Matrix Factorization and Matrix Concentration

Lester Mackey

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-99
http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-99.html

May 11, 2012

Matrix Factorization and Matrix Concentration

by

Lester Wayne Mackey II

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Electrical Engineering and Computer Sciences

and the Designated Emphasis

in

Communication, Computation, and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Chair
Professor Peter Bickel
Professor Bin Yu

Spring 2012

Matrix Factorization and Matrix Concentration

Abstract

Matrix Factorization and Matrix Concentration

by

Lester Wayne Mackey II

Doctor of Philosophy in Electrical Engineering and Computer Sciences

with the Designated Emphasis in

Communication, Computation, and Statistics

University of California, Berkeley

Professor Michael I. Jordan, Chair

Motivated by the constrained factorization problems of sparse principal components analysis (PCA) for gene expression modeling, low-rank matrix completion for recommender systems, and robust matrix factorization for video surveillance, this dissertation explores the modeling, methodology, and theory of matrix factorization.

We begin by exposing the theoretical and empirical shortcomings of standard deflation techniques for sparse PCA and developing alternative methodology more suitable for deflation with sparse "pseudo-eigenvectors." We then explicitly reformulate the sparse PCA optimization problem and derive a generalized deflation procedure that typically outperforms more standard techniques on real-world datasets.

We next develop a fully Bayesian matrix completion framework for integrating the complementary approaches of discrete mixed membership modeling and continuous matrix factorization. We introduce two Mixed Membership Matrix Factorization (M3F) models, develop highly parallelizable Gibbs sampling inference procedures, and find that M3F is both more parsimonious and more accurate than state-of-the-art baselines on real-world collaborative filtering datasets.

Our third contribution is Divide-Factor-Combine (DFC), a parallel divide-and-conquer framework for boosting the scalability of a matrix completion or robust matrix factorization algorithm while retaining its theoretical guarantees. Our experiments demonstrate the near-linear to super-linear speed-ups attainable with this approach, and our analysis shows that DFC enjoys high-probability recovery guarantees comparable to those of its base algorithm.

Finally, inspired by the analyses of matrix completion and randomized factorization procedures, we show how Stein's method of exchangeable pairs can be used to derive concentration inequalities for matrix-valued random elements. As an immediate consequence, we obtain analogues of classical moment inequalities and exponential tail inequalities for

independent and dependent sums of random matrices. We moreover derive comparable concentration inequalities for self-bounding matrix functions of dependent random elements.

To my grandparents:
Gertrude Mackey, Walter Mackey, James Bell, and Margaret Bell.

Contents

List of Figures

List of Tables

Acknowledgments

I can only begin to thank my advisor, Michael Jordan, whose warm encouragement and well-reasoned enthusiasm first convinced me to enroll at the University of California, Berkeley. Over the years, Mike has taught me to be a statistician, to think independently and freely, to balance theoretical rigor with practical relevance, and, perhaps most importantly, to never stop learning. This thesis is a tribute to Mike's guidance and support.

As much credit belongs to an exceptional set of colleagues and friends: Chap. 3 of this thesis can be traced back to a late night hotel room chat with David Weiss, Chap. 4 is the product of a second hotel colloquy with Ameet Talwalkar, and Chap. 5 has spawned a fortuitous collaboration with Joel Tropp, Richard Chen, and Brendan Farrell [47]. Equally valuable and equally treasured were my collaborations with Ariel Kleiner, Anne Shiu, John Duchi, Tamara Broderick, John Paisley, and The Ensemble (`http://the-ensemble.com/`) on various projects not reflected in these pages.

