Restricted Strong Convexity and Weighted Matrix Completion: Optimal Bounds with Noise

Sahand Negahban   SAHAND_N@EECS.BERKELEY.EDU

Martin J. Wainwright[*]   WAINWRIG@STAT.BERKELEY.EDU

Department of Electrical Engineering and Computer Sciences

University of California

Berkeley, CA 94720-1776, USA

Editor: Inderjit Dhillon

Abstract

We consider the matrix completion problem under a form of row/column weighted entrywise sampling, including the case of uniform entrywise sampling as a special case. We analyze the associated random observation operator, and prove that with high probability, it satisfies a form of restricted strong convexity with respect to weighted Frobenius norm. Using this property, we obtain as corollaries a number of error bounds on matrix completion in the weighted Frobenius norm under noisy sampling and for both exact and near low-rank matrices. Our results are based on measures of the "spikiness" and "low-rankness" of matrices that are less restrictive than the incoherence conditions imposed in previous work. Our technique involves an *M*-estimator that includes controls on both the rank and spikiness of the solution, and we establish non-asymptotic error bounds in weighted Frobenius norm for recovering matrices lying with ℓ_q-"balls" of bounded spikiness. Using information-theoretic methods, we show that no algorithm can achieve better estimates (up to a logarithmic factor) over these same sets, showing that our conditions on matrices and associated rates are essentially optimal.

Keywords: matrix completion, collaborative filtering, convex optimization

1. Introduction

Matrix completion problems correspond to reconstructing matrices, either exactly or approximately, based on observing a subset of their entries (Laurent, 2001; Deza and Laurent, 1997). In the simplest formulation of matrix completion, the observations are assumed to be uncorrupted, whereas a more general formulation (as considered in this paper) allows for noisiness in these observations. Matrix recovery based on only partial information is an ill-posed problem, and accurate estimates are possible only if the matrix satisfies additional structural constraints, with examples including bandedness, positive semidefiniteness, Euclidean distance measurements, Toeplitz, and low-rank structure (see the survey paper by Laurent (2001) and references therein for more background).

The focus of this paper is low-rank matrix completion based on noisy observations. This problem is motivated by a variety of applications where an underlying matrix is likely to have low-rank, or near low-rank structure. The archetypal example is the Netflix challenge, a version of the collaborative filtering problem, in which the unknown matrix is indexed by individuals and movies, and each observed entry of the matrix corresponds to the rating assigned to the associated movie

---

[*]. Also in the department of Statistics.

by the given individual. Since the typical person only watches a tiny number of movies (compared to the total Netflix database), it is only a sparse subset of matrix entries that are observed. In this context, one goal of collaborative filtering is to use the observed entries to make recommendations to a person regarding movies that they have *not* yet seen. We refer the reader to Srebro (2004) (and references therein) for further discussion and motivation for collaborative filtering and related problems.

In this paper, we analyze a method for approximate low-rank matrix recovery using an *M*-estimator that is a combination of a data term, and a weighted nuclear norm as a regularizer. The nuclear norm is the sum of the singular values of a matrix (Horn and Johnson, 1985), and has been studied in a body of past work, both on matrix completion and more general problems of low-rank matrix estimation (e.g., Fazel, 2002; Srebro, 2004; Srebro et al., 2005, 2004; Recht et al., 2010; Bach, 2008; Candes and Tao, 2010; Recht, 2011; Keshavan et al., 2010a,b; Negahban and Wainwright, 2011; Rohde and Tsybakov, 2011). A parallel line of work has studied computationally efficient algorithms for solving problems with nuclear norm constraints (e.g, Mazumber et al., 2010; Nesterov, 2007; Lin et al., 2009). Here we limit our detailed discussion to those papers that study various aspects of the matrix completion problem. Motivated by various problems in collaborative filtering, Srebro (2004) and Srebro et al. (2005) studied various aspects nuclear norm regularization, and established generalization error bounds under certain conditions. Candès and Recht (2009) studied the exact reconstruction of a low-rank matrix given perfect (noiseless) observations of a subset of entries, and provided sufficient conditions for exact recovery via nuclear norm relaxation, with later refinements provided by various authors (Candes and Tao, 2010; Recht, 2011; Gross, 2011). In particular, Gross (2011) recognized the utility of the Ahlswede-Winter matrix concentration bounds, and the simplest argument to date is provided by Recht (2011). In a parallel line of work, Keshavan et al. (2010a,b) have studied a method based on thresholding and singular value decomposition, and established various results on its behavior, both for noiseless and noisy matrix completion. Among other results, Rohde and Tsybakov (2011) establish prediction error bounds for matrix completion, a different metric than the matrix recovery problem of interest here. In recent work, Salakhutdinov and Srebro (2010) provided various motivations for the use of weighted nuclear norms, in particular showing that the standard nuclear norm relaxation can behave very poorly when the sampling is non-uniform. The analysis of this paper applies to both uniform and non-uniform sampling, as well as a form of reweighted nuclear norm as suggested by these authors, one which includes the ordinary nuclear norm as a special case. We provide a more detailed comparison between our results and some aspects of past work in Section 3.4.

As has been noted before (Candès and Plan, 2010), a significant theoretical challenge is that conditions that have proven very useful for sparse linear regression—among them the restricted isometry property—are *not* satisfied for the matrix completion problem. For this reason, it is natural to seek an alternative and less restrictive property that might be satisfied in the matrix completion setting. In recent work, Negahban et al. (2009) have isolated a weaker and more general condition known as *restricted strong convexity* (RSC), and proven that certain statistical models satisfy RSC with high probability when the associated regularizer satisfies a *decomposability* condition. When an *M*-estimator satisfies the RSC condition, it is relatively straightforward to derive non-asymptotic error bounds on parameter estimates (Negahban et al., 2009). The class of decomposable regularizers includes the nuclear norm as particular case, and the RSC/decomposability approach has been exploited to derive bounds for various matrix estimation problems, among them multi-task learning, autoregressive system identification, and compressed sensing (Negahban and Wainwright, 2011).

To date, however, an open question is whether or not an appropriate form of RSC holds for the matrix completion problem. If it did hold, then it would be possible to derive non-asymptotic error bounds (in Frobenius norm) for matrix completion based on noisy observations. Within this context, the main contribution of this paper is to prove that with high probability, a form of the RSC condition holds for the matrix completion problem, in particular over an interesting set of matrices $\mathfrak{C}$, as defined in Equation (4) to follow, that have both low nuclear/Frobenius norm ratio and low "spikiness". Exploiting this RSC condition then allows us to derive non-asymptotic error bounds on matrix recovery in weighted Frobenius norms, both for exactly and approximately low-rank matrices. The theoretical core of this paper consists of three main results. Our first result (Theorem 1) proves that the matrix completion loss function satisfies restricted strong convexity with high probability over the set $\mathfrak{C}$. Our second result (Theorem 2) exploits this fact to derive a non-asymptotic error bound for matrix recovery in the weighted Frobenius norm, one applicable to general matrices. We then specialize this result to the problem of estimating exactly low-rank matrices (with a small number of non-zero singular values), as well as near low-rank matrices characterized by relatively swift decay of their singular values. To the best of our knowledge, our results on near low-rank matrices are the first for approximate matrix recovery in the noisy setting, and as we discuss at more length in Section 3.4, our results on the exactly low-rank case are sharper than past work on the problem. Indeed, our final result (Theorem 3) uses information-theoretic techniques to establish that up to logarithmic factors, no algorithm can obtain faster rates than our method over the ℓ_q-balls of matrices with bounded spikiness treated in this paper.

The remainder of this paper is organized as follows. We begin in Section 2 with background and a precise formulation of the problem. Section 3 is devoted to a statement of our main results, and discussion of some of their consequences. In Sections 4 and Section 5, we prove our main results, with more technical aspects of the arguments deferred to appendices. We conclude with a discussion in Section 6.

2. Background and Problem Formulation

In this section, we introduce background on low-rank matrix completion problem, and also provide a precise statement of the problem studied in this paper.

2.1 Uniform and Weighted Sampling Models

Let $\Theta^* \in \mathbb{R}^{d_r \times d_c}$ be an unknown matrix, and consider an observation model in which we make n i.i.d. observations of the form

$$\widetilde{y}_i = \Theta^*_{j(i)k(i)} + \frac{\nu}{\sqrt{d_r d_c}} \widetilde{\xi}_i, \tag{1}$$

Here the quantities $\frac{\nu}{\sqrt{d_r d_c}} \widetilde{\xi}_i$ correspond to additive observation noises with variance appropriately scaled according to the matrix dimensions. In defining the observation model, one can either allow the Frobenius norm of Θ^* to grow with the dimension, as in done in other work (Candès and Plan, 2010; Keshavan et al., 2010b), or rescale the noise as we have done here. This choice is consistent with our assumption that Θ^* has constant Frobenius norm regardless of its rank or dimensions. With this scaling, each observation in the model (1) has a constant signal-to-noise ratio regardless of matrix dimensions.

In the simplest model, the row $j(i)$ and column $k(i)$ indices are chosen uniformly at random from the sets $\{1, 2, \ldots, d_r\}$ and $\{1, 2, \ldots, d_c\}$ respectively. In this paper, we consider a somewhat more general weighted sampling model. In particular, let $R \in \mathbb{R}^{d_r \times d_r}$ and $C \in \mathbb{R}^{d_c \times d_c}$ be diagonal matrices, with rescaled diagonals $\{R_j/d_r, j = 1, 2, \ldots, d_r\}$ and $\{C_k/d_c, k = 1, 2, \ldots, d_c\}$ representing probability distributions over the rows and columns of an $d_r \times d_c$ matrix. We consider the weighted sampling model in which we make a noisy observation of entry (j, k) with probability $R_j C_k/(d_r d_c)$, meaning that the row index $j(i)$ (respectively column index $k(i)$) is chosen according to the probability distribution R/d_r (respectively C/d_c). Note that in the special case that $R = \mathbf{1}_{d_r}$ and $C = \mathbf{1}_{d_c}$, the observation model (1) reduces to the usual model of uniform sampling.

We assume that each row and column is sampled with positive probability, in particular that there is some constant $1 \le L < \infty$ such that $R_a \ge 1/L$ and $C_b \ge 1/L$ for all rows and columns. However, apart from the constraints $\sum_{a=1}^{d_r} R_{aa} = d_r$ and $\sum_{b=1}^{d_c} C_{bb} = d_c$, we do not require that the row and column weights remain bounded as d_r and d_c tend to infinity.

2.2 The Observation Operator and Restricted Strong Convexity

We now describe an alternative formulation of the observation model (1) that, while statistically equivalent to the original, turns out to be more natural for analysis. For each $i = 1, 2, \ldots, n$, define the matrix

$$X^{(i)} = \sqrt{d_r d_c} \, \varepsilon_i \, e_{a(i)} e_{b(i)}^T,$$

where $\varepsilon_i \in \{-1, +1\}$ is a random sign, and consider the observation model

$$y_i = \langle\langle X^{(i)}, \Theta^* \rangle\rangle + \nu \xi_i, \qquad \text{for } i = 1, \ldots, n, \qquad (2)$$

where $\langle\langle A, B \rangle\rangle := \sum_{j,k} A_{jk} B_{jk}$ is the trace inner product, and ξ_i is an additive noise from the same distribution as the original model. The model (2) is can be obtained from the original model (1) by rescaling all terms by the factor $\sqrt{d_r d_c}$, and introducing the random signs ε_i. The rescaling has no statistical effect, and nor do the random signs, since the noise is symmetric (so that $\xi_i = \varepsilon_i \widetilde{\xi}_i$ has the same distribution as $\widetilde{\xi}_i$). Thus, the observation model (2) is statistically equivalent to the original one (1).

In order to specify a vector form of the observation model, let us define a linear operator $\mathfrak{X}_n : \mathbb{R}^{d_r \times d_c} \to \mathbb{R}^n$ via

$$[\mathfrak{X}_n(\Theta)]_i := \langle\langle X^{(i)}, \Theta \rangle\rangle, \quad \text{for } i = 1, 2, \ldots n.$$

We refer to $\mathfrak{X}_n$ as the *observation operator*, since it maps any matrix $\Theta \in \mathbb{R}^{d_r \times d_c}$ to an n-vector of samples. With this notation, we can write the observations (2) in a vectorized form as $y = \mathfrak{X}_n(\Theta^*) + \nu \xi$.

The reformulation (2) is convenient for various reasons. For any matrix $\Theta \in \mathbb{R}^{d_r \times d_c}$, we have $\mathbb{E}[\langle\langle X^{(i)}, \Theta \rangle\rangle] = 0$ and

$$\mathbb{E}\left[\langle\langle X^{(i)}, \Theta \rangle\rangle^2\right] = \sum_{j=1}^{d_r} \sum_{k=1}^{d_c} R_j \Theta_{jk}^2 C_k = \underbrace{\||\sqrt{R}\Theta\sqrt{C}\||_F^2}_{\||\Theta\||_{\omega(F)}^2},$$

where we have defined the *weighted Frobenius norm* $\||\cdot\||_{\omega(F)}$ in terms of the row R and column C weights. As a consequence, the signal-to-noise ratio in the observation model (2) is given by the ratio $\mathrm{SNR} = \frac{\||\Theta^*\||^2_{\omega(F)}}{v^2}$.

As shown by Negahban et al. (2009), a key ingredient in establishing error bounds for the observation model (2) is obtaining lower bounds on the restricted curvature of the sampling operator—in particular, to establish the existence of a constant $c > 0$, which may be arbitrarily small as long as it is positive, such that

$$\frac{\|\mathfrak{X}_n(\Theta)\|_2}{\sqrt{n}} \geq c \||\Theta\||_{\omega(F)}. \tag{3}$$

For sample sizes of interest for matrix completion ($n \ll d_r d_c$), one cannot expect such a bound to hold uniformly over all matrices $\Theta \in \mathbb{R}^{d_r \times d_c}$, even when rank constraints are imposed. Indeed, as noted by Candès and Plan (2010), the condition (3) is violated with high probability by the rank one matrix Θ^* such that $\Theta^*_{11} = 1$ with all other entries zero. Indeed, for a sample size $n \ll d_r d_c$, we have a vanishing probability of observing the entry Θ^*_{11}, so that $\mathfrak{X}_n(\Theta^*) = 0$ with high probability.

