Low-Rank Matrix Completion for Positive Semidefinite Matrices

Jaya Narasimhan

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2016-79
http://www.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-79.html

May 13, 2016

Copyright © 2016, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Acknowledgement

I would like to thank my advisor Professor Satish Rao for his continued guidance and mentorship throughout this project. I would also like to thank Tselil Schramm for introducing me to this topic and for all her suggestions and advice.

Low-Rank Matrix Completion for Positive Semidefinite Matrices

by Jaya Narasimhan

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor S. Rao
Research Advisor

(Date)

* * * * * * *

Professor P. Raghavendra
Second Reader

(Date)

Low-Rank Matrix Completion for Positive Semidefinite Matrices

Jaya Narasimhan

May 13, 2016

Abstract

In this report we lower the bounds on the number of required sampled entries for reconstructing low-rank positive semidefinite matrices through nuclear norm minimization. We show for an $n \times n$ matrix of rank r only $O(rn \log n \log \log n)$ samped entries are needed to complete the matrix as compared to previous work's $O(rn \log^2 n)$ entries. We obtain this result by using rejection sampling when constructing a dual variable certificate.

1 Introduction

In many applications, one desires to reconstruct a low-rank matrix from only a small sample of its entries. This problem appears in collaborative filtering [1], dimensionality reduction [2], and multi-class learning [3]. While many approximation algorithms and heuristics have been created to solve this, the general problem of finding the lowest rank matrix subject to equality constraints is NP-hard.

However, Candès and Recht [4] show that many low-rank matrices can be recovered exactly from sufficiently large sample sets by minimizing the nuclear norm of a matrix with entries equal to the samples. The nuclear norm of a matrix $\mathbf{X}$, symbolized as $\|\mathbf{X}\|_*$, is the sum of the singular values of the matrix. It is a convex function which means it can be minimized through semidefinite programming. The rank function of a matrix counts the number of nonvanishing singular values while the nuclear norm is their sum, much like how the ℓ_1 norm of a vector is a useful approximation for the the ℓ_0 norm of a vector.

The minimization of the nuclear norm has long been used as a heuristic for low-rank matrix completion in practice [5], but only recently has there been any theoretical basis for it. Recht, Fazel, and Parrilo [6] first provided the foundations by studying the behavior of nuclear norm minimization in the average case, and showed that it solved most instances of the problem with sufficiently large sample sizes. In [4] and Candès and Tao [7], these bounds were improved for most low-rank matrices. Keshavan, Montanari and Oh [8] showed how matrices could be reconstructed exactly in a special case by sampling a set of entries a polylogarithmic factor larger than the *intrinsic dimension* of a rank r matrix, which is the number of measurements required to represent a rank r matrix. Recht [9] and Gross [10] both present results lowering the bound on the required entries to reconstruct the matrix with minimal assumptions.

In this report we lower the bound in [9] even further for positive semi-definite matrices.

Not all matrices can be reconstructed exactly, however. Consider an $n \times n$ matrix with a single 1 entry and 0's in every other entry. To reconstruct this matrix with high probability one would have to sample all of its entries. Therefore, the matrices that can be reconstructed with a few samples need to have entries that provide similar amounts of information. To state this more precisely Candès and Recht in [4] introduce the following definition:

Definition 1.1. (Coherence) Let U be a subspace of $\mathbb{R}^n$ of dimension r and $\mathbf{P}_U$ is the projection onto S. Then the coherence of U is defined as:

$$\mu(U) \equiv \frac{n}{r} \max_{1 \leq i \leq n} \|\mathbf{P}_U \mathbf{e}_i\| \tag{1.1}$$

The smallest the coherence of a subspace can be is 1, which happens if all the vectors spanning U have entries $\pm\frac{1}{\sqrt{n}}$. The largest the coherence can be is $\frac{n}{r}$, which corresponds to U being spanned by at least one standard basis element. If the matrix has low coherence, then all entries can be expected to provide the same amount of information.

