Adaptive Strategy for Restricted-Sampling Noisy Low-Rank Matrix Completion

Daniel L. Pimentel-Alarcón, Robert D. Nowak
University of Wisconsin-Madison

***Abstract*—In this paper we propose a novel adaptive algorithm that provably performs low-rank matrix completion (LRMC) from restricted sets of observations, under ideal or noisy measurements, in lieu of coherence assumptions, with minimal sampling rates and optimal computational complexity. We discuss the main advantages of the adaptive setting of LRMC, and complement our theoretical analysis with experiments, illustrating the effectiveness of our algorithm.**

I. Introduction

The problem of low-rank matrix completion (LRMC) consists on estimating the missing entries in a rank-r matrix. This scenario arises in a wide range of practical applications, ranging from image processing [1] to collaborative filtering and recommender systems [2, 3], among others [4–8].

LRMC has been widely studied under a missing-at-random and bounded-coherence model [9–18], and recently under an adaptive setting [19], where one may select, on the go, which entries to observe. This is tightly related to the problem of identifying a low-rank approximation of a matrix from a subset of its entries [20–23].

These approaches assume that one may sample unrestrictedly, e.g., that one may choose to observe all entries of a given column, or assume access to certain characteristics of the matrix, e.g., its coherence, or the norms of its columns. This is a valid assumption for certain applications, e.g., when the matrix is stored in disk, and one may observe entries of the matrix with absolute freedom. However, this is not the case in many LRMC applications, where the entries one may choose to observe are typically limited. Take for example recommender systems, where obtaining a complete column equates to asking a single user (column) to evaluate every item (row). In these problems the number or rows can be very large, hence this can be an unreasonable thing to ask. Moreover, the combinations of rows that one may sample could be restricted. For example, some users (e.g., children) may be well suited to evaluate certain subsets of items (e.g., toys or candies), but unable to evaluate others (e.g. wines or home appliances). An other example arises in distributed settings, where at each location one may only sense a subset of all the information.

Motivated by scenarios like this, we propose a novel adaptive algorithm that provably performs LRMC from limited observations (e.g., only sampling few observations per column, which amounts, for example, to only asking each user about a few selected items), under ideal or noisy measurements, in lieu of coherence assumptions, with minimal sampling rates and optimal computational complexity.

Organization of the paper

In Section II we formally state the problem and present our algorithm and our main results, which we prove in Section IV. In Section III we discuss related work and the main advantages of using adaptive strategies for LRMC. Section V presents experiments that support our findings and demonstrate the effectiveness of our approach.

II. Model, algorithm and main results

Let $\mathbf{X}$ be the $d \times N$ data matrix of rank-r that we aim to reconstruct by observing a few selected subset of the entries of the noisy matrix $\mathbf{Y}$ given by

$$\mathbf{Y} \;=\; \mathbf{X} + \mathbf{Z},$$

where the elements of $\mathbf{Z}$ are i.i.d., zero-mean random variables with variance σ^2 and finite fourth moment. To characterize the sampling restrictions, let $\Upsilon \subset 2^{\{1,\dots,d\}}$ indicate the (unknown) sets of rows where one may choose to sample a column of $\mathbf{Y}$.

We will drop the coherence assumptions typically required for LRMC. Instead, our statements hold for *almost every* (a.e.) $\mathbf{X}$, with respect to the uniform measure on $\mathcal{X}$, the set of all $d\times N$ rank-r matrices. The zero-measure subset of $\mathcal{X}$ for which our statements do not hold is essentially that of pathological matrices, e.g., ones with zero rows or columns.

Our strategy (summarized in Algorithm 1) is mainly divided into an exploration phase and a reconstruction phase. In the exploration phase (which is the adaptive portion of the algorithm) we will search for a sampling matrix $\mathbf{\Omega}$ (compatible with the allowed observations in Υ). The matrix $\mathbf{\Omega}$ will encode the information of the entries that we will choose to observe in the reconstruction phase, where we will estimate $\mathbf{X}$ from the selected samples.

More precisely, let $\mathbf{\Omega}$ be a $d \times (d-r)$ matrix with exactly $r+1$ nonzero entries per column, and consider the following condition:

C1 Every matrix formed with a subset of n columns of $\mathbf{\Omega}$ has at least $n+r$ nonzero rows.

