Coherent Matrix Completion

Yudong Chen YDCHEN@UTEXAS.EDU
University of California, Berkeley, CA 94720, USA

Srinadh Bhojanapalli BSRINADH@UTEXAS.EDU
Sujay Sanghavi SANGHAVI@MAIL.UTEXAS.EDU
Rachel Ward RWARD@MATH.UTEXAS.EDU
The University of Texas at Austin, Austin, TX 78712, USA

Abstract

Matrix completion concerns the recovery of a low-rank matrix from a subset of its revealed entries, and nuclear norm minimization has emerged as an effective surrogate for this combinatorial problem. Here, we show that nuclear norm minimization can recover an arbitrary $n \times n$ matrix of rank r from $\mathcal{O}(nr\log^2(n))$ revealed entries, provided that revealed entries are drawn proportionally to the local row and column coherences (closely related to leverage scores) of the underlying matrix. Our results are order-optimal up to logarithmic factors, and extend existing results for nuclear norm minimization which require strong incoherence conditions on the types of matrices that can be recovered, due to assumed uniformly distributed revealed entries. We further provide extensive numerical evidence that a proposed *two-phase* sampling algorithm can perform nearly as well as local-coherence sampling and without requiring a priori knowledge of the matrix coherence structure. Finally, we apply our results to quantify how *weighted* nuclear norm minimization can improve on unweighted minimization given an arbitrary set of sampled entries.

1. Introduction

Low-rank matrix completion has been the subject of much recent study due to its application in myriad tasks: collaborative filtering, dimensionality reduction, clustering, and localization in sensor networks. Clearly, the problem is ill-posed in general; correspondingly, analytical work on the subject has focused on the joint development of algorithms, and sufficient conditions under which such algorithms are able recover the matrix.

While they differ in scaling/constant factors, all existing sufficient conditions (Candès & Recht, 2009; Recht, 2009; Keshavan et al., 2010; Gross, 2011; Jain et al., 2012) and (Negahban & Wainwright, 2012) (with a couple of exceptions we describe in Section 2) require that *(a)* the subset of observed elements should be uniformly randomly chosen, independent of the values of the matrix elements, and *(b)* the low-rank matrix be "incoherent" or "not spiky" – i.e. its row and column spaces should be diffuse, having low inner products with the standard basis vectors . Under these conditions, the matrix has been shown to be provably recoverable – via methods based on convex optimization (Candès & Recht, 2009; Recht, 2009; Gross, 2011), alternating minimization (Jain et al., 2012), iterative thresholding (Cai et al., 2010) etc. – using as few as $\Theta(nr\log n)$ observed elements for an $n \times n$ matrix of rank r.

Actually, the incoherence assumption *is required because of the uniform sampling:* coherent matrices are those which have most of their mass in a relatively small number of elements. By sampling entries uniformly and independently at random, most of the mass of a coherent low-rank matrix will be missed; this could (and *does*) throw off all existing recovery methods. One could imagine that if the sampling is dependent on the matrix, roughly in a way that elements with more mass are more likely to be observed, then it may be possible for *existing* methods to recover the full matrix.

In this paper, we show that the incoherence requirement can be eliminated completely, provided the sampling distribution is adapted to the matrix to be recovered in the right way. Specifically, we have the following results.

1. If the probability of an element being observed is dependent on the sum of the corresponding row and column leverage scores (local coherence parameters) of

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

the underlying matrix, then an *arbitrary* rank-r matrix can be exactly recovered from $\Theta(nr \log^2 n)$ observed elements with high probability, using nuclear norm minimization. In case all leverage scores are roughly equal, our results reduce to existing guarantees for incoherent matrices using uniform sampling. Our sample complexity bound $\Theta(nr \log^2 n)$ is optimal up to a single factor of $\log^2 n$, since the degrees of freedom in an $n \times n$ matrix of rank r is $2nr$. Our bounds are also near-optimal with respect to the local coherence parameters of the matrix.

2. We provide numerical evidence that an adaptive sampling strategy, which assumes no prior knowledge about the local coherences of the underlying matrix, can perform on par with the optimal sampling strategy in completing coherent matrices, and significantly outperforms uniform sampling. Specifically, we consider a two-phase sampling strategy whereby given a fixed budget of m samples, we first draw a fixed proportion of samples uniformly at random, and then draw the remaining samples according to the the local coherence structure of the resulting sampled matrix.

3. Using our theoretical results, we are able to quantify the benefit of *weighted* nuclear norm minimization over standard nuclear norm minimization, and provide a strategy for choosing the weights in such problems given non-uniformly distributed samples so as to reduce the sampling complexity of weighted nuclear norm minimization to that of standard nuclear norm. Our results give the first exact recovery guarantee for weighted nuclear norm minimization in (Salakhutdinov & Srebro, 2010; Negahban & Wainwright, 2012; Foygel et al., 2011), thus providing theoretical justification for its good empirical performance.

Our main theoretical results are achieved by a new analysis based on bounds involving the *weighted* $\ell_{\infty,2}$ matrix norm, defined as the maximum of the appropriately weighted row and column norms of the matrix. This differs from previous approaches that use ℓ_∞ or unweighted $\ell_{\infty,2}$ bounds (Gross, 2011; Chen, 2013). In some sense, using the weighted $\ell_{\infty,2}$-type bounds is natural for the analysis of low-rank matrices, because the rank is a property of the rows and columns of the matrix rather than its individual entries, and the weighted norm captures the relative importance of the rows/columns. Therefore, it is interesting to see if the techniques in this paper are relevant more generally, beyond the specific settings and algorithms considered here.

2. Related work

There is now a vast body of literature on matrix completion, and an even bigger body of literature on matrix approximations more generally; we restrict our related work review here to papers that are most directly related.

