Divide-and-Conquer Matrix Factorization

Lester Mackey[a] Ameet Talwalkar[a] Michael I. Jordan[a, b]

[a] Department of Electrical Engineering and Computer Science, UC Berkeley
[b] Department of Statistics, UC Berkeley

Abstract

This work introduces Divide-Factor-Combine (DFC), a parallel divide-and-conquer framework for noisy matrix factorization. DFC divides a large-scale matrix factorization task into smaller subproblems, solves each subproblem in parallel using an arbitrary base matrix factorization algorithm, and combines the subproblem solutions using techniques from randomized matrix approximation. Our experiments with collaborative filtering, video background modeling, and simulated data demonstrate the near-linear to super-linear speed-ups attainable with this approach. Moreover, our analysis shows that DFC enjoys high-probability recovery guarantees comparable to those of its base algorithm.

1 Introduction

The goal in matrix factorization is to recover a low-rank matrix from irrelevant noise and corruption. We focus on two instances of the problem: noisy matrix completion, i.e., recovering a low-rank matrix from a small subset of noisy entries, and noisy robust matrix factorization [2, 3, 4], i.e., recovering a low-rank matrix from corruption by noise and outliers of arbitrary magnitude. Examples of the matrix completion problem include collaborative filtering for recommender systems, link prediction for social networks, and click prediction for web search, while applications of robust matrix factorization arise in video surveillance [2], graphical model selection [4], document modeling [17], and image alignment [21].

These two classes of matrix factorization problems have attracted significant interest in the research community. In particular, convex formulations of noisy matrix factorization have been shown to admit strong theoretical recovery guarantees [1, 2, 3, 20], and a variety of algorithms (e.g., [15, 16, 23]) have been developed for solving both matrix completion and robust matrix factorization via convex relaxation. Unfortunately, these methods are inherently sequential and all rely on the repeated and costly computation of truncated SVDs, factors that limit the scalability of the algorithms.

To improve scalability and leverage the growing availability of parallel computing architectures, 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 any base matrix factorization algorithm, and combines the solutions to the subproblem using efficient techniques from randomized matrix approximation. The inherent parallelism of DFC allows for near-linear to superlinear speedups in practice, while our theory provides high-probability recovery guarantees for DFC comparable to those enjoyed by its base algorithm.

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. To illustrate the significant speed-up and robustness of DFC and to highlight the effectiveness of DFC ensembling, we present experimental results on collaborative filtering, video background modeling, and simulated data in Section 3. Our theoretical analysis follows in Section 4. There, we establish high-probability noisy recovery guarantees for DFC that rest upon a novel analysis of randomized matrix approximation and a new recovery result for noisy matrix completion.

Notation For $\mathbf{M} \in \mathbb{R}^{m\times n}$, we define $\mathbf{M}_{(i)}$ as the ith row vector and $\mathbf{M}_{ij}$ as the ijth entry. If $\text{rank}(\mathbf{M}) = r$, 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 and let $\|\cdot\|$ represent the ℓ_2 norm of a vector.

2 The Divide-Factor-Combine Framework

In this section, we present our 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

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

Our goal is to recover 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

Algorithms 1 and 2 summarize two canonical examples of the general Divide-Factor-Combine framework that we refer to as DFC-PROJ and DFC-NYS. Each algorithm has three simple steps:

(D step) **Divide input matrix into submatrices:** DFC-PROJ randomly partitions $\mathcal{P}_\Omega(\mathbf{M})$ into t l-column submatrices, $\{\mathcal{P}_\Omega(\mathbf{C}_1), \ldots, \mathcal{P}_\Omega(\mathbf{C}_t)\}$[1], 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 performs t parallel submatrix factorizations, while DFC-NYS performs two such parallel factorizations. Standard base MF algorithms output the low-rank approximations $\{\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t\}$ for DFC-PROJ and $\hat{\mathbf{C}}$, 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 $\hat{\mathbf{C}}_1$, while DFC-NYS forms the low-rank 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

Our divide-and-conquer algorithms rely on two methods that generate randomized low-rank approximations to an arbitrary matrix $\mathbf{M}$ from submatrices of $\mathbf{M}$.

[1] For ease of discussion, we assume that $\text{mod}(n, t) = 0$, and hence, $l = n/t$. Note that for arbitrary n and t, $\mathcal{P}_\Omega(\mathbf{M})$ 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: $\mathcal{P}_\Omega(\mathbf{M})$, $t$
$\{\mathcal{P}_\Omega(\mathbf{C}_i)\}_{1\le i\le t}$ = SAMPCOL($\mathcal{P}_\Omega(\mathbf{M})$, $t$)
do in parallel
    $\hat{\mathbf{C}}_1$ = BASE-MF-ALG($\mathcal{P}_\Omega(\mathbf{C}_1)$)
        $\vdots$
    $\hat{\mathbf{C}}_t$ = BASE-MF-ALG($\mathcal{P}_\Omega(\mathbf{C}_t)$)
end do
$\hat{\mathbf{L}}^{proj}$ = COLPROJECTION($\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t$)
```

Algorithm 2 DFC-NYS[a]

