On the Power of Adaptivity in Matrix Completion and Approximation

Akshay Krishnamurthy [*1] and Aarti Singh [†2]

[1]Computer Science Department, Carnegie Mellon University
[2]Machine Learning Department, Carnegie Mellon University

July 15, 2014

Abstract

We consider the related tasks of matrix completion and matrix approximation from missing data and propose adaptive sampling procedures for both problems. We show that adaptive sampling allows one to eliminate standard incoherence assumptions on the matrix row space that are necessary for passive sampling procedures. For exact recovery of a low-rank matrix, our algorithm judiciously selects a few columns to observe in full and, with few additional measurements, projects the remaining columns onto their span. This algorithm exactly recovers an $n \times n$ rank r matrix using $O(nr\mu_0 \log^2(r))$ observations, where μ_0 is a coherence parameter on the column space of the matrix. In addition to completely eliminating any row space assumptions that have pervaded the literature, this algorithm enjoys a better sample complexity than any existing matrix completion algorithm. To certify that this improvement is due to adaptive sampling, we establish that row space coherence is necessary for passive sampling algorithms to achieve non-trivial sample complexity bounds.

For constructing a low-rank approximation to a high-rank input matrix, we propose a simple algorithm that thresholds the singular values of a zero-filled version of the input matrix. The algorithm computes an approximation that is nearly as good as the best rank-r approximation using $O(nr\mu \log^2(n))$ samples, where μ is a slightly different coherence parameter on the matrix columns. Again we eliminate assumptions on the row space.

1 Introduction

While the cost of data acquisition has decreased significantly across the spectrum of scientific applications, it has failed to keep up with the increasing complexity of the systems and processes being studied. As a concrete example, routing optimization in communication networks or personalization in social networks involve making inferences at the granularity of individual nodes and links, and these tasks become more challenging as the networks grow. In other words, both the amount of data generated by the network *and* the complexity of statistical problems scale with the size of the network. This phenomenon is prevalent

[*]akshaykr@cs.cmu.edu
[†]aarti@cs.cmu.edu

across modern statistical applications and the result is that analysts face the challenge of making meaningful inferences from extremely undersampled datasets.

A number of innovative algorithmic ideas, including the compressive sensing paradigm [Candès and Wakin, 2008], show that one can reliabily perform inference in the presence of undersampling, but many of these methods are limited in that the algorithms do not *interact* with the acquisition mechanism. Adaptive sampling, where the inference algorithm interacts with the data measurement process, is a promising approach to tolerate more severe undersampling. Indeed this is true in several settings, where we are now aware of adaptive sampling methods that outperform all passive schemes [Haupt et al., 2011].

This paper proposes adaptive sampling algorithms for low rank matrix completion and matrix approximation. In the completion problem, we would like to exactly recover a low rank matrix after observing only a small fraction of its entries. In the approximation problem, rather than exact recovery, we aim to find a low rank matrix that approximates, in a precise sense, the input matrix, which need not be low rank. In both problems, we are only allowed to observe a small number of matrix entries, although these entries can be chosen sequentially and in a feedback-driven manner.

The thesis of our work is that adaptive sampling allows us to remove incoherence assumptions that have pervaded the literature. Previous analyses show that if the energy of the matrix is spread out fairly uniformly across its coordinates, then passive uniform-at-random samples suffice for completion or approximation. In contrast, our work shows that adaptive sampling algorithms can focus measurements appropriately to solve these problems even if the energy is non-uniformly distributed. Handling non-uniformity is essential in a variety of problems involving outliers, for example network monitoring problems with anomalous hosts, or recommendation problems with popular items or highly active users. This is a setting where passive algorithms fail, as we will show.

We make the following contributions:

1. For the matrix completion problem, we give a simple algorithm that exactly recovers an $n \times n$ rank r matrix using at most $O(nr\mu_0 \log^2(r))$ measurements where μ_0 is the coherence parameter on the column space of the matrix. This algorithm outperforms all existing results on matrix completion both in terms of sample complexity and in the fact that we place no assumptions on the row space of the matrix. The algorithm is extremely simple, runs in $O(nr^2)$ time, and can be implemented in one pass over the columns of the matrix.

2. We complement this sufficient condition with a lower bound showing that in the absence of row-space incoherence, *any* passive scheme must see $\Omega(n^2)$ entries. This concretely demonstrates the power of adaptivity in the matrix completion problem.

3. For matrix approximation, we analyze an algorithm that, after an adaptive sampling phase, approximates the input matrix by the top r ranks of an appropriately rescaled zero-filled version of the matrix. We show that with just $O(nr\mu \log^2(n))$ samples, this approximation is competitive with the best rank r approximation of the matrix. Here μ is a coherence parameter on each column of the matrix; as before we make no assumptions about the row space of the input. Again, this result significantly outperforms existing results on matrix approximation from passively collected samples.

This paper is organized as follows: we conclude this introduction with some basic definitions and then turn to related work in Section 2. The main results, consequences, and more detailed comparisons to existing work are given in Section 3. In Section 4, we provide some simulation that validate our theoretical results. We provide proofs in Section 5 and conclude with some future directions in Section 6. Some details are deferred to the appendix.

1.1 Preliminaries

Before proceeding, let us set up some notation used throughout the paper. We are interested in recovering, or approximating, a $d \times n$ matrix X from a set of at most M observations (We assume $d \leq n$.). We denote the columns of X by $x_1, \ldots, x_n \in \mathbb{R}^d$ and use t to index the columns. We use $x_t(i)$ to denote the ith coordinate of the column vector x_t.

We will frequently work with the truncated singular value decomposition (SVD) of X which is given by zero-ing out its smaller singular values. Specifically, write $X = U_r \Sigma_r V_r^T + U_{-r} \Sigma_{-r} V_{-r}^T$ where $[U_r, U_{-r}]$ (respectively $[V_r, V_{-r}]$) forms an orthonormal matrix and $\Sigma_r = \text{diag}(\sigma_1, \ldots, \sigma_r), \Sigma_{-r} = \text{diag}(\sigma_{r+1}, \ldots, \sigma_d)$ are diagonal matrices with $\sigma_1 \geq \ldots \geq \sigma_r \geq \sigma_{r+1} \geq \ldots \geq \sigma_d$. The truncated singular value decomposition is $X_r = U_r \Sigma_r V_r^T$, which is the best rank-r approximation to X both in Frobenius and spectral norm.

We also use capital letters to denote subspaces. We often overload notation by using the same symbol to refer to a subspace and any orthonormal basis for that subspace. Specifically, if $U \subset \mathbb{R}^d$ is a subspace with dimension r, we may sometimes use U to refer to a $d \times r$ matrix whose columns are an orthonormal basis for that subspace and vice versa. We use $U^\perp$ to denote the orthogonal complement to the subspace U and $\mathcal{P}_U$ to refer to the orthogonal projection operator onto U.

As we are dealing with missing data and sampling, we also need some notation for subsampling operations. Let $[d]$ denote the set $\{1, \ldots, d\}$ and let Ω be a list of m values from $[d]$, possibly with duplicates (One can think of Ω as a vector in $[d]^m$ and $\Omega(j)$ is the jth coordinate of this vector). If $x \in \mathbb{R}^d$, then $x_\Omega \in \mathbb{R}^m$ is the vector formed putting $x(i)$ in the jth coordinate if $\Omega(j) = i$ and $\mathcal{R}_\Omega x$ is a zero-filled rescaled version of x with $\mathcal{R}_\Omega x(i) = 0$ if $i \notin \Omega$ and $\mathcal{R}_\Omega x(i) = dx(i)/|\Omega|$ if $i \in \Omega$. In other words, $\mathcal{R}_\Omega$ is a $d \times d$ diagonal matrix with the (i,i)th entry equal to $d/|\Omega|$ if $i \in \Omega$ and zero otherwise.

For a r-dimensional subspace $U \subset \mathbb{R}^d$, $U_\Omega \in \mathbb{R}^{m \times r}$ is a *matrix* formed by doing a similar subsampling operation to the *rows* of any orthonormal basis for the subspace U, e.g. the jth row of U_Ω is the ith row of U if $\Omega(j) = i$. Note that U_Ω, and even the span of the columns of U_Ω, may not be uniquely defined, as they both depend on the choice of basis for U. Nevertheless, we will use $\mathcal{P}_{U_\Omega}$ to denote the projection onto the span of any single set of columns constructed by this subsampling operation.

In the matrix completion problem, where we aim for exact recovery, we require that X has rank at most r, meaning that $\sigma_{r+1} = \ldots = \sigma_n = 0$. Thus $X = X_r$, and our goal is to recover X_r exactly from a subset of entries. Specifically, we focus on the $0/1$ loss; given an estimator $\hat{X}$ for X, we would like to bound the probability of error:

$$R_{01}(\hat{X}) \triangleq \mathbb{P}\left(\hat{X} \neq X\right). \tag{1}$$

In the approximation problem, we relax the low rank assumption but are only interested in approximating the action of X_r. The goal is to find a rank r matrix $\hat{X}$ that minimizes:

$$R(\hat{X}) = \|X - \hat{X}\|_F.$$

The matrix X_r is the global minimizer (subject to the rank-r constraint), and our task is to approximate this low rank matrix effectively. Specifically, we will be interested in finding matrices $\hat{X}$ that satisfy excess risk bounds of the form:

$$R(\hat{X}) \triangleq \|X - \hat{X}\|_F \leq \|X - X_r\|_F + \epsilon \|X\|_F \tag{2}$$

Rescaling the excess risk term by $\|X\|_F$ is a form of normalization that has been used before in the matrix approximation literature [Frieze et al., 2004, Drineas et al., 2006a,b, Rudelson and Vershynin, 2007]. While

bounds of the form $(1+\epsilon)\|X - X_r\|_F$ may seem more appropriate when the bottom ranks are viewed as noise term, achieving such a bound seems to require highly accurate approximations of the SVD of the input matrix [Drineas et al., 2008], which is not possible given the extremely limited number of observations in our setting. Equation 2 can be interpreted by dividing by $\|X\|_F$, which shows that $\hat{X}$ captures almost as large a fraction of the energy of X as X_r does.

Apart from the observation budget M and the approximation rank r, the other main quantity governing the difficulty of these problems is the subspace coherence parameter. For a r dimensional subspace U of $\mathbb{R}^d$, define

$$\mu(U) = \frac{d}{r} \max_{i \in [d]} \|\mathcal{P}_U e_i\|_2^2,$$

which is a standard measure of subspace coherence [Recht, 2011]). The quantity $\mu_0 \triangleq \mu(U_r)$, which is bounded between 1 and d/r, measures how correlated the principal column space of the matrix X is with any single standard basis element. When this maximal correlation is small, the energy of the matrix is spread out fairly uniformly across the rows of the matrix, although it can be non-uniformly distributed across the columns. Without loss of generality we use the matrix column-space coherence μ_0 instead of the row-space analog, and we will see that the parameter μ_0 controls the sample complexity of our adaptive procedure.

