|S|}{n} + \ln|S|)} + \sqrt{6(\frac{|S|}{m} + \ln|S|)}\right) \qquad (6)
\end{aligned}$$

Taking the convex hull, scaling by M and rearranging terms:

Theorem 2. *For some universal constant K, the expected Rademacher complexity of matrices of trace-norm at most M, over uniform samplings of index pairs is at most (for $|S|/\ln n \geq n \geq m$): $R^{uniform}_{|S|}(\mathcal{X}[M]) \leq K \frac{M}{\sqrt{nm}} \sqrt{\frac{(n+m)\ln^{3/2} n}{|S|}}$*

Applying Theorem 2 of [11][5]:

Theorem 3. *For any L-Lipschitz loss function, target matrix Y, $\delta > 0$, $M > 0$ and sample sizes $|S| > n \log n$, and for a* uniformly *selected sample S of $|S|$ entries in Y, with probability at least $1-\delta$ over the sample selection, the following holds for all matrices $X \in \mathbb{R}^{n \times m}$ with $\frac{\|X\|_\Sigma}{\sqrt{nm}} \leq M$:*

$$\mathcal{D}(X;Y) < \mathcal{D}_S(X;Y) + KL\sqrt{\frac{M^2(n+m)\ln^{3/2} n - \log\delta}{|S|}}$$

Where K is a universal constant that does not depend on Y,n,m, the loss function, or any other quantity.

3.3 Low Max-Norm Matrices

Since the max-norm gives us a bound on the trace-norm, we can apply Theorems 2 and 3 also to matrices of bounded max-norm. However, when the max-norm is

[5] By bounding the zero-one sign-agreement error with the 1-Lipschitz function $g(x, y) = \max(0, \min(yx - 1, 1))$, which in turn is bounded by the margin sign-agreement error, generalization error bounds in terms of the margin can be obtained from bounds in terms of the Lipschitz constant.

bounded it is possible to obtain better bounds, avoiding the logarithmic terms, and more importantly, bounds that hold for *any* sampling distribution.

As we did for low trace-norm matrices, we bound the Rademacher complexity of low max-norm matrices by characterizing the unit ball of the max-norm $\mathcal{B}_{\max} = \{X | \|X\|_{\max} \leq 1\}$ as a convex hull. Unlike the trace-norm unit ball, we cannot exactly characterize the max-norm unit ball as a convex hull. However, using Grothendiek's Inequality we can bound the unit ball with the convex hull of rank-one sign matrices $\mathcal{X}_\pm = \{X \in \{\pm 1\}^{n\times m} | \operatorname{rank} X = 1\}$.

Theorem 4 (Grothendieck's Inequality [12, page 64]). *There is an absolute constant* $1.67 < K_G < 1.79$ *such that the following holds: Let* A_{ij} *be a real matrix, and suppose that* $|\sum_{i,j} A_{ij} s_i t_j| \leq 1$ *for every choice of reals with* $|s_i|, |t_j| \leq 1$ *for all* i, j*. Then* $\left|\sum_{i,j} a_{ij} \langle x_i, y_j\rangle\right| \leq K_G$*, for every choice of unit vectors* x_i, y_j *in a real Hilbert space.*

Corollary 2. $\mathrm{conv}\mathcal{X}_\pm \subset \mathcal{B}_{\max} \subset K_G \mathrm{conv}\mathcal{X}_\pm$

Proof. Noting that the dual norm to the max-norm is:

$$\|A\|^*_{max} = \max_{\|B\|_{\max}\leq 1} \langle A, B\rangle = \max_{x_i, y_j \in \mathbb{R}^k : |x_i|, |y_j| \leq 1} \sum_{i,j} a_{ij} x_i' y_j. \tag{7}$$

where the maximum is over any k, we can restate Grothendieck's inequality as $\|A\|^*_{max} \leq K_G \|A\|_{\infty\to 1}$ where $\|A\|_{\infty\to 1} = \max_{s_i, t_j \in \mathbb{R} : |s_i|, |t_j| \leq 1} \sum_{i,j} a_{ij} s_i t_j$. We also have $\|A\|_{\infty\to 1} \leq \|A\|^*_{max}$, and taking the duals:

$$\|A\|^*_{\infty\to 1} \geq \|A\|_{max} \geq K_G \|A\|^*_{\infty\to 1} \tag{8}$$