2.3 Controlling the Spikiness and Rank

Intuitively, one must exclude matrices that are overly "spiky" in order to avoid the phenomenon just described. Past work has relied on fairly restrictive matrix incoherence conditions (see Section 3.4 for more discussion), based on specific conditions on singular vectors of the unknown matrix Θ^*. In this paper, we formalize the notion of "spikiness" in a natural and less restrictive way—namely by comparing a weighted form of ℓ_∞-norm to the weighted Frobenius norm. In particular, for any non-zero matrix Θ, let us define (for any non-zero matrix) the *weighted spikiness ratio*

$$\alpha_{\mathrm{sp}}(\Theta) := \sqrt{d_r d_c} \ \frac{\||\Theta\||_{\omega(\infty)}}{\||\Theta\||_{\omega(F)}},$$

where $\||\Theta\||_{\omega(\infty)} := \|\sqrt{R}\Theta\sqrt{C}\|_\infty$ is the weighted elementwise ℓ_∞-norm. Note that this ratio is invariant to the scaling of Θ, and satisfies the inequalities $1 \leq \alpha_{\mathrm{sp}}(\Theta) \leq \sqrt{d_r d_c}$. We have $\alpha_{\mathrm{sp}}(\Theta) = 1$ for any non-zero matrix whose entries are all equal, whereas the opposite extreme $\alpha_{\mathrm{sp}}(\Theta) = \sqrt{d_r d_c}$ is achieved by the "maximally spiky" matrix that is zero everywhere except for a single position.

In order to provide a tractable measure of how close Θ is to a low-rank matrix, we define (for any non-zero matrix) the ratio

$$\beta_{\mathrm{ra}}(\Theta) := \frac{\||\Theta\||_{\omega(1)}}{\||\Theta\||_{\omega(F)}}$$

which satisfies the inequalities $1 \leq \beta_{\mathrm{ra}}(\Theta) \leq \sqrt{\min\{d_r, d_c\}}$. By definition of the (weighted) nuclear and Frobenius norms, note that $\beta_{\mathrm{ra}}(\Theta)$ is simply the ratio of the ℓ_1 to ℓ_2 norms of the singular values of the weighted matrix $\sqrt{R}\Theta\sqrt{C}$. This measure can also be upper bounded by the rank of Θ: indeed, since R and C are full-rank, we always have

$$\beta^2_{\mathrm{ra}}(\Theta) \leq \mathrm{rank}(\sqrt{R}\Theta\sqrt{C}) = \mathrm{rank}(\Theta),$$

with equality holding if all the non-zero singular values of $\sqrt{R}\Theta\sqrt{C}$ are identical.

3. Main Results and Their Consequences

We now turn to the statement of our main results, and discussion of their consequences. Section 3.1 is devoted to a result showing that a suitable form of restricted strong convexity holds for the random sampling operator $\mathfrak{X}_n$, as long as we restrict it to matrices Δ for which $\beta_{\mathrm{ra}}(\Delta)$ and $\alpha_{\mathrm{sp}}(\Delta)$ are not "overly large". In Section 3.2, we develop the consequences of the RSC condition for noisy matrix completion, and in Section 3.3, we prove that our error bounds are minimax-optimal up to logarithmic factors. In Section 3.4, we provide a detailed comparison of our results with past work.

3.1 Restricted Strong Convexity for Matrix Sampling

Introducing the convenient shorthand $d = \frac{1}{2}(d_r + d_c)$, let us define the constraint set

$$\mathfrak{C}(n; c_0) := \left\{ \Delta \in \mathbb{R}^{d_r \times d_c}, \Delta \neq 0 \mid \alpha_{\mathrm{sp}}(\Delta)\, \beta_{\mathrm{ra}}(\Delta) \leq \frac{1}{c_0 L} \sqrt{\frac{n}{d \log d}} \right\}, \tag{4}$$

where c_0 is a universal constant. Note that as the sample size n increases, this set allows for matrices with larger values of the spikiness and/or rank measures, $\alpha_{\mathrm{sp}}(\Delta)$ and $\beta_{\mathrm{ra}}(\Delta)$ respectively.

Theorem 1 *There are universal constants (c_0, c_1, c_2, c_3) such that as long as $n > c_3\, d \log d$, we have*

$$\frac{\|\mathfrak{X}_n(\Delta)\|_2}{\sqrt{n}} \geq \frac{1}{8} \|\!|\Delta|\!\|_{\omega(F)} \left\{ 1 - \frac{128\, \alpha_{\mathrm{sp}}(\Delta) L}{\sqrt{n}} \right\} \qquad \text{for all } \Delta \in \mathfrak{C}(n; c_0) \tag{5}$$

with probability greater than $1 - c_1 \exp(-c_2 d \log d)$.

Roughly speaking, this bound guarantees that the observation operator captures a substantial component of any matrix $\Delta \in \mathfrak{C}(n; c_0)$ that is not overly spiky. More precisely, as long as $\frac{128 L \alpha_{\mathrm{sp}}(\Delta)}{\sqrt{n}} \leq \frac{1}{2}$, the bound (5) implies that

$$\frac{\|\mathfrak{X}_n(\Delta)\|_2^2}{n} \geq \frac{1}{256} \|\!|\Delta|\!\|_{\omega(F)}^2 \qquad \text{for any } \Delta \in \mathfrak{C}(n; c_0). \tag{6}$$

This bound can be interpreted in terms of *restricted strong convexity* (Negahban et al., 2009). In particular, given a vector $y \in \mathbb{R}^n$ of noisy observations, consider the quadratic loss function

$$\mathcal{L}(\Theta; y) = \frac{1}{2n} \|y - \mathfrak{X}_n(\Theta)\|_2^2.$$

Since the Hessian matrix of this function is given by $\mathfrak{X}_n{}^* \mathfrak{X}_n / n$, the bound (6) implies that the quadratic loss is strongly convex in a restricted set of directions Δ.

As discussed previously, the worst-case value of the "spikiness" measure is $\alpha_{\mathrm{sp}}(\Delta) = \sqrt{d_r d_c}$, achieved for a matrix that is zero everywhere except a single position. In this most degenerate of cases, the combination of the constraints $\frac{\alpha_{\mathrm{sp}}(\Delta)}{\sqrt{n}} < 1$ and the membership condition $\Delta \in \mathfrak{C}(n; c_0)$ imply that even for a rank one matrix (so that $\beta_{\mathrm{ra}}(\Delta) = 1$), we need sample size $n \gg d^2$ for Theorem 1 to provide a non-trivial result, as is to be expected.

3.2 Consequences for Noisy Matrix Completion

We now turn to some consequences of Theorem 1 for matrix completion in the noisy setting. In particular, assume that we are given n i.i.d. samples from the model (2), and let $\widehat{\Theta}$ be some estimate of the unknown matrix Θ^*. Our strategy is to exploit the lower bound (5) in application to the error matrix $\widehat{\Theta} - \Theta^*$, and accordingly, we need to ensure that it has relatively low-rank and spikiness. Based on this intuition, it is natural to consider the estimator

$$\widehat{\Theta} \in \arg \min_{\||\Theta\||_{\omega(\infty)} \leq \frac{\alpha^*}{\sqrt{d_r d_c}}} \left\{ \frac{1}{2n} \|y - \mathfrak{X}_n(\Theta)\|_2^2 + \lambda_n \||\Theta\||_{\omega(1)} \right\}, \tag{7}$$

where $\alpha^* \geq 1$ is a measure of spikiness, and the regularization parameter $\lambda_n > 0$ serves to control the nuclear norm of the solution. In the special case when both R and C are identity matrices (of the appropriate dimensions), this estimator is closely related to the standard one considered in past work on the problem, with the only difference between the additional ℓ_∞-norm constraint. In the more general weighted case, an M-estimator of the form (7) using the weighted nuclear norm (but without the elementwise constraint) was recently suggested by Salakhutdinov and Srebro (2010), who provided empirical results to show superiority of the weighted nuclear norm over the standard choice for the Netflix problem.

Past work on matrix completion has focused on the case of exactly low-rank matrices. Here we consider the more general setting of approximately low-rank matrices, including the exact setting as a particular case. We begin by stating a general upper bound that applies to any matrix Θ^*, and involves a natural decomposition into estimation and approximation error terms. The only relevant quantity is the signal-to-noise ratio, as measured by the ratio of the Frobenius norm of Θ^* to the noise variance, so that we allow the noise variance to be free, while assuming that $\||\widetilde{\Delta}\||_{\omega(F)}$ remains bounded.

Theorem 2 *Suppose that $n \geq Ld \log d$, and consider any solution $\widehat{\Theta}$ to the weighted SDP* (7) *using regularization parameter*

$$\lambda_n \geq 2\nu \||\frac{1}{n} \sum_{i=1}^n \xi_i R^{-\frac{1}{2}} X^{(i)} C^{-\frac{1}{2}}\||_{\mathrm{op}}, \tag{8}$$

and define $\lambda_n^* = \max\{\lambda_n, L\sqrt{\frac{d \log d}{n}}\}$. Then with probability greater than $1 - c_2 \exp(-c_2 \log d)$, for each $r = 1, \ldots, d_r$, the error $\widetilde{\Delta} = \widehat{\Theta} - \Theta^*$ satisfies

$$\||\widetilde{\Delta}\||_{\omega(F)}^2 \leq c_1 \, \alpha^* \, \lambda_n^* \left[\sqrt{r} \||\widetilde{\Delta}\||_{\omega(F)} + \sum_{j=r+1}^{d_r} \sigma_j(\sqrt{R}\Theta^*\sqrt{C}) \right] + \frac{c_1 (\alpha^* L)^2}{n}. \tag{9}$$

Apart from the trailing $O(n^{-1})$ the term, the bound (9) shows a natural splitting into two terms. The first can be interpreted as the *estimation error* associated with a rank r matrix, whereas the second term corresponds to *approximation error*, measuring how far $\sqrt{R}\Theta^*\sqrt{C}$ is from a rank r matrix. Of course, the bound holds for any choice of r, and in the corollaries to follow, we choose r optimally so as to balance the estimation and approximation error terms.

In order to provide concrete rates using Theorem 2, it remains to address two issues. First, we need to specify an explicit choice of λ_n by bounding the operator norm of the noise matrix

$\frac{1}{n}\sum_{i=1}^{n}\xi_i\sqrt{R}X^{(i)}\sqrt{C}$, and secondly, we need to understand how to choose the parameter r so as to achieve the tightest possible bound. When Θ^* is exactly low-rank, then it is obvious that we should choose $r = \text{rank}(\Theta^*)$, so that the approximation error vanishes—more specifically, so that $\sum_{j=r+1}^{d_r}\sigma_j(\sqrt{R}\Theta^*\sqrt{C})_j = 0$. Doing so yields the following result:

Corollary 1 (Exactly low-rank matrices) *Suppose that the noise sequence $\{\xi_i\}$ is i.i.d., zero-mean and sub-exponential, and Θ^* has rank at most r, Frobenius norm at most 1, and spikiness at most $\alpha_{sp}(\Theta^*) \leq \alpha^*$. If we solve the SDP (7) with $\lambda_n = 4\nu\sqrt{\frac{d\log d}{n}}$ then there is a numerical constant c_1' such that*

$$|\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_{\omega(F)}^2 \leq c_1'\,(\nu^2 \vee L^2)\,(\alpha^*)^2\,\frac{rd\log d}{n} + \frac{c_1(\alpha^* L)^2}{n} \tag{10}$$

with probability greater than $1 - c_2\exp(-c_3\log d)$.

Note that this rate has a natural interpretation: since a rank r matrix of dimension $d_r \times d_c$ has roughly $r(d_r + d_c)$ free parameters, we require a sample size of this order (up to logarithmic factors) so as to obtain a controlled error bound. An interesting feature of the bound (10) is the term $\nu^2 \vee 1 = \max\{\nu^2, 1\}$, which implies that we do not obtain exact recovery as $\nu \to 0$. As we discuss at more length in Section 3.4, under the mild spikiness condition that we have imposed, this behavior is unavoidable due to lack of identifiability within a certain radius, as specified in the set $\mathfrak{C}$. For instance, consider the matrix Θ^* and the perturbed version $\widetilde{\Theta} = \Theta^* + \frac{1}{\sqrt{d_r d_c}}e_1 e_1^T$. With high probability, we have $\mathfrak{X}_n(\Theta^*) = \mathfrak{X}_n(\widetilde{\Theta})$, so that the observations—even if they were noiseless—fail to distinguish between these two models. These types of examples, leading to non-identifiability, cannot be overcome without imposing fairly restrictive matrix incoherence conditions, as we discuss at more length in Section 3.4.

As with past work (Candès and Plan, 2010; Keshavan et al., 2010b), Corollary 1 applies to the case of matrices that have exactly rank r. In practical settings, it is more realistic to assume that the unknown matrix is not exactly low-rank, but rather can be well approximated by a matrix with low rank. One way in which to formalize this notion is via the ℓ_q-"ball" of matrices

$$\mathbb{B}_q(\rho_q) := \left\{ \Theta \in \mathbb{R}^{d_r \times d_c} \mid \sum_{j=1}^{\min\{d_r,d_c\}} |\sigma_j(\sqrt{R}\Theta\sqrt{C})|^q \leq \rho_q \right\}. \tag{11}$$

For $q = 0$, this set corresponds to the set of matrices with rank at most $r = \rho_0$, whereas for values $q \in (0, 1]$, it consists of matrices whose (weighted) singular values decay at a relatively fast rate. By applying Theorem 2 to this matrix family, we obtain the following corollary:

Corollary 2 (Estimation of near low-rank matrices) *Suppose that the noise $\{\xi_i\}$ is zero-mean and sub-exponential, Consider a matrix $\Theta^* \in \mathbb{B}_q(\rho_q)$ with spikiness at most $\alpha_{sp}(\Theta^*) \leq \alpha^*$, and Frobenius norm at most one. With the same choice of λ_n as Corollary 1, there is a universal constant c_1' such that*

$$|\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_{\omega(F)}^2 \leq c_1\rho_q\left((\nu^2 \vee L^2)(\alpha^*)^2\frac{d\log d}{n}\right)^{1-\frac{q}{2}} + \frac{c_1(\alpha^* L)^2}{n} \tag{12}$$

with probability greater than $1 - c_2\exp(-c_3\log d)$.

Note that this result is a strict generalization of Corollary 1, to which it reduces in the case $q = 0$. (When $q = 0$, we have $\rho_0 = r$ so that the bound has the same form.) Note that the price that we pay for approximately low rank is a smaller exponent—namely, $1 - q/2$ as opposed to 1 in the case $q = 0$. The proof of Corollary 2 is based on a more subtle application of Theorem 2, one which chooses the effective rank r in the bound (9) so as to trade off between the estimation and approximation errors. In particular, the choice $r \asymp \rho_q (\frac{n}{d \log d})^{q/2}$ turns out to yield the optimal trade-off, and hence the given error bound (12).