In classical results on matrix completion, the parameter $\mu_0' \triangleq \max\{\mu(U_r), \mu(V_r)\}$ instead governs the sample complexity. When μ_0' is small, both principal subspaces are incoherent, so that the energy of the matrix is uniformly spread across the entries. Informally, this means that a random sample of entries captures the salient features of the matrix, and, indeed, the number of uniform-at-random samples necessary and sufficient for exact recovery scales linearly with μ_0' [Recht, 2011].

Such an incoherence assumption does not translate appropriately to the approximate recovery problem, since the matrix is no longer low rank, but some measure of uniformity is still necessary. One one hand, the statistics literature typically assumes that the matrix X can be decomposed into an incoherent low rank matrix and a stochastic perturbation [Negahban and Wainwright, 2012, Koltchinskii et al., 2011]. On the other hand, classical results on matrix approximation make no stochastic assumptions, but also do not need uniformity, as they do not consider the missing data setting [Frieze et al., 2004, Drineas et al., 2006a]. As we aim to bridge these two lines of research by considering matrix approximation with missing data, we remove the stochastic assumption. We instead turn to an alternative assumption to ensure that the high ranks of the matrix are well-behaved under sampling.

We parameterize the problem by a quantity related to the usual definition of incoherence:

$$\mu = \max_{t \in [n]} \frac{d||x_t||_\infty^2}{||x||_2^2},$$

which is the maximal column coherence. Here, we make no stochastic assumptions, but notice that this is a restriction on the higher ranks of the matrix. We also make no assumptions about the row space of the matrix[1].

2 Related Work

The literature on low rank matrix approximation is extremely vast and we do not attempt to cover all of the existing ideas. Instead, we focus on the most relevant lines of work to our specific problems. We briefly mention some related work on adaptive sensing.

[1] As before this could equivalently be the column space with assumption on the maximal row coherence. Without loss of generality we parametrize columns and column spaces throughout this paper.

For matrix completion, a series of papers provide better and better analyses of the nuclear norm minimization procedure, finally showing that $M \asymp (n+d)r\mu_0' \log(n)$ uniform-at-random observations are sufficient to exactly recover a rank r matrix with high probability [Candès and Tao, 2010, Candès and Recht, 2009, Gross, 2011, Recht, 2011, Chen, 2013]. These results involve the parameter μ_0', implying that both matrix subspaces must be incoherent for strong guarantees. Recently, two papers have relaxed the row-space incoherence assumption with adaptive sampling [Krishnamurthy and Singh, 2013, Chen et al., 2014]. Our analysis leads to a better sample complexity than both of these results. We defer a more detailed discussion to after Theorem 1.

A number of authors have studied matrix completion with noise and under weaker assumptions. The most prominent difference between our work and all of these is a relaxation of the main incoherence assumptions. Both Candès and Plan [2010], and Keshavan et al. [2010] require that both the row and column space of the matrix of interest is highly incoherent. Negahban and Wainwright [2012] instead use a notion of *spikiness*, but that too places assumptions on the row space of interest. Koltchinskii et al. [2011] consider matrices with bounded entries, which is related to the spikiness assumption. In comparison, our results make essentially no assumptions about the row space, leading to substantially more generality. This is the thesis of our work; one can eliminate row space assumptions in matrix recovery problems through adaptive sampling.

Another close line of work is on matrix sparsification [Achlioptas and Mcsherry, 2007, Arora et al., 2006, Achlioptas et al., 2013]. Here, the goal is to zero out a large number of entries of a given matrix while preserving global properties such as the principal subspace. The main difference from the matrix completion literature is that the entire matrix is observed, and this allows one to relax incoherence assumptions. The only result from this literature that does not require knowledge of the matrix is a random sampling scheme of Achlioptas and Mcsherry [2007], but it is only competitive with matrix completion results when the input matrix has entries of fairly constant magnitude Koltchinskii et al. [2011]. Interestingly, this requirement is essentially the same as the spikiness assumption of Negahban and Wainwright [2012] and the bounded magnitude assumption of Koltchinskii et al. [2011].

Several techniques have been proposed for matrix approximation in the fully observed setting, optimizing computational complexity or other objectives. A particularly relevant series of papers is on the column subset selection (CSS) problem, where the span of several judiciously chosen columns is used as to approximate the principal subspace. One of the best approaches involves sampling columns according to the statistical leverage scores, which are the norms of the rows of the $n \times r$ matrix formed by the top r right singular vectors [Boutsidis et al., 2009, 2011, Drineas et al., 2008]. Unfortunately, this strategy does not seem to apply in the missing data setting, as the distribution used to sample columns – which are subsequently used to approximate the matrix – depends on the unobserved input matrix. Approximating this distribution seems to require a very accurate estimate of the matrix itself, and this initial estimate would suffice for the matrix approximation problem. This difficulty also arises with volume sampling [Guruswami and Sinop, 2012], another popular approach to CSS; the sampling distribution depends on the input matrix and we are not aware of strategies for approximating this distribution in the missing data setting.

In terms of adaptive sampling, a number of methods for recovery of sparse, possibly structured, signals have been shown to outperform passive methods [Haupt et al., 2011, Malloy and Nowak, 2011, Tánczos and Castro, 2013, Balakrishnan et al., 2012, Krishnamurthy et al., 2013]. While having their share of differences, these methods can all be viewed as either binary search or local search methods, that iteratively discard irrelevant coordinates and focus measurements on the remainder. In particular, these methods rely heavily on the sparsity and structure of the input signal, and extensions to other settings have been elusive. While a low rank matrix is sparse in its eigenbasis, the search-style techniques from the signal processing community do not seem to leverage this structure effectively and these approaches do not appear to be applicable to our setting.

Algorithm 1: Adaptive Matrix Completion ($X \in \mathbb{R}^{d \times n}, m$)

1. Let $U = \emptyset$.
2. Randomly draw entries $\Omega \subset [d]$ of size m uniformly with replacement.
3. For each column x_t of X ($t \in [N]$):
   (a) If $||x_{t\Omega} - \mathcal{P}_{U_\Omega} x_{t\Omega}||_2^2 > 0$:
      i. Fully observe x_t and add to U (orthogonalize U).
      ii. Randomly draw a new set Ω of size m uniformly with replacement.
   (b) Otherwise $\hat{x}_t \leftarrow U(U_\Omega^T U_\Omega)^{-1} U_\Omega x_{t\Omega}$.
4. Return $\hat{X}$ with columns $\hat{x}_t$.

Some of these adaptive sampling efforts focus specifically on recovering or approximating highly structured matrices, which is closely related to our setting. Tánczos and Castro [2013] and Balakrishnan et al. [2012] consider variants of biclustering, which is equivalent to recovering a rank-one binary matrix from noisy observations. Singh et al. [2012] recover noisy ultrametric matrices while Krishnamurthy et al. [2012] uses a similar idea to find hierarchical clustering from adaptively sampled similarities. All of these results can be viewed as matrix completion or approximation, but impose significantly more structure on the target matrix than we do here. For this reason, many of these algorithmic ideas also do not appear to be useful in our setting.

3 Results

In this section we develop the main theoretical contributions of this manuscript. We first turn to the matrix completion problem, where we improve the results of Krishnamurthy and Singh [2013] and show that $O(dr + nr\mu_0 \log^2 r)$ samples suffice to recover a rank r matrix whose column space has coherence bounded by μ_0. We complement this result with some necessary conditions on passive and adaptive matrix completion algorithms. Then, we turn to the low rank approximation problem, where we describe a simple algorithm and show that it achieve the excess risk bound in Equation 2 with $O(nr\mu \log^2(n)/\epsilon^4)$ samples. We also provide a detailed comparison of this result with prior work.

3.1 Matrix Completion

Our algorithm for the matrix completion problem is identical to the algorithm of Krishnamurthy and Singh [2013]. The procedure, whose pseudocode is displayed in Algorithm 1, streams the columns of the matrix X into memory and iteratively adds directions to an estimate for the column space of X. The algorithm maintains a subspace U and, when processing the tth column x_t, estimates the norm of $\mathcal{P}_{U^\perp} x_t$ using only a few entries of x_t. We will ensure that, with high probability, this estimate will be non-zero if and only if x_t contains a new direction. If the estimate is non-zero, the algorithm asks for the remaining entries of x_t and adds the new direction to the subspace U. Otherwise, x_t lies in U and we will see that the algorithm already has sufficient information to complete the column x_t.

Therefore, the key ingredient of the algorithm is the estimator for the projection onto the orthogonal complement of the subspace U. This quantity is estimated as follows. Using a list of m locations Ω from

$[d]$, we downsample both x_t and an orthonormal basis U to $x_{t\Omega}$ and U_Ω. We then use $\|x_{t\Omega} - \mathcal{P}_{U_\Omega} x_{t\Omega}\|^2$ as our estimate. It is easy to see that this estimator leads to a test with one-sided error, since the estimator is identically zero if $x_t \in U$. In our analysis, we establish a relative-error deviation bound, which allows us to control the error of our test for energy outside of U.

A subtle but critical aspect of the algorithm is the choice of Ω. The list Ω always has m elements, and each element is sampled uniformly with replacement from $[d]$. More importantly, we only resample Ω when we add a direction to U. This ensures that the algorithm does not employ too much randomness, which would lead to an undesireable logarithmic dependence on n.

The analysis of this test statistic and the reconstruction procedure leads to the following guarantee on the performance of the algorithm, whose proof is deferred to Section 5.

Theorem 1. *Let $X \in \mathbb{R}^{d\times n}$ be a matrix of rank r whose column space U has coherence $\mu(U) \le \mu_0$. Then the output of Algorithm 1 has risk:*

$$R_{01}(\hat{X}) \le 10r^2 \exp\left\{-\sqrt{\frac{m}{32r\mu_0}}\right\} \tag{3}$$

provided that $m \ge 4r\mu_0 \log(2r/\delta)$. Equivalently, whenever $m \ge 32r\mu_0 \log^2(10r^2/\delta)$, we have $R_{01}(\hat{X}) \le \delta$. The sample complexity is $dr + nm$ and the running time is $O(nmr + r^3 m + dr^2)$.