Our main result is the following theorem:

Theorem 1.2. *Let* $\mathbf{M}$ *be an* $n \times n$ *positive semidefinite matrix of rank* r *with singular value decomposition* $\mathbf{U\Sigma U}^*$. *We assume that*

A0 *The row and column spaces have coherence bounded by* $\mu(\mathbf{U}) \leq \mu_0$ *for some positive* μ_0

A1 *The matrix* $\mathbf{UU}^*$ *has a maximum entry bounded in absolute value by* $\mu_1\sqrt{r/(n^2)}$ *for some positive* μ_1

Suppose m *entries of* $\mathbf{M}$ *are observed, the locations sampled uniformly at random. If*

$$m \geq O(\mu_0 rn \log(n) \log\log n) \tag{1.2}$$

the minimizer to the problem

$$\begin{array}{ll} \textit{minimize} & \|\mathbf{X}\|_* \\ \textit{subject to} & X_{i,j} = M_{i,j}, (i,j) \in \Omega \end{array} \tag{1.3}$$

is unique and equal to $\mathbf{M}$ *with high probability, where* Ω *is the set of sampled indices.*

The assumptions **A0** and **A1** were introduced in [4] and [9]. In addition to **A1**, Candès and Tao [7] require a much stronger incoherence condition than **A0** with additional conditions on r. Keshavan *et. al* [8] require conditions on the ratio between the largest and smallest singular values of $\mathbf{M}$ and the rank r of the matrix. In [9] Recht presents a more compact proof using just **A0** and **A1** and requires $O(\mu_0 rn \log^2(n))$ entries. Gross [10] shows a similar result using tools from quantum information theory and improves the bound for matrices that are incoherent to every basis to $O(\mu_0 rn \log n)$

From the coupon collectors problem, at least $n \log n$ entries have to be sampled to guarantee that every column and every row is sampled at least once. Candès and Tao show that $O(\mu_0 rn \log n)$ entries are necessary for completion in [7] as well. The bound in Theorem 1.2 is then within a factor of $\log\log n$ of optimal.

As in [4] and [9], our proof relies on the construction of a dual variable to certify that $\mathbf{M}$ is the exact minimizer of equation 1.3. The dual is constructed in [9] through an iterative process guaranteeing that at each iteration, the dual variable gets closer and closer in Frobenius norm to $\mathbf{UU}^*$ and a low spectral norm. Here we instead reject samples that do not fit those criteria since we are constructing the solution ourselves. Because of this we use rejection sampling on each column of the dual variable to lower its spectral norm. The columns of the dual are independent so we reject any sampled column that exceeds a threshold norm. We show that the sampled columns do not exceed their expected norm with high probability and therefore guarantee with a certain number of iterations that the dual variable fulfills the necessary conditions.

2 Preliminaries

Throughout this report, I will follow the convention that matrices are in bold, vectors are bold lowercase, and scalars and entries are unbolded lowercase. The transpose of matrices and vectors will be indicated with a star—$\mathbf{X}^*$ is the transpose of the matrix $\mathbf{X}$.

There will also be a few different matrix norms used throughout this report. The nuclear norm of a matrix is denoted $\|\mathbf{X}\|_*$. The spectral norm (the largest singular value) is $\|\mathbf{X}\|$ and the Frobenius norm is $\|\mathbf{X}\|_F$. The Euclidean inner product between two matrices is $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathrm{Tr}(\mathbf{X}^*\mathbf{Y})$. The maximum entry of a matrix in absolute value is denoted as $\|\mathbf{X}\|_\infty$. The norm of vectors is always the ℓ_2 norm denoted as $\|\mathbf{x}\|$.

Linear transformations that operate on matrices will be denoted with calligraphic letters. The spectral norm of these operators will be denoted as $\|\mathcal{A}\| = \sup_{\mathbf{X}:\|\mathbf{X}\|_F \leq 1} \|\mathcal{A}(\mathbf{X})\|_F$.