Although we have stated our results in terms of bounds on the weighted squared Frobenius norm $\|\|\Theta\|\|_{\omega(F)}^2 = \|\|\sqrt{R}\Theta\sqrt{C}\|\|_F^2$, our assumed lower bound on the entries R and C implies that $\|\|\Theta\|\|_{\omega(F)}^2 \geq \frac{\|\|\Theta\|\|_F^2}{L^2}$. Consequently, as long as each row and column is sampled a constant fraction of the time, our results also yield bounds on the Frobenius norm. In some applications, certain rows and columns might be heavily sampled, meaning that some entries of R and/or C could be relatively large. Since we require *only a lower bound* on the row/column sampling frequencies, our Frobenius norm bounds would not degrade if some rows and/or columns were heavily sampled. In contrast, a RIP-type analysis would not be valid in this setting, since heavy sampling means that the Frobenius norm could not be uniformly bounded from above.

In order to illustrate the sharpness of our theory, let us compare the predictions of our two corollaries to the empirical behavior of the *M*-estimator. In particular, we applied the nuclear norm SDP to simulated data, using Gaussian observation noise with variance $v^2 = 0.25$ and the uniform sampling model. In all cases, we solved the nuclear norm SDP using a non-smooth optimization procedure due to Nesterov (2007), via our own implementation in MATLAB. For a given problem size d, we ran $T = 25$ trials and computed the squared Frobenius norm error $\|\|\widehat{\Theta} - \Theta^*\|\|_F^2$ averaged over the trials.

Figure 1 shows the results in the case of exactly low-rank matrices ($q = 0$), with the matrix rank given by $r = \lceil \log^2(d) \rceil$. Panel (a) shows plots of the mean-squared Frobenius error versus the raw sample size, for three different problem sizes with the number of matrix elements sizes $d^2 \in \{40^2, 60^2, 80^2, 100^2\}$. These plots show that the *M*-estimator is consistent, since each of the curves decreases to zero as the sample size n increases. Note that the curves shift to the right as the matrix dimension d increases, reflecting the natural intuition that larger matrices require more samples. Based on the scaling predicted by Corollary 1, we expect that the mean-squared Frobenius error should exhibit the scaling $\|\|\widehat{\Theta} - \Theta^*\|\|_F^2 \asymp \frac{rd \log d}{n}$. Equivalently, if we plot the MSE versus the *rescaled sample size* $N := \frac{n}{rd \log d}$, then all the curves should be relatively well aligned, and decay at the rate $1/N$. Panel (b) of Figure 1 shows the same simulation results re-plotted versus this rescaled sample size. Consistent with the prediction of Corollary 1, all four plots are now relatively well-aligned. Figure 2 shows the same plots for the case of approximately low-rank matrices ($q = 0.5$). Again, consistent with the prediction of Corollary 2, we see qualitatively similar behavior in the plots of the MSE versus sample size (panel (a)), and the rescaled sample size (panel (b)).

3.3 Information-theoretic Lower Bounds

The results of the previous section are achievable results, based on a particular polynomial-time estimator. It is natural to ask how these bounds compare to the fundamental limits of the problem, meaning the best performance achievable by any algorithm. As various authors have noted (Candès and Plan, 2010; Keshavan et al., 2010b), a parameter counting argument indicates that roughly $n \approx r(d_r + d_c)$ samples are required to estimate an $d_r \times d_c$ matrix with rank r. This calculation can
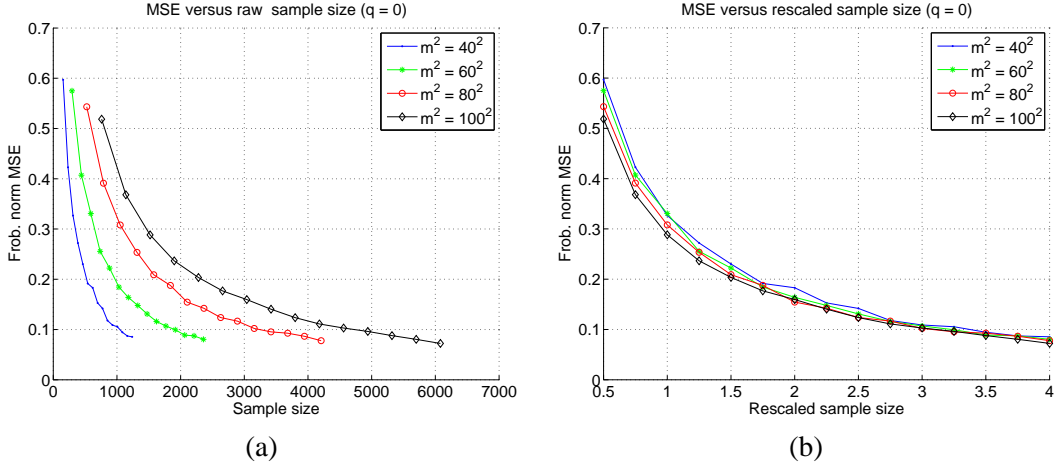
Figure 1: Plots of the mean-squared error in Frobenius norm for $q = 0$. Each curve corresponds to a different problem size $d^2 \in \{40^2, 60^2, 80^2, 100^2\}$. (a) MSE versus the raw sample size n. As expected, the curves shift to the right as d increases, since more samples should be required to achieve a given MSE for larger problems. (b) The same MSE plotted versus the rescaled sample size $n/(rd \log d)$. Consistent with Corollary 1, all the plots are now fairly well-aligned.

be made more formal by metric entropy calculations for the Grassman manifold (e.g., Szarek, 1983); see also Rohde and Tsybakov (2011) for results on approximation numbers for the more general ℓ_q-balls of matrices. Such calculations, while accounting for the low-rank conditions, do *not* address the additional "spikiness" constraints that are essential to the setting of matrix completion. It is conceivable that these additional constraints could lead to a substantial volume reduction in the allowable class of matrices, so that the scalings suggested by parameter counting or metric entropy calculation for Grassman manifolds would be overly conservative.

Accordingly, in this section, we provide a direct and constructive argument to lower bound the minimax rates of Frobenius norm over classes of matrices that are near low-rank and not overly spiky. This argument establishes that the bounds established in Corollaries 1 and 2 are sharp up to logarithmic factors, meaning that no estimator performs substantially better than the one considered here. More precisely, consider the matrix classes

$$\widetilde{\mathbb{B}}(\rho_q) = \left\{ \Theta \in \mathbb{R}^{d \times d} \mid \sum_{j=1}^{d} \sigma_j(\Theta)^q \leq \rho_q, \ \alpha_{\mathrm{sp}}(\Theta) \leq \sqrt{32 \log d} \right\},$$

corresponding to square $d \times d$ matrices that are near low-rank (belonging to the ℓ_q-balls previously defined (11)), and have a logarithmic spikiness ratio. The following result applies to the *minimax risk* in Frobenius norm, namely the quantity

$$\mathfrak{M}_n(\widetilde{\mathbb{B}}(\rho_q)) := \inf_{\widetilde{\Theta}} \sup_{\Theta^* \in \widetilde{\mathbb{B}}(\rho_q)} \mathbb{E}\left[\|\|\widetilde{\Theta} - \Theta^*\|\|_F^2\right],$$

where the infimum is taken over all estimators $\widetilde{\Theta}$ that are measurable functions of n samples.
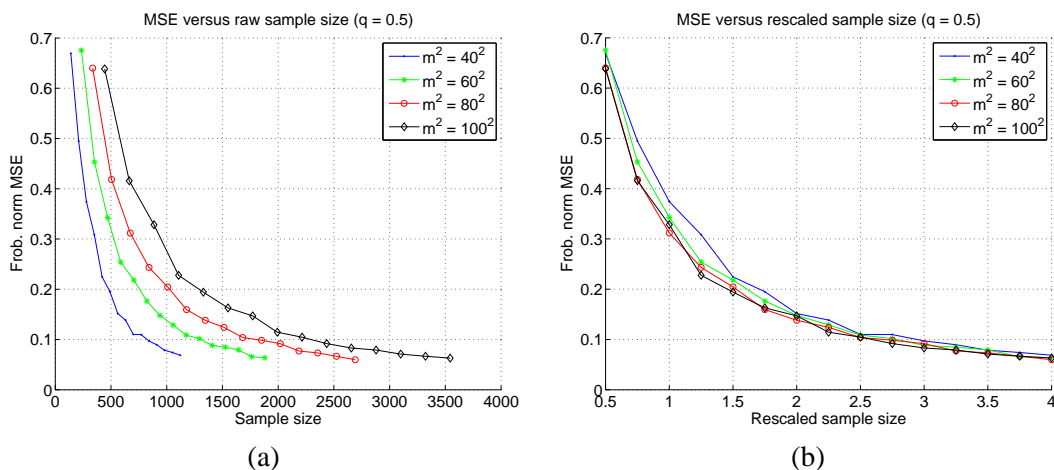
Figure 2: Plots of the mean-squared error in Frobenius norm for $q = 0.5$. Each curve corresponds to a different problem size $d^2 \in \{40^2, 60^2, 80^2, 100^2\}$. (a) MSE versus the raw sample size n. As expected, the curves shift to the right as d increases, since more samples should be required to achieve a given MSE for larger problems. (b) The same MSE plotted versus the rescaled sample size $n/(\rho_q^{\frac{1}{1-q/2}} d \log d)$. Consistent with Corollary 2, all the plots are now fairly well-aligned.

Theorem 3 *There is a universal numerical constant $c_5 > 0$ such that*

$$\mathfrak{M}_n(\widetilde{\mathbb{B}}(\rho_q)) \geq c_5 \, \min\left\{\rho_q \left(\frac{\nu^2 d}{n}\right)^{1-\frac{q}{2}}, \, \frac{\nu^2 d^2}{n}\right\}.$$

The term of primary interest in this bound is the first one—namely, $\rho_q \left(\frac{\nu^2 d}{n}\right)^{1-\frac{q}{2}}$. It is the dominant term in the bound whenever the ℓ_q-radius satisfies the bound

$$\rho_q \leq \left(\frac{\nu^2 d}{n}\right)^{\frac{q}{2}} d. \tag{13}$$

In the special case $q = 0$, corresponding the exactly low-rank case, the bound (13) always holds, since it reduces to requiring that the rank $r = \rho_0$ is less than or equal to d. In these regimes, Theorem 3 establishes that the upper bounds obtained in Corollaries 1 and 2 are minimax-optimal up to factors logarithmic in matrix dimension d.

3.4 Comparison to Other Work

We now turn to a detailed comparison of our bounds to those obtained in past work on noisy matrix completion, in particular the papers by Candès and Plan (2010) (hereafter CP) and Keshavan et al. (2010b) (hereafter KMO). Both papers considered only the case of exactly low-rank matrices, corresponding to the special case of $q = 0$ in our notation. Since neither paper provided results for the general case of near-low rank matrices, nor the general result (with estimation and approximation

errors) stated in Theorem 2, our discussion is mainly limited to comparing Corollary 1 to their results. So as to simplify discussion, we restate all results under the scalings used in this paper[1] (i.e., with $|\!|\!|\Theta^*|\!|\!|_F = 1$).

3.4.1 COMPARISON OF RATES

Under the strong incoherence conditions required for exact matrix recovery (see below for discussion), Theorem 7 in CP give an bound on $|\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F$ that depends on the Frobenius norm of the potentially adversarial error matrix $\Xi \in \mathbb{R}^{d_1 \times d_2}$, as defined by the noise variables $[\Xi]_{j(i)\,k(i)} = \widetilde{\xi}_i$ in our case. In the special case of stochastic noise, under the observation model (1) and the scalings of our paper, as long as $n > d$, where $d = d_1 + d_2$—a condition certainly required for Frobenius norm consistency—we have $|\!|\!|\Xi|\!|\!|_F = \Theta(\nu\sqrt{n}/d)$ with high probability. Given this scaling, the CP upper bound takes the form

$$|\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F \lesssim \nu \left\{ \sqrt{d} + \frac{\sqrt{n}}{d} \right\}.$$

Note that if the noise standard deviation ν tends to zero while the sample size n, matrix size p and rank r all remain fixed, then this bound guarantees that the Frobenius error tends to zero. This behavior as $\nu \to 0$ is intuitively reasonable, given that their proof technique is an extrapolation from the case of exact recovery for noiseless observations ($\nu = 0$). However, note that for any fixed noise deviation $\nu > 0$, the first term increases to infinity as the matrix dimension d increases, whereas the second term actually grows as the sample size n increases. Consequently, the CP results do not guarantee statistical consistency, unlike the bounds proved here.

Turning to a setting with adversarial noise, suppose that the error vector has Frobenius norm at most δ. A modification of our analysis yields error bounds of the form $|\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F \lesssim \left\{ \frac{d^2}{\sqrt{n}}\delta + \sqrt{\frac{rd\log d}{n}} \right\}$. In the setting of square matrices with $\delta \geq \sqrt{\frac{r\log d}{d}}$, our result yields an upper bound tighter by a factor of order $\sqrt{d}$ better than those presented in CP. Last, as pointed out by a reviewer, the CP analysis does yield bounds for approximately low-rank matrices, in particular by writing $\Theta^* = \Pi_r(\Theta^*) + \Delta$, where Π_r is the Frobenius norm projection onto the space of rank r matrices, and $\Delta = \Theta^* - \Pi_r(\Theta^*)$ is the approximation error. With this notation, their analysis guarantees error bounds of the form $\sqrt{d}|\!|\!|\Delta|\!|\!|_F$ with high probability, which is a weaker guarantee than our bound whenever $|\!|\!|\Delta|\!|\!|_F \geq c\sqrt{\frac{r\log d}{n}}$ and $n = \Omega(d\log d)$.

Keshavan et al. (2010b) analyzed alternative methods based on trimming and applying the SVD. For Gaussian noise, their methods guarantee bounds (with high probability) of the form

$$|\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F \lesssim \nu \min\left\{ \alpha\sqrt{\frac{d_2}{d_1}}, \kappa^2(\Theta^*) \right\} \sqrt{\frac{rd_2}{n}}, \tag{14}$$

where d_2/d_1 is the aspect ratio of Θ^*, and $\kappa(\Theta^*) = \frac{\sigma_{\max}(\Theta^*)}{\sigma_{\min}(\Theta^*)}$ is the condition number of Θ^*. This result is more directly comparable to our Corollary 1; apart from the additional factor involving either the aspect ratio or the condition number, it is sharper since it does not involve the factor $\log d$ present in our bound. For a fixed noise standard deviation ν, the bound (14) guarantees statistical

---

1. The paper CP and KMO use two different sets of scaling, one with $|\!|\!|\Theta^*|\!|\!|_F = \Theta(d)$ and the other with $|\!|\!|\Theta^*|\!|\!|_F = \sqrt{r}$, so that some care is required in converting between results.

consistency as long as $\frac{rd_2}{n}$ tends to zero. The most significant differences are the presence of the aspect ratio d_2/d_1 or the condition number $\kappa(\Theta^*)$ in the upper bound (14). The aspect ratio is a quantity that can be as small as one, or as large as d_2, so that the pre-factor in the bound (14) can scale in a dimension-dependent way. Similarly, for any matrix with rank larger than one, the condition number can be made arbitrarily large. For instance, in the rank two case, define a matrix with $\sigma_{\max}(\Theta^*) = \sqrt{1-\delta^2}$ and $\sigma_{\min}(\Theta^*) = \delta$, and consider the behavior as $\delta \to 0$. In contrast, our bounds are invariant to both the aspect ratio and the condition number of Θ^*.

