Concentration-Based Guarantees for Low-Rank Matrix Reconstruction

Rina Foygel RINA@UCHICAGO.EDU
Department of Statistics, University of Chicago

Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

We consider the problem of approximately reconstructing a partially-observed, approximately low-rank matrix. This problem has received much attention lately, mostly using the trace-norm as a surrogate to the rank. Here we study low-rank matrix reconstruction using both the trace-norm, as well as the less-studied max-norm, and present reconstruction guarantees based on existing analysis on the Rademacher complexity of the unit balls of these norms. We show how these are superior in several ways to recently published guarantees based on specialized analysis.

Keywords: Matrix completion, low-rank matrices, trace norm, nuclear norm, max norm, Rademacher complexity

1. Introduction

We consider the problem of (approximately) reconstructing an (approximately) low-rank matrix based on observing a random subset of entries. That is, we observe s randomly chosen entries of an unknown matrix $Y \in \mathbb{R}^{n\times m}$, where we assume either Y is of rank at most r, or there exists $X \in \mathbb{R}^{n\times m}$ of rank at most r that is close to Y. Based on these s observations, we would like to construct a matrix $\hat{X}$ that is as close as possible to Y.

There has been much interest recently in computationally efficient methods for reconstructing a partially-observed, possibly noisy, low-rank matrix, and on accompanying guarantees on the quality of the reconstruction and the required number of observations. Since directly searching for a low-rank matrix minimizing the empirical reconstruction error is NP-hard (Chistov and Grigoriev, 1984), most work has focused on using the trace-norm (a.k.a. nuclear norm, or Schatten-1-norm) as a surrogate for the rank. The trace-norm of a matrix is the sum (i.e. ℓ_1-norm) of its singular values, and thus relaxing the rank (i.e. the number of non-zero singular values) to the trace-norm is akin to relaxing the sparsity of a vector to its ℓ_1-norm, as is frequently done in compressed sensing. The analysis of the quality of reconstruction has also been largely driven by ideas coming from compressed sensing, typically studying the optimality conditions of the empirical optimization problem, and often requiring various "incoherence"-type assumptions on the underlying low-rank matrix.

In this paper we provide simple guarantees on approximate low-rank matrix reconstruction using a different surrogate regularizer: the $\gamma_{2:\ell_1\to\ell_\infty}$ norm, which we refer to simply as the "max-norm". This regularizer was first suggested by Srebro et al. (2005), though it has not received much attention since. Here we show how this regularizer can yield guarantees that are superior in some ways to recent state-of-the-art. In particular, we show that when the entries are uniformly bounded, i.e. $|X|_\infty = \mathbf{O}(1)$ (this corresponds to the "no spikiness" assumption of Negahban and Wainwright (2010), and is also assumed by Koltchinskii et al. (2010) and in the approximate reconstruction guarantee of Keshavan et al. (2010)), then the max-norm regularized predictor requires a sample size of

$$s = \mathbf{O}\left(\frac{r(n+m)}{\epsilon} \cdot \frac{\sigma^2+\epsilon}{\epsilon} \cdot \log^3(1/\epsilon)\right) \tag{1}$$

to achieve mean-squared reconstruction error $\frac{1}{nm}|\hat{X} - Y|_2^2 = \sigma^2 + \epsilon$, where σ^2 is the the mean-squared-error of the best rank-r approximation of Y—that is, $\sigma^2 = \frac{1}{nm}|X - Y|_2^2$, where X is the rank-r approximation. When Y is exactly low-rank (the noiseless case), $\sigma^2 = 0$ and the sample complexity is $\mathbf{O}\left(\frac{r(n+m)}{\epsilon} \cdot \log^3(1/\epsilon)\right)$. Compared to the three recent similar bounds mentioned above, this guarantee avoids the extra logarithmic dependence on the dimensionality, as well as the assumption of independent noise, but has a slightly worse dependence on ϵ. We emphasize that we do not make any assumptions about the noise, nor about incoherence properties of the underlying low-rank matrix X.

We also provide a guarantee on the mean-absolute-error of the reconstruction, and discuss guarantees for reconstruction using the trace-norm as a surrogate. Using the trace-norm allows us to provide mean-absolute-error guarantees also for matrices where the magnitudes are *not* uniformly bounded (i.e. "spiky" matrices). We further show that a spikiness assumption is necessary for squared-error approximate reconstruction of low-rank matrices, regardless of the estimator used.

Instead of focusing on optimality conditions as in previous work, our guarantees follow from generic generalization guarantees based on the Rademacher complexity, and an analysis of the Rademacher complexity of the max-norm and trace-norm balls conducted by Srebro and Shraibman (2005). To obtain the desired low rank reconstruction guarantees, we combine these with bounds on the max-norm and trace-norm in terms of the rank. The point we make here is that these fairly simple arguments, mostly based on the work of Srebro and Shraibman (2005), are enough to obtain guarantees that are in many ways better and more general than those presented in recent years.

Notation. We use $|M|$ to denote the elementwise norms of a matrix M: $|M|_1 = \sum_{ij} |M_{ij}|$, $|M|_2$ is the Frobenius norm, and $|M|_\infty = \max_{ij} |M_{ij}|$. We discuss $n \times m$ matrices, and without loss of generality always assume $n \geq m$.

2. The Max-Norm and Trace-Norm

We will consider the following two matrix norms, which are both surrogates for the rank:

Definition 1 *The* **trace-norm** *of a matrix $X \in \mathbb{R}^{n\times m}$ is given by:*

$$\|X\|_\Sigma = \sum(\textit{singular values of } X) = \min_{U,V:X=UV^T} |U|_2 |V|_2 .$$

Definition 2 *The* **max-norm** *of a matrix $X \in \mathbb{R}^{n \times m}$ is given by:*

$$\|X\|_{max} = \min_{U,V:X=UV^T} \left(\max_i |U_{(i)}|_2 \right) \left(\max_j |V_{(j)}|_2 \right) ,$$

where $U_{(i)}$ and $V_{(j)}$ denote the i^{th} row of U and the j^{th} row of V, respectively.

Both the trace-norm and the max-norm are semi-definite representable (Fazel et al., 2002; Srebro et al., 2005). Consequently, optimization problems involving a constraint on the trace-norm or max-norm, a linear or quadratic objective, and possibly additional linear constraints, are solvable using semi-definite programming. We will consider estimators which are solutions to such problems.

Srebro and Shraibman (2005) and later Sherstov (2007) studied the max-norm and trace-norm as surrogates for the rank in a classification setting, where one is only concerned with the signs of the underlying matrix. They showed that a sign matrix might be realizable with low rank, but realizing it with unit margin might require exponentially high max-norm or trace-norm. Based on this analysis, they argued that the max-norm and trace-norm *cannot* be used to obtain reconstruction guarantees for sign matrices of low rank matrices.

Here, we show that in a regression setting, the situation is quite different, and the max-norm and trace-norm *are* good convex surrogates for the rank. The specific relationship between these surrogates and the rank is determined by how we control the scale of the matrix X (i.e. the magnitude of its entries). This will be made explicit in the next section, but for now we state the bounds on the trace-norm and max-norm in terms of the rank which we will leverage in Section 3.

By bounding the ℓ_1 norm of the singular values (i.e. the trace-norm) by their ℓ_2 norm (i.e. the Frobenius norm) and the number of non-zero values (the rank), we obtain the following relationship between the trace-norm and Frobenius norm:

$$|X|_2 \leq \|X\|_\Sigma \leq \sqrt{\text{rank}(X)} \cdot |X|_2 \ . \tag{2}$$

Interpreting the Frobenius norm as specifying the *average* entry magnitude, $\frac{1}{nm}|X|_2^2$, we can view the above as upper bounding the trace-norm with the square root of the rank, when the average entry magnitude is fixed.

An analagous bound for the max norm, substituting ℓ_∞ norm (maximal entry magnitude) for Frobenius norm (average entry magnitude), can be obtained as follows:

Lemma 3 *For any $X \in \mathbb{R}^{n \times m}$, $|X|_\infty \leq \|X\|_{\max} \leq \sqrt{\text{rank}(X)} \cdot |X|_\infty$.*

Proof Consider the minimizing factorization $X = UV^T$ and let X_{ij} be the largest magnitude entry in X, then: $\|X\|_{\max} \geq \left|U_{(i)}\right| \cdot \left|V_{(j)}\right| \geq |X_{ij}| = |X|_\infty$.

To obtain the upper bound we first write the max-norm as (Lee et al., 2008):

$$\|X\|_{\max} = \sup_{p,q} \|\text{diag}(p) X \text{diag}(q)^2\|_\Sigma \ , \tag{3}$$

where the supremum is over nonnegative unit vectors p, q. We can now continue using (2):

$$\leq \sup_{p,q} \sqrt{\text{rank}(\text{diag}(p) X \text{diag}(q))} \cdot |\text{diag}(p) X \text{diag}(q)|_2$$

$$\leq \sup_{p,q} \sqrt{\text{rank} X} \cdot \sqrt{\sum_{ij} p_i^2 q_j^2 X_{ij}^2} = \sqrt{\text{rank} X}\, |X|_\infty \ .$$

■

3. Reconstruction Guarantees

The theorems below provide reconstructions guarantees, first under the a mean-absolute-error reconstruction measure (Theorem 4) and then under a mean-squared-error reconstruction measure (Theorem 6). Since the guarantees are for *approximate* reconstruction, we must impose some notion of scale. In other words, we can think of measuring the error relative to the scale of the data—if Y is multiplied by some constant, then obviously the reconstruction error would also be multiplied by this constant. In the theorems below we refer to two notions of scale: the *average* squared magnitude of matrix entries, i.e. $\frac{1}{nm}\left|X\right|_2^2$, and the *maximal* magnitude of matrix entries, i.e. $\left|X\right|_\infty$. For simplicity and without loss of generality, the results are stated for unit scale.

An issue to take note of is whether the s observed entries of Y are chosen with or without replacement, i.e. whether we choose a set S of entries uniformly at random over all sets of exactly s entries (no replacements), or whether we make s independent uniform choices of entries, possibly observing the same entry twice. Our results apply in both cases.