Fix a matrix $\mathbf{M}$ obeying assumptions **A0** and **A1**, with singular value decomposition $\mathbf{UU}^*$. Let $\mathbf{u}_k$ be the kth column of $\mathbf{U}$. We can now define a subspace T as the linear space formed by the span of $\mathbf{u}_k\mathbf{y}$ and $\mathbf{x}\mathbf{u}_k^T$ from $1 \leq k \leq r$ where $\mathbf{x}$ and $\mathbf{y}$ are arbitrary vectors. The projection operator onto T is defined as:

$$\mathcal{P}_\mathcal{T}(\mathbf{Z}) = \mathbf{P}_U \mathbf{Z}$$

The projection operator onto $T_\perp$ is then:

$$\mathcal{P}_{\mathcal{T}_\perp}(\mathbf{Z}) = (\mathcal{I} - \mathcal{P}_\mathcal{T})(\mathbf{Z})$$

It will be of use to note that any unit vector $\mathbf{z} \in T$ has a small maximum entry value, using **A1**.

$$z_j = \sum \alpha_i \mathbf{u}_i j \leq \sum \alpha_i \sqrt{\frac{r}{n^2}} \leq \sqrt{\frac{r}{n^2}}$$

2.1 Useful Theorems

The following theorems will be useful in the main argument.

Theorem 2.1 (Standard Bernstein Inequality)**.** *Let $X_1 \dots X_n$ be independent, zero-mean random variables. If $|X_i| \leq M$ for all i, then for positive τ*

$$\mathbb{P}\left[\sum_i X_i > \tau\right] \leq \exp\left(\frac{\tau^2/2}{\sum_j \mathbb{E}[X_j^2] + M\tau/3}\right)$$

Theorem 2.2 (Matrix Bernstein: Rectangular Case)**.** *Let $\{\mathbf{Z}_k\}$ be a sequence of independent random matrices with mean $\mathbf{0}$ and of dimension $d_1 \times d_2$. Assume that for surely all k*

$$\|\mathbf{Z}_k\| \leq M$$

Let

$$\sigma^2 := \max\left(\|\sum_k \mathbb{E}(\mathbf{Z}_k \mathbf{Z}_k^*)\|, \|\sum_k \mathbb{E}(\mathbf{Z}_k^* \mathbf{Z}_k)\|\right)$$

Then for all $\tau \geq 0$,

$$\mathbb{P}\left[\|\sum_k \mathbf{Z}_k\| \geq \tau\right] \leq (d_1 + d_2) \cdot \exp\left(\frac{-\tau^2/2}{\sigma^2 + M\tau/3}\right)$$

3 Setup

We define a sampling operator, $\mathcal{R}_\Omega$ which samples each entry of a column with probability $\frac{k}{n}$ and scales it by $\frac{n}{k}$. In this way, each entry is sampled independently from the other entries.

if we construct a dual variable $\mathbf{Y}$ in the range of $\mathcal{R}_\Omega$ that satisfies the two following conditions, we will show that the minimizer of equation 1.3 is $\mathbf{M}$ because $\mathbf{Y}$ will be a dual certificate.

$$\|\mathcal{P}_\mathcal{T}(\mathbf{Y}) - \mathbf{U}\mathbf{U}*\|_F \leq \sqrt{\frac{r}{2n}} \tag{3.1}$$

$$\|\mathcal{P}_{\mathcal{T}_\perp}(Y)\| \leq \frac{1}{2} \tag{3.2}$$

4 Constructing a Dual Certificate.

First, we start by using the construction of the dual found in [9]. We partition samples into p partitions of size $q \geq \frac{512}{3} \max \mu_0, \mu_1^2 r n \beta \log(2n)$. Let $\mathcal{R}_{\Omega_l}$ be the sampling operator that samples indices of the lth partition.