```
Input: $\mathcal{P}_\Omega(\mathbf{M})$, $l$, $d$
$\mathcal{P}_\Omega(\mathbf{C})$ , $\mathcal{P}_\Omega(\mathbf{R})$ = SAMPCOLROW($\mathcal{P}_\Omega(\mathbf{M})$, $l$, $d$)
do in parallel
    $\hat{\mathbf{C}}$ = BASE-MF-ALG($\mathcal{P}_\Omega(\mathbf{C})$)
    $\hat{\mathbf{R}}$ = BASE-MF-ALG($\mathcal{P}_\Omega(\mathbf{R})$)
end do
$\hat{\mathbf{L}}^{nys}$ = GENNYSTRÖM ($\hat{\mathbf{C}}$, $\hat{\mathbf{R}}$)
```

[a] When $\mathbf{Q}$ is a submatrix of $\mathbf{M}$ we abuse notation and define $\mathcal{P}_\Omega(\mathbf{Q})$ as the corresponding submatrix of $\mathcal{P}_\Omega(\mathbf{M})$.

Column Projection This approximation, introduced by Frieze et al. [7], is derived from column sampling of $\mathbf{M}$. We begin by sampling $l < n$ columns uniformly without replacement and let $\mathbf{C}$ be the $m \times l$ matrix of sampled columns. Then, column projection uses $\mathbf{C}$ to generate a "matrix projection" approximation [13] of $\mathbf{M}$ as follows:

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

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

Generalized Nyström Method The standard Nyström method is often used to speed up large-scale learning applications involving symmetric positive semidefinite (SPSD) matrices [24] and has been generalized for arbitrary real-valued matrices [8]. In particular, after sampling columns to obtain $\mathbf{C}$, imagine that we independently sample $d < m$ rows uniformly without replacement. Let $\mathbf{R}$ be the $d \times n$ matrix of sampled rows and $\mathbf{W}$ be the $d \times l$ matrix formed from the intersection of the sampled rows and columns. Then, the generalized Nyström method uses $\mathbf{C}, \mathbf{W}$, and $\mathbf{R}$ to compute an "spectral reconstruction" approximation [13] of $\mathbf{M}$ as follows:

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

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 $\mathbf{U}_W^{\top}\mathbf{R}$.

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, since the outputs of standard MF algorithms are returned in factored form. Indeed, the column projection step of DFC-PROJ requires only O$(mk^2 + lk^2)$ time for $k \triangleq \max_i k_{C_i}$: 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 $\bar{k} \triangleq \max(k_C, k_R)$. Hence, DFC divides the expensive task of matrix factorization into smaller subproblems that can be executed in parallel and efficiently combines the low-rank, factored results.

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 [14]. As such, we propose ensemble variants of the DFC algorithms that demonstrably reduce recovery 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-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. 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 3 highlights the empirical effectiveness of ensembling.

3 Experimental Evaluation

We now explore the accuracy and speed-up of DFC on a variety of simulated and real-world datasets. We use state-of-the-art matrix factorization algorithms in our experiments: the Accelerated Proximal Gradient (APG) algorithm of [23] as our base noisy MC algorithm and the APG algorithm of [15] as our base noisy RMF algorithm. In all experiments, we use the default parameter settings suggested by [23] and [15], measure recovery error via root mean square error (RMSE), and report parallel running times for DFC. We moreover compare against two baseline methods: APG used on the full matrix $\mathbf{M}$ and PARTITION, which performs matrix factorization on t submatrices just like DFC-PROJ but omits the final column projection step.

3.1 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, e.g., [2, 12, 25]. 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 recovered low-rank matrix, and all reported results are averages over five trials.

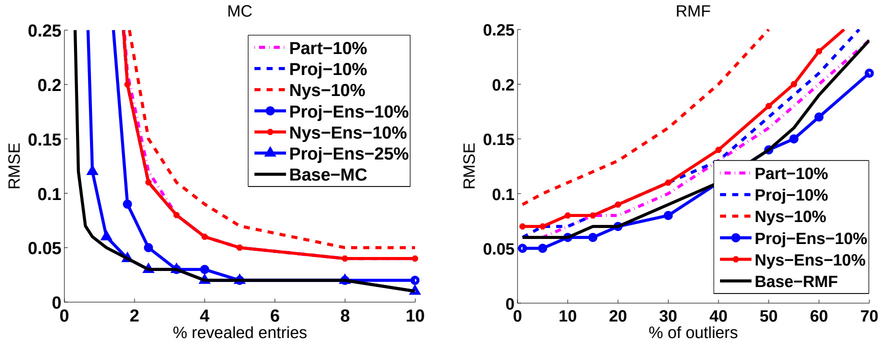


Figure 1: Recovery error of DFC relative to base algorithms.

We first explored the recovery error of DFC as a function of s, using $(m = 10\text{K}, r = 10)$ with varying observation sparsity for MC and $(m = 1\text{K}, r = 10)$ with a varying percentage of outliers for RMF. The results are summarized in Figure 1.[2] In both MC and RMF, the gaps in recovery between APG and DFC are small when sampling only 10% of rows and columns. Moreover, DFC-PROJ-ENS in particular consistently outperforms PARTITION and DFC-NYS-ENS and matches the performance of APG for most settings of s.

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 recovery 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.

[2] In the left-hand plot of Figure 1, the lines for Proj-10% and Proj-Ens-10% overlap.

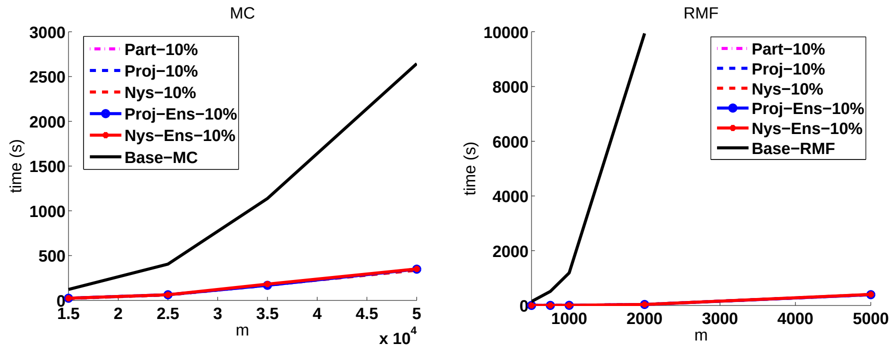


Figure 2: Speed-up of DFC relative to base algorithms.

3.2 Collaborative Filtering

Collaborative filtering for recommender systems is one prevalent real-world application of noisy matrix completion. A collaborative filtering dataset 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 datasets: MovieLens 10M[3] ($m = 4$K, $n = 6$K, $s > 10$M) and the Netflix Prize dataset[4] ($m = 18$K, $n = 480$K, $s > 100$M). To generate test sets drawn from the training distribution, for each dataset, 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 3.2. Notably, DFC-Proj, DFC-Proj-Ens, and DFC-Nys-Ens all outperform Partition, and DFC-Proj-Ens performs comparably to APG while providing a nearly linear parallel time speed-up. 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 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.

Table 1: Performance of DFC relative to APG on collaborative filtering tasks.

Method	**MovieLens 10M**		**Netflix**	
	RMSE	**Time**	**RMSE**	**Time**
APG	0.8005	294.3s	0.8433	2653.1s