I was privileged to be surrounded and inspired each day by the gifted minds of the Statistical Artificial Intelligence Laboratory, by a small army of roommates (Rob Carroll, Jean Han, Leo Meyerovich, Andy Konwinski, Kuang Chen, Kurtis Heimerl, Alice Lin, Jesse Trutna, Tyson Condie, Fabian Wauthier, Kurt Miller, Garvesh Raskutti, Percy Liang, Dave Golland, and Andre Wibisono), and by good friends scattered about the Bay Area. I give special thanks to Ben Hindman for teaching me how to weather a snow storm, to the Turing Machines/Floppy Disks for giving me a reason to run, to Ankur Mehta for always convincing me to "do stuff," and to Veritas for teaching me about the truth.

I was blessed with many sources of support from outside of Berkeley. I thank Carl Seger, Maria Klawe, and David Walker for introducing me to the world of computer science research; the AT&T Labs Fellowship Program and the National Defense Science and Engineering Fellowship Program for funding my years of study; and Bob Bell and Yehuda Koren, my AT&T Labs mentors, for generously imparting their wisdom and advice.

Finally, I thank the Lord, my parents, my sisters, Angela and Dawn, and my extraordinary girlfriend, Lilly, for sustaining me, encouraging me, and walking with me throughout this five year journey. This thesis is a testament to their love.

Chapter 1

Introduction

The goal in matrix factorization is to approximate a target matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ by a product of two lower dimensional *factor matrices*, $\mathbf{A} \in \mathbb{R}^{m \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times n}$, where the common dimension r is typically far smaller than m or n. Here, and throughout, we measure the quality of approximation through the Frobenius norm $\|\cdot\|_F$ over matrix differences. When $\mathbf{M}$ is fully observed and $\mathbf{A}$ and $\mathbf{B}$ are unconstrained, this problem has a well-known optimal solution, given by the truncated singular value decomposition of $\mathbf{M}$. More precisely, to minimize the reconstruction error $\|\mathbf{M} - \mathbf{AB}\|_F$ over all factor matrices with common dimension r, it suffices to choose $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r$ and $\mathbf{B} = \mathbf{V}_r^\top$, where $\mathbf{\Sigma}_r \in \mathbb{R}^{r \times r}$ is a diagonal matrix of the r largest singular values of $\mathbf{M}$, and $\mathbf{U}_r \in \mathbb{R}^{m \times r}$ and $\mathbf{V}_r \in \mathbb{R}^{n \times r}$ are the corresponding left and right singular vectors of $\mathbf{M}$.

Unfortunately, the demands of many real-world factorization problems are incompatible with this complete-information, unconstrained-optimization setting, and additional constraints must be imposed that render the matrix factorization problem far more challenging. Consider the following three classes of modern matrix factorization problems:

1. In the setting of *sparse principal components analysis* [33, 9, 82, 83, 34, 85, 17, 16, 55, 54, 73], $\mathbf{M}$ is a centered data matrix of m observations over n variables, and each row of $\mathbf{B}$ is constrained to have relatively few non-zero entries. Such cardinality constraints arise naturally in biology and finance, where sparse factor vectors depending on fewer variables offer the promise of greater interpretability and more practical relevance. These same constraints, however, render the matrix factorization problem NP-hard [54].

2. In the setting of *matrix completion* or *dyadic data prediction* [30], one observes only a small subset of the entries of $\mathbf{M}$ and aims to estimate the missing entries. Such missing data problems arise naturally in the domains of collaborative filtering for recommender systems, link prediction for social networks, and click prediction for web search. While matrix factorization techniques offer state of the art performance for matrix completion tasks [see, e.g., 38], they lack closed-form solutions, and their objectives may be plagued by local minima.

3. In the *robust matrix factorization* problem [12], also known as robust PCA [10], we observe a corrupted version of $\mathbf{M}$ where some entries have been replaced by outliers, and the locations of those entries are unknown. This problem, which finds diverse motivations in video surveillance [10], graphical model selection [12], document modeling [53], and image alignment [63], is strictly harder than the matrix completion problem, in which the locations of unobserved entries are known in advance.

Our understanding of matrix factorization in each of these constrained settings has grown rapidly in recent years, but, in each case, significant room remains for the development of

1. More accurate and parsimonious models of matricial data
2. Computationally efficient algorithms for large-scale or real-time factorization problems
3. Theoretical justification for existing methodology.