Let $\mathbf{W}_0 = \mathbf{U}\mathbf{U}^*$ and $\mathbf{Y}_l = \frac{n^2}{q} \sum_j^l \mathcal{R}_{\Omega_l}(\mathbf{W}_{j-1})$, with $\mathbf{W_1} = \mathbf{U}\mathbf{V}^* - \mathcal{P}_\mathcal{T}(\mathbf{Y}_l)$. Using the definitions of $\mathbf{W}_l$ and Theorem 3.4 from [9] we have

$$\|\mathbf{W}_l\|_F = \|\mathbf{W}_{l-1} - \frac{n^2}{q}\mathbf{W}_{l-1}\|_F = \|(\mathcal{P}_\mathcal{T} - \frac{n^2}{q}\mathcal{P}_\mathcal{T}\mathcal{R}_{\Omega_l}\mathcal{P}_\mathcal{T})(\mathbf{W}_{l-1})\|_F \leq \frac{1}{2}\|\mathbf{W}_{l-1}\|_F$$

which implies that $\|\mathbf{W}_l\| \leq 2^{-l}\|\mathbf{W}_0\|_F = 2^{-l}\sqrt{r}$ for $l = 1, \dots p$.

If we set $p = 10 \log\log n$, then the frobenius and spectral norms become:

$$\|\mathbf{U}\mathbf{U}^* - \mathcal{P}_\mathcal{T}(\mathbf{Y}_p)\|_F \leq \frac{1}{2^p}\|\mathbf{U}\mathbf{U}^*\|_F \leq \frac{1}{\log^{10} n}\|\mathbf{U}\mathbf{U}^*\|_F \tag{4.1}$$

$$\|\mathcal{P}_{\mathcal{T}_\perp}\mathbf{Y}_p\| < \sqrt{\frac{32u_1^2 rn\beta\log(2n)}{3q}} < \frac{1}{4} \tag{4.2}$$

Let $\|z\|_c$ be the the maximum ℓ_2 norm of a column of $\mathbf{Z}$. After the initial p rounds, the maximum column measure of the $\mathbf{Y}$'s is also reduced:

$$\|W_p\|_c \leq \frac{1}{2^p}\|\mathbf{U}\mathbf{U}^*\|_c \leq \frac{1}{\log^{10} n}\|\mathbf{U}\mathbf{U}^*\|_c \leq \frac{1}{\log^{10} n}\|\mathbf{U}\mathbf{U}^*\|_c \tag{4.3}$$

For the second part of the construction, we continue sampling $\mathbf{W}_l$'s but we sample column by column. If a column has too large of an ℓ_2 norm we reject it. We show below that the probability that a column is rejected is $< \frac{1}{2}$. We now need to see how many rounds of this rejection sampling we need to guarantee that the spectral norm decreases to $\frac{1}{4}$.

We also have the following theorem which states that the norms of columns sampled with $O(r)$ samples do not deviate from their expected value with high probability.

Theorem 4.1. *Let $k = \frac{r}{\alpha}$, $\alpha < \frac{1}{2}$. For $\mathbf{Z} \in T$, if we have k entries sampled in expectation*

$$E[\|(\mathcal{P}_\mathcal{T}\mathcal{R}_{\Omega_l}\mathcal{P}_\mathcal{T} - \mathcal{P}_\mathcal{T})\mathbf{Z}\|_F] \leq \alpha\|\mathbf{Z}\|_F \tag{4.4}$$

Moreover, with high probability, if $k = 100\frac{\mu_0 r\log n}{\alpha}$, we get that

$$\|(\mathcal{P}_\mathcal{T}\mathcal{R}_{\Omega_l}\mathcal{P}_\mathcal{T} - \mathcal{P}_\mathcal{T})\mathbf{Z}\|_F \leq \alpha\|\mathbf{Z}\|_F \leq 3\alpha\|\mathbf{Z}\|_F \tag{4.5}$$

And finally if M is a positive semi-definite matrix, for a sufficiently small constant α, we have

$$\|(\mathcal{P}_\mathcal{T}\mathcal{R}_{\Omega_l}\mathcal{P}_\mathcal{T} - \mathcal{P}_\mathcal{T})\| \leq \frac{1}{2} \tag{4.6}$$