Exact Completion, Incoherent Matrices, Random Samples: The first algorithm and theoretical guarantees for the exact recovery of a low-rank matrix from a subset of elements appeared in (Candès & Recht, 2009); there it was shown that algorithm (1) above works when the low-rank matrix is incoherent, and the sampling is uniform random and independent of the matrix. Subsequent works have refined provable completion results for incoherent matrices under the uniform random sampling model, both via nuclear norm minimization (Candès & Tao, 2010; Recht, 2009; Gross, 2011; Chen, 2013), and other methods like SVD followed by local descent (Keshavan et al., 2010), alternating minimization (Jain et al., 2012) etc, and also with both sparse errors and additive noise (Candes & Plan, 2010; Chen et al., 2013; Chandrasekaran et al., 2011).

Matrix approximations via sub-sampling: Weighted sampling methods have been widely considered in the related context of matrix *sparsification*, where one aims to approximate a large dense matrix with a sparse matrix. The strategy of element-wise matrix sparsification was introduced in (Achlioptas & Mcsherry, 2007). They propose and provide bounds for the L_2 *element-wise sampling* model, where entries of the matrix are sampled with probability proportional to their squared magnitude. These bounds were later refined in (Drineas & Zouzias, 2011). Alternatively, (Arora et al., 2006) proposed the L_1 *entry-wise sampling* model, where entries are sampled with probabilities proportional to their magnitude. This model was further investigated in (Achlioptas et al., 2013) and argued to be almost always preferable to L_2 sampling.

Closely related to the matrix sparsification problem is the matrix column selection problem, where one aims to find the "best" k column subset of a matrix to use as an approximation. State-of-the-art algorithms for column subset selection (Boutsidis et al., 2009; Mahoney, 2011) involve randomized sampling strategies whereby columns are selected proportionally to their *statistical leverage scores* – the squared Euclidean norms of projections of the canonical unit vectors on the column singular subspaces. The statistical leverage scores of a matrix can be approximated efficiently, faster than the time needed to compute an SVD (Drineas et al., 2012). Statistical leverage scores have been used extensively in statistical regression analysis for outlier detection (Chatterjee & Hadi, 1986). More recently, statistical leverage scores were used in the context of graph sparsification under the name of graph resistance (Spielman & Srivastava, 2011). The sampling distribution we use for the matrix completion in this paper is based on statistical leverage scores. As shown in Section 4.1, sampling as such outperforms both L_1 and L_2 entry-wise sampling,

at least in the context of matrix completion.

Weighted sampling in compressed sensing: This paper is similar in spirit to recent work in compressed sensing which shows that sparse recovery guarantees traditionally requiring mutual incoherence can be extended to systems which are only *weakly* incoherent, without any loss of approximation power, provided measurements from the sensing basis are subsampled according to their coherence with the sparsity basis. This notion of *local coherence sampling* seems to have originated in (Rauhut & Ward, 2012) in the context of sparse orthogonal polynomial expansions, and has found applications in uncertainty quantification (Yang & Karniadakis, 2013), interpolation with spherical harmonics (Burq et al., 2012), and MRI compressive imaging (Krahmer & Ward, 2012).

Finally, closely related to our paper is the recent work by (Krishnamurthy & Singh, 2013), which considers matrix completion where only the row space is allowed to be coherent. Their proposed algorithm selects columns to observe in their entirety and requires a total of $O(r^2 n \log r)$ observed entries, which is quadratic in r.

2.1. Organization

We present our main results for coherent matrix completion in Section 3. In Section 4 we propose a two-phase algorithm that requires no prior knowledge about the underlying matrix coherence structure. In Section 5 we provide guarantees for weighted nuclear norm minimization. We provide the proof of the main theorem in the appendix.

3. Main Results

The results in this paper hold for what is arguably the most popular approach to matrix completion: nuclear norm minimization. If the true matrix is M with entries M_{ij}, and the set of observed elements is Ω, this method guesses as the completion the optimum of the convex program:

$$\begin{aligned} \min_X \quad & \|X\|_* \\ \text{s.t.} \quad & X_{ij} = M_{ij} \text{ for } (i,j) \in \Omega. \end{aligned} \tag{1}$$

where the "nuclear norm" $\|\cdot\|_*$ of a matrix is the sum of its singular values[1]. Throughout, we use the standard notation $f(n) = \Theta(g(n))$ to mean that $cg(n) \le f(n) \le Cg(n)$ for some positive constants c, C, where $n := \max\{n_1, n_2\}$.

We focus on the setting where matrix entries are revealed from an underlying probability distribution. To introduce the distribution of interest, we first need a definition.

Definition 3.1. *For an $n_1 \times n_2$ real-valued matrix M of rank r with SVD given by $U\Sigma V^\top$, the* **local coherences**[2] *– μ_i for any row i, and ν_j for any column j - are defined by the following relations*

$$\begin{aligned} \left\|U^\top e_i\right\| &= \sqrt{\frac{\mu_i r}{n_1}} \quad , \quad i = 1, \dots, n_1 \\ \left\|V^\top e_j\right\| &= \sqrt{\frac{\nu_j r}{n_2}} \quad , \quad j = 1, \dots, n_2. \end{aligned} \tag{2}$$

Note that the μ_i, ν_js are non-negative, and since U and V have orthonormal columns we always have $\sum_i \mu_i r/n_1 = \sum_j \nu_j r/n_2 = r$.

We are ready to state our main result, the theorem below.