We now note that $\|A\|_{\infty\to 1} = \max_{B\in\mathcal{X}_\pm} \langle A, B\rangle$ and so $\mathcal{X}_\pm$ is the unit ball of $\|A\|^*_{\infty\to 1}$ and (8) establishes the Corollary. □

The class of rank-one sign matrices is a finite class of size $|\mathcal{X}_\pm| = 2^{n+m-1}$, and so its empirical Rademacher complexity (for any sample) can be bounded by $\hat{R}_S(\mathcal{X}_\pm) < \sqrt{7\frac{2(n+m)+\log|S|}{|S|}}$ [9]. Taking the convex hull of this class and scaling by $2M$ we have (for $2 < |S| < nm$):

Theorem 5. *The Rademacher complexity of matrices of max-norm at most* M*, for any index-pair distribution, is bounded by*[6]*:* $R_{|S|}(\mathcal{X}^{\max}[M]) \leq 12M\sqrt{\frac{n+m}{|S|}}$

Theorem 6. *For any L-Lipschitz loss function, any matrix* Y*,* any *distribution* $\mathcal{P}$ *of index pairs* (i, j)*,* $n, m > 2$*,* $\delta > 0$ *and* $M > 0$*, with probability at least* $1-\delta$ *over choosing a set* S *of* $|S|$ *index pairs according to* $\mathcal{P}$*, for all matrices* X *with* $\|X\|_{\max} \leq M$*:*

$$\mathcal{D}_\mathcal{P}(X; Y) < \mathcal{D}_S(X; Y) + 17\sqrt{\frac{M^2(n+m) - \log\delta}{|S|}}$$

[6] For large enough n, m, the constant 12 can be reduced to $K_G\sqrt{8\ln 2} < 4.197$.

4 Between the Max-Norm and the Rank

We have already seen that the max-norm bounds the trace-norm, and so any low max-norm approximation is also a low trace-norm approximation. Although the max-norm does not bound the rank (e.g. the identity matrix has max-norm one but rank n), using random projections, a low max-norm matrix can be approximated by a low rank matrix [3]. Ben David *et al* [4] used this to show that $\mathrm{dc}(Y) = O(\mathrm{mc}^2(Y) \log n)$. Here, we present a slightly more general analysis, for any Lipschitz continuous loss function.

Lemma 5. *For any $X \in R^{n\times m}$ and any $\|X\|_{\max} > \epsilon > 0$, there exists X' such that $|X - X'|_\infty < \epsilon$ and* $\operatorname{rank} X \leq 9(\|X\|_{\max}/\epsilon)^2 \log(3nm)$*.*

Proof. Set $M = \|X\|_{\max}$ and let $X = UV'$ with $\|U\|^2_{2\to\infty} = \|V\|^2_{2\to\infty} = M$. Let $A \in \mathbb{R}^{k\times d}$ be a random matrix with independent normally distributed entries, then for any u, v with $u' = Auandv' = Av$ we have [3]:

$$\Pr\left(1-\varepsilon\right)|u-v|^2 \leq |u'-v'|^2 \leq \left(1+\varepsilon\right)|u-v|^2 \geq 1 - 2e^{-k(\varepsilon^2-\varepsilon^3)/4} \tag{9}$$

Set $\varepsilon = \frac{2\epsilon}{3M}$ and $k = 4\ln(3nm)/\varepsilon^2 = 9(M/\epsilon)^2 \ln(3nm)$. Taking a union bound over all pairs (U_i, V_j) of rows of U and V, as well as all pairs $(U_i, 0)$ and $(V_j, 0)$, we get that with positive probability, for all i, j, $\left|U_i' - V_j'\right|^2$, $|U_i'|^2$ and $\left|V_j'\right|^2$ are all within $(1\pm\varepsilon)$ of $|U_i - V_j|^2$, $|U_i|^2 \leq M$ and $|V_j|^2 \leq M$, respectively. Expressing $U_i'V_j'$ in terms of these norms yields $U_iV_j - 3M\varepsilon/2 \leq U_i'V_j' \leq U_iV_j + 3M\varepsilon/2$, and so $|UV' - X|_\infty \leq 3M\varepsilon/2 = \epsilon$ and $\operatorname{rank} UV \leq k = 9(M/\epsilon)^2 \ln(3nm)$. □