3.4.2 COMPARISON OF MATRIX CONDITIONS

We now turn to a comparison of the various *matrix incoherence assumptions* invoked in the analysis of CP and KMO, and comparison to our spikiness condition. As before, for clarity, we specialize our discussion to the square case ($d_r = d_c = d$), since the rectangular case is not essentially different. The matrix incoherence conditions are stated in terms of the singular value decomposition $\Theta^* = U\Sigma V^T$ of the target matrix. Here $U \in \mathbb{R}^{d \times r}$ and $V \in \mathbb{R}^{d \times r}$ are matrices of the left and right singular vectors respectively, satisfying $U^T U = V^T V = I_{r \times r}$, whereas $\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix of the singular values. The purpose of matrix incoherence is to enforce that the left and right singular vectors should not be aligned with the standard basis. Among other assumptions, the CP analysis imposes the incoherence conditions

$$\|UU^T - \frac{r}{d}I_{d \times d}\|_\infty \le \mu \frac{\sqrt{r}}{d}, \qquad \|VV^T - \frac{r}{d}I_{d \times d}\|_\infty \le \mu \frac{\sqrt{r}}{d}, \quad \text{and} \quad \|UV^T\|_\infty \le \mu \frac{\sqrt{r}}{d}, \qquad (15)$$

for some constant $\mu > 0$. Parts of the KMO analysis impose the related incoherence condition

$$\max_{j=1,\ldots,d} |UU^T|_{jj} \le \mu_0 \frac{r}{d}, \quad \text{and} \max_{j=1,\ldots,d} |VV^T|_{jj} \le \mu_0 \frac{r}{d}. \qquad (16)$$

Both of these conditions ensure that the singular vectors are sufficiently "spread-out", so as not to be aligned with the standard basis.

A remarkable property of conditions (15) and (16) is that they exhibit *no dependence* on the singular values of Θ^*. If one is interested only in exact recovery in the noiseless setting, then this lack of dependence is reasonable. However, if approximate recovery is the goal—as is necessarily the case in the more realistic setting of noisy observations—then it is clear that a minimal set of sufficient conditions should also involve the singular values, as is the case for our spikiness measure $\alpha_{sp}(\Theta^*)$. The following example gives a concrete demonstration of an instance where our conditions are satisfied, so that approximate recovery is possible, whereas the incoherence conditions are violated.

Example. Let $\Gamma \in \mathbb{R}^{d \times d}$ be a positive semidefinite symmetric matrix with rank $r-1$, Frobenius norm $\|\|\Gamma\|\|_F = 1$ and $\|\Gamma\|_\infty \le c_0/d$. For a scalar parameter $t > 0$, consider the matrix

$$\Theta^* := \Gamma + te_1e_1^T$$

where $e_1 \in \mathbb{R}^d$ is the canonical basis vector with one in its first entry, and zero elsewhere. By construction, the matrix Θ^* has rank at most r. Moreover, as long as $t = O(1/d)$, we are guaranteed that our spikiness measure satisfies the bound $\alpha_{sp}(\Theta^*) = O(1)$. Indeed, we have $\|\|\Theta^*\|\|_F \ge \|\|\Gamma\|\|_F - t = 1 - t$, and hence

$$\alpha_{sp}(\Theta^*) = \frac{d\|\Theta^*\|_\infty}{\|\|\Theta^*\|\|_F} \le \frac{d(\|\Gamma\|_\infty + t)}{1 - t} \le \frac{c_0 + dt}{1 - t} = O(1).$$

Consequently, for any choice of Γ as specified above, Corollary 1 implies that the SDP will recover the matrix Θ^* up to a tolerance $O(\sqrt{\frac{rd\log d}{n}})$. This captures the natural intuition that "poisoning" the matrix Γ with the term $te_1^T e_1$ should have essentially no effect, as long as t is not too large.

On the other hand, suppose that we choose the matrix Γ such that its $r-1$ eigenvectors are orthogonal to e_1. In this case, we have $\Theta^* e_1 = te_1$, so that e_1 is also an eigenvector of Θ^*. Letting $U \in \mathbb{R}^{d \times r}$ be the matrix of eigenvectors, we have $e_1^T U U^T e_1 = 1$. Consequently, for any fixed μ (or μ_0) and rank $r \ll d$, conditions (15) and (16) are violated.

$$\diamondsuit$$

4. Proofs for Noisy Matrix Completion

We now turn to the proofs of our results. This section is devoted to the results that apply directly to noisy matrix completion, in particular the achievable result given in Theorem 2, its associated Corollaries 1 and 2, and the information-theoretic lower bound given in Theorem 3. The proof of Theorem 1 is provided in Section 5 to follow.

4.1 A Useful Transformation

We begin by describing a transformation that is useful both in these proofs, and the later proof of Theorem 1. In particular, we consider the mapping $\Theta \mapsto \Gamma := \sqrt{R}\Theta\sqrt{C}$, as well as the modified observation operator $\mathfrak{X}_n' : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ with elements

$$[\mathfrak{X}_n'(\Gamma)]_i = \langle\!\langle \widetilde{X}^{(i)}, \Gamma \rangle\!\rangle, \quad \text{for } i = 1, 2, \ldots, n,$$

where $\widetilde{X}^{(i)} := R^{-1/2} X^{(i)} C^{-1/2}$. Note that $\mathfrak{X}_n'(\Gamma) = \mathfrak{X}_n(\Theta)$ by construction, and moreover

$$\|\!|\Gamma|\!\|_F = \|\!|\Theta|\!\|_{\omega(F)}, \quad \|\!|\Gamma|\!\|_1 = \|\!|\Theta|\!\|_{\omega(1)}, \quad \text{and} \quad \|\!|\Gamma|\!\|_\infty = \|\!|\Theta|\!\|_{\omega(\infty)},$$

which implies that

$$\beta_{\mathrm{ra}}(\Theta) = \underbrace{\frac{\|\!|\Gamma|\!\|_1}{\|\!|\Gamma|\!\|_F}}_{\beta_{\mathrm{ra}}'(\Gamma)}, \quad \text{and} \quad \alpha_{\mathrm{sp}}(\Theta) = \underbrace{\frac{d\,\|\!|\Gamma|\!\|_\infty}{\|\!|\Gamma|\!\|_F}}_{\alpha_{\mathrm{sp}}'(\Gamma)}.$$

Based on this change of variables, let us define a modified version of the constraint set (4) as follows

$$\mathfrak{C}'(n;c_0) = \left\{ 0 \neq \Gamma \in \mathbb{R}^{d \times d} \mid \alpha_{\mathrm{sp}}'(\Gamma)\,\beta_{\mathrm{ra}}'(\Gamma) \leq \frac{1}{c_0}\sqrt{\frac{n}{d\log d}} \right\}. \tag{17}$$

In this new notation, the lower bound (5) from Theorem 1 can be re-stated as

$$\frac{\|\mathfrak{X}_n'(\Gamma)\|_2}{\sqrt{n}} \geq \frac{1}{8}\|\!|\Gamma|\!\|_F\left\{1 - \frac{128L\alpha_{\mathrm{sp}}'(\Gamma)}{\sqrt{n}}\right\} \quad \text{for all } \Gamma \in \mathfrak{C}'(n;c_0). \tag{18}$$

4.2 Proof of Theorem 2

We now turn to the proof of Theorem 2. Defining the estimate $\widehat{\Gamma} := \sqrt{R}\widehat{\Theta}\sqrt{C}$, we have

$$\widehat{\Gamma} \in \arg\min_{\|\Gamma\|_\infty \leq \frac{\alpha^*}{\sqrt{d_r d_c}}} \left\{ \frac{1}{2n}\|y - \mathfrak{X}_n'(\Gamma)\|_2^2 + \lambda_n\|\!|\Gamma|\!\|_1 \right\}, \tag{19}$$

and our goal is to upper bound the ordinary Frobenius norm $|\!|\!|\widehat{\Gamma} - \Gamma^*|\!|\!|_F$.

We now state a useful technical result. Parts (a) and (b) of the following lemma were proven by Recht et al. (2010), and Negahban and Wainwright (2011), respectively. We recall that we adopt the shorthand $\widehat{\Delta} = \widehat{\Gamma} - \Gamma^*$ throughout the analysis.

Lemma 1 *Let $(\widetilde{U}, \widetilde{V})$ represent a pair of r-dimensional subspaces of left and right singular vectors of Γ^*. Then there exists a matrix decomposition $\widehat{\Delta} = \widehat{\Delta}' + \widehat{\Delta}''$ of the error $\widehat{\Delta}$ such that*

(a) The matrix $\widehat{\Delta}'$ satisfies the constraint $\mathrm{rank}(\widehat{\Delta}') \leq 2r$, and

(b) Given the choice (8), the nuclear norm of $\widehat{\Delta}''$ is bounded as

$$|\!|\!|\widehat{\Delta}''|\!|\!|_1 \;\leq\; 3|\!|\!|\widehat{\Delta}'|\!|\!|_1 + 4 \sum_{j=r+1}^{d_r} \sigma_j(\Gamma^*). \tag{20}$$

Note that the bound (20), combined with triangle inequality, implies that

$$|\!|\!|\widehat{\Delta}|\!|\!|_1 \leq |\!|\!|\widehat{\Delta}'|\!|\!|_1 + |\!|\!|\widehat{\Delta}''|\!|\!|_1 \;\leq\; 4|\!|\!|\widehat{\Delta}'|\!|\!|_1 + 4 \sum_{j=r+1}^{d_r} \sigma_j(\Gamma^*)$$

$$\leq 8\sqrt{r}|\!|\!|\widehat{\Delta}|\!|\!|_F + 4 \sum_{j=r+1}^{d_r} \sigma_j(\Gamma^*) \tag{21}$$

where the second inequality uses the fact that $\mathrm{rank}(\widehat{\Delta}') \leq 2r$.

We now split into two cases, depending on whether or not the error $\widehat{\Delta}$ belongs to the set $\mathfrak{C}'(n; c_0)$.

4.2.1 CASE 1

First suppose that $\widehat{\Delta} \notin \mathfrak{C}'(n; c_0)$. In this case, by the definition (17), we have

$$|\!|\!|\widehat{\Delta}|\!|\!|_F^2 \leq c_0 L \left(\sqrt{d_r d_c} \|\widehat{\Delta}\|_\infty \right) |\!|\!|\widehat{\Delta}|\!|\!|_1 \sqrt{\frac{d \log d}{n}}$$

$$\leq 2c_0 L \alpha^* |\!|\!|\widehat{\Delta}|\!|\!|_1 \sqrt{\frac{d \log d}{n}},$$

since $\|\widehat{\Delta}\|_\infty \leq \|\Gamma^*\|_\infty + \|\widehat{\Gamma}\|_\infty \leq \frac{2\alpha^*}{\sqrt{d_r d_c}}$. Now applying the bound (21), we obtain

$$|\!|\!|\widehat{\Delta}|\!|\!|_F^2 \leq 2c_0 L \alpha^* \sqrt{\frac{d \log d}{n}} \Big\{ 8\sqrt{r}|\!|\!|\widehat{\Delta}|\!|\!|_F + 4 \sum_{j=r+1}^{d_r} \sigma_j(\Gamma^*) \Big\}. \tag{22}$$

4.2.2 CASE 2

Otherwise, we must have $\widehat{\Delta} \in \mathfrak{C}'(n; c_0)$. Recall the reformulated lower bound (18). On one hand, if $\frac{128 L \alpha'_{\mathrm{sp}}(\widehat{\Delta})}{\sqrt{n}} > 1/2$, then we have

$$|\!|\!|\widehat{\Delta}|\!|\!|_F \leq \frac{256 L \sqrt{d_r d_c} \|\widehat{\Delta}\|_\infty}{\sqrt{n}} \;\leq\; \frac{512 L \alpha^*}{\sqrt{n}}. \tag{23}$$

On the other hand, if $\frac{128 L \alpha'_{\mathrm{sp}}(\widehat{\Delta})}{\sqrt{n}} \leq 1/2$, then from the bound (18), we have

$$\frac{\|\mathfrak{X}_n{}'(\widehat{\Delta})\|_2}{\sqrt{n}} \geq \frac{\|\|\widehat{\Delta}\|\|_F}{16} \tag{24}$$

with high probability. Note that $\widehat{\Gamma}$ is optimal and Γ^* is feasible for the convex program (19), so that we have the basic inequality

$$\frac{1}{2n}\|y - \mathfrak{X}_n{}'(\widehat{\Gamma})\|_2^2 + \lambda_n \|\|\widehat{\Gamma}\|\|_1 \leq \frac{1}{2n}\|y - \mathfrak{X}_n{}'(\Gamma^*)\|_2^2 + \lambda_n \|\|\Gamma^*\|\|_1.$$

Some algebra then yields the inequality

$$\frac{1}{2n}\|\mathfrak{X}_n{}'(\widehat{\Delta})\|_2^2 \leq \nu \langle\!\langle \widehat{\Delta}, \frac{1}{n}\sum_{i=1}^{n}\xi_i \widetilde{X}^{(i)} \rangle\!\rangle + \lambda_n \|\|\Gamma^*\|\|_1 - \lambda_n \|\|\Gamma^* + \widehat{\Delta}\|\|,$$

Substituting the lower bound (24) into this inequality yields

$$\frac{\|\|\widehat{\Delta}\|\|_F^2}{512} \leq \nu \langle\!\langle \widehat{\Delta}, \frac{1}{n}\sum_{i=1}^{n}\xi_i \widetilde{X}^{(i)} \rangle\!\rangle + \lambda_n \|\|\Gamma^*\|\|_1 - \lambda_n \|\|\Gamma^* + \widehat{\Delta}\|\|.$$

From this point onwards, the proof is identical (apart from constants) to Theorem 1 in Negahban and Wainwright (2011), and we obtain that there is a numerical constant c_1 such that

$$\|\|\widehat{\Delta}\|\|_F^2 \leq c_1 \alpha^* \lambda_n \left\{ \sqrt{r}\|\|\widehat{\Delta}\|\|_F + \sum_{j=r+1}^{d_r} \sigma_j(\Gamma^*) \right\}. \tag{25}$$

4.2.3 PUTTING TOGETHER THE PIECES

Summarizing our results, we have shown that with high probability, one of the three bounds (22), (23) or (25) must hold. Since $\alpha^* \geq 1$, these claims can be summarized in the form

$$\|\|\widehat{\Delta}\|\|_F^2 \leq c_1 \, \max\left\{\lambda_n, \sqrt{\frac{d \log d}{n}}\right\} \left[\sqrt{r}\|\|\widehat{\Delta}\|\|_F + \sum_{j=r+1}^{d_r} \sigma_j(\Gamma^*)\right].$$

for a universal positive constant c_1. Translating this result back to the original co-ordinate system ($\Gamma^* = \sqrt{R}\Theta^*\sqrt{C}$) yields the claim (9).