This dissertation presents contributions to each of these core areas. Chapters 2 and 3 present modeling improvements in the settings of sparse PCA and dyadic data prediction, respectively. In analogy to the PCA setting, the sparse PCA problem is often solved by iteratively alternating between two subtasks: cardinality-constrained rank-one variance maximization and matrix deflation. While the former has received a great deal of attention in the literature, the latter is seldom analyzed and is typically borrowed without justification from the PCA context. In Chapter 2, we demonstrate that the standard PCA deflation procedure is seldom appropriate for the sparse PCA setting. To rectify the situation, we first develop several deflation alternatives better suited to the cardinality-constrained context. We then reformulate the sparse PCA optimization problem to explicitly reflect the maximum *additional* variance objective on each round. The result is a generalized deflation procedure that typically outperforms more standard techniques on real-world datasets.

Discrete mixed membership modeling is a popular, complementary alternative to continuous latent factor modeling (i.e., matrix factorization) for analyzing the interactions between two populations. While latent factor models typically demonstrate greater predictive accuracy, mixed membership models better capture the heterogeneous nature of objects and their interactions. In Chapter 3, we develop a fully Bayesian framework for integrating the two approaches into unified Mixed Membership Matrix Factorization (M3F) models. We introduce two M3F models, derive highly parallelizable Gibbs sampling inference procedures, and validate our methods on the EachMovie, MovieLens, and Netflix Prize collaborative filtering datasets. We find that, even when fitting fewer parameters, the M3F models outperform state-of-the-art latent factor approaches in all experiments, yielding the greatest gains in accuracy on sparsely-rated, high-variance items.

Chapter 4 is devoted to the design of scalable but provably accurate methods for matrix completion and robust matrix factorization. Many modern matrix factorization methods boast strong theoretical guarantees but scale poorly due to expensive subroutines. To

address this shortcoming, we introduced Divide-Factor-Combine (DFC), a parallel divide-and-conquer framework that 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.

Fundamental to our analysis in Chapter 4 – and to the analyses of many matrix completion procedures – are matrix concentration inequalities that characterize the fluctuations of a random matrix about its mean. In Chapter 5, we will show how Steins method of exchangeable pairs can be used to derive concentration inequalities for matrix-valued random elements. When applied to a sum of independent random matrices, this approach yields matrix generalizations of the classical inequalities due to Hoeffding, Bernstein, and Khintchine. The same technique delivers bounds for sums of dependent random matrices and more general matrix functionals of dependent random elements.

Chapter 2

Deflation Methods for Sparse PCA

2.1 Introduction

Principal component analysis (PCA) is a popular change of variables technique used in data compression, predictive modeling, and visualization. The goal of PCA is to extract several principal components, linear combinations of input variables that together best account for the variance in a data set. Often, PCA is formulated as an eigenvalue decomposition problem: each eigenvector of the sample covariance matrix of a data set corresponds to the *loadings* or coefficients of a principal component. A common approach to solving this partial eigenvalue decomposition is to iteratively alternate between two subproblems: rank-one variance maximization and matrix deflation. The first subproblem involves finding the maximum-variance loadings vector for a given sample covariance matrix or, equivalently, finding the leading eigenvector of the matrix. The second involves modifying the covariance matrix to eliminate the influence of that eigenvector.

A primary drawback of PCA is its lack of sparsity. Each principal component is a linear combination of all variables, and the loadings are typically non-zero. Sparsity is desirable as it often leads to more interpretable results, reduced computation time, and improved generalization. Sparse PCA [33, 9, 82, 83, 34, 85, 17, 16, 55, 54, 73] injects sparsity into the PCA process by searching for "pseudo-eigenvectors", sparse loadings that explain a maximal amount variance in the data.