Partition-25%	0.8146	77.4s	0.8451	689.1s
Partition-10%	0.8461	36.0s	0.8492	289.2s
DFC-Nys-25%	0.8449	77.2s	0.8832	890.9s
DFC-Nys-10%	0.8769	53.4s	0.9224	487.6s
DFC-Nys-Ens-25%	0.8085	84.5s	0.8486	964.3s
DFC-Nys-Ens-10%	0.8327	63.9s	0.8613	546.2s
DFC-Proj-25%	0.8061	77.4s	0.8436	689.5s
DFC-Proj-10%	0.8272	36.1s	0.8484	289.7s
DFC-Proj-Ens-25%	0.7944	77.4s	0.8411	689.5s
DFC-Proj-Ens-10%	0.8119	36.1s	0.8433	289.7s

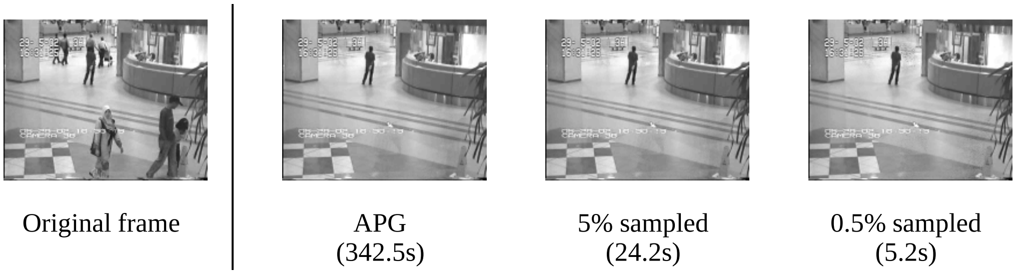
3.3 Background Modeling

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

[3] `http://www.grouplens.org/`

[4] `http://www.netflixprize.com/`

background variations, e.g., changes in illumination, are outliers ($\mathbf{S}_0$). We evaluate DFC on two videos: 'Hall' (200 frames of size 176×144) contains significant foreground variation and was studied by [2], while 'Lobby' (1546 frames of size 168×120) includes many changes in illumination (a smaller video with 250 frames was studied by [2]). We focused on DFC-PROJ-ENS, due to its superior performance in previous experiments, and measured the RMSE between the background model recovered by DFC and that of APG. On both videos, DFC-PROJ-ENS recovered 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 runtime 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,' the RMSE of DFC-PROJ-ENS-4% was 0.64, and the speed-up over APG was more than 20X, i.e., the runtime reduced from 16557s to 792s.



Original frame

APG (342.5s)

5% sampled (24.2s)

0.5% sampled (5.2s)

Figure 3: Sample 'Hall' recovery by APG, DFC-PROJ-ENS-5%, and DFC-PROJ-ENS-.5%.

4 Theoretical Analysis

Having investigated the empirical advantages of DFC, we now show that DFC admits high-probability recovery guarantees comparable to those of its base algorithm.

4.1 Matrix Coherence

Since not all matrices can be recovered from missing entries or gross outliers, recent theoretical advances have studied sufficient conditions for accurate noisy MC [3, 12, 20] and RMF [1, 25]. Most prevalent among these are *matrix coherence* conditions, which limit the extent to which the singular vectors of a matrix are correlated with the standard basis. Letting $\mathbf{e}_i$ be the ith column of the standard basis, we define two standard notions of coherence [22]:

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:*

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

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 any $\mu > 0$, we will call a matrix $\mathbf{L}$ (μ, r)*-coherent* if $\mathrm{rank}(\mathbf{L}) = r$, $\max(\mu_0(\mathbf{U}_L), \mu_0(\mathbf{V}_L)) \leq \mu$, and $\mu_1(\mathbf{L}) \leq \sqrt{\mu}$. Our analysis will focus on base MC and RMF algorithms that express their recovery 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 recovery properties.

4.2 DFC Master Theorem

We now show that the same coherence conditions that allow for accurate MC and RMF also imply high-probability recovery 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 $\|\mathcal{P}_\Omega(\mathbf{Z}_0)\|_F \leq \Delta$. We further fix any $\epsilon, \delta \in (0, 1]$ and define $A(\mathbf{X})$ as the event that a matrix $\mathbf{X}$ is $(\frac{r\mu^2}{1-\epsilon/2}, r)$-coherent. Then, our Thm. 3 provides a generic recovery bound for DFC when used in combination with an arbitrary base algorithm. The proof requires a novel, coherence-based analysis of column projection and random column sampling. These results of independent interest are presented in Appendix A.

Theorem 3. *Choose $t = n/l$ and $l \ge cr\mu \log(n)\log(2/\delta)/\epsilon^2$, where c is a fixed positive constant, and fix any $c_e \ge 0$. Under the notation of Algorithm 1, if a base MF algorithm yields $\mathbf{P}\Big(\|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F > c_e\sqrt{ml}\Delta \mid A(\mathbf{C}_{0,i})\Big) \le \delta_C$ for each i, where $\mathbf{C}_{0,i}$ is the corresponding partition of $\mathbf{L}_0$, then, with probability at least $(1-\delta)(1-t\delta_C)$,* DFC-Proj *guarantees*

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

Under Algorithm 2, if a base MF algorithm yields $\mathbf{P}\Big(\|\mathbf{C}_0 - \hat{\mathbf{C}}\|_F > c_e\sqrt{ml}\Delta \mid A(\mathbf{C})\Big) \le \delta_C$ and $\mathbf{P}\Big(\|\mathbf{R}_0 - \hat{\mathbf{R}}\|_F > c_e\sqrt{dn}\Delta \mid A(\mathbf{R})\Big) \le \delta_R$ for $d \ge cl\mu_0(\hat{\mathbf{C}})\log(m)\log(4/\delta)/\epsilon^2$, then, with probability at least $(1-\delta)(1-\delta-0.2)(1-\delta_C-\delta_R)$, DFC-Nys *guarantees*

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

To understand the conclusions of Thm. 3, consider a typical base algorithm which, when applied to $\mathcal{P}_\Omega(\mathbf{M})$, recovers an estimate $\hat{\mathbf{L}}$ satisfying $\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F \le c_e\sqrt{mn}\Delta$ with high probability. Thm. 3 asserts that, with appropriately reduced probability, DFC-Proj exhibits the same recovery error scaled by an adjustable factor of $2+\epsilon$, while DFC-Nys exhibits a somewhat smaller error scaled by $2+3\epsilon$.[5] The key take-away then 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 recovery in the noisy setting and exact recovery in the noiseless setting ($\Delta = 0$), even when entries are missing or grossly corrupted. The next two sections demonstrate how Thm. 3 can be applied to derive specific DFC recovery guarantees for noisy MC and noisy RMF. In these sections, we let $\bar{n} \triangleq \max(m,n)$.

4.3 Consequences for Noisy MC

Our first corollary of Thm. 3 shows that DFC retains the high-probability recovery guarantees of a standard MC solver while operating on matrices of much smaller dimension. Suppose that a base MC algorithm solves the following convex optimization problem, studied in [3]:

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

Then, Cor. 4 follows from a novel guarantee for noisy convex MC, proved in the appendix.

Corollary 4. *Suppose that $\mathbf{L}_0$ is (μ, r)-coherent and that s entries of $\mathbf{M}$ are observed, with locations Ω distributed uniformly. Define the oversampling parameter*

$$\beta_s \triangleq \frac{s(1-\epsilon/2)}{32\mu^2 r^2 (m+n)\log^2(m+n)},$$

and fix any target rate parameter $1 < \beta \le \beta_s$. Then, if $\|\mathcal{P}_\Omega(\mathbf{M}) - \mathcal{P}_\Omega(\mathbf{L}_0)\|_F \le \Delta$ a.s., it suffices to choose $t = n/l$ and

$$l \ge \max\left(\frac{n\beta}{\beta_s} + \sqrt{\frac{n(\beta-1)}{\beta_s}}, cr\mu\frac{\log(n)\log(2/\delta)}{\epsilon^2}\right), \quad d \ge \max\left(\frac{m\beta}{\beta_s} + \sqrt{\frac{m(\beta-1)}{\beta_s}}, cl\mu_0(\hat{\mathbf{C}})\frac{\log(m)\log(4/\delta)}{\epsilon^2}\right)$$

to achieve

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

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

with probability at least

$$\textbf{DFC-Proj:} \quad (1-\delta)(1-5t\log(\bar{n})\bar{n}^{2-2\beta}) \ge (1-\delta)(1-\bar{n}^{3-2\beta})$$

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

respectively, with c as in Thm. 3 and c'_e a positive constant.

[5] Note that 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.

Notably, Cor. 4 allows for the fraction of columns and rows sampled to decrease as the oversampling parameter β_s increases with m and n. In the best case, $\beta_s = \Theta(mn/[(m+n)\log^2(m+n)])$, and Cor. 4 requires only $\mathrm{O}(\frac{n}{m}\log^2(m+n))$ sampled columns and $\mathrm{O}(\frac{m}{n}\log^2(m+n))$ sampled rows. In the worst case, $\beta_s = \Theta(1)$, and Cor. 4 requires the number of sampled columns and rows to grow linearly with the matrix dimensions. As a more realistic intermediate scenario, consider the setting in which $\beta_s = \Theta(\sqrt{m+n})$ and thus a vanishing fraction of entries are revealed. In this setting, only $\mathrm{O}(\sqrt{m+n})$ columns and rows are required by Cor. 4.

4.4 Consequences for Noisy RMF

Our next corollary shows that DFC retains the high-probability recovery guarantees of a standard RMF solver while operating on matrices of much smaller dimension. Suppose that a base RMF algorithm solves the following convex optimization problem, studied in [25]:

$$\text{minimize}_{\mathbf{L},\mathbf{S}} \quad \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \quad \text{subject to} \quad \|\mathbf{M}-\mathbf{L}-\mathbf{S}\|_F \le \Delta,$$

with $\lambda = 1/\sqrt{\bar{n}}$. Then, Cor. 5 follows from Thm. 3 and the noisy RMF guarantee of [25, Thm. 2].

Corollary 5. *Suppose that $\mathbf{L}_0$ is (μ, r)-coherent and 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 $4\beta_s - 3/\rho_s \le \beta' \le \beta_s$. Then, if $\|\mathbf{M}-\mathbf{L}_0-\mathbf{S}_0\|_F \le \Delta$ a.s., it suffices to choose $t = n/l$ and

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

to have

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

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

with probability at least

$$\textbf{DFC-Proj:} \ (1-\delta)(1-tc_p\bar{n}^{-\beta}) \ge (1-\delta)(1-c_p\bar{n}^{1-\beta})$$

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

respectively, with c as in Thm. 3 and ρ_r, c''_e, and c_p positive constants.

Note that Cor. 5 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 high-probability noisy recovery.

5 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. We note that the contemporaneous work of [19] addresses the computational burden of noiseless RMF by reformulating a standard convex optimization problem to internally incorporate random projections. The differences between DFC and the approach of [19] highlight some of the main advantages of this work: i) DFC can be used in combination with any underlying MF algorithm, ii) DFC is trivially parallelized, and iii) DFC provably maintains the recovery guarantees of its base algorithm, even in the presence of noise.

References

[1] 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.

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

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

[4] 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.

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

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

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

[8] 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.

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

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

[11] D. Hsu, S. M. Kakade, and T. Zhang. Dimension-free tail inequalities for sums of random matrices. `arXiv:1104.1672v3[math.PR]`, 2011.

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

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

[14] S. Kumar, M. Mohri, and A. Talwalkar. Ensemble Nyström method. In *NIPS*, 2009.

[15] 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, 2009.

[16] 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.

[17] 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.

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

[19] 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.

[20] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. `arXiv:1009.2118v2[cs.IT]`, 2010.

[21] 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.

[22] B. Recht. A simpler approach to matrix completion. `arXiv:0910.0651v2[cs.IT]`, 2009.

[23] 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.

[24] C.K. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *NIPS*, 2000.

[25] Z. Zhou, X. Li, J. Wright, E. J. Candès, and Y. Ma. Stable principal component pursuit. `arXiv: 1001.2363v1[cs.IT]`, 2010.