There exist plenty of matrices $\mathbf{\Omega}$ satisfying **C1**, for example

$$\mathbf{\Omega} \;=\; \begin{bmatrix} \mathbf{1} \\ \hline \\ \mathbf{I} \\ \\ \end{bmatrix} \begin{matrix} \} r \\ \\ \} d-r, \end{matrix} \tag{1}$$

where $\mathbf{1}$ denotes a block of all 1's and $\mathbf{I}$ the identity matrix. In addition, let $\boldsymbol{\omega}_i \subset \{1,\dots,d\}$ denote the set of nonzero rows in the i^{th} column of $\mathbf{\Omega}$.

In the exploration phase, we will propose a matrix $\mathbf{\Omega}$ satisfying **C1**, and verify whether it is compatible with the allowed observations, i.e., whether

$$\{\boldsymbol{\omega}_i\}_{i=1}^{d-r} \subset \Upsilon. \tag{2}$$

Notice that Υ is unknown, but we can test whether $\boldsymbol{\omega}_i \in \Upsilon$ by simply trying to sample a column of $\mathbf{Y}$ in $\boldsymbol{\omega}_i$. If (2) is satisfied, we will proceed to the recovery phase. Otherwise, we will propose new matrices $\mathbf{\Omega}$ satisfying **C1** until we find one that also satisfies (2). If there exists no such matrix $\mathbf{\Omega}$, $\mathbf{X}$ cannot be recovered by any means (see Corollary 1 in [25]).

Once we have a matrix $\mathbf{\Omega}$ satisfying **C1** and (2), we will proceed to the recovery phase, where we will observe $N_i \geq r$ columns of $\mathbf{Y}$ in the rows of $\boldsymbol{\omega}_i$ to obtain an $(r+1) \times N_i$ matrix $\mathbf{Y}_{\boldsymbol{\omega}_i}$. We will use $\mathbf{Y}_{\boldsymbol{\omega}_i}$ to estimate $S_{\boldsymbol{\omega}_i} \subset \mathbb{R}^{r+1}$, the restriction of the subspace S spanned by the columns of $\mathbf{X}$ to the rows in $\boldsymbol{\omega}_i$. In the noiseless setting, $N_i = r$ linearly independent columns will uniquely and perfectly determine $S_{\boldsymbol{\omega}_i}$, and in the presence of noise, the larger N_i, the better our estimate of $S_{\boldsymbol{\omega}_i}$ will be.

To estimate $S_{\boldsymbol{\omega}_i}$, we will compute $\hat{\mathbf{a}}_{\boldsymbol{\omega}_i}$, the $(r+1)^{th}$ singular vector of $\frac{1}{N_i}\mathbf{Y}_{\boldsymbol{\omega}_i}\mathbf{Y}_{\boldsymbol{\omega}_i}^{\mathsf{T}}$ (the sample covariance matrix of $\mathbf{Y}_{\boldsymbol{\omega}_i}$). This will be our estimate of a nonzero vector $\mathbf{a}_{\boldsymbol{\omega}_i} \in \ker S_{\boldsymbol{\omega}_i}$, (which characterizes $S_{\boldsymbol{\omega}_i}$, as $S_{\boldsymbol{\omega}_i}$ is an r-dimensional subspace of $\mathbb{R}^{r+1}$). We will do this for every column in $\mathbf{\Omega}$, thus obtaining estimates of different portions of S (characterized by $\{\hat{\mathbf{a}}_{\boldsymbol{\omega}_i}\}_{i=1}^{d-r}$). We will next stitch together these portions obtain an estimate of the whole subspace S. To this end, we will construct the vector $\hat{\mathbf{a}}_i \in \mathbb{R}^d$ with the entries of $\hat{\mathbf{a}}_{\boldsymbol{\omega}_i}$ in the rows of $\boldsymbol{\omega}_i$, and zeros elsewhere. Doing this for every $\hat{\mathbf{a}}_{\boldsymbol{\omega}_i}$, we will obtain the matrix $\hat{\mathbf{A}} = [\hat{\mathbf{a}}_1 \ \cdots \ \hat{\mathbf{a}}_{d-r}]$. The subspace $\hat{S} := \ker \hat{\mathbf{A}}^{\mathsf{T}}$ will be our estimate of S.