4.3 Proof of Corollary 1

When Θ^* (and hence $\sqrt{R}\Theta^*\sqrt{C}$) has rank $r < d_r$, then we have $\sum_{j=r+1}^{d_r}\sigma_j(\sqrt{R}\Theta^*\sqrt{C}) = 0$. Consequently, the bound (9) reduces to $\|\|\widetilde{\Delta}\|\|_{\omega(F)} \leq c_1 \alpha^* \lambda_n^* \sqrt{r}$. To complete the proof, it suffices to show that

$$\mathbb{P}\big[\|\|\frac{1}{n}\sum_{i=1}^{n}\xi_i R^{-1/2}X^{(i)}C^{-1/2}\|\|_2 \geq c_1 \nu \sqrt{\frac{d \log d}{n}}\big] \leq c_2 \exp(-c_2 d \log d).$$

We do so via the Ahlswede-Winter matrix bound, as stated in Appendix F. Defining the random matrix $Y^{(i)} := \xi_i R^{-1/2} X^{(i)} C^{-1/2}$, we first note that ξ_i is sub-exponential with parameter 1, and $|R^{-1/2} X^{(i)} C^{-1/2}|$ has a single entry with magnitude at most $L\sqrt{d_r d_c}$, which implies that

$$\|Y^{(i)}\|_{\psi_1} \le L\nu \sqrt{d_r d_c} \le 2\nu L d.$$

(Here $\|\cdot\|_{\psi_1}$ denotes the Orlicz norm (Ledoux and Talagrand, 1991) of a random variable, as defined by the function $\psi_1(x) = \exp(x) - 1$; see Appendix F). Moreover, we have

$$\mathbb{E}[(Y^{(i)})^T Y^{(i)}] = \nu^2 \mathbb{E}\Big[\frac{d_r d_c}{R_{j(i)} C_{k(i)}} e_{k(i)} e_{j(i)}^T e_{j(i)} e_{k(i)}^T\Big]$$

$$= \nu^2 \mathbb{E}\Big[\frac{d_r d_c}{R_{j(i)} C_{k(i)}} e_{k(i)} e_{k(i)}^T\Big]$$

$$= \nu^2 d_r I_{d_c \times d_c}.$$

so that $\||\mathbb{E}[(Y^{(i)})^T Y^{(i)}]\||_2 \le 2\nu^2 d$, recalling that $2d = d_r + d_c \ge d_r$. The same bound applies to $\||\mathbb{E}[Y^{(i)}(Y^{(i)})^T]\||_2$, so that applying Lemma 7 with $t = n\delta$, we conclude that

$$\mathbb{P}\Big[\||\frac{1}{n}\sum_{i=1}^n \xi_i R^{-1/2} X^{(i)} C^{-1/2}\||_2 \ge \delta\Big] \le (d_r \times d_c) \max\big\{\exp(-n\delta^2/(16\nu^2 d)), \exp(-\frac{n\delta}{4\nu L d})\big\}.$$

Since $\sqrt{d_r d_c} \le d_r + d_c = 2d$, if we set $\delta^2 = c_1^2 \nu^2 \frac{d\log d}{n}$ for a sufficiently large constant c_1, the result follows. (Here we also use the assumption that $n = \Omega(Ld\log d)$, so that the term $\sqrt{\frac{d\log d}{n}}$ is dominant.)

4.4 Proof of Corollary 2

For this corollary, we need to determine an appropriate choice of r so as to optimize the bound (9). To ease notation, let us make use of the shorthand notation $\Gamma^* = \sqrt{R}\Theta^*\sqrt{C}$. With the singular values of Γ^* ordered in non-increasing order, fix some threshold $\tau > 0$ to be determined, and set $r = \max\{j \mid \sigma_j(\Gamma^*) > \tau\}$. This choice ensures that

$$\sum_{j=r+1}^{d_r} \sigma_j(\Gamma^*) = \tau \sum_{j=r+1}^{d_r} \frac{\sigma_j(\Gamma^*)}{\tau} \le \tau \sum_{j=r+1}^{d_r} \Big(\frac{\sigma_j(\Gamma^*)}{\tau}\Big)^q \le \tau^{1-q}\rho_q.$$

Moreover, we have $r\tau^q \le \sum_{j=1}^r \{\sigma_j(\Gamma^*)\}^q \le \rho_q$, which implies that $\sqrt{r} \le \sqrt{\rho_q}\tau^{-q/2}$. Substituting these relations into the upper bound (9) leads to

$$\||\widetilde{\Delta}\||_{\omega(F)}^2 \le c_1\,\alpha^*\,\lambda_n^*\,\big[\sqrt{\rho_q}\tau^{-q/2}\||\widetilde{\Delta}\||_{\omega(F)} + \tau^{1-q}\rho_q\big\}.$$

In order to obtain the sharpest possible upper bound, we set $\tau = \alpha^*\lambda_n^*$. Following some algebra, we find that there is a universal constant c_1 such that

$$\||\widetilde{\Delta}\||_{\omega(F)}^2 \le c_1\rho_q\big((\alpha^*)^2(\lambda_n^*)^2\big)^{1-\frac{q}{2}}.$$

As in the proof of Corollary 1, it suffices to choose $\lambda_n = \Omega(\nu\sqrt{\frac{d\log d}{n}})$, so that $\lambda_n^* = O\sqrt{(\nu^2 + L)\frac{d\log d}{n}}$, from which the claim follows.

4.5 Proof of Theorem 3

Our proof of this lower bound exploits a combination of information-theoretic methods (Yu, 1997; Yang and Barron, 1999), which allow us to reduce to a multiway hypothesis test, and an application of the probabilistic method so as to construct a suitably large packing set. By Markov's inequality, it suffices to prove that

$$\sup_{\Theta^* \in \widetilde{\mathbb{B}}(\rho_q)} \mathbb{P}\left[\||\widehat{\Theta} - \Theta^*\||_F^2 \geq \frac{\delta^2}{4} \right] \geq \frac{1}{2}.$$

In order to do so, we proceed in a standard way—namely, by reducing the estimation problem to a testing problem over a suitably constructed packing set contained within $\widetilde{\mathbb{B}}(\rho_q)$. In particular, consider a set $\{\Theta^1, \ldots, \Theta^{M(\delta)}\}$ of matrices, contained within $\widetilde{\mathbb{B}}(\rho_q)$, such that $\||\Theta^k - \Theta^\ell\||_F \geq \delta$ for all $\ell \neq k$. To ease notation, we use M as shorthand for $M(\delta)$ through much of the argument. Suppose that we choose an index $V \in \{1, 2, \ldots, M\}$ uniformly at random (u.a.r.), and we are given observations $y \in \mathbb{R}^n$ from the observation model (2) with $\Theta^* = \Theta^V$. Then triangle inequality yields the lower bound

$$\sup_{\Theta^* \in \widetilde{\mathbb{B}}(\rho_q)} \mathbb{P}\left[\||\widehat{\Theta} - \Theta^*\||_F \geq \frac{\delta}{2} \right] \geq \mathbb{P}[\widehat{V} \neq V].$$

If we condition on $\mathfrak{X}_n$, a variant of Fano's inequality yields

$$\mathbb{P}[\widehat{V} \neq V \mid \mathfrak{X}_n] \geq 1 - \frac{\left(\binom{M}{2}\right)^{-1} \sum_{\ell \neq k} D(\Theta^k \| \Theta^\ell) + \log 2}{\log M}, \tag{26}$$

where $D(\Theta^k \| \Theta^\ell)$ denotes the Kullback-Leibler divergence between the distributions of $(y | \mathfrak{X}_n, \Theta^k)$ and $(y | \mathfrak{X}_n, \Theta^\ell)$. In particular, for additive Gaussian noise with variance ν^2, we have

$$D(\Theta^k \| \Theta^\ell) = \frac{1}{2\nu^2} \| \mathfrak{X}_n(\Theta^k) - \mathfrak{X}_n(\Theta^\ell) \|_2^2,$$

and moreover,

$$\mathbb{E}_{\mathfrak{X}_n}\left[D(\Theta^k \| \Theta^\ell) \right] = \frac{1}{2\nu^2} \||\Theta^k - \Theta^\ell\||_F^2.$$

Combined with the bound (26), we obtain the bound

$$\begin{aligned}
\mathbb{P}[\widehat{V} \neq V] &= \mathbb{E}_{\mathfrak{X}_n}\left\{ \mathbb{P}[\widehat{V} \neq V \mid \mathfrak{X}_n] \right\} \\
&\geq 1 - \frac{\left(\binom{M}{2}\right)^{-1} \sum_{\ell \neq k} \frac{n}{2\nu^2} \||\Theta^k - \Theta^\ell\||_F^2 + \log 2}{\log M},
\end{aligned} \tag{27}$$

The remainder of the proof hinges on the following technical lemma, which we prove in Appendix A.

Lemma 2 *Let $d \geq 10$ be a positive integer, and let $\delta > 0$. Then for each $r = 1, 2, \ldots, d$, there exists a set of d-dimensional matrices $\{\Theta^1, \ldots, \Theta^M\}$ with cardinality $M = \lfloor \frac{1}{4} \exp\left(\frac{rd}{128}\right) \rfloor$ such that each*

matrix has rank r, and moreover

$$\|\!|\Theta^\ell|\!\|_F = \delta \qquad \text{for all } \ell = 1,2,\ldots,M,$$
$$\|\!|\Theta^\ell - \Theta^k|\!\|_F \geq \delta \qquad \text{for all } \ell \neq k,$$
$$\alpha_{\text{sp}}(\Theta^\ell) \leq \sqrt{32\log d} \quad \text{for all } \ell = 1,2,\ldots,M, \text{ and}$$
$$\|\!|\Theta^\ell|\!\|_2 \leq \frac{4\delta}{\sqrt{r}} \quad \text{for all } \ell = 1,2,\ldots,M.$$

We now show how to use this packing set in our Fano bound. To avoid technical complications, we assume throughout that $rd > 1024\log 2$. Note that packing set from Lemma 2 satisfies $\|\!|\Theta^k - \Theta^\ell|\!\|_F \leq 2\delta$ for all $k \neq \ell$, and hence from Fano bound (27), we obtain

$$\mathbb{P}[\widehat{V} \neq V] \geq 1 - \frac{2\frac{n\delta^2}{v^2} + \log 2}{\frac{rd}{128} - \log 4}$$
$$\geq 1 - \frac{2\frac{n\delta^2}{v^2} + \log 2}{\frac{rd}{256}}$$
$$\geq 1 - \frac{512\frac{n\delta^2}{v^2} + 256\log 2}{rd}.$$

If we now choose $\delta^2 = \frac{v^2}{2048}\frac{rd}{n}$, then

$$\mathbb{P}[\widehat{V} \neq V] \geq 1 - \frac{\frac{rd}{4} + 256\log 2}{rd} \geq \frac{1}{2},$$

where the final inequality again uses the bound $rd \geq 1024\log 2$.

In the special case $q = 0$, the proof is complete, since the elements Θ^ℓ all have rank $r = R_0$, and satisfy the bound $\alpha_{\text{sp}}(\Theta^\ell) \leq \sqrt{32\log d}$. For $q \in (0,1]$, consider the matrix class $\widetilde{\mathbb{B}}(\rho_q)$, and let us set $r = \min\{d, \lceil \rho_q(\frac{d}{n})^{-\frac{q}{2}}\rceil\}$ in Lemma 2. With this choice, since each matrix Θ^ℓ has rank r, we have

$$\sum_{j=1}^p \sigma_i(\Theta^\ell)^q \leq r\left(\frac{\delta}{\sqrt{r}}\right)^q = r\left(\frac{1}{2048}\sqrt{\frac{d}{n}}\right)^q \leq \rho_q,$$

so that we are guaranteed that $\Theta^\ell \in \widetilde{\mathbb{B}}(\rho_q)$. Finally, we note that

$$\frac{rd}{n} \geq \min\left\{\rho_q\left(\frac{d}{n}\right)^{1-\frac{q}{2}}, \frac{d^2}{n}\right\},$$

so that we conclude that the minimax error is lower bounded by

$$\frac{1}{4096}\min\left\{\rho_q\left(\frac{v^2d}{n}\right)^{1-\frac{q}{2}}, \frac{v^2d^2}{n}\right\}$$

for dr sufficiently large. (At the expense of a worse pre-factor, the same bound holds for all $d \geq 10$.)

5. Proof of Theorem 1

We now turn to the proof that the sampling operator in weighted matrix completion satisfies restricted strong convexity over the set $\mathfrak{C}$, as stated in Theorem 1. In order to lighten notation, we prove the theorem in the case $d_r = d_c$. In terms of rates, this is a worst-case assumption, effectively amounting to replacing both d_r and d_c by the worst-case $\max\{d_r, d_c\}$. However, since our rates are driven by $d = \frac{1}{2}(d_r + d_c)$ and we have the inequalities

$$\frac{1}{2}\max\{d_r, d_c\} \leq \frac{1}{2}(d_r + d_c) \leq \max\{d_r, d_c\},$$

this change has only an effect on the constant factors. The proof can be extended to the general setting $d_r \neq d_c$ by appropriate modifications if these constant factors are of interest.

5.1 Reduction to Simpler Events

In order to prove Theorem 1, it is equivalent to show that, with high probability, we have

$$\frac{\|\mathfrak{X}_n{}'(\Gamma)\|_2}{\sqrt{n}} \geq \frac{1}{8}\|\|\Gamma\|\|_F - \frac{48L\, d\, \|\Gamma\|_\infty}{\sqrt{n}} \qquad \text{for all } \Gamma \in \mathfrak{C}'(n; c_0). \tag{28}$$

The remainder of the proof is devoted to studying the "bad" event

$$\mathcal{E}(\mathfrak{X}_n{}') := \left\{ \exists \quad \Gamma \in \mathfrak{C}'(n; c_0) \mid \left| \frac{\|\mathfrak{X}_n{}'(\Gamma)\|_2}{\sqrt{n}} - \|\|\Gamma\|\|_F \right| > \frac{7}{8}\|\|\Gamma\|\|_F + \frac{48L\, d\, \|\Gamma\|_\infty}{\sqrt{n}} \right\}.$$

Suppose that $\mathcal{E}(\mathfrak{X}_n{}')$ does *not* hold: then we have

$$\left| \frac{\|\mathfrak{X}_n{}'(\Gamma)\|_2}{\sqrt{n}} - \|\|\Gamma\|\|_F \right| \leq \frac{7}{8}\|\|\Gamma\|\|_F + \frac{48L\, d\, \|\Gamma\|_\infty}{\sqrt{n}} \qquad \text{for all } \Gamma \in \mathfrak{C}'(n; c_0),$$

which implies that the bound (28) holds. Consequently, in terms of the "bad" event, the claim of Theorem 1 is implied by the tail bound $\mathbb{P}[\mathcal{E}(\mathfrak{X}_n{}')] \leq 16\exp(-c'd\log d)$.