A Analysis of Randomized Approximation Algorithms

In this section, we will establish several key properties of randomized approximation algorithms under standard coherence assumptions that 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.

A.1 Conservation of Incoherence

The following lemma bounds the μ_0 and μ_1-coherence of a uniformly sampled submatrix in terms of the coherence of the full matrix. These properties will allow for accurate submatrix completion or outlier removal using standard MC and RMF algorithms. Its proof is given in Sec. B.

Lemma 6. *Let $\mathbf{L} \in \mathbb{R}^{m\times n}$ be a rank-r matrix and $\mathbf{L}_C \in \mathbb{R}^{m\times l}$ be 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 Thm. 7, then*

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

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 $1-\delta/n$.

A.2 Randomized ℓ_2 Regression

Our next theorem shows that projection based on uniform column sampling leads to near optimal estimation in matrix regression when the covariate matrix has small coherence. The result builds upon the randomized ℓ_2 regression work of [6] and the matrix concentration analysis of [11] and immediately gives rise to estimation guarantees for column projection and the generalized Nyström method. The proof of Thm. 7 will be given in Sec. C.

Theorem 7. *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 $1-\delta-0.2$.

A first consequence of Thm. 7 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 $r\log n$. Our result generalizes Thm. 1 of [6] by providing guarantees relative to an arbitrary low-rank approximation. The proof is given in Sec. D.

Corollary 8. *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 $1-\delta$.

Thm. 7 and Cor. 8 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 Sec. E.

Corollary 9. *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 Cor. 8, 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 $(1-\delta)(1-\delta'-0.2)$.

B Proof of Lemma 6

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 11 with $\beta = 1/\mu_0(\mathbf{V}_L)$, $p_j = 1/n$ for all j, and $\mathbf{D} = \mathbf{I}\sqrt{n/l}$. When $\mathrm{rank}(\mathbf{L}_C) = \mathrm{rank}(\mathbf{L})$, Lemma 1 of [18] 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 11, 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 $\mathrm{rank}(\mathbf{L}_C) = \mathrm{rank}(\mathbf{L}) = r$, 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,

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

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

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

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

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 11 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 11, note that $\mathbf{P}_{U_L} = \mathbf{P}_{U_{L_C}}$ implies $\mathbf{U}_L \mathbf{U}_L^\top \mathbf{U}_{L_C} = \mathbf{U}_{L_C}$. We thus observe that,

$$\begin{aligned}
\mathbf{U}_{L_C} \mathbf{V}_{L_C}^\top &= \mathbf{U}_{L_C} \mathbf{\Sigma}_{L_C}^{-1} \mathbf{U}_{L_C}^\top \mathbf{L}_C \\
&= \mathbf{U}_{L_C} \mathbf{\Sigma}_{L_C}^{-1} \mathbf{U}_{L_C}^\top \mathbf{U}_L \mathbf{\Sigma}_L \mathbf{V}_l^\top \\
&= \mathbf{U}_L \mathbf{U}_L^\top \mathbf{U}_{L_C} \mathbf{\Sigma}_{L_C}^{-1} \mathbf{U}_{L_C}^\top \mathbf{U}_L \mathbf{\Sigma}_L \mathbf{V}_l^\top \,.
\end{aligned}$$

Letting $\mathbf{B} = \mathbf{U}_L^\top \mathbf{U}_{L_C} \mathbf{\Sigma}_{L_C}^{-1} \mathbf{U}_{L_C}^\top \mathbf{U}_L \mathbf{\Sigma}_L$, we have

$$\begin{aligned}
\mu_1(\mathbf{L}_C) &= \sqrt{\frac{ml}{r}} \max_{\substack{1 \le i \le m \\ 1 \le j \le l}} |\mathbf{e}_{i,m}^\top \mathbf{U}_{L_C} \mathbf{V}_{L_C}^\top \mathbf{e}_{j,l}| \\
&= \sqrt{\frac{ml}{r}} \max_{\substack{1 \le i \le m \\ 1 \le j \le l}} |\mathbf{e}_{i,m}^\top \mathbf{U}_L \mathbf{B} \mathbf{V}_l^\top \mathbf{e}_{j,l}| \\
&= \sqrt{\frac{ml}{r}} \max_{\substack{1 \le i \le m \\ 1 \le j \le l}} |\mathbf{e}_{i,m}^\top \mathbf{U}_L \mathbf{B} \mathbf{V}_L^\top \mathbf{e}_{j,n}| \\
&= \sqrt{\frac{ml}{r}} \max_{\substack{1 \le i \le m \\ 1 \le j \le l}} |\operatorname{Tr}\big[\mathbf{e}_{i,m}^\top \mathbf{U}_L \mathbf{B} \mathbf{V}_L^\top \mathbf{e}_{j,n}\big]| \\
&= \sqrt{\frac{ml}{r}} \max_{\substack{1 \le i \le m \\ 1 \le j \le l}} |\operatorname{Tr}\big[\mathbf{B} \mathbf{V}_L^\top \mathbf{e}_{j,n} \mathbf{e}_{i,m}^\top \mathbf{U}_L\big]| \\
&\le \sqrt{\frac{ml}{r}} \|\mathbf{B}\|_2 \max_{\substack{1 \le i \le m \\ 1 \le j \le l}} \|\mathbf{V}_L^\top \mathbf{e}_{j,n} \mathbf{e}_{i,m}^\top \mathbf{U}_L\|_* \,,
\end{aligned}$$

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

$$\begin{aligned}
\mu_1(\mathbf{L}_C) &\le \sqrt{\frac{ml}{r}}\|\mathbf{B}\|_2 \max_{\substack{1\le i\le m\\ 1\le j\le l}} \|\mathbf{P}_{U_L}\mathbf{e}_{i,m}\|\|\mathbf{P}_{V_L}\mathbf{e}_{j,n}\| \\
&= \sqrt{\frac{mlr^2}{mnr}}\|\mathbf{B}\|_2 \left(\sqrt{\frac{m}{r}} \max_{1\le i\le m} \|\mathbf{P}_{U_L}\mathbf{e}_{i,m}\|\right)\left(\sqrt{\frac{n}{r}} \max_{1\le j\le l} \|\mathbf{P}_{V_L}\mathbf{e}_{j,n}\|\right) \\
&\le \sqrt{\frac{mlr^2}{mnr}}\|\mathbf{B}\|_2 \left(\sqrt{\frac{m}{r}} \max_{1\le i\le m} \|\mathbf{P}_{U_L}\mathbf{e}_{i,m}\|\right)\left(\sqrt{\frac{n}{r}} \max_{1\le j\le n} \|\mathbf{P}_{V_L}\mathbf{e}_{j,n}\|\right) \\
&= \sqrt{\frac{lr}{n}}\|\mathbf{B}\|_2 \sqrt{\mu_0(\mathbf{U}_L)\mu_0(\mathbf{V}_L)}\,,
\end{aligned}$$

by the definitition of μ_0-coherence.

Next, we notice that

$$\begin{aligned}
\mathbf{B}^\top\mathbf{B} &= \mathbf{\Sigma}_L\mathbf{U}_L^\top\mathbf{U}_{L_C}\mathbf{\Sigma}_{L_C}^{-1}\mathbf{U}_{L_C}^\top\mathbf{U}_L\mathbf{U}_L^\top\mathbf{U}_{L_C}\mathbf{\Sigma}_{L_C}^{-1}\mathbf{U}_{L_C}^\top\mathbf{U}_L\mathbf{\Sigma}_L \\