Corollary 3. *For any L-Lipschitz continuous loss function, any matrices X, Y, and any $\|X\|_{\max} > \epsilon > 0$, there exists X' such that $\mathcal{D}(X'; Y) \leq \mathcal{D}(X; Y) + \epsilon$ and* $\operatorname{rank} X' \leq 9\,\|X\|^2_{\max} (L/\epsilon)^2 \log(3nm)$*.*

Corollary 4. *For any sign matrix Y, $dc(Y) \leq 10mc^2(Y) \log(3nm)$.*

Proof. For $X \in \mathrm{SP}^1(Y)$, setting $\epsilon = \sqrt{0.9}$ ensures $\operatorname{sign} X' = \operatorname{sign} X = Y$. □

Using Lemma 5 and Theorem 1 it is possible to obtain a generalization error bound similar to that of Theorem 6, but with additional log-factors. More interestingly, Corollary 4 allows us to bound the number of matrices with low max-complexity[7]:

Lemma 6. $\log|\{Y \in \{\pm 1\}^{n\times m} | mc(Y) \leq M\}| < 10M^2(n+m) \log(3nm) \log(\frac{m}{M^2})$

Noting that $Y \in \{\pm 1\}^{n\times m}$ with at most M "1"s in each row has $\mathrm{mc}(Y) \leq M$ establishes that this bound is tight up to logarithmic factors:

Lemma 7. *For $M^2 < n/2$,* $\log|\{Y \in \{\pm 1\}^{n\times n} | mc(Y) \leq M\}| \geq M^2 n \log(n/M^2)$

[7] A tighter analysis, allowing the random projection to switch a few signs, can reduce the bound to $40M^2(n+m)\log^2(m/M^2)$.

A Gap Between $\mathbf{dc}(\boldsymbol{Y})$ and $\mathbf{mc}(\boldsymbol{Y})$ We have seen that $\mathrm{dc}(Y)$ can be bounded in terms of $\mathrm{mc}^2(Y)$ and that both yield similar generalization error bounds. We now consider the inverse relationship: can $\mathrm{mc}^2(Y)$ be bounded in terms of $\mathrm{dc}(Y)$?

The Hadamard matrix $H_p \in \mathbb{R}^{n\times n}$ $(n = 2^p)$ is an example of a matrix with a polynomial gap between $\mathrm{mc}^2(H_p) = n$ and $\sqrt{n} \leq \mathrm{rank}(H_p) < n^{0.8}$. This gap still leaves open the possibility of a weaker polynomial bound. The triangular matrix $T_n \in \{\pm 1\}^{n\times n}$ with $+1$ on and above the diagonal and -1 below it, exhibits a non-polynomial gap: $\mathrm{dc}(T_n) = 2$ while $\mathrm{mc}(T_n) = \theta(\log n)$ [5, Theorem 6.1]. But we may ask if there is a polynomial relation with logarithmic factors in n. In order to show that $\mathrm{mc}(Y)$ is not polynomially bounded by $\mathrm{dc}(Y)$, even with $\mathrm{poly}\log n$ factors, we examine tensor exponents[8] of triangular matrices (note that $H_1 = T_2$, and so $H_p = T_2^{\otimes p}$, up to row and column permutations).

Theorem 7. *For any $r > 0$, there exists an $n \times n$ sign matrix Y such that $mc(Y) > (dc(Y)log(n))^r$.*

To prove the Theorem, we will use the following known results:

Lemma 8. *For any four matrices A, B, C, D: $(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$.*

Theorem 8 ([5, Theorem 4.1]). *Let Y be a sign matrix, and let $Y = UDV$ be its SVD. If the matrix UV has the same signs as Y then $\frac{\|Y\|_\Sigma}{\sqrt{nm}} \leq mc(Y)$. If in addition all the rows of the matrix $U\sqrt{D}$, and all the columns of the matrix $\sqrt{D}V$ have equal length, then $\frac{\|Y\|_\Sigma}{\sqrt{nm}} = mc(Y)$.*