Once S is known, $\mathbf{X}$ can be optimally recovered observing only r entries per column. To see this, let $\mathbf{U}$ be a basis of S, and select an arbitrary set $\boldsymbol{\upsilon} \in \Upsilon$ with exactly r elements. We will use the subscript $\boldsymbol{\upsilon}$ to denote restriction to the rows in $\boldsymbol{\upsilon}$. Since the coefficients of column $\mathbf{x}$ in the basis $\mathbf{U}$ are given by $\boldsymbol{\theta} = (\mathbf{U}_{\boldsymbol{\upsilon}}^{\mathsf{T}}\mathbf{U}_{\boldsymbol{\upsilon}})^{-1}\mathbf{U}_{\boldsymbol{\upsilon}}^{\mathsf{T}}\mathbf{x}_{\boldsymbol{\upsilon}}$, we can recover the unobserved entries of this column as $\mathbf{x} = \mathbf{U}\boldsymbol{\theta}$. We may thus observe $\mathbf{Y}$ on the r rows in $\boldsymbol{\upsilon}$ to obtain $\mathbf{Y}_{\boldsymbol{\upsilon}}$, and project $\mathbf{Y}_{\boldsymbol{\upsilon}}$ onto $\hat{S}$ to obtain $\hat{\mathbf{X}} := \hat{\mathbf{U}}(\hat{\mathbf{U}}_{\boldsymbol{\upsilon}}^{\mathsf{T}}\hat{\mathbf{U}}_{\boldsymbol{\upsilon}})^{-1}\hat{\mathbf{U}}_{\boldsymbol{\upsilon}}^{\mathsf{T}}$, our estimate of $\mathbf{X}$ given $\mathbf{Y}_{\boldsymbol{\upsilon}}$.

Our first result states that Algorithm 1 will recover the subspace S with arbitrary accuracy, as long as N_i is sufficiently large to overcome the noise and achieve the desired level of precision, and

A1 There exists a matrix $\mathbf{\Omega}$ satisfying **C1** and (2).

Theorem 1. *Let **A1** hold, and let $\hat{S}$ be as in Algorithm 1. Then for* a.e. $\mathbf{X}$, $\hat{S} \to S$ *as N_i grows.*

To present our next result, define $\mathbf{X}^\star$ as the optimal estimator of $\mathbf{X}$ given $\mathbf{Y}_{\boldsymbol{\upsilon}}$, i.e.,

$$\mathbf{X}^\star \ := \ \underset{\mathbf{X}' \subset S}{\arg\min} \|\mathbf{Y}_{\boldsymbol{\upsilon}} - \mathbf{X}'_{\boldsymbol{\upsilon}}\|.$$

Algorithm 1: Adaptive Linear Algorithm for Restricted Completion Over Noise (ALARCON)

- Take a matrix $\mathbf{\Omega}$ satisfying **C1**.
- **while** $\mathbf{\Omega}$ fails to satisfy (2) **do**
  - $\mathbf{\Omega}$ = new matrix satisfying **C1**.
- **for** $i = 1$ **to** $d - r$ **do**
  - Observe $N_i \geq r$ columns of $\mathbf{Y}$ in the nonzero rows of $\boldsymbol{\omega}_i$ to obtain the $(r+1) \times N_i$ matrix $\mathbf{Y}_{\boldsymbol{\omega}_i}$.
  - Compute $\hat{\mathbf{a}}_{\boldsymbol{\omega}_i} = (r+1)^{th}$ singular vector of $\frac{1}{N_i}\mathbf{Y}_{\boldsymbol{\omega}_i}\mathbf{Y}_{\boldsymbol{\omega}_i}^{\mathsf{T}}$.
  - Construct $\hat{\mathbf{a}}_i \in \mathbb{R}^d$ with the entries of $\hat{\mathbf{a}}_{\boldsymbol{\omega}_i}$ in the nonzero rows of $\boldsymbol{\omega}_i$, and zeros elsewhere.
- Construct $\hat{\mathbf{A}} = [\hat{\mathbf{a}}_1 \ \cdots \ \hat{\mathbf{a}}_{d-r}]$.
- Take $\hat{\mathbf{U}}$, a basis of $\hat{S} = \ker \hat{\mathbf{A}}^{\mathsf{T}}$.
- Take $\boldsymbol{\upsilon} \in \Upsilon$ with exactly r elements.
- Observe $\mathbf{Y}$ on the rows in $\boldsymbol{\upsilon}$ to obtain $\mathbf{Y}_{\boldsymbol{\upsilon}}$.
- Estimate $\hat{\mathbf{X}} = \hat{\mathbf{U}}(\hat{\mathbf{U}}_{\boldsymbol{\upsilon}}^{\mathsf{T}}\hat{\mathbf{U}}_{\boldsymbol{\upsilon}})^{-1}\hat{\mathbf{U}}_{\boldsymbol{\upsilon}}^{\mathsf{T}}$.