To the best of our knowledge, this result provides the strongest guarantee for the matrix completion problem. The vast majority of results require both incoherent row and column spaces and are therefore considerably more restrictive than ours [Candès and Tao, 2010, Candès and Recht, 2009, Gross, 2011, Recht, 2011, Chen, 2013]. For example, Recht shows that by solving the nuclear norm minimization program, one can recover X exactly, provided that the number of measurements exceeds $32(d+n)r \max\{\mu_0', \mu_1^2\} \log^2(n)$ where recall that μ_0' upper bounds the coherence of both the row and column space, and μ_1 provides another incoherence-type assumption (which can be removed [Chen, 2013]). Our result improves on his not only in relaxing the row space incoherence assumption, but also in terms of sample complexity, as we remove the logarithmic dependence on problem dimension.

As another example, Gittens [2011] showed that Nystrom method can recover a rank r matrix from randomly sampling $O(r \log r)$ columns. While his result matches ours in terms of sample complexity, he analyzes positive-semidefinite matrices with incoherent principal subspace, which translates to assuming that both row and column spaces are incoherent. Again, in relaxing this assumption, our result is substantially more general.

We mention two papers that allow coherent row spaces. The first is the paper of Krishnamurthy and Singh [2013], that gives a weaker analysis of Algorithm 1 resulting in a polynomially worse dependence on r. The other is the two-phase algorithm of Chen et al. [2014] based on local coherence sampling. Their algorithm requires $O((n + d)r\mu_0 \log(n))$ samples which is weaker than our guarantee in that it has a slightly super-linear dependence on problem dimension. An interesting consequence of Theorem 1 is that the amortized number of samples per column is *completely independent* of the problem dimension.

Regarding computational considerations, the algorithm operates in one pass over the columns, and need only store the matrix in condensed form, which requires $O((n + d)r)$ space. Specifically, the algorithm maintains a (partial) basis for column space and the coefficients for representing each column by that basis, which leads to an optimally condensed representation. Moreover, the computational complexity of the algorithm is *linear* in the matrix dimensions d, n with mild polynomial dependence on the rank r. For this run-time analysis, we work in a computational model where accessing any entry of the matrix is a constant-time operation, which allows us to circumvent the $\Omega(dn)$ time it would otherwise take to read the input. In

comparison, the two standard algorithms for matrix completion, the iterative Singular Value Thresholding Algorithm [Cai et al., 2010] and alternating least-squares [Jain et al., 2013, Hardt, 2013], are significantly slower than Algorithm 1, not only due to their iterative nature, but also in per-iteration running time.

3.2 Necessary Conditions for Matrix Completion

We now establish a lower bound on any passive sampling algorithm for the matrix completion problem. Our lower bound shows that if the matrix has coherent row space, then any passive sampling scheme followed by any recovery algorithm requires $\sim dn$ samples.

To formalize our lower bound we fix a sampling budget M and consider an estimator to be a sampling distribution q over $\{(i,j)|i \in [d], j \in [n]\}^M$ and a (possibly randomized) function $f : \{(\Omega, X_\Omega)\} \to \mathbb{R}^{d\times n}$ that maps a set of indices and values to a $d \times n$ matrix. Let $\mathcal{Q}$ denote the set of all such sampling distributions and let $\mathcal{F}$ denote the set of all such estimators. Lastly let $\mathcal{X}$ denote the set of all $d \times n$ rank r matrices with column incoherence at most μ_0. We consider the minimax probability of error:

$$R^\star = \inf_{f\in\mathcal{F}} \inf_{q\in\mathcal{Q}} \sup_{X\in\mathcal{X}} \mathbb{P}_{\Omega\sim q}\left[f(\Omega, X_\Omega \neq X\right]$$

where the probability also accounts for potential randomness in the estimator f. Note that since we make no assumptions about the distribution q other than excluding adaptive distributions, this setup subsumes essentially all passive sampling strategies including uniform-at-random, deterministic, and distributions sampling entire columns. The one exception is the bernoulli sampling model, where each entry (i,j) is observed with probability q_{ij} independently of all other entries, although we believe a similar lower bound holds there.

The following theorem lower bounds success probability of any passive strategy and consequently gives a necessary condition on the sample complexity.

Theorem 2. *The minimax risk $R^\star$ satisfies:*

$$R^\star \geq \frac{1}{2} - \left\lceil \frac{m}{(1-\frac{r-1}{r\mu_0})d} \right\rceil \frac{1}{2(n-r)}, \tag{4}$$

which approaches $1/2$ whenever:

$$m = o\left((dn-dr)(1+\frac{1}{r\mu_0}-\frac{1}{\mu_0})\right). \tag{5}$$

As a concrete instantiation of the theorem, if μ_0 is bounded from below by any constant $c > 1$ (which is possible whenever $r \leq d/c$), then the bound approaches $1/2$ whenever $m = o(d(n-r))$. Thus all passive algorithms must have sample complexity that is quadratic in the problem dimension. In contrast, Theorem 1 ensures that Algorithm 1 has nearly linear sample complexity, which is a significant improvement over passive algorithms.

The literature contains several other necessary conditions on the sample complexity for matrix completion. A fairly simple argument shows that without any form of incoherence, one requires $\Omega(dn)$ samples to recover even a rank one matrix that is non-zero in just one entry. This argument applies to both passive and adaptive sampling strategies and shows that some measure of incoherence is necessary. With both row and column incoherence, but still under uniform sampling, Candès and Tao [2010] prove that $\Omega(\mu_0' nr\log(n))$ observations are necessary to recover a $n \times n$ matrix.

One can relax the incoherence assumption by non-uniform passive sampling, although the sampling distribution is matrix-specific as it depends on the local coherence structure [Chen et al., 2014]. Unfortunately,

Algorithm 2: Low Rank Approximation (X, m_1, m_2)

1. Pass 1: For each column, observe Ω_t of size m_1 uniformly at random with replacement and estimate $\hat{c}_t = \frac{d}{m_1}||x_{t,\Omega_t}||_2^2$. Estimate $\hat{f} = \sum_t \hat{c}_t$.
2. Pass 2: Set $\tilde{X} = 0 \in \mathbb{R}^{d\times n}$.
   (a) For each column x_t, sample $m_{2,t} = m_2 n\hat{c}_2/\hat{f}$ observations $\Omega_{2,t}$ uniformly at random with replacement.
   (b) Update $\tilde{X} = \tilde{X} + (\mathcal{R}_{\Omega_{2,t}} x_t)e_t^T$.
3. Compute the SVD of $\tilde{X}$ and output $\hat{X}$ which is formed by the top-r ranks of $\tilde{X}$.

one cannot compute the appropriate sampling distribution, before taking any measurements. Our result shows that in the absence of row-space incoherence, there is no universal passive sampling scheme that can achieve a non-trivial sample complexity. Thus adaptive sampling is necessary to relax the incoherence assumption while retaining near-optimal sample complexity.

Finally, a parameter counting argument shows that even adaptive sampling requires $\Omega((d+n)r)$ samples. Each entry of a rank r matrix can be expressed as a polynomial of the left and right singular vectors and the singular values, so the observations lead to a polynomial system in $(d+n)r + r$ variables. If $M < (d+n)r - r^2$ (there are $r(r+1)$ orthonormality constraints), then this system is underdetermined, and since it has one solution, it must have infinitely many, so that recovery is impossible. Consequently, our algorithm is nearly optimal, and significantly outperforms any passive sampling strategy.

3.3 Low Rank Approximation

For the matrix approximation problem, we propose an adaptive sampling algorithm to obtain a low-rank approximation to X. The algorithm (see Algorithm 2 for pseudocode) makes two passes through the columns of the matrix. In the first pass, it subsamples each column uniformly at random and estimates each column norm and the matrix Frobenius norm. In the second pass, the algorithm samples additional observations from each column, and for each t, places the rescaled zero-filled vector $\mathcal{R}_{\Omega_{2,t}} x_t$ into the tth column of a new matrix $\tilde{X}$, which is a preliminary estimate of the input, X. Once the initial estimate $\tilde{X}$ is computed, the algorithm zeros out all but the top r ranks of $\tilde{X}$ to form $\hat{X}$. We will show that $\hat{X}$ has low excess risk, when compared with the best rank-r approximation, X_r.

A crucial feature of the second pass is that the number of samples per column is proportional to the squared norm of that column. Of course this sampling strategy is only possible if the column norms are known, motivating the first pass of the algorithm, where we estimate precisely this sampling distribution. This feature allows the algorithm to tolerate highly non-uniform column norms, as it focuses measurements on high-energy columns, and leads to significantly better approximation. This idea has been used before, although only in the exactly low-rank case [Chen et al., 2014].

For the main performance guarantee, we only assume that the matrix has incoherent columns, that is $d\|x_t\|_\infty^2/\|x_t\|_2^2 \le \mu$ for each column x_t. In particular we make no additional assumptions about the high-rank structure of the matrix. We have the following theorem:

Theorem 3. *Set $m_1 \ge 32\mu\log(n/\delta)$ and assume $n \ge d$ and that X has μ-incoherent columns. With*

probability $\geq 1 - 2\delta$*, Algorithm 2 computes an approximation* $\hat{X}$ *such that:*

$$\|X - \hat{X}\|_F \leq \|X - X_r\|_F + \|X\|_F \left(6\sqrt{\frac{r\mu}{m_2}} \log\left(\frac{d+n}{\delta}\right) + \left(6\sqrt{\frac{r\mu}{m_2}} \log\left(\frac{d+n}{\delta}\right)\right)^{1/2}\right)$$

using $n(m_1 + m_2)$ *samples. In other words, the output* $\hat{X}$ *satisfies* $\|X - \hat{X}\|_F \leq \|X - X_r\|_F + \epsilon\|X\|_F$ *with probability* $\geq 1 - 2\delta$ *and with sample complexity:*

$$32n\mu \log(n/\delta) + \frac{576}{\epsilon^4} nr\mu \log^2\left(\frac{d+n}{\delta}\right). \tag{6}$$

The proof is deferred to Section 5. The theorem shows that the matrix $\hat{X}$ serves as nearly as good an approximation to X as X_r. Specifically, with $O(nr\mu \log^2(d+n))$ observations, one can compute a suitable approximation to X. The running time of the algorithm is dominated by the cost of computing the truncated SVD, which is at most $O(d^2 n)$.

While the dependence between the number of samples and the problem parameters n, r, and μ is quite mild and matches existing matrix completion results, the dependence on the error ϵ in Equation 6 seems undesirable. This dependence arises from our translation of a bound on $\|\tilde{X} - X\|_2$ into a bound on $\|\hat{X} - X\|_F$, which results in the $m_2^{-1/4}$-dependence in the error bound. We are not aware of better results in the general setting, but a number of tighter translations are possible under various assumptions, and these can result in better guarantees. We mention just two such results here.