Theorem 4 *For any $M, Y \in \mathbb{R}^{n\times m}$ where M is of rank at most r:*

a. **Entry magnitudes bounded on-average.** *Consider the estimator*[1]

$$\hat{X}(S) = \arg\min_{\|X\|_\Sigma \leq \sqrt{rnm}} \sum_{(i,j)\in S} \left|Y_{ij} - X_{ij}\right| .$$

If $\frac{1}{nm}|M|_2^2 \leq 1$ and $s \geq \mathbf{O}\left(\frac{r(n+m)\log(n)}{\epsilon^2}\right)$, then in expectation over a sample S chosen either uniformly over sets of size s (without replacements) or by choosing s entries uniformly and independently (with replacements):

$$\frac{1}{nm}|Y - \hat{X}(S)|_1 \leq \frac{1}{nm}|Y - M|_1 + \epsilon .$$

b. **Entry magnitudes bounded uniformly.** *Consider the estimator*

$$\hat{X}(S) = \arg\min_{\|X\|_{\max} \leq \sqrt{r}} \sum_{(i,j)\in S} \left|Y_{ij} - X_{ij}\right| .$$

If $\left|M\right|_\infty \leq 1$ and $s \geq \mathbf{O}\left(\frac{r(n+m)}{\epsilon^2}\right)$, then in expectation over a sample S of size s chosen either with or without replacements as above:

$$\frac{1}{nm}|Y - \hat{X}(S)|_1 \leq \frac{1}{nm}|Y - M|_1 + \epsilon .$$

Remark 5 *The above results can also be shown to hold in high probability over the sample S, rather than in expectation. Specifically, to ensure that the results of Theorem 4 hold with probability at least $1 - n^{-\beta}$ (for sampling with replacement) or $1 - n^{-(\beta-2)}$ (for sampling without replacement), it is sufficient to change the sample size requirement to $s \geq \mathbf{O}\left(\frac{r(n+m)\log(n)+\beta\log(n)}{\epsilon^2}\right)$ (in the trace-norm case) or $s \geq \mathbf{O}\left(\frac{r(n+m)+\beta\log(n)}{\epsilon^2}\right)$ (in the max-norm case).*

[1] If S is chosen with replacements, it is a multiset, and the summation $\sum_{(i,j)\in S}$ should be interpreted as summation with repetitions.

Theorem 6 *For any $Y = M + Z \in \mathbb{R}^{n\times m}$ where $|Z|_\infty \le \sqrt{\frac{rn}{\log n}}$ and M is of rank at most r with $|M|_\infty \le 1$, denote $\sigma^2 = \frac{1}{nm}|Z|_2^2$. Consider the estimator*

$$\hat{X}(S) = \arg \min_{\|X\|_{\max} \le \sqrt{r}} \sum_{(i,j)\in S} (Y_{ij} - X_{ij})^2 \ . \tag{4}$$

If $s \ge \mathbf{O}\left(\frac{r(n+m)}{\epsilon} \cdot \frac{\sigma^2+\epsilon}{\epsilon} \cdot (\log^3(r/\epsilon) + \beta)\right)$, then, with probability at least $1 - n^{-\beta}$ over a sample S of size s chosen with replacement, or with probability at least $1 - n^{-(\beta-2)}$ over a sample S of size s chosen without replacement,

$$\frac{1}{nm}|Y - \hat{X}(S)|_2^2 \le \sigma^2 + \epsilon \ . \tag{5}$$

If we instead use the estimator:

$$\hat{X}(S) = \arg \min_{\substack{\|X\|_{\max} \le \sqrt{r} \\ |X|_\infty \le 1}} \sum_{(i,j)\in S} (Y_{ij} - X_{ij})^2 \ , \tag{6}$$

then we obtain (5) when $s \ge \mathbf{O}\left(\frac{r(n+m)}{\epsilon} \cdot \frac{\sigma^2+\epsilon}{\epsilon} \cdot (\log^3(1/\epsilon) + \beta)\right)$.

The estimator (6) is SDP-representable, though potentially more cumbersome.

Remark 7 *The requirement on the maximal magnitude of the error in Theorem 6, $|Z|_\infty \le \sqrt{\frac{rn}{\log n}}$, is very generous, and easily holds with high probability for sub-exponential noise. A stricter requirement, e.g. $\mathbf{O}(\sqrt{r \log n})$, which still holds with high probability for subgaussian noise, yields a guarantee with exponentially high probability $1 - e^{-n/\log n}$, without a sample-complexity dependence on β.*

Remark 8 *A guarantee similar to Theorem 6 can be obtained if we can ensure $\|M\|_{max} \le A$, for some A, without requiring $|M|_\infty \le 1$. For $\hat{X}(S) = \arg\min_{\|X\|_{max} \le A} \sum_{ij \in S}(Y_{ij} - X_{ij})^2$, we have (5) with a sample of size*

$$s \ge \mathbf{O}\left(\frac{A^2(n+m)}{\epsilon} \cdot \frac{\sigma^2+\epsilon}{\epsilon} \cdot (\log^3(A^2/\epsilon) + \beta)\right) \ .$$

In Section 4.2.3, we will see how certain incoherence assumptions used in previous bounds yield a bound on $\|M\|_{max}$, and compare the max-norm based reconstruction guarantee to the previously published results.

In Theorems 4 and 6 we do not assume the noise, i.e. the entries of $Z = Y - M$, are independent or zero-mean—in fact, we make no assumption on Z, other than the very generous upper bound $|Z|_\infty \le \sqrt{\frac{rn}{\log n}}$ discussed above. When entries of Z can be arbitrary, it is not possible to ensure reconstruction of M (e.g. we can set things up so Y actually has lower rank then M, and so it is impossible to identify M). Consequently, in Theorems 4 and 6 we instead bound the excess error in predicting Y itself. If entries of Z *are* independent and zero-mean, then we may give the following guarantee about reconstructing the underlying matrix M:

Theorem 9 *For $(i,j) \in [n] \times [m]$, let $\mathcal{F}_{(i,j)}$ be any mean-zero distribution. Suppose that the observed entries of Y are given by $Y_{(i_t,j_t)} = M_{(i_t,j_t)} + Z_t$ for $t = 1, 2, \ldots, s$, where $(i_t, j_t) \overset{iid}{\sim} Unif([n] \times [m])$ and $Z_t|(i_t, j_t) \sim \mathcal{F}_{(i_t,j_t)}$ independently for each t. That is, the noise is independent and zero-mean (though its distribution is allowed to depend on the location of the observation), the sample is drawn with replacement, and if an entry of the matrix is observed more than once, then the noise on the entry is drawn independently each time.*

Assume $|M|_\infty \leq 1$, $\mathrm{rank}(M) \leq r$, and $\sup_{t\in[s]} |Z_t| \leq \mathbf{o}\left(\sqrt{\frac{rn}{\log n}}\right)$ with high probability. Denote

$$\sigma^2 = \frac{1}{nm}\sum_{i,j} E_{Z_{ij}\sim\mathcal{F}_{ij}}(Z_{ij}^2) \ .$$

For the estimator given in Equation (4), with high probability over the sample S of size $s \geq \mathbf{O}\left(\frac{r(n+m)}{\epsilon} \cdot \frac{\sigma^2+\epsilon}{\epsilon} \cdot \log^3(r/\epsilon)\right)$,

$$\frac{1}{nm}|M - \hat{X}(S)|_2^2 \leq \epsilon \ . \tag{7}$$

Alternatively, is S is sampled uniformly without replacements, with the same assumptions and sample size, and as long as $s \leq \frac{K+1}{e}(nm)^{1-\frac{1}{K+1}}$, we have $\frac{1}{nm}|M - \hat{X}(S)|_2^2 \leq 4K\epsilon$.

Remark 10 *When sampling without replacement, we imposed both a lower bound and an upper bound on the sample size. For these two bounds to be compatible (in an asymptotic sense) for a fixed K, we need $m = \Omega(n^a)$ for some positive power a, and make ϵ arbitrarily small. Alternately, we can set $K = \mathbf{O}(\log n)$, ensuring the upper bound on s always holds (since $s \leq nm$ necessarily), yielding $\frac{1}{nm}|M - \hat{X}(S)|_2^2 \leq \epsilon$ whenever $s \geq \mathbf{O}\left(\frac{r(n+m)\log(n)}{\epsilon} \cdot \frac{\sigma^2+\epsilon}{\epsilon} \cdot \log^3(r/\epsilon)\right)$.*

The remainder of this Section is organized as follows: In Section 3.1, we prove Theorems 4 and 6 in the case where the sample is drawn without replacement. In Section 3.2, we discuss possible bounds of the mean-squared-error, as in Theorem 6, but using the trace-norm. In Section 3.3, we compare sampling with and without replacement, establishing Theorems 4 and 6 also for sampling with replacement. In Section 3.4, we turn to the setting of independent mean-zero noise, and prove Theorem 9 in both the sampling-with-replacement and sampling-without-replacement settings.

3.1. Proof of Theorems 4 and 6 when S is drawn with replacement

We first establish the Theorems for a sample chosen i.i.d. with replacements. In this case, following Srebro and Shraibman (2005), we may view matrix reconstruction as a prediction problem, by regarding a matrix $X \in \mathbb{R}^{n\times m}$ as a function $[n] \times [m] \to \mathbb{R}$. Each observation in the training set consists of a covariate $(i,j) \in [n] \times [m]$ and an observed noisy response $Y_{ij} \in \mathbb{R}$. Here, we assume that the distribution over $[n] \times [m]$ is uniform, and the joint distribution over (i,j) and its response is determined by the unknown Y. The hypothesis class is then a set of matrices bounded in either trace-norm or max-norm, and for a particular hypothesis $X \in \mathbb{R}^{n\times m}$, the averaged error $\frac{1}{nm}|Y - X|_1$ or $\frac{1}{nm}|Y - X|_2^2$ is equal to the expected loss $L(X) = \mathbf{E}_{ij}\left[\mathrm{loss}(X_{ij}, Y_{ij})\right]$ under either the absolute-error or squared-error loss, respectively.

Srebro and Shraibman (2005) established bounds on the Rademacher complexity of the trace-norm and max-norm balls. For any sample of size s, the empirical Rademacher complexity of the max-norm ball is bounded by

$$\hat{\mathcal{R}}_s\left(\left\{X \in \mathbb{R}^{n\times m} \mid \|X\|_{\max} \le A\right\}\right) \le 12\sqrt{\frac{A^2(n+m)}{s}} \ . \tag{8}$$

Although the empirical Rademacher complexity of the trace-norm ball might be fairly high, the *expected* Rademacher complexity, for a random sample of s independent *uniformly* chosen index pairs (with replacements) can be bounded as

$$\mathbf{E}\left[\hat{\mathcal{R}}_s\left(\left\{X \in \mathbb{R}^{n\times m} \mid \|X\|_{\Sigma} \le A\right\}\right)\right] \le K\sqrt{\frac{\frac{A^2}{nm}(n+m)\log(n)}{s}} \tag{9}$$

for some numeric constant K (this is a slightly better bound then the one given by Srebro and Shraibman (2005), and is proved in Appendix B).