In analogy to the PCA setting, many authors attempt to solve the sparse PCA problem by iteratively alternating between two subtasks: cardinality-constrained rank-one variance maximization and matrix deflation. The former is an NP-hard problem, and a variety of relaxations and approximate solutions have been developed in the literature [17, 16, 55, 54, 73, 82, 83]. The latter subtask has received relatively little attention and is typically borrowed without justification from the PCA context. In this chapter, we demonstrate that the standard PCA deflation procedure is seldom appropriate for the sparse PCA setting. To rectify the situation, we first develop several heuristic deflation alternatives with more desirable properties [48]. We then reformulate the sparse PCA optimization problem to

explicitly reflect the maximum *additional* variance objective on each round. The result is a generalized deflation procedure that typically outperforms more standard techniques on real-world datasets.

The remainder of the chapter is organized as follows. In Section 2.2 we discuss matrix deflation as it relates to PCA and sparse PCA. We examine the failings of typical PCA deflation in the sparse setting and develop several alternative deflation procedures. In Section 2.3, we present a reformulation of the standard iterative sparse PCA optimization problem and derive a generalized deflation procedure to solve the reformulation. Finally, in Section 2.4, we demonstrate the utility of our newly derived deflation techniques on real-world datasets.

Notation

I is the identity matrix. $\mathbb{S}_+^p$ is the set of all symmetric, positive semidefinite matrices in $\mathbb{R}^{p\times p}$. $\mathbf{Card}(x)$ represents the cardinality of or number of non-zero entries in the vector x.

2.2 Deflation methods

A *matrix deflation* modifies a matrix to eliminate the influence of a given eigenvector, typically by setting the associated eigenvalue to zero (see [80] for a more detailed discussion). We will first discuss deflation in the context of PCA and then consider its extension to sparse PCA.

Hotelling's deflation and PCA

In the PCA setting, the goal is to extract the r leading eigenvectors of the sample covariance matrix, $A_0 \in \mathbb{S}_+^p$, as its eigenvectors are equivalent to the loadings of the first r principal components. Hotelling's deflation method [69] is a simple and popular technique for sequentially extracting these eigenvectors. On the t-th iteration of the deflation method, we first extract the leading eigenvector of A_{t-1},

$$x_t = \underset{x:x^Tx=1}{\operatorname{argmax}}\, x^T A_{t-1} x \tag{2.1}$$

and we then use Hotelling's deflation to annihilate x_t:

$$A_t = A_{t-1} - x_t x_t^T A_{t-1} x_t x_t^T. \tag{2.2}$$

The deflation step ensures that the $t+1$-st leading eigenvector of A_0 is the leading eigenvector of A_t. The following proposition explains why.

Proposition 1. *If $\lambda_1 \geq \ldots \geq \lambda_p$ are the eigenvalues of $A \in \mathbb{S}_+^p$, $x_1, \ldots, x_p$ are the corresponding eigenvectors, and $\hat{A} = A - x_j x_j^T A x_j x_j^T$ for some $j \in 1, \ldots, p$, then $\hat{A}$ has eigenvectors $x_1, \ldots, x_p$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_{j-1}, 0, \lambda_{j+1}, \ldots, \lambda_p$.*

PROOF.

$$\hat{A}x_j = Ax_j - x_j x_j^T A x_j x_j^T x_j = Ax_j - x_j x_j^T A x_j = \lambda_j x_j - \lambda_j x_j = 0 x_j.$$
$$\hat{A}x_i = Ax_i - x_j x_j^T A x_j x_j^T x_i = Ax_i - 0 = \lambda_i x_i, \forall i \neq j.$$

□

Thus, Hotelling's deflation preserves all eigenvectors of a matrix and annihilates a selected eigenvalue while maintaining all others. Notably, this implies that Hotelling's deflation preserves positive-semidefiniteness. In the case of our iterative deflation method, annihilating the t-th leading eigenvector of A_0 renders the $t+1$-st leading eigenvector dominant in the next round.