Theorem 9 ([5]). *Denote by T_n the triangular $n \times n$ matrix and $T_n = UDV$ its SVD decomposition, then UV is signed as T_n and all the rows of the matrix $U\sqrt{D}$, and all the columns of the matrix $\sqrt{D}V$ have equal length.*

Proof of Theorem 7 To prove the theorem we first show that if two matrices A and B satisfy the properties that are guarantied by Theorem 9 for triangular matrices, then the tensor product $A \otimes B$ also satisfies this properties. And thus tensor products of triangular matrices have these properties. This follows from the following applications of Lemma 8:

1. Let $U_A D_A V_A = A$ and $U_B D_B V_B = B$ be the SVD of A and B respectively, then $(U_A \otimes U_B)(D_A \otimes D_B)(V_A \otimes V_B)$ is the SVD of $A \otimes B$, since if v_A is a eigenvector of AA^t with eigenvalue μ_A and v_B is an eigenvector of BB^t with eigenvalue μ_B then

$$\begin{aligned}(A\otimes B)(A\otimes B)^t(v_A\otimes v_B) &= (AA^t)\otimes(BB^t)(v_A\otimes v_B)\\ &= (AA^t v_A)\otimes(BB^t v_B) = \mu_A v_A \otimes \mu_B v_B = \mu_A\mu_B(v_A\otimes v_B).\end{aligned}$$

Thus $v_A \otimes v_B$ is an eigenvector of $(A \otimes B)(A \otimes B)^t$ with eigenvalue $\mu_A\mu_B$.

[8] $A \otimes B$ and $A^{\otimes p}$ denotes tensor products and exponentiation.

2. If the matrix $U_A V_A$ has the same signs as A, and the matrix $U_B V_B$ as the same signs as B then the matrix $(U_A \otimes U_B)(V_A \otimes V_B) = (U_A V_A) \otimes (V_A V_B)$ has the same signs as $A \otimes B$, since the sings of the tensor product is determined only by the signs of the matrices in the product.
3. If the rows of $U_A \sqrt{D_A}$ have equal length and so does the rows of $U_B \sqrt{D_B}$, and equivalently the columns of $\sqrt{D_A} V_A$ and $\sqrt{D_B} V_B$, then the rows of the matrix $(U_A \otimes U_B)\sqrt{D_A \otimes D_B}$, and the columns of the matrix $\sqrt{D_A \otimes D_B}(V_A \otimes V_B)$ have equal length, since rows (equiv. columns) of $P \otimes Q$ are tensor products of rows (equiv. columns) in P and Q.

For any $t > 0$ and integer $p > 0$, let $k = 2^{2^t}$ and $n = 2^{p2^t}$ and consider $T_k^{\otimes p} \in \{\pm 1\}^{n \times n}$. By the above considerations and Theorems 8 and 9, $\mathrm{mc}(T_k^{\otimes p}) = \mathrm{mc}(T_k)^p \geq (c2^t)^p$ for some $c > 0$, while $\mathrm{dc}(T_k^{\otimes p}) = \mathrm{dc}(T_k)^p \leq 2^p$. For any $r > 0$ we can choose $t = p > \max(6r, -2\log c)$ and so:

$$(\mathrm{dc}(T_k^{\otimes p}) log(n))^r \leq 2^{r(p+t+\log p)} < 2^{2tp} < 2^{p(t+\log c)} \leq \mathrm{mc}(T_k^{\otimes p}) \qquad \square$$

Matrices with Bounded Entries We note that a large gap between the max-complexity and the dimensional-complexity is possible only when the low-rank matrix realizing the dimensional-complexity has entries of vastly varying magnitudes: For a rank-k matrix X with entries bounded by R, Awerbuch and Kleinberg's *Barycentric spanner* [13] construction can be used to obtain a factorization $X = UV', U \in R^{n \times k}, V \in R^{m \times k}$, such that the entries of U and V are bounded by $\sqrt{R}$. This establishes that $\|X\|_{\max} \leq |X|_\infty \operatorname{rank} X$. Now, if $X \in \mathrm{SP}(Y)$ with $\operatorname{rank} X = k$ and $\frac{\max_{ij} |X_{ij}|}{\min_{ij} |X_{ij}|} \leq R$, we can scale X to obtain $X' \in \mathrm{SP}^1(Y)$ with $\|X'\|_{\max} \leq |X'|_\infty \operatorname{rank} X' \leq Rk$.