Proposition 4. *Under the same assumptions as Theorem 3, suppose further that* X *has rank at most* r*. Then with probability* $\geq 1 - 2\delta$*:*

$$\|X - \hat{X}\|_F \leq 20\|X\|_F \sqrt{\frac{r\mu}{m_2}} \log\left(\frac{d+n}{\delta}\right)$$

This proposition tempers the dependence on the error ϵ from $1/\epsilon^4$ to $1/\epsilon^2$ in the event that the input matrix has rank at most r. This gives a relative error guarantee for Algorithm 2 on the matrix completion problem, which improves on the one implied by Theorem 3. Note that this guarantee is weaker than Theorem 1, but Algorithm 2 is much more robust to relaxations of the low rank assumption as demonstrated in Theorem 3.

A similarly mild dependence on ϵ can be derived under the assumption that $X = A + R$, A has rank r and R is some perturbation, which has the flavor of existing noisy matrix completion results. Here, it is natural to recover the parameter A rather than the top r ranks of X and we have the following parameter recovery guarantee for Algorithm 2:

Proposition 5. *Let* $X = A + R$ *where* A *has rank at most* r*. Suppose further that* X *has* μ*-incoherent columns and set* $m_1 \geq 32\mu \log(n/\delta)$*. Then with probability* $\geq 1 - 2\delta$*:*

$$\|\hat{X} - A\|_F \leq 20\sqrt{\frac{r\mu}{m_2}} \log\left(\frac{d+n}{\delta}\right) (\|A\|_F + \|R_\Omega\|_F) + \sqrt{8r}\|R\|_2 \tag{7}$$

where the number of samples is $n(m_1 + m_2)$ *and* Ω *is the set of all entries observed over the course of the algorithm.*

To interpret this bound, let $\|A\|_F = 1$, and let R be a random matrix whose entries are independently drawn from a Gaussian distribution with variance $\sigma^2/(dn)$. Note that this normalization for the variance is appropriate in the high-dimensional setting where $n, d \to \infty$, since we keep the signal-to-noise ratio $\|A\|_F^2/\|R\|_F^2 = 1/\sigma^2$ constant. In this setting, the last term can essentially be ignored, since by the standard bound on the spectral norm of a Gaussian matrix, $\|R\|_2 = O(\sigma\sqrt{\frac{1}{d}}\log((n+d)/\delta))$ which will be lower order [Achlioptas and Mcsherry, 2007]. We can also bound $\|R_\Omega\|_F \leq O(\sigma\sqrt{\frac{m_1+m_2}{d}}\log((n+d)/\delta))$ using a Gaussian tail bound. With $m_1 \leq m_2$ we arrive at:

$$\|\hat{X} - A\|_F \leq c_\star \left(\sqrt{\frac{r\mu}{m_2}} + \sigma\sqrt{\frac{r\mu}{d}}\right)\log^2\left(\frac{d+n}{\delta}\right),$$

where $c_\star$ is some positive constant. In the high dimensional setting, when $r\mu = \tilde{o}(d)$, this shows that Algorithm 2 consistently recovers A as long as $m_2 = \tilde{\omega}(r\mu)$. This second condition implies that the total number of samples uses is $\tilde{\omega}(nr\mu)$[2].

3.3.1 Comparison with Matrix Completion Results

The closest result to Theorem 3 is the result of Koltchinskii et al. [2011] who consider a soft-thresholding procedure and bound the approximation error in squared-Frobenius norm. They assume that the matrix has bounded entrywise ℓ_∞ norm and give an entrywise squared-error guarantee of the form:

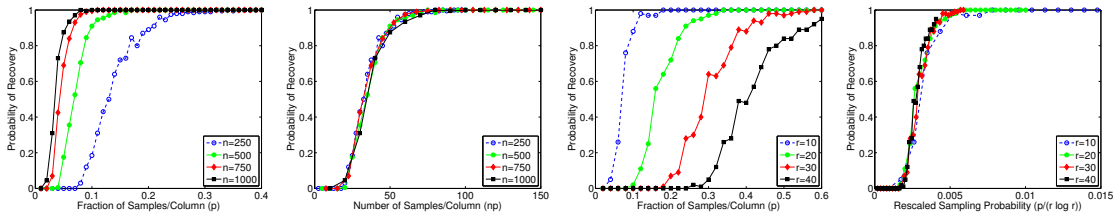
$$\|\hat{X} - X\|_F^2 \leq \|X - X_r\|_F^2 + cdn\|X\|_\infty^2 \frac{nr\log(d+n)}{M} \tag{8}$$

where M is the total number of samples and c is a constant. Their bound is quite similar to ours in the relationship between the number of samples and the target rank r. However, since $dn\|X\|_\infty^2 \geq \|X\|_F^2$, their bound is significantly worse in the event that the energy of the matrix is concentrated on a few columns.

To make this concrete, fix $\|X\|_F = 1$ and let us compare the matrix where every entry is $\frac{1}{\sqrt{dn}}$ with the matrix where one column has all entries equal to $\frac{1}{\sqrt{d}}$. In the former, the error term in the squared-Frobenius error bound of Koltchinskii et al. is $nr\log(d+n)/M$ while our bound on Frobenius error is, modulo logarithmic factors, the square root of this quantity. In this example, the two results are essentially equivalent. For the second matrix, their bound deteriorates significantly to $n^2 r\log(d+n)/M$ while our bound remains the same. Thus our algorithm is particularly suited to handle matrices with non-uniform column norms.

Apart from adaptive sampling, the difference between our procedure and the algorithm of Koltchinskii et al. [2011] is a matter of soft- versus hard-thresholding of the singular values of the zero-filled matrix. In the setting of Proposition 5, soft thresholding seems more appropriate, as the choice of regularization parameter allows one to trade off the amount of signal and noise captured in $\hat{X}$. While in practice one could replace the hard thresholding step with soft thresholding in our algorithm, there are some caveats with the theoretical analysis. First, soft-thresholding does not ensure that $\hat{X}$ will be at most rank r, so it is not suitable for the matrix approximation problem. Second, the resulting error guarantee depends on the sampling distribution, which cannot be translated to the Frobenius norm unless the distribution is quite uniform [Negahban and Wainwright, 2012, Koltchinskii et al., 2011]. Thus the soft-thresholding procedure does not give a Frobenius-norm error guarantee in the non-uniform setting that we are most interested in.

[2]The notation $\tilde{o}(\cdot), \tilde{\omega}(\cdot)$ is the Bachmann-Landau asymptotic notation but suppressing logarithmic factors.



(a) (b) (c) (d)

Figure 1: (a): Probability of success of Algorithm 1 versus fraction of samples per column ($p = m/d$) with $r = 10, \mu_0 = 1$. (b): Data from (a) plotted against samples per column, m. (c): Probability of success of Algorithm 1 versus fraction of samples per column ($p = m/d$) with $n = 500, \mu_0 = 1$. (d): Data from (c) plotted against rescaled sample probability $p/(r \log r)$.

The majority of other results on low rank matrix completion focus on parameter recovery rather than approximation [Negahban and Wainwright, 2012, Keshavan et al., 2010, Candès and Plan, 2010]. It is therefore best to compare with Proposition 5, where we show that Algorithm 2 consistently recovers the parameter, A. These results exhibit similar dependence between the number of samples and the problem parameters n, r, ϵ but hold under different notions of uniformity, such as spikiness, boundedness, or incoherence. Our result agrees with these existing results but holds under a much weaker notion of uniformity.

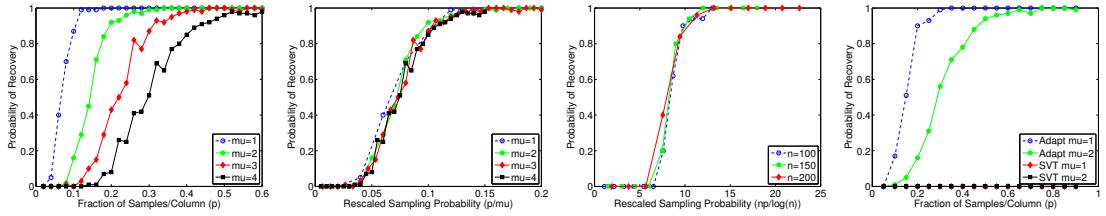
Lastly, we emphasize the effect of adaptive sampling in our bound. We do not need *any* uniformity assumption over the columns of the input matrix X. All existing works on noisy low rank matrix completion or matrix approximation from missing data have some assumption of this form, be it incoherence [Keshavan et al., 2010, Candès and Plan, 2010], spikiness [Negahban and Wainwright, 2012], or bounded ℓ_∞ norm [Koltchinskii et al., 2011]. The detailed comparison with the result of Koltchinskii et al. gives a precise characterization of this effect and shows that in the absence of such uniformity, our adaptive sampling algorithm enjoys a significantly lower sample complexity.

In the event of uniformity, our algorithm performs similarly to existing ones. Specifically, we obtain the same relationship between the total number of samples M, the problem dimensions n, d and the target rank r. If we knew *a priori* that the matrix had near-uniform column lengths, we could simply omit the first pass of the algorithm, sample uniformly in the second pass and avoid the need for any adaptivity.

4 Simulations

We perform a number of simulations to analyze the empirical performance of both Algorithms 1 and 2. The first set of simulations, in Figures 1 and 2, examine the behavior of Algorithm 1. We work with square matrices where the column space is spanned by binary vectors, constructed so that the matrix has the appropriate rank and coherence. The row space is spanned by either random gaussian vectors in the case of incoherent row space or a random collection of standard basis elements if we want high coherence.

In the first two figures (1(a) and 1(b)) we study the algorithms dependence on the matrix dimension. For various matrix sizes, we record the probability of exact recovery as we vary the number of samples alloted to the algorithm. We plot the probability of recovery as a function of the fraction of samples per column, denote by p, (Figure 1(a)) and as a function of the total samples per column m (Figure 1(b)). It is clear from the simulations that p can decrease with matrix dimension while still ensuring exact recovery. On the other hand, the curves in the second figure line up, demonstrating that the number of samples per column remains



(a) (b) (c) (d)

Figure 2: (a): Probability of success of Algorithm 1 versus fraction of samples per column ($p = m/d$) with $n = 500, r = 10$. (b): Data from (a) plotted against rescaled sampling probability p/μ_0. (c): Probability of success of SVT versus rescaled sampling probability $np/\log(n)$ with $r = 5, \mu_0 = 1$. (d): Probability of sucess of Algorithm 1 and SVT versus sampling probability for matrices with highly coherent row space with $r = 5, n = 100$.