Hotelling's deflation and sparse PCA

In the sparse PCA setting, we seek r sparse loadings which together capture the maximum amount of variance in the data. Most authors [17, 55, 82, 73] adopt the additional constraint that the loadings be produced in a sequential fashion. To find the first such "pseudo-eigenvector", we can consider a cardinality-constrained version of Eq. (2.1):

$$x_1 = \underset{x: x^T x = 1, \mathbf{Card}(x) \leq k_1}{\operatorname{argmax}} x^T A_0 x. \tag{2.3}$$

That leaves us with the question of how to best extract subsequent pseudo-eigenvectors. A common approach in the literature [17, 55, 82, 73] is to borrow the iterative deflation method of the PCA setting. Typically, Hotelling's deflation is utilized by substituting an extracted pseudo-eigenvector for a true eigenvector in the deflation step of Eq. (2.2). This substitution, however, is seldom justified, for the properties of Hotelling's deflation, discussed in Section 2.2, depend crucially on the use of a true eigenvector.

To see what can go wrong when Hotelling's deflation is applied to a non-eigenvector, consider the following example.

Example. Let $C = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, a 2×2 matrix. The eigenvalues of C are $\lambda_1 = 2.6180$ and $\lambda_2 = .3820$. Let $x = (1, 0)^T$, a sparse pseudo-eigenvector, and $\hat{C} = C - xx^T Cxx^T$, the corresponding deflated matrix. Then $\hat{C} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ with eigenvalues $\hat{\lambda_1} = 1.6180$ and $\hat{\lambda_2} = -.6180$. Thus, Hotelling's deflation does not in general preserve positive-semidefiniteness when applied to a non-eigenvector.

That $\mathbb{S}_+^p$ is not closed under pseudo-eigenvector Hotelling's deflation is a serious failing, for most iterative sparse PCA methods assume a positive-semidefinite matrix on each iteration. A second, related shortcoming of pseudo-eigenvector Hotelling's deflation is its failure to render a pseudo-eigenvector orthogonal to a deflated matrix. If A is our matrix of interest, x is our pseudo-eigenvector with variance $\lambda = x^T Ax$, and $\hat{A} = A - xx^T Axx^T$ is our deflated

matrix, then $\hat{A}x = Ax - xx^TAxx^Tx = Ax - \lambda x$ is zero iff x is a true eigenvector. Thus, even though the "variance" of x w.r.t. $\hat{A}$ is zero ($x^T\hat{A}x = x^TAx - x^Txx^TAxx^Tx = \lambda - \lambda = 0$), "covariances" of the form $y^T\hat{A}x$ for $y \neq x$ may still be non-zero. This violation of the Cauchy-Schwarz inequality betrays a lack of positive-semidefiniteness and may encourage the reappearance of x as a component of future pseudo-eigenvectors.

Alternative deflation techniques

In this section, we will attempt to rectify the failings of pseudo-eigenvector Hotelling's deflation by considering several alternative deflation techniques better suited to the sparse PCA setting. Note that any deflation-based sparse PCA method (e.g. [17, 55, 82, 73]) can utilize any of the deflation techniques discussed below.

Projection deflation

Given a data matrix $Y \in \mathbb{R}^{n\times p}$ and an arbitrary unit vector in $x \in \mathbb{R}^p$, an intuitive way to remove the contribution of x from Y is to project Y onto the orthocomplement of the space spanned by x: $\hat{Y} = Y(I - xx^T)$. If A is the sample covariance matrix of Y, then the sample covariance of $\hat{Y}$ is given by $\hat{A} = (I - xx^T)A(I - xx^T)$, which leads to our formulation for projection deflation:

Projection deflation

$$A_t = A_{t-1} - x_tx_t^TA_{t-1} - A_{t-1}x_tx_t^T + x_tx_t^TA_{t-1}x_tx_t^T = (I - x_tx_t^T)A_{t-1}(I - x_tx_t^T) \qquad (2.4)$$