5 Between the Trace-Norm and the Max-Norm or Rank

The generalization error bounds highlight an important distinction between the trace-norm and the other two measures: the trace-norm is an *on average* measure of complexity, and leads to generalization error bounds only with respect to a uniform sampling distribution. This is not an artifact of the proof techniques. To establish this, consider:

Lemma 9. *For any $k < n$ and $Y \in \{\pm 1\}^{n \times n}$ such that $Y_{ij} = 1$ for $i > k$ or $j > k$ (i.e. except on the leading $k \times k$ submatrix): $tc(Y) \leq \|Y\|_\Sigma / n \leq k^{3/2}/n + \sqrt{2}$*

Proof. Write $Y = X_1 + X_2$ where X_1 is 0 on the leading $k \times k$ submatrix and 1 elsewhere: $\|Y\|_\Sigma \leq \|X_1\|_\Sigma + \|X_2\|_\Sigma \leq \sqrt{\operatorname{rank} X_1}\, \|X_1\|_{\mathrm{Fro}} + \sqrt{\operatorname{rank} X_2}\, \|X_2\|_{\mathrm{Fro}} \leq \sqrt{k}k + \sqrt{2}n$. $\square$

Corollary 5. $|\{Y \in \{\pm 1\}^{n \times n} | tc(Y) \leq M\}| \geq 2^{((M-\sqrt{2})n)^{4/3}}$

Consider fitting an $n\times n$ binary target matrix, where entries are sampled only in the leading $n^{2/3}\times n^{2/3}$ submatrix. A matrix X with $\|X\|_{\Sigma}/n < 3$ is sufficient to get all possible values in the submatrix, and so even with $|S| = \Theta(n^{4/3})$ we cannot expect to generalize even when $\|X\|_{\Sigma}/n$ is constant.

Using Lemma 9 we can also describe matrices Y with large gaps between $\mathrm{tc}(Y)$ and both $\mathrm{mc}(Y)$ and $\mathrm{dc}(Y)$. An $n \times n$ sign matrix with a Hadamard matrix in the leading $k \times k$ subspace and ones elsewhere provides an example where $\mathrm{mc}(Y) = \Theta((\mathrm{tc}(Y)n)^{1/3})$, e.g. $\mathrm{tc}(Y) < 3$ and $\mathrm{tc}(Y) = n^{1/3}$. Counting arguments ensure a similar gap with $\sqrt{\mathrm{dc}(Y)}$. We show that this gap is tight:

Theorem 10. *For every $n \times n$ sign matrix Y, $mc(Y) \leq 3(tc(Y)n)^{1/3}$.*

Recall that $1 \leq \mathrm{tc}(Y) \leq \mathrm{mc}(Y) \leq \sqrt{n}$. The bound in meaningful even for matrices with large $\mathrm{tc}(Y)$, up to $\sqrt{n}/27$. To prove the Theorem, we first show:

Lemma 10. *Let $X \in \mathbb{R}^{n\times n}$ with $\|X\|_{\Sigma} = M$, then X can be expressed as $X = B + R + C$, where $\|B\|_{\max} \leq (M^{1/3}$, R has at most $M^{2/3}$ rows that are non-zero and C has at most $M^{2/3}$ columns that are non-zero. Furthermore, for every i, j, at most one of B_{ij}, R_{ij} and C_{ij} is non-zero.*

Proof. Let $X = UV'$ be a decomposition of X s.t. $\|U\|^2_{\mathrm{Fro}} = \|V\|^2_{\mathrm{Fro}} = M$. At most $M^{2/3}$ of the rows of U and $M^{2/3}$ of the rows of V have squared norms greater than $M^{1/3}$. Let $R_{ij} = X_{ij}$ when $|U_i|^2 > M^{1/3}$ and zero otherwise. Let $C_{ij} = X_{ij} - R_{ij}$ when $|V_j|^2 > M^{1/3}$, zero otherwise. Let $B = X - R - C$. Zeroing the rows of U and V with squared norms greater than $M^{1/3}$ leads to a factorization of B with maximal squared row-norm $M^{1/6}$, establishing $\|B\|_{\max} \leq M^{1/3}$. □

