Journal of Machine Learning Research 16 (2015) 913-960 Submitted 10/13; Revised 9/14; Published 4/15

Distributed Matrix Completion and Robust Factorization

Lester Mackey† LMACKEY@STANFORD.EDU
Stanford University
Department of Statistics
390 Serra Mall
Stanford, CA 94305

Ameet Talwalkar† ATALWALKAR@GMAIL.COM
University of California, Los Angeles
Computer Science Department
4732 Boelter Hall
Los Angeles, CA 90095

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
University of California, Berkeley
Department of Electrical Engineering and Computer Science and Department of Statistics
465 Soda Hall
Berkeley, CA 94720

Editor: Nathan Srebro
† These authors contributed equally.

Abstract

If learning methods are to scale to the massive sizes of modern data sets, it is essential for the field of machine learning to embrace parallel and distributed computing. Inspired by the recent development of matrix factorization methods with rich theory but poor computational complexity and by the relative ease of mapping matrices onto distributed architectures, we introduce a scalable divide-and-conquer framework for noisy matrix factorization. We present a thorough theoretical analysis of this framework in which we characterize the statistical errors introduced by the "divide" step and control their magnitude in the "conquer" step, so that the overall algorithm enjoys high-probability estimation guarantees comparable to those of its base algorithm. We also present experiments in collaborative filtering and video background modeling that demonstrate the near-linear to superlinear speed-ups attainable with this approach.

Keywords: collaborative filtering, divide-and-conquer, matrix completion, matrix factorization, parallel and distributed algorithms, randomized algorithms, robust matrix factorization, video surveillance

1. Introduction

The scale of modern scientific and technological data sets poses major new challenges for computational and statistical science. Data analyses and learning algorithms suitable for modest-sized data sets are often entirely infeasible for the terabyte and petabyte data sets that are fast becoming the norm. There are two basic responses to this challenge. One response is to abandon algorithms that have superlinear complexity, focusing attention on simplified algorithms that—in the setting of massive data—may achieve satisfactory results

because of the statistical strength of the data. While this is a reasonable research strategy, it requires developing suites of algorithms of varying computational complexity for each inferential task and calibrating statistical and computational efficiencies. There are many open problems that need to be solved if such an effort is to bear fruit.

The other response to the massive data problem is to retain existing algorithms but to apply them to subsets of the data. To obtain useful results under this approach, one embraces parallel and distributed computing architectures, applying existing base algorithms to multiple subsets of the data in parallel and then combining the results. Such a divide-and-conquer methodology has two main virtues: (1) it builds directly on algorithms that have proven their value at smaller scales and that often have strong theoretical guarantees, and (2) it requires little in the way of new algorithmic development. The major challenge, however, is in preserving the theoretical guarantees of the base algorithm once one embeds the algorithm in a computationally-motivated divide-and-conquer procedure. Indeed, the theoretical guarantees often refer to subtle statistical properties of the data-generating mechanism (e.g., sparsity, information spread, and near low-rankedness). These may or may not be retained under the "divide" step of a putative divide-and-conquer solution. In fact, we generally would expect subsampling operations to damage the relevant statistical structures. Even if these properties are preserved, we face the difficulty of combining the intermediary results of the "divide" step into a final consilient solution to the original problem. The question, therefore, is whether we can design divide-and-conquer algorithms that manage the tradeoffs relating these statistical properties to the computational degrees of freedom such that the overall algorithm provides a scalable solution that retains the theoretical guarantees of the base algorithm.

In this paper,[1] we explore this issue in the context of an important class of machine learning algorithms—the matrix factorization algorithms underlying a wide variety of practical applications, including collaborative filtering for recommender systems , e.g., Koren et al. (2009) and the references therein, link prediction for social networks (Hoff, 2005), click prediction for web search (Das et al., 2007), video surveillance (Candès et al., 2011), graphical model selection (Chandrasekaran et al., 2009), document modeling (Min et al., 2010), and image alignment (Peng et al., 2010). We focus on two instances of the general matrix factorization problem: noisy matrix completion (Candès and Plan, 2010), where the goal is to recover a low-rank matrix from a small subset of noisy entries, and noisy robust matrix factorization (Candès et al., 2011; Chandrasekaran et al., 2009), where the aim is to recover a low-rank matrix from corruption by noise and outliers of arbitrary magnitude. These two classes of matrix factorization problems have attracted significant interest in the research community.

Various approaches have been proposed for scalable noisy matrix factorization problems, in particular for noisy matrix completion, though the vast majority tackle rank-constrained non-convex formulations of these problems with no assurance of finding optimal solutions (Zhou et al., 2008; Gemulla et al., 2011; Recht and Ré, 2011; F. Niu et al., 2011; Yu et al., 2012). In contrast, convex formulations of noisy matrix factorization relying on the nuclear norm have been shown to admit strong theoretical estimation guarantees (Agarwal et al., 2011; Candès et al., 2011; Candès and Plan, 2010; Negahban and Wainwright, 2012),

1. A preliminary form of this work appears in Mackey et al. (2011).

and a variety of algorithms (e.g., Lin et al., 2009b; Ma et al., 2011; Toh and Yun, 2010) have been developed for solving both matrix completion and robust matrix factorization via convex relaxation. Unfortunately, however, all of these methods are inherently sequential, and all rely on the repeated and costly computation of truncated singular value decompositions (SVDs), factors that severely limit the scalability of the algorithms. Moreover, previous attempts at reducing this computational burden have introduced approximations without theoretical justification (Mu et al., 2011).

To address this key problem of noisy matrix factorization in a scalable and theoretically sound manner, we propose a divide-and-conquer framework for large-scale matrix factorization. Our framework, entitled Divide-Factor-Combine (DFC), randomly divides the original matrix factorization task into cheaper subproblems, solves those subproblems in parallel using a base matrix factorization algorithm for nuclear norm regularized formulations, and combines the solutions to the subproblems using efficient techniques from randomized matrix approximation. We develop a thoroughgoing theoretical analysis for the DFC framework, linking statistical properties of the underlying matrix to computational choices in the algorithms and thereby providing conditions under which statistical estimation of the underlying matrix is possible. We also present experimental results for several DFC variants demonstrating that DFC can provide near-linear to superlinear speed-ups in practice. Indeed, DFC naturally handles massive data sets that are too large to fit on a single machine, as DFC's minimal communication footprint is particularly well-suited for distributed computing environments.

The remainder of the paper is organized as follows. In Section 2, we define the setting of noisy matrix factorization and introduce the components of the DFC framework. Secs. 3, 4, and 5 present our theoretical analysis of DFC, along with a new analysis of convex noisy matrix completion and a novel characterization of randomized matrix approximation algorithms. To illustrate the practical speed-up and robustness of DFC, we present experimental results on collaborative filtering, video background modeling, and simulated data in Section 6. Finally, we conclude in Section 7.

Notation: For a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, we define $\mathbf{M}_{(i)}$ as the ith row vector, $\mathbf{M}^{(j)}$ as the jth column vector, and $\mathbf{M}_{ij}$ as the ijth entry. If $\operatorname{rank}(\mathbf{M}) = r$, we write the compact singular value decomposition (SVD) of $\mathbf{M}$ as $\mathbf{U}_M \mathbf{\Sigma}_M \mathbf{V}_M^\top$, where $\mathbf{\Sigma}_M$ is diagonal and contains the r non-zero singular values of $\mathbf{M}$, and $\mathbf{U}_M \in \mathbb{R}^{m \times r}$ and $\mathbf{V}_M \in \mathbb{R}^{n \times r}$ are the corresponding left and right singular vectors of $\mathbf{M}$. We define $\mathbf{M}^+ = \mathbf{V}_M \mathbf{\Sigma}_M^{-1} \mathbf{U}_M^\top$ as the Moore-Penrose pseudoinverse of $\mathbf{M}$ and $\mathbf{P}_M = \mathbf{M}\mathbf{M}^+$ as the orthogonal projection onto the column space of $\mathbf{M}$. We let $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|_*$ respectively denote the spectral, Frobenius, and nuclear norms of a matrix, $\|\cdot\|_\infty$ denote the maximum entry of a matrix, and $\|\cdot\|$ represent the ℓ_2 norm of a vector.

2. The Divide-Factor-Combine Framework

In this section, we present a general divide-and-conquer framework for scalable noisy matrix factorization. We begin by defining the problem setting of interest.

2.1 Noisy Matrix Factorization (MF)

In the setting of noisy matrix factorization, we observe a subset of the entries of a matrix $\mathbf{M} = \mathbf{L}_0 + \mathbf{S}_0 + \mathbf{Z}_0 \in \mathbb{R}^{m \times n}$, where $\mathbf{L}_0$ has rank $r \ll m, n$, $\mathbf{S}_0$ represents a sparse matrix of outliers of arbitrary magnitude, and $\mathbf{Z}_0$ is a dense noise matrix. We let Ω represent the locations of the observed entries and $\mathcal{P}_\Omega$ be the orthogonal projection onto the space of $m \times n$ matrices with support Ω, so that

$$(\mathcal{P}_\Omega(\mathbf{M}))_{ij} = \mathbf{M}_{ij}, \text{ if } (i,j) \in \Omega \quad \text{and} \quad (\mathcal{P}_\Omega(\mathbf{M}))_{ij} = 0 \text{ otherwise.}^2$$

Our goal is to estimate the low-rank matrix $\mathbf{L}_0$ from $\mathcal{P}_\Omega(\mathbf{M})$ with error proportional to the noise level $\Delta \triangleq \|\mathbf{Z}_0\|_F$. We will focus on two specific instances of this general problem:

- **Noisy Matrix Completion (MC):** $s \triangleq |\Omega|$ entries of $\mathbf{M}$ are revealed uniformly without replacement, along with their locations. There are no outliers, so that $\mathbf{S}_0$ is identically zero.
- **Noisy Robust Matrix Factorization (RMF):** $\mathbf{S}_0$ is identically zero save for s outlier entries of arbitrary magnitude with unknown locations distributed uniformly without replacement. All entries of $\mathbf{M}$ are observed, so that $\mathcal{P}_\Omega(\mathbf{M}) = \mathbf{M}$.

2.2 Divide-Factor-Combine

The Divide-Factor-Combine (DFC) framework divides the expensive task of matrix factorization into smaller subproblems, executes those subproblems in parallel, and then efficiently combines the results into a final low-rank estimate of $\mathbf{L}_0$. We highlight three variants of this general framework in Algorithms 1, 2, and 3. These algorithms, which we refer to as DFC-Proj, DFC-RP, and DFC-Nys, differ in their strategies for division and recombination but adhere to a common pattern of three simple steps:

(D step) Divide input matrix into submatrices: DFC-Proj and DFC-RP randomly partition $\mathcal{P}_\Omega(\mathbf{M})$ into t l-column submatrices, $\{\mathcal{P}_\Omega(\mathbf{C}_1), \ldots, \mathcal{P}_\Omega(\mathbf{C}_t)\}$,[3] while DFC-Nys selects an l-column submatrix, $\mathcal{P}_\Omega(\mathbf{C})$, and a d-row submatrix, $\mathcal{P}_\Omega(\mathbf{R})$, uniformly at random.

(F step) Factor each submatrix in parallel using any base MF algorithm: DFC-Proj and DFC-RP perform t parallel submatrix factorizations, while DFC-Nys performs two such parallel factorizations. Standard base MF algorithms output the following low-rank approximations: $\{\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t\}$ for DFC-Proj and DFC-RP; $\hat{\mathbf{C}}$ and $\hat{\mathbf{R}}$ for DFC-Nys. All matrices are retained in factored form.

(C step) Combine submatrix estimates: DFC-Proj generates a final low-rank estimate $\hat{\mathbf{L}}^{proj}$ by projecting $[\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t]$ onto the column space of $\hat{\mathbf{C}}_1$, DFC-RP uses random projection to compute a rank-k estimate $\hat{\mathbf{L}}^{rp}$ of $[\hat{\mathbf{C}}_1 \cdots \hat{\mathbf{C}}_t]$ where k is the median rank of the returned subproblem estimates, and DFC-Nys forms the low-rank

2. When $\mathbf{Q}$ is a submatrix of $\mathbf{M}$ we abuse notation and let $\mathcal{P}_\Omega(\mathbf{Q})$ be the corresponding submatrix of $\mathcal{P}_\Omega(\mathbf{M})$.
3. For ease of discussion, we assume that t evenly divides n so that $l = n/t$. In general, $\mathcal{P}_\Omega(\mathbf{M})$ can always be partitioned into t submatrices, each with either $\lfloor n/t \rfloor$ or $\lceil n/t \rceil$ columns.

Algorithm 1 DFC-Proj

Input: $\mathcal{P}_\Omega(\mathbf{M})$, t
$\{\mathcal{P}_\Omega(\mathbf{C}_i)\}_{1\leq i\leq t} = \textsc{SampCol}(\mathcal{P}_\Omega(\mathbf{M}),\, t)$
do in parallel
$\hat{\mathbf{C}}_1 = \textsc{Base-MF-Alg}(\mathcal{P}_\Omega(\mathbf{C}_1))$
$\vdots$
$\hat{\mathbf{C}}_t = \textsc{Base-MF-Alg}(\mathcal{P}_\Omega(\mathbf{C}_t))$
end do
$\hat{\mathbf{L}}^{proj} = \textsc{ColProjection}(\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t)$

Algorithm 2 DFC-RP

Input: $\mathcal{P}_\Omega(\mathbf{M})$, t
$\{\mathcal{P}_\Omega(\mathbf{C}_i)\}_{1\leq i\leq t} = \textsc{SampCol}(\mathcal{P}_\Omega(\mathbf{M}),\, t)$
do in parallel
$\hat{\mathbf{C}}_1 = \textsc{Base-MF-Alg}(\mathcal{P}_\Omega(\mathbf{C}_1))$
$\vdots$
$\hat{\mathbf{C}}_t = \textsc{Base-MF-Alg}(\mathcal{P}_\Omega(\mathbf{C}_t))$
end do
$k = \text{median}_{i\in\{1\ldots t\}}\big(\text{rank}(\hat{\mathbf{C}}_i)\big)$
$\hat{\mathbf{L}}^{proj} = \textsc{RandProjection}(\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t, k)$

Algorithm 3 DFC-Nys

Input: $\mathcal{P}_\Omega(\mathbf{M})$, l, d
$\mathcal{P}_\Omega(\mathbf{C}), \mathcal{P}_\Omega(\mathbf{R}) = \textsc{SampColRow}(\mathcal{P}_\Omega(\mathbf{M}),\, l,\, d)$
do in parallel
$\hat{\mathbf{C}} = \textsc{Base-MF-Alg}(\mathcal{P}_\Omega(\mathbf{C}))$
$\hat{\mathbf{R}} = \textsc{Base-MF-Alg}(\mathcal{P}_\Omega(\mathbf{R}))$
end do
$\hat{\mathbf{L}}^{nys} = \textsc{GenNyström}(\hat{\mathbf{C}},\, \hat{\mathbf{R}})$

estimate $\hat{\mathbf{L}}^{nys}$ from $\hat{\mathbf{C}}$ and $\hat{\mathbf{R}}$ via the generalized Nyström method. These matrix approximation techniques are described in more detail in Section 2.3.

2.3 Randomized Matrix Approximations

Underlying the C step of each DFC algorithm is a method for generating randomized low-rank approximations to an arbitrary matrix $\mathbf{M}$.

Column Projection: DFC-Proj (Algorithm 1) uses the column projection method of Frieze et al. (1998). Suppose that $\mathbf{C}$ is a matrix of l columns sampled uniformly and without replacement from the columns of $\mathbf{M}$. Then, column projection generates a "matrix projection" approximation (Kumar et al., 2009a) of $\mathbf{M}$ via

$$\mathbf{L}^{proj} = \mathbf{C}\mathbf{C}^{+}\mathbf{M} = \mathbf{U}_C\mathbf{U}_C^{\top}\mathbf{M}.$$

In practice, we do not reconstruct $\mathbf{L}^{proj}$ but rather maintain low-rank factors, e.g., $\mathbf{U}_C$ and $\mathbf{U}_C^{\top}\mathbf{M}$.

Random Projection: The celebrated result of Johnson and Lindenstrauss (1984) shows that random low-dimensional embeddings preserve Euclidean geometry. Inspired by this result, several random projection algorithms (e.g., Papadimitriou et al., 1998; Liberty, 2009; Rokhlin et al., 2009) have been introduced for approximating a matrix by projecting it onto a random low-dimensional subspace (see Halko et al. 2011 for further discussion). DFC-RP (Algorithm 2) uses such a random projection method due to Halko et al. (2011). Given a

target low-rank parameter k, let $\mathbf{G}$ be an $n \times (k+p)$ standard Gaussian matrix $\mathbf{G}$, where p is an oversampling parameter. Next, let $\mathbf{Y} = (\mathbf{M}\mathbf{M}^\top)^q\mathbf{M}\mathbf{G}$, and define $\mathbf{Q} \in \mathbb{R}^{m\times k}$ as the top k left singular vectors of $\mathbf{Y}$. The random projection approximation of $\mathbf{M}$ is then given by

$$\mathbf{L}^{rp} = \mathbf{Q}\mathbf{Q}^+\mathbf{M}.$$

We work with an implementation (Tygert, 2009) of a numerically stable variant of this algorithm described in Algorithm 4.4 of Halko et al. (2011). Moreover, the parameters p and q are typically set to small positive constants (Tygert, 2009; Halko et al., 2011), and we set $p = 5$ and $q = 2$.

Generalized Nyström Method: The Nyström method was developed for the discretization of integral equations (Nyström, 1930) and has since been used to speed up large-scale learning applications involving symmetric positive semidefinite matrices (Williams and Seeger, 2000). DFC-Nys (Algorithm 3) makes use of a generalization of the Nyström method for arbitrary real matrices (Goreinov et al., 1997). Suppose that $\mathbf{C}$ consists of l columns of $\mathbf{M}$, sampled uniformly without replacement, and that $\mathbf{R}$ consists of d rows of $\mathbf{M}$, independently sampled uniformly and without replacement. Let $\mathbf{W}$ be the $d \times l$ matrix formed by sampling the corresponding rows of $\mathbf{C}$.[4] Then, the generalized Nyström method computes a "spectral reconstruction" approximation (Kumar et al., 2009a) of $\mathbf{M}$ via

$$\mathbf{L}^{nys} = \mathbf{C}\mathbf{W}^+\mathbf{R} = \mathbf{C}\mathbf{V}_W\mathbf{\Sigma}_W^+\mathbf{U}_W^\top\mathbf{R}\,.$$

As with $\mathbf{M}^{proj}$, we store low-rank factors of $\mathbf{L}^{nys}$, such as $\mathbf{C}\mathbf{V}_W\mathbf{\Sigma}_W^+$ and $\mathbf{U}_W^\top\mathbf{R}$.

Algorithm	**Factorization (Per Iteration)**		**Combine Step**	
	Serial	**Parallel**	**Serial**	**Parallel**
Base Alg	$O(mn\hat{k})$	$O(mn\hat{k})$	-	-
DFC-Proj	$O(tml\hat{k})$	$O(ml\hat{k})$	$O(tm\hat{k}^2)$	$O(m\hat{k}^2)$
DFC-RP	$O(tml\hat{k})$	$O(ml\hat{k})$	$O(tm\hat{k}^2 + n\hat{k})$	$O(m\hat{k}^2 + tm\hat{k} + n\hat{k})$
DFC-Nys	$O((ml + nd)\hat{k})$	$O(\max(ml, nd)\hat{k})$	$O(m\hat{k}^2)$	$O(m\hat{k}^2)$

Table 1: Summary of running time complexity of DFC variants in contrast to many standard start-of-the-art MF algorithms. This running time analysis assumes that $l \leq m \leq n$ and that all low-rank matrices considered have rank $\hat{k}$. See Section 2.4 for a more detailed analysis.

2.4 Running Time of DFC

Many state-of-the-art MF algorithms have $\Omega(mnk_M)$ per-iteration time complexity due to the rank-k_M truncated SVD performed on each iteration. DFC significantly reduces the per-iteration complexity to $O(mlk_{C_i})$ time for $\mathbf{C}_i$ (or $\mathbf{C}$) and $O(ndk_R)$ time for $\mathbf{R}$. The cost of combining the submatrix estimates is even smaller when using column projection or the generalized Nyström method, since the outputs of standard MF algorithms are returned

4. This choice is arbitrary: $\mathbf{W}$ could also be defined as a submatrix of $\mathbf{R}$.

in factored form. Indeed, if we define $k' \triangleq \max_i k_{C_i}$, then the column projection step of DFC-Proj requires only $O(mk'^2 + lk'^2)$ time: $O(mk'^2 + lk'^2)$ time for the pseudoinversion of $\hat{\mathbf{C}}_1$ and $O(mk'^2 + lk'^2)$ time for matrix multiplication with each $\hat{\mathbf{C}}_i$ in parallel. Similarly, the generalized Nyström step of DFC-Nys requires only $O(l\bar{k}^2 + d\bar{k}^2 + \min(m,n)\bar{k}^2)$ time, where $\bar{k} \triangleq \max(k_C, k_R)$.

DFC-RP also benefits from the factored form of the outputs of standard MF algorithms. Assuming that p and q are positive constants, the random projection step of DFC-RP requires $O(mkt + mkk' + lkk' + nk)$ time where k is the low-rank parameter of $\mathbf{Q}$: $O(nk)$ time to generate $\mathbf{G}$, $O(mkk' + lkk' + mkt)$ to compute $\mathbf{Y}$ in parallel, $O(mk^2)$ to compute the SVD of $\mathbf{Y}$, and $O(mk'^2 + lk'^2)$ time for matrix multiplication with each $\hat{\mathbf{C}}_i$ in parallel in the final projection step. Note that the running time of the random projection step depends on t (even when executed in parallel) and thus has a larger complexity than the column projection and generalized Nyström variants. Nevertheless, the random projection step need be performed only once and thus yields a significant savings over the repeated computation of SVDs required by typical base algorithms.