The next theorem states that the completion given by Algorithm 1 will be arbitrarily close to the optimal completion $\mathbf{X}^\star$ if N_i is sufficiently large to achieve the desired precision.

Theorem 2. *Let **A1** hold, and let $\hat{\mathbf{X}}$ be the output of Algorithm 1. Then for* a.e. $\mathbf{X}$, $\hat{\mathbf{X}} \to \mathbf{X}^\star$ *as N_i grows.*

As a direct consequence of these results, we obtain the following specialization. It shows that in the absence of noise, Algorithm 1 will perfectly recover $\mathbf{X}$ in linear time (in the ambient dimension d) using only $r(d - r + N)$ samples (the the minimum required for completion).

Theorem 3. *Let **A1** hold, and suppose $\mathbf{Z} = 0$. Let $\hat{\mathbf{X}}$ be the output of Algorithm 1 with $N_i = r$. Then $\hat{\mathbf{X}} = \mathbf{X}$ for* a.e. $\mathbf{X}$.

III. What we gain by being adaptive

Unsurprisingly, adaptive sampling brings advantages to LRMC, the most relevants being sample and computational complexity. We will use this section to give a brief discussion on these topics, and compare our results with related work.

It is known that $\mathcal{O}(\log d)$ *random* samples per column are necessary to guarantee that LRMC is *possible*, but completing a matrix with such few random samples may be computationally prohibitive, as it may require solving a complex polynomial system of equations [24].

On the other hand, several algorithms have been shown to complete with high probability (w.h.p.) incoherent matrices with as little as $\mathcal{O}(r\mu \log d)$ random samples per column (μ being the coherence parameter indicating the alignment of the matrix [9]) using convex optimization [9–14], iterative

thresholding [15, 16] and alternating minimization [17, 18], among others. If the distribution of information over the matrix is known, nuclear norm minimization has also been shown to complete coherent matrices with as little as $\mathcal{O}(r \log^2 d)$ random samples per column [14]. These sampling assumptions are sufficient, but not necessary, and are sometimes unverifiable or unjustified in practice, as typically neither μ nor the distribution of information over the matrix are known.

In the adaptive setting, the strategy in [19] samples entire columns of an incoherent matrix, and requires either $\mathcal{O}(r^{3/2} \log r)$ additional samples per column to complete the rest of the matrix in the noiseless case, or $\mathcal{O}(r^{3/2}\text{polylog}d)$ in the presence of noise.

In contrast, a simple count of the degrees of freedom in a $d \times N$ matrix of rank-r shows that $r(d - r + N)$ samples are necessary for completion, which is exactly the number of samples required by Algorithm 1. Thus our adaptive strategy achieves the minimum sampling required for completion.

In addition, Algorithm 1 operates with as little as $r + 1$ samples per column (the minimum required, as $\mathbf{X}$ is rank-r, so observing columns with at least $r + 1$ entries is necessary for completion [24]). Therefore, Algorithm 1 works even on the minimal sampling regime.

Furthermore, our approach drops all coherence assumptions, and works with probability 1, as opposed to w.h.p.

On the other hand, in the noiseless setting, the fastest algorithm that we know of has a computational complexity of $\mathcal{O}(dr^3 \mu \log^3 d)$ [18]. On this end, Algorithm 1 requires to compute: (i) the singular value decomposition (SVD) of $d-r$ small matrices, of size $(r+1) \times r$, to obtain $\{\hat{\mathbf{a}}_{\boldsymbol{\omega}_i}\}_{i=1}^{d-r}$, (ii) the SVD of a sparse $d \times (d-r)$ matrix (with only $r+1$ nonzero entries per column), to obtain a basis of $\hat{S} = \ker \hat{\mathbf{A}}^\mathsf{T}$, and (iii) some matrix multiplications to estimate $\mathbf{X}$ once $\hat{S}$ is known.

This gives Algorithm 1 a computational complexity of $\mathcal{O}(dr^3)$, i.e., Algorithm 1 achieves linear dependency on d (ideal, as d is usually large). We thus conclude that adaptive sampling also brings computational advantages to LRMC.

IV. Proofs

Let $\mathbf{X}_{\boldsymbol{\omega}_i}$ be the $(r+1) \times N_i$ submatrix of $\mathbf{X}$ corresponding to $\mathbf{Y}_{\boldsymbol{\omega}_i}$. Notice that for a.e. $\mathbf{X}$, $\ker \mathbf{X}_{\boldsymbol{\omega}_i}^\mathsf{T}$ is a 1-dimensional subspace of $\mathbb{R}^{r+1}$.