Proof. We want to sample k entries from each column vector $\mathbf{z}$ of the matrix $\mathbf{Z}$. Let $\hat{\mathbf{z}}$ be the sampled column of $\mathbf{Z}$. We want $\mathcal{P}_\mathcal{T}(\hat{\mathbf{z}})$ to be close to $\mathcal{P}_\mathcal{T}(\mathbf{z})$. Because $\mathbf{U}$ is a basis for the space, we only have to check that $\|\mathbf{U}^*\hat{\mathbf{z}}\|_F$ is close to $\|\mathbf{U}^*\mathbf{z}\|_F$. Each entry will be sampled independently with probability $\frac{k}{n}$. Let X_i be a random variable that is $\frac{n}{k}$ if entry z_i of the column vector is selected with $p = \frac{k}{n}$ and 0 otherwise. Let $\alpha_j = \sum z_i u_{ij}$ and $\hat{\alpha}_j = \sum X_i z_i u_{ij}$. The squared error in the norm of the column $\mathbf{z}$ is then $\sum_{j=1}^r(\alpha_j - \hat{\alpha}_j)^2$. We can write this sum as a sum of independent random variables as follows.

$$\sum_{j=1}^r(\alpha_j - \hat{\alpha}_j)^2] = \sum_{j=1}^r\sum_{i=1}^n(z_i u_{ij}(1-X_i))^2 = \sum_{i=1}^n z_i^2(1-X_i)^2\sum_{j=1}^r u_{ij}^2 \leq \frac{\mu_0 r}{n}\sum_{i=1}^n z_i^2(1-X_i)^2 \tag{4.7}$$

Using the fact that $\mathbb{E}[(1-X_i)^2] = \frac{n}{k}$, we obtain the following.

$$\mathbb{E}[\sum_{j=1}^r(\alpha_j - \hat{\alpha}_j)^2] \leq \frac{\mu_0 r}{k}\|\mathbf{z}\|$$

We can choose $k = \frac{\mu_0 r}{\alpha}$, and due to $\mathbf{U}$ being a basis, and reasoning column by column yields

$$\mathbb{E}[\|\mathbf{U}\mathbf{U}^*(\mathbf{Z} - \hat{\mathbf{Z}})]\|_F] \leq \alpha\|\mathbf{Z}\|_F$$

This is equation 4.4.

When $k = \frac{100\mu_0 r\log n}{\alpha}$, we observe that each the total variance of the sum of random variables is 4.7 is at most $E[(\frac{\mu_0 r}{n})^2\sum_i|z_i|^4(1-X_i)^4]$ which is at most $(\frac{\mu_0 r}{n})^2(\frac{n}{k})^3|z_i|^4\sum_i v_i^2 v_i^2$, where $v_i = \frac{z_i}{|z_i|}$. We know that $|v_i| \leq \sqrt{r}n$ since $\mathbf{z}$ is in T. Thus, we get that the variance is at most $(\frac{r}{k})^3\|\mathbf{z}\|^4$. Moreover, the maximum value of any term is $\frac{r}{k}\|\mathbf{z}\|^2$. Thus, we get that the deviation τ is at most $2\max(\alpha\|\mathbf{z}\|^2, \sqrt{2\alpha^3}\|\mathbf{z}\|^2)$ with high probability ($> 1 - n^{-100}$) if $k = 100\frac{r\log n}{\alpha}$. This yields equation 4.5.