A summary of these running times is presented in Table 1.

2.5 Ensemble Methods

Ensemble methods have been shown to improve performance of matrix approximation algorithms, while straightforwardly leveraging the parallelism of modern many-core and distributed architectures (Kumar et al., 2009b). As such, we propose ensemble variants of the DFC algorithms that demonstrably reduce estimation error while introducing a negligible cost to the parallel running time. For DFC-Proj-Ens, rather than projecting only onto the column space of $\hat{\mathbf{C}}_1$, we project $[\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t]$ onto the column space of each $\hat{\mathbf{C}}_i$ in parallel and then average the t resulting low-rank approximations. For DFC-RP-Ens, rather than projecting only onto a column space derived from a single random matrix $\mathbf{G}$, we project $[\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t]$ onto t column spaces derived from t random matrices in parallel and then average the t resulting low-rank approximations. For DFC-Nys-Ens, we choose a random d-row submatrix $\mathcal{P}_\Omega(\mathbf{R})$ as in DFC-Nys and independently partition the columns of $\mathcal{P}_\Omega(\mathbf{M})$ into $\{\mathcal{P}_\Omega(\mathbf{C}_1), \ldots, \mathcal{P}_\Omega(\mathbf{C}_t)\}$ as in DFC-Proj and DFC-RP. After running the base MF algorithm on each submatrix, we apply the generalized Nyström method to each $(\hat{\mathbf{C}}_i, \hat{\mathbf{R}})$ pair in parallel and average the t resulting low-rank approximations. Section 6 highlights the empirical effectiveness of ensembling.

3. Roadmap of Theoretical Analysis

While DFC in principle can work with any base matrix factorization algorithm, it offers the greatest benefits when united with accurate but computationally expensive base procedures. Convex optimization approaches to matrix completion and robust matrix factorization (e.g., Lin et al., 2009b; Ma et al., 2011; Toh and Yun, 2010) are prime examples of this class, since they admit strong theoretical estimation guarantees (Agarwal et al., 2011; Candès et al., 2011; Candès and Plan, 2010; Negahban and Wainwright, 2012) but suffer from poor computational complexity due to the repeated and costly computation of truncated SVDs. Section 6 will provide empirical evidence that DFC provides an attractive framework to

improve the scalability of these algorithms, but we first present a thorough theoretical analysis of the estimation properties of DFC.

Over the course of the next three sections, we will show that the same assumptions that give rise to strong estimation guarantees for standard MF formulations also guarantee strong estimation properties for DFC. While these results represent an important first step toward understanding the theoretical behavior of DFC, we will see that certain gaps remain between our theoretical characterization and the practical performance of DFC. We will reflect on these gaps and the attendant opportunities for tightened theoretical analysis in Section 6.4. In the remainder of this section, we first introduce these standard assumptions and then present simplified bounds to build intuition for our theoretical results and our underlying proof techniques.

3.1 Standard Assumptions for Noisy Matrix Factorization

Since not all matrices can be recovered from missing entries or gross outliers, recent theoretical advances have studied sufficient conditions for accurate noisy MC (Candès and Plan, 2010; Keshavan et al., 2010; Negahban and Wainwright, 2012) and RMF (Agarwal et al., 2011; Zhou et al., 2010). Informally, these conditions capture the degree to which information about a single entry is "spread out" across a matrix. The ease of matrix estimation is correlated with this spread of information. The most prevalent set of conditions are *matrix coherence* conditions, which limit the extent to which the singular vectors of a matrix are correlated with the standard basis. However, there exist classes of matrices that violate the coherence conditions but can nonetheless be recovered from missing entries or gross outliers. Negahban and Wainwright (2012) define an alternative notion of *matrix spikiness* in part to handle these classes.

3.1.1 Matrix Coherence

Letting $\mathbf{e}_i$ be the ith column of the standard basis, we define two standard notions of coherence (Recht, 2011):

Definition 1 (μ_0-Coherence) *Let $\mathbf{V} \in \mathbb{R}^{n \times r}$ contain orthonormal columns with $r \leq n$. Then the μ_0-coherence of $\mathbf{V}$ is:*

$$\mu_0(\mathbf{V}) \triangleq \frac{n}{r} \max_{1 \leq i \leq n} \|\mathbf{P}_V \mathbf{e}_i\|^2 = \frac{n}{r} \max_{1 \leq i \leq n} \|\mathbf{V}_{(i)}\|^2 .$$

Definition 2 (μ_1-Coherence) *Let $\mathbf{L} \in \mathbb{R}^{m \times n}$ have rank r. Then, the μ_1-coherence of $\mathbf{L}$ is:*

$$\mu_1(\mathbf{L}) \triangleq \sqrt{\tfrac{mn}{r}} \max_{ij} |\mathbf{e}_i^\top \mathbf{U}_L \mathbf{V}_L^\top \mathbf{e}_j| .$$

For conciseness, we extend the definition of μ_0-coherence to an arbitrary matrix $\mathbf{L} \in \mathbb{R}^{m \times n}$ with rank r via $\mu_0(\mathbf{L}) \triangleq \max(\mu_0(\mathbf{U}_L), \mu_0(\mathbf{V}_L))$. Further, for any $\mu > 0$, we will call a matrix $\mathbf{L}$ (μ, r)-*coherent* if $\text{rank}(\mathbf{L}) = r$, $\mu_0(\mathbf{L}) \leq \mu$, and $\mu_1(\mathbf{L}) \leq \sqrt{\mu}$. Our analysis in Section 4 will focus on base MC and RMF algorithms that express their estimation guarantees in terms of the (μ, r)-coherence of the target low-rank matrix $\mathbf{L}_0$. For such algorithms, lower values of μ correspond to better estimation properties.

3.1.2 Matrix Spikiness

The matrix spikiness condition of Negahban and Wainwright (2012) captures the intuition that a matrix is easier to estimate if its maximum entry is not much larger than its average entry (in the root mean square sense):

Definition 3 (Spikiness) *The spikiness of* $\mathbf{L} \in \mathbb{R}^{m \times n}$ *is:*

$$\alpha(\mathbf{L}) \triangleq \sqrt{mn}\|\mathbf{L}\|_\infty / \|\mathbf{L}\|_F.$$

We call a matrix α*-spiky if* $\alpha(\mathbf{L}) \leq \alpha$.

Our analysis in Section 5 will focus on base MC algorithms that express their estimation guarantees in terms of the α-spikiness of the target low-rank matrix $\mathbf{L}_0$. For such algorithms, lower values of α correspond to better estimation properties.

3.2 Prototypical Estimation Bounds

We now present a prototypical estimation bound for DFC. Suppose that a base MC algorithm solves the *noisy nuclear norm heuristic*, studied in Candès and Plan (2010):

$$\text{minimize}_{\mathbf{L}} \quad \|\mathbf{L}\|_* \quad \text{subject to} \quad \|\mathcal{P}_\Omega(\mathbf{M} - \mathbf{L})\|_F \leq \Delta,$$

and that, for simplicity, $\mathbf{M}$ is square. The following prototype bound, derived from a new noisy MC guarantee in Theorem 10, describes the behavior of this estimator under matrix coherence assumptions. Note that the bound implies exact recovery in the noiseless setting, i.e., when $\Delta = 0$.

Proto-Bound 1 (MC under Incoherence) *Suppose that* $\mathbf{L}_0$ *is* (μ, r)*-coherent,* s *entries of* $\mathbf{M} \in \mathbb{R}^{n \times n}$ *are observed uniformly at random where* $s = \Omega(\mu r n \log^2(n))$*, and* $\|\mathbf{M} - \mathbf{L}_0\|_F \leq \Delta$*. If* $\hat{\mathbf{L}}$ *solves the noisy nuclear norm heuristic, then*

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F \leq f(n)\Delta$$

with high probability, where f *is a function of* n.

Now we present a corresponding prototype bound for DFC-Proj, a simplified version of our Corollary 14, under precisely the same coherence assumptions. Notably, this bound i) preserves accuracy with a flexible $(2 + \epsilon)$ degradation in estimation error over the base algorithm, ii) allows for speed-up by requiring only a vanishingly small fraction of columns to be sampled (i.e., $l/n \to 0$) whenever $s = \omega(n \log^2(n))$ entries are revealed, and iii) maintains exact recovery in the noiseless setting.

Proto-Bound 2 (DFC-MC under Incoherence) *Suppose that* $\mathbf{L}_0$ *is* (μ, r)*-coherent,* s *entries of* $\mathbf{M} \in \mathbb{R}^{n \times n}$ *are observed uniformly at random, and* $\|\mathbf{M} - \mathbf{L}_0\|_F \leq \Delta$*. Then it suffices to choose*

$$l \geq c \frac{\mu^2 r^2 n^2 \log^2(n)}{s\epsilon^2}$$

random columns suffice to have

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq (2+\epsilon) f(n) \Delta$$

with high probability when the noisy nuclear norm heuristic is used as a base algorithm, where f is the same function of n defined in Proto. 1 and c is a fixed positive constant.

The proof of Proto. 2, and indeed of each of our main DFC results, consists of three high-level steps:

1. *Bound coherence of submatrices*: Recall that the F step of DFC operates by applying a base MF algorithm to submatrices. We show that, with high probability, uniformly sampled submatrices are only moderately more coherent and moderately more spiky than the matrix from which they are drawn. This allows for accurate estimation of submatrices using base algorithms with standard coherence or spikiness requirements. The conservation of incoherence result is summarized in Lemma 4, while the conservation of non-spikiness is presented in Lemma 17.

2. *Bound error of randomized matrix approximations*: The error introduced by the C step of DFC depends on the framework variant. Drawing upon tools from randomized ℓ_2 regression (Drineas et al., 2008), randomized matrix multiplication (Drineas et al., 2006a,b), and matrix concentration (Hsu et al., 2012), we show that the same assumptions on the spread of information responsible for accurate MC and RMF also yield high fidelity reconstructions for column projection (Corollary 6 and Theorem 18) and the Nyström method (Corollary 7 and Corollary 8). We additionally present general approximation guarantees for random projection due to Halko et al. (2011) in Corollary 9. These results give rise to "master theorems" for coherence (Theorem 12) and spikiness (Theorem 20) that generically relate the estimation error of DFC to the error of any base algorithm.

3. *Bound error of submatrix factorizations*: The final step combines a master theorem with a base estimation guarantee applied to each DFC subproblem. We study both new (Theorem 10) and established bounds (Theorem 11 and Corollary 19) for MC and RMF and prove that DFC submatrices satisfy the base guarantee preconditions with high probability. We present the resulting coherence-based estimation guarantees for DFC in Corollary 14 and Corollary 16 and the spikiness-based estimation guarantee in Corollary 22.

The next two sections present the main results contributing to each of these proof steps, as well as their consequences for MC and RMF. Section 4 presents our analysis under coherence assumptions, while Section 5 contains our spikiness analysis.

4. Coherence-based Theoretical Analysis

This section presents our analysis of DFC under standard coherence assumptions encountered in the MC and RMF literature.

4.1 Coherence Analysis of Randomized Approximation Algorithms

We begin our coherence-based analysis by characterizing the behavior of randomized approximation algorithms under standard coherence assumptions. The derived properties will aid us in deriving DFC estimation guarantees. Hereafter, $\epsilon \in (0, 1]$ represents a prescribed error tolerance, and $\delta, \delta' \in (0, 1]$ denote target failure probabilities.

4.1.1 Conservation of Incoherence

Our first result bounds the μ_0 and μ_1-coherence of a uniformly sampled submatrix in terms of the coherence of the full matrix. This conservation of incoherence allows for accurate submatrix completion or submatrix outlier removal when using standard MC and RMF algorithms. Its proof is given in Section B.

Lemma 4 (Conservation of Incoherence) *Let $\mathbf{L} \in \mathbb{R}^{m\times n}$ be a rank-r matrix and define $\mathbf{L}_C \in \mathbb{R}^{m\times l}$ as a matrix of l columns of $\mathbf{L}$ sampled uniformly without replacement. If $l \geq cr\mu_0(\mathbf{V}_L)\log(n)\log(1/\delta)/\epsilon^2$, where c is a fixed positive constant defined in Corollary 6, then*

i) $\operatorname{rank}(\mathbf{L}_C) = \operatorname{rank}(\mathbf{L})$

ii) $\mu_0(\mathbf{U}_{L_C}) = \mu_0(\mathbf{U}_L)$

iii) $\mu_0(\mathbf{V}_{L_C}) \leq \dfrac{\mu_0(\mathbf{V}_L)}{1-\epsilon/2}$

iv) $\mu_1^2(\mathbf{L}_C) \leq \dfrac{r\mu_0(\mathbf{U}_L)\mu_0(\mathbf{V}_L)}{1-\epsilon/2}$

all hold jointly with probability at least $1 - \delta/n$.

4.1.2 Column Projection Analysis

Our next result shows that projection based on uniform column sampling leads to near optimal estimation in matrix regression when the covariate matrix has small coherence. This statement will immediately give rise to estimation guarantees for column projection and the generalized Nyström method.

Theorem 5 (Subsampled Regression under Incoherence) *Given a target matrix $\mathbf{B} \in \mathbb{R}^{p\times n}$ and a rank-r matrix of covariates $\mathbf{L} \in \mathbb{R}^{m\times n}$, choose $l \geq 3200r\mu_0(\mathbf{V}_L)\log(4n/\delta)/\epsilon^2$, let $\mathbf{B}_C \in \mathbb{R}^{p\times l}$ be a matrix of l columns of $\mathbf{B}$ sampled uniformly without replacement, and let $\mathbf{L}_C \in \mathbb{R}^{m\times l}$ consist of the corresponding columns of $\mathbf{L}$. Then,*

$$\|\mathbf{B} - \mathbf{B}_C\mathbf{L}_C^{+}\mathbf{L}\|_F \leq (1+\epsilon)\|\mathbf{B} - \mathbf{B}\mathbf{L}^{+}\mathbf{L}\|_F$$

with probability at least $1 - \delta - 0.2$.

Fundamentally, Theorem 5 links the notion of coherence, common in matrix estimation communities, to the randomized approximation concept of *leverage score sampling* (Mahoney and Drineas, 2009). The proof of Theorem 5, given in Section A, builds upon the

randomized ℓ_2 regression work of Drineas et al. (2008) and the matrix concentration results of Hsu et al. (2012) to yield a subsampled regression guarantee with better sampling complexity than that of Drineas et al. (2008, Theorem 5).

A first consequence of Theorem 5 shows that, with high probability, column projection produces an estimate nearly as good as a given rank-r target by sampling a number of columns proportional to the coherence and $r \log n$.

Corollary 6 (Column Projection under Incoherence) *Given a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ and a rank-r approximation $\mathbf{L} \in \mathbb{R}^{m \times n}$, choose $l \geq cr\mu_0(\mathbf{V}_L) \log(n) \log(1/\delta)/\epsilon^2$, where c is a fixed positive constant, and let $\mathbf{C} \in \mathbb{R}^{m \times l}$ be a matrix of l columns of $\mathbf{M}$ sampled uniformly without replacement. Then,*

$$\|\mathbf{M} - \mathbf{C}\mathbf{C}^+\mathbf{M}\|_F \leq (1+\epsilon)\|\mathbf{M} - \mathbf{L}\|_F$$

with probability at least $1 - \delta$.

Our result generalizes Theorem 1 of Drineas et al. (2008) by providing improved sampling complexity and guarantees relative to an *arbitrary* low-rank approximation. Notably, in the "noiseless" setting, when $\mathbf{M} = \mathbf{L}$, Corollary 6 guarantees exact recovery of $\mathbf{M}$ with high probability. The proof of Corollary 6 is given in Section C.

4.1.3 Generalized Nyström Analysis

Theorem 5 and Corollary 6 together imply an estimation guarantee for the generalized Nyström method relative to an arbitrary low-rank approximation $\mathbf{L}$. Indeed, if the matrix of sampled columns is denoted by $\mathbf{C}$, then, with appropriately reduced probability, $\mathrm{O}(\mu_0(\mathbf{V}_L) r \log n)$ columns and $\mathrm{O}(\mu_0(\mathbf{U}_C) r \log m)$ rows suffice to match the reconstruction error of $\mathbf{L}$ up to any fixed precision. The proof can be found in Section D.

Corollary 7 (Generalized Nyström under Incoherence) *Given a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ and a rank-r approximation $\mathbf{L} \in \mathbb{R}^{m \times n}$, choose $l \geq cr\mu_0(\mathbf{V}_L) \log(n) \log(1/\delta)/\epsilon^2$ with c a constant as in Corollary 6, and let $\mathbf{C} \in \mathbb{R}^{m \times l}$ be a matrix of l columns of $\mathbf{M}$ sampled uniformly without replacement. Further choose $d \geq cl\mu_0(\mathbf{U}_C) \log(m) \log(1/\delta')/\epsilon^2$, and let $\mathbf{R} \in \mathbb{R}^{d \times n}$ be a matrix of d rows of $\mathbf{M}$ sampled independently and uniformly without replacement. Then,*

$$\|\mathbf{M} - \mathbf{C}\mathbf{W}^+\mathbf{R}\|_F \leq (1+\epsilon)^2\|\mathbf{M} - \mathbf{L}\|_F$$

with probability at least $(1 - \delta)(1 - \delta' - 0.2)$.

Like the generalized Nyström bound of Drineas et al. (2008, Theorem 4) and unlike our column projection result, Corollary 7 depends on the coherence of the submatrix $\mathbf{C}$ and holds only with probability bounded away from 1. Our next contribution shows that we can do away with these restrictions in the noiseless setting, where $\mathbf{M} = \mathbf{L}$.

Corollary 8 (Noiseless Generalized Nyström under Incoherence) *Let $\mathbf{L} \in \mathbb{R}^{m \times n}$ be a rank-r matrix. Choose $l \geq 48r\mu_0(\mathbf{V}_L) \log(4n/(1-\sqrt{1-\delta}))$ and $d \geq 48r\mu_0(\mathbf{U}_L) \log(4m/(1-\sqrt{1-\delta}))$. Let $\mathbf{C} \in \mathbb{R}^{m \times l}$ be a matrix of l columns of $\mathbf{L}$ sampled uniformly without replacement, and let $\mathbf{R} \in \mathbb{R}^{d \times n}$ be a matrix of d rows of $\mathbf{L}$ sampled independently and uniformly without replacement. Then,*

$$\mathbf{L} = \mathbf{C}\mathbf{W}^+\mathbf{R}$$

with probability at least $1 - \delta$.

This result may appear surprising at first sight, since only vanishingly small fractions of rows and columns may participate in the generalized Nyström reconstruction. The intuition for the method's success that when the rank of $\mathbf{L}$ is small, only a small number of well-chosen rows and columns are needed to reconstruct the row and column space of $\mathbf{L}$ and that, when $\mathbf{L}$ is incoherent, uniform random sampling is likely produce well-chosen rows and columns. The proof of Corollary 8, given in Section E, adapts a strategy of Talwalkar and Rostamizadeh (2010) developed for the analysis of positive semidefinite matrices.

4.1.4 Random Projection Analysis

We next present an estimation guarantee for the random projection method relative to an arbitrary low-rank approximation $\mathbf{L}$. The result implies that using a random matrix with oversampled columns proportional to $r \log(1/\delta)$ suffices to match the reconstruction error of $\mathbf{L}$ up to any fixed precision with probability $1 - \delta$. The result is a direct consequence of the random projection analysis of Halko et al. (2011, Theorem 10.7), and the proof can be found in Section F.

Corollary 9 (Random Projection) *Given a matrix* $\mathbf{M} \in \mathbb{R}^{m \times n}$ *and a rank-r approximation* $\mathbf{L} \in \mathbb{R}^{m \times n}$ *with* $r \geq 2$*, choose an oversampling parameter*

$$p \geq 242 \; r \log(7/\delta)/\epsilon^2.$$

Draw an $n \times (r + p)$ *standard Gaussian matrix* $\mathbf{G}$ *and define* $\mathbf{Y} = \mathbf{MG}$*. Then, with probability at least* $1 - \delta$*,*

$$\|\mathbf{M} - \mathbf{P}_Y \mathbf{M}\|_F \leq (1 + \epsilon)\|\mathbf{M} - \mathbf{L}\|_F.$$

Moreover, define $\mathbf{L}^{rp}$ *as the best rank-r approximation of* $\mathbf{P}_Y \mathbf{M}$ *with respect to the Frobenius norm. Then, with probability at least* $1 - \delta$*,*

$$\|\mathbf{M} - \mathbf{L}^{rp}\|_F \leq (2 + \epsilon)\|\mathbf{M} - \mathbf{L}\|_F.$$

We note that, in contrast to Corollary 6 and Corollary 7, Corollary 9 does not depend on the coherence of $\mathbf{L}$ and hence can be fruitfully applied even in the absence of an incoherence assumption. We demonstrate such a use case in Section 5. We note moreover that past empirical studies have demonstrated excellent estimation error with $p \leq 10$ irrespective of the target matrix rank (Halko et al., 2011); bridging the gap between theory and practice in this instance represents an interesting open problem.

4.2 Base Algorithm Guarantees

As prototypical examples of the coherence-based estimation guarantees available for noisy MC and noisy RMF, consider the following two theorems. The first bounds the estimation error of a convex optimization approach to noisy matrix completion, under the assumptions of incoherence and uniform sampling.