Theorem 3.2. *Let $M = (M_{ij})$ be an $n_1 \times n_2$ matrix with local coherence parameters $\{\mu_i, \nu_j\}$, and suppose that its entries M_{ij} are observed only over a subset of elements $\Omega \subset [n_1] \times [n_2]$. There are universal constants $c_0, c_1, c_2 > 0$ such that if each element (i,j) is independently observed with probability p_{ij}, and p_{ij} satisfies*

$$\begin{aligned} p_{ij} &\ge \min\left\{ c_0 \frac{(\mu_i + \nu_j) r \log^2(n_1+n_2)}{\min\{n_1, n_2\}} \ , \ 1 \right\} \\ p_{ij} &\ge \frac{1}{\min\{n_1, n_2\}^{10}}, \end{aligned} \tag{3}$$

then M is the unique optimal solution to the nuclear norm minimization problem (1) with probability at least $1 - c_1(n_1+n_2)^{-c_2}$.

We will refer to the sampling strategy (3) as *local coherence* sampling. Note that the expected number of observed entries is $\sum_{i,j} p_{ij}$, and this satisfies

$$\begin{aligned} \sum_{i,j} p_{ij} &\ge \max\left\{ c_0 \frac{r \log^2(n_1+n_2)}{\min\{n_1,n_2\}} \sum_{i,j} (\mu_i + \nu_j), \sum_{i,j} \frac{1}{n^{10}} \right\} \\ &= 2c_0 \max\{n_1, n_2\}\, r \log^2(n_1+n_2), \end{aligned}$$

independent of the coherence, or indeed any other property, of the matrix. Hoeffding's inequality implies that the actual number of observed entries sharply concentrates around its expectation, leading to the following corollary:

Corollary 3.3. *Let $M = (M_{ij})$ be an $n_1 \times n_2$ matrix with local coherence parameters $\{\mu_i, \nu_j\}$. Draw a subset of its entries by local coherence sampling according to the procedure described in Theorem 3.2. There are universal constants $c_1', c_2' > 0$ such that the following holds with probability at least $1 - c_1'(n_1+n_2)^{-c_2'}$: the number m of revealed entries is bounded by*

$$m \le 3c_0 \max\{n_1, n_2\}\, r \log^2(n_1+n_2)$$

[1] This becomes the trace norm for positive-definite matrices. It is now well-recognized to be a convex surrogate for rank minimization.

[2] In the matrix sparsification literature (Drineas et al., 2012; Boutsidis et al., 2009) and beyond, the quantities $\left\|U^\top e_i\right\|^2$ and $\left\|V^\top e_j\right\|^2$ are referred to as the *leverage scores* of M.

and M is the unique optimal solution to the nuclear norm minimization program (1).

We now provide comments and discussion.

(A) Roughly speaking, the condition given in (3) ensures that entries in important rows/columns (indicated by large local coherences μ_i and ν_j) of the matrix should be observed more often. Note that Theorem 3.2 only stipulates that an *inequality* relation hold between p_{ij} and $\{\mu_i, \nu_j\}$. This allows for there to be some discrepancy between the sampling distribution and the local coherences. It also has the natural interpretation that the more the sampling distribution $\{p_{ij}\}$ is "aligned" to the local coherence pattern of the matrix, the fewer observations are needed.

(B) Sampling based on local coherences provides close to the optimal number of sampled elements required for exact recovery (when sampled with any distribution). In particular, assume $n_1 = n_2 = n$ and recall that the number of degrees of freedom of an $n \times n$ matrix with rank r is $2nr(1 - r/2n)$. Hence, regardless how the entries are sampled, a minimum of $\Theta(nr)$ entries is required to recover the matrix. Theorem 3.2 matches this lower bound, with an additional $O(\log^2(n))$ factor.

(C) Our work improves on existing results *even* in the case of uniform sampling and uniform incoherence. Recall that the original work of (Candès & Recht, 2009), and subsequent works (Candès & Tao, 2010; Recht, 2009; Gross, 2011) give recovery guarantees based on two parameters of the matrix M: a global *incoherence parameter* μ_0 which is a uniform bound on the (above-defined) local coherences – i.e. every $\mu_i \leq \mu_0$ and every $\nu_j \leq \mu_0$ – and a *joint incoherence parameter* μ_{str} defined by $\|UV^\top\|_\infty = \sqrt{\frac{r\mu_{str}}{n_1 n_2}}$. With these definitions, the current state of the art states that if the *uniform* sampling probability satisfies

$$p_{ij} \equiv p \geq c\frac{\max\{\mu_0, \mu_{str}\} r \log^2 n}{n} \quad ,$$

when $n_1 = n_2 = n$, where c is a constant, then M will be the unique optimum of (1) with high probability. A direct corollary of our work improves on this result, by removing the need for extra constraints on the joint incoherence; in particular, it is easy to see that our theorem implies that a uniform sampling probability of $p \geq c\frac{\mu_0 r \log^2 n}{n}$ – that is, with no μ_{str} – guarantees recovery of M with high probability. Note that in general, μ_{str} can be as high as $\mu_0 r$; our corollary thus removes this sub-optimal dependence on the rank and on the joint incoherence. This improvement was recently observed in (Chen, 2013).

Remark 3.4. Suppose $n_1 = n_2 = n$. If the column space of M is incoherent with $\max_i \mu_i \leq \mu_0$ and the row space is arbitrary, then one can randomly pick $\Theta(\mu_0 r \log n)$ rows of M and observe all their entries, and compute the local coherences of the space spanned by these rows. These parameters will be equal to the ν_j's of M with high probability. Based on these values, we can perform non-uniform sampling according to (3) and *exactly* recover M. This procedure does not require any prior knowledge about the local coherences of M. It uses a total of $\Theta(\mu_0 rn \log^2 n)$ samples. This improves on the $\Theta(\mu_0^2 r^2 n \log r)$ sample complexity in (Krishnamurthy & Singh, 2013), which is quadratic in $\mu_0 r$. We prove this remark in the supplement.

Algorithm 1 Two-phase sampling for coherent matrix completion

input Sampled matrix $\mathcal{P}_\Omega(M)$, rank parameter r, and m, β such that $|\Omega| = \beta m$.