&= \mathbf{\Sigma}_L\mathbf{U}_L^\top\mathbf{U}_{L_C}\mathbf{\Sigma}_{L_C}^{-1}\mathbf{U}_{L_C}^\top\mathbf{U}_{L_C}\mathbf{\Sigma}_{L_C}^{-1}\mathbf{U}_{L_C}^\top\mathbf{U}_L\mathbf{\Sigma}_L \\
&= \mathbf{\Sigma}_L\mathbf{U}_L^\top\mathbf{U}_{L_C}\mathbf{\Sigma}_{L_C}^{-2}\mathbf{U}_{L_C}^\top\mathbf{U}_L\mathbf{\Sigma}_L \\
&= \mathbf{\Sigma}_L\mathbf{U}_L^\top(\mathbf{L}_C\mathbf{L}_C^\top)^+\mathbf{U}_L\mathbf{\Sigma}_L \\
&= \mathbf{\Sigma}_L\mathbf{U}_L^\top(\mathbf{U}_L\mathbf{\Sigma}_L\mathbf{V}_l^\top\mathbf{V}_l\mathbf{\Sigma}_L\mathbf{U}_L^\top)^+\mathbf{U}_L\mathbf{\Sigma}_L \\
&= \mathbf{\Sigma}_L\mathbf{U}_L^\top\mathbf{U}_L\mathbf{\Sigma}_L^{-1}(\mathbf{V}_l^\top\mathbf{V}_l)^{-1}\mathbf{\Sigma}_L^{-1}\mathbf{U}_L^\top\mathbf{U}_L\mathbf{\Sigma}_L \\
&= (\mathbf{V}_l^\top\mathbf{V}_l)^{-1},
\end{aligned}$$

where the penultimate equality follows from $\mathbf{U}_L$ having orthogonal columns and $\mathbf{\Sigma}_L\mathbf{V}_l^\top\mathbf{V}_l\mathbf{\Sigma}_L$ having full rank. The proof of Lemma 11 established that the smallest singular value of $\frac{n}{l}\mathbf{V}_l^\top\mathbf{V}_l = \mathbf{V}_l^\top\mathbf{D}\mathbf{D}\mathbf{V}_l$ is lower bounded by $1-\epsilon/2$ and hence that $\|\mathbf{B}^\top\mathbf{B}\|_2 \le \frac{n}{l(1-\epsilon/2)}$ and $\|\mathbf{B}\|_2 \le \sqrt{\frac{n}{l(1-\epsilon/2)}}$. Thus, we conclude that $\mu_1(\mathbf{L}_C) \le \sqrt{r\mu_0(\mathbf{U}_L)\mu_0(\mathbf{V}_L)}/\sqrt{1-\epsilon/2}$.

C Proof of Theorem 7

We now give a proof of Thm. 7. 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 [10, Sec. 6] implies the same conclusions when columns and rows are sampled without replacement.

Our proof of Thm. 7 will require a strengthened version of the randomized ℓ_2 regression work of [6, Thm. 5]. The proof of Thm. 5 of [6] relies heavily on the fact that $\|\mathbf{AB}-\mathbf{GH}\|_F \le \frac{\epsilon}{2}\|\mathbf{A}\|_F\|\mathbf{B}\|_F$ 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 [11], modified to allow for slack in the probabilities, shows that a related claim holds with probability $1-\delta$ for arbitrary $\delta \in (0,1]$.

Lemma 10 (Sec. 3.4.3 of [11])**.** *Given matrices $\mathbf{A} \in \mathbb{R}^{m\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$ with $r \ge \max(\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B}))$, an error tolerance $\epsilon \in (0,1]$, and a failure probability $\delta \in (0,1]$, define probabilities p_j satisfying*

$$p_j \ge \frac{\beta}{Z}\|\mathbf{A}_{(j)}\|\|\mathbf{B}_{(j)}\|, \quad Z = \sum_j \|\mathbf{A}_{(j)}\|\|\mathbf{B}_{(j)}\|, \quad \textit{and} \quad \textstyle\sum_{j=1}^k p_j = 1 \tag{1}$$

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 $l \ge 48r\log(4r/(\beta\delta))/(\beta\epsilon^2)$ columns are selected in i.i.d. trials in which the j-th column is chosen with probability p_j, and let $\mathbf{H} \in \mathbb{R}^{l\times n}$ be a matrix containing the corresponding rows of $\mathbf{B}$. 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 $1-\delta$,

$$\|\mathbf{AB}-\mathbf{GDDH}\|_2 \le \frac{\epsilon}{2}\|\mathbf{A}\|_2\|\mathbf{B}\|_2.$$

Using Lemma 10, we now establish a stronger version of Lemma 1 of [6]. 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

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

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 $t \in \{1,\dots,l\}$. 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 11. *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 $1-\delta$:*

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

Proof By Lemma 10, for all $1 \leq i \leq r$,

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

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 [6]. □

Lemma 11 immediately yields improved sampling complexity for the randomized ℓ_2 regression of [6]:

Proposition 12. *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 $1-\delta-0.2$:*

$$\|\mathbf{B} - \mathbf{BSD}(\mathbf{LSD})^+\mathbf{L}\|_F \leq (1+\epsilon)\|\mathbf{B} - \mathbf{BL}^+\mathbf{L}\|_F.$$

Proof The proof is identical to that of Thm. 5 of [6] once Lemma 11 is substituted for Lemma 1 of [6]. □

A typical application of Prop. 12 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. (2). Here, we will take advantage of the slack term, β, allowed in the sampling probabilities of Eq. (2) 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 Thm. 7, we first notice that $n \geq r\mu_0(\mathbf{V}_L)$ and hence

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

whenever $\beta \geq 1/\mu_0(\mathbf{V}_L)$. Thus, we may apply Prop. 12 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

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

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

D Proof of Corollary 8

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

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

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

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,[6] and let $[\mathbf{L}_{C_1}, \cdots, \mathbf{L}_{C_b}]$ be the corresponding partition of $\mathbf{L}_C$. Since

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

we may apply Prop. 12 independently for each i to yield

$$\|\mathbf{M} - \mathbf{C}_i\mathbf{L}_{C_i}^+\mathbf{L}\|_F \leq (1+\epsilon)\|\mathbf{M} - \mathbf{M}\mathbf{L}^+\mathbf{L}\|_F \leq (1+\epsilon)\|\mathbf{M} - \mathbf{L}\|_F \tag{3}$$

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

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

for each i. Hence, if

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

fails to hold, then, for each i, Eq. (3) also fails to hold. The desired conclusion therefore must hold with probability at least $1 - 0.45^b = 1 - \delta$.