Theorem 10 (Noisy MC under Incoherence) *Suppose that* $\mathbf{L}_0 \in \mathbb{R}^{m \times n}$ *is* (μ, r)*-coherent and that, for some target rate parameter* $\beta > 1$*,*

$$s \geq 32\mu r(m+n)\beta \log^2(m+n)$$

entries of $\mathbf{M}$ *are observed with locations* Ω *sampled uniformly without replacement. Then, if* $m \leq n$ *and* $\|\mathcal{P}_\Omega(\mathbf{M}) - \mathcal{P}_\Omega(\mathbf{L}_0)\|_F \leq \Delta$ *a.s., the minimizer* $\hat{\mathbf{L}}$ *of the problem*

$$\text{minimize}_{\mathbf{L}} \quad \|\mathbf{L}\|_* \quad \text{subject to} \quad \|\mathcal{P}_\Omega(\mathbf{M} - \mathbf{L})\|_F \leq \Delta. \tag{1}$$

satisfies

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F \leq 8\sqrt{\frac{2m^2n}{s} + m + \frac{1}{16}}\Delta \leq c_e\sqrt{mn}\Delta$$

with probability at least $1 - 4\log(n)n^{2-2\beta}$ *for* c_e *a positive constant.*

A similar estimation guarantee was obtained by Candès and Plan (2010) under stronger assumptions. We give the proof of Theorem 10 in Section J.

The second result, due to Zhou et al. (2010) and reformulated for a generic rate parameter β, as described in Candès et al. (2011, Section 3.1), bounds the estimation error of a convex optimization approach to noisy RMF, under the assumptions of incoherence and uniformly distributed outliers.

Theorem 11 (Noisy RMF under Incoherence, Zhou et al. 2010, Theorem 2) *Suppose that* $\mathbf{L}_0$ *is* (μ, r)*-coherent and that the support set of* $\mathbf{S}_0$ *is uniformly distributed among all sets of cardinality* s*. Then, if* $m \leq n$ *and* $\|\mathbf{M} - \mathbf{L}_0 - \mathbf{S}_0\|_F \leq \Delta$ *a.s., there is a constant* c_p *such that with probability at least* $1 - c_p n^{-\beta}$*, the minimizer* $(\hat{\mathbf{L}}, \hat{\mathbf{S}})$ *of the problem*

$$\text{minimize}_{\mathbf{L},\mathbf{S}} \quad \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \quad \text{subject to} \quad \|\mathbf{M} - \mathbf{L} - \mathbf{S}\|_F \leq \Delta \quad \text{with} \quad \lambda = 1/\sqrt{n} \tag{2}$$

satisfies $\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F^2 + \|\mathbf{S}_0 - \hat{\mathbf{S}}\|_F^2 \leq c_e'^2 mn\Delta^2$*, provided that*

$$r \leq \frac{\rho_r m}{\mu \log^2(n)} \quad \text{and} \quad s \leq (1 - \rho_s\beta)mn$$

for target rate parameter $\beta > 2$*, and positive constants* ρ_r, ρ_s*, and* c_e'*.*

4.3 Coherence Master Theorem

We now show that the same coherence conditions that allow for accurate MC and RMF also imply high-probability estimation guarantees for DFC. To make this precise, we let $\mathbf{M} = \mathbf{L}_0 + \mathbf{S}_0 + \mathbf{Z}_0 \in \mathbb{R}^{m \times n}$, where $\mathbf{L}_0$ is (μ, r)-coherent and $\|\mathcal{P}_\Omega(\mathbf{Z}_0)\|_F \leq \Delta$. Then, our next theorem provides a generic bound on the estimation error of DFC used in combination with an arbitrary base algorithm. The proof, which builds upon the results of Section 4.1, is given in Section G.

Theorem 12 (Coherence Master Theorem) *Choose $t = n/l$, $l \geq cr\mu \log(n)\log(2/\delta)/\epsilon^2$, where c is a fixed positive constant, and $p \geq 242\ r\log(14/\delta)/\epsilon^2$. Under the notation of Algorithms 1 and 2, let $\{\mathbf{C}_{0,1}, \cdots, \mathbf{C}_{0,t}\}$ be the corresponding partition of $\mathbf{L}_0$. Then, with probability at least $1-\delta$, $\mathbf{C}_{0,i}$ is $(\frac{r\mu^2}{1-\epsilon/2}, r)$-coherent for all i, and*

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^*\|_F \leq (2+\epsilon)\sqrt{\textstyle\sum_{i=1}^t \|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F^2},$$

where $\hat{\mathbf{L}}^*$ is the estimate returned by either DFC-Proj *or* DFC-RP*.*

Under the notation of Algorithm 3, let $\mathbf{C}_0$ and $\mathbf{R}_0$ be the corresponding column and row submatrices of $\mathbf{L}_0$. If in addition $d \geq cl\mu_0(\hat{\mathbf{C}})\log(m)\log(4/\delta)/\epsilon^2$, then, with probability at least $(1-\delta)(1-\delta-0.2)$, DFC-Nys *guarantees that $\mathbf{C}_0$ and $\mathbf{R}_0$ are $(\frac{r\mu^2}{1-\epsilon/2}, r)$-coherent and that*

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{nys}\|_F \leq (2+3\epsilon)\sqrt{\|\mathbf{C}_0 - \hat{\mathbf{C}}\|_F^2 + \|\mathbf{R}_0 - \hat{\mathbf{R}}\|_F^2}.$$

Remark 13 *The* DFC-Nys *guarantee requires the number of rows sampled to grow in proportion to $\mu_0(\hat{\mathbf{C}})$, a quantity always bounded by μ in our simulations. Here and in the consequences to follow, the* DFC-Nys *result can be strengthened in the noiseless setting ($\Delta = 0$) by utilizing Corollary 8 in place of Corollary 7 in the proof of Theorem 12.*

When a target matrix is incoherent, Theorem 12 asserts that – with high probability for DFC-Proj and DFC-RP and with fixed probability for DFC-Nys – the estimation error of DFC is not much larger than the error sustained by the base algorithm on each subproblem. Because Theorem 12 further bounds the coherence of each submatrix, we can use any coherence-based matrix estimation guarantee to control the estimation error on each subproblem. The next two sections demonstrate how Theorem 12 can be applied to derive specific DFC estimation guarantees for noisy MC and noisy RMF. In these sections, we let $\bar{n} \triangleq \max(m,n)$.

4.4 Consequences for Noisy MC

As a first consequence of Theorem 12, we will show that DFC retains the high-probability estimation guarantees of a standard MC solver while operating on matrices of much smaller dimension. Suppose that a base MC algorithm solves the convex optimization problem of Eq. (1). Then, Corollary 14 follows from the Coherence Master Theorem (Theorem 12) and the base algorithm guarantee of Theorem 10.

Corollary 14 (DFC-MC under Incoherence) *Suppose that $\mathbf{L}_0$ is (μ, r)-coherent and that s entries of $\mathbf{M}$ are observed, with locations Ω distributed uniformly. Fix any target rate parameter $\beta > 1$. Then, if $\|\mathcal{P}_\Omega(\mathbf{M}) - \mathcal{P}_\Omega(\mathbf{L}_0)\|_F \leq \Delta$ a.s., and the base algorithm solves the optimization problem of Eq. (1), it suffices to choose $t = n/l$,*

$$l \geq c\mu^2 r^2 (m+n) n\beta \log^2(m+n)/(s\epsilon^2), \quad d \geq cl\mu_0(\hat{\mathbf{C}})(2\beta-1)\log^2(4\bar{n})\bar{n}/(n\epsilon^2),$$

and $p \geq 242\ r\log(14\bar{n}^{2\beta-2})/\epsilon^2$ to achieve

$$\textbf{DFC-Proj}: \ \|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq (2+\epsilon)c_e\sqrt{mn}\Delta$$

DFC-RP : $\|\mathbf{L}_0 - \hat{\mathbf{L}}^{rp}\|_F \leq (2+\epsilon)c_e\sqrt{mn}\Delta$

DFC-Nys : $\|\mathbf{L}_0 - \hat{\mathbf{L}}^{nys}\|_F \leq (2+3\epsilon)c_e\sqrt{ml+dn}\Delta$

with probability at least

DFC-Proj / DFC-RP : $1 - (5t\log(\bar{n})+1)\bar{n}^{2-2\beta} \geq 1 - \bar{n}^{3-2\beta}$

DFC-Nys : $1 - (10\log(\bar{n})+2)\bar{n}^{2-2\beta} - 0.2$,

respectively, with c as in Theorem 12 and c_e as in Theorem 10.

Remark 15 *Corollary 14 allows for the fraction of columns and rows sampled to decrease as the number of revealed entries, s, increases. Only a vanishingly small fraction of columns ($l/n \to 0$) and rows ($d/\bar{n} \to 0$) need be sampled whenever $s = \omega((m+n)\log^2(m+n))$.*

To understand the conclusions of Corollary 14, consider the base algorithm of Theorem 10, which, when applied to $\mathcal{P}_\Omega(\mathbf{M})$, recovers an estimate $\hat{\mathbf{L}}$ satisfying $\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F \leq c_e\sqrt{mn}\Delta$ with high probability. Corollary 14 asserts that, with appropriately reduced probability, DFC-Proj and DFC-RP exhibit the same estimation error scaled by an adjustable factor of $2+\epsilon$, while DFC-Nys exhibits a somewhat smaller error scaled by $2+3\epsilon$.

The key take-away is that DFC introduces a controlled increase in error and a controlled decrement in the probability of success, allowing the user to interpolate between maximum speed and maximum accuracy. Thus, DFC can quickly provide near-optimal estimation in the noisy setting and exact recovery in the noiseless setting ($\Delta = 0$), even when entries are missing. The proof of Corollary 14 can be found in Section H.

4.5 Consequences for Noisy RMF

Our next corollary shows that DFC retains the high-probability estimation guarantees of a standard RMF solver while operating on matrices of much smaller dimension. Suppose that a base RMF algorithm solves the convex optimization problem of Eq. (2). Then, Corollary 16 follows from the Coherence Master Theorem (Theorem 12) and the base algorithm guarantee of Theorem 11.

Corollary 16 (DFC-RMF under Incoherence) *Suppose that $\mathbf{L}_0$ is (μ, r)-coherent with*

$$r^2 \leq \frac{\min(m,n)\rho_r}{2\mu^2\log^2(\bar{n})}$$

for a positive constant ρ_r. Suppose moreover that the uniformly distributed support set of $\mathbf{S}_0$ has cardinality s. For a fixed positive constant ρ_s, define the undersampling parameter

$$\beta_s \triangleq \left(1 - \frac{s}{mn}\right)/\rho_s,$$

and fix any target rate parameter $\beta > 2$ with rescaling $\beta' \triangleq \beta\log(\bar{n})/\log(m)$ satisfying $4\beta_s - 3/\rho_s \leq \beta' \leq \beta_s$. Then, if $\|\mathbf{M} - \mathbf{L}_0 - \mathbf{S}_0\|_F \leq \Delta$ a.s., and the base algorithm solves

the optimization problem of Eq. (2), it suffices to choose $t = n/l$,

$$l \geq \max\left(\frac{cr^2\mu^2\beta\log^2(2\bar{n})}{\epsilon^2\rho_r}, \frac{4\log(\bar{n})\beta(1-\rho_s\beta_s)}{m(\rho_s\beta_s-\rho_s\beta')^2}\right),$$
$$d \geq \max\left(\frac{cl\mu_0(\hat{\mathbf{C}})\beta\log^2(4\bar{n})}{\epsilon^2}, \frac{4\log(\bar{n})\beta(1-\rho_s\beta_s)}{n(\rho_s\beta_s-\rho_s\beta')^2}\right)$$

and $p \geq 242\ r\log(14\bar{n}^{\beta})/\epsilon^2$ *to have*

$$\textbf{DFC-Proj}: \ \|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq (2+\epsilon)c'_e\sqrt{mn}\Delta$$

$$\textbf{DFC-RP}: \ \|\mathbf{L}_0 - \hat{\mathbf{L}}^{rp}\|_F \leq (2+\epsilon)c'_e\sqrt{mn}\Delta$$

$$\textbf{DFC-Nys}: \ \|\mathbf{L}_0 - \hat{\mathbf{L}}^{nys}\|_F \leq (2+3\epsilon)c'_e\sqrt{ml+dn}\Delta$$

with probability at least

$$\textbf{DFC-Proj / DFC-RP}: \ 1-(t(c_p+1)+1)\bar{n}^{-\beta} \geq 1-c_p\bar{n}^{1-\beta}$$

$$\textbf{DFC-Nys}: \ 1-(2c_p+3)\bar{n}^{-\beta}-0.2,$$

respectively, with c *as in Theorem 12 and* ρ_r, c'_e, *and* c_p *as in Theorem 11.*

Note that Corollary 16 places only very mild restrictions on the number of columns and rows to be sampled. Indeed, l and d need only grow poly-logarithmically in the matrix dimensions to achieve estimation guarantees comparable to those of the RMF base algorithm (Theorem 11). Hence, DFC can quickly provide near-optimal estimation in the noisy setting and exact recovery in the noiseless setting ($\Delta = 0$), even when entries are grossly corrupted. The proof of Corollary 16 can be found in Section I.

5. Theoretical Analysis under Spikiness Conditions

This section presents our analysis of DFC under standard spikiness assumptions from the MC and RMF literature.

5.1 Spikiness Analysis of Randomized Approximation Algorithms

We begin our spikiness analysis by characterizing the behavior of randomized approximation algorithms under standard spikiness assumptions. The derived properties will aid us in developing DFC estimation guarantees. Hereafter, $\epsilon \in (0,1]$ represents a prescribed error tolerance, and $\delta, \delta' \in (0,1]$ designates a target failure probability.

5.1.1 Conservation of Non-Spikiness

Our first lemma establishes that the uniformly sampled submatrices of an α-spiky matrix are themselves nearly α-spiky with high probability. This property will allow for accurate submatrix completion or outlier removal using standard MC and RMF algorithms. Its proof is given in Section K.

Lemma 17 (Conservation of Non-Spikiness) *Let $\mathbf{L}_C \in \mathbb{R}^{m\times l}$ be a matrix of l columns of $\mathbf{L} \in \mathbb{R}^{m\times n}$ sampled uniformly without replacement. If $l \geq \alpha^4(\mathbf{L})\log(1/\delta)/(2\epsilon^2)$, then*

$$\alpha(\mathbf{L}_C) \leq \frac{\alpha(\mathbf{L})}{\sqrt{1-\epsilon}}$$

with probability at least $1-\delta$.

5.1.2 Column Projection Analysis

Our first theorem asserts that, with high probability, column projection produces an approximation nearly as good as a given rank-r target by sampling a number of columns proportional to the spikiness and $r\log(mn)$.

Theorem 18 (Column Projection under Non-Spikiness) *Given a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ and a rank-r, α-spiky approximation $\mathbf{L} \in \mathbb{R}^{m\times n}$, choose*

$$l \geq 8r\alpha^4\log(2mn/\delta)/\epsilon^2,$$

and let $\mathbf{C} \in \mathbb{R}^{m\times l}$ be a matrix of l columns of $\mathbf{M}$ sampled uniformly without replacement. Then,

$$\|\mathbf{M} - \mathbf{L}^{proj}\|_F \leq \|\mathbf{M}-\mathbf{L}\|_F + \epsilon$$

with probability at least $1-\delta$, whenever $\|\mathbf{M}\|_\infty \leq \alpha/\sqrt{mn}$.

The proof of Theorem 18 builds upon the randomized matrix multiplication work of Drineas et al. (2006a,b) and will be given in Section L.

5.2 Base Algorithm Guarantee

The next result, a reformulation of Negahban and Wainwright (2012, Corollary 1), is a prototypical example of a spikiness-based estimation guarantee for noisy MC. Corollary 19 bounds the estimation error of a convex optimization approach to noisy matrix completion, under non-spikiness and uniform sampling assumptions.

Corollary 19 (Noisy MC under Non-Spikiness) (Negahban and Wainwright, 2012) *Suppose that $\mathbf{L}_0 \in \mathbb{R}^{m\times n}$ is α-spiky with rank r and $\|\mathbf{L}_0\|_F \leq 1$ and that $\mathbf{Z}_0 \in \mathbb{R}^{m\times n}$ has i.i.d. zero-mean, sub-exponential entries with variance ν^2/mn. If, for an oversampling parameter $\beta > 0$,*

$$s \geq \alpha^2\beta r(m+n)\log(m+n)$$

entries of $\mathbf{M} = \mathbf{L}_0 + \mathbf{Z}_0$ are observed with locations Ω sampled uniformly with replacement, then any solution $\hat{\mathbf{L}}$ of the problem

$$\text{minimize}_{\mathbf{L}} \quad \frac{mn}{2s}\|\mathcal{P}_\Omega(\mathbf{M}-\mathbf{L})\|_F^2 + \lambda\|\mathbf{L}\|_* \quad \text{subject to} \quad \|\mathbf{L}\|_\infty \leq \frac{\alpha}{\sqrt{mn}} \tag{3}$$
$$\text{with} \quad \lambda = 4\nu\sqrt{(m+n)\log(m+n)/s}$$

satisfies

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F^2 \leq c_1\max(\nu^2, 1)/\beta$$

with probability at least $1 - c_2\exp(-c_3\log(m+n))$ for positive constants c_1, c_2, and c_3.

5.3 Spikiness Master Theorem

We now show that the same spikiness conditions that allow for accurate MC also imply high-probability estimation guarantees for DFC. To make this precise, we let $\mathbf{M} = \mathbf{L}_0 + \mathbf{Z}_0 \in \mathbb{R}^{m\times n}$, where $\mathbf{L}_0$ is α-spiky with rank r and that $\mathbf{Z}_0 \in \mathbb{R}^{m\times n}$ has i.i.d. zero-mean, sub-exponential entries with variance ν^2/mn. We further fix any $\epsilon, \delta \in (0, 1]$. Then, our Theorem 20 provides a generic bound on estimation error for DFC when used in combination with an arbitrary base algorithm. The proof, which builds upon the results of Section 5.1, is deferred to Section M.

Theorem 20 (Spikiness Master Theorem) *Choose $t = n/l$, $l \geq 13r\alpha^4 \log(4mn/\delta)/\epsilon^2$, and $p \geq 242\, r \log(14/\delta)/\epsilon^2$. Under the notation of Algorithms 1 and 2, let $\{\mathbf{C}_{0,1}, \cdots, \mathbf{C}_{0,t}\}$ be the corresponding partition of $\mathbf{L}_0$. Then, with probability at least $1-\delta$,* DFC-Proj *and* DFC-RP *guarantee that $\mathbf{C}_{0,i}$ is $(\sqrt{1.25}\alpha)$-spiky for all i and that*

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq 2\sqrt{\textstyle\sum_{i=1}^t \|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F^2} + \epsilon \quad \textit{and}$$
$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{rp}\|_F \leq (2+\epsilon)\sqrt{\textstyle\sum_{i=1}^t \|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F^2}$$

whenever $\|\hat{\mathbf{C}}_i\|_\infty \leq \sqrt{1.25}\alpha/\sqrt{ml}$ for all i.

Remark 21 *The factor of $\sqrt{1.25}$ can be replaced with the smaller term $\sqrt{1+\epsilon/(4\sqrt{r})}$.*

When a target matrix is non-spiky, Theorem 20 asserts that, with high probability, the estimation error of DFC is not much larger than the error sustained by the base algorithm on each subproblem. Theorem 20 further bounds the spikiness of each submatrix with high probability, and hence we can use any spikiness-based matrix estimation guarantee to control the estimation error on each subproblem. The next section demonstrates how Theorem 20 can be applied to derive specific DFC estimation guarantees for noisy MC.

5.4 Consequences for Noisy MC

Our corollary of Theorem 20 shows that DFC retains the high-probability estimation guarantees of a standard MC solver while operating on matrices of much smaller dimension. Suppose that a base MC algorithm solves the convex optimization problem of Eq. (3). Then, Corollary 22 follows from the Spikiness Master Theorem (Theorem 20) and the base algorithm guarantee of Corollary 19.

Corollary 22 (DFC-MC under Non-Spikiness) *Suppose that $\mathbf{L}_0 \in \mathbb{R}^{m\times n}$ is α-spiky with rank r and $\|\mathbf{L}_0\|_F \leq 1$ and that $\mathbf{Z}_0 \in \mathbb{R}^{m\times n}$ has i.i.d. zero-mean, sub-exponential entries with variance ν^2/mn. Let c_1, c_2, and c_3 be positive constants as in Corollary 19. If s entries of $\mathbf{M} = \mathbf{L}_0 + \mathbf{Z}_0$ are observed with locations Ω sampled uniformly with replacement, and the base algorithm solves the optimization problem of Eq. (3), then it suffices to choose $t = n/l$,*

$$l \geq 13(c_3+1)\sqrt{\frac{(m+n)\log(m+n)\beta}{s}} nr\alpha^4 \log(4mn)/\epsilon^2,$$

and $p \geq 242\ r\log(14(m+l)^{c_3})/\epsilon^2$ *to achieve*

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq 2\sqrt{c_1 \max((l/n)\nu^2, 1)/\beta} + \epsilon \quad \textit{and}$$
$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{rp}\|_F \leq (2+\epsilon)\sqrt{c_1 \max((l/n)\nu^2, 1)/\beta}$$

with respective probability at least $1-(t+1)(c_2+1)\exp(-c_3\log(m+l))$*, if the base algorithm of Eq. (3) is used with* $\lambda = 4\nu\sqrt{(m+n)\log(m+n)/s}$*.*

Remark 23 *Corollary 22 allows for the fraction of columns sampled to decrease as the number of revealed entries, s, increases. Only a vanishingly small fraction of columns* $(l/n \to 0)$ *need be sampled whenever* $s = \omega((m+n)\log^3(m+n))$*.*

To understand the conclusions of Corollary 22, consider the base algorithm of Corollary 19, which, when applied to $\mathbf{M}$, recovers an estimate $\hat{\mathbf{L}}$ satisfying

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F \leq \sqrt{c_1 \max(\nu^2, 1)/\beta}$$

with high probability. Corollary 14 asserts that, with appropriately reduced probability, DFC-RP exhibits the same estimation error scaled by an adjustable factor of $2+\epsilon$, while DFC-Proj exhibits at most twice this error plus an adjustable factor of ϵ. Hence, DFC can quickly provide near-optimal estimation for non-spiky matrices as well as incoherent matrices, even when entries are missing. The proof of Corollary 22 can be found in Section N.