1: Compute the rank-r SVD of $\mathcal{P}_\Omega(M)$, $\tilde{U}\tilde{\Sigma}\tilde{V}^\top$.

2: Estimate the local coherences by $\tilde{\mu_i} = \frac{n_1}{r}\left\|\tilde{U}^\top e_i\right\|^2$ and $\tilde{\nu_j} = \frac{n_2}{r}\left\|\tilde{V}^\top e_j\right\|^2$.

3: Generate a set of $(1-\beta)m$ new samples $\tilde{\Omega}$ distributed as $\tilde{p}_{ij} = \min\left\{c_0\frac{(\tilde{\mu_i}+\tilde{\nu_j})r\log^2(n_1+n_2)}{\min\{n_1,n_2\}}, 1\right\}$.

4: $\hat{M} = \arg\min \|X\|_* \text{ s.t } \mathcal{P}_{\Omega\cup\tilde{\Omega}}(X) = \mathcal{P}_{\Omega\cup\tilde{\Omega}}(M)$.

output Completed matrix $\hat{M}$.

4. A two-phase sampling procedure

We have seen that one can exactly recover an arbitrary $n \times n$ low-rank matrix using $\Theta(nr\log^2(n))$ entries if sampled in accordance with the local coherences. In practical applications of matrix completion, even when the user is free to choose how to sample the matrix entries, she likely will not be privy to the local coherence parameters $\{\mu_i, \nu_j\}$. In this section we propose a two-phase sampling procedure, described below and in Table 1, which assumes no a priori knowledge about the matrix coherence structure, yet is observed to be competitive with the "oracle" local coherence distribution (3).

Consider a total budget of m samples, and a set of sampled indices Ω such that $|\Omega| = \beta m$, where $\beta \in [0, 1]$. Let $\mathcal{P}_\Omega()$ be the sampling operator which maps the matrix entries not in Ω to 0. The first step of the algorithm is to take the rank-r SVD of $\mathcal{P}_\Omega(M)$, $\tilde{U}\tilde{\Sigma}\tilde{V}^\top$, where $\tilde{U}, \tilde{V} \in \mathbb{R}^{n\times r}$ and $\tilde{\Sigma} \in \mathbb{R}^{r\times r}$. Use the local coherences $\tilde{\mu_i}, \tilde{\nu_j}$ of $\tilde{U}, \tilde{V}$ respectively as estimates for the local coherences of M. Let

$$\tilde{p}_{ij} = \min\left\{c_0\frac{(\tilde{\mu_i}+\tilde{\nu_j})r\log^2(n_1+n_2)}{\min\{n_1,n_2\}}, 1\right\}. \quad (4)$$

Now generate the remaining $(1-\beta)m$ samples of matrix M according to this distribution (4). Let $\tilde{\Omega}$ denote the new set of samples. Using the combined set of samples $\mathcal{P}_{\Omega\cup\tilde{\Omega}}(M)$ as constraints, run the nuclear norm minimization program (1). Let $\hat{M}$ be the optimum of this program.

To understand the performance of the two-phase algorithm, assume that the initial set of $m_1 = \beta m$ samples $\mathcal{P}_\Omega(M)$ are generated uniformly at random. If the underlying matrix M is incoherent, then already the algorithm will recover M if $m_1 = \Theta(\max\{n_1, n_2\} r \log^2(n_1+n_2))$. On the other hand, if M is *highly* coherent, having almost all energy concentrated on just a few entries, then the estimated local coherences (4) from uniform sampling will be poor and hence the recovery algorithm suffers. Between these two extremes, there is reason to believe that the two-phase sampling procedure will provide a better estimate to the underlying matrix than if all m entries were sampled uniformly. Indeed, numerical experiments suggest that the two-phase procedure can indeed significantly outperform uniform sampling for completing coherent matrices.

4.1. Numerical experiments

We now analyze the performance of the two-phase sampling procedure outlined in Algorithm 1 through numerical experiments. For this, we consider rank-5 matrices of size 500×500 of the form $M = DUV^\top D$, where the entries of matrices U and V are i.i.d. Gaussian $\mathcal{N}(0,1)$ and D is a diagonal matrix with power-law decay, $D_{ii} = i^{-\alpha}, 1 \leq i \leq 500$. We refer to such constructions as *power-law* matrices. The parameter α adjusts the coherence of the matrix with $\alpha = 0$ being incoherent and $\alpha = 1$ corresponding to maximal coherence $\mu_0 = \Theta(n)$.

We normalize M to make $\|M\|_F = 1$. Figure 1 plots the number of samples required for successful recovery (y-axis) for different values of α (x-axis) using Algorithm 1 with initial samples Ω taken i.i.d. uniform. Successful recovery is defined as when at least 95% of trials have relative error in the Frobenius norm not exceeding 0.01. To put the results in perspective, we plot it in Figure 1 against the performance of pure uniform sampling, as well as other popular sampling distributions from the matrix sparsification literature (Achlioptas & Mcsherry, 2007; Achlioptas et al., 2013; Arora et al., 2006; Drineas & Zouzias, 2011), namely, in step 3 of the algorithm, sampling proportional to entry ($\tilde{p}_{ij} \propto |\tilde{M}_{ij}|$) and sampling proportional to entry squared ($\tilde{p}_{ij} \propto \tilde{M}_{ij}^2$), as opposed to sampling from the distribution (4). In all cases, the estimated matrix $\tilde{M}$ is constructed from the rank r SVD of $\mathcal{P}_\Omega(M)$, $\tilde{M} = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$. Performance of nuclear norm minimization using samples generated according to the "oracle" distribution (3) serves as baseline for the best possible recovery, as theoretically justified by Theorem 3.2. We use an Augmented Lagrangian Method (ALM) based solver by (Chen & Ganesh, 2009) to solve the convex optimization program (1).