To bound the spectral norm of this process, we again view the sampling as applied to each column of a matrix $\mathbf{Z}$ in T, and call it $\mathbf{Z}_j$. We note that $\mathbf{Z}_j\mathbf{Z}_j^*$ is the matrix whose (j,j) entry corresponds to equation 4.7 for column j. Thus, we have that $E[\|\sum_j \mathbf{Z}_j\mathbf{Z}_j^*\|]$ is bounded by the maximum of these n values as the sum is diagonal. Using the previous argument, each is at most $3\alpha\|\mathbf{z}_j\|_F$, with high probability. Moreover, with high

probability $\|\mathbf{Z}_j\|$ is bounded by the square root of this quantity. Thus by the Matrix Bernstein inequality, we can conlude that $\|\mathcal{P}_\mathcal{T}\mathcal{R}_\Omega\mathcal{P}_\mathcal{T} - \mathcal{P}_\mathcal{T})(\mathbf{Z})\| \leq 5\alpha\|\mathbf{Z}\|$ as long as $k > 100\frac{r\log n}{\alpha}$ with high probability ($> 1 - n^{-c}$ for some constant c.) Equation 4.6 follows by choosing $\alpha < 1/10$.

□

We now set $s = O(kn)$, the expected number of entries sampled in each round, and let $\mathbf{Y'}_i = (\mathcal{R}_{\Omega_l}\mathcal{P}_\mathcal{T}\mathbf{W'}_{i-1})$. However, we zero out or **reject**, any column, j, where $\|(\mathcal{P}_\mathcal{T}\mathbf{Y'}_i - \mathbf{W'}_{i-1})^{(j)}\| > \|\mathbf{W'}^{(j)}_{i-1}\|$. We continue this rejection sampling for $i = 1 \dots 10\log n$ rounds and then we have the final $\mathbf{W}_i$ satisfying equations 3.1 and 3.2. Now we argue that the Frobenius norm and the spectral norm both drop in these remaining rounds.

Theorem 4.2. *For each column with probability $1 - \alpha$ the remaining Frobenius norm drops.*

Proof. By Markov and 4.1 we know that on expectation the column norm is decreased if $\alpha < \frac{1}{2}$ by Markov.

□

Theorem 4.3. *For $10\log n$ steps of the rejection sampling, all the column norms will be small.*

Proof. There will be $\log n$ successes in $10\log n$ trials wiih high probability if $\alpha < \frac{1}{2}$.

□

Finally, we need to argue that $\|\sum_i \mathbf{Y'}_i\|$ is small. This trivially follows from the equation 4.3 and the following theorem, which will be proven below.

Theorem 4.4. *Let $\mathbf{Z}$ be an $n \times n$ positve semi-definite matrix in T and $\mathcal{R}_\Omega$ the sampling operator that samples $k = \frac{\mu_0}{\alpha}r$ samples per column of a matrix. Let $\|\mathbf{z}\|_c$ be the norm of the largest column of $\mathbf{Z}$. Then*

$$\|(\mathcal{R}_\Omega - \mathcal{I})(\mathbf{Z})\| \leq \alpha\log^{3/2}(n)\|\mathbf{z}\|_c\sqrt{10\mu_0\frac{n}{r}} \tag{4.8}$$

with high probability with $0 < \alpha < 1$ for some α.

5 Proof of Theorem 4.4

Proof. First we show that the operator norm can be bounded above by the maximum column norm

$$\|\mathbf{Z}\| = \sup_{\|\mathbf{x}\|=1,\|\mathbf{y}\|=1}\sum_{a,b} Z_{ab}y_a x_b \leq \left(\sum_{a,b} Z_{ab}^2 y_a^2\right)^{1/2}\left(\sum_{a,b} x_b^2\right)^{1/2} \leq \sqrt{n}\max_a\left(\sum_b Z_{ab}^2\right)^{1/2} \leq n\|\mathbf{z}\|_c$$