Note that when x_t is a true eigenvector of A_{t-1} with eigenvalue λ_t, projection deflation reduces to Hotelling's deflation:

$$\begin{aligned} A_t &= A_{t-1} - x_tx_t^TA_{t-1} - A_{t-1}x_tx_t^T + x_tx_t^TA_{t-1}x_tx_t^T \\ &= A_{t-1} - \lambda_tx_tx_t^T - \lambda_tx_tx_t^T + \lambda_tx_tx_t^T \\ &= A_{t-1} - x_tx_t^TA_{t-1}x_tx_t^T. \end{aligned}$$

However, in the general case, when x_t is not a true eigenvector, projection deflation maintains the desirable properties that were lost to Hotelling's deflation. For example, positive-semidefiniteness is preserved:

$$\forall y, y^TA_ty = y^T(I - x_tx_t^T)A_{t-1}(I - x_tx_t^T)y = z^TA_{t-1}z$$

where $z = (I - x_tx_t^T)y$. Thus, if $A_{t-1} \in \mathbb{S}_+^p$, so is A_t. Moreover, A_t is rendered left and right orthogonal to x_t, as $(I - x_tx_t^T)x_t = x_t - x_t = 0$ and A_t is symmetric. Projection deflation therefore annihilates all covariances with x_t: $\forall v, v^TA_tx_t = x_t^TA_tv = 0$.

Schur complement deflation

Since our goal in matrix deflation is to eliminate the influence, as measured through variance and covariances, of a newly discovered pseudo-eigenvector, it is reasonable to consider the conditional variance of our data variables given a pseudo-principal component. While this conditional variance is non-trivial to compute in general, it takes on a simple closed form when the variables are normally distributed. Let $x \in \mathbb{R}^p$ be a unit vector and $W \in \mathbb{R}^p$ be a Gaussian random vector, representing the joint distribution of the data variables. If W has covariance matrix Σ, then (W, Wx) has covariance matrix $V = \begin{pmatrix} \Sigma & \Sigma x \\ x^T\Sigma & x^T\Sigma x \end{pmatrix}$, and $Var(W|Wx) = \Sigma - \frac{\Sigma x x^T \Sigma}{x^T \Sigma x}$ whenever $x^T\Sigma x \neq 0$ [20]. That is, the conditional variance is the Schur complement of the vector variance $x^T\Sigma x$ in the full covariance matrix V. By substituting sample covariance matrices for their population counterparts, we arrive at a new deflation technique:

Schur complement deflation

$$A_t = A_{t-1} - \frac{A_{t-1}x_t x_t^T A_{t-1}}{x_t^T A_{t-1} x_t} \tag{2.5}$$

Schur complement deflation, like projection deflation, preserves positive-semidefiniteness. To see this, suppose $A_{t-1} \in \mathbb{S}_+^p$. Then, $\forall v, v^T A_t v = v^T A_{t-1} v - \frac{v^T A_{t-1} x_t x_t^T A_{t-1} v}{x_t^T A_{t-1} x_t} \geq 0$ as $v^T A_{t-1} v x_t^T A_{t-1} x_t - (v^T A_{t-1} x_t)^2 \geq 0$ by the Cauchy-Schwarz inequality and $x_t^T A_{t-1} x_t \geq 0$ as $A_{t-1} \in \mathbb{S}_+^p$.

Furthermore, Schur complement deflation renders x_t left and right orthogonal to A_t, since A_t is symmetric and $A_t x_t = A_{t-1}x_t - \frac{A_{t-1}x_t x_t^T A_{t-1} x_t}{x_t^T A_{t-1} x_t} = A_{t-1}x_t - A_{t-1}x_t = 0$.