We now show that in order to establish a tail bound on $\mathcal{E}(\mathfrak{X}_n{}')$, it suffices to bound the probability of some simpler events $\mathcal{E}(\mathfrak{X}_n{}'; D)$, defined below. Since the definition of the set $\mathfrak{C}'(n; c_0)$ and event $\mathcal{E}(\mathfrak{X}_n{}')$ is invariant to rescaling of Γ, we may assume without loss of generality that $\|\Gamma\|_\infty = \frac{1}{d}$. The remaining degrees of freedom in the set $\mathfrak{C}'(n; c_0)$ can be parameterized in terms of the quantities $D = \|\|\Gamma\|\|_F$ and $\rho = \|\|\Gamma\|\|_1$. For any $\Gamma \in \mathfrak{C}'(n; c_0)$ with $\|\Gamma\|_\infty = \frac{1}{d}$ and $\|\|\Gamma\|\|_F \leq D$, we have $\|\|\Gamma\|\|_1 \leq \rho(D)$, where

$$\rho(D) := \frac{D^2}{c_0 L \sqrt{\frac{d\log d}{n}}}.$$

For each radius $D > 0$, consider the set

$$\mathfrak{B}(D) := \left\{ \Gamma \in \mathfrak{C}'(n; c_0) \mid \|\Gamma\|_\infty = \frac{1}{d}, \ \|\|\Gamma\|\|_F \leq D, \ \|\|\Gamma\|\|_1 \leq \rho(D) \right\}, \tag{29}$$

and the associated event

$$\mathcal{E}(\mathfrak{X}_n{}'; D) := \left\{ \exists \quad \Gamma \in \mathfrak{B}(D) \mid \left| \frac{\|\mathfrak{X}_n{}'(\Gamma)\|_2}{\sqrt{n}} - \|\|\Gamma\|\|_F \right| \geq \frac{3}{4}D + \frac{48L}{\sqrt{n}} \right\}.$$

The following lemma shows that it suffices to upper bound the probability of the event $\mathcal{E}(\mathfrak{X}_n{}'; D)$ for each fixed $D > 0$.

Lemma 3 *Suppose that are universal constants (c_1, c_2) such that*

$$\mathbb{P}[\mathcal{E}(\mathfrak{X}_n'; D)] \leq c_1 \exp(-c_2 n D^2) \tag{30}$$

for each fixed $D > 0$. Then there is a universal constant c_2' such that

$$\mathbb{P}[\mathcal{E}(\mathfrak{X}_n')] \leq c_1 \frac{\exp(-c_2' d \log d)}{1 - \exp(-c_2' d \log d)}.$$

The proof of this claim, provided in Appendix B, follows by a peeling argument.

5.2 Bounding the Probability of $\mathcal{E}(\mathfrak{X}_n'; D)$

Based on Lemma 3, it suffices to prove the tail bound (30) on the event $\mathcal{E}(\mathfrak{X}_n'; D)$ for each fixed $D > 0$. Let us define

$$Z_n(D) := \sup_{\Gamma \in \overline{\mathfrak{B}}(D)} \left| \frac{\|\mathfrak{X}_n'(\Gamma)\|_2}{\sqrt{n}} - \|\Gamma\|_F \right|,$$

where

$$\overline{\mathfrak{B}}(D) := \left\{ \Gamma \in \mathfrak{C}'(n; c_0) \mid \|\Gamma\|_\infty \leq \frac{1}{d}, \ \|\Gamma\|_F \leq D, \ \|\Gamma\|_1 \leq \rho(D) \right\}. \tag{31}$$

(The only difference from $\mathfrak{B}(D)$ is that we have relaxed to the inequality $\|\Gamma\|_\infty \leq \frac{1}{d}$.) In the remainder of this section, we prove that there are universal constants (c_1, c_2) such that

$$\mathbb{P}\left[Z_n(D) \geq \frac{3}{4} D + \frac{48L}{\sqrt{n}} \right] \leq c_1 \exp\left(-c_2 \frac{n D^2}{L^2}\right) \quad \text{for each fixed } D > 0. \tag{32}$$

This tail bound means that the condition of Lemma 3 is satisfied, and so completes the proof of Theorem 1.

In order to prove (32), we begin with a discretization argument. Let $\Gamma^1, \ldots, \Gamma^{N(\delta)}$ be a δ-covering of $\overline{\mathfrak{B}}(D)$ in the Frobenius norm. By definition, given an arbitrary $\Gamma \in \overline{\mathfrak{B}}(D)$, there exists some index $k \in \{1, \ldots, N(\delta)\}$ and a matrix $\Delta \in \mathbb{R}^{d \times d}$ with $\|\Delta\|_F \leq \delta$ such that $\Gamma = \Gamma^k + \Delta$. Therefore, we have

$$
\begin{aligned}
\frac{\|\mathfrak{X}_n'(\Gamma)\|_2}{\sqrt{n}} - \|\Gamma\|_F &= \frac{\|\mathfrak{X}_n'(\Gamma^k + \Delta)\|_2}{\sqrt{n}} - \|\Gamma^k + \Delta\|_F \\
&\leq \frac{\|\mathfrak{X}_n'(\Gamma^k)\|_2}{\sqrt{n}} + \frac{\|\mathfrak{X}_n'(\Delta)\|_2}{\sqrt{n}} - \|\Gamma^k\|_F + \|\Delta\|_F \\
&\leq \left| \frac{\|\mathfrak{X}_n'(\Gamma^k)\|_2}{\sqrt{n}} - \|\Gamma^k\|_F \right| + \frac{\|\mathfrak{X}_n'(\Delta)\|_2}{\sqrt{n}} + \delta,
\end{aligned}
$$

where we have used the triangle inequality. Following the same steps establishes that this inequality holds for the absolute value of the difference.

Moreover, since $\Delta = \Gamma^k - \Gamma$ with both Γ^k and Γ belonging to $\overline{\mathfrak{B}}(D)$, we have $\|\Delta\|_1 \leq 2\rho(D)$ and $\|\Delta\|_\infty \leq \frac{2}{d}$, where we have used the definition (29). Putting together the pieces, we conclude that

$$Z_n(D) \ \leq \ \delta \ + \ \max_{k=1,\ldots,N(\delta)} \left| \frac{\|\mathfrak{X}_n'(\Gamma^k)\|_2}{\sqrt{n}} - \|\Gamma^k\|_F \right| \ + \ \sup_{\Delta \in \mathfrak{D}(\delta, R)} \left| \frac{\|\mathfrak{X}_n'(\Delta)\|_2}{\sqrt{n}} \right|, \tag{33}$$

where

$$\mathfrak{D}(\delta, R) := \left\{ \Delta \in \mathbb{R}^{d \times d} \mid \|\Delta\|_F \leq \delta, \ |\|\Delta\||_1 \leq 2\rho(D), \ \|\Delta\|_\infty \leq \frac{2}{d} \right\}.$$

Note that the bound (33) holds for any choice of $\delta > 0$. We establish the tail bound (32) with the choice $\delta = D/8$, and using the following two lemmas. The first lemma provides control of the maximum over the covering set:

Lemma 4 *As long $d \geq 10$, we have*

$$\max_{k=1,\ldots,N(D/8)} \left| \frac{\|\mathfrak{X}_n'(\Gamma^k)\|_2}{\sqrt{n}} - \||\Gamma^k\||_F \right| \leq \frac{D}{8} + \frac{48L}{\sqrt{n}}$$

with probability greater than $1 - c\exp\left(-\frac{nD^2}{2048L^2}\right)$.

See Appendix C for the proof of this claim.

Our second lemma, proved in Appendix D, provides control over the final term in the upper bound (33).

Lemma 5

$$\sup_{\Delta \in \mathfrak{D}(\frac{D}{8}, R)} \left| \frac{\|\mathfrak{X}_n'(\Delta)\|_2}{\sqrt{n}} \right| \leq \frac{D}{2}$$

with probability at least $1 - 2\exp\left(-\frac{nD^2}{8192L^2}\right)$.

Combining these two lemmas with the upper bound (33) with $\delta = D/8$, we obtain

$$Z_n(D) \leq \frac{D}{8} + \frac{D}{8} + \frac{48L}{\sqrt{n}} + \frac{D}{2}$$
$$\leq \frac{3D}{4} + \frac{48L}{\sqrt{n}}$$

with probability at least $1 - 4\exp\left(-\frac{nD^2}{8192}\right)$, thereby establishing the tail bound (32) and completing the proof of Theorem 1.

6. Discussion

In this paper, we have established error bounds for the problem of weighted matrix completion based on partial and noisy observations. We proved both a general result, one which applies to any matrix, and showed how it yields corollaries for both the cases of exactly low-rank and approximately low-rank matrices. A key technical result is establishing that the matrix sampling operator satisfies a suitable form of restricted strong convexity (Negahban et al., 2009) over a set of matrices with controlled rank and spikiness. Since more restrictive properties such as RIP do not hold for matrix completion, this RSC ingredient is essential to our analysis. Our proof of the RSC condition relied on a number of techniques from empirical process and random matrix theory, including concentration of measure, contraction inequalities and the Ahlswede-Winter bound. Using information-theoretic

methods, we also proved that up to logarithmic factors, our error bounds cannot be improved upon by any algorithm, showing that our method is essentially minimax-optimal.

There are various open questions that remain to be studied. Although our analysis applies to both uniform and non-uniform sampling models, it is limited to the case where each row (or column) is sampled with a certain probability. It would be interesting to consider extensions to settings in which the sampling probability differed from entry to entry, as investigated empirically by Salakhutdinov and Srebro (2010). Although we have focused on least-squares losses in this paper, the notion of restricted strong convexity applies to more general loss functions. Indeed, it should be possible to combine the results of this paper with Proposition 2 in Negahban et al. (2009) so as to obtain bounds for matrix completion with general losses. Lastly, although this paper has focused on statistical consequences, the RSC property also has implications for the fast convergence of gradient-type algorithms for solving matrix completion problems (Agarwal et al., 2011).

Acknowledgments

Appendix A. Proof of Lemma 2

We proceed via the probabilistic method, in particular by showing that a random procedure succeeds in generating such a set with probability at least $1/2$. Let $M' = \exp\left(\frac{rd}{128}\right)$, and for each $\ell = 1, \ldots, M'$, we draw a random matrix $\widetilde{\Theta}^\ell \in \mathbb{R}^{d \times d}$ according to the following procedure:

(a) For rows $i = 1, \ldots, r$ and for each column $j = 1, \ldots, d$, choose each $\widetilde{\Theta}^\ell_{ij} \in \{-1, +1\}$ uniformly at random, independently across (i, j).

(b) For rows $i = r+1, \ldots, d$, set $\widetilde{\Theta}^\ell_{ij} = 0$.

We then let $Q \in \mathbb{R}^{d \times d}$ be a random unitary matrix, and define $\Theta^\ell = \frac{\delta}{\sqrt{rd}} Q \widetilde{\Theta}^\ell$ for all $\ell = 1, \ldots, M'$. The remainder of the proof analyzes the random set $\{\Theta^1, \ldots, \Theta^{M'}\}$, and shows that it contains a subset of size at least $M = M'/4$ that has properties (a) through (d) with probability at least $1/2$.

By construction, each matrix $\widetilde{\Theta}^\ell$ has rank at most r, and Frobenius norm $\|\|\widetilde{\Theta}^\ell\|\|_F = \sqrt{rd}$. Since Q is unitary, the rescaled matrices Θ^ℓ have Frobenius norm $\|\|\Theta^\ell\|\|_F = \delta$. We now prove that

$$\|\|\Theta^\ell - \Theta^k\|\|_F \geq \delta \qquad \text{for all } \ell \neq k$$

with probability at least $7/8$. Again, since Q is unitary, it suffices to show that $\|\|\widetilde{\Theta}^\ell - \widetilde{\Theta}^k\|\|_F \geq \sqrt{rd}$ for any pair $\ell \neq k$. We have

$$\frac{1}{rd} \|\|\widetilde{\Theta}^k - \widetilde{\Theta}^\ell\|\|_F^2 = \frac{1}{rd} \sum_{i=1}^r \sum_{j=1}^d \left(\widetilde{\Theta}^\ell_{ij} - \widetilde{\Theta}^k_{ij}\right)^2.$$

This is a sum of rd i.i.d. variables, each bounded by 4. The mean of the sum is 2, so that the Hoeffding bound implies that

$$\mathbb{P}\left[\frac{1}{rd} \|\|\widetilde{\Theta}^k - \widetilde{\Theta}^\ell\|\|_F^2 \leq 2 - t\right] \leq 2 \exp\left(-rd t^2/32\right).$$

Since there are less than $(M')^2$ pairs of matrices in total, setting $t = 1$ yields

$$\mathbb{P}\Big[\min_{\ell,k=1,\ldots,M'} \frac{\|\|\widetilde{\Theta}^\ell - \widetilde{\Theta}^k\|\|_F^2}{rd} \geq 1\Big] \geq 1 - 2\exp\big(-\frac{rd}{32} + 2\log M'\big) \geq \frac{7}{8},$$

where we have used the facts $\log M' = \frac{rd}{128}$ and $d \geq 10$. Recalling the definition of Θ^ℓ, we conclude that

$$\mathbb{P}\Big[\min_{\ell,k=1,\ldots,M'} \|\|\Theta^\ell - \Theta^k\|\|_F^2 \geq \delta^2\Big] \geq \frac{7}{8}.$$

We now establish bounds on $\alpha_{\mathrm{sp}}(\Theta^\ell)$ and $\|\|\Theta^\ell\|\|_2$. We first prove that for any fixed index $\ell \in \{1, 2, \ldots, M'\}$, our construction satisfies

$$\mathbb{P}\Big[\alpha_{\mathrm{sp}}(\Theta^\ell) \leq \sqrt{32\log d}\Big] \geq \frac{3}{4}. \tag{34}$$

Indeed, for any pair of indices (i, j), we have $|\Theta_{ij}^\ell| = |\langle q_i, v_j \rangle|$, where $q_i \in \mathbb{R}^d$ is drawn from the uniform distribution over the d-dimensional sphere, and $\|v_j\|_2 = \sqrt{r}\,\frac{\delta}{\sqrt{rd}} = \frac{\|\|\Theta^\ell\|\|_F}{\sqrt{d}}$. By Levy's theorem for concentration on the sphere (Ledoux, 2001), we have

$$\mathbb{P}\big[|\langle q_i, v_j \rangle| \geq t\big] \leq 2\exp\big(-\frac{d^2}{8\|\|\Theta^\ell\|\|_F^2}t^2\big).$$

Setting $t = s/d$ and taking the union bound over all d^2 indices, we obtain

$$\mathbb{P}\big[d\,\|\Theta^\ell\|_\infty \geq s\big] \leq 2\exp\Big(-\frac{1}{8\|\|\Theta^\ell\|\|_F^2}s^2 + 2\log d\Big).$$

This probability is less than $1/2$ for $s = \|\|\Theta^\ell\|\|_F\sqrt{32\log d}$ and $d \geq 2$, which establishes the intermediate claim (34).