Next, we want the norms of the sampled columns to not exceed the norm of the maximum column with high probability. We will show this using the Standard Bernstein Inequality. Let $\hat{\mathbf{z}}$ be the jth column of $(\mathcal{R}_\Omega - \mathcal{I})(\mathbf{Z})$

$$\|\hat{\mathbf{z}}\|^2 = \sum_{i=1}^{n}(X_{ij}Z_{ij} - Z_{ij})^2$$

where $X_{ij} = \frac{n}{k}$ with probability $\frac{k}{n}$ and 0 otherwise. Note that this value is a sum of independent, zero-mean random variables. Let $Y_i = (1 - X_i)^2 Z_{ij}^2$. Let $\mathbf{v} = \frac{\mathbf{z}}{\|\mathbf{z}\|}$ and note that the maximum entry is $\sqrt{\mu_0\frac{n}{r}}$ since $\mathbf{Z} \in T$. Therefore

$$\max_i Y_i = \|\mathbf{z}\|^2\max_i\left(v_i\frac{n}{k}\right)^2 \leq \frac{n}{r}\|\mathbf{z}\|^2\alpha^2$$

The variance of the norm is similarly calculated

$$\sum_i \mathbb{E}[Y_i^2] = \|\mathbf{z}\|^4\sum_i (v_i)^4\left(\frac{n}{k}\right)^3 \leq \|\mathbf{z}\|^4\frac{\mu_0 n}{r}\left(\frac{n}{k}\right)^3\sum_i(v_i^2) \leq \|\mathbf{z}\|^4\alpha\left(\frac{n}{k}\right)^2$$

Thus, by the Bernstein Inequality, we have $\|\hat{\mathbf{z}}\|^2 \leq 10\mu_0 \frac{n}{r}\alpha^2\|\mathbf{z}\|_c^2 \log n$ with high probability. The $10 \log n$ factor is added to make sure the event occurs with high probability.

Now that we have an upper bound on the norms of the estimated columns of $\mathbf{Z}$ with high probability, we can use the Matrix Bernstein inequality to bound $\|(\mathcal{R}_\Omega - \mathcal{I})(\mathbf{Z})\|$. Let $\mathbf{Z}_j$ be the matrix which is zero except for the jth column, which has each entry sampled with probability $\frac{k}{n}$ from $\mathbf{Z}$'s jth column. Let

$$\sigma^2 = \|\sum_j \mathbf{Z}_j^* \mathbf{Z}_j\| = 10\mu_0\alpha^2 \frac{n}{r}\|\mathbf{z}\| \log n$$

since the sum is a diagonal matrix with column norms on the diagonal.

Using the matrix Bernstein Inequality on the $\mathbf{Z}_j$'s with $\tau \geq \alpha \log^{3/2}(n)\|\mathbf{z}\|_c \sqrt{10\mu_0 \frac{n}{r}}$ yields equation 4.8 with high probability. □

6 Discussion

The results here extend from the use of rejection sampling independent columns. We strongly suspect that the number of entries required can be reduced to $O(rn \log n)$ for positive semidefinite matrices by extending this rejection sampling scheme. They key is to guarantee that the spectral norm of $\mathcal{P}_{\mathcal{T}_\perp}(\mathbf{Y})$ is not too large and does not increase later on in the construction. A beginning is to use a refined Matrix Bernstain type bound to prove a version of Theorem 4.4 where the $\log^{3/2} n$ is replaced by a $\log n$. This would allow us to forgo the use of the sampling in [9] entirely.

Moreover, we may be able to extend this line of reasoning for general matrices but we would have to understand how to reject both columns and rows if the norm of either became too large. It also may be possible to eliminate assumption **A1** as no one has shown that an assumption similar to **A1** is necessary for matrix completion. However all prior results impose such an assumption so it may be necessary after all.

References

[1] J. D. M. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, pp. 713–719, 2005.

[2] N. Linial, E. London, and Y. Rabinovich, "The geometry of graphs and some of its algorithmic applications," *Combinatorica*, vol. 15, no. 2, pp. 215–245, 1995.

[3] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[4] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[5] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *American Control Conference, 2001. Proceedings of the 2001*, vol. 6, pp. 4734–4739, IEEE, 2001.