6. Experimental Evaluation

We now explore the accuracy and speed-up of DFC on a variety of simulated and real-world data sets. We use the Accelerated Proximal Gradient (APG) algorithm of Toh and Yun (2010) as our base noisy MC algorithm[5] and the APG algorithm of Lin et al. (2009b) as our base noisy RMF algorithm. In order to provide a fair comparison with baseline algorithms, we perform all experiments on an x86-64 architecture using a single 2.60 Ghz core and 30GB of main memory. In practice, one will typically run DFC jobs in a distributed fashion across a cluster; our released code supports this standard use case. We use the default parameter settings suggested by Toh and Yun (2010) and Lin et al. (2009b), and measure estimation error via root mean square error (RMSE). To achieve a fair running time comparison, we execute each subproblem in the F step of DFC in a serial fashion on the same machine using a single core. Since, in practice, each of these subproblems would be executed in parallel, the *parallel running time* of DFC is calculated as the time to complete the D and C steps of DFC plus the running time of the longest running subproblem in the F step. We compare DFC with two baseline methods: the base algorithm APG applied to the full matrix $\mathbf{M}$ and Partition, which carries out the D and F steps of DFC-Proj but omits the final C step (projection). We denote a particular sampling method along with the size of its partitions as '*method-xx*%,' e.g., Proj-25% refers to DFC-Proj with partitioned submatrices containing 25% of the columns of the full matrix (i.e., $t = 4$). For Partition, DFC-Proj, and DFC-RP, we orient our data matrices such that $n \geq m$ and partition by column. Moreover, for DFC-RP we set $p = 5$ and $q = 2$.

5. Our experiments with the Augmented Lagrange Multiplier (ALM) algorithm of Lin et al. (2009a) as a base algorithm (not reported) yield comparable relative speedups and performance for DFC.

6.1 Simulations

For our simulations, we focused on square matrices ($m = n$) and generated random low-rank and sparse decompositions, similar to the schemes used in related work (Candès et al., 2011; Keshavan et al., 2010; Zhou et al., 2010). We created $\mathbf{L}_0 \in \mathbb{R}^{m \times m}$ as a random product, $\mathbf{A}\mathbf{B}^\top$, where $\mathbf{A}$ and $\mathbf{B}$ are $m \times r$ matrices with independent $\mathcal{N}(0, \sqrt{1/r})$ entries such that each entry of $\mathbf{L}_0$ has unit variance. $\mathbf{Z}_0$ contained independent $\mathcal{N}(0, 0.1)$ entries. In the MC setting, s entries of $\mathbf{L}_0 + \mathbf{Z}_0$ were revealed uniformly at random. In the RMF setting, the support of $\mathbf{S}_0$ was generated uniformly at random, and the s corrupted entries took values in $[0, 1]$ with uniform probability. For each algorithm, we report error between $\mathbf{L}_0$ and the estimated low-rank matrix, and all reported results are averages over ten trials.

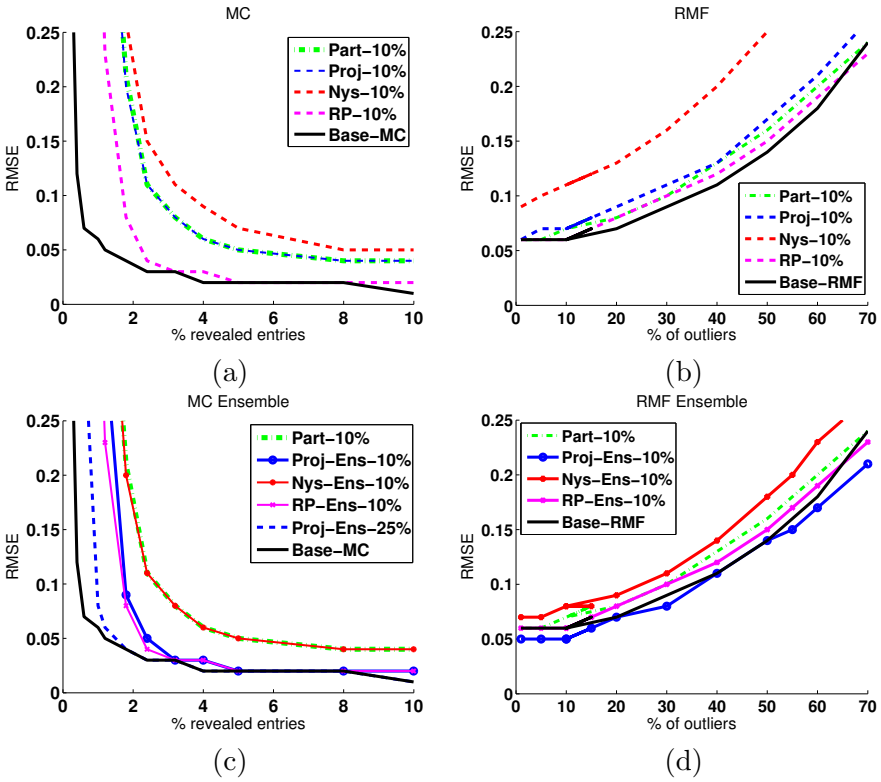

Figure 1: Recovery error of DFC relative to base algorithms.

We first explored the estimation error of DFC as a function of s, using (m = 10K, $r = 10$) with varying observation sparsity for MC and (m = 1K, $r = 10$) with a varying percentage of outliers for RMF. The results are summarized in Figure 1. In both MC and RMF, the gaps in estimation between APG and DFC are small when sampling only 10% of rows and columns. Moreover, of the standard DFC algorithms, DFC-RP performs the best, as shown in Figures 1(a) and (b). Ensembling improves the performance of DFC-Nys and DFC-Proj, as shown in Figures 1(c) and (d), and DFC-Proj-Ens in particular consistently outperforms Partition and DFC-Nys-Ens, slightly outperforms DFC-RP, and matches the performance of APG for most settings of s. In practice we observe that $\mathbf{L}^{rp}$ equals the optimal (with respect to the spectral or Frobenius norm) rank-k approximation

of $[\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t]$, and thus the performance of DFC-RP consistently matches that of DFC-RP-Ens. We therefore omit the DFC-RP-Ens results in the remainder this section.

We next explored the speed-up of DFC as a function of matrix size. For MC, we revealed 4% of the matrix entries and set $r = 0.001 \cdot m$, while for RMF we fixed the percentage of outliers to 10% and set $r = 0.01 \cdot m$. We sampled 10% of rows and columns and observed that estimation errors were comparable to the errors presented in Figure 1 for similar settings of s; in particular, at all values of n for both MC and RMF, the errors of APG and DFC-Proj-Ens were nearly identical. Our timing results, presented in Figure 2, illustrate a near-linear speed-up for MC and a superlinear speed-up for RMF across varying matrix sizes. Note that the timing curves of the DFC algorithms and Partition all overlap, a fact that highlights the minimal computational cost of the final matrix approximation step.

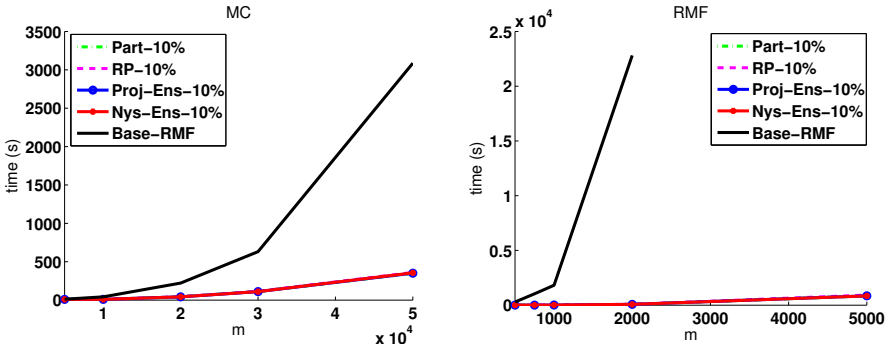

Figure 2: Speed-up of DFC relative to base algorithms.

6.2 Collaborative Filtering

Collaborative filtering for recommender systems is one prevalent real-world application of noisy matrix completion. A collaborative filtering data set can be interpreted as the incomplete observation of a ratings matrix with columns corresponding to users and rows corresponding to items. The goal is to infer the unobserved entries of this ratings matrix. We evaluate DFC on two of the largest publicly available collaborative filtering data sets: MovieLens 10M (`http://www.grouplens.org/`) with $m = 10$K, $n = 72$K, $s > 10$M, and the Netflix Prize data set (`http://www.netflixprize.com/`) with $m = 18$K, $n = 480$K, $s > 100$M. To generate test sets drawn from the training distribution, for each data set, we aggregated all available rating data into a single training set and withheld test entries uniformly at random, while ensuring that at least one training observation remained in each row and column. The algorithms were then run on the remaining training portions and evaluated on the test portions of each split. The results, averaged over three train-test splits, are summarized in Table 2. Notably, DFC-Proj, DFC-Proj-Ens, DFC-Nys-Ens, and DFC-RP all outperform Partition, and DFC-Proj-Ens performs comparably to APG while providing a nearly linear parallel time speed-up. Similar to the simulation results presented in Figure 1, DFC-RP performs the best of the standard DFC algorithms, though DFC-Proj-Ens slightly outperforms DFC-RP. Moreover, the poorer performance of DFC-Nys can be in part explained by the asymmetry of these problems. Since these matrices have many more columns than rows, MF on column submatrices is inherently

Method	MovieLens 10M		Netflix	
	RMSE	Time	RMSE	Time
Base algorithm (APG)	**0.8005**	**552.3s**	**0.8433**	**4775.4s**
Partition-25%	0.8146	146.2s	0.8451	1274.6s
Partition-10%	0.8461	56.0s	0.8491	548.0s
DFC-Nys-25%	0.8449	141.9s	0.8832	1541.2s
DFC-Nys-10%	0.8776	82.5s	0.9228	797.4s
DFC-Nys-Ens-25%	0.8085	153.5s	0.8486	1661.2s
DFC-Nys-Ens-10%	0.8328	96.2s	0.8613	909.8s
DFC-Proj-25%	0.8061	146.3s	0.8436	1274.8s
DFC-Proj-10%	0.8270	56.0s	0.8486	548.1s
DFC-Proj-Ens-25%	**0.7944**	**146.3s**	**0.8411**	**1274.8s**
DFC-Proj-Ens-10%	**0.8117**	**56.0s**	**0.8434**	**548.1s**
DFC-RP-25%	0.8027	147.4s	0.8438	1283.6s
DFC-RP-10%	0.8074	56.2s	0.8448	550.1s

Table 2: Performance of DFC relative to base algorithm APG on collaborative filtering tasks.

easier than MF on row submatrices, and for DFC-Nys, we observe that $\hat{\mathbf{C}}$ is an accurate estimate while $\hat{\mathbf{R}}$ is not.

6.3 Background Modeling in Computer Vision

Background modeling has important practical ramifications for detecting activity in surveillance video. This problem can be framed as an application of noisy RMF, where each video frame is a column of some matrix ($\mathbf{M}$), the background model is low-rank ($\mathbf{L}_0$), and moving objects and background variations, e.g., changes in illumination, are outliers ($\mathbf{S}_0$). We evaluate DFC on two videos (treating each frame as a row): 'Hall' (200 frames of size 176×144) contains significant foreground variation and was studied by Candès et al. (2011), while 'Lobby' (1546 frames of size 168×120) includes many changes in illumination (a smaller video with 250 frames was studied by Candès et al. 2011). We focused on DFC-Proj-Ens, due to its superior performance in previous experiments, and measured the RMSE between the background model estimated by DFC and that of APG. On both videos, DFC-Proj-Ens estimated nearly the same background model as the full APG algorithm in a small fraction of the time. On 'Hall,' the DFC-Proj-Ens-5% and DFC-Proj-Ens-0.5% models exhibited RMSEs of 0.564 and 1.55, quite small given pixels with 256 intensity values. The associated running time was reduced from 342.5s for APG to real-time (5.2s for a 13s video) for DFC-Proj-Ens-0.5%. Snapshots of the results are presented in Figure 3. On 'Lobby,'

Original frame | APG (342.5s) | 5% sampled (24.2s) | 0.5% sampled (5.2s)

Figure 3: Sample 'Hall' estimation by APG, DFC-Proj-Ens-5%, and DFC-Proj-Ens-.5%.

the RMSE of DFC-Proj-Ens-4% was 0.64, and the speed-up over APG was more than 20X, i.e., the running time reduced from 16557s to 792s.

6.4 From Theory to Practice

Our experimental results suggest that the theoretical error bounds of Secs. 4 and 5 can be further tightened. In particular, our master theorems Theorems 12 and 20 guarantee that DFC-Proj-Ens and DFC-RP are never more than a constant factor worse than Partition, yet in both real data experiments and simulations we observe significant gains in accuracy over Partition due to the incorporation of projection and ensembling. Moreover, our theory gives rise to comparable estimation guarantees for DFC-Nys, albeit under stronger assumptions as noted in Remark 13. This is a surprising fact given that DFC-Nys may make use of only a vanishingly small subset of all available matrix entries; however, we find that for data sets with high noise levels, methods that make use of all available data like DFC-Proj and DFC-RP are unsurprisingly more accurate than DFC-Nys. We view addressing these gaps between theory and practice as important directions for future work.

7. Conclusions

To improve the scalability of existing matrix factorization algorithms while leveraging the ubiquity of parallel computing architectures, we introduced, evaluated, and analyzed DFC, a divide-and-conquer framework for noisy matrix factorization with missing entries or outliers. DFC is trivially parallelized and particularly well suited for distributed environments given its low communication footprint. Moreover, DFC provably maintains the estimation guarantees of its base algorithm, even in the presence of noise, and yields linear to super-linear speedups in practice. A number of natural follow-up questions suggest themselves:

- Can the sampling complexities and conclusions of our theoretical analyses be strengthened? For example, can the $(2+\epsilon)$ approximation guarantees of our master theorems be sharpened to $(1+\epsilon)$? More generally, can we close the gaps between theory and practice described in Section 6.4?

- How does DFC compare empirically with scalable heuristics for MC and RMF that have little theoretical backing (see, e.g., Zhou et al., 2008; Gemulla et al., 2011; Recht and Ré, 2011; F. Niu et al., 2011; Yu et al., 2012; Mu et al., 2011)? Is improved performance obtained by pairing DFC with base algorithms lacking theoretical guarantees but displaying other practical benefits?

- Which algorithmic refinements lead to enhanced performance for DFC? For instance, could ensemble variants of DFC be improved by learning combination weights in a manner analogous to that of Kumar et al. (2009b)? In the matrix completion setting, could one use held-out entries to determine the optimal dimension (via rows or via columns) for partitioning in DFC-Proj or DFC-RP?

These open questions are fertile ground for future work.

Acknowledgments

Lester Mackey gratefully acknowledges the support of DARPA through the National Defense Science and Engineering Graduate Fellowship Program. Ameet Talwalkar gratefully acknowledges support from NSF award No. 1122732.

Appendix A. Proof of Theorem 5: Subsampled Regression under Incoherence

We now give a proof of Theorem 5. While the results of this section are stated in terms of i.i.d. with-replacement sampling of columns and rows, a concise argument due to Hoeffding (1963, Section 6) implies the same conclusions when columns and rows are sampled without replacement.

Our proof of Theorem 5 will require a strengthened version of the randomized ℓ_2 regression work of Drineas et al. (2008, Theorem 5). The proof of Theorem 5 of Drineas et al. (2008) relies heavily on the fact that $\|\mathbf{AB} - \mathbf{GH}\|_F \leq \frac{\epsilon}{2}\|\mathbf{A}\|_F\|\mathbf{B}\|_F$ with probability at least 0.9, when $\mathbf{G}$ and $\mathbf{H}$ contain sufficiently many rescaled columns and rows of $\mathbf{A}$ and $\mathbf{B}$, sampled according to a particular non-uniform probability distribution. A result of Hsu et al. (2012), modified to allow for slack in the probabilities, establishes a related claim with improved sampling complexity.[6]

Lemma 24 (Hsu et al. 2012, Example 4.3) *Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times k}$ with $r \geq \operatorname{rank}(\mathbf{A})$, an error tolerance $\epsilon \in (0, 1]$, and a failure probability $\delta \in (0, 1]$, define probabilities p_j satisfying*

$$p_j \geq \frac{\beta}{Z}\|\mathbf{A}_{(j)}\|^2, \quad Z = \sum_j \|\mathbf{A}_{(j)}\|^2, \quad \text{and} \quad \textstyle\sum_{j=1}^k p_j = 1$$

6. The general conclusion of (Hsu et al., 2012, Example 4.3) is incorrectly stated as noted in Hsu (2012). However, the original statement is correct in the special case when a matrix is multiplied by its own transpose, which is the case of interest here.

for some $\beta \in (0,1]$. *Let* $\mathbf{G} \in \mathbb{R}^{m\times l}$ *be a column submatrix of* $\mathbf{A}$ *in which exactly* $l \geq 48r\log(4r/(\beta\delta))/(\beta\epsilon^2)$ *columns are selected in i.i.d. trials in which the* j*-th column is chosen with probability* p_j. *Further, let* $\mathbf{D} \in \mathbb{R}^{l\times l}$ *be a diagonal rescaling matrix with entry* $\mathbf{D}_{tt} = 1/\sqrt{lp_j}$ *whenever the* j*-th column of* $\mathbf{A}$ *is selected on the* t*-th sampling trial, for* $t = 1,\ldots,l$. *Then, with probability at least* $1-\delta$,

$$\|\mathbf{A}\mathbf{A}^\top - \mathbf{G}\mathbf{D}\mathbf{D}\mathbf{G}^\top\|_2 \leq \frac{\epsilon}{2}\|\mathbf{A}\|_2^2.$$

Using Lemma 24, we now establish a stronger version of Lemma 1 of Drineas et al. (2008). For a given $\beta \in (0,1]$ and $\mathbf{L} \in \mathbb{R}^{m\times n}$ with rank r, we first define column sampling probabilities p_j satisfying

$$p_j \geq \frac{\beta}{r}\|(\mathbf{V}_L)_{(j)}\|^2 \quad \text{and} \quad \textstyle\sum_{j=1}^n p_j = 1. \tag{4}$$

We further let $\mathbf{S} \in \mathbb{R}^{n\times l}$ be a random binary matrix with independent columns, where a single 1 appears in each column, and $\mathbf{S}_{jt} = 1$ with probability p_j for each $t \in \{1,\ldots,l\}$. Moreover, let $\mathbf{D} \in \mathbb{R}^{l\times l}$ be a diagonal rescaling matrix with entry $\mathbf{D}_{tt} = 1/\sqrt{lp_j}$ whenever $\mathbf{S}_{jt} = 1$. Postmultiplication by $\mathbf{S}$ is equivalent to selecting l random columns of a matrix, independently and with replacement. Under this notation, we establish the following lemma:

Lemma 25 *Let* $\epsilon \in (0,1]$, *and define* $\mathbf{V}_l^\top = \mathbf{V}_L^\top\mathbf{S}$ *and* $\Gamma = (\mathbf{V}_l^\top\mathbf{D})^+ - (\mathbf{V}_l^\top\mathbf{D})^\top$. *If* $l \geq 48r\log(4r/(\beta\delta))/(\beta\epsilon^2)$ *for* $\delta \in (0,1]$ *then with probability at least* $1-\delta$*:*

$$\begin{aligned}
\operatorname{rank}(\mathbf{V}_l) &= \operatorname{rank}(\mathbf{V}_L) = \operatorname{rank}(\mathbf{L}) \\
\|\Gamma\|_2 &= \|\mathbf{\Sigma}^{-1}_{V_l^\top D} - \mathbf{\Sigma}_{V_l^\top D}\|_2 \\
(\mathbf{LSD})^+ &= (\mathbf{V}_l^\top\mathbf{D})^+\mathbf{\Sigma}_L^{-1}\mathbf{U}_L^\top \\
\|\mathbf{\Sigma}^{-1}_{V_l^\top D} - \mathbf{\Sigma}_{V_l^\top D}\|_2 &\leq \epsilon/\sqrt{2}.
\end{aligned}$$

Proof By Lemma 24, for all $1 \leq i \leq r$,

$$\begin{aligned}
|1-\sigma_i^2(\mathbf{V}_l^\top\mathbf{D})| &= |\sigma_i(\mathbf{V}_L^\top\mathbf{V}_L) - \sigma_i(\mathbf{V}_l^\top\mathbf{D}\mathbf{D}\mathbf{V}_l)| \\
&\leq \|\mathbf{V}_L^\top\mathbf{V}_L - \mathbf{V}_L^\top\mathbf{S}\mathbf{D}\mathbf{D}\mathbf{S}^\top\mathbf{V}_L\|_2 \\
&\leq \epsilon/2\|\mathbf{V}_L^\top\|_2^2 = \epsilon/2,
\end{aligned}$$

where $\sigma_i(\cdot)$ is the i-th largest singular value of a given matrix. Since $\epsilon/2 \leq 1/2$, each singular value of $\mathbf{V}_l$ is positive, and so $\operatorname{rank}(\mathbf{V}_l) = \operatorname{rank}(\mathbf{V}_L) = \operatorname{rank}(\mathbf{L})$. The remainder of the proof is identical to that of Lemma 1 of Drineas et al. (2008). ■

Lemma 25 immediately yields improved sampling complexity for the randomized ℓ_2 regression of Drineas et al. (2008):

Proposition 26 *Suppose* $\mathbf{B} \in \mathbb{R}^{p\times n}$ *and* $\epsilon \in (0,1]$. *If* $l \geq 3200r\log(4r/(\beta\delta))/(\beta\epsilon^2)$ *for* $\delta \in (0,1]$, *then with probability at least* $1-\delta-0.2$*:*

$$\|\mathbf{B} - \mathbf{B}\mathbf{S}\mathbf{D}(\mathbf{L}\mathbf{S}\mathbf{D})^+\mathbf{L}\|_F \leq (1+\epsilon)\|\mathbf{B} - \mathbf{B}\mathbf{L}^+\mathbf{L}\|_F.$$

Proof The proof is identical to that of Theorem 5 of Drineas et al. (2008) once Lemma 25 is substituted for Lemma 1 of Drineas et al. (2008). ■

A typical application of Prop. 26 would involve performing a truncated SVD of $\mathbf{M}$ to obtain the *statistical leverage scores*, $\|(\mathbf{V}_L)_{(j)}\|^2$, used to compute the column sampling probabilities of Eq. (4). Here, we will take advantage of the slack term, β, allowed in the sampling probabilities of Eq. (4) to show that uniform column sampling gives rise to the same estimation guarantees for column projection approximations when $\mathbf{L}$ is sufficiently incoherent.