Finally, we turn to property (d). For each fixed ℓ, by definition of Θ^ℓ and the unitary nature of Q, we have $\|\|\Theta^\ell\|\|_2 = \frac{\delta}{\sqrt{rd}}\|\|U\|\|_2$, where $U \in \{-1, +1\}^{r \times d}$ is a random matrix with i.i.d. Rademacher (and hence sub-Gaussian) entries. Known results on sub-Gaussian matrices (Vershynin, 2012) yield

$$\mathbb{P}\Big[\frac{\delta}{\sqrt{rd}}\|\|U\|\|_2 \leq \frac{2\delta}{\sqrt{rd}}\big(\sqrt{r} + \sqrt{d}\big)\Big] \geq 1 - 2\exp\big(-\frac{1}{4}(\sqrt{r} + \sqrt{d})^2\big) \geq \frac{3}{4}$$

for $d \geq 10$. Since $r \leq d$, we conclude that

$$\mathbb{P}\Big[\|\|\Theta^\ell\|\|_2 \leq \frac{4\delta}{\sqrt{r}}\Big] \geq \frac{3}{4}. \tag{35}$$

By combining the bounds (34) and (35), we find that for each fixed $\ell = 1, \ldots, M'$, we have

$$\mathbb{P}\Big[\|\|\Theta^\ell\|\|_2 \leq \frac{4\delta}{\sqrt{r}}, \ \frac{\alpha_{\mathrm{sp}}(\Theta^\ell)}{\|\|\Theta\|\|_F} \leq \sqrt{32\log d}\Big] \geq \frac{1}{2}. \tag{36}$$

Consider the event $\mathcal{E}$ that there exists a subset $S \subset \{1, \ldots, M'\}$ of cardinality $M = \frac{1}{4}M'$ such that

$$\|\|\Theta^\ell\|\|_2 \leq 4\frac{\delta}{\sqrt{n}}, \quad \text{and} \quad \frac{\alpha_{\mathrm{sp}}(\Theta^\ell)}{\|\|\Theta\|\|_F} \leq \sqrt{32\log d} \qquad \text{for all } \ell \in S.$$

By the bound (36), we have

$$\mathbb{P}[\mathcal{E}] \geq \sum_{k=M}^{M'} \binom{M'}{k} (1/2)^k.$$

Since we have chosen $M < M'/2$, we are guaranteed that $\mathbb{P}[\mathcal{E}] \geq 1/2$, thereby completing the proof.

Appendix B. Proof of Lemma 3

We first observe that for any $\Gamma \in \mathcal{C}'(n; c_0)$ with $\|\Gamma\|_\infty = \frac{1}{d}$, we have

$$\|\|\Gamma\|\|_F^2 \geq c_0 \|\|\Gamma\|\|_1 \sqrt{\frac{d \log d}{n}} \geq c_0 \|\|\Gamma\|\|_F \sqrt{\frac{d \log d}{n}},$$

whence $\|\|\Gamma\|\|_F \geq c_0 \sqrt{\frac{d \log d}{n}}$. Accordingly, recalling the definition (29), it suffices to restrict our attention to sets $\mathfrak{B}(D)$ with $D \geq \mu := c_0 \sqrt{\frac{d \log d}{n}}$. For $\ell = 1, 2, \ldots$ and $\alpha = \frac{7}{6}$, define the sets

$$\mathbb{S}_\ell := \left\{ \Gamma \in \mathcal{C}'(n; c_0) \mid \|\Gamma\|_\infty = \frac{1}{d}, \quad \alpha^{\ell-1} \mu \leq \|\|\Gamma\|\|_F \leq \alpha^\ell \mu, \text{ and } \|\|\Gamma\|\|_1 \leq \rho(\alpha^\ell \mu) \right\}. \tag{37}$$

From the definition (29), note that by construction, we have $\mathbb{S}_\ell \subset \mathfrak{B}(\alpha^\ell \mu)$.

Now if the event $\mathcal{E}(\mathfrak{X}_n')$ holds for some matrix Γ, then this matrix Γ must belong to some set $\mathbb{S}_\ell$. When $\Gamma \in \mathbb{S}_\ell$, then we are guaranteed the existence of a matrix $\Gamma \in \mathfrak{B}(\alpha^\ell \mu)$ such that

$$\begin{aligned}
\left| \frac{\|\mathfrak{X}_n'(\Gamma)\|_2}{\sqrt{n}} - \|\|\Gamma\|\|_F \right| &> \frac{7}{8} \|\|\Gamma\|\|_F + \frac{48L}{\sqrt{n}} \\
&\geq \frac{7}{8} \alpha^{\ell-1} \mu + \frac{48L}{\sqrt{n}} \\
&= \frac{3}{4} \alpha^\ell \mu + \frac{48L}{\sqrt{n}},
\end{aligned}$$

where the final equality follows since $\alpha = 7/6$. Thus, we have shown that when the violating matrix $\Gamma \in \mathbb{S}_\ell$, then event $\mathcal{E}(\mathfrak{X}_n'; \alpha^\ell \mu)$ must hold. Since any violating matrix must fall into some set $\mathbb{S}_\ell$, the union bound implies that

$$\begin{aligned}
\mathbb{P}[\mathcal{E}(\mathfrak{X}_n')] &\leq \sum_{\ell=1}^\infty \mathbb{P}[\mathcal{E}(\mathfrak{X}_n'; \alpha^\ell \mu)] \\
&\leq c_1 \sum_{\ell=1}^\infty \exp\left(- c_2 n \alpha^{2\ell} \mu^2\right) \\
&\leq c_1 \sum_{\ell=1}^\infty \exp\left(- 2c_2 \log(\alpha) \ell n \mu^2\right) \\
&\leq c_1 \frac{\exp(-c_2' n \mu^2)}{1 - \exp(-c_2' n \mu^2)}.
\end{aligned}$$

Since $n\mu^2 = \Omega(d \log d)$, the claim follows.

Appendix C. Proof of Lemma 4

For a fixed matrix Γ, define the function $F_\Gamma(\mathfrak{X}_n') = \frac{1}{\sqrt{n}}\|\mathfrak{X}_n'(\Gamma)\|_2$. We prove the lemma in two parts: first, we establish that for any fixed Γ, the function F_Γ satisfies the tail bound

$$\mathbb{P}\left[|F_\Gamma(\mathfrak{X}_n') - \|\Gamma\|_F| \geq \delta + \frac{48L}{\sqrt{n}}\right] \leq 4\exp\left(-\frac{n\delta^2}{4L^2}\right). \tag{38}$$

We then show that there exists a δ-covering of $\overline{\mathfrak{B}}(D)$ such that

$$\log N(\delta) \leq 36\big(\rho(D)/\delta\big)^2 d. \tag{39}$$

Combining the tail bound (38) with the union bound, we obtain

$$\mathbb{P}\left[\max_{k=1,\dots,N(\delta)} |F_\Gamma(\mathfrak{X}_n') - \|\Gamma^k\|_F| \geq \delta + \frac{16L}{\sqrt{n}}\right] \leq 4\exp\left(-\frac{n\delta^2}{4L^2} + \log N(\delta)\right)$$

$$\leq 4\exp\left\{-\frac{n\delta^2}{4L^2} + 36\big(\rho(D)/\delta\big)^2 d\right\}$$

where the final inequality uses the bound (39). Since Lemma 4 is based on the choice $\delta = D/8$, it suffices to show that

$$\frac{nD^2}{512L^2} \geq 36\left(\rho(D)/(D/8)\right)^2 d$$

$$= 36\left(\frac{8D}{c_0 L}\sqrt{\frac{n}{d\log d}}\right)^2 d$$

$$= \frac{2304 D^2}{c_0^2 L^2}\frac{n}{\log d}.$$

Noting that the terms involving D^2, L^2, and n both cancel out, we see that for any fixed c_0, this inequality holds once $\log d$ is sufficiently large. By choosing c_0 sufficiently large, we can ensure that it holds for all $d \geq 2$.

It remains to establish the two intermediate claims (38) and (39).

C.1 Upper Bounding the Covering Number (39)

We start by proving the upper bound (39) on the covering number. To begin, let $\widetilde{N}(\delta)$ denote the δ-covering number (in Frobenius norm) of the nuclear norm ball $\mathbb{B}_1(\rho(D)) = \{\Delta \in \mathbb{R}^{d\times d} \mid \|\|\Delta\|\|_1 \leq \rho(D)\}$, and let $N(\delta)$ be the covering number of the set $\overline{\mathfrak{B}}(D)$. We first claim that $N(\delta) \leq \widetilde{N}(\delta)$. Let $\{\Gamma^1,\dots,\Gamma^{\widetilde{N}(\delta)}\}$ be a δ-cover of $\mathbb{B}_1(\rho(D))$, From Equation (31), note that the set $\overline{\mathfrak{B}}(D)$ is contained within $\mathbb{B}_1(\rho(D))$; in particular, it is obtained by intersecting the latter set with the set

$$S := \{\Delta \in \mathbb{R}^{d\times d} \mid \|\Delta\|_\infty \leq \frac{1}{d}, \ \|\|\Delta\|\|_F \leq D\}.$$

Letting Π_S denote the projection operator under Frobenius norm onto this set, we claim that $\{\Pi_S(\Gamma^j), j = 1, \ldots, \widetilde{N}(\delta)\}$ is a δ-cover of $\overline{\mathfrak{B}}(D)$. Indeed, since S is non-empty, closed and convex, the projection operator is non-expansive (Bertsekas, 1995), and thus for any $\Gamma \in \overline{\mathfrak{B}}(D) \subset S$, we have

$$\||\Pi_S(\Gamma^j) - \Gamma\||_F = \||\Pi_S(\Gamma^j) - \Pi_S(\Gamma)\||_F \leq \||\Gamma^j - \Gamma\||_F,$$

which establishes the claim.

We now upper bound $\widetilde{N}(\delta)$. Let $G \in \mathbb{R}^{d \times d}$ be a random matrix with i.i.d. $N(0,1)$ entries. By Sudakov minoration (cf. Theorem 5.6 in Pisier, 1989), we have

$$\sqrt{\log \widetilde{N}(\delta)} \leq \frac{3}{\delta} \mathbb{E}\Big[\sup_{\||\Delta\||_1 \leq \rho(D)} \langle\langle G, \Delta \rangle\rangle \Big]$$

$$\leq \frac{3\rho(D)}{\delta} \mathbb{E}\big[\||G\||_2\big],$$

where the second inequality follows from the duality between the nuclear and operator norms. From known results on the operator norms of Gaussian random matrices (Davidson and Szarek, 2001), we have the upper bound $\mathbb{E}[\||G\||_2] \leq 2\sqrt{d}$, so that

$$\sqrt{\log \widetilde{N}(\delta)} \leq \frac{6\rho(D)}{\delta} \sqrt{d},$$

thereby establishing the bound (39).

C.2 Establishing the Tail Bound (38)

Recalling the definition of the operator $\mathfrak{X}_n'$, we have

$$F_\Gamma(\mathfrak{X}_n') = \frac{1}{\sqrt{n}} \Big\{ \sum_{i=1}^n \langle\langle \widetilde{X}^{(i)}, \Gamma \rangle\rangle^2 \Big\}^{1/2}$$

$$= \frac{1}{\sqrt{n}} \sup_{\|u\|_2 = 1} \sum_{i=1}^n u_i \langle\langle \widetilde{X}^{(i)}, \Gamma \rangle\rangle$$

$$= \frac{1}{\sqrt{n}} \sup_{\|u\|_2 = 1} \sum_{i=1}^n u_i Y_i$$

where we have defined the random variables $Y_i := \langle\langle \widetilde{X}^{(i)}, \Gamma \rangle\rangle$. Note that each Y_i is zero-mean, and bounded by $2L$ since

$$|Y_i| = |\langle\langle \widetilde{X}^{(i)}, \Gamma \rangle\rangle|$$

$$\leq \big(\sum_{a,b} |\widetilde{X}^{(i)}|_{ab} \big) \|\Gamma\|_\infty \leq 2L.$$

where we have used the facts that $\|\Gamma\|_\infty \leq 2/d$, and $\sum_{a,b} |\widetilde{X}^{(i)}|_{ab} \leq L\, d$, by definition of the matrices $\widetilde{X}^{(i)}$.

Therefore, applying Corollary 4.8 from Ledoux (2001), we conclude that

$$\mathbb{P}\big[|F_\Gamma(\mathfrak{X}_n') - \mathbb{E}[F_\Gamma(\mathfrak{X}_n')]| \geq \delta + \frac{32L}{\sqrt{n}}\big] \leq 4 \exp\big(-\frac{n\delta^2}{4L^2}\big).$$

The same corollary implies that

$$\left| \sqrt{\mathbb{E}[F_\Gamma^2(\mathfrak{X}_n{}')]} - \mathbb{E}[F_\Gamma(\mathfrak{X}_n{}')] \right| \leq \frac{16L}{\sqrt{n}}.$$

Since $\mathbb{E}[F_\Gamma^2(\mathfrak{X}_n{}')] = \|\|\Gamma\|\|_F^2$, the tail bound (38) follows.

Appendix D. Proof of Lemma 5

From the proof of Lemma 4, recall the definition $F_\Gamma(\mathfrak{X}_n{}') = \frac{1}{\sqrt{n}} \|\mathfrak{X}_n{}'(\Gamma)\|_2$ where $\mathfrak{X}_n{}'$ is the random sampling operator defined by the n matrices $(\widetilde{X}^{(1)}, \dots, \widetilde{X}^{(n)})$. Using this notation, our goal is to bound the function

$$G(\mathfrak{X}_n{}') := \sup_{\Delta \in \mathfrak{D}(\delta, R)} F_\Delta(\mathfrak{X}_n{}'),$$

where we recall that $\mathfrak{D}(\delta, R) := \left\{ \Delta \in \mathbb{R}^{d_r \times d_c} \mid \|\|\Delta\|\|_F \leq \delta, \ \|\|\Delta\|\|_1 \leq 2\rho(D), \ \|\Delta\|_\infty \leq \frac{2}{d} \right\}$. Ultimately, we will set $\delta = \frac{D}{8}$, but we use δ until the end of the proof for compactness in notation.