Let $\mathbf{x}$ denote an arbitrary column of $\mathbf{X}$, and $\mathbf{x}_{\boldsymbol{\omega}_i} \in \mathbb{R}^{r+1}$ denote the restriction of $\mathbf{x}$ to the rows of $\boldsymbol{\omega}_i$. Letting $\mathbf{a}_{\boldsymbol{\omega}_i}$ be a nonzero vector of $\ker \mathbf{X}_{\boldsymbol{\omega}_i}^\mathsf{T}$, it is clear that $\langle \mathbf{a}_{\boldsymbol{\omega}_i}, \mathbf{x}_{\boldsymbol{\omega}_i} \rangle = 0$. Defining $\mathbf{a}_i$ as the vector in $\mathbb{R}^d$ with the entries of $\mathbf{a}_{\boldsymbol{\omega}_i}$ in the rows of $\boldsymbol{\omega}_i$, and zeros elsewhere, we have that $\langle \mathbf{a}_i, \mathbf{x} \rangle = 0$. Since this is true for every i, and $\mathbf{x}$ was arbitrary, letting $\mathbf{A} := [\mathbf{a}_1 \ \cdots \ \mathbf{a}_{d-r}]$ we conclude that $\mathbf{A}^\mathsf{T}\mathbf{X} = 0$, which implies $\mathbf{X} \in \ker \mathbf{A}^\mathsf{T}$.

If $\mathbf{A}$ is full-rank, then $\ker \mathbf{A}^\mathsf{T} = S$, the subspace spanned by the columns of $\mathbf{X}$. This will be the case if and only if the $(d-r)$ columns in $\mathbf{A}$ are linearly independent, which by Lemma 2 in [25], will be the case if and only if every matrix formed with a subset of n columns of $\mathbf{A}$ has at least $n + r$ nonzero rows. Since for a.e. $\mathbf{X}$, an entry of $\mathbf{A}$ will be nonzero if and only if the corresponding entry of $\boldsymbol{\Omega}$ is nonzero (Lemma 1 in [25]), we have shown the following.

Lemma 1. *For* a.e. $\mathbf{X}$, $\ker \mathbf{A}^\mathsf{T} = S$ *if and only if* ***C1*** *holds.*

Next observe that $\text{span}\{\mathbf{X}_{\boldsymbol{\omega}_i}\} = \text{span}\{\mathbf{X}_{\boldsymbol{\omega}_i}\mathbf{X}_{\boldsymbol{\omega}_i}^\mathsf{T}\}$. With this in mind, let $(\mathbf{W}, \boldsymbol{\Lambda}, \mathbf{V})$ denote the SVD of $\frac{1}{N_i}\mathbf{X}_{\boldsymbol{\omega}_i}\mathbf{X}_{\boldsymbol{\omega}_i}^\mathsf{T}$, such that $\mathbf{W} = [\mathbf{U}_{\boldsymbol{\omega}_i} \ \mathbf{a}_{\boldsymbol{\omega}_i}]$, where $\mathbf{U}_{\boldsymbol{\omega}_i}$ spans the same subspace as $\mathbf{X}_{\boldsymbol{\omega}_i}$, and $\mathbf{a}_i$ is a unitary vector in $\ker \mathbf{X}_{\boldsymbol{\omega}_i}^\mathsf{T}$.

Similarly, let $(\hat{\mathbf{W}}, \hat{\boldsymbol{\Sigma}}, \hat{\mathbf{V}})$ denote the SVD of $\frac{1}{N_i}\mathbf{Y}_{\boldsymbol{\omega}_i}\mathbf{Y}_{\boldsymbol{\omega}_i}^\mathsf{T}$, and let $\hat{\mathbf{W}} = [\hat{\mathbf{U}}_{\boldsymbol{\omega}_i} \ \hat{\mathbf{a}}_{\boldsymbol{\omega}_i}]$. Recall that by assumption, $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$, where $\text{cov}(\mathbf{Z}) = \sigma^2\mathbf{I}$, so by the strong law of large numbers (all convergences are almost surely with respect to the probability measure on $\mathbf{Z}$), as N_i grows,

$$\tfrac{1}{N_i}\mathbf{Y}_{\boldsymbol{\omega}_i}\mathbf{Y}_{\boldsymbol{\omega}_i}^\mathsf{T} \ \longrightarrow \ \tfrac{1}{N_i}\mathbf{X}_{\boldsymbol{\omega}_i}\mathbf{X}_{\boldsymbol{\omega}_i}^\mathsf{T} + \sigma^2\mathbf{I}.$$