To prove Theorem 5, we first notice that $n \geq r\mu_0(\mathbf{V}_L)$ and hence

$$\begin{aligned} l &\geq 3200 r\mu_0(\mathbf{V}_L)\log(4r\mu_0(\mathbf{V}_L)/\delta)/\epsilon^2 \\ &\geq 3200 r\log(4r/(\beta\delta))/(\beta\epsilon^2) \end{aligned}$$

whenever $\beta \geq 1/\mu_0(\mathbf{V}_L)$. Thus, we may apply Prop. 26 with $\beta = 1/\mu_0(\mathbf{V}_L) \in (0,1]$ and $p_j = 1/n$ by noting that

$$\frac{\beta}{r}\|(\mathbf{V}_L)_{(j)}\|^2 \leq \frac{\beta}{r}\frac{r}{n}\mu_0(\mathbf{V}_L) = \frac{1}{n} = p_j$$

for all j, by the definition of $\mu_0(\mathbf{V}_L)$. By our choice of probabilities, $\mathbf{D} = \mathbf{I}\sqrt{n/l}$, and hence

$$\|\mathbf{B} - \mathbf{B}_C\mathbf{L}_C^+\mathbf{L}\|_F = \|\mathbf{B} - \mathbf{B}_C\mathbf{D}(\mathbf{L}_C\mathbf{D})^+\mathbf{L}\|_F \leq (1+\epsilon)\|\mathbf{B} - \mathbf{B}\mathbf{L}^+\mathbf{L}\|_F$$

with probability at least $1 - \delta - 0.2$, as desired.

Appendix B. Proof of Lemma 4: Conservation of Incoherence

Since for all $n > 1$,

$$c\log(n)\log(1/\delta) = (c/4)\log(n^4)\log(1/\delta) \geq 48\log(4n^2/\delta) \geq 48\log(4r\mu_0(\mathbf{V}_L)/(\delta/n))$$

as $n \geq r\mu_0(\mathbf{V}_L)$, claim *i* follows immediately from Lemma 25 with $\beta = 1/\mu_0(\mathbf{V}_L)$, $p_j = 1/n$ for all j, and $\mathbf{D} = \mathbf{I}\sqrt{n/l}$. When $\text{rank}(\mathbf{L}_C) = \text{rank}(\mathbf{L})$, Lemma 1 of Mohri and Talwalkar (2011) implies that $\mathbf{P}_{U_{L_C}} = \mathbf{P}_{U_L}$, which in turn implies claim *ii*.

To prove claim *iii* given the conclusions of Lemma 25, assume, without loss of generality, that $\mathbf{V}_l$ consists of the first l rows of $\mathbf{V}_L$. Then if $\mathbf{L}_C = \mathbf{U}_L\mathbf{\Sigma}_L\mathbf{V}_l^\top$ has $\text{rank}(\mathbf{L}_C) = \text{rank}(\mathbf{L}) = r$, the matrix $\mathbf{V}_l$ must have full column rank. Thus we can write

$$\begin{aligned} \mathbf{L}_C^+\mathbf{L}_C &= (\mathbf{U}_L\mathbf{\Sigma}_L\mathbf{V}_l^\top)^+\mathbf{U}_L\mathbf{\Sigma}_L\mathbf{V}_l^\top \\ &= (\mathbf{\Sigma}_L\mathbf{V}_l^\top)^+\mathbf{U}_L^+\mathbf{U}_L\mathbf{\Sigma}_L\mathbf{V}_l^\top \\ &= (\mathbf{\Sigma}_L\mathbf{V}_l^\top)^+\mathbf{\Sigma}_L\mathbf{V}_l^\top \\ &= (\mathbf{V}_l^\top)^+\mathbf{\Sigma}_L^+\mathbf{\Sigma}_L\mathbf{V}_l^\top \\ &= (\mathbf{V}_l^\top)^+\mathbf{V}_l^\top \\ &= \mathbf{V}_l(\mathbf{V}_l^\top\mathbf{V}_l)^{-1}\mathbf{V}_l^\top, \end{aligned}$$

where the second and third equalities follow from $\mathbf{U}_L$ having orthonormal columns, the fourth and fifth result from $\mathbf{\Sigma}_L$ having full rank and $\mathbf{V}_l$ having full column rank, and the sixth follows from $\mathbf{V}_l^\top$ having full row rank.

Now, denote the right singular vectors of $\mathbf{L}_C$ by $\mathbf{V}_{L_C} \in \mathbb{R}^{l \times r}$. Observe that $\mathbf{P}_{V_{L_C}} = \mathbf{V}_{L_C}\mathbf{V}_{L_C}^\top = \mathbf{L}_C^+\mathbf{L}_C$, and define $\mathbf{e}_{i,l}$ as the ith column of $\mathbf{I}_l$ and $\mathbf{e}_{i,n}$ as the ith column of $\mathbf{I}_n$. Then we have,

$$\begin{aligned}
\mu_0(\mathbf{V}_{L_C}) &= \frac{l}{r}\max_{1\le i\le l}\|\mathbf{P}_{V_{L_C}}\mathbf{e}_{i,l}\|^2 \\
&= \frac{l}{r}\max_{1\le i\le l}\mathbf{e}_{i,l}^\top\mathbf{L}_C^+\mathbf{L}_C\mathbf{e}_{i,l} \\
&= \frac{l}{r}\max_{1\le i\le l}\mathbf{e}_{i,l}^\top(\mathbf{V}_l^\top)^+\mathbf{V}_l^\top\mathbf{e}_{i,l} \\
&= \frac{l}{r}\max_{1\le i\le l}\mathbf{e}_{i,l}^\top\mathbf{V}_l(\mathbf{V}_l^\top\mathbf{V}_l)^{-1}\mathbf{V}_l^\top\mathbf{e}_{i,l} \\
&= \frac{l}{r}\max_{1\le i\le l}\mathbf{e}_{i,n}^\top\mathbf{V}_L(\mathbf{V}_l^\top\mathbf{V}_l)^{-1}\mathbf{V}_L^\top\mathbf{e}_{i,n},
\end{aligned}$$

where the final equality follows from $\mathbf{V}_l^\top\mathbf{e}_{i,l} = \mathbf{V}_L^\top\mathbf{e}_{i,n}$ for all $1 \le i \le l$.

Now, defining $\mathbf{Q} = \mathbf{V}_l^\top\mathbf{V}_l$ we have

$$\begin{aligned}
\mu_0(\mathbf{V}_{L_C}) &= \frac{l}{r}\max_{1\le i\le l}\mathbf{e}_{i,n}^\top\mathbf{V}_L\mathbf{Q}^{-1}\mathbf{V}_L^\top\mathbf{e}_{i,n} \\
&= \frac{l}{r}\max_{1\le i\le l}\mathrm{Tr}\left[\mathbf{e}_{i,n}^\top\mathbf{V}_L\mathbf{Q}^{-1}\mathbf{V}_L^\top\mathbf{e}_{i,n}\right] \\
&= \frac{l}{r}\max_{1\le i\le l}\mathrm{Tr}\left[\mathbf{Q}^{-1}\mathbf{V}_L^\top\mathbf{e}_{i,n}\mathbf{e}_{i,n}^\top\mathbf{V}_L\right] \\
&\le \frac{l}{r}\|\mathbf{Q}^{-1}\|_2\max_{1\le i\le l}\|\mathbf{V}_L^\top\mathbf{e}_{i,n}\mathbf{e}_{i,n}^\top\mathbf{V}_L\|_*\,,
\end{aligned}$$

by Hölder's inequality for Schatten p-norms. Since $\mathbf{V}_L^\top\mathbf{e}_{i,n}\mathbf{e}_{i,n}^\top\mathbf{V}_L$ has rank one, we can explicitly compute its trace norm as $\|\mathbf{V}_L^\top\mathbf{e}_{i,n}\|^2 = \|\mathbf{P}_{V_L}\mathbf{e}_{i,n}\|^2$. Hence,

$$\begin{aligned}
\mu_0(\mathbf{V}_{L_C}) &\le \frac{l}{r}\|\mathbf{Q}^{-1}\|_2\max_{1\le i\le l}\|\mathbf{P}_{V_L}\mathbf{e}_{i,n}\|^2 \\
&\le \frac{l}{r}\frac{r}{n}\|\mathbf{Q}^{-1}\|_2\left(\frac{n}{r}\max_{1\le i\le n}\|\mathbf{P}_{V_L}\mathbf{e}_{i,n}\|^2\right) \\
&= \frac{l}{n}\|\mathbf{Q}^{-1}\|_2\mu_0(\mathbf{V}_L)\,,
\end{aligned}$$

by the definition of μ_0-coherence. The proof of Lemma 25 established that the smallest singular value of $\frac{n}{l}\mathbf{Q} = \mathbf{V}_l^\top\mathbf{D}\mathbf{D}\mathbf{V}_l$ is lower bounded by $1 - \frac{\epsilon}{2}$ and hence $\|\mathbf{Q}^{-1}\|_2 \le \frac{n}{l(1-\epsilon/2)}$. Thus, we conclude that $\mu_0(\mathbf{V}_{L_C}) \le \mu_0(\mathbf{V}_L)/(1-\epsilon/2)$.

To prove claim *iv* under Lemma 25, we note that

$$\begin{aligned}
\mu_1(\mathbf{L}_C) &= \sqrt{\frac{ml}{r}} \max_{\substack{1\le i\le m\\ 1\le j\le l}} |\mathbf{e}_{i,m}^\top \mathbf{U}_{L_C} \mathbf{V}_{L_C}^\top \mathbf{e}_{j,l}| \\
&\le \sqrt{\frac{ml}{r}} \max_{1\le i\le m} \|\mathbf{U}_{L_C}^\top \mathbf{e}_{i,m}\| \max_{1\le j\le l} \|\mathbf{V}_{L_C}^\top \mathbf{e}_{j,l}\| \\
&= \sqrt{r}\left(\sqrt{\frac{m}{r}} \max_{1\le i\le m} \|\mathbf{P}_{U_{L_C}} \mathbf{e}_{i,m}\|\right)\left(\sqrt{\frac{l}{r}} \max_{1\le j\le l} \|\mathbf{P}_{V_{L_C}} \mathbf{e}_{j,l}\|\right) \\
&= \sqrt{r\mu_0(\mathbf{U}_{L_C})\mu_0(\mathbf{V}_{L_C})} \le \sqrt{r\mu_0(\mathbf{U}_L)\mu_0(\mathbf{V}_L)/(1-\epsilon/2)}
\end{aligned}$$

by Hölder's inequality for Schatten p-norms, the definition of μ_0-coherence, and claims *ii* and *iii*.

Appendix C. Proof of Corollary 6: Column Projection under Incoherence

Fix $c = 48000/\log(1/0.45)$, and notice that for $n > 1$,

$$48000\log(n) \ge 3200\log(n^5) \ge 3200\log(16n).$$

Hence $l \ge 3200r\mu_0(\mathbf{V}_L)\log(16n)(\log(\delta)/\log(0.45))/\epsilon^2$.

Now partition the columns of $\mathbf{C}$ into $b = \log(\delta)/\log(0.45)$ submatrices, $\mathbf{C} = [\mathbf{C}_1, \cdots, \mathbf{C}_b]$, each with $a = l/b$ columns,[7] and let $[\mathbf{L}_{C_1}, \cdots, \mathbf{L}_{C_b}]$ be the corresponding partition of $\mathbf{L}_C$. Since

$$a \ge 3200r\mu_0(\mathbf{V}_L)\log(4n/0.25)/\epsilon^2,$$

we may apply Prop. 26 independently for each i to yield

$$\|\mathbf{M} - \mathbf{C}_i\mathbf{L}_{C_i}^+\mathbf{L}\|_F \le (1+\epsilon)\|\mathbf{M} - \mathbf{M}\mathbf{L}^+\mathbf{L}\|_F \le (1+\epsilon)\|\mathbf{M} - \mathbf{L}\|_F \tag{5}$$

with probability at least 0.55, since $\mathbf{M}\mathbf{L}^+$ minimizes $\|\mathbf{M} - \mathbf{Y}\mathbf{L}\|_F$ over all $\mathbf{Y} \in \mathbb{R}^{m\times m}$.

Since each $\mathbf{C}_i = \mathbf{C}\mathbf{S}_i$ for some matrix $\mathbf{S}_i$ and $\mathbf{C}^+\mathbf{M}$ minimizes $\|\mathbf{M} - \mathbf{C}\mathbf{X}\|_F$ over all $\mathbf{X} \in \mathbb{R}^{l\times n}$, it follows that

$$\|\mathbf{M} - \mathbf{C}\mathbf{C}^+\mathbf{M}\|_F \le \|\mathbf{M} - \mathbf{C}_i\mathbf{L}_{C_i}^+\mathbf{L}\|_F,$$

for each i. Hence, if

$$\|\mathbf{M} - \mathbf{C}\mathbf{C}^+\mathbf{M}\|_F \le (1+\epsilon)\|\mathbf{M} - \mathbf{L}\|_F,$$

fails to hold, then, for each i, Eq. (5) also fails to hold. The desired conclusion therefore must hold with probability at least $1 - 0.45^b = 1 - \delta$.

7. For simplicity, we assume that b divides l evenly.

Appendix D. Proof of Corollary 7: Generalized Nyström Method under Incoherence

With $c = 48000/\log(1/0.45)$ as in Corollary 6, we notice that for $m > 1$,

$$48000\log(m) = 16000\log(m^3) \geq 16000\log(4m).$$

Therefore,

$$\begin{aligned} d &\geq 16000r\mu_0(\mathbf{U}_C)\log(4m)(\log(\delta')/\log(0.45))/\epsilon^2 \\ &\geq 3200r\mu_0(\mathbf{U}_C)\log(4m/\delta')/\epsilon^2, \end{aligned}$$

for all $m > 1$ and $\delta' \leq 0.8$. Hence, we may apply Theorem 5 and Corollary 6 in turn to obtain

$$\|\mathbf{M} - \mathbf{C}\mathbf{W}^+\mathbf{R}\|_F \leq (1+\epsilon)\|\mathbf{M} - \mathbf{C}\mathbf{C}^+\mathbf{M}\|_F \leq (1+\epsilon)^2\|\mathbf{M} - \mathbf{L}\|$$

with probability at least $(1-\delta)(1-\delta'-0.2)$ by independence.

Appendix E. Proof of Corollary 8: Noiseless Generalized Nyström Method under Incoherence

Since $\text{rank}(\mathbf{L}) = r$, $\mathbf{L}$ admits a decomposition $\mathbf{L} = \mathbf{Y}^\top\mathbf{Z}$ for some matrices $\mathbf{Y} \in \mathbb{R}^{r\times m}$ and $\mathbf{Z} \in \mathbb{R}^{r\times n}$. In particular, let $\mathbf{Y}^\top = \mathbf{U}_L\mathbf{\Sigma}_L^{\frac{1}{2}}$ and $\mathbf{Z} = \mathbf{\Sigma}_L^{\frac{1}{2}}\mathbf{V}_L^\top$. By block partitioning $\mathbf{Y}$ and $\mathbf{Z}$ as $\mathbf{Y} = \begin{bmatrix}\mathbf{Y}_1 & \mathbf{Y}_2\end{bmatrix}$ and $\mathbf{Z} = \begin{bmatrix}\mathbf{Z}_1 & \mathbf{Z}_2\end{bmatrix}$ for $\mathbf{Y}_1 \in \mathbb{R}^{r\times d}$ and $\mathbf{Z}_1 \in \mathbb{R}^{r\times l}$, we may write $\mathbf{W} = \mathbf{Y}_1^\top\mathbf{Z}_1, \mathbf{C} = \mathbf{Y}^\top\mathbf{Z}_1$, and $\mathbf{R} = \mathbf{Y}_1^\top\mathbf{Z}$. Note that we assume that the generalized Nyström approximation is generated from sampling the first l columns and the first d rows of $\mathbf{L}$, which we do without loss of generality since the rows and columns of the original low-rank matrix can always be permuted to match this assumption.

Prop. 27 shows that, like the Nyström method (Kumar et al., 2009a), the generalized Nyström method yields exact recovery of $\mathbf{L}$ whenever $\text{rank}(\mathbf{L}) = \text{rank}(\mathbf{W})$. The same result was established in Wang et al. (2009) with a different proof.

Proposition 27 *Suppose* $r = \text{rank}(\mathbf{L}) \leq \min(d, l)$ *and* $\text{rank}(\mathbf{W}) = r$. *Then* $\mathbf{L} = \mathbf{L}^{nys}$.

Proof By appealing to our factorized block decomposition, we may rewrite the generalized Nyström approximation as $\mathbf{L}^{nys} = \mathbf{C}\mathbf{W}^+\mathbf{R} = \mathbf{Y}^\top\mathbf{Z}_1(\mathbf{Y}_1^\top\mathbf{Z}_1)^+\mathbf{Y}_1^\top\mathbf{Z}$. We first note that $\text{rank}(\mathbf{W}) = r$ implies that $\text{rank}(\mathbf{Y}_1) = r$ and $\text{rank}(\mathbf{Z}_1) = r$ so that $\mathbf{Z}_1\mathbf{Z}_1^\top$ and $\mathbf{Y}_1\mathbf{Y}_1^\top$ are full-rank. Hence, $(\mathbf{Y}_1^\top\mathbf{Z}_1)^+ = \mathbf{Z}_1^\top(\mathbf{Z}_1\mathbf{Z}_1^\top)^{-1}(\mathbf{Y}_1\mathbf{Y}_1^\top)^{-1}\mathbf{Y}_1$, yielding

$$\mathbf{L}^{nys} = \mathbf{Y}^\top\mathbf{Z}_1\mathbf{Z}_1^\top(\mathbf{Z}_1\mathbf{Z}_1^\top)^{-1}(\mathbf{Y}_1\mathbf{Y}_1^\top)^{-1}\mathbf{Y}_1\mathbf{Y}_1^\top\mathbf{Z} = \mathbf{Y}^\top\mathbf{Z} = \mathbf{L}.$$

■

Prop. 27 allows us to lower bound the probability of exact recovery with the probability of randomly selecting a rank-r submatrix. As $\text{rank}(\mathbf{W}) = r$ iff both $\text{rank}(\mathbf{Y}_1) = r$ and $\text{rank}(\mathbf{Z}_1) = r$, it suffices to characterize the probability of selecting full rank submatrices of $\mathbf{Y}$ and $\mathbf{Z}$. Following the treatment of the Nyström method in Talwalkar and Rostamizadeh

(2010), we note that $\mathbf{\Sigma}_L^{-\frac{1}{2}}\mathbf{Z} = \mathbf{V}_L^\top$ and hence that $\mathbf{Z}_1^\top \mathbf{\Sigma}_{\mathbf{L}}^{-\frac{1}{2}} = \mathbf{V}_l$ where $\mathbf{V}_l \in \mathbb{R}^{l\times r}$ contains the first l components of the leading r right singular vectors of $\mathbf{L}$. It follows that $\text{rank}(\mathbf{Z}_1) = \text{rank}(\mathbf{Z}_1^\top \mathbf{\Sigma}_L^{-\frac{1}{2}}) = \text{rank}(\mathbf{V}_l)$. Similarly, $\text{rank}(\mathbf{Y}_1) = \text{rank}(\mathbf{U}_d)$ where $\mathbf{U}_d \in \mathbb{R}^{d\times r}$ contains the first d components of the leading r left singular vectors of $\mathbf{L}$. Thus, we have

$$\mathbf{P}(\text{rank}(\mathbf{Z}_1) = r) = \mathbf{P}(\text{rank}(\mathbf{V}_l) = r) \qquad \text{and} \tag{6}$$

$$\mathbf{P}(\text{rank}(\mathbf{Y}_1) = r) = \mathbf{P}(\text{rank}(\mathbf{U}_d) = r). \tag{7}$$

Next we can apply the first result of Lemma 25 to lower bound the RHSs of Eq. (6) and Eq. (7) by selecting $\epsilon = 1$, $\mathbf{S}$ such that its diagonal entries equal 1, and $\beta = \frac{1}{\mu_0(\mathbf{V}_L)}$ for the RHS of Eq. (6) and $\beta = \frac{1}{\mu_0(\mathbf{U}_L)}$ for the RHS of Eq. (7). In particular, given the lower bounds on d and l in the statement of the corollary, the RHSs are each lower bounded by $\sqrt{1-\delta}$. Furthermore, by the independence of row and column sampling and Eq. (6) and Eq. (7), we see that

$$\begin{aligned} 1-\delta &\leq \mathbf{P}(\text{rank}(\mathbf{U}_d) = r)\mathbf{P}(\text{rank}(\mathbf{V}_l) = r) \\ &= \mathbf{P}(\text{rank}(\mathbf{Y}_1) = r)\mathbf{P}(\text{rank}(\mathbf{Z}_1) = r) \\ &= \mathbf{P}(\text{rank}(\mathbf{W}) = r). \end{aligned}$$

Finally, Prop. 27 implies that

$$\mathbf{P}(\mathbf{L} = \mathbf{L}^{nys}) \geq \mathbf{P}(\text{rank}(\mathbf{W}) = r) \geq 1-\delta,$$

which proves the statement of the theorem.