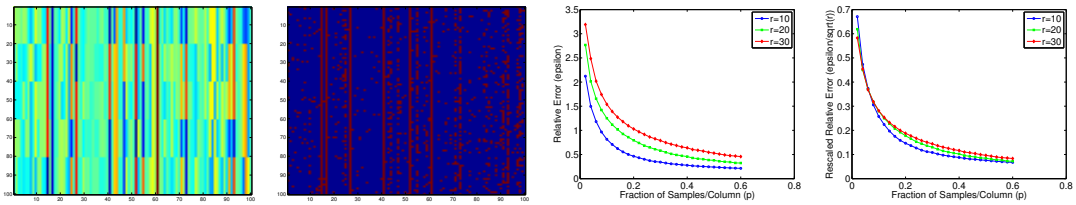
fixed for fixed probability of recovery. This behavior is predicted by Theorem 1, which shows that the total number of samples scales linearly with dimension, so that the number of samples per column remains constant.

In Figures 1(c) and 1(d) we show the results of a similar simulation, instead varying the matrix rank r, with dimension fixed at 500. The first figure shows that the fraction of samples per column must increase with rank to ensure successful recovery while second shows that the ratio $p/(r \log r)$ governs the probability of success. Figures 2(a) and 2(b) similarly confirm a linear dependence between the incoherence parameter μ_0 and the sample complexity. Notice that the empirical dependence on rank is actually a better than what is predicted by Theorem 1, which suggests that $r \log^2 r$ is the appropriate scaling. Our theorem does seem to capture the correct dependence on the coherence parameter.

In the last two plots we compare Algorithm 1 against the Singular Value Thresholding algorithm (SVT) of Cai et al. [2010]. The SVT algorithm is a non-adaptive iterative algorithm for nuclear norm minimization from a set of uniform-at-random observations. In Figure 2(c), we show that the success probability is governed by $np/\log(n)$, which is predicted by the existing analysis of the nuclear norm minimization program. This dependence is worse than for Algorithm 1, whose success probability is governed by np as demonstrated in Figure 1(b). Finally, in Figure 2(d), we record success probability versus sample complexity on matrices with maximally coherent row spaces. The simulation shows that our algorithm can tolerate coherent row spaces while the SVT algorithm cannot.

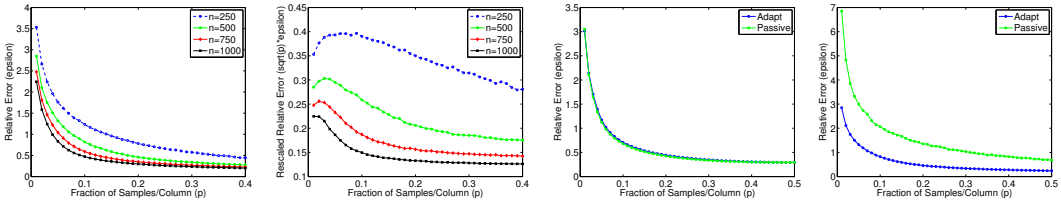
For Algorithm 2, we display the results of a similar set of simulations in Figures 3 and 4. Here, we construct low rank matrices whose column spaces are spanned by binary vectors and whose columns are also constant in magnitude on their support. The length of the columns is distributed either log-normally, resulting in non-uniform column lengths, or uniformly between 0.9 and 1.1. We then corrupt this low rank matrix by adding a gaussian matrix whose entries have variance $\frac{1}{dn}$. In Figure 3(a) we show a matrix constructed via this process and in Figure 3(b) we show the set of entries sampled by Algorithm 2 on this input. From the plots, it is clear that the algorithm focuses its measurements on the columns with high energy, while using very few samples to capture the columns with lower energy.

In Figure 3(c), we plot the relative error, which is the ϵ in Equation 2, as a function of the average fraction of samples per column (averaged over columns, as we are using non-uniform sampling) for 500×500 matrices of varying rank. In the next plot, Figure 3(d), we rescale the relative error by $\sqrt{r}$, to capture the dependence on rank predicted by Theorem 3. As we increase the number of observations, the relative error



(a) (b) (c) (d)

Figure 3: (a): An example matrix with with highly non-uniform colum norms and (b) the sampling pattern of Algorithm 2. (c): Relative error as a function of samping probability p for different target rank r ($\mu = 1$). (d): The same data where the y-axis is instead $\epsilon/\sqrt{r}$.



(a) (b) (c) (d)

Figure 4: (a): Relative error of Algorithm 2 as a function of sampling probability p for different size matrices with fixed target rank $r = 10$ and $\mu = 1$. (b): The same data where the y-axis is instead $\sqrt{p}\epsilon$. (c): Relative error for adaptive and passive sampling on matrices with uniform column lengths (column coherence $\mu = 1$ and column norms are uniform from $[0.9, 1.1]$). (c): Relative error for adaptive and passive sampling on matrices with highly nonuniform column lengths (column coherence $\mu = 1$ and column norms are from a standard Log-Normal distribution).

decreases quite rapidly. Moreover, the algorithm needs more observations as the target rank r increases. Qualitatively both of these effects are predicted by Theorem 3. Lastly, the fact that the curves in Figure 3(d) nearly line up suggests that the relative error ϵ does scale with $\sqrt{r}$.

In Figure 4(a), we plot the relative error as a function of the average fraction of samples, p, per column for different matrix sizes. We rescale this data by plotting the y-axis in terms of $\sqrt{p}\epsilon$ (Figure 4(b)). From the first plot, we see that the error quickly decays, while a smaller fraction of samples are needed for larger problems. In the second plot, we see that rescaling the error by $\sqrt{p}$ has the effect of flattening out all of the curves, which suggests that the relationship between ϵ and the number of samples is indeed $\epsilon\sqrt{p} \asymp 1$ or that $\epsilon \asymp \frac{1}{\sqrt{p}}$. This phenomenon is predicted by Proposition 5.

In the last set of simulations, we compare our algorithm with an algorithm that first performs uniform sampling and then hard thresholds the singular values to build a rank r approximation. In Figure 4(c), we use matrices with uniform column norms, and observe that both algorithms perform comparably. However, in Figure 4(d), when the column norms are highly non-uniform, we see that Algorithm 2 dramatically outperforms the passive sampling approach. This confirms our claim that adaptive sampling leads to better approximation when the energy of the matrix is not uniformly distributed.

5 Proofs

In this section we provide the proofs of our main theorems. We defer some concentration results and some details to the appendix.

5.1 Proof of Theorem 1

The proof is identical to the proof by Krishnamurthy and Singh [2013], with an improved concentration of measure argument. We reproduce some of the details here.

The main step in the proof analyzes of the test based on the projection $\|x_{t\Omega} - \mathcal{P}_{U_\Omega} x_{t\Omega}\|_2^2$. Using various versions of Bernstein's inequality we are able to prove the following theorem, which builds off of Balzano et al. [2010] and Krishnamurthy and Singh [2013].

Theorem 6. *Let U be an r-dimensional subspace of $\mathbb{R}^d$ and $y = x + v$ where $x \in U$ and $v \in U^\perp$. Fix $\delta > 0$ and $m \geq \max\{\frac{8}{3} r\mu(U)\log(2d/\delta), 4\mu(v)\log(1/\delta)\}$ and let Ω be an index set of m entries sampled uniformly with replacement from $[d]$. With probability $\geq 1 - 4\delta$:*

$$\frac{m(1-\alpha) - r\mu(U)\frac{\beta}{1-\gamma}}{d}\|v\|_2^2 \leq \|y_\Omega - \mathcal{P}_{U_\Omega} y_\Omega\|_2^2 \leq (1+\alpha)\frac{m}{d}\|v\|_2^2 \tag{9}$$

where $\alpha = \sqrt{2\frac{\mu(v)}{m}\log(1/\delta)} + \frac{2\mu(v)}{3m}\log(1/\delta), \beta = (1 + 2\log(1/\delta))^2$, and $\gamma = \sqrt{\frac{8r\mu(U)}{3m}\log(2d/\delta)}$.

This result showcases much stronger concentration of measure than the result of Balzano et al. [2010]. The main difference is in the definitions of α and β, which in their work have worse dependence on the coherence parameter $\mu(v)$. Some of these improvements were established by Krishnamurthy and Singh [2013], but our result further improves the dependence on β, which will play out into our stronger sample complexity guarantee for the matrix completion algorithm. In terms of proof, we use scalar, vector, and matrix Bernstein's inequality to control the terms in the decomposition:

$$\|y_\Omega - \mathcal{P}_{U_\Omega} y_\Omega\|_2^2 \geq \|v_\Omega\|_2^2 - \|(U_\Omega^T U_\Omega)^{-1}\|_2 \|U_\Omega^T v_\Omega\|_2^2.$$

The decomposition is valid provided that $U_\Omega^T U_\Omega$ is invertible, which we will account for.

The above result, followed by some algebraic manipulations, yields the following corollary, which we use in the analysis of the Algorithm 1:

Corollary 7. *Suppose that $\tilde{U}$ is a subspace of U and $x_t \in U$ but $x_t \notin \tilde{U}$. Observe a set of coordinates $\Omega \subset [d]$ of m entries sampled uniformly at random with replacement. If $m \geq 32 r\mu_0 \log^2(2r/\delta)$ then with probability $\geq 1 - 4\delta$, $\|x_{t\Omega} - \mathcal{P}_{\tilde{U}_\Omega} x_{t\Omega}\|_2 > 0$. If $x_t \in \tilde{U}$, then conditioned on the fact that $U_\Omega^T U_\Omega$ is invertible, $\|x_{t\Omega} - \mathcal{P}_{\tilde{U}_\Omega} x_{t\Omega}\|_2 = 0$ with probability 1.*

Proof. The second statement follows from the fact that if $x_t \in \tilde{U}$, then $x_{t\Omega} \in \tilde{U}_\Omega$, so the projection onto the orthogonal complement is identically zero. As for the first statement, we apply Theorem 6, noting that the conditions on m are satisfied.

We now verify that the lower bound is strictly positive. We will use the fact that any vector v in U has coherence $\mu(v) \leq r\mu_0$ and similarly any subspace $\tilde{U} \subset U$ has $\dim(\tilde{U})\mu(\tilde{U}) \leq r\mu_0$. Plugging in m into the definition α, γ, and using the previous facts, we see that $\alpha < 1/2$ and $\gamma < 1/3$. We are left with:

$$\|x_{t\Omega} - \mathcal{P}_{\tilde{U}_\Omega} x_{t\Omega}\|_2^2 \geq \frac{1}{d}\left(\frac{m}{2} - \frac{3r\mu\beta}{2}\right)$$

and the lower bound is strictly positive whenever $3r\mu\beta \leq m$. Plugging in the definition of β, we see that this relation is also satisfied, concluding the proof. $\square$

We are now ready to prove Theorem 1. First notice that our estimate U for the column space is always a subspace of the true column space, since we only ever add in fully observed vectors that live in the column space. Also notice that we only resample the set Ω at most $r+1$ times, since the matrix is exactly rank r, and we only resample when we find a linearly independent column. Thus with probability $1-(r+1)\delta$, by application of Lemma 12 from the appendix, all of the matrices $\tilde{U}_\Omega^T \tilde{U}_\Omega$ are invertible.