To prove the Theorem, let $X \in \mathrm{SP}^1(Y)$ with $\mathrm{tc}(Y) = \|X\|_{\Sigma}/n$ and let $X = B + R + C$ as in Lemma 10, and note that $B + \operatorname{sign} R + \operatorname{sign} C \in \mathrm{SP}^1(Y)$ ($\operatorname{sign} R, \operatorname{sign} C$ are zero where R, C are zero). Writing $(\operatorname{sign} R) = I(\operatorname{sign} R)$ establishes $\|\operatorname{sign} R\|_{\max} \leq \|I\|_{2\to\infty} \|\operatorname{sign} R\|_{2\to\infty} = 1\sqrt{\|X\|^{2/3}_{\Sigma}} = \|X\|^{1/3}_{\Sigma}$ and similarly $\|\operatorname{sign} C\|_{\max} \leq \|X\|^{1/3}_{\Sigma}$. Using the convexity of the max-norm:

$$mc(Y) \leq \|B + \operatorname{sign} R + \operatorname{sign} C\|_{\Sigma} \leq 3\,\|X\|^{1/3}_{\Sigma} = 3(n\mathrm{tc}(Y))^{1/3} \qquad \square$$

Since $\mathrm{dc}(Y) = O(\mathrm{mc}^2(Y)log(n))$, Theorem 10 also provides a tight (up to log factors) bounds on the possible gap between dc and tc.

Using Lemma 6, Theorem 10 provides a non-trivial upper bound on the number of sign matrices with low trace-complexity, but a gap of $\sqrt[3]{M^2/n}$ still remains between this upper bound and the lower bound of Corollary 5:

Corollary 6. $\log|\{Y|tc(Y) \leq M\}| < 7M^{2/3}n^{5/3}\log(3nm)\log(n/M^2)$

6 Discussion

The initial motivation for the study reported here was to obtain a better understanding and a theoretical foundation for “Maximum Margin Matrix Factorization” (MMMF) [1], i.e. learning with low trace-norm and low max-norm matrices. We see as the main product of this study not the generalization error bounds as numerical bounds, but rather the relationships between the three measures, and the way in which they control the "complexity", as measured in terms of their generalization ability. The generalization error bounds display the similar roles of $\operatorname{rank} X$, $\|X\|_{\max}^2$ and $\|X\|_{\Sigma}^2 / nm$ in controlling complexity and highlight the main difference between the trace-norm and the other two measures. We note the interesting structure of the two hierarchies of classes of low dimensional-complexity and max-complexity matrices: Any class of matrices with bounded max-complexity is a subset of a class of matrices with bounded dimensional-complexity of "roughly" the same size (logarithm of size differs only by logarithmic factors). But this class of bounded dimensional-complexity matrices includes matrices with very high max-complexity.

Open Issues Although we show that the dimensional-complexity can not bound the max-complexity, it might still be the case that changing a few entries of a low-dimensional-complexity matrix is enough to get to to a low-max-complexity matrix. Beyond sign matrices, we can ask whether for any X and ϵ there exists X' with $\|X'\|_{\max}^2 \leq O(\operatorname{rank} X (1/\epsilon)^2 \operatorname{poly} \log n)$ and $\delta(X, X') \leq \epsilon$ for some error measure δ. Theorem 7 precludes this possibility for $\delta(X, X') = |X - X'|_{\infty}$, but it is possible that such a relationship holds for, e.g., $\delta(X, X') = \frac{1}{nm} \sum_{ij} |X_{ij} - X'_{ij}|$. Such results might tell us that when enough discrepancy is allowed, approximating with the rank is not very different then approximating with the max-norm. On the other hand, it would be interesting to understand if, for example, the matrices $T_t^{\otimes p}$ do not have any low max-norm matrix in their vicinity.

Throughout the paper we have largely ignored log-factors, but these can be very significant. For example, tighter bounds on the number of low max-complexity matrices can help us understand questions like the median max-complexity over all matrices.