Since the absolute error loss, $\mathrm{loss}(x, y) = |x - y|$, is 1-Lipschitz, these Rademacher complexity bounds immediately imply (Bartlett and Mendelson, 2001):

$$\frac{1}{nm}\left|Y - \hat{X}(S)\right|_1 \le \inf_{\|X\|_{\max}\le A}\left(\frac{1}{nm}|Y - X|_1\right) + 24\sqrt{\frac{A^2(n+m)}{s}} \tag{10}$$

for $\hat{X}(S) = \arg\min_{\|X\|_{\max}\le A}\sum_{(i,j)\in S}|Y_{ij} - X_{ij}|$, and:

$$\frac{1}{nm}\left|Y - \hat{X}(S)\right|_1 \le \inf_{\|X\|_{\Sigma}\le A}\left(\frac{1}{nm}|Y - X|_1\right) + 2K\sqrt{\frac{\frac{A^2}{nm}(n+m)\log(n)}{s}} \tag{11}$$

for $\hat{X}(S) = \arg\min_{\|X\|_{\Sigma}\le A}\sum_{(i,j)\in S}|Y_{ij} - X_{ij}|$. These provide guarantees on reconstructing matrices with bounded max-norm or trace-norm. Choosing $A = \sqrt{r}$ for the max-norm and $A = \sqrt{rnm}$ for the trace-norm, Theorem 4 (for sampling with replacement) follows from Equation (2) and Lemma 3. (Remark 5 follows from the results of Bartlett and Mendelson (2001) with identical arguments for the sampling-with-replacement case.)

In order to obtain Theorem 6, we use a recent bound on the excess error with respect to a *smooth* (rather then Lipschitz) loss function, such as the squared loss. Specifically, Theorem 1 of Srebro et al. (2010) states that, for a class of predictors $X : \mathcal{I} \to [-B, B]$ and a loss function bonded by b with second derivative bounded by H, with probability at least $1 - \delta$ over a random sample of size s,

$$L(\hat{X}) \le L^* + O\left(\sqrt{L^*\tilde{\mathcal{R}}_s} + \tilde{\mathcal{R}}_s\right) \ , \tag{12}$$

$$L^* = \inf_X L(X) \ ,$$

$$\tilde{\mathcal{R}}_s = H\mathcal{R}_s^2\log^3\left(\frac{B}{\mathcal{R}_s}\right) + \frac{b\log(\log(s)/\delta)}{s} \ , \tag{13}$$

where the infimum is over predictors in the class, $\hat{X}$ is the empirical error minimizer in the class, and $\mathcal{R}_s$ is an upper bound on the Rademacher complexity for all samples of size s.

In our case, for the class $\{X | \|X\|_{\max} \leq A\}$ and the squared loss, we have $B = \sup_X \sup_{ij} |X_{ij}| = \sup_X |X|_\infty \leq \sup_X \|X\|_{\max} \leq A$ and $b = \sup_X |X - Y|_\infty^2 \leq \sqrt{\frac{4A^2(n+m)}{\log(n+m)}}$, when we assume $|Z|_\infty \leq A\sqrt{\frac{n+m}{\log(n+m)}}$. Applying the bound (8) on the Rademacher complexity yields:

$$\tilde{\mathcal{R}}_s = \mathbf{O}\left(\frac{A^2(n+m)}{s}\log^3\left(\frac{s}{n}\right) + \frac{A^2(n+m)\log\log s}{s\log(n+m)} + \frac{A^2(n+m)\log(1/\delta)}{s\log n}\right) \tag{14}$$

$$= \mathbf{O}\left(\frac{A^2(n+m)}{s}\left(\log^3\left(\frac{s}{n+m}\right) + \frac{\log(1/\delta)}{\log n}\right)\right) . \tag{15}$$

Here the last inequality uses the fact that $s \leq n^2$, while the next-to-last inequality assumes $s \geq e^3(n+m)$, and applies the fact that $x^2\log^3(1/x)$ is an increasing function for $x < e^{-1.5}$, where in this case $x = \sqrt{\frac{n+m}{s}}$.

Remark 8 follows immediately. The first claim in Theorem 6 follows when we assume $|M|_\infty \leq 1$ and $\text{rank}(M) \leq r$ and set $A = \sqrt{r}$ (since, by Lemma 3, $\|M\|_{\max} \leq A$). If we instead consider the class $\{X : \|X\|_{\max} \leq \sqrt{r}, |X|_\infty \leq 1\}$, then in the notation of (12), we may define $B = 1$ instead of $B = A = \sqrt{r}$, and thus obtain

$$\tilde{\mathcal{R}}_s = \mathbf{O}\left(\frac{r(n+m)}{s}\left(\log^3\left(\frac{s}{r(n+m)}\right) + \frac{\log(1/\delta)}{\log n}\right)\right) , \tag{16}$$

which yields the second claim of Theorem 6.

Finally, we prove the claim Remark 7. If instead we assume $|Z|_\infty \leq \sqrt{r\log n}$, then in the the notation of (12), we may define $b = r\log n$ instead of $b = \frac{4A^2n}{\log n} = \frac{4rn}{\log n}$, and thus obtain

$$\tilde{\mathcal{R}}_s = \mathbf{O}\left(\frac{r(n+m)}{s}\left(\log^3\left(\frac{s}{(n+m)}\right) + \frac{\log n \cdot \log(1/\delta)}{n+m}\right)\right) . \tag{17}$$

For $\delta \leq e^{-n/logn}$, the second term is dominated by the first; therefore the sample complexity no longer depends on β.

3.2. Bounds on ℓ_2 error using the trace norm

In Theorem 4, we saw that for mean-absolute-error matrix reconstruction, using the trace-norm instead of the max-norm allows us to forgo a bound on the spikiness, and rely only on the average squared magnitude $\frac{1}{nm}|Y|_2^2$. One might hope that we can similarly get a squared-error reconstruction guarantee using the trace-norm and without a spikiness bound that was required in Theorem 6. Unfortunately, this is not possible.

In fact, as the following example demonstrates, it is not possible to reconstruct a low-rank matrix to within much-better-then-trivial squared-error without a spikiness assumption, and relying only on $\frac{1}{nm}|Y|_2 \leq 1$. Specifically, consider an $n \times m$ matrix

$$Y = \sqrt{m/r}\left(A \,|\, 0_{n\times(m-r)}\right)$$

where $A \in \{\pm 1\}^{n \times r}$ is an arbitrary sign matrix. The matrix Y has rank at most r and average squared magnitude $\frac{1}{nm}|Y|_2^2 = 1$ (but maximal squared magnitude $|Y|_\infty^2 = m/r$). Now, with even half the entries observed (i.e. $s = nm/2$), we have no way of reconstructing the unobserved entries of A, as any values we choose for these entries would be consistent with the rank-r assumption, yielding an expected average squared error of at least $1/2$. We can conclude that regardless of the estimator, controlling the average squared magnitude is not enough here, and we cannot expect to obtain a squared-error reconstruction guarantee based on $\frac{1}{nm}|Y|_2^2$, even if we use the trace-norm.

We note that if $|M|_\infty, |Y|_\infty = \mathbf{O}(1)$, then the squared-loss in the relevant regime has a bounded Lipschitz constants, and Theorem 4a applies. In particular, if $|M|_\infty, |Y|_\infty \leq 1$, then we can consider the estimator

$$\hat{X}(S) = \arg \min_{\substack{\|X\|_\Sigma \leq \sqrt{rnm} \\ |X|_\infty \leq 1}} \sum_{(i,j)\in S} (Y_{ij} - X_{ij})^2 \ . \tag{18}$$

Since we now only need to consider X where $|X_{ij} - Y_{ij}| \leq 2$, the squared-loss in the relevant domain is 4-Lipschitz. We can therefore use the standard generalization results for Lipschitz loss as in Theorem 4, and obtain that with high probability over a sample of size

$$s \geq \mathbf{O}\left(\frac{r(n+m)\log n}{\epsilon^2}\right) \ , \tag{19}$$

we have $\frac{1}{nm}|Y - \hat{X}(S)|_2^2 \leq \sigma^2 + \epsilon$. However, this result gives a dependence on ϵ that is quadratic, as opposed to the more favorable dependence (at least when $\epsilon = \Omega(\sigma^2)$) of Theorem 6.

We believe that, when $|M|_\infty, |Y|_\infty \leq \mathbf{O}(1)$, it is possible to improve the dependence on ϵ to a dependence similar to that of Theorem 6 (this would require a more delicate analysis then that of Srebro et al. (2010), as their techniques rely on bounding the worst-case Rademacher complexity). But even this would not give any advantage over the max-norm, since the bound on $|M|_\infty$ could not be relaxed, while an additional factor of $\log n$ would be introduced into the sample complexity (coming from the Rademacher complexity calculation for the trace-norm). It seems then, that at least in terms of the quantities and conditions considered in this paper, as well as elsewhere in the low-rank reconstruction literature we are familiar with, there is no theoretical advantage for the trace-norm over the max-norm in terms of squared-error approximate reconstruction, though there could be an advantage for the max-norm in avoiding a logarithmic factor.

3.3. Sampling with or without replacement in Theorems 4 and 6

Theorems 4 and 6 give results that hold for either sampling with replacement or sampling without replacement. When an entry of the matrix Y is sampled twice, the same value is observed each time—no new information about the matrix is observed, and so intuitively, sampling without replacement should yield strictly better results than sampling with replacement. The two lemmas below, proved in the Appendix, establish that sampling without replacement is indeed as at least as good as sampling with replacement (up to a constant).