E Proof of Corollary 9

With $c = 48000/\log(1/0.45)$ as in Cor. 8, 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 Thm. 7 and Cor. 8 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.

F Proof of Theorem 3

Let $\mathbf{L}_0 = [\mathbf{C}_{0,1}, \ldots, \mathbf{C}_{0,t}]$ and $\hat{\mathbf{L}} = [\hat{\mathbf{C}}_1, \ldots, \hat{\mathbf{C}}_t]$. Define G as the event $\|\mathbf{L}_0 - \hat{\mathbf{L}}^{proj}\|_F \leq (2+\epsilon)c_e\sqrt{mn}\Delta$, H as the event $\|\hat{\mathbf{L}} - \hat{\mathbf{L}}^{proj}\|_F \leq (1+\epsilon)\|\mathbf{L}_0 - \hat{\mathbf{L}}\|_F$, and B_i as the event $\|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F \leq c_e\sqrt{ml}\Delta$, for each $i \in \{1, \ldots, t\}$. When H holds, we have that

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

by the triangle inequality, and hence

$$\mathbf{P}(G) \geq \mathbf{P}(\textstyle\bigcap_i B_i \cap H \cap \bigcap_i A(\mathbf{C}_{0,i})) = \mathbf{P}(\bigcap_i B_i \mid H \cap \bigcap_i A(\mathbf{C}_{0,i}))\mathbf{P}(H \cap \bigcap_i A(\mathbf{C}_{0,i})).$$

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 6, while H holds with probability at least $1 - \delta/2$ by Thm. 7. Hence, by the union bound,

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

[6] For simplicity, we assume that b divides l evenly.

Further, by a union bound and our base MF assumption,

$$\mathbf{P}(\bigcap_i B_i \mid H \cap \bigcap_i A(\mathbf{C}_{0,i})) \geq 1 - \sum_i \mathbf{P}(B_i^c \mid A(\mathbf{C}_{0,i})) \geq 1 - t\delta_C$$

yielding the desired bound on $\mathbf{P}(G)$.

To prove the second statement, we redefine $\hat{\mathbf{L}}$ and write it in block notation as:

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

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

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

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

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

imply that $A(\mathbf{C})$ and $A(\mathbf{R})$ hold with probability at least $1-\delta/(2n)$ and $1-\delta/(4n)$ respectively by Lemma 6, while K holds with probability at least $(1-\delta/2)(1-\delta/4-0.2)$ by Cor. 9. 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 $n > 1$ and $\delta \leq 0.8$. Further, by a union bound and our base MF assumption,

$$\begin{aligned} \mathbf{P}(J) &\geq \mathbf{P}(B_C \cap B_R \mid K \cap A(\mathbf{C}) \cap A(\mathbf{R}))\mathbf{P}(K \cap A(\mathbf{C}) \cap A(\mathbf{R})) \\ &\geq (1-\delta_C-\delta_R)(1-\delta)(1-\delta-0.2). \end{aligned}$$

G Proof of Corollary 4

Cor. 4 is based on a new noisy MC theorem, which we prove in Sec. I. A similar recovery guarantee is obtained by [3] under stronger assumptions.

Theorem 13. *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 $\hat{\mathbf{L}}$ to the problem

$$\text{minimize}_{\mathbf{L}} \quad \|\mathbf{L}\|_* \quad \text{subject to} \quad \|\mathcal{P}_\Omega(\mathbf{M}-\mathbf{L})\|_F \leq \Delta \tag{4}$$

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.

We begin by proving the DFC-PROJ bound. For each $i \in \{1,\ldots,t\}$, let B_i be the event that $\|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F > c_e'\sqrt{ml}\Delta$ and 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 $\mathbf{C}_{0,i}$,

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

Then, by Thm. 3, it suffices to establish that

$$\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) \leq (4\log(\bar{n})+1)\bar{n}^{2-2\beta}$$

for each i. By Thm. 13 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 $\mathrm{E}s_i = \frac{sl}{n}$ and hence satisfies Hoeffding's inequality for the hypergeometric distribution [10, Sec. 6]:

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

It therefore follows that

$$\begin{aligned}\mathbf{P}(D_i) &= \mathbf{P}\left(s_i < \mathrm{E}s_i - s\left(\frac{l}{n} - \frac{32\mu' r(m+l)\beta'\log^2(m+l)}{s}\right)\right) \\ &= \mathbf{P}\left(s_i < \mathrm{E}s_i - s\left(\frac{l}{n} - \frac{\beta(m+l)\log^2(m+l)}{\beta_s(m+n)\log^2(m+n)}\frac{\log(\bar{n})}{\log(\max(m,l))}\right)\right) \\ &\leq \mathbf{P}\left(s_i < \mathrm{E}s_i - s\left(\frac{l}{n} - \frac{\beta}{\beta_s}\right)\right) \\ &\leq \mathbf{P}\left(s_i < \mathrm{E}s_i - s\sqrt{\frac{\beta-1}{n\beta_s}}\right) \\ &\leq \exp\left(-2s\frac{\beta-1}{n\beta_s}\right) \leq \exp(-2\log(\bar{n})(\beta-1)) = \bar{n}^{2-2\beta}\end{aligned}$$

by our assumptions on s and l. Hence, $\mathbf{P}(B_i \mid A(\mathbf{C}_{0,i})) \leq (4\log(\bar{n})+1)\bar{n}^{2-2\beta}$ for each i, and the DFC-PROJ result follows from Thm. 3.

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 $\|\mathbf{R}_0 - \hat{\mathbf{R}}\|_F > c_e'\sqrt{dn}\Delta$. Reasoning identical to that above yields $\mathbf{P}(B_C \mid A(\mathbf{C})) \leq (4\log(\bar{n}) + 1)\bar{n}^{2-2\beta}$ and $\mathbf{P}(B_R \mid A(\mathbf{R})) \leq (4\log(\bar{n})+1)\bar{n}^{2-2\beta}$. Thus, the DFC-NYS bound also follows from Thm. 3.

H Proof of Corollary 5

Cor. 5 is based on the following theorem of Zhou et al. [25], reformulated for a generic rate parameter β, as described in [2, Section 3.1].

Theorem 14 (Thm. 2 of [25])**.** *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 $(\hat{\mathbf{L}}, \hat{\mathbf{S}})$ to the problem*

$$\textit{minimize}_{\mathbf{L},\mathbf{S}} \quad \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \quad \textit{subject to} \quad \|\mathbf{M} - \mathbf{L} - \mathbf{S}\|_F \leq \Delta \tag{5}$$

with $\lambda = 1/\sqrt{n}$ 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

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

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