References

1. Srebro, N., Rennie, J., Jaakkola, T.: Maximum margin matrix factorization. In: Advances In Neural Information Processing Systems 17. (2005)
2. Srebro, N., Alon, N., Jaakkola, T.: Generalization error bounds for collaborative prediction with low-rank matrices. In: Advances In Neural Information Processing Systems 17. (2005)
3. Arriaga, R.I., Vempala, S.: An algorithmic theory of learning: Robust concepts and random projection. In: Proc. of the 40th Foundations of Computer Science. (1999)
4. Ben-David, S., Eiron, N., Simon, H.U.: Limitations of learning via embeddings in euclidean half spaces. JMLR **3** (2002) 441–461
5. Forster, J., Schmitt, N., Simon, H.U., Suttorp, T.: Estimating the optimal margins of embeddings in euclidean half spaces. Machine Learning **51** (2003) 263–281
6. Forster, J., Simon, H.U.: On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes uniform distribution. In: Proceedings of the 13th International Conference on Algorithmic Learning Theory, Springer-Verlag (2002) 128–138

7. Linial, N., Mendelson, S., Schechtman, G., Shraibman, A.: Complexity measures of sign matrices. `www.cs.huji.ac.il/~nati/PAPERS` (2004)
8. Alon, N., Frankl, P., Rödel, V.: Geometrical realization of set systems and probabilistic communication complexity. In: Proceedings of the 26th Annual Symposium on the Foundations of Computer Science (FOCS). (1985) 227–280
9. Srebro, N.: Learning with Matrix Factorization. PhD thesis, Massachusetts Institute of Technology (2004)
10. Seginer, Y.: The expected norm of random matrices. Comb. Probab. Comput. **9** (2000) 149–166
11. Panchenko, D., Koltchinskii, V.: Empirical margin distributions and bounding the generalization error of combined classifiers. Annals of Statistics **30** (2002)
12. Pisier, G.: Factorization of linear operators and geometry of Banach spaces. Volume 60. Conference Board of the Mathemacial Sciences (1986)
13. Awerbuch, B., Kleinberg, R.: Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In: Proceedings of the 36th ACM Symposium on Theory of Computing (STOC). (2004)

A Consolidating Signs of Repeated Points

We show that for any function class and distribution, the Rademacher complexity can be bounded from above by consolidating all random signs corresponding to the same point into a single sign. We first show that consolidating a single sign can only increase the Rademacher complexity:

Lemma 11. *For any function class $\mathcal{F}$ and sample $S = (x_1, \ldots, x_n)$ with $x_1 = x_2$:*

$$\mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right] \le \mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\sigma_2 2f(x_2) + \sum_{i=3}^{n}\sigma_i f(x_i)\right|\right]$$

where σ_i are i.i.d. unbiased signs.

Proof. We first note that removing x_1, x_2 can only decrease the expectation:

$$\mathop{\mathbf{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right] = \mathop{\mathbf{E}}_{\sigma_{3:n}}\left[\mathop{\mathbf{E}}_{\sigma_{1,2}}\left[\sup_{f\in\mathcal{F}}\left|\sigma_1 f(x_1) + \sigma_2 f(x_2) + \sum_{i=3}^{n}\sigma_i f(x_i)\right|\right]\right]$$
$$\ge \mathop{\mathbf{E}}_{\sigma_{3:n}}\left[\sup_{f\in\mathcal{F}}\left|\mathop{\mathbf{E}}_{\sigma_{1,2}}\left[\sigma_1 f(x_1) + \sigma_2 f(x_2)\right] + \sum_{i=3}^{n}\sigma_i f(x_i)\right|\right] = \mathop{\mathbf{E}}_{\sigma_{3:n}}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=3}^{n}\sigma_i f(x_i)\right|\right]$$

Using this inequality we can now calculate:

$$\mathop{\mathbf{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right] \le \frac{1}{2}\mathop{\mathbf{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right] + \frac{1}{2}\mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\sigma_2 2f(x_2) + \sum_{i=3}^{n}\sigma_i f(x_i)\right|\right]$$

Subtracting the first term on the right-hand side from the original left-hand side gives us the desired inequality. □

By iteratively consolidating identical sample points, we get:

Lemma 12 (Sign Consolidation). *For any function class $\mathcal{F}$ and sample $S = (x_1, \ldots, x_n)$, denote by s_x the number of times a sample appears in the class, and let σ_x be i.i.d. unbiased random signs. Then:*

$$\mathcal{R}_S(\mathcal{F}) \le \mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{|S|}\sum_{x\in S}\sigma_x s_x f(x)\right|\right]$$