Before stating the lemmas, we briefly introduce some notation. Let $L(X)$ denote the loss for an estimated matrix X; that is, $L(X) = \frac{1}{nm}|Y - X|_1$ or $L(X) = \frac{1}{nm}|Y - X|_2^2$, as appropriate. Let $\hat{L}_S(X)$ denote the empirical loss, $\hat{L}_S(X) = \sum_{(i,j)\in S} |Y_{ij} - X_{ij}|^p$ (where $p \in \{1,2\}$ and the sum includes repeated elements in S). Let $\mathcal{D}^s$ and $\mathcal{D}^s_{w/o}$ denote the distributions of a sample of size s drawn uniformly at random from the matrix, either with or without replacement, respectively.

Lemma 11 *Let $\mathcal{X}$ denote any class of matrices, with $\mathcal{D}^s$ and $\mathcal{D}^s_{w/o}$ defined as above. Then*

$$E_{S\sim\mathcal{D}^s_{w/o}}\left[\sup_{X\in\mathcal{X}} L(X) - \hat{L}_S(X)\right] \leq E_{S\sim\mathcal{D}^s}\left[\sup_{X\in\mathcal{X}} L(X) - \hat{L}_S(X)\right] .$$

Lemma 12 *Let $\mathcal{X}$ denote any class of matrices, with $\mathcal{D}^s$ and $\mathcal{D}^s_{w/o}$ defined as above. Then for any $c \in \mathbb{R}$, and for any function g,*

$$P_{S\sim\mathcal{D}^s_{w/o}}\left\{\left(\sup_{X\in\mathcal{X}} g(L(X)) - \hat{L}_S(X)\right) \geq c\right\} \leq 4s \cdot P_{S\sim\mathcal{D}^s}\left\{\left(\sup_{X\in\mathcal{X}} g(L(X)) - \hat{L}_S(X)\right) \geq c\right\} .$$

For the ℓ_1-loss case, the Rademacher bounds (10) and (11) are derived from Bartlett and Mendelson (2001) by bounding $E_{S\sim\mathcal{D}^s}\left(\sup_{X\in\mathcal{X}} L(X) - \hat{L}_S(X)\right)$ (or by bounding $P_{S\sim\mathcal{D}^s}\left(\sup_{X\in\mathcal{X}} L(X) - \hat{L}_S(X) \geq c\right)$, for the proof of Remark 5). By Lemma 11, the same bound then holds for the same expectation taken over $S \sim \mathcal{D}^s_{w/o}$, and therefore (10) and (11) must hold for this case as well. This implies that the results of Theorem 4 (and Remark 5) hold for sampling without replacement as well as sampling with replacement.

Similarly, for the ℓ_2-loss case, the Rademacher bound (12) is derived in Srebro et al. (2010) by bounding $\sup_{X\in\mathcal{X}} L(X) - \sqrt{a \cdot L(X)} - \hat{L}_S(X)$ for some constant a, with probability at least $1-\delta$ over $S \sim \mathcal{D}^s$. Defining $g(L) = L - \sqrt{a \cdot L}$, the same bound must therefore hold with probability at least $1 - 4s\delta \geq 1 - 4n^2\delta$ over $S \sim \mathcal{D}^s_{w/o}$, and therefore (12) holds for this case also. This implies that the results of Theorem 6 (and the subsequent remarks) hold for sampling without replacement as well as sampling with replacement.

3.4. Proof of Theorem 9: independent errors in the ℓ_2-loss setting.

First, we prove the theorem when sampling with replacement. For a matrix X, let $L(X)$ denote the expected squared error for a randomly sampled entry, that is,

$$L(X) = \frac{1}{nm}\sum_{(i,j)} E((Y_{ij} - X_{ij})^2) = \frac{1}{nm}\sum_{(i,j)} E_{Z\sim\mathcal{F}_{(i,j)}}((Z + M_{ij} - X_{ij})^2) .$$

Now write $\sigma^2 = \frac{1}{nm}\sum_{(i,j)} E_{Z\sim\mathcal{F}_{(i,j)}}(Z^2)$. Then $L(M) = \sigma^2$.

Then, for any sample S, given $\hat{X}(S)$ which is a random matrix depending on some observed sample, the expected loss (over a future observation of an entry in the matrix) of $\hat{X}(S)$ satisfies the following (due to the fact that noise in a future observation of the matrix has zero mean and is independent from $\hat{X}(S)$):

$$\begin{aligned} L(\hat{X}(S)) &= E_{(i,j)}\left((Y_{ij} - \hat{X}(S)_{ij})^2 \middle| \hat{X}(S)\right) = E_{(i,j),Z\sim\mathcal{F}_{ij}}\left((Z + M_{ij} - \hat{X}(S)_{ij})^2 \middle| \hat{X}(S)\right) \\ &= E_{(i,j),Z\sim\mathcal{F}_{ij}}\left(Z^2 + (M_{ij} - \hat{X}(S)_{ij})^2 \middle| \hat{X}(S)\right) = E_{(i,j),Z\sim\mathcal{F}_{ij}}\left(Z^2\right) + \frac{1}{nm}|M - \hat{X}(S)|_2^2 \\ &= \sigma^2 + \frac{1}{nm}|M - \hat{X}(S)|_2^2 . \end{aligned}$$

Therefore, following the same reasoning as the proof of Theorem 6 (and Remark 7, we have that if $s \geq \mathbf{O}\left(\frac{r(n+m)}{\epsilon} \cdot \frac{\sigma^2+\epsilon}{\epsilon} \cdot \log^3(r/\epsilon)\right)$, then with high probability,

$$L(\hat{X}) \leq \sigma^2 + \epsilon \ .$$

Applying the work above, we obtain

$$\frac{1}{nm}|M - \hat{X}(S)|_2^2 \leq \epsilon \ . \tag{20}$$

Now we turn to sampling without replacement. We first state a lemma which is proved in the appendix. (Notation: here $\mathcal{D}^s$ and $\mathcal{D}^s_{w/o}$ again denote sampling with or without replacement, but in this context $\mathcal{D}^s$ represents sampling with replacement when the noise is added independently each time an entry is sampled, as in the statement of Theorem 9.)

Lemma 13 *Let $\mathcal{X}$ denote any class of matrices, with $\mathcal{D}^s$ and $\mathcal{D}^s_{w/o}$ defined as above. For any c, if s satisfies $s \leq \frac{K+1}{e}(nm)^{1-\frac{1}{K+1}}$, then*

$$P_{S\sim\mathcal{D}^s_{w/o}}\left(\sup_{X\in\mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq c\right) \leq 4K \cdot P_{S\sim\mathcal{D}^s}\left(\sup_{X\in\mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq (2K)^{-1}c\right) \ .$$

As in the proof of the sampling-without-replacement case of Theorem 6, this is sufficient to show that $\frac{1}{nm}|M - \hat{X}(S)|_2^2 \leq 4K \cdot \epsilon$ with high probability for the stated sample complexity, as long as we also have that $s \leq \frac{K+1}{e}(nm)^{1-\frac{1}{K+1}}$.

4. Comparison to prior work

Suppose $Y = M + Z$ where $\text{rank}(M) \leq r$ and Z is a "noise" matrix of average squared magnitude $\sigma^2 = \frac{1}{nm}|Z|_2^2$, and we observe random entries of Y. One might then consider different types of reconstruction guarantees, requiring different assumptions on M, Z and the sampling distribution:

$$\begin{array}{ll}
\text{Exact recovery of } M: & \hat{X}(S) = M \ . \\
\text{Near-exact recovery of } M: & \frac{1}{nm}|\hat{X}(S) - M|_2^2 \leq \epsilon \cdot \sigma^2 \ . \\
\text{Approximate recovery of } M: & \frac{1}{nm}|\hat{X}(S) - M|_2^2 \leq \epsilon \cdot \text{scale}(M) \ . \\
\text{Approximate recovery of } Y: & \frac{1}{nm}|\hat{X}(S) - Y|_2^2 \leq \sigma^2 + \epsilon \cdot \text{scale}(M) \ .
\end{array}$$

Exact or near-exact recovery require strong incoherence-type assumptions on the matrix M, and is not possible for arbitrary low-rank matrices (see, e.g. Candès and Recht (2009)). Here we do not make any such assumptions, and show that approximate recovery is still possible. Such approximate recovery must be relative to some measure of the scale of M, and we discuss results relative to both the maximal magnitude, $\text{scale}(M) = |M|_\infty^2$, and the average squared magnitude $\text{scale}(M) = \frac{1}{nm}|M|_2^2$. Although not actually guaranteeing the same type of "recovery", in Section 4.2 we nevertheless compare the sample complexity required for our approximate recovery results to the best sample complexity guarantee for exact and near-exact recovery (obtained by Recht (2009) and Keshavan et al. (2010), respectively), and comment on the differences between the required assumptions on M.

More directly comparable to our results are recent results by Keshavan et al. (2010), Negahban and Wainwright (2010) and Koltchinskii et al. (2010) on approximate recovery of M. These give essentially the same type of guarantee as in Theorem 9, and also rely on $|M|_\infty^2$ as a measure of scale. In Section 4.1 we compare our guarantee to these results, discussing the different dependence on the various parameters and different assumptions on the noise. (Note that both types of results appear in Keshavan et al. (2010); in Section 4.1, we refer to the approximate recovery result stated in Theorem 1.1 of their paper, while in Section 4.2, we refer to the near-exact recovery result stated in Theorem 1.2 of their paper.)

Recovery of M, whether exact, near-exact, or approximate, also requires the noise to be independent and zero-mean, otherwise M might not be identifiable. All prior matrix reconstruction results we are aware of work in this setting. Approximate recovery of M also immediately implies an excess error bound on approximate recovery of Y. However, we also provide excess error bounds for approximate recovery of Y, that do *not* assume independent nor zero-mean noise (Theorems 4 and 6). That is, we provide reconstruction guarantees in a significantly less restrictive setting compared to other matrix reconstruction guarantees.

Another difference between different results is whether entries are sampled with or without replacement, and if replacement is allowed, whether the error is per-entry (i.e. repeat observations of the same entry are identical) or per-observation (i.e. repeat observations of the same entry are each corrupted independently). However, as we show in Sections 3.3 and 3.4, and as has also been shown for exact recovery (Recht, 2009), these differences do not significantly alter the quality of reconstruction or the required sample size.

The most common algorithm for low-rank matrix recovery in the literature is squared-error minimization subject to a penalty on trace norm. All the methods cited here prove results about some variation of this approach, with the exception of a recent result by Keshavan et al. (2010), which applies to the output of the local search procedure OptSpace. In contrast, our results are mostly for error minimization subject to a *max-norm* constraint.