Appendix F. Proof of Corollary 9: Random Projection

Our proof rests upon the following random projection guarantee of Halko et al. (2011):

Theorem 28 (Halko et al. 2011, Theorem 10.7) *Given a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ and a rank-r approximation $\mathbf{L} \in \mathbb{R}^{m\times n}$ with $r \geq 2$, choose an oversampling parameter $p \geq 4$, where $r + p \leq \min(m, n)$. Draw an $n \times (r+p)$ standard Gaussian matrix $\mathbf{G}$, let $\mathbf{Y} = \mathbf{MG}$. For all $u, t \geq 1$,*

$$\|\mathbf{M} - \mathbf{P}_Y\mathbf{M}\|_F \leq (1 + t\sqrt{12r/p})\|\mathbf{M} - \mathbf{M}_r\|_F + ut \cdot \frac{e\sqrt{r+p}}{p+1}\|\mathbf{M} - \mathbf{M}_r\|$$

with probability at least $1 - 5t^{-p} - 2e^{-u^2/2}$.

Fix $(u, t) = (\sqrt{2\log(7/\delta)}, e)$, and note that

$$1 - 5e^{-p} - 2e^{-u^2/2} = 1 - 5e^{-p} - 2\delta/7 \geq 1 - \delta,$$

since $p \geq \log(7/\delta)$. Hence, Theorem 28 implies that

$$\begin{aligned}\|\mathbf{M} - \mathbf{P}_Y\mathbf{M}\|_F &\leq (1 + e\sqrt{12r/p})\|\mathbf{M} - \mathbf{M}_r\|_F + \frac{e^2\sqrt{2(r+p)\log(7/\delta)}}{p+1}\|\mathbf{M} - \mathbf{M}_r\|_2 \\ &\leq \left(1 + e\sqrt{12r/p} + \frac{e^2\sqrt{2(r+p)\log(7/\delta)}}{p+1}\right)\|\mathbf{M} - \mathbf{L}\|_F \\ &\leq \left(1 + e\sqrt{12r/p} + e^2\sqrt{2r\log(7/\delta)/p}\right)\|\mathbf{M} - \mathbf{L}\|_F \\ &\leq \left(1 + 11\sqrt{2r\log(7/\delta)/p}\right)\|\mathbf{M} - \mathbf{L}\|_F \leq (1+\epsilon)\|\mathbf{M} - \mathbf{L}\|_F\end{aligned}$$

with probability at least $1 - \delta$, where the second inequality follows from $\|\mathbf{M} - \mathbf{M}_r\|_2 \leq \|\mathbf{M} - \mathbf{M}_r\|_F \leq \|\mathbf{M} - \mathbf{L}\|_F$, the third follows from $\sqrt{r+p}\sqrt{p} \leq (p+1)\sqrt{r}$ for all r and p, and the final follows from our choice of $p \geq 242\ r\log(7/\delta)/\epsilon^2$.

Next, we note, as in the proof of Theorem 9.3 of Halko et al. (2011), that

$$\|\mathbf{P}_Y\mathbf{M} - \mathbf{L}^{rp}\|_F \leq \|\mathbf{P}_Y\mathbf{M} - \mathbf{P}_Y\mathbf{M}_r\|_F \leq \|\mathbf{M} - \mathbf{M}_r\|_F \leq \|\mathbf{M} - \mathbf{L}\|_F\,.$$

The first inequality holds because $\mathbf{L}^{rp}$ is by definition the best rank-r approximation to $\mathbf{P}_Y\mathbf{M}$ and $\text{rank}(\mathbf{P}_Y\mathbf{M}_r) \leq r$. The second inequality holds since

$$\|\mathbf{M} - \mathbf{M}_r\|_F = \|\mathbf{P}_Y(\mathbf{M} - \mathbf{M}_r)\|_F + \|\mathbf{P}_Y^{\perp}(\mathbf{M} - \mathbf{M}_r)\|_F\,.$$

The final inequality holds since $\mathbf{M}_r$ is the best rank-r approximation to $\mathbf{M}$ and $\text{rank}(\mathbf{L}) = r$. Moreover, by the triangle inequality,

$$\begin{aligned}\|\mathbf{M} - \mathbf{L}^{rp}\|_F &\leq \|\mathbf{M} - \mathbf{P}_Y\mathbf{M}\|_F + \|\mathbf{P}_Y\mathbf{M} - \mathbf{L}^{rp}\|_F \\ &\leq \|\mathbf{M} - \mathbf{P}_Y\mathbf{M}\|_F + \|\mathbf{M} - \mathbf{L}\|_F\,. \qquad (8)\end{aligned}$$

Combining Eq. (8) with the first statement of the corollary yields the second statement.

Appendix G. Proof of Theorem 12: Coherence Master Theorem

G.1 Proof of DFC-Proj and DFC-RP Bounds

Let $\mathbf{L}_0 = [\mathbf{C}_{0,1}, \ldots, \mathbf{C}_{0,t}]$ and $\tilde{\mathbf{L}} = [\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t]$. Define $A(\mathbf{X})$ as the event that a matrix $\mathbf{X}$ is $(\frac{r\mu^2}{1-\epsilon/2}, r)$-coherent and K as the event $\|\tilde{\mathbf{L}} - \hat{\mathbf{L}}^{proj}\|_F \leq (1+\epsilon)\|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F$. When K holds, we have that

$$\begin{aligned}\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F &\leq \|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F + \|\tilde{\mathbf{L}} - \hat{\mathbf{L}}^{proj}\|_F \leq (2+\epsilon)\|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F \\ &= (2+\epsilon)\sqrt{\textstyle\sum_{i=1}^t \|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F^2},\end{aligned}$$

by the triangle inequality, and hence it suffices to lower bound $\mathbf{P}(K \cap \bigcap_i A(\mathbf{C}_{0,i}))$. Our choice of l, with a factor of $\log(2/\delta)$, implies that each $A(\mathbf{C}_{0,i})$ holds with probability at least $1 - \delta/(2n)$ by Lemma 4, while K holds with probability at least $1 - \delta/2$ by Corollary 6. Hence, by the union bound,

$$\mathbf{P}(K \cap \textstyle\bigcap_i A(\mathbf{C}_{0,i})) \geq 1 - \mathbf{P}(K^c) - \sum_i \mathbf{P}(A(\mathbf{C}_{0,i})^c) \geq 1 - \delta/2 - t\delta/(2n) \geq 1 - \delta.$$

An identical proof with Corollary 9 substituted for Corollary 6 yields the random projection result.

G.2 Proof of DFC-Nys Bound

To prove the generalized Nyström result, we redefine $\tilde{\mathbf{L}}$ and write it in block notation as:

$$\tilde{\mathbf{L}} = \begin{bmatrix} \hat{\mathbf{C}}_1 & \hat{\mathbf{R}}_2 \\ \hat{\mathbf{C}}_2 & \mathbf{L}_{0,22} \end{bmatrix}, \quad \text{where} \quad \hat{\mathbf{C}} = \begin{bmatrix} \hat{\mathbf{C}}_1 \\ \hat{\mathbf{C}}_2 \end{bmatrix}, \quad \hat{\mathbf{R}} = \begin{bmatrix} \hat{\mathbf{R}}_1 & \hat{\mathbf{R}}_2 \end{bmatrix}$$

and $\mathbf{L}_{0,22} \in \mathbb{R}^{(m-d)\times(n-l)}$ is the bottom right submatrix of $\mathbf{L}_0$. We further redefine K as the event $\|\tilde{\mathbf{L}} - \hat{\mathbf{L}}^{nys}\|_F \leq (1+\epsilon)^2 \|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F$. As above,

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{nys}\|_F \leq \|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F + \|\tilde{\mathbf{L}} - \hat{\mathbf{L}}^{nys}\|_F \leq (2 + 2\epsilon + \epsilon^2)\|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F \leq (2+3\epsilon)\|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F,$$

when K holds, by the triangle inequality. Our choices of l and

$$d \geq cl\mu_0(\hat{\mathbf{C}}) \log(m) \log(4/\delta)/\epsilon^2 \geq cr\mu \log(m) \log(1/\delta)/\epsilon^2$$

imply that $A(\mathbf{C})$ and $A(\mathbf{R})$ hold with probability at least $1 - \delta/(2n)$ and $1 - \delta/(4n)$ respectively by Lemma 4, while K holds with probability at least $(1-\delta/2)(1-\delta/4-0.2)$ by Corollary 7. Hence, by the union bound,

$$\begin{aligned} \mathbf{P}(K \cap A(\mathbf{C}) \cap A(\mathbf{R})) &\geq 1 - \mathbf{P}(K^c) - \mathbf{P}(A(\mathbf{C})^c) - \mathbf{P}(A(\mathbf{R})^c) \\ &\geq 1 - (1 - (1-\delta/2)(1-\delta/4-0.2)) - \delta/(2n) - \delta/(4n) \\ &\geq (1-\delta/2)(1-\delta/4-0.2) - 3\delta/8 \\ &\geq (1-\delta)(1-\delta-0.2) \end{aligned}$$

for all $n \geq 2$ and $\delta \leq 0.8$.

Appendix H. Proof of Corollary 14: DFC-MC under Incoherence

H.1 Proof of DFC-Proj and DFC-RP Bounds

We begin by proving the DFC-Proj bound. Let G be the event that

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq (2+\epsilon) c_e \sqrt{mn}\Delta,$$

H be the event that

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq (2+\epsilon)\sqrt{\textstyle\sum_{i=1}^t \|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F^2},$$

$A(\mathbf{X})$ be the event that a matrix $\mathbf{X}$ is $(\frac{r\mu^2}{1-\epsilon/2}, r)$-coherent, and, for each $i \in \{1, \ldots, t\}$, B_i be the event that $\|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F > c_e\sqrt{ml}\Delta$.

Note that, by assumption,

$$\begin{aligned} l &\geq c\mu^2 r^2 (m+n) n\beta \log^2(m+n)/(s\epsilon^2) \geq cr\mu \log(n) 2\beta \log(m+n)/\epsilon^2 \\ &\geq cr\mu \log(n)((2\beta-2)\log(\bar{n}) + \log(2))/\epsilon^2 = cr\mu \log(n) \log(2\bar{n}^{2\beta-2})/\epsilon^2. \end{aligned}$$

Hence the Coherence Master Theorem (Theorem 12) guarantees that, with probability at least $1-\bar{n}^{2-2\beta}$, H holds and the event $A(\mathbf{C}_{0,i})$ holds for each i. Since G holds whenever H holds and B_i^c holds for each i, we have

$$\begin{aligned}
\mathbf{P}(G) &\geq \mathbf{P}(H \cap \textstyle\bigcap_i B_i^c) \geq \mathbf{P}(H \cap \bigcap_i A(\mathbf{C}_{0,i}) \cap \bigcap_i B_i^c) \\
&= \mathbf{P}(H \cap \textstyle\bigcap_i A(\mathbf{C}_{0,i}))\mathbf{P}(\bigcap_i B_i^c \mid H \cap \bigcap_i A(\mathbf{C}_{0,i})) \\
&= \mathbf{P}(H \cap \textstyle\bigcap_i A(\mathbf{C}_{0,i}))(1 - \mathbf{P}(\bigcup_i B_i \mid H \cap \bigcap_i A(\mathbf{C}_{0,i}))) \\
&\geq (1-\bar{n}^{2-2\beta})(1 - \textstyle\sum_i \mathbf{P}(B_i \mid A(\mathbf{C}_{0,i}))) \\
&\geq 1 - \bar{n}^{2-2\beta} - \textstyle\sum_i \mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})).
\end{aligned}$$

To prove our desired claim, it therefore suffices to show

$$\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) \leq 4\log(\bar{n})\bar{n}^{2-2\beta} + \bar{n}^{-2\beta} \leq 5\log(\bar{n})\bar{n}^{2-2\beta}$$

for each i.

For each i, let D_i be the event that $s_i < 32\mu' r(m+l)\beta' \log^2(m+l)$, where s_i is the number of revealed entries in $\mathbf{C}_{0,i}$,

$$\mu' \triangleq \frac{\mu^2 r}{1-\epsilon/2}, \qquad \text{and} \qquad \beta' \triangleq \frac{\beta \log(\bar{n})}{\log(\max(m,l))}.$$

By Theorem 10 and our choice of β',

$$\begin{aligned}
\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) &\leq \mathbf{P}(B_i \mid A(\mathbf{C}_{0,i}), D_i^c) + \mathbf{P}(D_i \mid A(\mathbf{C}_{0,i})) \\
&\leq 4\log(\max(m,l))\max(m,l)^{2-2\beta'} + \mathbf{P}(D_i) \\
&\leq 4\log(\bar{n})\bar{n}^{2-2\beta} + \mathbf{P}(D_i).
\end{aligned}$$

Further, since the support of $\mathbf{S}_0$ is uniformly distributed and of cardinality s, the variable s_i has a hypergeometric distribution with $\mathbf{E}(s_i) = \frac{sl}{n}$ and hence satisfies Hoeffding's inequality for the hypergeometric distribution (Hoeffding, 1963, Section 6):

$$\mathbf{P}(s_i \leq \mathbf{E}(s_i) - st) \leq \exp\bigl(-2st^2\bigr).$$

Since, by assumption,

$$s \geq c\mu^2 r^2 (m+n) n\beta \log^2(m+n)/(l\epsilon^2) \geq 64\mu' r(m+l)n\beta' \log^2(m+l)/l,$$

and

$$sl^2/n^2 \geq c\mu^2 r^2 (m+n) l\beta \log^2(m+n)/(n\epsilon^2) \geq 4\log(\bar{n})\beta,$$

it follows that

$$\begin{aligned}
\mathbf{P}(D_i) &= \mathbf{P}\left(s_i < \mathbf{E}(s_i) - s\left(\frac{l}{n} - \frac{32\mu' r(m+l)\beta' \log^2(m+l)}{s}\right)\right) \\
&\leq \mathbf{P}\left(s_i < \mathbf{E}(s_i) - s\left(\frac{l}{n} - \frac{l}{2n}\right)\right) = \mathbf{P}\left(s_i < \mathbf{E}(s_i) - s\frac{l}{2n}\right) \\
&\leq \exp\left(-\frac{sl^2}{2n^2}\right) \leq \exp(-2\log(\bar{n})\beta) = \bar{n}^{-2\beta}.
\end{aligned}$$

Hence, $\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) \leq 4\log(\bar{n})\bar{n}^{2-2\beta} + \bar{n}^{-2\beta}$ for each i, and the DFC-Proj result follows.

Since, $p \geq 242\ r\log(14\bar{n}^{2\beta-2})/\epsilon^2$, the DFC-RP bound follows in an identical manner from the Coherence Master Theorem (Theorem 12).

H.2 Proof of DFC-Nys Bound

For DFC-Nys, let B_C be the event that $\|\mathbf{C}_0 - \hat{\mathbf{C}}\|_F > c_e\sqrt{ml}\Delta$ and B_R be the event that $\|\mathbf{R}_0 - \hat{\mathbf{R}}\|_F > c_e\sqrt{dn}\Delta$. The Coherence Master Theorem (Theorem 12) and our choice of

$$d \geq cl\mu_0(\hat{\mathbf{C}})(2\beta-1)\log^2(4\bar{n})\bar{n}/(n\epsilon^2) \geq cl\mu_0(\hat{\mathbf{C}})\log(m)\log(4\bar{n}^{2\beta-2})/\epsilon^2$$

guarantee that, with probability at least $(1-\bar{n}^{2-2\beta})(1-\bar{n}^{2-2\beta}-0.2) \geq 1-2\bar{n}^{2-2\beta}-0.2$,

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{nys}\|_F \leq (2+3\epsilon)\sqrt{\|\mathbf{C}_0-\hat{\mathbf{C}}\|_F^2 + \|\mathbf{R}_0-\hat{\mathbf{R}}\|_F^2},$$

and both $A(\mathbf{C})$ and $A(\mathbf{R})$ hold. Moreover, since

$$d \geq cl\mu_0(\hat{\mathbf{C}})(2\beta-1)\log^2(4\bar{n})\bar{n}/(n\epsilon^2) \geq c\mu^2r^2(m+n)\bar{n}\beta\log^2(m+n)/(s\epsilon^2),$$

reasoning identical to the DFC-Proj case yields $\mathbf{P}(B_C \mid A(\mathbf{C})) \leq 4\log(\bar{n})\bar{n}^{2-2\beta}+\bar{n}^{-2\beta}$ and $\mathbf{P}(B_R \mid A(\mathbf{R})) \leq 4\log(\bar{n})\bar{n}^{2-2\beta}+\bar{n}^{-2\beta}$, and the DFC-Nys bound follows as above.

Appendix I. Proof of Corollary 16: DFC-RMF under Incoherence

I.1 Proof of DFC-Proj and DFC-RP Bounds

We begin by proving the DFC-Proj bound. Let G be the event that

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq (2+\epsilon)c_e'\sqrt{mn}\Delta$$

for the constant c_e' defined in Theorem 11, H be the event that

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq (2+\epsilon)\sqrt{\textstyle\sum_{i=1}^t \|\mathbf{C}_{0,i}-\hat{\mathbf{C}}_i\|_F^2},$$

$A(\mathbf{X})$ be the event that a matrix $\mathbf{X}$ is $(\frac{r\mu^2}{1-\epsilon/2}, r)$-coherent, and, for each $i \in \{1,\dots,t\}$, B_i be the event that $\|\mathbf{C}_{0,i}-\hat{\mathbf{C}}_i\|_F > c_e'\sqrt{ml}\Delta$.

We may take $\rho_r \leq 1$, and hence, by assumption,

$$l \geq cr^2\mu^2\beta\log^2(2\bar{n})/(\epsilon^2\rho_r) \geq cr\mu\log(n)\log(2\bar{n}^\beta)/\epsilon^2.$$

Hence the Coherence Master Theorem (Theorem 12) guarantees that, with probability at least $1-\bar{n}^{-\beta}$, H holds and the event $A(\mathbf{C}_{0,i})$ holds for each i. Since G holds whenever H holds and B_i^c holds for each i, we have

$$\begin{aligned}
\mathbf{P}(G) &\geq \mathbf{P}(H \cap \textstyle\bigcap_i B_i^c) \geq \mathbf{P}(H \cap \bigcap_i A(\mathbf{C}_{0,i}) \cap \bigcap_i B_i^c) \\
&= \mathbf{P}(H \cap \textstyle\bigcap_i A(\mathbf{C}_{0,i}))\mathbf{P}(\bigcap_i B_i^c \mid H \cap \bigcap_i A(\mathbf{C}_{0,i})) \\
&= \mathbf{P}(H \cap \textstyle\bigcap_i A(\mathbf{C}_{0,i}))(1-\mathbf{P}(\bigcup_i B_i \mid H \cap \bigcap_i A(\mathbf{C}_{0,i}))) \\
&\geq (1-\bar{n}^{-\beta})(1-\textstyle\sum_i \mathbf{P}(B_i \mid A(\mathbf{C}_{0,i}))) \\
&\geq 1-\bar{n}^{-\beta}-\textstyle\sum_i \mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})).
\end{aligned}$$

To prove our desired claim, it therefore suffices to show

$$\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) \leq (c_p+1)\bar{n}^{-\beta}$$

for each i.