It follows that $\hat{\mathbf{W}}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{V}}^\mathsf{T} \longrightarrow \mathbf{W}(\boldsymbol{\Lambda} + \sigma^2\mathbf{I})\mathbf{V}^\mathsf{T}$, which implies $\hat{\mathbf{a}}_{\boldsymbol{\omega}_i} \to \mathbf{a}_{\boldsymbol{\omega}_i}$. We thus have that $\hat{\mathbf{A}} \to \mathbf{A}$. Since $\boldsymbol{\Omega}$ satisfies **C1**, we know by Lemma 1 that $\ker \mathbf{A}^\mathsf{T} = S$, hence $\ker \hat{\mathbf{A}}^\mathsf{T} \to S$, which concludes the proof of Theorem 1.

In order to show Theorem 2, notice that the optimal estimator of $\mathbf{X}$ given $\mathbf{Y}_{\boldsymbol{\upsilon}}$ is given by $\mathbf{X}^\star = \mathbf{B}\mathbf{Y}_{\boldsymbol{\upsilon}}$, with $\mathbf{B} := \mathbf{U}(\mathbf{U}_{\boldsymbol{\upsilon}}^\mathsf{T}\mathbf{U}_{\boldsymbol{\upsilon}})^{-1}\mathbf{U}_{\boldsymbol{\upsilon}}^\mathsf{T}$. Let $\hat{\mathbf{B}} = \hat{\mathbf{U}}(\hat{\mathbf{U}}_{\boldsymbol{\upsilon}}^\mathsf{T}\hat{\mathbf{U}}_{\boldsymbol{\upsilon}})^{-1}\hat{\mathbf{U}}_{\boldsymbol{\upsilon}}^\mathsf{T}$. By the same arguments as before, $\hat{\mathbf{A}} \to \mathbf{A}$ as N_i grows, which implies $\hat{\mathbf{B}} \to \mathbf{B}$. We thus conclude that $\hat{\mathbf{X}} \to \mathbf{X}^\star$, as desired.

We point out that though very similar, $\mathbf{B}$ is not exactly the projection operator onto S (given by $\mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}$). The operator $\mathbf{B}$ will essentially receive an r-dimensional vector $\mathbf{x}_{\boldsymbol{\upsilon}}$, compute its coefficients in the basis $\mathbf{U}$ (through the projection-like operation: $\boldsymbol{\theta} = (\mathbf{U}_{\boldsymbol{\upsilon}}^\mathsf{T}\mathbf{U}_{\boldsymbol{\upsilon}})^{-1}\mathbf{U}_{\boldsymbol{\upsilon}}^\mathsf{T}\mathbf{x}_{\boldsymbol{\upsilon}}$), and then use those coefficients to recover the unobserved entries ($\mathbf{x} = \mathbf{U}\boldsymbol{\theta}$).

In the noiseless setting, $\mathbf{X}^\star = \mathbf{X}$, and $\hat{\mathbf{A}} = \mathbf{A}$, which implies $\hat{\mathbf{X}} = \mathbf{X}$, thus showing Theorem 3.

V. Experiments

Our results show that Algorithm 1 will estimate $\mathbf{X}$ efficiently and accurately from a few cleverly selected noisy samples. To support these findings, we simulated the setup above with $d = 100$, $r = 10$, and different values for the number of columns N_i and the noise level σ.

To this end, we first generated matrices $\mathbf{U} \in \mathbb{R}^{d \times r}$, $\boldsymbol{\Theta} \in \mathbb{R}^{N \times r}$ and $\mathbf{Z} \in \mathbb{R}^{d \times N}$, with $N = N_i(d-r)$. Next we obtained the low-rank matrix $\mathbf{X} = \mathbf{U}\boldsymbol{\Theta}^\mathsf{T}$, and the noisy matrix $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$ (which we then normalized). The entries of $\mathbf{U}$ and $\boldsymbol{\Theta}$ were drawn i.i.d., $\mathcal{N}(0, 1)$, and the entries of $\mathbf{Z}$ were drawn i.i.d., $\mathcal{N}(0, \sigma^2)$.