4.1. Comparison With Recent Approximate Recovery Guarantees

Negahban and Wainwright (2010) and Koltchinskii et al. (2010) recently presented guarantees on approximate recovery using trace-norm regularization, in a setting very similar to our Theorem 9. Earlier work by Keshavan et al. (2010) uses a low-rank SVD approximation to $\tilde{Y}_S$ in the same setting to also obtain an approximate recovery guarantee. (Here Y_S is the matrix consisting of all observed entries of Y, with zeros elsewhere, and $\tilde{Y}_S$ is the same matrix with overrepresented rows and columns removed.) In particular, each of the three guarantees provide an ϵ-approximate reconstruction of M relative to $|M|_\infty^2$. That is, when $|M|_\infty \leq 1$ as in Theorem 9, they provide the exact same guarantee $\frac{1}{nm}\left|\hat{X}(S) - M\right| \leq \epsilon$. (Negahban and Wainwright state the result relative to $\frac{1}{nm}|M|_2^2$, but have a linear dependence on the "spikiness" $\frac{|M|_\infty}{|M|_2/\sqrt{nm}}$, effectively giving a guarantee relative to $|M|_\infty^2$).

Specifically, assuming $|M|_\infty = 1$ without loss of generality, Negahban and Wainwright and Koltchinskii et al. assume the noise is independent and subgaussian (or subexponential) with variance $\mathbf{O}(\sigma^2)$, and require a sample size of:

$$s \geq \mathbf{O}\left(\frac{rn\log(n)}{\epsilon} \cdot (1+\sigma^2)\right) . \tag{21}$$

where the sample is drawn with replacement—in particular, an entry (i, j) of the matrix which is sampled multiple times gives multiple independent estimates of M_{ij}.

Keshavan et al. give a result on approximate recovery which holds with no assumption on the noise, but requires additional assumptions such as i.i.d. noise to be a meaningful bound. The estimator used is the rank-r SVD approximation to $\tilde{Y}_S$, defined above. Specifically, they show that, for sufficiently large sample size, with high probability, $\frac{1}{\sqrt{nm}}|\hat{X}(S) - M|_2 \le \mathbf{O}\left(\frac{nr\sqrt{n/m}}{s} + \frac{nmr}{s^2}\|\tilde{Z}_S\|_2^2\right)$, where $\tilde{Z}_S$ is defined in the same way as $\tilde{Y}_S$. For this bound to be meaningful, there must be some distributional assumption on Z—otherwise, we could have $\|Z_S\|_2 \approx |Z_S|_2 = \mathbf{O}(\sqrt{s})$, and the bound on mean error would actually increase with $\frac{nm}{s}$, and is thus not a meaningful bound. In the presence of i.i.d. subgaussian noise, however, Keshavan et al. show that with high probability, $\|\tilde{Z}_S\|_2^2 \le \frac{\sigma^2(\sqrt{n/m})s\log(s)}{m}$. Using this, approximate recovery of M is obtained for sample complexity

$$s \geq \mathbf{O}\left(\frac{rn}{\epsilon} \cdot (\sqrt{n/m}) \cdot \left(1 + \log(n)\sigma^2\right)\right) , \tag{22}$$

where the sample is drawn *without* replacement. Therefore we may regard Keshavan et al.'s result as bounding error under the assumption of i.i.d. subgaussian noise (or perhaps some weaker assumption that gives the same result, such as independent subgaussian noise that might not be i.i.d., or similar). The guarantees (22) and (21) are therefore quite similar, even though they are for fairly different methods, with (22) being better when $\sigma^2 = \mathbf{o}(1)$ but worse for highly rectangular matrices.

Comparing our Theorems 6 and 9 to the above, the advantages of our results are:

- We avoid the extra logarithmic dependence on n.
- Even in order to guarantee recovery of M, we assume only a much milder condition on the noise: that noise is mean-zero, and that with high probability, $|Z_S|_\infty \le \sqrt{\frac{rn}{\log n}}$. We do not assume the noise is identically distributed, nor subgaussian or subexponential.
- We provide a guarantee on the excess error of recovering Y, even when the noise is *not* zero-mean nor independent.

The deficiency of our result is a possible slower rate of error decrease: when $\sigma > 0$ and $\epsilon = \mathbf{o}(\sigma^2)$ (i.e. to get "estimation error" significantly lower then the "approximation error"), our sample complexity scales as $\tilde{\mathbf{O}}(1/\epsilon^2)$ compared to just $\mathbf{O}(1/\epsilon)$ in the other results. We do not know if this difference represents a real consequence of not assuming zero-mean independent noise in our analysis, or just looseness in the proof. Our results also include an additional $\log^3(1/\epsilon)$ factor, which we believe is purely an artifact of the proof technique.

A strength of our analysis, as compared to that of Negahban and Wainwright and Koltchinskii et al., is that the cases of sampling with and without replacement are both covered, including the case of per-entry noise when sampling with replacement, while the results of Negahban and Wainwright and Koltchinskii et al. are for sampling with replacement with per-observation noise. This is an important improvement because in many applications, the observed entries are drawn from a fixed matrix which was randomly generated, meaning that it is not possible to obtain multiple independent observations of any M_{ij}.

4.2. Comparison of results on exact and near-exact recovery

The results of Recht and of Keshavan et al. show that exact or near-exact recovery of the underlying low-rank matrix M can be obtained with high probability, when strong conditions on M are assumed, and when the observations are either noiseless (for Recht's exact recovery result) or are corrupted by i.i.d. subgaussian noise (for Keshavan et al.'s near-exact recovery result).

These results cannot be directly compared to the results we obtain in this paper, because the guarantees on recovery given by this work and by our work are fundamentally different—for instance, the error bound ϵ has completely different meanings in our definitions of near-exact recovery and approximate recovery above. These two incomparable types of guarantees are linked to very different conditions on the data—exact and near-exact recovery cannot be obtained without strict assumptions about how the observations are generated.

Nonetheless, one comparison between these methods which can be made, is in the magnitude of the required sample complexities to obtain some meaningful bound via each result—exact recovery for Recht's result, near-exact recovery for Keshavan et al.'s result, and approximate recovery relative to $|M|_\infty^2$ for our result. The rest of this section is organized as follows: we summarize the results in the literature in Section 4.2.1, compare sample complexities in Section 4.2.2, and describe how incoherence is sufficient but not necessary for approximate recovery relative to $\frac{1}{nm}|M|_2^2$ (instead of $|M|_\infty^2$) in Sections 4.2.3 and 4.2.4.

4.2.1. Details on exact and near-exact results in the literature

Let $M = U\Sigma V^T$ be a reduced SVD of M. Let κ be the condition number of Σ. Define also the incoherence parameters for matrix M (Candès and Recht, 2009):

$$\mu_0 = \max\left\{\frac{n}{r}\cdot\max_i |U_{(i)}|_2^2, \frac{m}{r}\cdot\max_j |V_{(j)}|_2^2\right\} ,$$

$$\mu_1 = \sqrt{\frac{nm}{r}}\cdot\max_{i,j}|U_{(i)}^T V_{(j)}| ,$$

where $U_{(i)}$ denotes the ith row of U and $V_{(j)}$ denotes the jth row of V.

Suppose that M has low incoherence parameters and $Z = 0$. Improving on the earlier results of Candès and Recht (2009) and Candes and Tao (2010), Recht proves that $\hat{X}(S) = M$ (that is, exact recovery is obtained) with high probability if

$$s \geq \mathbf{O}\left(rn\max\{\mu_0,\mu_1^2\}\log^2 n\right) . \tag{23}$$

In the case of noisy observations, Keshavan et al. give conditions on low ℓ_2 error in recovery (with high probability) in the setting of i.i.d. subgaussian noise with incoherent M, improving on Candes and Plan (2010) earlier work on the noisy case. (More precisely, Keshavan et al. give a result which holds with no assumption on the noise, but requires additional assumptions such as i.i.d. noise to be a meaningful bound. We therefore regard their result as assuming i.i.d. subgaussian noise—see the discussion of their approximate reconstruction result above in Section 4.1.) Their OptSpace algorithm is a method for finding the rank-r matrix $\hat{X}$ minimizing squared error on the observed entries. Let $\hat{X}(S)$

denote the matrix recovered by this algorithm. When the entries of Z are i.i.d. subgaussian, Keshavan et al. show that, with high probability, if s satisfies

$$s \geq \mathbf{O}\left(rn\kappa^4 \cdot \max\left\{ \frac{1}{\epsilon} \log\left(\frac{rn\kappa^4}{\epsilon} \right), r\kappa^2\mu_0^2, r\kappa^2\mu_1^2 \right\} \right) , \tag{24}$$

then $|\hat{X}(S) - M|_2^2 \leq |Z|_2^2 \cdot \epsilon$. (For simplicity of the comparison, we use a slightly relaxed form of their required sample complexity, and ignore $\sqrt{n/m}$ in their error and sample bounds.)

4.2.2. Comparing sample complexities

Ignoring the dependence on ϵ, which as we discussed earlier is in any case incomparable between approximate and exact and near-exact recovery, our sample complexity for approximate recovery using the max-norm is $\mathbf{O}(rn)$. Even with "perfect" incoherence parameters, this a factor of $\log^2(n)$ less then the sample complexity established by Recht for exact recovery (23), and a factor of r less then the sample complexity established by Keshavan et al. for near-exact recovery (24). Of course, "bad" incoherence parameters may sharply increase the sample complexity for exact or near-exact recovery, but do not affect our sample complexity for approximate recovery.

4.2.3. Approximate recovery relative to average signal magnitude, in the presence of incoherence conditions

It is interesting to note that the incoherence assumptions, used by Recht and by Keshavan et al., enable approximate recovery with the max-norm relative to the average magnitude $\frac{1}{nm}|M|_2^2$, and not only the maximal magnitude, as in Theorem 6. This is based on the following observation:

Lemma 14 *Let $M \in \mathbb{R}^{n\times m}$ and let κ and μ_0 be defined as before. Then*

$$\|M\|_{\max} \leq \min\{\kappa, \sqrt{r}\}\mu_0\sqrt{r} \cdot \frac{|M|_2}{\sqrt{nm}} .$$

In particular, by Lemma 3, the above expression is also an upper bound for $|M|_\infty$.