Define $\bar{m} \triangleq \max(m, l)$ and $\beta'' \triangleq \beta \log(\bar{n})/\log(\bar{m}) \leq \beta'$. By assumption,

$$r \leq \frac{\rho_r m}{2\mu^2 r \log^2(\bar{n})} \leq \frac{\rho_r m(1-\epsilon/2)}{\mu^2 r \log^2(\bar{m})} \quad \text{and} \quad r \leq \frac{\rho_r l \epsilon^2}{c\mu^2 r \beta \log^2(2\bar{n})} \leq \frac{\rho_r l(1-\epsilon/2)}{\mu^2 r \log^2(\bar{m})}.$$

Hence, by Theorem 11 and the definitions of β' and β'',

$$\begin{aligned}
\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) &\leq \mathbf{P}\big(B_i \mid A(\mathbf{C}_{0,i}), s_i \leq (1-\rho_s\beta'')ml\big) + \mathbf{P}\big(s_i > (1-\rho_s\beta'')ml \mid A(\mathbf{C}_{0,i})\big) \\
&\leq c_p \bar{n}^{-\beta''} + \mathbf{P}\big(s_i > (1-\rho_s\beta'')ml\big) \\
&\leq c_p \bar{n}^{-\beta} + \mathbf{P}\big(s_i > (1-\rho_s\beta')ml\big),
\end{aligned}$$

where s_i is the number of corrupted entries in $\mathbf{C}_{0,i}$. Further, since the support of $\mathbf{S}_0$ is uniformly distributed and of cardinality s, the variable s_i has a hypergeometric distribution with $\mathbf{E}(s_i) = \frac{sl}{n}$ and hence satisfies Bernstein's inequality for the hypergeometric (Hoeffding, 1963, Section 6):

$$\mathbf{P}(s_i \geq \mathbf{E}(s_i) + st) \leq \exp\big(-st^2/(2\sigma^2 + 2t/3)\big) \leq \exp\big(-st^2 n/4l\big),$$

for all $0 \leq t \leq 3l/n$ and $\sigma^2 \triangleq \frac{l}{n}(1-\frac{l}{n}) \leq \frac{l}{n}$. It therefore follows that

$$\begin{aligned}
\mathbf{P}\big(s_i > (1-\rho_s\beta')ml\big) &= \mathbf{P}\left(s_i > \mathbf{E}(s_i) + s\left(\frac{(1-\rho_s\beta')ml}{s} - \frac{l}{n}\right)\right) \\
&= \mathbf{P}\left(s_i > \mathbf{E}(s_i) + s\frac{l}{n}\left(\frac{(1-\rho_s\beta')}{(1-\rho_s\beta_s)} - 1\right)\right) \\
&\leq \exp\left(-s\frac{l}{4n}\left(\frac{(1-\rho_s\beta')}{(1-\rho_s\beta_s)} - 1\right)^2\right) \\
&= \exp\left(-\frac{ml}{4}\frac{(\rho_s\beta_s - \rho_s\beta')^2}{(1-\rho_s\beta_s)}\right) \leq \bar{n}^{-\beta}
\end{aligned}$$

by our assumptions on s and l and the fact that $\frac{l}{n}\left(\frac{(1-\rho_s\beta')}{(1-\rho_s\beta_s)} - 1\right) \leq 3l/n$ whenever $4\beta_s - 3/\rho_s \leq \beta'$. Hence, $\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) \leq (c_p+1)\bar{n}^{-\beta}$ for each i, and the DFC-Proj result follows.

Since, $p \geq 242\, r\log(14\bar{n}^{\beta})/\epsilon^2$, the DFC-RP bound follows in an identical manner from the Coherence Master Theorem (Theorem 12).

I.2 Proof of DFC-Nys Bound

For DFC-Nys, let B_C be the event that $\|\mathbf{C}_0 - \hat{\mathbf{C}}\|_F > c'_e\sqrt{ml}\Delta$ and B_R be the event that $\|\mathbf{R}_0 - \hat{\mathbf{R}}\|_F > c'_e\sqrt{dn}\Delta$. The Coherence Master Theorem (Theorem 12) and our choice of $d \geq cl\mu_0(\hat{\mathbf{C}})\beta\log^2(4\bar{n})/\epsilon^2$ guarantee that, with probability at least $(1-\bar{n}^{-\beta})(1-\bar{n}^{-\beta}-0.2) \geq 1 - 2\bar{n}^{-\beta} - 0.2$,

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{nys}\|_F \leq (2+3\epsilon)\sqrt{\|\mathbf{C}_0 - \hat{\mathbf{C}}\|_F^2 + \|\mathbf{R}_0 - \hat{\mathbf{R}}\|_F^2},$$

and both $A(\mathbf{C})$ and $A(\mathbf{R})$ hold. Moreover, since

$$d \geq cl\mu_0(\hat{\mathbf{C}})\beta \log^2(4\bar{n})/\epsilon^2 \geq c\mu^2 r^2 \beta \log^2(\bar{n})/(\epsilon^2 \rho_r),$$

reasoning identical to the DFC-Proj case yields

$$\mathbf{P}(B_C \mid A(\mathbf{C})) \leq (c_p + 1)\bar{n}^{-\beta} \quad \text{and} \quad \mathbf{P}(B_R \mid A(\mathbf{R})) \leq (c_p + 1)\bar{n}^{-\beta},$$

and the DFC-Nys bound follows as above.

Appendix J. Proof of Theorem 10: Noisy MC under Incoherence

In the spirit of Candès and Plan (2010), our proof will extend the noiseless analysis of Recht (2011) to the noisy matrix completion setting. As suggested in Gross and Nesme (2010), we will obtain strengthened results, even in the noiseless case, by reasoning directly about the without-replacement sampling model, rather than appealing to a with-replacement surrogate, as done in Recht (2011).

For $\mathbf{U}_{L_0}\mathbf{\Sigma}_{L_0}\mathbf{V}_{L_0}^\top$ the compact SVD of $\mathbf{L}_0$, we let $T = \{\mathbf{U}_{L_0}\mathbf{X} + \mathbf{Y}\mathbf{V}_{L_0}^\top : \mathbf{X} \in \mathbb{R}^{r\times n}, \mathbf{Y} \in \mathbb{R}^{m\times r}\}$, $\mathcal{P}_T$ denote orthogonal projection onto the space T, and $\mathcal{P}_{T^\perp}$ represent orthogonal projection onto the orthogonal complement of T. We further define $\mathcal{I}$ as the identity operator on $\mathbb{R}^{m\times n}$ and the spectral norm of an operator $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ as $\|\mathcal{A}\|_2 = \sup_{\|\mathbf{X}\|_F \leq 1} \|\mathcal{A}(\mathbf{X})\|_F$.

We begin with a theorem providing sufficient conditions for our desired estimation guarantee.

Theorem 29 *Under the assumptions of Theorem 10, suppose that*

$$\frac{mn}{s}\left\|\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T - \frac{s}{mn}\mathcal{P}_T\right\|_2 \leq \frac{1}{2} \tag{9}$$

and that there exists a $\mathbf{Y} = \mathcal{P}_\Omega(\mathbf{Y}) \in \mathbb{R}^{m\times n}$ satisfying

$$\|\mathcal{P}_T(\mathbf{Y}) - \mathbf{U}_{L_0}\mathbf{V}_{L_0}^\top\|_F \leq \sqrt{\frac{s}{32mn}} \quad \textit{and} \quad \|\mathcal{P}_{T^\perp}(\mathbf{Y})\|_2 < \frac{1}{2}. \tag{10}$$

Then,

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F \leq 8\sqrt{\frac{2m^2n}{s} + m + \frac{1}{16}}\Delta \leq c_e''\sqrt{mn}\Delta.$$

Proof We may write $\hat{\mathbf{L}}$ as $\mathbf{L}_0 + \mathbf{G} + \mathbf{H}$, where $\mathcal{P}_\Omega(\mathbf{G}) = \mathbf{G}$ and $\mathcal{P}_\Omega(\mathbf{H}) = \mathbf{0}$. Then, under Eq. (9),

$$\|\mathcal{P}_\Omega\mathcal{P}_T(\mathbf{H})\|_F^2 = \left\langle \mathbf{H}, \mathcal{P}_T\mathcal{P}_\Omega^2\mathcal{P}_T(\mathbf{H})\right\rangle \geq \langle \mathbf{H}, \mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T(\mathbf{H})\rangle \geq \frac{s}{2mn}\|\mathcal{P}_T(\mathbf{H})\|_F^2.$$

Furthermore, by the triangle inequality, $0 = \|\mathcal{P}_\Omega(\mathbf{H})\|_F \geq \|\mathcal{P}_\Omega\mathcal{P}_T(\mathbf{H})\|_F - \|\mathcal{P}_\Omega\mathcal{P}_{T^\perp}(\mathbf{H})\|_F$. Hence, we have

$$\sqrt{\frac{s}{2mn}}\|\mathcal{P}_T(\mathbf{H})\|_F \leq \|\mathcal{P}_\Omega\mathcal{P}_T(\mathbf{H})\|_F \leq \|\mathcal{P}_\Omega\mathcal{P}_{T^\perp}(\mathbf{H})\|_F \leq \|\mathcal{P}_{T^\perp}(\mathbf{H})\|_F \leq \|\mathcal{P}_{T^\perp}(\mathbf{H})\|_*, \tag{11}$$

where the penultimate inequality follows as $\mathcal{P}_\Omega$ is an orthogonal projection operator.

Next we select $\mathbf{U}_\perp$ and $\mathbf{V}_\perp$ such that $[\mathbf{U}_{L_0}, \mathbf{U}_\perp]$ and $[\mathbf{V}_{L_0}, \mathbf{V}_\perp]$ are orthonormal and $\left\langle \mathbf{U}_\perp \mathbf{V}_\perp^\top, \mathcal{P}_{T^\perp}(\mathbf{H}) \right\rangle = \|\mathcal{P}_{T^\perp}(\mathbf{H})\|_*$ and note that

$$\begin{aligned}
\|\mathbf{L}_0 + \mathbf{H}\|_* &\geq \left\langle \mathbf{U}_{L_0}\mathbf{V}_{L_0}^\top + \mathbf{U}_\perp \mathbf{V}_\perp^\top, \mathbf{L}_0 + \mathbf{H} \right\rangle \\
&= \|\mathbf{L}_0\|_* + \left\langle \mathbf{U}_{L_0}\mathbf{V}_{L_0}^\top + \mathbf{U}_\perp \mathbf{V}_\perp^\top - \mathbf{Y}, \mathbf{H} \right\rangle \\
&= \|\mathbf{L}_0\|_* + \left\langle \mathbf{U}_{L_0}\mathbf{V}_{L_0}^\top - \mathcal{P}_T(\mathbf{Y}), \mathcal{P}_T(\mathbf{H}) \right\rangle + \left\langle \mathbf{U}_\perp \mathbf{V}_\perp^\top, \mathcal{P}_{T^\perp}(\mathbf{H}) \right\rangle - \langle \mathcal{P}_{T^\perp}(\mathbf{Y}), \mathcal{P}_{T^\perp}(\mathbf{H}) \rangle \\
&\geq \|\mathbf{L}_0\|_* - \|\mathbf{U}_{L_0}\mathbf{V}_{L_0}^\top - \mathcal{P}_T(\mathbf{Y})\|_F \|\mathcal{P}_T(\mathbf{H})\|_F + \|\mathcal{P}_{T^\perp}(\mathbf{H})\|_* - \|\mathcal{P}_{T^\perp}(\mathbf{Y})\|_2 \|\mathcal{P}_{T^\perp}(\mathbf{H})\|_* \\
&> \|\mathbf{L}_0\|_* + \frac{1}{2}\|\mathcal{P}_{T^\perp}(\mathbf{H})\|_* - \sqrt{\frac{s}{32mn}}\|\mathcal{P}_T(\mathbf{H})\|_F \\
&\geq \|\mathbf{L}_0\|_* + \frac{1}{4}\|\mathcal{P}_{T^\perp}(\mathbf{H})\|_F
\end{aligned}$$

where the first inequality follows from the variational representation of the trace norm, $\|\mathbf{A}\|_* = \sup_{\|\mathbf{B}\|_2 \leq 1} \langle \mathbf{A}, \mathbf{B} \rangle$, the first equality follows from the fact that $\langle \mathbf{Y}, \mathbf{H} \rangle = 0$ for $\mathbf{Y} = \mathcal{P}_\Omega(\mathbf{Y})$, the second inequality follows from Hölder's inequality for Schatten p-norms, the third inequality follows from Eq. (10), and the final inequality follows from Eq. (11).

Since $\mathbf{L}_0$ is feasible for Eq. (1), $\|\mathbf{L}_0\|_* \geq \|\hat{\mathbf{L}}\|_*$, and, by the triangle inequality, $\|\hat{\mathbf{L}}\|_* \geq \|\mathbf{L}_0 + \mathbf{H}\|_* - \|\mathbf{G}\|_*$. Since $\|\mathbf{G}\|_* \leq \sqrt{m}\|\mathbf{G}\|_F$ and $\|\mathbf{G}\|_F \leq \|\mathcal{P}_\Omega(\hat{\mathbf{L}} - \mathbf{M})\|_F + \|\mathcal{P}_\Omega(\mathbf{M} - \mathbf{L}_0)\|_F \leq 2\Delta$, we conclude that

$$\begin{aligned}
\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F^2 &= \|\mathcal{P}_T(\mathbf{H})\|_F^2 + \|\mathcal{P}_{T^\perp}(\mathbf{H})\|_F^2 + \|\mathbf{G}\|_F^2 \\
&\leq \left(\frac{2mn}{s} + 1\right)\|\mathcal{P}_{T^\perp}(\mathbf{H})\|_F^2 + \|\mathbf{G}\|_F^2 \\
&\leq 16\left(\frac{2mn}{s} + 1\right)\|\mathbf{G}\|_*^2 + \|\mathbf{G}\|_F^2 \\
&\leq 64\left(\frac{2m^2 n}{s} + m + \frac{1}{16}\right)\Delta^2.
\end{aligned}$$

Hence

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F \leq 8\sqrt{\frac{2m^2 n}{s} + m + \frac{1}{16}}\Delta \leq c_e'' \sqrt{mn}\Delta$$

for some constant c_e'', by our assumption on s. ■

To show that the sufficient conditions of Theorem 29 hold with high probability, we will require four lemmas. The first establishes that the operator $\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T$ is nearly an isometry on T when sufficiently many entries are sampled.

Lemma 30 *For all $\beta > 1$,*

$$\frac{mn}{s}\left\|\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_T - \frac{s}{mn}\mathcal{P}_T\right\|_2 \leq \sqrt{\frac{16\mu r(m+n)\beta \log(n)}{3s}}$$

with probability at least $1 - 2n^{2-2\beta}$ provided that $s > \frac{16}{3}\mu r(n+m)\beta \log(n)$.

The second states that a sparsely but uniformly observed matrix is close to a multiple of the original matrix under the spectral norm.

Lemma 31 *Let* $\mathbf{Z}$ *be a fixed matrix in* $\mathbb{R}^{m\times n}$. *Then for all* $\beta > 1$,

$$\left\| \left(\frac{mn}{s} \mathcal{P}_\Omega - \mathcal{I} \right)(\mathbf{Z}) \right\|_2 \leq \sqrt{\frac{8\beta mn^2 \log(m+n)}{3s}} \|\mathbf{Z}\|_\infty$$

with probability at least $1 - (m+n)^{1-\beta}$ *provided that* $s > 6\beta m \log(m+n)$.

The third asserts that the matrix infinity norm of a matrix in T does not increase under the operator $\mathcal{P}_T \mathcal{P}_\Omega$.

Lemma 32 *Let* $\mathbf{Z} \in T$ *be a fixed matrix. Then for all* $\beta > 2$

$$\left\| \frac{mn}{s} \mathcal{P}_T \mathcal{P}_\Omega(\mathbf{Z}) - \mathbf{Z} \right\|_\infty \leq \sqrt{\frac{8\beta\mu r(m+n)\log(n)}{3s}} \|\mathbf{Z}\|_\infty$$

with probability at least $1 - 2n^{2-\beta}$ *provided that* $s > \frac{8}{3}\beta\mu r(m+n)\log(n)$.

These three lemmas were proved in Recht (2011, Theorem 6, Theorem 7, and Lemma 8) under the assumption that entry locations in Ω were sampled *with* replacement. They admit identical proofs under the sampling without replacement model by noting that the referenced Noncommutative Bernstein Inequality (Recht, 2011, Theorem 4) also holds under sampling without replacement, as shown in Gross and Nesme (2010).

Lemma 30 guarantees that Eq. (9) holds with high probability. To construct a matrix $\mathbf{Y} = \mathcal{P}_\Omega(\mathbf{Y})$ satisfying Eq. (10), we consider a sampling with batch replacement scheme recommended in Gross and Nesme (2010) and developed in Chen et al. (2011). Let $\tilde{\Omega}_1, \ldots, \tilde{\Omega}_p$ be independent sets, each consisting of q random entry locations sampled without replacement, where $pq = s$. Let $\tilde{\Omega} = \cup_{i=1}^p \tilde{\Omega}_i$, and note that there exist p and q satisfying

$$q \geq \frac{128}{3}\mu r(m+n)\beta\log(m+n) \quad \text{and} \quad p \geq \frac{3}{4}\log(n/2).$$

It suffices to establish Eq. (10) under this batch replacement scheme, as shown in the next lemma.

Lemma 33 *For any location set* $\Omega_0 \subset \{1,\ldots,m\} \times \{1,\ldots,n\}$, *let* $A(\Omega_0)$ *be the event that there exists* $\mathbf{Y} = \mathcal{P}_{\Omega_0}(\mathbf{Y}) \in \mathbb{R}^{m\times n}$ *satisfying Eq. (10). If* $\Omega(s)$ *consists of* s *locations sampled uniformly without replacement and* $\tilde{\Omega}(s)$ *is sampled via batch replacement with* p *batches of size* q *for* $pq = s$, *then* $\mathbf{P}(A(\tilde{\Omega}(s))) \leq \mathbf{P}(A(\Omega(s)))$.

Proof As sketched in Gross and Nesme (2010)

$$\begin{aligned}
\mathbf{P}\Big(A(\tilde{\Omega(s)})\Big) &= \sum_{i=1}^s \mathbf{P}(|\tilde{\Omega}| = i)\mathbf{P}(A(\tilde{\Omega}(i)) \mid |\tilde{\Omega}| = i) \\
&\leq \sum_{i=1}^s \mathbf{P}(|\tilde{\Omega}| = i)\mathbf{P}(A(\Omega(i))) \\
&\leq \sum_{i=1}^s \mathbf{P}(|\tilde{\Omega}| = i)\mathbf{P}(A(\Omega(s))) = \mathbf{P}(A(\Omega(s))),
\end{aligned}$$

since the probability of existence never decreases with more entries sampled without replacement and, given the size of $\tilde{\Omega}$, the locations of $\tilde{\Omega}$ are conditionally distributed uniformly (without replacement). ■

We now follow the construction of Recht (2011) to obtain $\mathbf{Y} = \mathcal{P}_{\tilde{\Omega}}(\mathbf{Y})$ satisfying Eq. (10). Let $\mathbf{W}_0 = \mathbf{U}_{L_0}\mathbf{V}_{L_0}^\top$ and define $\mathbf{Y}_k = \frac{mn}{q}\sum_{j=1}^{k}\mathcal{P}_{\tilde{\Omega}_j}(\mathbf{W}_{j-1})$ and $\mathbf{W}_k = \mathbf{U}_{L_0}\mathbf{V}_{L_0}^\top - \mathcal{P}_T(\mathbf{Y}_k)$ for $k = 1, \ldots, p$. Assume that

$$\frac{mn}{q}\left\|\mathcal{P}_T\mathcal{P}_{\tilde{\Omega}_k}\mathcal{P}_T - \frac{q}{mn}\mathcal{P}_T\right\|_2 \leq \frac{1}{2} \tag{12}$$

for all k. Then

$$\|\mathbf{W}_k\|_F = \left\|\mathbf{W}_{k-1} - \frac{mn}{q}\mathcal{P}_T\mathcal{P}_{\tilde{\Omega}_k}(\mathbf{W}_{k-1})\right\|_F = \left\|(\mathcal{P}_T - \frac{mn}{q}\mathcal{P}_T\mathcal{P}_{\tilde{\Omega}_k}\mathcal{P}_T)(\mathbf{W}_{k-1})\right\|_F \leq \frac{1}{2}\|\mathbf{W}_{k-1}\|_F$$

and hence $\|\mathbf{W}_k\|_F \leq 2^{-k}\|\mathbf{W}_0\|_F = 2^{-k}\sqrt{r}$. Since

$$p \geq \frac{3}{4}\log(n/2) \geq \frac{1}{2}\log_2(n/2) \geq \log_2\sqrt{32rmn/s}\,,$$

$\mathbf{Y} \triangleq \mathbf{Y}_p$ satisfies the first condition of Eq. (10).

The second condition of Eq. (10) follows from the assumptions

$$\left\|\mathbf{W}_{k-1} - \frac{mn}{q}\mathcal{P}_T\mathcal{P}_{\tilde{\Omega}_k}(\mathbf{W}_{k-1})\right\|_\infty \leq \frac{1}{2}\|\mathbf{W}_{k-1}\|_\infty \tag{13}$$

$$\left\|\left(\frac{mn}{q}\mathcal{P}_{\tilde{\Omega}_k} - \mathcal{I}\right)(\mathbf{W}_{k-1})\right\|_2 \leq \sqrt{\frac{8mn^2\beta\log(m+n)}{3q}}\|\mathbf{W}_{k-1}\|_\infty \tag{14}$$

for all k, since Eq. (13) implies $\|\mathbf{W}_k\|_\infty \leq 2^{-k}\|\mathbf{U}_{L_0}\mathbf{V}_{L_0}^\top\|_\infty$, and thus

$$\begin{aligned}
\|\mathcal{P}_{T^\perp}(\mathbf{Y}_p)\|_2 &\leq \sum_{j=1}^{p}\left\|\frac{mn}{q}\mathcal{P}_{T^\perp}\mathcal{P}_{\tilde{\Omega}_j}(\mathbf{W}_{j-1})\right\|_2 = \sum_{j=1}^{p}\left\|\mathcal{P}_{T^\perp}(\frac{mn}{q}\mathcal{P}_{\tilde{\Omega}_j}(\mathbf{W}_{j-1}) - \mathbf{W}_{j-1})\right\|_2 \\
&\leq \sum_{j=1}^{p}\left\|(\frac{mn}{q}\mathcal{P}_{\tilde{\Omega}_j} - \mathcal{I})(\mathbf{W}_{j-1})\right\|_2 \\
&\leq \sum_{j=1}^{p}\sqrt{\frac{8mn^2\beta\log(m+n)}{3q}}\|\mathbf{W}_{j-1}\|_\infty \\
&= 2\sum_{j=1}^{p}2^{-j}\sqrt{\frac{8mn^2\beta\log(m+n)}{3q}}\|\mathbf{U}_W\mathbf{V}_W^\top\|_\infty < \sqrt{\frac{32\mu rn\beta\log(m+n)}{3q}} < 1/2
\end{aligned}$$

by our assumption on q. The first line applies the triangle inequality; the second holds since $\mathbf{W}_{j-1} \in T$ for each j; the third follows because $\mathcal{P}_{T^\perp}$ is an orthogonal projection; and the final line exploits (μ, r)-coherence.