Figure 1 suggests that the two-phase algorithm performs comparably to the theoretically optimal coherence-based distribution (3), despite not having access to the underlying local coherences, in the regime of mild to moderate coherence $\alpha \leq 0.7$. While the entrywise sampling strategies perform comparable for low values of α, the number of samples for successful recovery increases for $\alpha > 0.6$. Completion from purely uniformly sampled entries requires significantly more samples at higher values of α.

Choosing β: Recall that the parameter β in Algorithm 1 is the fraction of number of uniform samples to the total number of samples. Figure 2(a) plots the number of samples required for successful recovery (y-axis) as β (x-axis) varies from 0.1 to 1 for different values of α. $\beta = 1$ reduces to purely uniform sampling, and for small values of β, the local coherences estimated in (4) will be far from the actual local coherences. Then, as expected, the sample complexity goes up for β near 0 and $\beta = 1$. We find that setting $\beta \approx 2/3$ results in the lowest sample complexity. Surprisingly, even taking $\beta = 0.9$ as opposed to pure uniform sampling $\beta = 1$ results in a significant decrease in the sample complexity (Figure 2(b)). That is, even budgeting just a small fraction of samples to be drawn from the estimated local coherences can significantly improve the success rate in low-rank matrix recovery as long as the underlying matrix is not completely coherent. In applications like collaborative filtering, this would imply (assuming that the local coherences are smaller ($\alpha \leq 0.5$)) that incentivizing just a small fraction of users to rate a few selected movies according to the estimated local coherence distribution obtained by previous samples has the potential to greatly improve the quality of the recovered matrix of preferences.

In Figure 3 we compare the performance of the two-phase algorithm for different values of the matrix dimension n, and notice for each n a phase transition occurring at $\Theta(n \log(n))$ samples. In Figure 4 we consider the scenario where the samples are noisy and compare the performance of Algorithm 1 to uniform sampling and the theoretically-optimal local coherence sampling from Theorem 3.2. Specifically we assume that the samples are generated from $M + Z$ where Z is a Gaussian noise matrix. We consider two values for the noise $\sigma \stackrel{\text{def}}{=} \|Z\|_F / \|M\|_F$: $\sigma = 0.1$ and $\sigma = 0.2$. The figures plot error in Frobenius norm $\|M - \hat{M}\|_F$ (y-axis), vs total number of samples m (x-axis). These plots demonstrate the robustness of the algorithm to noise and once again show that sampling with estimated coherences can be as good as sampling with exact coherences for matrix recovery using nuclear norm minimization for $\alpha \leq 0.7$.

5. Weighted Nuclear Norm Minimization

Theorem 3.2 suggests that the more a set of observed entries are aligned with the local coherences of a matrix, the better will be the performance of nuclear norm minimization. Interestingly, Theorem 3.2 can be used in a reverse

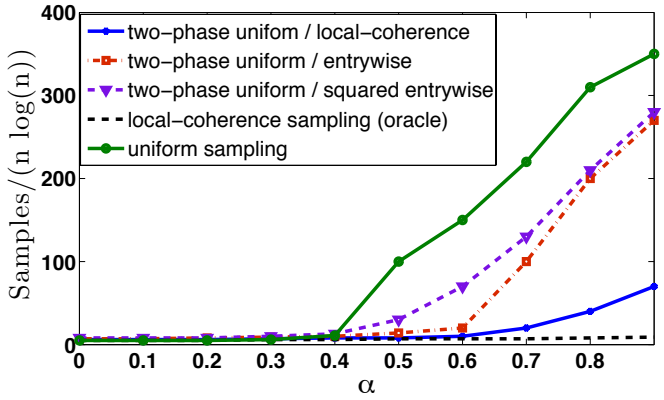


Figure 1. **Performance of Algorithm 1 for power-law matrices:** We consider rank-5 matrices of form $M = DUV^\top D$, where entries of the matrices U and V are generated from a Gaussian distribution $\mathcal{N}(0,1)$ and D is a diagonal matrix with $D_{ii} = \frac{1}{i^\alpha}$. Higher values of α correspond to higher coherence. The above simulations are run with two-phase parameter $\beta = 2/3$. Sampling (3) gives the best results of successful recovery using $10n\log(n)$ samples for all values of α in accordance with Theorem 3.2. Surprisingly, sampling according to (4) with estimated local coherences has almost the same sample complexity for $\alpha \leq 0.7$. Sampling proportional to entry and entry squared perform as well for low values of α, but their sample complexity increases quickly for $\alpha > 0.6$.

way: *one may adjust the local coherences to align with a given set of observations*. Here we demonstrate an application of this idea in quantifying the benefit of *weighted* nuclear norm minimization for non-uniform sampling.

In many applications of matrix completion, the revealed entries are given to us, and distributed non-uniformly among the rows and columns. As observed by (Salakhutdinov & Srebro, 2010), standard unweighted nuclear norm minimization (1) is inefficient in this setting. They propose to instead use weighted nuclear norm minimization:

$$\hat{X} = \arg\min_X \ \|RXC\|_* \quad \text{s.t. } X_{ij} = M_{ij}, \text{ for } (i,j) \in \Omega, \tag{5}$$

where $R = \text{diag}(R_1, R_2, \ldots, R_{n_1})$ and $C = \text{diag}(C_1, \ldots, C_{n_2})$ are diagonal weight matrices with positive diagonal entries.