Proof First, observe that

$$\|M\|_{\max} \leq \max_{i,j} |(U\Sigma)_{(i)}|_2 \cdot |V_{(j)}|_2 \leq \sigma_1 \cdot \max_{i,j} |U_{(i)}|_2 \cdot |V_{(j)}|_2 \leq \sigma_1 \cdot \frac{\mu_0 r}{\sqrt{nm}} .$$

Also,

$$\sigma_1 \leq \kappa\sqrt{\sigma_r^2} \leq \frac{\kappa}{\sqrt{r}}\sqrt{\sigma_1^2 + \cdots + \sigma_r^2} = \frac{\kappa|M|_2}{\sqrt{r}} \text{ and } \sigma_1 \leq \sqrt{\sigma_1^2 + \cdots + \sigma_r^2} = |M|_2 .$$

■

Now, based on Remark 8, if $\frac{1}{nm}|M|_2^2 \leq 1$ (and with a mild bound on $|Z|_\infty$), with high probability over a sample of size

$$s \geq \mathbf{O}\left(\frac{rn}{\epsilon} \cdot \frac{\sigma^2 + \epsilon}{\epsilon} \cdot \min\{\kappa^2, r\}\mu_0^2 \cdot \log^3\left(\frac{\mu_0^2 r}{\epsilon} \right) \right) , \tag{25}$$

we have $|Y - \hat{X}(S)|_2^2 \leq \sigma^2 + \epsilon$. Up to log factors and the dependence on ϵ, this sample complexity is at most as much as the sample complexity required by Keshavan et al., given in (24).

4.2.4. Approximate recovery relative to average signal magnitude, in the absence of incoherence conditions

We make note of several special cases where using max-norm and the concentration result, and bounding excess error relative to $\frac{1}{nm}|M|_2^2$, may compare more favorably to other methods than the results above would indicate.

- If $U = V$ (that is, M is symmetric), then $\mu_1 = \mu_0\sqrt{r}$ and so our sample complexity compares more favorably to the sample complexities obtained by Recht and Keshavan et al. (which both involve μ_1^2).
- Our sample complexity uses Lemma 14 to bound $\|M\|_{\max}$ relative to $\frac{1}{\sqrt{nm}}|M|_2$. An example where $\kappa = 1$ and $\|M\|_{\max} \ll \frac{\mu_0\sqrt{r}|M|_2}{\sqrt{nm}}$ (i.e. the bound in Lemma 14 is extremely loose) is the case where the spiky columns of U do not align with the spiky columns of V, for example writing $n = m = N + 1$ we have:

$$M = \begin{pmatrix} 1 & 0 \\ 0 & N^{-1/2} \\ 0 & N^{-1/2} \\ \dots & \dots \\ 0 & N^{-1/2} \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ N^{-1/2} & 0 \\ N^{-1/2} & 0 \\ \dots & \dots \\ N^{-1/2} & 0 \end{pmatrix}^T = \begin{pmatrix} N^{-1/4} & 0 \\ 0 & N^{-1/4} \\ 0 & N^{-1/4} \\ \dots & \dots \\ 0 & N^{-1/4} \end{pmatrix} \cdot \begin{pmatrix} 0 & N^{-1/4} \\ N^{-1/4} & 0 \\ N^{-1/4} & 0 \\ \dots & \dots \\ N^{-1/4} & 0 \end{pmatrix}^T .$$

  Since the left-hand factorization is an SVD of M (omitting $\Sigma = I_2$), we therefore have $\mu_0\sqrt{r} \cdot \frac{|M|_2}{\sqrt{nm}} = 1$ while the right-hand factorization shows that $\|M\|_{\max} \leq \frac{1}{\sqrt{n-1}}$.
- Large condition numbers κ can often lead to the same situation, in which the max norm is far lower than the bound implied by Lemma 14. For example, if low-rank M is a matrix where $\|M\|_{\max} \approx \frac{\kappa\mu_0\sqrt{r}\cdot|M|_2}{\sqrt{nm}}$, but if we perturb M slightly and add an extremely low singular value, then κ becomes extremely high while $\|M\|_{\max}$ is only slightly perturbed.

5. Summary

We presented low rank matrix reconstruction guarantees based on an existing analysis of the Rademacher complexity of low trace-norm and low max-norm matrices, and carefully compared these to other recently presented results. We view the main contributions of this papers as:

- Following a string of results on low-rank matrix reconstruction, showing that an existing Rademacher complexity analysis combined with simple arguments on the relationship between the rank, max-norm, and trace-norm, can yield guarantees that are in several ways better, and relying on weaker assumptions.

- Pointing out that the max-norm can yield superior reconstruction guarantees over the more commonly used trace-norm.
- Studying the issue of sampling with and without replacement, and establishing rigorous generic results relating the two settings. This has been done before for exact recovery (Recht, 2009), but is done here for the more delicate situation of approximate recovery of either M or Y.

The main deficiency of our approach is a worse dependence on the approximation parameter ϵ, when $\sigma > 0$ (i.e. the approximately low rank case) and $\epsilon = \mathbf{o}(\sigma^2)$ (i.e. estimation error less then approximation error). Although this dependence is tight for general classes with bounded Rademacher complexity, we do not know if it can be improved in Theorem 6. In particular, we do not know whether the less favorable dependence is a consequence of not relying on zero-mean i.i.d. noise, or not relying on M having low-rank (instead of only assuming low max-norm), or on relying only on the Rademacher complexity of the class of low max-norm matrices—perhaps better bounds can be obtained with a more careful analysis.

References

P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. In *Computational Learning Theory*, pages 224–240. Springer, 2001.

E.J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6): 925–936, 2010.

E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

E.J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

A. Chistov and D. Grigoriev. Complexity of quantifier elimination in the theory of algebraically closed fields. In *Proceedings of the 11th Symposium on Mathematical Foundations of Computer Science*, volume 176 of *Lecture Notes in Computer Science*, pages 17–31. Springer, 1984.

M. Fazel, H. Hindi, and S.P. 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. IEEE, 2002.

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

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

T. Lee, A. Shraibman, and R. Spalek. A direct product theorem for discrepancy. In *23rd Annual IEEE Conference on Computational Complexity (CCC'08)*, pages 71–80. IEEE, 2008.

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

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

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: 2056–2064, 2010.

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

A.A. Sherstov. Halfspace matrices. In *22nd Annual IEEE Conference on Computational Complexity (CCC'07)*, pages 83–95. IEEE, 2007.

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

N. Srebro, J.D.M. Rennie, and T.S. Jaakkola. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems*, 17:1329–1336, 2005.

N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. *Advances in Neural Information Processing Systems*, 23:2199–2207, 2010.

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

Appendix A. Proof of Sampling-Without-Replacement Lemmas

Proof *(Lemmas 11 and 12).* Let $\mathbb{S}_r = \{S \in \mathcal{X}^s : \text{ each } x \in \mathcal{X} \text{ appears at most } r \text{ times in } S\}$. Let $S \sim \mathcal{D}^s_r$ denote a sample S drawn uniformly from $\mathbb{S}_r$. In particular, $\mathcal{D}^s_0 = \mathcal{D}^s_{w/o}$ and $\mathcal{D}^s_s = \mathcal{D}^s$. By Lemma 15 (proved below), for any r,

$$E_{S\sim\mathcal{D}^s_{w/o}}\left(\sup_{h\in\mathcal{H}} L(h) - \hat{L}_S(h)\right) \le E_{S\sim\mathcal{D}^s_r}\left(\sup_{h\in\mathcal{H}} L(h) - \hat{L}_S(h)\right) ,$$

$$P_{S\sim\mathcal{D}^s_{w/o}}\left(\sup_{h\in\mathcal{H}} g(L(h)) - \hat{L}_S(h) \ge c\right) \le r! \cdot P_{S\sim\mathcal{D}^s_r}\left(\sup_{h\in\mathcal{H}} g(L(h)) - \hat{L}_S(h) \ge c\right) .$$

Taking the first inequality with $r = s$, this completes the proof for Lemma 11.

Now we complete the proof of Lemma 12. Take $S \sim \mathcal{D}^s$ and write $S = \{e_1, \dots, e_s\}$. For any $i_1 < i_2 < \dots < i_{K+1}$,

$$P\left(e_{i_1} = e_{i_2} = \dots = e_{i_{K+1}}\right) = \frac{1}{(nm)^K} ,$$

and so for any K with $(K+1)! \ge 2s$, the probability that any entry of the matrix appears at least $(K+1)$ times in S is bounded by

$$\binom{s}{K+1} \cdot \frac{1}{(nm)^K} \le \frac{s^{K+1}}{(K+1)!(nm)^K} \le \frac{s}{(K+1)!} \le \tfrac{1}{2} .$$

Fix the smallest K such that $(K+1)! \ge 2s$. This implies $K! < 2s$. We then have

$$\begin{aligned}
&P_{S\sim\mathcal{D}^s_{w/o}}\left(\sup_{h\in\mathcal{H}} g(L(h)) - \hat{L}_S(h) \ge c\right)\\
&\le K! \cdot P_{S\sim\mathcal{D}^s_K}\left(\sup_{h\in\mathcal{H}} g(L(h)) - \hat{L}_S(h) \ge c\right)\\
&\le K! \cdot \left(P_{S\sim\mathcal{D}^s}\left(\text{each } x \in \mathcal{X} \text{ appears } \le K \text{ times in } S\right)\right)^{-1} \cdot P_{S\sim\mathcal{D}^s}\left(\sup_{h\in\mathcal{H}} g(L(h)) - \hat{L}_S(h) \ge c\right)\\
&\le 2K! \cdot P_{S\sim\mathcal{D}^s}\left(\sup_{h\in\mathcal{H}} g(L(h)) - \hat{L}_S(h) \ge c\right)\\
&\le 4s \cdot P_{S\sim\mathcal{D}^s}\left(\sup_{h\in\mathcal{H}} g(L(h)) - \hat{L}_S(h) \ge c\right) .
\end{aligned}$$

This is completes the proof for Lemma 12.