Our approach is a standard one: first show concentration of G around its expectation $\mathbb{E}[G(\mathfrak{X}_n{}')]$, and then upper bound the expectation. We show concentration via a bounded difference inequality; since G is a symmetric function of its arguments, it suffices to establish the bounded difference property with respect to the first co-ordinate. In order to do so, consider a second operator $\widetilde{\mathfrak{X}_n{}'}$ defined by the matrices $(Z^{(1)}, \widetilde{X}^{(2)}, \dots, \widetilde{X}^{(n)})$, differing from $\mathfrak{X}_n{}'$ only in the first matrix. Given the pair $(\mathfrak{X}_n{}', \widetilde{\mathfrak{X}_n{}'})$, we have

$$\begin{aligned}
G(\mathfrak{X}_n{}') - G(\widetilde{\mathfrak{X}_n{}'}) &= \sup_{\Delta \in \mathfrak{D}(\delta, R)} F_\Delta(\mathfrak{X}_n{}') - \sup_{\Theta \in \mathfrak{D}(\delta, R)} F_\Theta(\widetilde{\mathfrak{X}_n{}'}) \\
&\leq \sup_{\Delta \in \mathfrak{D}(\delta, R)} \left[F_\Delta(\mathfrak{X}_n{}') - F_\Delta(\widetilde{\mathfrak{X}_n{}'}) \right] \\
&\leq \sup_{\Delta \in \mathfrak{D}(\delta, R)} \frac{1}{\sqrt{n}} \|\mathfrak{X}_n{}'(\Delta) - \widetilde{\mathfrak{X}_n{}'}(\Delta)\|_2 \\
&= \sup_{\Delta \in \mathfrak{D}(\delta, R)} \frac{1}{\sqrt{n}} \left| \langle\!\langle \widetilde{X}^{(1)} - Z^{(1)}, \ \Delta \rangle\!\rangle \right|.
\end{aligned}$$

For any fixed $\Delta \in \mathfrak{D}(\delta, R)$, we have

$$\left| \langle\!\langle \widetilde{X}^{(1)} - Z^{(1)}, \ \Delta \rangle\!\rangle \right| \leq 2Ld \, \|\Delta\|_\infty \leq 4L,$$

where we have used the fact that the matrix $\widetilde{X}^{(1)} - Z^{(1)}$ is non-zero in at most two entries with values upper bounded by $2Ld$. Combining the pieces yields $G(\mathfrak{X}_n{}') - G(\widetilde{\mathfrak{X}_n{}'}) \leq \frac{4L}{\sqrt{n}}$. Since the same argument can be applied with the roles of $\mathfrak{X}_n{}'$ and $\widetilde{\mathfrak{X}_n{}'}$ interchanged, we conclude that $|G(\mathfrak{X}_n{}') - G(\widetilde{\mathfrak{X}_n{}'})| \leq \frac{4L}{\sqrt{n}}$. Therefore, by the bounded differences variant of the Azuma-Hoeffding inequality (Ledoux, 2001), we have

$$\mathbb{P}\left[|G(\mathfrak{X}_n{}') - \mathbb{E}[G(\mathfrak{X}_n{}')]| \geq t \right] \leq 2\exp\left(-\frac{nt^2}{32L^2} \right). \tag{40}$$

Next we bound the expectation. First applying Jensen's inequality, we have

$$(\mathbb{E}[G(\mathfrak{X}_n')])^2 \le \mathbb{E}[G^2(\mathfrak{X}_n')]$$

$$= \mathbb{E}\Big[\sup_{\Delta \in \mathfrak{D}(\delta,R)} \frac{1}{n} \sum_{i=1}^n \langle\langle \widetilde{X}^{(i)}, \Delta \rangle\rangle^2 \Big]$$

$$= \mathbb{E}\Big[\sup_{\Delta \in \mathfrak{D}(\delta,R)} \Big\{ \frac{1}{n} \sum_{i=1}^n \big[\langle\langle \widetilde{X}^{(i)}, \Delta \rangle\rangle^2 - \mathbb{E}[\langle\langle \widetilde{X}^{(i)}, \Delta \rangle\rangle^2] \big] + \||\Delta\||_F^2 \Big\} \Big]$$

$$\le \mathbb{E}\Big[\sup_{\Delta \in \mathfrak{D}(\delta,R)} \Big\{ \frac{1}{n} \sum_{i=1}^n \big[\langle\langle \widetilde{X}^{(i)}, \Delta \rangle\rangle^2 - \mathbb{E}[\langle\langle \widetilde{X}^{(i)}, \Delta \rangle\rangle^2] \big] \Big\} \Big] + \delta^2,$$

where we have used the fact that $\mathbb{E}[\langle\langle \widetilde{X}^{(i)}, \Delta \rangle\rangle^2 = \||\Delta\||_F^2 \le \delta^2$. Now a standard symmetrization argument (Ledoux and Talagrand, 1991) yields

$$\mathbb{E}_{\mathfrak{X}_n'}[G^2(\mathfrak{X}_n')] \le 2 \mathbb{E}_{\mathfrak{X}_n', \varepsilon} \Big[\sup_{\Delta \in \mathfrak{D}(\delta,R)} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle\langle \widetilde{X}^{(i)}, \Delta \rangle\rangle^2 \Big] + \delta^2,$$

where $\{\varepsilon_i\}_{i=1}^n$ is an i.i.d. Rademacher sequence. Since $|\langle\langle \widetilde{X}^{(i)}, \Delta \rangle\rangle| \le 2L$ for all i, the Ledoux-Talagrand contraction inequality (p. 112, Ledoux and Talagrand, 1991) implies that

$$\mathbb{E}[G^2(\mathfrak{X}_n')] \le 16L\, \mathbb{E}\Big[\sup_{\Delta \in \mathfrak{D}(\delta,R)} \Big\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle\langle \widetilde{X}^{(i)}, \Delta \rangle\rangle \Big\} \Big] + \delta^2.$$

By the duality between operator and nuclear norms, we have

$$\Big| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle\langle \widetilde{X}^{(i)}, \Delta \rangle\rangle \Big| \le \||\frac{1}{n} \sum_{i=1}^n \varepsilon_i \widetilde{X}^{(i)}\||_2\, \||\Delta\||_1,$$

and hence, since $\||\Delta\||_1 \le \rho(D)$ for all $\Delta \in \mathfrak{D}(\delta, R)$, we have

$$\mathbb{E}[G^2(\mathfrak{X}_n')] \le 16L\rho(D)\, \mathbb{E}\Big[\||\frac{1}{n} \sum_{i=1}^n \varepsilon_i \widetilde{X}^{(i)}\||_2 \Big] + \delta^2. \tag{41}$$

It remains to bound the operator norm $\mathbb{E}\big[\||\frac{1}{n} \sum_{i=1}^n \varepsilon_i \widetilde{X}^{(i)}\||_2 \big]$. The following lemma, proved in Appendix E, provides a suitable upper bound:

Lemma 6 *We have the upper bound*

$$\mathbb{E}\Big[\||\frac{1}{n} \sum_{i=1}^n \varepsilon_i \widetilde{X}^{(i)}\||_2 \Big] \le 10 \max\Big\{ \sqrt{\frac{d \log d}{n}}, \frac{L d \log d}{n} \Big\}.$$

Thus, as long as $n = \Omega(L d \log d)$, combined with the earlier bound (41), we conclude that

$$\mathbb{E}[G(\mathfrak{X}_n')] \le \sqrt{\mathbb{E}[G^2(\mathfrak{X}_n')]} \le \Big[160 L \rho(D) \sqrt{\frac{d \log d}{n}} + \delta^2 \Big]^{1/2},$$

where we have used the fact that $L \ge 1$. By definition of $\rho(D)$, we have

$$160 L^2 \rho(D) \sqrt{\frac{d \log d}{n}} = \frac{160}{c_0} D^2 \le \Big(\frac{5D}{16} \Big)^2,$$

where the final inequality can be guaranteed by choosing c_0 sufficiently large.

Consequently, recalling our choice $\delta = D/8$ and using the inequality $\sqrt{a^2 + b^2} \leq |a| + |b|$, we obtain

$$\mathbb{E}[G(\mathfrak{X}_n{}')] \leq \frac{5}{16}D + \frac{D}{8} = \frac{7}{16}D.$$

Finally, setting $t = \frac{D}{16}$ in the concentration bound (40) yields

$$G(\mathfrak{X}_n{}') \leq \frac{D}{16} + \frac{7}{16}D = \frac{D}{2}$$

with probability at least $1 - 2\exp\left(-c'\frac{nD^2}{L^2}\right)$ as claimed.

Appendix E. Proof of Lemma 6

We prove this lemma by applying a form of Ahlswede-Winter matrix bound (2002), as stated in Appendix F, to the matrix $Y^{(i)} := \varepsilon_i \widetilde{X}^{(i)}$. We first compute the quantities involved in Lemma 7. Note that $Y^{(i)}$ is a zero-mean random matrix, and satisfies the bound

$$\|\|Y^{(i)}\|\|_2 = d\frac{1}{\sqrt{R_{j(i)}}\sqrt{C_{k(i)}}}\|\|\varepsilon_i\, e_{j(i)}\, e_{k(i)}^T\|\|_2 \leq Ld.$$

Let us now compute the quantities σ_i in Lemma 7. We have

$$\mathbb{E}\left[(Y^{(i)T})Y^{(i)}\right] = \mathbb{E}\left[\frac{d^2}{R_{j(i)}\,C_{k(i)}}e_{k(i)}e_{k(i)}^T\right] = dI_{d \times d}$$

and similarly, $\mathbb{E}\left[Y^{(i)}\,(Y^{(i)})^T\right] = dI_{d \times d}$, so that

$$\sigma_i^2 = \max\left\{\|\|\mathbb{E}\left[Y^{(i)}\,(Y^{(i)})^T\right]\|\|_2,\ \|\|\mathbb{E}\left[(Y^{(i)})^T\,Y^{(i)}\right]\|\|_2\right\} = d.$$

Thus, applying Lemma 7 yields the tail bound

$$\mathbb{P}\left[\|\|\sum_{i=1}^n \varepsilon_i\widetilde{X}^{(i)}\|\|_2 \geq t\right] \leq 2d\,\max\left\{\exp(-\frac{t^2}{4nd}),\ \exp(-\frac{t}{2Ld})\right\}.$$

Setting $t = n\delta$, we obtain

$$\mathbb{P}\left[\|\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\widetilde{X}^{(i)}\|\|_2 \geq \delta\right] \leq 2d\,\max\left\{\exp(-\frac{n\delta^2}{4d}),\exp(-\frac{n\delta}{2Ld})\right\}.$$

Recall that for any non-negative random variable T, we have $\mathbb{E}[T] = \int_0^\infty \mathbb{P}[T \geq s]ds$. Applying this fact to $T := \|\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\widetilde{X}^{(i)}\|\|_2$ and integrating the tail bound, we obtain

$$\mathbb{E}\left[\|\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\widetilde{X}^{(i)}\|\|_2\right] \leq 10\,\max\left\{\sqrt{\frac{d\log d}{n}},\ \frac{Ld\log d}{n}\right\}.$$

Appendix F. Ahlswede-Winter Matrix Bound

Here we state a Bernstein version of the Ahlswede-Winter (2002) tail bound for the operator norm of a sum of random matrices. The version here is a slight weakening (but sufficient for our purposes) of a result due to Recht (2011); we also refer the reader to the chapter of Vershynin (2012), and the strengthened results provided by Tropp (2010).

Let $Y^{(i)}$ be independent $d_r \times d_c$ zero-mean random matrices such that $\|\|Y^{(i)}\|\|_2 \le M$, and define $\sigma_i^2 := \max\left\{ \|\|\mathbb{E}[(Y^{(i)})^T Y^{(i)}]\|\|_2, \quad \|\|\mathbb{E}[Y^{(i)}(Y^{(i)})^T]\|\|_2 \right\}$ as well as $\sigma^2 := \sum_{i=1}^n \sigma_i^2$.

Lemma 7 *We have*

$$\mathbb{P}\left[\|\|\sum_{i=1}^n Y^{(i)}\|\|_2 \ge t\right] \le (d_r \times d_c) \max\left\{ \exp(-t^2/(4\sigma^2)), \, \exp(-\frac{t}{2M}) \right\}.$$

As noted by Vershynin (2009), the same bound also holds under the assumption that each $Y^{(i)}$ is sub-exponential with parameter $M = \|Y^{(i)}\|_{\psi_1}$. Here we are using the Orlicz norm

$$\|Z\|_{\psi_1} := \inf\{t > 0 \mid \mathbb{E}[\psi(|Z|/t)] < \infty\},$$

defined by the function $\psi_1(x) = \exp(x) - 1$, as is appropriate for sub-exponential variables (e.g., see Ledoux and Talagrand, 1991).

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Annals of Statistics*, 2011. To appear.

R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, March 2002.

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

D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.

E. Candès 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. *Found. Comput. Math.*, 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.

K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pages 317–336. Elsevier, Amsterdam, NL, 2001.

M. Deza and M. Laurent. *Geometry of Cuts and Metric Embeddings*. Springer-Verlag, New York, 1997.

M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford, 2002. Available online: http://faculty.washington.edu/mfazel/thesis-final.pdf.

D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, March 2011.

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

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, June 2010a.

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

M. Laurent. Matrix completion problems. In *The Encyclopedia of Optimization*, pages 221—229. Kluwer Academic, 2001.

M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. Technical Report UILU-ENG-09-2214, Univ. Illinois, Urbana-Champaign, July 2009.

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

S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2009. To appear in Statistical Science.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 2007/76, CORE, Universit'e catholique de Louvain, 2007.

G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, UK, 1989.

B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12: 3413–3430, 2011.

B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 39(2):887–930, 2011.

R. Salakhutdinov and N. Srebro. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. Technical Report abs/1002.2780v1, Toyota Institute of Technology, 2010.

N. Srebro. *Learning with Matrix Factorizations*. PhD thesis, MIT, 2004. Available online: http://ttic.uchicago.edu/ nati/Publications/thesis.pdf.

N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2004.

N. Srebro, N. Alon, and T. S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2005.

S. J. Szarek. The finite dimensional basis problem with an appendix on nets of grassmann manifolds. *Acta Mathematica*, 151:153–179, 1983.

J. Tropp. User-friendly tail bounds for matrix martingales. Technical report, Caltech, April 2010.

R. Vershynin. A note on sums of independent random matrices after Ahlswede-Winter. Technical report, Univ. Michigan, December 2009.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*, 2012. Available at `http://www-personal.umich.edu/ ~romanv/papers/non-asymptotic-rmt-plain.pdf`.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

B. Yu. Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, Berlin, 1997.