We ran Algorithm 1 (where we observe a few selected entries of $\mathbf{Y}$ to estimate $\mathbf{X}$) with $\boldsymbol{\Omega}$ as in (1). We repeated this experiment 100 trials, and recorded the accuracy of our algorithm as a function of N_i and σ. The results are summarized in Figure 1.

Theorem 2 states that $\hat{\mathbf{X}} \to \mathbf{X}^\star$ as N_i grows, where $\mathbf{X}^\star$ is the optimal estimator of $\mathbf{X}$ from the observed data. This is illustrated in Figure 1. Of course, how large N_i needs to be will depend on the desired level of precision, and the noise level. To investigate this, we used the results above to compute

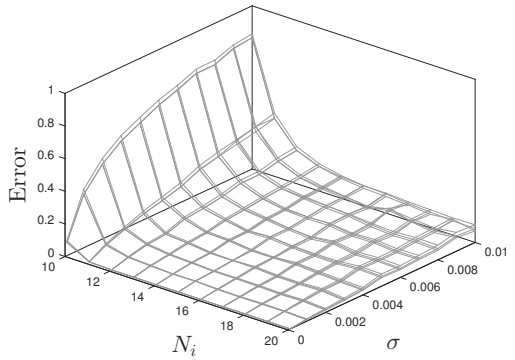

Fig. 1. Completion error of Algorithm 1 as a function of N_i and σ. For each pair (N_i, σ), we simultaneously present the three normalized errors: $\|\mathbf{X} - \hat{\mathbf{X}}\|/\|\hat{\mathbf{X}}\|$, $\|\mathbf{X}^\star - \hat{\mathbf{X}}\|/\|\hat{\mathbf{X}}\|$ and $\|\mathbf{U} - \hat{\mathbf{U}}\|/\|\hat{\mathbf{U}}\|$, where $\mathbf{X}^\star$ is the optimal estimator of $\mathbf{X}$ from the observed data, and $\hat{\mathbf{X}}$, $\hat{\mathbf{U}}$ are the estimators of Algorithm 1. Theorem 1 states that $\hat{\mathbf{U}} \to \mathbf{U}$ as N_i grows. Once $\mathbf{U}$ is known, $\mathbf{X}$ can be estimated optimally (Theorem 2 states that $\hat{\mathbf{X}} \to \mathbf{X}^\star$). It is thus not surprising that all these error quantities are very close.

the minimum N_i for which $\|\mathbf{X}^\star - \hat{\mathbf{X}}\|/\|\hat{\mathbf{X}}\| \leq \sqrt{\sigma}$. The results are shown in Figure 2.

Our next experiments involve a comparison with three non-adaptive algorithms: iterative thresholding [16], alternating minimization [17] and EM [26]. We first tried completing a noiseless matrix observed in the same entries as the ones sampled by Algorithm 1, with the setup described above. Unfortunately none of these algorithms succeeded at this task. This supports theoretical results, which show that even when non-adaptive LRMC is theoretically possible using only the minimum required $r + 1$ samples per column, this may be computationally prohibitive [24].

Finally, we studied the behavior of these algorithms at completing a noiseless 100×100, rank-10 (same as before) as a function of ℓ, the number of random samples per column. Figure 2 shows that in the best-case scenario, these algorithms may succeed with about 30 random samples per column. Nonetheless, for settings like this, theoretical results of these algorithms require all entries to be observed to guarantee the correctness and uniqueness of a result.

In contrast, our adaptive strategy provably achieves theoretical optimal sampling rates, using computational efficient algorithms, even in the presence of noise.

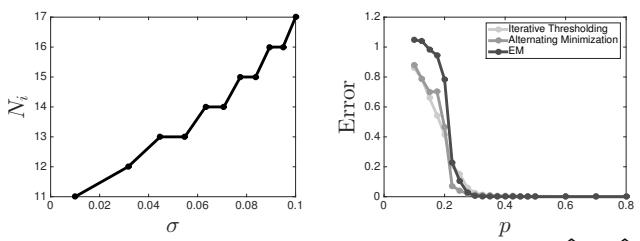

Fig. 2. **Left**: minimum number of columns N_i for which $\|\mathbf{X}^\star - \hat{\mathbf{X}}\|/\|\hat{\mathbf{X}}\| \leq \sqrt{\sigma}$, where $\mathbf{X}^\star$ is the optimal estimator of $\mathbf{X}$ from the observed data, and $\hat{\mathbf{X}}$ is the output of Algorithm 1. Theorem 2 states that $\hat{\mathbf{X}} \to \mathbf{X}^\star$ as N_i grows. **Right**: completion error of a noiseless matrix as a function of $p := \ell/d$ (the fraction of random samples per column) for three non-adaptive algorithms. Displaying the lowest completion error (best-case scenario) out of 100 trials.