We conclude by bounding the probability of any assumed event failing. Lemma 30 implies that Eq. (9) fails to hold with probability at most $2n^{2-2\beta}$. For each k, Eq. (12) fails to hold with probability at most $2n^{2-2\beta}$ by Lemma 30, Eq. (13) fails to hold with probability at most $2n^{2-2\beta}$ by Lemma 32, and Eq. (14) fails to hold with probability at most $(m+n)^{1-2\beta}$ by Lemma 31. Hence, by the union bound, the conclusion of Theorem 29 holds with probability at least

$$1 - 2n^{2-2\beta} - \frac{3}{4}\log(n/2)(4n^{2-2\beta} + (m+n)^{1-2\beta}) \geq 1 - \frac{15}{4}\log(n)n^{2-2\beta} \geq 1 - 4\log(n)n^{2-2\beta}.$$

Appendix K. Proof of Lemma 17: Conservation of Non-Spikiness

By assumption,

$$\mathbf{L}_C\mathbf{L}_C^\top = \sum_{a=1}^{l} \mathbf{L}^{(j_a)}(\mathbf{L}^{(j_a)})^\top$$

where $\{j_1, \ldots, j_l\}$ are random indices drawn uniformly and without replacement from $\{1, \ldots, n\}$. Hence, we have that

$$\begin{aligned}\mathrm{E}\Big[\|\mathbf{L}_C\|_F^2\Big] = \mathrm{E}\Big[\mathrm{Tr}\Big[\mathbf{L}_C\mathbf{L}_C^\top\Big]\Big] &= \mathrm{Tr}\left[\mathrm{E}\left[\sum_{a=1}^{l}\mathbf{L}^{(j_a)}(\mathbf{L}^{(j_a)})^\top\right]\right] \\ &= \mathrm{Tr}\left[\sum_{a=1}^{l}\frac{1}{n}\sum_{j=1}^{n}\mathbf{L}^{(j)}(\mathbf{L}^{(j)})^\top\right] = \frac{l}{n}\mathrm{Tr}\Big[\mathbf{L}\mathbf{L}^\top\Big] = \frac{l}{n}\|\mathbf{L}\|_F^2.\end{aligned}$$

Since $\|\mathbf{L}^{(j)}\|^4 \leq m^2\|\mathbf{L}\|_\infty^4$ for all $j \in \{1, \ldots, n\}$, Hoeffding's inequality for sampling without replacement (Hoeffding, 1963, Section 6) implies

$$\begin{aligned}\mathbf{P}\Big((1-\epsilon)(l/n)\|\mathbf{L}\|_F^2 \geq \|\mathbf{L}_C\|_F^2\Big) &\leq \exp\Big(-2\epsilon^2\|\mathbf{L}\|_F^4 l^2/(n^2 l m^2\|\mathbf{L}\|_\infty^4)\Big) \\ &= \exp\big(-2\epsilon^2 l/\alpha^4(\mathbf{L})\big) \leq \delta,\end{aligned}$$

by our choice of l. Hence,

$$\sqrt{l}\frac{1}{\|\mathbf{L}_C\|_F} \leq \frac{\sqrt{n}}{\sqrt{1-\epsilon}}\frac{1}{\|\mathbf{L}\|_F}$$

with probability at least $1-\delta$. Since, $\|\mathbf{L}_C\|_\infty \leq \|\mathbf{L}\|_\infty$ almost surely, we have that

$$\alpha(\mathbf{L}_C) = \frac{\sqrt{ml}\|\mathbf{L}_C\|_\infty}{\|\mathbf{L}_C\|_F} \leq \frac{\sqrt{mn}\|\mathbf{L}\|_\infty}{\sqrt{1-\epsilon}\|\mathbf{L}\|_F} = \frac{\alpha(\mathbf{L})}{\sqrt{1-\epsilon}}$$

with probability at least $1-\delta$ as desired.

Appendix L. Proof of Theorem 18: Column Projection under Non-Spikiness

We now give a proof of Theorem 18. While the results of this section are stated in terms of i.i.d. with-replacement sampling of columns and rows, a simple argument due to (Hoeffding,

1963, Section 6) implies the same conclusions when columns and rows are sampled without replacement.

Our proof builds upon two key results from the randomized matrix approximation literature. The first relates column projection to randomized matrix multiplication:

Theorem 34 (Theorem 2 of Drineas et al. 2006b) *Let $\mathbf{G} \in \mathbb{R}^{m\times l}$ be a matrix of l columns of $\mathbf{A} \in \mathbb{R}^{m\times n}$, and let r be a nonnegative integer. Then,*

$$\|\mathbf{A} - \mathbf{G}_r\mathbf{G}_r^+\mathbf{A}\|_F \leq \|\mathbf{A} - \mathbf{A}_r\|_F + \sqrt{r}\|\mathbf{A}\mathbf{A}^\top - (n/l)\mathbf{G}\mathbf{G}^\top\|_F.$$

The second allows us to bound $\|\mathbf{A}\mathbf{A}^\top - (n/l)\mathbf{G}\mathbf{G}^\top\|_F$ in probability when entries are bounded:

Lemma 35 (Lemma 2 of Drineas et al. 2006a) *Given a failure probability $\delta \in (0, 1]$ and matrices $\mathbf{A} \in \mathbb{R}^{m\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$ with $\|\mathbf{A}\|_\infty \leq b$ and $\|\mathbf{B}\|_\infty \leq b$, suppose that $\mathbf{G}$ is a matrix of l columns drawn uniformly with replacement from $\mathbf{A}$ and that $\mathbf{H}$ is a matrix of the corresponding l rows of $\mathbf{B}$. Then, with probability at least $1 - \delta$,*

$$|(\mathbf{A}\mathbf{B})_{ij} - (n/l)(\mathbf{G}\mathbf{H})_{ij}| \leq \frac{kb^2}{\sqrt{l}}\sqrt{8\log(2mn/\delta)} \quad \forall i, j.$$

Under our assumption, $\|\mathbf{M}\|_\infty$ is bounded by $\alpha/\sqrt{mn}$. Hence, Lemma 35 with $\mathbf{A} = \mathbf{M}$ and $\mathbf{B} = \mathbf{M}^\top$ guarantees

$$\|\mathbf{M}\mathbf{M}^\top - (n/l)\mathbf{C}\mathbf{C}^\top\|_F^2 \leq \frac{m^2n^2\alpha^4 8\log(2mn/\delta)}{m^2n^2l} \leq \epsilon^2/r$$

with probability at least $1 - \delta$, by our choice of l.

Now, Theorem 34 implies that

$$\begin{aligned}\|\mathbf{M} - \mathbf{C}\mathbf{C}^+\mathbf{M}\|_F \leq \|\mathbf{M} - \mathbf{C}_r\mathbf{C}_r^+\mathbf{M}\|_F &\leq \|\mathbf{M} - \mathbf{M}_r\|_F + \sqrt{r}\|\mathbf{M}\mathbf{M}^\top - (n/l)\mathbf{C}\mathbf{C}^\top\|_F \\ &\leq \|\mathbf{M} - \mathbf{L}\|_F + \epsilon\end{aligned}$$

with probability at least $1 - \delta$, as desired.

Appendix M. Proof of Theorem 20: Spikiness Master Theorem

Define $A(\mathbf{X})$ as the event that a matrix $\mathbf{X}$ is $(\alpha\sqrt{1 + \epsilon/(4\sqrt{r})})$-spiky. Since $\sqrt{1 + \epsilon/(4\sqrt{r})} \leq \sqrt{1.25}$ for all $\epsilon \in (0, 1]$ and $r \geq 1$, $\mathbf{X}$ is $(\sqrt{1.25}\alpha)$-spiky whenever $A(\mathbf{X})$ holds.

Let $\mathbf{L}_0 = [\mathbf{C}_{0,1}, \ldots, \mathbf{C}_{0,t}]$ and $\tilde{\mathbf{L}} = [\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t]$, and define H as the event $\|\tilde{\mathbf{L}} - \hat{\mathbf{L}}^{proj}\|_F \leq \|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F + \epsilon$. When H holds, we have that

$$\begin{aligned}\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq \|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F + \|\tilde{\mathbf{L}} - \hat{\mathbf{L}}^{proj}\|_F &\leq 2\|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F + \epsilon \\ &= 2\sqrt{\textstyle\sum_{i=1}^t \|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F^2} + \epsilon,\end{aligned}$$

by the triangle inequality, and hence it suffices to lower bound $\mathbf{P}(H \cap \bigcap_i A(\mathbf{C}_{0,i}))$.

By assumption,

$$l \geq 13r\alpha^4\log(4mn/\delta)/\epsilon^2 \geq \alpha^4\log(2n/\delta)/(2\tilde{\epsilon}^2)$$

where $\tilde{\epsilon} \triangleq \epsilon/(5\sqrt{r})$. Hence, for each i, Lemma 17 implies that $\alpha(\mathbf{C}_{0,i}) \leq \alpha/\sqrt{1-\tilde{\epsilon}}$ with probability at least $1-\delta/(2n)$. Since

$$(1-\epsilon/(5\sqrt{r}))(1+\epsilon/(4\sqrt{r})) = 1+\epsilon(1-\epsilon/\sqrt{r})/(20\sqrt{r}) \geq 1$$

it follows that

$$\frac{1}{\sqrt{1-\tilde{\epsilon}}} = \frac{1}{\sqrt{1-\epsilon/(5\sqrt{r})}} \leq \sqrt{1+\epsilon/(4\sqrt{r})},$$

so that each event $A(\mathbf{C}_{0,i})$ also holds with probability at least $1-\delta/(2n)$.

Our assumption that $\|\hat{\mathbf{C}}_i\|_\infty \leq \sqrt{1.25}\alpha/\sqrt{mn}$ for all i implies that $\|\tilde{\mathbf{L}}\|_\infty \leq \sqrt{1.25}\alpha/\sqrt{mn}$. Our choice of l, with a factor of $\log(4mn/\delta)$, therefore implies that H holds with probability at least $1-\delta/2$ by Theorem 18. Hence, by the union bound,

$$\mathbf{P}(H \cap \textstyle\bigcap_i A(\mathbf{C}_{0,i})) \geq 1 - \mathbf{P}(H^c) - \sum_i \mathbf{P}(A(\mathbf{C}_{0,i})^c) \geq 1 - \delta/2 - t\delta/(2n) \geq 1-\delta.$$

To establish the DFC-RP bound, redefine H as the event $\|\tilde{\mathbf{L}} - \mathbf{L}^{rp}\|_F \leq (2+\epsilon)\|\mathbf{L}_0 - \tilde{\mathbf{L}}\|_F$. Since $p \geq 242\, r\log(14/\delta)/\epsilon^2$, H holds with probability at least $1-\delta/2$ by Corollary 9, and the DFC-RP bound follows as above.

Appendix N. Proof of Corollary 22: Noisy MC under Non-Spikiness

N.1 Proof of DFC-Proj Bound

We begin by proving the DFC-Proj bound. Let G be the event that

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq 2\sqrt{c_1 \max((l/n)\nu^2, 1)/\beta} + \epsilon,$$

H be the event that

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq 2\sqrt{\textstyle\sum_{i=1}^t \|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F^2} + \epsilon,$$

$A(\mathbf{X})$ be the event that a matrix $\mathbf{X}$ is $(\sqrt{1.25}\alpha)$-spiky, and, for each $i \in \{1,\ldots,t\}$, B_i be the event that $\|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F^2 > (l/n)c_1 \max\big((l/n)\nu^2, 1\big)/\beta$.

By definition, $\|\hat{\mathbf{C}}_i\|_\infty \leq \sqrt{1.25}\alpha/\sqrt{ml}$ for all i. Furthermore, we have assumed that

$$\begin{aligned} l &\geq 13(c_3+1)\sqrt{\frac{(m+n)\log(m+n)\beta}{s}} nr\alpha^4 \log(4mn)/\epsilon^2 \\ &\geq 13r\alpha^4(\log(4mn) + c_3\log(m+n))/\epsilon^2 \geq 13r\alpha^4 \log(4mn(m+l)^{c_3})/\epsilon^2. \end{aligned}$$

Hence the Spikiness Master Theorem (Theorem 20) guarantees that, with probability at least $1-\exp(-c_3\log(m+l))$, H holds and the event $A(\mathbf{C}_{0,i})$ holds for each i. Since G holds whenever H holds and B_i^c holds for each i, we have

$$\begin{aligned} \mathbf{P}(G) &\geq \mathbf{P}(H \cap \textstyle\bigcap_i B_i^c) \geq \mathbf{P}(H \cap \bigcap_i A(\mathbf{C}_{0,i}) \cap \bigcap_i B_i^c) \\ &= \mathbf{P}(H \cap \textstyle\bigcap_i A(\mathbf{C}_{0,i}))\mathbf{P}(\bigcap_i B_i^c \mid H \cap \bigcap_i A(\mathbf{C}_{0,i})) \\ &= \mathbf{P}(H \cap \textstyle\bigcap_i A(\mathbf{C}_{0,i}))(1 - \mathbf{P}(\bigcup_i B_i \mid H \cap \bigcap_i A(\mathbf{C}_{0,i}))) \\ &\geq (1-\exp(-c_3\log(m+l)))(1 - \textstyle\sum_i \mathbf{P}(B_i \mid A(\mathbf{C}_{0,i}))) \\ &\geq 1 - (c_2+1)\exp(-c_3\log(m+l)) - \textstyle\sum_i \mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})). \end{aligned}$$

To prove our desired claim, it therefore suffices to show

$$\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) \leq (c_2 + 1)\exp(-c_3 \log(m + l))$$

for each i.

For each i, let D_i be the event that $s_i < 1.25\alpha^2\beta(n/l)r(m + l)\log(m + l)$, where s_i is the number of revealed entries in $\mathbf{C}_{0,i}$. Since $\text{rank}(\mathbf{C}_{0,i}) \leq \text{rank}(\mathbf{L}_0) = r$ and $\|\mathbf{C}_{0,i}\|_F \leq \|\mathbf{L}_0\|_F \leq 1$, Corollary 19 implies that

$$\begin{aligned}\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) &\leq \mathbf{P}(B_i \mid A(\mathbf{C}_{0,i}), D_i^c) + \mathbf{P}(D_i \mid A(\mathbf{C}_{0,i})) \\ &\leq c_2 \exp(-c_3 \log(m + l)) + \mathbf{P}(D_i). \end{aligned} \tag{15}$$

Further, since the support of $\mathbf{S}_0$ is uniformly distributed and of cardinality s, the variable s_i has a hypergeometric distribution with $\mathbf{E}(s_i) = \frac{sl}{n}$ and hence satisfies Hoeffding's inequality for the hypergeometric distribution (Hoeffding, 1963, Section 6):

$$\mathbf{P}(s_i \leq \mathbf{E}(s_i) - st) \leq \exp\bigl(-2st^2\bigr).$$

Our assumption on l implies that

$$\begin{aligned}\frac{l}{n} &\geq 169(c_3 + 1)^2\alpha^8\beta\frac{n}{ls}r^2(m + n)\log(m + n)\log^2(4mn)/\epsilon^4 \\ &\geq 1.25\alpha^2\beta\frac{n}{ls}r(m + l)\log(m + l) + \sqrt{c_3 \log(m + l)/(2s)},\end{aligned}$$

and therefore

$$\begin{aligned}\mathbf{P}(D_i) &= \mathbf{P}\left(s_i < \mathbf{E}(s_i) - s\left(\frac{l}{n} - 1.25\alpha^2\beta\frac{n}{ls}r(m + l)\log(m + l)\right)\right) \\ &= \mathbf{P}\left(s_i < \mathbf{E}(s_i) - s\sqrt{c_3 \log(m + l)/(2s)}\right) \\ &\leq \exp(-2sc_3 \log(m + l)/(2s)) = \exp(-c_3 \log(m + l)).\end{aligned}$$

Combined with Eq. (15), this yields $\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) \leq (c_2 + 1)\exp(-c_3 \log(m + l))$ for each i, and the DFC-Proj result follows.

N.2 Proof of DFC-RP Bound

Let G be the event that

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{rp}\|_F \leq (2 + \epsilon)\sqrt{c_1 \max((l/n)\nu^2, 1)/\beta}$$

and H be the event that

$$\|\mathbf{L}_0 - \hat{\mathbf{L}}^{rp}\|_F \leq (2 + \epsilon)\sqrt{\textstyle\sum_{i=1}^t \|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F^2}.$$

Since $p \geq 242\ r\log(14(m + l)^{c_3})/\epsilon^2$, the DFC-RP bound follows in an identical manner from the Spikiness Master Theorem (Theorem 20).

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. In *International Conference on Machine Learning*, 2011.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

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

V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Sparse and low-rank matrix decompositions. In *Allerton Conference on Communication, Control, and Computing*, 2009.

Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi. Robust matrix completion and corrupted columns. In *International Conference on Machine Learning*, 2011.

A. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *WWW*, 2007.

P. Drineas, R. Kannan, and M. W. Mahoney. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM J. Comput.*, 36(1):158–183, 2006a.

P. Drineas, R. Kannan, and M. W. Mahoney. Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM J. Comput.*, 36(1):132–157, 2006b.

P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30:844–881, 2008.

B. Recht F. Niu, C. Ré, and S. J. Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In *NIPS*, 2011.

A. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. In *Foundations of Computer Science*, 1998.

R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *KDD*, 2011.

S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin. A theory of pseudoskeleton approximations. *Linear Algebra and its Applications*, 261(1-3):1 – 21, 1997.

D. Gross and V. Nesme. Note on sampling without replacing from a finite collection of matrices. *CoRR*, abs/1001.2738, 2010.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

P. D. Hoff. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100:286–295, March 2005.

D. Hsu. `http://www.cs.columbia.edu/~djhsu/papers/randmatrix-errata.txt`, 2012.

D. Hsu, S. Kakade, and T. Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electron. Commun. Probab.*, 17:no. 14, 1–13, 2012.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 99:2057–2078, 2010.

Y. Koren, R. M. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

S. Kumar, M. Mohri, and A. Talwalkar. On sampling-based approximate spectral decomposition. In *International Conference on Machine Learning*, 2009a.

S. Kumar, M. Mohri, and A. Talwalkar. Ensemble Nyström method. In *Advances in Neural Information Processing Systems*, 2009b.

E. Liberty. *Accelerated Dense Random Projections*. Ph.D. thesis, computer science department, Yale University, New Haven, CT, 2009.

Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. UIUC Technical Report UILU-ENG-09-2215, 2009a.

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

S. Ma, D. Goldfarb, and L. Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353, 2011.

L. Mackey, A. Talwalkar, and M. I. Jordan. Divide-and-conquer matrix factorization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1134–1142. 2011.

M. W. Mahoney and P. Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

K. Min, Z. Zhang, J. Wright, and Y. Ma. Decomposing background topics from keywords by principal component pursuit. In *Conference on Information and Knowledge Management*, 2010.

M. Mohri and A. Talwalkar. Can matrix coherence be efficiently and accurately estimated? In *Conference on Artificial Intelligence and Statistics*, 2011.

Y. Mu, J. Dong, X. Yuan, and S. Yan. Accelerated low-rank visual recovery by random projection. In *Conference on Computer Vision and Pattern Recognition*, 2011.

S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.*, 13:1665–1697, 2012.

E. J. Nyström. Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. *Acta Mathematica*, 54(1):185–204, 1930.

C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala. Latent Semantic Indexing: a probabilistic analysis. In *Principles of Database Systems*, 1998.

Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In *Conference on Computer Vision and Pattern Recognition*, 2010.

B. Recht. Simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12:3413–3430, 2011.

B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. In *Optimization Online*, 2011.

V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for Principal Component Analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009.

A. Talwalkar and A. Rostamizadeh. Matrix coherence and the Nyström method. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, 2010.

K. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 6(3):615–640, 2010.

M. Tygert. http://www.mathworks.com/matlabcentral/fileexchange/21524-principal-component-analysis, 2009.

J. Wang, Y. Dong, X. Tong, Z. Lin, and B. Guo. Kernel Nyström method for light transport. *ACM Transactions on Graphics*, 28(3), 2009.

C.K. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, 2000.

H.-F. Yu, C.-J. Hsieh, S. Si, and I. Dhillon. Scalable coordinate descent approaches to parallel matrix factorization for recommender systems. In *ICDM*, 2012.

Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the netflix prize. In *Proceedings of the 4th international conference on Algorithmic Aspects in Information and Management*, 2008.

Z. Zhou, X. Li, J. Wright, E. J. Candès, and Y. Ma. Stable principal component pursuit. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 1518 –1522, 2010.