[6] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[7] E. J. Candès and T. Tao, "The power of convex relaxation: near-optimal matrix completion," *IEEE Trans. Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[8] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[9] B. Recht, "A simpler approach to matrix completion," *Journal of Machine Learning Research*, vol. 12, pp. 3413–3430, 2011.

[10] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Information Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

A Proof that M is the minimum

Here I will outline the proof in [9] that shows that $\mathbf{M}$ is the unique and exact minimizer to equation 1.3 with high probability. Assume we have dual variable $\mathbf{Y}$ satisfying equations 3.1 and 3.2

It shown in [9] that for any $\mathbf{Z} \in \ker \mathcal{R}_\Omega$

$$\|\mathcal{P}_{\mathcal{T}_\perp}(\mathbf{Z})\|_F > \sqrt{\frac{2r}{n}}\|\mathcal{P}_{\mathcal{T}}(\mathbf{Z})\|_F$$

Since, we use $O(rn \log n \log\log n)$ rather than $O(rn \log^2 n)$ samples, we get a slightly weaker bound.

$$\|\mathcal{P}_{\mathcal{T}_\perp}(\mathbf{Z})\|_F > \sqrt{\frac{2r \log\log n}{n \log n}}\|\mathcal{P}_{\mathcal{T}}(\mathbf{Z})\|_F$$

For any $\mathbf{Z} \in \ker \mathcal{R}_\Omega$, choose $\mathbf{U}_\perp$ such that $[\mathbf{U}, \mathbf{U}_\perp]$ is a unitary matrix and that $\langle \mathbf{U}_\perp \mathbf{U}_\perp^*, \mathcal{P}_{\mathcal{T}_\perp}(\mathbf{Z})\rangle = \|\mathcal{P}_{\mathcal{T}_\perp}(\mathbf{Z})\|_*$. Note that $\langle \mathbf{Y}, \mathbf{Z}\rangle = 0$ for all $\mathbf{Z} \in \ker \mathcal{R}_\Omega$. The nuclear norm can also be defined as $\|\mathbf{A}\|_* = \sup_{\|\mathbf{B}\| \le 1} \langle \mathbf{A}, \mathbf{B}\rangle$. Note that for any matrix $\mathbf{X}$, $\|\mathbf{X}\|_F \le \|\mathbf{X}\|_*$.

$$\|\mathbf{M} + \mathbf{Z}\|_* \ge \langle \mathbf{U}\mathbf{U}^* + \mathbf{U}_\perp \mathbf{U}_\perp^*, \mathbf{M} + \mathbf{Z}\rangle$$

This follows from the variational characteristics. Continuing on we get:

$$= \|\mathbf{M}\|_* + \langle \mathbf{U}\mathbf{U}^* + \mathbf{U}_\perp \mathbf{U}_\perp^*, \mathbf{Z}\rangle$$

$$= \|\mathbf{M}\|_* + \langle \mathbf{U}\mathbf{U}^* - \mathcal{P}_{\mathcal{T}}(\mathbf{Y}), \mathcal{P}_{\mathcal{T}}(\mathbf{Z})\rangle + \langle \mathbf{U}_\perp \mathbf{U}_\perp^* - \mathcal{P}_{\mathcal{T}_\perp}(\mathbf{Y}), \mathcal{P}_{\mathcal{T}}(\mathbf{Z})\rangle$$

$$\|\mathbf{M}\|_* - \sqrt{\frac{r}{2n \log n}}\|\mathcal{P}_{\mathcal{T}}(\mathbf{Z})\|_F + \frac{1}{2}\|\mathcal{P}_{\mathcal{T}_\perp}(\mathbf{Z})\|_* \ge \|\mathbf{M}\|_*$$

The last inequality differs from that in [9] by a log n factor to compensate for the smaller lower bound on $\mathcal{P}_{\mathcal{T}_\perp}(\mathbf{Z})\|_F$. The correspondingly small upper bound on $\mathcal{P}_{\mathcal{T}}(\mathbf{Z})$ is obtained by running $O(\log\log n)$ more interations of our rejection sampling process.