We now provide a theoretical guarantee for this method, and quantify its advantage over unweighted nuclear norm minimization. Suppose M satisfies the standard incoherence condition $\max_{i,j}\{\mu_i, \nu_j\} \leq \mu_0$. Let $\lfloor x \rfloor$ denote the largest integer not exceeding x. Under this setting, we have the following (proved in the supplementary materials):

Theorem 5.1. *Without lost of generality, assume $R_1 \leq R_2 \leq \cdots \leq R_{n_1}$ and $C_1 \leq C_2 \leq \cdots \leq C_{n_2}$. There exist universal constants c_0, c_1, c_2 such that M is the unique optimum to (5) with probability at least $1 - c_1(n_1+n_2)^{-c_2}$ provided $p_{ij} \geq \frac{1}{\min\{n_1,n_2\}^{10}}$ and*

$$p_{ij} \geq c_0 \left(\frac{R_i^2}{\sum_{i'=1}^{\lfloor n_1/(\mu_0 r)\rfloor} R_{i'}^2} + \frac{C_j^2}{\sum_{j'=1}^{\lfloor n_2/(\mu_0 r)\rfloor} C_{j'}^2} \right) \log^2 n. \tag{6}$$

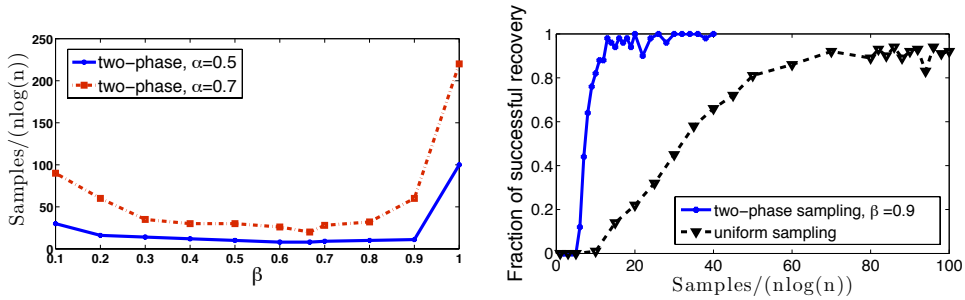
We prove this theorem by drawing a connection between the weighted nuclear norm and the local incoherence parameters (2). Define the scaled matrix $\bar{M} := RMC$. Observe that the program (5) is equivalent to first solving the following *unweighted* problem with scaled observations

$$\bar{X} = \arg\min_X \ \|X\|_* \quad \text{s.t. } X_{ij} = \bar{M}_{ij}, \text{ for } (i,j) \in \Omega, \tag{7}$$

and then setting $\hat{X} = R^{-1}\bar{X}C^{-1}$. In other words, through the weighted nuclear norm, we convert the problem of completing M to that of completing $\bar{M}$. Therefore, **if we can choose the weights R and C such that the local incoherence parameters of $\bar{M}$, denoted as $\{\bar{\mu}_i, \bar{\nu}_j\}$, are aligned with the non-uniform observations in a way that roughly satisfies condition (3), then we gain in sample complexity compared to the unweighted approach.** We now quantify this more precisely for a particular class of matrix completion problems.

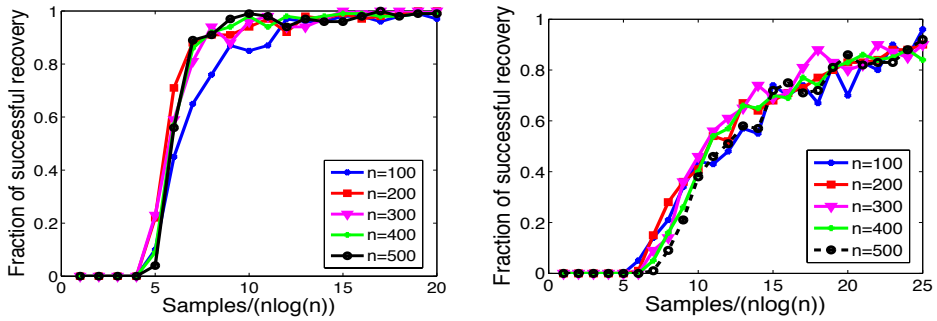
Comparison to unweighted nuclear norm. Suppose $n_1 = n_2 = n$ and the observation probabilities have a product form: $p_{ij} = p_i^r p_j^c$, with $p_1^r \leq p_2^r \leq \cdots \leq p_n^r$ and $p_1^c \leq p_2^c \leq \cdots \leq p_n^c$. If we choose $R_i = \sqrt{\frac{1}{n} p_i^r \sum_{j'} p_{j'}^c}$ and $C_j = \sqrt{\frac{1}{n} p_j^c \sum_{i'} p_{i'}^r}$ (which is suggested by the condition (6)), Theorem 5.1 asserts that the following is sufficient for recovery of M:

$$p_j^c \sum_{i=1}^{\lfloor n/(\mu_0 r)\rfloor} p_i^r \gtrsim \log^2 n, \forall j; \quad p_i^r \sum_{j=1}^{\lfloor n/(\mu_0 r)\rfloor} p_j^c \gtrsim \log^2 n, \forall i. \tag{8}$$



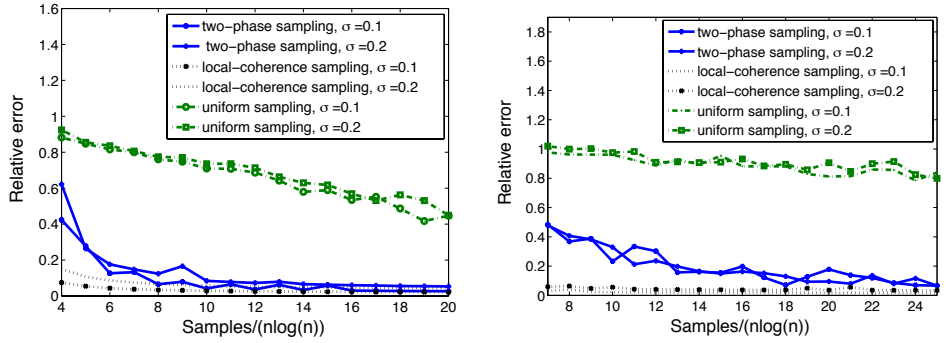
(a) **(b)**

Figure 2. We consider power-law matrices with parameter $\alpha = 0.5$ and $\alpha = 0.7$. **(a):** This plot shows that Algorithm 1 successfully recovers coherent low-rank matrices with fewest samples ($\approx 10n\log(n)$) when the proportion of initial samples drawn from the uniform distribution is in the range $\beta \in [0.5, 0.8]$. In particular, the sampling complexity is significantly lower than that for uniform sampling ($\beta = 1$). **(b):** Even by drawing 90% of the samples uniformly and using the estimated local coherences to sample the remaining 10% samples, one observes a marked improvement in the rate of recovery.