When processing the tth column, one of two things can happen. Either x_t lives in our current estimate for the column space, in which case we know from the above corollary that with probability 1, $\|x_{t\Omega} - \mathcal{P}_{U_\Omega} x_{t\Omega}\|^2 = 0$. This holds since we have already conditioned on the fact that $U_\Omega^T U_\Omega$ is invertible. When this happens we do not obtain additional samples and just need to ensure that we reconstruct x_t, which we will see below. If x_t does not live in U, then with probability $\geq 1-4\delta$ the estimated projection is strictly positive, in which case we fully observe the new direction x_t and augment our subspace estimate. In fact, this failure probability includes the event that $U_\Omega^T U_\Omega$ is not invertible.

Since X has rank at most r, this latter case can happen no more than r times, and via a union bound, the failure probability is $\leq 4r\delta + \delta$. Here the last factor of δ ensures that the last subsampled projection operator is well behaved. In other words, with probability $\geq 1-4r\delta-\delta$, our estimate U at the end of the algorithm is exactly the column space of X.

The vectors that were not fully observed are recovered exactly as long as $(U_\Omega^T U_\Omega)^{-1}$ is invertible. This follows from the fact that, if $x_t \in U$, we can write $x_t = U\alpha_t$ and we have:

$$\hat{x}_t = U(U_\Omega^T U_\Omega)^{-1} U_\Omega^T U_\Omega \alpha_t = U\alpha_t = x_t$$

We already accounted for the probability that these matrices are invertible. We showed above that the total failure probability is $\leq 5r\delta$ and solving for δ in the sample complexity in Corollary 7, we find that:

$$\delta \leq 2r \exp\left\{-\sqrt{\frac{m}{32r\mu_0}}\right\},$$

which gives the risk bound.

For the running time, per column, the dominating computational costs involve the projection $\mathcal{P}_{\tilde{U}_\Omega}$ and the reconstruction procedure. The projection involves several matrix multiplications and the inversion of a $r \times r$ matrix, which need not be recomputed on every iteration. Ignoring the matrix inversion, this procedure takes at most $O(mr)$ time per column, since the vector and the projector are subsampled to m-dimensions, for a total running time of $O(nmr)$. At most r times, we must recompute $(U_\Omega^T U_\Omega)^{-1}$, which takes $O(r^2 m)$, contributing a factor of $O(r^3 m)$ to the total running time. Finally, we run the Gram-Schmidt process once over the course of the algorithm, which takes $O(dr^2)$ time.

5.2 Proof of Theorem 2

The proof of the necessary condition in Theorem 2 is based on a standard reduction-to-testing style argument. The high-level architecture is to consider a subset $\mathcal{X}' \subset \mathcal{X}$ of inputs and lower bound the Bayes risk.

Specifically, if we fix a prior π supported on $\mathcal{X}'$,

$$\begin{aligned} R^\star &= \inf_{f\in\mathcal{F}} \inf_{q\in\mathcal{Q}} \max_{X\in\mathcal{X}} \mathbb{P}_{\Omega\sim q}[f(\Omega, X_\Omega) \neq X] \\ &\geq \inf_{f\in\mathcal{F}} \inf_{q\in\mathcal{Q}} \mathbb{E}_{\Omega\sim q, X\sim\pi}[\mathbb{P}_f[f(\Omega, X_\Omega) \neq X]] \\ &\geq \inf_{f\in\mathcal{F}} \min_{\Omega:|\Omega|=m} \mathbb{E}_{X\sim\pi}[\mathbb{P}_f[f(\Omega, X_\Omega) \neq X]] \end{aligned}$$

The first step is a standard one in information theoretic lower bounds and follows from the fact that the maximum dominates any expectation over the same set. The second step is referred to as Yao's Minimax Principle in the analysis of randomized algorithms, which says that one need only consider deterministic algorithms if the input is randomized. It is easily verified by the fact that in the second line, the inner expression is linear in q, so it is minimized on the boundary of the simplex, which is a deterministic choice of Ω. We use $\mathbb{P}_f$ to emphasize that f can be randomized, although it will suffice to consider deterministic f.

Let π be the uniform distribution over $\mathcal{X}' \subset \mathcal{X}$. The minimax risk is lower bounded by:

$$R^\star \geq 1 - \max_{\Omega} \mathbb{E}_{X\sim\pi} |\{X' \in \mathcal{X}' | X'_\Omega = X_\Omega\}|^{-1}$$

since if there is more than one matrix in $\mathcal{X}'$ that agrees with X on Ω, the best any estimator could do is guess. Notice that since X is drawn uniformly, this is equivalent to considering an f that deterministically picks on matrix $X' \in \mathcal{X}'$ that agrees with the observations.

To upper bound the second term, define $\mathcal{U}_\Omega = \{X \in \mathcal{X}' : |\{X' \in \mathcal{X}' | X'_\Omega = X_\Omega\}| = 1\}$ which is the set of matrices that are uniquely identified by the entries Ω. Also set $\mathcal{N}_\Omega = \mathcal{X}' \setminus \mathcal{U}_\Omega$, which is the set of matrices that are not uniquely identified by Ω. We may write:

$$\max_{\Omega} \mathbb{E}_{X\sim\pi} |\{X' \in \mathcal{X}' | X'_\Omega = X_\Omega\}|^{-1} \leq \max_{\Omega} \frac{1}{2} + \frac{|\mathcal{U}_\Omega|}{2|\mathcal{X}'|}$$

Since if $X \in \mathcal{N}_\Omega$, there are at least two matrices that agree on those observations, so the best estimator is correct with probability no more than $1/2$.

We now turn to constructing a set $\mathcal{X}'$. Set $l = \frac{d}{r\mu_0}$. The left singular vectors $u_1, \ldots, u_{r-1}$ will be constant on $\{1, \ldots, l\}, \{l+1, \ldots, 2l\}$ etc. while the first $r-1$ right singular vectors $v_1, \ldots, v_{r-1}$ will be the first $r-1$ standard basis elements. We are left with:

$$d - (r-1)l = d - \frac{r-1}{r}\frac{d}{\mu_0} \triangleq dc_1,$$

coordinates where we will attempt to hide the last left singular vector. Here we defined $c_1 = 1 - \frac{r-1}{r\mu_0}$, which is not a constant, but will ease the presentation. For u_r, we pick l coordinates out of the dc_1 remaining, pick a sign for each and let u_r have constant magnitude on those coordinates. There are $2^l \binom{dc_1}{l}$ possible choices for this vector. The last right singular vector is one of the $n-r$ remaining standard basis vectors. Notice that our choice of l ensures that every matrix in this family meets the column space incoherence condition.

To upper bound $|\mathcal{U}_\Omega|$ notice that since u_r can have both positive and negative signs, a matrix is uniquely identified only if all of the entries corresponding to the last singular vector are observed. Thus observations in the tth column only help to identify matrices whose last rank was hidden in that column. If we use m_t observations on the tth column, we uniquely identify $2^l \binom{m_t}{l}$ matrices, where $\binom{m_t}{l} = 0$ if $m_t < l$. In total we have:

$$|\mathcal{X}'| = (n-r)2^l \binom{dc_1}{l} \qquad \text{and} \qquad |\mathcal{U}_\Omega| = 2^l \sum_{i=r}^{n} \binom{m_i}{l}$$

We are free to choose m_i to maximize $|\mathcal{U}_\Omega|$ subject to the constraints $m_i \leq dc_1$ and $\sum_i m_i \leq m$, the total sensing budget. Optimizing over m_i is a convex maximization problem with linear constraints, and consequently the solution is on the boundary. By symmetry, this means that that best sampling pattern is to observe columns in their entirety and devote the remaining observations to one more column. With m observations, we can observe $\frac{m}{c_1 n}$ columns fully, leading to the bounds:

$$|\mathcal{U}_\Omega| \leq 2^l \lceil \frac{m}{c_1 n} \rceil \binom{nc_1}{l}, \qquad \text{and} \qquad \frac{|\mathcal{U}_\Omega|}{|\mathcal{X}'|} \leq \lceil \frac{m}{c_1 n} \rceil \frac{1}{n_2 - r},$$

which, after plugging in for c_1, leads to the lower bound on the risk.

5.3 Proof of Theorem 3

To prove the main approximation theorem, we must analyze the three phases of the algorithm. The analysis of the first phase is fairly straightforward: we show that under the incoherence assumption, one can compute a reliable estimate of each column norm from a very small number of measurements per column. For the second phase, we show that by sampling according to the re-weighted distribution using the column-norm estimates, the matrix $\tilde{X}$ is close to X in spectral norm. We then translate this spectral norm guarantee into a approximation guarantee for $\hat{X} = \tilde{X}_r$.

Let us start with this translation. We use a lemma of Achlioptas and Mcsherry [2007].

Lemma 8 (Achlioptas and Mcsherry [2007])**.** *Let A and N be any matrices and write $\hat{A} = A + N$. Then:*

$$\|A - \hat{A}_k\|_2 \leq \|A - A_k\|_2 + 2\|N_k\|_2$$
$$\|A - \hat{A}_k\|_F \leq \|A - A_k\|_F + \|N_k\|_F + 2\sqrt{\|N_k\|_F \|A_k\|_F}$$

The lemma states that if $\hat{A} - A$ is small, then the top k ranks of $\hat{A}$ is nearly as good an approximation to A as is the top k ranks of A itself. Notice that all of the error terms only depend on rank-k matrices. We will use this lemma with $\tilde{X}$ and X and of course with the target rank as r. We will soon show that $\|X - \tilde{X}\|_2 \leq \epsilon \|X\|_F$, which implies:

$$\begin{aligned}
\|X - \hat{X}\|_F &\leq \|X - X_r\| + \|(X - \tilde{X})_r\|_F + 2\sqrt{\|(X - \tilde{X})_r\|_F \|X_r\|_F} \\
&\leq \|X - X_r\| + \sqrt{r}\|X - \tilde{X}\|_2 + 2\sqrt{\sqrt{r}\|X - \tilde{X}\|_2 \|X\|_F} \\
&\leq \|X - X_r\| + \|X\|_F \left(\sqrt{r}\epsilon + 2r^{1/4}\epsilon^{1/2}\right) \qquad (10)
\end{aligned}$$

So if we can obtain a bound on $\|X - \tilde{X}\|_2$ of that form, we will have proved the theorem.