VI. Conclusions

In this paper we present an adaptive, provable LRMC algorithm that only uses limited sets of observations, under ideal or noisy measurements, in lieu of coherence assumptions, with minimal sampling rates and optimal computational complexity, thus showing that the adaptive sampling setting brings several advantages over passive sampling.

References

[1] K. Weinberger and L. Saul, *Unsupervised learning of image manifolds by semidefinite programming*, International Jrnl. of Computer Vision, 2006.

[2] D. Goldberg, D. Nichols, B. Oki and D. Terry, *Using collaborative filtering to weave an information tapestry*, ACM Communications, 1992.

[3] J. Rennie and N. Srebro, *Fast maximum margin matrix factorization for collaborative prediction*, Intl. Conference on Machine Learning, 2005.

[4] Z. Liu and L. Vandenberghe, *Interior-point method for nuclear norm approximation with application to system identification*, SIAM Journal on Matrix Analysis and Applications, 2009.

[5] P. Biswas, T. Lian, T. Wang and Y. Ye, *Semidefinite programming based algorithms for sensor network localization*, ACM Transactions on Sensor Networks, 2006.

[6] C. Tomasi and T. Kanade, *Shape and motion from image streams under orthography: a factorization method*, Intl. Jrnl. of Computer Vision, 1992.

[7] Y. Amit, M. Fink, N. Srebro and S. Ullman, *Uncovering shared structures in multiclass classification*, Intl. Conference on Machine Learning, 2007.

[8] A. Argyriou, T. Evgeniou and M. Pontil, *Multi-task feature learning*, Neural Information Processing Systems, 2007.

[9] E. Candès and B. Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics, 2009.

[10] E. Candès and T. Tao, *The power of convex relaxation: near-optimal matrix completion*, IEEE Transactions on Information Theory, 2010.

[11] E. Candès and Y. Plan, *Matrix Completion With Noise*, Proceedings of the IEEE, 2010.

[12] B. Recht, *A simpler approach to matrix completion*, Journal of Machine Learning Research, 2011.

[13] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Transactions on Information Theory, 2011.

[14] Y. Chen, S. Bhojanapalli, S. Sanghavi and R. Ward, *Coherent matrix completion*, International Conference on Machine Learning, 2014.

[15] J. Cai, E. Candès and and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization, 2010.

[16] E. Chunikhina, R. Raich and T. Nguyen, *Performance analysis for matrix completion via iterative hard-thresholded SVD*, IEEE Statistical Signal Processing Workshop, 2014.

[17] P. Jain, P. Netrapalli and S. Sanghavi, *Low-rank matrix completion using alternating minimization*, ACM Symp. on Theory Of Computing, 2013.

[18] M. Hardt, *Understanding alternating minimization for matrix completion*, IEEE Annual Symp. on Foundations of Computer Science, 2014.

[19] A. Krishnamurthy and A. Singh, *Low-rank matrix and tensor completion via adaptive sampling*, Neural Information Processing Systems, 2013.

[20] A. Frieze, R. Kannan, and S. Vempala, *Fast Monte-Carlo algorithms for finding low-rank approximations*, IEEE Annual Symposium of Foundations of Computer Science, 1998.

[21] P. Drineas and R. Kannan, *Pass efficient algorithms for approximating large matrices*, Annual ACM-SIAM Symp. on Discrete Algorithms, 2003.

[22] P. Drineas, A. Frieze, R. Kannan, S. Vempala and V. Vinay, *Clustering large graphs via the singular value decomposition*, Journal of Machine Learning, 2004.

[23] D. Achlioptas and F. Mcsherry, *Fast computation of low-rank matrix approximations*, Journal of the ACM, 2007.

[24] D. Pimentel-Alarcón, N. Boston and R. Nowak, *A characterization of deterministic sampling patterns for low-rank matrix completion*, Allerton, 2015.

[25] D. Pimentel-Alarcón, N. Boston and R. Nowak, *Deterministic conditions for subspace identifiability from incomplete sampling*, IEEE International Symposium on Information Theory, 2015.

[26] D. Pimentel-Alarcón, L. Balzano and R. Nowak, *On the sample complexity of subspace clustering with missing data*, IEEE Statistical Signal Processing Workshop, 2014.