We begin by proving the DFC-PROJ bound. For each $i \in \{1, \ldots, t\}$, let B_i be the event that $\|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\|_F > c_e''\sqrt{ml}\Delta$, and further define $\bar{m} \triangleq \max(m,l)$ and

$$\beta'' \triangleq \beta\log(\bar{n})/\log(\bar{m}) \leq \beta'.$$

Then, by Thm. 3, it suffices to establish that

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

for each i. By Thm. 14 and the definitions of β' and β'',

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

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 $\mathrm{E}s_i = \frac{sl}{n}$ and hence satisfies Bernstein's inequality for the hypergeometric [10, Sec. 6]:

$$\mathbf{P}(s_i \ge \mathrm{E}s_i + st) \le \exp\bigl(-st^2/(2\sigma^2 + 2t/3)\bigr) \le \exp\bigl(-st^2 n/4l\bigr),$$

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

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

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

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 $\|\mathbf{R}_0 - \hat{\mathbf{R}}\|_F > c_e''\sqrt{dn}\Delta$. Reasoning identical to that above yields $\mathbf{P}(B_C \mid A(\mathbf{C})) \le (c_p+1)\bar{n}^{-\beta}$ and $\mathbf{P}(B_R \mid A(\mathbf{R})) \le (c_p+1)\bar{n}^{-\beta}$. Thus, the DFC-Nys bound also follows from Thm. 3.

I Proof of Theorem 13

In the spirit of [3], our proof will extend the noiseless analysis of [22] to the noisy matrix completion setting. As suggested in [9], 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 [22].

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 $\|\mathcal{A}\|_2 = \sup_{\|\mathbf{X}\|_F \le 1}\|\mathcal{A}(\mathbf{X})\|_F$.

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

Theorem 15. *Under the assumptions of Thm. 13, suppose that*

$$\frac{mn}{s}\left\|\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T - \frac{s}{mn}\mathcal{P}_T\right\|_2 \le \frac{1}{2} \tag{6}$$

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

$$\|\mathcal{P}_T(\mathbf{Y}) - \mathbf{U}_{L_0}\mathbf{V}_{L_0}^\top\|_F \le \sqrt{\frac{s}{32mn}} \quad \textit{and} \quad \|\mathcal{P}_{T^\perp}(\mathbf{Y})\|_2 < \frac{1}{2}. \tag{7}$$

Then,

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

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. (6),

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

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

$$\sqrt{\frac{s}{2mn}} \|\mathcal{P}_T(\mathbf{H})\|_F \leq \|\mathcal{P}_\Omega \mathcal{P}_T(\mathbf{H})\|_F \leq \|\mathcal{P}_\Omega \mathcal{P}_{T^\perp}(\mathbf{H})\|_F \leq \|\mathcal{P}_{T^\perp}(\mathbf{H})\|_F \leq \|\mathcal{P}_{T^\perp}(\mathbf{H})\|_*, \quad (8)$$

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

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

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. (7), and the final inequality follows from Eq. (8).

Since $\mathbf{L}_0$ is feasible for Eq. (4), $\|\mathbf{L}_0\|_* \geq \|\hat{\mathbf{L}}\|_*$, 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 Thm. 15 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 16. *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 17. *Let* $\mathbf{Z}$ *be a fixed matrix in* $\mathbb{R}^{m\times n}$*. Then for all* $\beta > 1$,

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

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

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

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 [22, Thm. 3.4, Thm. 3.5, and Lemma 3.6] 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 [22, Thm. 3.2] also holds under sampling without replacement, as shown in [9].

Lemma 16 guarantees that Eq. (6) holds with high probability. To construct a matrix $\mathbf{Y} = \mathcal{P}_\Omega(\mathbf{Y})$ satisfying Eq. (7), we consider a sampling with batch replacement scheme recommended in [9] and developed in [5]. 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

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

It suffices to establish Eq. (7) under this batch replacement scheme, as shown in the next lemma.

Lemma 19. *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. (7). 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* $\mathbf{P}(A(\tilde{\Omega}(s))) \le \mathbf{P}(A(\Omega(s)))$.

Proof As sketched in [9]

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

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 [22] to obtain $\mathbf{Y} = \mathcal{P}_{\tilde{\Omega}}(\mathbf{Y})$ satisfying Eq. (7). 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

$$\frac{mn}{q}\left\|\mathcal{P}_T\mathcal{P}_{\tilde{\Omega}_k}\mathcal{P}_T - \frac{q}{mn}\mathcal{P}_T\right\|_2 \le \frac{1}{2} \tag{9}$$

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. (7).

The second condition of Eq. (7) follows from the assumptions

$$\left\|\mathbf{W}_{k-1} - \frac{mn}{q}\mathcal{P}_T\mathcal{P}_{\tilde{\Omega}_k}(\mathbf{W}_{k-1})\right\|_\infty \leq \frac{1}{2}\|\mathbf{W}_{k-1}\|_\infty \tag{10}$$

$$\left\|\left(\frac{mn}{q}\mathcal{P}_{\tilde{\Omega}_k} - \mathcal{I}\right)(\mathbf{W}_{k-1})\right\|_2 \leq \sqrt{\frac{8mn^2\beta\log(m+n)}{3q}}\|\mathbf{W}_{k-1}\|_\infty \tag{11}$$

for all k, since Eq. (10) 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 16 implies that Eq. (6) fails to hold with probability at most $2n^{2-2\beta}$. For each k, Eq. (9) fails to hold with probability at most $2n^{2-2\beta}$ by Lemma 16, Eq. (10) fails to hold with probability at most $2n^{2-2\beta}$ by Lemma 18, and Eq. (11) fails to hold with probability at most $(m+n)^{1-2\beta}$ by Lemma 17. Hence, by the union bound, the conclusion of Thm. 15 holds with probability at least

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