As for Propositions 4 and 5, the translation uses the first inequality of Achlioptas and Mcsherry [2007]. If X is rank r, the matrix $\hat{X} - X$ has rank at most $2r$, which means that:

$$\|\hat{X} - X\|_F \leq \sqrt{2r}\|\hat{X} - X\|_2 \leq 2\sqrt{2r}\|\tilde{X} - X\|_2 \leq 2\sqrt{2r}\epsilon\|X\|_F$$

For the second proposition, we first bound $\|\hat{X} - M\|_2$ and then use the same argument.

$$\begin{aligned}
\|\hat{X} - M\|_2 &\leq \|\hat{X} - X\|_2 + \|R\|_2 \leq \|X - X_r\|_2 + 2\epsilon\|X\|_F + \|R\|_2 \\
&\leq 2\|R\|_2 + 2\epsilon(\|M\|_F + \|R_\Omega\|_F).
\end{aligned}$$

To arrive at the second line, we use the fact that X_r is the best rank r approximation to X, so $\|X - X_r\|_2 \leq \|X - M\|_2 = \|R\|_2$. We also use the triangle inequality on the term $\|X\|_F$, but use the fact that since the algorithm never looked at X on Ω^C it is fair to set $R_{\Omega^C} = 0$.

Let us now turn to the first phase. In our analysis of the Algorithm 1, we proved that the norm of an incoherent vector can be approximated by subsampling. Specifically, Lemma 10 shows that with high probability, the estimates $\hat{c}_t$ once appropriately rescaled are trapped between $(1-\alpha)c_t$ and $(1+\alpha)c_t$ where $\alpha = \sqrt{2\mu/m_1 \log(n/\delta)} + \frac{2\mu}{3m_1}\log(n/\delta)$. The same is of course true for $\hat{f}$. Setting $m_1 \geq 32\mu \log(n/\delta)$ we find that $\alpha \leq 1/2$, meaning that by using in total $32n\mu\log(n/\delta)$ samples in the first phase, we approximate the target sampling distribution to within a multiplicative factor of $1/2$ with probability $\geq 1 - \delta$.

For the second pass, we must show that $\tilde{X}$ is close to X in spectral norm. Some calculations, that we defer to the appendix, give the following lemma:

Lemma 9. *Provided that* $(1-\alpha)c_t \leq \frac{d}{m_1}\hat{c}_t \leq (1+\alpha)c_t$ *and* $(1-\alpha)f \leq \frac{d}{m_1}\hat{f} \leq (1+\alpha)f$*, with probability* $\geq 1 - \delta$*:*

$$\|\tilde{X} - X\|_2 \leq \|X\|_F \sqrt{\frac{1+\alpha}{1-\alpha}} \left(\sqrt{\frac{4}{m_2} \max\left(\frac{d}{n}, \mu\right) \log\left(\frac{d+n}{\delta}\right)} + \frac{4}{3}\sqrt{\frac{d\mu}{m_2 n}} \log\left(\frac{d+n}{\delta}\right) \right)$$

The adaptive sampling procedure has a dramatic effect on the bound in Lemma 9. If one sampled uniformly across the columns, then both terms grows with the squared norm of the largest column rather than with the average squared norms, which is much weaker when the energy of the matrix is concentrated on a few columns. This is precisely when the row space is coherent.

To wrap up, recall that $1 \leq \mu \leq d$ and $n \geq d$. Setting $m_1 \geq 32\mu\log(n/\delta)$ so that $\alpha \leq 1/2$, the bound in Lemma 9 is dominated by:

$$\|\tilde{X} - X\|_2 \leq \|X\|_F \frac{10}{\sqrt{3}} \sqrt{\frac{\mu}{m_2}} \log\left(\frac{d+n}{\delta}\right).$$

Returning to Equation 10 we can now substitute in for ϵ and conclude the proof.

6 Discussion

This paper considers the two related problems of low rank matrix completion and matrix approximation. In both problems, we show how to use adaptive sampling to overcome uniformity assumptions that have pervaded the literature. Our algorithms focus measurements on interesting columns (in the former, the columns that contain new directions and in the latter, the high energy columns) and have performance guarantees that are significantly better than any known passive algorithms in the absence of uniformity. Moreover, they are competitive with state-of-the-art passive algorithms in the presence of uniformity. Our algorithms are conceptually simple, easy to implement, and fairly scalable.

There are several interesting directions for future work and we mention two here. First, while we did discuss a lower bound on adaptive algorithms for matrix completion, we do not have a lower bound on the performance of adaptive algorithms for the matrix approximation problem. Such a bound would gives us a better understanding on the fundamental limits of the matrix approximation problem. More broadly, we are only beginning to understand the power of adaptive sampling and active learning in unsupervised settings and it would be interesting, both theoretically and practically, to develop this line of work further.

Acknowledgements

This research is supported in part by NSF under grants IIS-1116458 and CAREER award IIS-1252412. AK is supported in part by an NSF Graduate Research Fellowship.

References

Dimitris Achlioptas and Frank Mcsherry. Fast computation of low-rank matrix approximations. *Journal of the ACM*, April 2007.

Dimitris Achlioptas, Zohar S. Karnin, and Edo Liberty. Near-optimal entrywise sampling for data matrices. In *Advances in Neural Information Processing Systems*, 2013.

Sanjeev Arora, Elad Hazan, and Satyen Kale. A fast random sampling algorithm for sparsifying matrices. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. 2006.

Sivaraman Balakrishnan, Mladen Kolar, Alessandro Rinaldo, and Aarti Singh. Recovering block-structured activations using compressive measurements. *arXiv:1209.3431*, 2012.

Laura Balzano, Benjamin Recht, and Robert Nowak. High-dimensional matched subspace detection when data are missing. In *IEEE International Symposium on Information Theory*. IEEE, June 2010.

Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. *ACM-SIAM Symposium on Discrete Algorithms*, January 2009.

Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near optimal column-based matrix reconstruction. In *IEEE Symposium on Foundations of Computer Science*, October 2011.

Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, January 2010.

Emmanuel J Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, June 2010.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, April 2009.

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, May 2010.

Emmanuel J. Candès and Michael .B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, March 2008.

Yudong Chen. Incoherence-optimal matrix completion. *arXiv:1310.0154*, October 2013.

Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Coherent matrix completion. In *International Conference on Machine Learning*, 2014.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo Algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, January 2006a.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, 2006b.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, January 2008.

Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM*, November 2004.

Alex Gittens. The spectral norm error of the naive Nystrom extension. *arXiv:1110.5305*, October 2011.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, March 2011.

Venkatesan Guruswami and Ali Kemal Sinop. Optimal column-based low-rank matrix reconstruction. In *ACM-SIAM symposium on Discrete Algorithms*. SIAM, January 2012.

Moritz Hardt. Understanding alternating minimization for matrix completion. *arXiv:1312.0925*, December 2013.

Jarvis Haupt, Rui Castro, and Robert Nowak. Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory*, 2011.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *ACM Symposium on Theory of Computing*, June 2013.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, June 2010.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, October 2011.

Akshay Krishnamurthy and Aarti Singh. Low-rank matrix and tensor completion via adaptive sampling. *Advances in Neural Information Processing Systems*, 2013.

Akshay Krishnamurthy, Sivaraman Balakrishnan, Min Xu, and Aarti Singh. Efficient active algorithms for hierarchical clustering. In *International Conference on Machine Learning*, 2012.

Akshay Krishnamurthy, James Sharpnack, and Aarti Singh. Recovering graph-structured activations using adaptive compressive measurements. *arXiv:1305.0213*, 2013.

Matthew Malloy and Robert Nowak. Sequential analysis in high-dimensional multiple testing and sparse recovery. *IEEE International Symposium on Information Theory*, 2011.

Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *The Journal of Machine Learning Research*, March 2012.

Benjamin Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, February 2011.

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 2007.

Aarti Singh, Akshay Krishnamurthy, Sivaraman Balakrishnan, and Min Xu. Completion of high-rank ultrametric matrices using selective entries. In *IEEE International Conference on Signal Processing and Communications*, July 2012.

Ervin Tánczos and Rui Castro. Adaptive sensing for estimation of structured sparse signals. *arXiv:1311.7118*, 2013.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, August 2011.

A Proof of Theorem 6

For completeness we provide the entire proof of Theorem 6 although apart from the improved concentration bounds, the proof is similar to that of Balzano *et al.* Balzano et al. [2010].

We begin with the decomposition:

$$\|y_\Omega - \mathcal{P}_{U_\Omega} y_\Omega\|_2^2 = \|v_\Omega\|_2^2 - v_\Omega^T U_\Omega (U_\Omega^T U_\Omega)^{-1} U_\Omega^T v_\Omega. \tag{11}$$

Next, let $W_\Omega^T W_\Omega = (U_\Omega^T U_\Omega)^{-1}$, which is valid provided that $U_\Omega^T U_\Omega$ is invertible (which we will subsequently ensure). We have:

$$v_\Omega^T U_\Omega (U_\Omega^T U_\Omega)^{-1} U_\Omega^T v_\Omega = \|W_\Omega U_\Omega^T v_\Omega\|_2^2 \leq \|W_\Omega|_2^2 \|U_\Omega^T v_\Omega\|_2^2 = \|(U_\Omega^T U_\Omega)^{-1}\| \|U_\Omega^T v_\Omega\|_2^2,$$

which means that:

$$\|v_\Omega\|_2^2 - \|(U_\Omega^T U_\Omega)^{-1}\| \|U_\Omega^T v_\Omega\|^2 \leq \|y_\Omega - \mathcal{P}_{U_\Omega} y_\Omega\|_2^2 \leq \|v_\Omega\|_2^2. \tag{12}$$

The theorem now follows from three lemmas, which control the quantities in the above inequalities. The first lemma is identical to the one in Krishnamurthy and Singh Krishnamurthy and Singh [2013] while the third is from Balzano *et al.* Balzano et al. [2010]. The second one improves on both of the similar results from those to works.