■

Lemma 15 *Using the notation of the proof above, for any r,*

$$E_{S\sim\mathcal{D}^s_{w/o}}\left(\sup_{h\in\mathcal{H}} L(h) - \hat{L}_S(h)\right) \le E_{S\sim\mathcal{D}^s_r}\left(\sup_{h\in\mathcal{H}} L(h) - \hat{L}_S(h)\right) ,$$

$$P_{S\sim\mathcal{D}^s_{w/o}}\left(\sup_{h\in\mathcal{H}} g(L(h)) - \hat{L}_S(h) \ge c\right) \le r! \cdot P_{S\sim\mathcal{D}^s_r}\left(\sup_{h\in\mathcal{H}} g(L(h)) - \hat{L}_S(h) \ge c\right) .$$

Proof

Write $\Omega = [n] \times [m]$. Let $\alpha(S)$ be any function of the sample S, where S may contain repeated entries. Assume that, for any $S, S_1, \dots, S_r$ of equal size such that $r \cdot S = S_1 + \cdots + S_r$, $\alpha(\cdot)$ satisfies the following for some function $a(r)$:

$$a(r) \cdot \alpha(S) \leq \sum_{i=1}^{r} \alpha(S_i) \ . \tag{26}$$

Consider all samples from Ω, drawn with replacement. For a sample set S of size s, for $i = 1, \dots, s$, let $N_i(S)$ equal the number of elements of Ω appearing exactly i times in S, which obeys $\sum_i iN_i(S) = s$. We call $\mathbf{N}(S) = (N_1(S), \dots, N_s(S))$ the multiplicity vector of S; note that, when convenient, we might write $\mathbf{N}(S)$ to have length greater than s (filling the last terms with zeros). From this point on, we will regard these samples as ordered lists, and assume that in any sample, S is ordered in the format

$$(\omega_1^1, \dots, \omega_{N_1(S)}^1, \omega_1^2, \omega_1^2, \dots, \omega_{N_2(S)}^2, \omega_{N_2(S)}^2, \omega_1^3, \omega_1^3, \omega_1^3, \dots) \ ,$$

where for any i we might permute the ω_j^i's.

Let $\mathbf{N}$ be any multiplicity vector, of the form $(N_1, \dots, N_r, 0, \dots, 0)$ for some $r \leq s$. Let $\mathbf{N}'$ and $\mathbf{N}''$ be multiplicity vectors derived from $\mathbf{N}$ as follows:

$$N_i' = \begin{cases} N_1 + rN_r, & i = 1 \\ N_i, & 2 \leq i \leq r-1 \\ 0, & i \geq r \end{cases} \ , \quad N_i'' = \begin{cases} N_i, & 1 \leq i \leq r-1 \\ 0, & i \geq r \end{cases}$$

Define $s = \sum_i iN_i$. Note that $\sum_i iN_i' = s$ and $\sum_i iN_i'' = s - rN_r$.

Let $\mathbb{S} = \{S : \mathbf{N}(S) = \mathbf{N}\}$, $\mathbb{S}' = \{S : \mathbf{N}(S) = \mathbf{N}'\}$, $\mathbb{S}'' = \{S : \mathbf{N}(S) = \mathbf{N}''\}$. We will first prove that $E_{S' \sim Unif(\mathbb{S}')}\left[\alpha(S')\right] \leq E_{S \sim Unif(\mathbb{S})}\left[\alpha(S)\right]$, and then induct on r.

First consider $\mathbb{S}'$. We have

$$|\mathbb{S}'| E_{S' \sim Unif(\mathbb{S}')}\left[\alpha(S')\right] = \sum_{S' \in \mathbb{S}'} \left[\alpha(S')\right] = \sum_{S'' \in \mathbb{S}''} \sum_{\substack{A_1, \dots, A_r \subset \Omega \backslash S'' \\ |A_j| = N_r \\ A_j\text{'s disjoint}}} \left[\alpha(S'' + A_1 + \cdots + A_r)\right] \ .$$

The last equality arises when, starting with some $S' \in \mathbb{S}'$, we recall that S'' is an ordered sample set beginning with the $N_1 + rN_r$ elements which appear exactly once. Let S'' be the first N_1 elements of S', then let A_1 be the next N_r elements of S', let A_2 be the next N_r elements of S', etc.

Next consider $\mathbb{S}$. As before, we have

$$|\mathbb{S}| E_{S \sim Unif(\mathbb{S})}\left[\alpha(S)\right] = \sum_{S \in \mathbb{S}} \left[\alpha(S)\right] = \sum_{S'' \in \mathbb{S}''} \sum_{\substack{A \subset \Omega \backslash S \\ |A| = N_r}} \left[\alpha(S'' + r \cdot A)\right] \ .$$

By counting how many times each choice of A appears in the sum below, and then rescaling accordingly, we get

$$= \left(\frac{(nm - N_1 - \cdots - N_r)!}{(nm - N_1 - \cdots - N_{r-1} - rN_r)!} \right)^{-1} r^{-1} \sum_{S'' \in \mathbb{S}''} \sum_{\substack{A_1, \dots, A_r \subset \Omega \backslash S'' \\ |A_j| = N_r \\ A_j\text{'s disjoint}}} \sum_j \left[\alpha(S'' + r \cdot A_j)\right]$$

$$\geq \left(\frac{(nm - N_1 - \cdots - N_r)!}{(nm - N_1 - \cdots - N_{r-1} - rN_r)!} \right)^{-1} \frac{a(r)}{r} \sum_{S'' \in \mathbb{S}''} \sum_{\substack{A_1,\ldots,A_r \subset \Omega \setminus S'' \\ |A_j| = N_r \\ A_j\text{'s disjoint}}} \alpha(S'' + A_1 + \cdots + A_r) \; .$$

To summarize so far, we have

$$|\mathbb{S}| E_{S \sim Unif(\mathbb{S})} \left[\alpha(S) \right] \geq \left(\frac{(nm - N_1 - \cdots - N_r)!}{(nm - N_1 - \cdots - N_{r-1} - rN_r)!} \right)^{-1} \cdot \frac{a(r)}{r} |\mathbb{S}'| E_{S' \sim Unif(\mathbb{S}')} \left[\alpha(S') \right] \; .$$

Next, we see that (since sample sets are treated as ordered)

$$|\mathbb{S}| = \frac{(nm)!}{(nm - N_1 - \cdots - N_r)!}, \; |\mathbb{S}'| = \frac{(nm)!}{(nm - N_1 - \cdots - N_{r-1} - rN_r)!}$$

Therefore,

$$E_{S \sim Unif(\mathbb{S})} \left[\alpha(S) \right] \geq \frac{a(r)}{r} \cdot E_{S' \sim Unif(\mathbb{S}')} \left[\alpha(S') \right] \; .$$

By inducting over r, we then see that

$$E_{S \sim Unif(\mathbb{S})} \left[\alpha(S) \right] \geq \frac{\prod_{i=1}^r a(i)}{r!} \cdot E_{S \sim \mathcal{D}^s_{w/o}} \left[\alpha(S) \right] \; ,$$

where $\mathbb{S} = \{S : \mathbf{N}(S) = \mathbf{N}\}$ for any multiplicity vector $\mathbf{N} = (N_1, \ldots, N_r, 0, \ldots, 0)$. Therefore,

$$E_{S \sim \mathcal{D}^s_r} \left[\alpha(S) \right] \geq \frac{\prod_{i=1}^r a(i)}{r!} \cdot E_{S \sim \mathcal{D}^s_{w/o}} \left[\alpha(S) \right] \; ,$$

Finally, we observe that if $\alpha(S) = \sup_{h \in \mathcal{H}} L(h) - \hat{L}_S(h)$, then $\alpha(S)$ satisfies (26) with $a(r) = r$, while if $\alpha(S) = \mathbb{I}\left\{ \sup_{h \in \mathcal{H}} g(L(h)) - \hat{L}_S(h) \geq c \right\}$, then $\alpha(S)$ satisfies (26) with $a(r) = 1$. This concludes the proof.

■

Proof *(Lemma 13.)*

Suppose $s \leq \frac{K+1}{e} (nm)^{1 - \frac{1}{K+1}}$. Then, as in the proof of Lemma 12,

$$P\left(\text{any entry is sampled more than } K \text{ times}\right) \leq \binom{s}{K} \cdot \frac{1}{(nm)^{K-1}}$$

$$\leq \frac{s^{K+1}}{(K+1)!(nm)^K} \leq \frac{((K+1)/e)^{K+1} (nm)^K}{(K+1)!(nm)^K} \leq \frac{1}{2}, \text{ by Stirling's approximation.}$$

We show below that, for any c,

$$P_{S \sim \mathcal{D}^s_{w/o}} \left(\sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq c \right) \leq 2K \cdot P_{S \sim \mathcal{D}^s_K} \left(\sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq (2K)^{-1} c \right) \; ,$$

where $\mathcal{D}^s_{w/o}$ and $\mathcal{D}^s_K$ are defined as in the proof of Lemmas 11 and 12, except with the independent noise model. As in the proof of Lemmas 11 and 12, this implies that

$$P_{S \sim \mathcal{D}^s_{w/o}} \left(\sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq c \right) \leq 4K \cdot P_{S \sim \mathcal{D}^s} \left(\sup_{X \in \mathcal{X}} g(L(X)) - \hat{L}_S(X) \geq (2K)^{-1} c \right) \; .$$

We now prove that, for any c,

$$P_{S\sim\mathcal{D}^s_{w/o}}\left(\sup_{X\in\mathcal{X}} g(L(X))-\hat{L}_S(X)\geq c\right)\leq 2K\cdot P_{S\sim\mathcal{D}^s_K}\left(\sup_{X\in\mathcal{X}} g(L(X))-\hat{L}_S(X)\geq (2K)^{-1}c\right)\ .$$

Write $\Omega=[n]\times[m]$. Consider all samples from Ω, drawn with replacement. When a particular (i,j) is drawn multiple times, then the observed values at that entry of the matrix follow the independent noise model as described in the statement of Theorem 9.