(a) **(b)**

Figure 3. **(a) & (b):Scaling of sample complexity of Algorithm 1 with n:** We consider power-law matrices (with $\alpha = 0.5$ in plot **(a)** and 0.7 in plot **(b)**). The plots suggest that the sample complexity of Algorithm 1 scales roughly as $\Theta(n\log(n))$.



(a) **(b)**

Figure 4. **(a) & (b):Performance of Algorithm 1 with noisy samples:** We consider power-law matrices (with $\alpha = 0.5$ in plot **(a)** and $\alpha = 0.7$ in plot **(b)**), perturbed by a Gaussian noise matrix Z with $\|Z\|_F/\|M\|_F = \sigma$. The plots consider two different noise levels, $\sigma = 0.1$ and $\sigma = 0.2$. We compare two-phase sampling (Algorithm 1) with $\beta = 2/3$, sampling from the exact local coherences, and uniform sampling. Algorithm 1 has error almost as low as the local-coherence sampling without requiring any a priori knowledge of the low-rank matrix, while uniform sampling suffers dramatically.

We compare this condition to that required by unweighted nuclear norm minimization: by Thm. 3.2, the latter requires

$$p_i^r p_j^c \gtrsim \frac{\mu_0 r}{n} \log^2 n, \quad \forall i,j.$$

That is, the weighted approach succeeds under much less restrictive conditions. In particular, the unweighted approach imposes a condition on the *least* sampled row and column, whereas condition (8) shows that the weighted approach can use the heavily sampled rows/columns to assist the less sampled. This benefit is most significant precisely when the observations are very non-uniform.

The weighted nuclear norm approach is shown to be empirically successful in (Salakhutdinov & Srebro, 2010). There they propose to weigh the rows (columns, resp.) by the square root of the corresponding row (column, resp) marginals, which coincides with the R and C chosen according to our theory in the last paragraph.

We remark that Theorem 5.1 is the first exact recovery guarantee for weighted nuclear norm minimization. It provides an explanation, complementary to those in (Salakhutdinov & Srebro, 2010; Foygel et al., 2011; Negahban & Wainwright, 2012), for why the weighted approach is advantageous over the unweighted approach for non-uniform observations. It also serves as a testament to the power of Theorem 3.2 as a general result on the relationship between sampling and local coherence.

6. Proof Outline for Theorem 3.2

The proof proceeds by constructing a dual certificate Y that obeys certain sub-gradient optimality conditions and certifies the optimality of M to (1). One of the major differences between our proof and existing ones is in validating one of the optimality conditions, namely, that $\|Y\|$ is small. In previous work, this is done by bounding $\|Y\|$ by $\|Y'\|_\infty := \sum_{i,j} |Y'_{ij}|$ for a certain matrix Y', which eventually leads to the standard incoherence conditions. Here, we derive a new bound using the *weighted*-$\ell_{\infty,2}$ norm of Y', which is the maximum of the weighted row and column norms of Y', with the weights depending on the local coherences μ_i and ν_j. We turn to the details below.

Define the projections $P_T Z := UU^\top Z + ZVV^\top - UU^\top VZZ^\top$ and $P_{T^\perp} Z := Z - P_T Z$, and let $R_\Omega Z$ be the matrix with $(R_\Omega Z)_{ij} = Z_{ij}/p_{ij}$ if $(i,j) \in \Omega$ and zero otherwise. As usual, $\|Z\|_F$ and $\|Z\|$ are the Frobenius norm and spectral norm of the matrix Z, and $\|\mathcal{A}\|_{op}$ is the operator norm of the mapping $\mathcal{A}$. Using standard convex analysis, we show that M is the unique optimum to (1) if

1. $\|P_T R_\Omega P_T - P_T\|_{op} \le \frac{1}{2}$, and
2. there exists some Y obeying (a) $Y_{ij} = 0, \forall (i,j) \notin \Omega$, (b) $\left\|P_T Y - UV^\top\right\|_F \le \frac{1}{4n^5}$, and (c) $\|P_{T^\perp} Y\| \le \frac{1}{2}$.

We proceed to show that condition 1 above holds with high probability (w.h.p.) assuming only the local bounds (3) on sampling and incoherence. We then construct Y using the Golfing Scheme (Gross, 2011), setting $W_0 := 0$,

$$W_k := W_{k-1} + R_{\Omega_k} P_T(UV^\top - P_T W_{k-1}), k \in [k_0],$$

and $Y = W_{k_0}$, where the Ω_k's are $k_0 := 20 \log n$ i.i.d. random index sets with $\mathbb{P}((i,j) \in \Omega_k) = 1 - (1 - p_{ij})^{1/k_0}$ and R_{Ω_k} is defined analogously to R_Ω. Y satisfies condition 2(a) above. Setting $\Delta_k = UV^\top - P_T W_k$, we verify

$$\|\Delta_{k_0}\|_F \le \left(\prod_k \|P_T - P_T R_{\Omega_k} P_T\|_{op}\right) \left\|UV^\top\right\|_F,$$

which implies the condition 2(b) using the condition 1.