Lemma 10. *With the same notations as in Theorem 6, with probability* $\geq 1 - 2\delta$*:*

$$(1-\alpha)\frac{m}{d}\|v\|_2^2 \leq \|v_\Omega\|_2^2 \leq (1+\alpha)\frac{m}{d}\|v\|_2^2 \tag{13}$$

Proof. The proof is an application of Bernstein's inequality (Theorem 13). Let $\Omega(i)$ denote the ith coordinate in the sample and let $X_i = v_{\Omega(i)}^2 - \frac{1}{d}\|v\|_2^2$ so that $\sum_{i=1}^m X_i = \|v_\Omega\|_2^2 - \frac{m}{d}\|v\|_2^2$. The variance and absolute bounds are:

$$\sigma^2 = \sum_{i=1}^m \mathbb{E} X_i^2 \leq \frac{m}{n}\sum_{i=1}^n v_i^4 \leq \frac{m}{n}\|v\|_\infty^2 \|v\|_2^2, \qquad R = \max \|X_i\| \leq \|v\|_\infty^2.$$

Bernstein's Inequality then shows that:

$$\mathbb{P}\left(\left|\sum_{i=1}^m X_i\right| \geq t\right) \leq 2\exp\left(\frac{-t^2}{2\|v\|_\infty^2(\frac{m}{d}\|v\|_2^2 + \frac{1}{3}t}\right).$$

Setting $t = \alpha \frac{m}{d} \|v\|_2^2$ and using the definition $\mu(v) = d\|v\|_\infty^2 / \|v\|_2^2$ this bound becomes:

$$\mathbb{P}\left(\left| \sum_{i=1}^{m} X_i \right| \geq \alpha \frac{m}{d} \|v\|_2^2 \right) \leq 2 \exp\left(\frac{-\alpha^2}{2\mu(v)(1+\alpha/3)} \right)$$

And plugging in the definition of α ensures that the probability is upper bounded by 2δ. □

Lemma 11. *With the same notation as Theorem 6 and provided that $m \geq 4\mu(v)\log(1/\delta)$, with probability at least $1-\delta$:*

$$\|U_\Omega^T v_\Omega\|_2^2 \leq \beta \frac{m}{d} \frac{r\mu(U)}{d} \|v\|_2^2 \tag{14}$$

Proof. The proof is an application of the vector version of Bernstein's inequality (Proposition 14. Let $u_i \in \mathbb{R}^r$ denote the ith row of an orthonormal basis for U and set $X_i = u_{\Omega(i)} v_{\Omega(i)}$. Since $v \in U^\perp$, the X_is are centered so we are left to compute the variance:

$$\sum_{i=1}^{m} \mathbb{E}\|X_i\|^2 = \frac{m}{d} \sum_{j=1}^{d} \|u_j v_j\|^2 \leq \frac{m}{d} \frac{r\mu(U)}{d} \|v\|_2^2 = V$$

Applying Proposition 14 and re-arranging, we have that with probability at least $1-\delta$:

$$\|U_\Omega^T v_\Omega\|_2 \leq \sqrt{V} + \sqrt{4V\log(1/\delta)} = \sqrt{\frac{m}{d}\frac{r\mu}{d}} \|v\|_2 \left(1 + 2\sqrt{\log(1/\delta)}\right)$$

As long as:

$$t = \sqrt{4V\log(1/\delta)} \leq V(\max_i \|X_i\|)^{-1}$$

Since $\max_i \|X_i\| \leq \|v\|_\infty \sqrt{r\mu/d}$ and using the incoherence assumption on v this condition translates to $m \geq 4\mu(v)\log(1/\delta)$. Squaring the above inequality proves the lemma. □

Lemma 12 (Balzano et al. [2010])**.** *Let $\delta > 0$ and $m \geq \frac{8}{3} r\mu(U)\log(2r/\delta)$. Then*

$$\|(U_\Omega^T U_\Omega)^{-1}\|_2 \leq \frac{d}{(1-\gamma)m} \tag{15}$$

with probability at least $1-\delta$ provided that $\gamma < 1$. In particular $U_\Omega^T U_\Omega$ is invertible.

B Proof of Lemma 9

Under the uniform at random sampling model, we will apply the non-commutative Bernstein inequality (Proposition 16) to bound $\|\tilde{X} - X\|_2$. Recall that for each column x_t, we observe a set of $m_{2,t} = m_2 n \frac{\hat{c}_t}{\hat{f}}$ observations and form the zero-filled vector y_t define dby:

$$y_t = \frac{d}{m_{2,t}} \sum_{s=1}^{m_{2,t}} x_t(i_s) e_{i_s}$$

where $\{i_s\}_{s=1}^{m_{2,t}}$ are the observations. Since the set of observations is sampled with replacement (although duplicates in each half of the sample are thrown out), each entry of y_t occurs with probability $d/m_{2,t}$, so y_t is an unbiased estimate of x_t. So we will apply the rectangular Matrix Bernstein inequality to $y_t e_t^T - x_t e_t^T$. Moreover:

$$\|y_t e_t^T - x_t e_t^T\| \leq \|y_t\|\|e_t\| + \|x_t\| \leq \left(1 + \sqrt{\frac{d\mu}{m_{2,t}}}\right)\|x_t\| \leq 2\sqrt{\frac{d\mu}{m_{2,t}}}\|x_t\|$$

which follows by the triangle inequality, Cauchy-Schwarz and the chain of inequalities:

$$\|y_t\|_2 \leq \sqrt{m_{2,t}}\|y_t\|_\infty \leq \frac{d}{\sqrt{m_{2,t}}}\|x_t\|_\infty \leq \sqrt{\frac{d\mu}{m_{2,t}}}\|x_t\|_2$$

When we plug in for $m_{2,t}$ we get:

$$\|y_t e_t^T - x_t e_t^T\| \leq 2\sqrt{\frac{d\mu}{m_2 n}\frac{c_t}{\hat{c}_t}\hat{f}} \leq 2\|X\|_F\sqrt{\frac{d\mu}{m_2 n}\frac{1+\alpha}{1-\alpha}}$$

where α is the error bound from the first phase of the algorithm.

As for the variance terms in Proposition 16, both turn out to be quite small as we will soon see. For the first term:

$$\begin{aligned}\|\sum_{t=1}^n \mathbb{E}e_t y_t^T y_t e_t^T - e_t x_t^T x_t e_t^T\| &= \|\sum_{t=1}^n e_t e_t^T(\mathbb{E}\|y_t\|^2 - \|x_t\|^2)\| = \\ &= \|\sum_{t=1}^n e_t e_t^T(\frac{d}{m_{2,t}} - 1)\|x_t\|^2\| \leq 2d\max_{t\in[n]}\frac{\|x_t\|^2}{m_{2,t}}\end{aligned}$$

The first equality is straightforward while the second follows from linearity of expectation and the fact that each coordinate of y_t is non-zero with probability $m_{2,t}/d$. The third line follows from the fact that applying the sum leads to an $n \times n$ diagonal matrix with $\frac{d}{m_{2,t}}\|x_t\|^2$ on the diagonal. When we use our definition of $m_{2,t}$ this becomes:

$$\|\sum_{t=1}^n \mathbb{E}e_t y_t^T y_t e_t^T\| \leq \frac{2d}{m_2 n}\|X\|_F^2\frac{1+\alpha}{1-\alpha}$$

For the second term, we have:

$$\begin{aligned}\|\sum_{t=1}^n \mathbb{E}y_t e_t^T e_t y_t^T - \mathbb{E}x_t e_t^T e_t x_t^T\| = \|\sum_{t=1}^n \mathbb{E}y_t y_t^T - x_t x_t^T\| &= \|\sum_{t=1}^n(\frac{d}{m_{2,t}} - 1)\text{diag}(x_t(1)^2,\ldots,x_t(d)^2)\| \\ &\leq \max_{i\in[d]}\sum_{t=1}^n\frac{2d}{m_{2,t}}x_t(i)^2 \leq \sum_{i=1}^n\frac{2\mu}{m_{2,t}}\|x_t\|_2^2 \leq \|X\|_F^2\frac{2\mu}{m_2}\frac{1+\alpha}{1-\alpha}\end{aligned}$$

Here the first equality is trivial while the second one uses the fact that off diagonals of $y_t y_t^T$ are unbiased for $x_t x_t^T$ and hence we are left with a diagonal matrix. To arrive at the second line we note that the spectral norm a diagonal matrix is simply the largest diagonal entry. Then we apply the incoherence assumption and final our sampling distribution.

At this point we may apply the inequality which states that with probability $\geq 1-\delta$:

$$\|\sum_{t=1}^n y_t e_t^T - x_t e_t^T\| \leq \|X\|_F\sqrt{\frac{1+\alpha}{1-\alpha}}\left(\sqrt{\frac{4}{m_2}\max(\frac{d}{n},\mu)\log(\frac{d+n}{\delta})} + \frac{4}{3}\sqrt{\frac{d\mu}{m_2 n}}\log(\frac{d+n}{\delta})\right)$$

C Some Concentration Inequalities

Here we collect a number of concentration inequalities used in our proofs.

Proposition 13 (Scalar Bernstein)**.** *Let $X_1, \ldots, X_n$ be independent, centered scalar random variables with $\sigma^2 = \sum_{i=1}^n \mathbb{E}[X_i^2]$ and $R = \max_i |X_i|$. Then:*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left\{\frac{-t^2}{2\sigma^2 + \frac{2}{3}Rt}\right\} \tag{16}$$

Proposition 14 (Vector Bernstein Gross [2011])**.** *Let $X_1, \ldots, X_n$ be independent centered random vectors with $\sum_{i=1}^n \mathbb{E}\|X_i\|_2^2 \leq V$. Then for any $t \leq V(\max_i \|X_i\|_2)^{-1}$:*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\|_2 \geq \sqrt{V} + t\right) \leq \exp\left\{\frac{-t^2}{4V}\right\} \tag{17}$$

Proposition 15 (Matrix Bernstein Tropp [2011])**.** *Let $X_1, \ldots, X_n$ be independent, random, self-adjoint matrices with dimension d satisfying:*

$$\mathbb{E}X_k = 0 \qquad \textit{and} \qquad \|X_k\|_2 \leq R \textit{ almost surely.}$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^n X_k\right\| \geq t\right) \leq d \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right) \qquad \textit{where} \qquad \sigma^2 = \left\|\sum_{k=1}^n \mathbb{E}X_k^2\right\|$$

Proposition 16 (Rectangular Matrix Bernstein Tropp [2011])**.** *Let $X_1, \ldots, X_n$ be independent random matrices with dimension $d_1 \times d_2$ satisfying:*

$$\mathbb{E}X_k = 0 \qquad \textit{and} \qquad \|X_k\|_2 \leq R \textit{ almost surely.}$$

Define:

$$\sigma^2 = \max\left\{\left\|\sum_{k=1}^n \mathbb{E}(X_k X_k^T)\right\|_2, \left\|\sum_{k=1}^n \mathbb{E}(X_k^T X_k)\right\|_2\right\}.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^n X_k\right\|_2 \geq t\right) \leq (d_1 + d_2) \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$