For a sample set S of size s, for $i=1,\dots,s$, define $\mathbf{N}(S)$ as in the proof of Lemma 15. Let $\mathbf{N}$ be any multiplicity vector, of the form $(N_1,\dots,N_r,0,\dots,0)$ for some $r\leq s$. Let $\mathbf{M}$ be a multiplicity vector defined from $\mathbf{N}$ as follows:

$$\mathbf{M}=(M_i)_i,\ \text{where}\ M_i=N_{2i-1}+2N_i+N_{i+1}\ .$$

Now take any $A_1,A_2,\dots,A_{2r},B_2,\dots,B_{2r}\subset[n]\times[m]$, all disjoint, with $|A_i|=|B_i|=N_i$ for all i. Define $B_1=A_1$, and

$$S_A=\sum_{i=1}^{2r}\left(\sum_{j=1}^{i}A_i^{(j)}\right)\ ,\ S_B=\sum_{i=1}^{2r}\left(\sum_{j=1}^{i}B_i^{(j)}\right)\ .$$

Note that $\mathbf{N}(S_A)=\mathbf{N}(S_B)=\mathbf{N}$. Now define

$$T_1=\sum_{i=1}^{2r}\left(\sum_{j=1}^{\lfloor\frac{i}{2}\rfloor}A_i^{(j)}+\sum_{j=\lfloor\frac{i}{2}\rfloor+1}^{i}B_i^{(j)}\right)\ ,\ T_2=\sum_{i=1}^{2r}\left(\sum_{j=1}^{\lfloor\frac{i}{2}\rfloor}B_i^{(j)}+\sum_{j=\lfloor\frac{i}{2}\rfloor+1}^{i}A_i^{(j)}\right)\ .$$

Note that $\mathbf{N}(T_1)=\mathbf{N}(T_2)=\mathbf{M}$, and that up to reordering, $S_A+S_B=T_1+T_2$. We treat T_1 and T_2 as functions of (S_A,S_B).

Write $\alpha_c(S)=\mathbb{I}\left\{\sup_{X\in\mathcal{X}} g(L(X))-\hat{L}_S(X)\geq c\right\}$. Then α satisfies the following whenever $|S_1|=|S_2|$:

$$\frac{1}{2}\left(\alpha_{2c}(S_1)+\alpha_{2c}(S_2)\right)\leq\alpha_c(S_1+S_2)\leq\alpha_c(S_1)+\alpha_c(S_2)\ .$$

Therefore,

$$\begin{aligned}
&2E_{S\sim Unif(\mathbf{N})}\left(\alpha_c(S)\right)\\
&= (\#(S_A,S_B) \text{ pairs as above})^{-1} \sum_{(S_A,S_B) \text{ as above}} \alpha_c(S_A)+\alpha_c(S_B)\\
&\geq (\#(S_A,S_B) \text{ pairs as above})^{-1} \sum_{(S_A,S_B) \text{ as above}} \alpha_c(S_A+S_B)\\
&= (\#(S_A,S_B) \text{ pairs as above})^{-1} \sum_{(S_A,S_B) \text{ as above}} \alpha_c(T_1+T_2)\\
&\geq (\#(S_A,S_B) \text{ pairs as above})^{-1} \sum_{(S_A,S_B) \text{ as above}} \frac{1}{2}\left(\alpha_{2c}(T_1)+\alpha_{2c}(T_2)\right)\\
&= (\#(S_A,S_B) \text{ pairs as above})^{-1} \sum_{(S_A,S_B) \text{ as above}} \alpha_{2c}(T_1)\\
&= (\#(S_A,S_B) \text{ pairs as above})^{-1} \sum_{T:\mathbf{N}(T)=\mathbf{M}} \alpha_{2c}(T)\cdot(\#(S_A,S_B) \text{ pairs such that } T=T_1)
\end{aligned}$$

We also have the following (note that here we treat samples as unordered, unlike in the proofs of Lemmas 11 and 12):

$$(\#(S_A,S_B) \text{ pairs as above}) = \binom{nm}{N_1,N_2,N_2,N_3,N_3,\ldots,N_{2r},N_{2r}},$$

and for any T with $\mathbf{N}(T)=\mathbf{M}$,

$$(\#(S_A,S_B) \text{ pairs such that } T=T_1) = \prod_{i=1}^{r}\binom{M_i}{N_{2i-1},N_{2i},N_{2i},N_{2i+1}}.$$

Finally,

$$(\#T:\mathbf{N}(T)=\mathbf{M}(T)) = \binom{nm}{M_1,M_2,\ldots,M_r},$$

and therefore, continuing from above,

$$\begin{aligned}
&2E_{S\sim Unif(\mathbf{N})}\left(\alpha_c(S)\right)\\
&= (\#(S_A,S_B) \text{ pairs as above})^{-1} \sum_{T:\mathbf{N}(T)=\mathbf{M}} \alpha_{2c}(T)\cdot(\#(S_A,S_B) \text{ pairs such that } T=T_1)\\
&= (\#T:\mathbf{N}(T)=\mathbf{M})^{-1} \sum_{T:\mathbf{N}(T)=\mathbf{M}} \alpha_{2c}(T)\\
&= E_{T\sim Unif(\mathbf{M})}(\alpha_{2c}(T))\,.
\end{aligned}$$

Inducting over r, we see that for any $\mathbf{N}=(N_1,\ldots,N_r,0,\ldots,0,$

$$2^{K(r)}E_{S\sim Unif(\mathbf{N})}(\alpha_c(S)) \geq E_{S\sim\mathcal{D}^s_{w/o}}(\alpha_{2^{K(r)}}(S))\,,$$

where $K(r)$ is the number of times that the operation $x \mapsto \lceil x/2 \rceil$ must be applied iteratively to r to obtain 1; note that $2^{K(r)} \le 2r$. Therefore,

$$P_{S\sim\mathcal{D}^s_{w/o}}\left(\sup_{X\in\mathcal{X}} g(L(x)) - \hat{L}_S(X) \ge c\right) \le 2r P_{S\sim\mathcal{D}^s_r}\left(\sup_{X\in\mathcal{X}} g(L(x)) - \hat{L}_S(X) \ge (2r)^{-1}c\right) .$$

■

Appendix B. The Rademacher Complexity of the Trace-Norm Ball

Srebro and Shraibman (2005) established that for a sample $S = \{(i_1, j_1), \ldots, (i_s, j_s)\}$ of s index-pairs, the empirical Rademacher complexity of the trace-norm ball, viewed a predictor of entries, is given by:

$$\begin{aligned}\hat{\mathcal{R}}_s\left(\left\{(i,j) \mapsto X_{ij} \,\middle|\, X \in \mathbb{R}^{n\times m}, \|X\|_\Sigma \le A\right\}\right) &= \mathbf{E}_\xi\left[\sup_{\|X\|_\Sigma \le A} \frac{1}{s}\sum_{t=1}^{s} \xi_t X_{(i_t,j_t)}\right] \\ &= \frac{A}{s}\mathbf{E}_\xi\left[\left\|\sum_{t=1}^{s} \xi_t e_{i_t,j_t}\right\|_2\right] , \qquad (27)\end{aligned}$$

where the expectations is over independent uniformly distributed random variables $\xi_1, \ldots, \xi_t \in \pm 1$, $\|X\|_2$ is the spectral norm (maximal singular value) of X, and $e_{i,j} = e_i e_j^T$ is a matrix with a single 1 at location (i,j) and zeros elsewhere. Analyzing the Rademacher complexity then amounts to analyzing the expected spectral norm of the random matrix $Q = \sum_{t=1}^{s} \xi_t e_{i_t,j_t}$.

The worst-case Rademacher complexity, i.e. the supermum of (27) over all samples S, is $\frac{1}{\sqrt{s}}$, and does not lead to meaningful generalization results. Indeed, if we could meaningfully bound the worst-case Rademacher complexity, we could guarantee learning under arbitrary sampling distributions over index-pairs, but this is not the case—we know that trace-norm regularization can fail when entries are not sampled uniformly (Salakhutdinov and Srebro, 2010).

Instead, we focus on bounding the *expected* Rademacher complexity, i.e. the expectation of (27) when entries in S are chosen independently from a *uniform* distribution over index pairs. Srebro and Shraibman (2005) bounded the expected Rademacher complexity by $\mathbf{O}\left(\frac{A}{\sqrt{nm}}\sqrt{\frac{(n+m)\log^{3/2} n}{s}}\right)$ using a bound of Seginer (2000) on the spectral norm of a matrix with fixed magnitudes and random signs, combined with arguments bounding the number of observations in each row and column. Here we present a much simpler analysis, reducing the logarithmic factor from $\log^{3/2}(n)$ to $\log(n)$, using a recent result of Tropp (2010).

We now proceed to bounding $\mathbf{E}\left[\|Q\|_2\right]$, where the expectation is over the sample S and the random signs ξ_t. Denote $P_t = \xi_t e_{i_t,j_t}$, we have $Q = \sum_t P_t$ and P_t are i.i.d. zero-mean random matrices (recall that now both ξ_t and (i_t, j_t) are random). Theorem 6.1 of Tropp (2010), combined with Remarks 6.3 and 6.5, allows us to bound the expected spectral norm of such a sum of independent random matrices by:

$$\mathbf{E}\left[\|Q\|\right] = \mathbf{O}\left(\sigma\sqrt{\log(n+m)} + R\log(n+m)\right) , \qquad (28)$$

where $\|P_t\|_2 \leq R$ (almost surely) and

$$\sigma^2 = \max\left(\left\| \sum \mathbf{E}\left[P_t^T P_t\right] \right\|_2, \left\| \sum \mathbf{E}\left[P_t P_t^T\right] \right\|_2 \right).$$

For each t, P_t is just a matrix with a single $+1$ or -1, hence $\|P_t\| \leq 1$. The matrix $P_t P_t^T \in \mathbb{R}^{n\times n}$ is equal to $e_{i,i}$ with probability $\frac{1}{n}$, hence $\mathbf{E}\left[P_t P_t^T\right] = \frac{1}{n} I_n$ and $\left\|\sum \mathbf{E}\left[P_t P_t^T\right]\right\|_2 = \left\|\frac{s}{n} I_n\right\| = \frac{s}{n}$. Symmetrically, $\left\|\sum \mathbf{E}\left[P_t^T P_t\right]\right\|_2 = \frac{s}{m}$ and so $\sigma^2 = \frac{s}{nm}\max(n,m)$. Plugging σ and T into (28) we have:

$$\mathbf{E}\left[\|Q\|_2\right] = O\left(\sqrt{\frac{s(n+m)\log(n+m)}{nm}} + \log(n+m)\right) = O\left(\sqrt{\frac{s(n+m)\log(n+m)}{nm}}\right) \tag{29}$$

where in the second inequality we assume $s \geq m$. Plugging (29) into (27) we get:

$$\mathbf{E}\left[\hat{\mathcal{R}}_s\left(\left\{(i,j) \to X_{ij} \,\middle|\, X \in \mathbb{R}^{n\times m}, \|X\|_\Sigma \leq A\right\}\right)\right] = O\left(\frac{A}{\sqrt{nm}}\sqrt{\frac{(n+m)\log(n+m)}{s}}\right) \tag{30}$$