It remains to validate the condition 2(c), which is the most innovative part of our proof. We need the following definitions of weighted $\ell_{\infty,2}$ and ℓ_∞ norms

$$\|Z\|_{\mu(\infty,2)} := \max_{i,j} \left\{ \sqrt{\frac{n}{\mu_i r} \sum_b Z_{ib}^2}, \sqrt{\frac{n}{\nu_j r} \sum_a Z_{aj}^2} \right\},$$

$$\|Z\|_{\mu(\infty)} := \max_{i,j} |Z_{ij}| \sqrt{\frac{n}{\mu_i r}} \sqrt{\frac{n}{\nu_j r}}.$$

We show, crucially, that these norms have the following concentration properties

$$\|(R_\Omega - I) Z\| \le \frac{c}{\sqrt{c_0}} \left(\|Z\|_{\mu(\infty)} + \|Z\|_{\mu(\infty,2)}\right),$$

$$\|(P_T R_\Omega - P_T) Z\|_{\mu(\infty,2)} \le \frac{1}{2} \left(\|Z\|_{\mu(\infty)} + \|Z\|_{\mu(\infty,2)}\right),$$

$$\|(P_T R_\Omega - P_T) Z\|_{\mu(\infty)} \le \frac{1}{2} \|Z\|_{\mu(\infty)},$$

which hold w.h.p. for a fixed Z. Using the first inequality above, we can obtain

$$\|P_{T^\perp}(Y)\| \le \frac{c}{\sqrt{c_0}} \sum_{k=1}^{k_0} \left(\|\Delta_{k-1}\|_{\mu(\infty)} + \|\Delta_{k-1}\|_{\mu(\infty,2)}\right).$$

We then apply the next two inequalities to show

$$\|\Delta_k\|_{\mu(\infty)} \le (1/2)^k \left\|UV^\top\right\|_{\mu(\infty)},$$

$$\|\Delta_k\|_{\mu(\infty,2)} \le (1/2)^k \left(2k \left\|UV^\top\right\|_{\mu(\infty)} + \|UV\|_{\mu(\infty,2)}\right).$$

for each k. The theorem follows from combining the last three display equations and expressing $\|UV^\top\|_{\mu(\infty,2)}$ and $\|UV^\top\|_{\mu(\infty)}$ in terms of $\{\mu_i, \nu_j\}$.

Acknowledgements

We would like to thank Petros Drineas, Michael Mahoney and Aarti Singh for helpful discussions. R. Ward was supported by an NSF CAREER award, AFOSR Young Investigator Program award, and ONR Grant N00014-12-1-0743.

References

Achlioptas, D. and Mcsherry, F. Fast computation of low-rank matrix approximations. *J. ACM*, 54(2):9, 2007.

Achlioptas, D., Karnin, Z., and Liberty, E. Matrix entry-wise sampling: Simple is best. http://cs-www.cs.yale.edu/homes/el327/papers/matrixSampling.pdf, 2013.

Arora, S., Hazan, E., and Kale, S. A fast random sampling algorithm for sparsifying matrices. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pp. 272–279. Springer, 2006.

Boutsidis, C., Mahoney, M., and Drineas, P. An improved approximation algorithm for the column subset selection problem. In *SODA*, pp. 968–977, 2009.

Burq, N., Dyatlov, S., Ward, R., and Zworski, M. Weighted eigenfunction estimates with applications to compressed sensing. *SIAM J. Math. Anal.*, 44(5):3481–3501, 2012.

Cai, J., Candès, E., and Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM J. Optimiz.*, 20(4):1956–1982, 2010.

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

Candès, E. and Recht, B. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9, 2009.

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

Candès, E., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

Chandrasekaran, V., Sanghavi, S., Parrilo, P., and Willsky, A. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Chatterjee, S. and Hadi, A. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986.

Chen, M. and Ganesh, A. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, October 2009. URL http://perception.csl.illinois.edu/matrix-rank/sample_code.html.

Chen, Y. Incoherence-optimal matrix completion. *arXiv preprint arXiv:1310.0154*, 2013.

Chen, Y., Jalali, A., Sanghavi, S., and Caramanis, C. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7), 2013.

Drineas, P. and Zouzias, A. A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality. *Information Processing Letters*, 111(8):385–389, 2011.

Drineas, P., Magdon-Ismail, M., Mahoney, M., and Woodruff, D. Fast approximation of matrix coherence and statistical leverage. *JMLR*, 13:3475, 2012.

Foygel, R., Salakhutdinov, R., Shamir, O., and Srebro, N. Learning with the weighted trace-norm under arbitrary sampling distributions. *arXiv:1106.4251*, 2011.

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

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. *arXiv preprint arXiv:1212.0467*, 2012.

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

Krahmer, F. and Ward, R. Beyond incoherence: Stable and robust sampling strategies for compressive imaging. *arXiv preprint arXiv:1210.2380*, 2012.

Krishnamurthy, A. and Singh, A. Low-rank matrix and tensor completion via adaptive sampling. *arXiv preprint arXiv:1304.4672v2*, 2013.

Mahoney, M. Randomized algorithms for matrices & data. *Foundations & Trends in Machine learning*, 3(2), 2011.

Negahban, S. and Wainwright, M. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *JMLR*, 98888:1665–1697, 2012.

Rauhut, H. and Ward, R. Sparse Legendre expansions via l1-minimization. *Journal of Approximation Theory*, 164 (5):517–533, 2012.

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

Salakhutdinov, R. and Srebro, N. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. *arXiv preprint arXiv:1002.2780*, 2010.

Spielman, D. and Srivastava, N. Graph sparsification by effective resistances. *SIAM J. Comp.*, 40(6):1913, 2011.

Tropp, J. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

Yang, X. and Karniadakis, G. Reweighted l1 minimization method for stochastic elliptic differential equations. *Journal of Computational Physics*, 2013.