Additionally, Schur complement deflation reduces to Hotelling's deflation when x_t is an eigenvector of A_{t-1} with eigenvalue $\lambda_t \neq 0$:

$$\begin{aligned} A_t &= A_{t-1} - \frac{A_{t-1}x_t x_t^T A_{t-1}}{x_t^T A_{t-1} x_t} \\ &= A_{t-1} - \frac{\lambda_t x_t x_t^T \lambda_t}{\lambda_t} \\ &= A_{t-1} - x_t x_t^T A_{t-1} x_t x_t^T. \end{aligned}$$

While we motivated Schur complement deflation with a Gaussianity assumption, the technique admits a more general interpretation as a column projection of a data matrix. Suppose $Y \in \mathbb{R}^{n \times p}$ is a mean-centered data matrix, $x \in \mathbb{R}^p$ has unit norm, and $\hat{Y} = (I - \frac{Yxx^TY^T}{\|Yx\|^2})Y$, the projection of the columns of Y onto the orthocomplement of the space spanned by the pseudo-principal component, Yx. If Y has sample covariance matrix A, then the sample covariance of $\hat{Y}$ is given by $\hat{A} = \frac{1}{n}Y^T(I - \frac{Yxx^TY^T}{\|Yx\|^2})^T(I - \frac{Yxx^TY^T}{\|Yx\|^2})Y = \frac{1}{n}Y^T(I - \frac{Yxx^TY^T}{\|Yx\|^2})Y = A - \frac{Axx^TA}{x^TAx}$.

Orthogonalized deflation

While projection deflation and Schur complement deflation address the concerns raised by performing a single deflation in the non-eigenvector setting, new difficulties arise when we attempt to sequentially deflate a matrix with respect to a *series* of non-orthogonal pseudo-eigenvectors.

Whenever we deal with a sequence of non-orthogonal vectors, we must take care to distinguish between the variance explained by a vector and the *additional* variance explained, given all previous vectors. These concepts are equivalent in the PCA setting, as true eigenvectors of a matrix are orthogonal, but, in general, the vectors extracted by sparse PCA will not be orthogonal. The additional variance explained by the t-th pseudo-eigenvector, x_t, is equivalent to the variance explained by the component of x_t orthogonal to the space spanned by all previous pseudo-eigenvectors, $q_t = x_t - \mathcal{P}_{t-1}x_t$, where $\mathcal{P}_{t-1}$ is the orthogonal projection onto the space spanned by $x_1, \ldots, x_{t-1}$. On each deflation step, therefore, we only want to eliminate the variance associated with q_t. Annihilating the full vector x_t will often lead to "double counting" and could re-introduce components parallel to previously annihilated vectors. Consider the following example:

Example. Let $C_0 = I$. If we apply projection deflation w.r.t. $x_1 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})^T$, the result is $C_1 = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$, and x_1 is orthogonal to C_1. If we next apply projection deflation to C_1 w.r.t. $x_2 = (1, 0)^T$, the result, $C_2 = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$, is no longer orthogonal to x_1.

The authors of [73] consider this issue of non-orthogonality in the context of Hotelling's deflation. Their modified deflation procedure is equivalent to Hotelling's deflation (Eq. (2.2)) for $t = 1$ and can be easily expressed in terms of a running Gram-Schmidt decomposition for $t > 1$:

Orthogonalized Hotelling's deflation (OHD)

$$\begin{aligned} q_t &= \frac{(I - Q_{t-1}Q_{t-1}^T)x_t}{\left\|(I - Q_{t-1}Q_{t-1}^T)x_t\right\|} \\ A_t &= A_{t-1} - q_t q_t^T A_{t-1} q_t q_t^T \end{aligned} \tag{2.6}$$

where $q_1 = x_1$, and $q_1, \ldots, q_{t-1}$ form the columns of Q_{t-1}. Since $q_1, \ldots, q_{t-1}$ form an orthonormal basis for the space spanned by $x_1, \ldots, x_{t-1}$, we have that $Q_{t-1}Q_{t-1}^T = \mathcal{P}_{t-1}$, the aforementioned orthogonal projection.

Since the first round of OHD is equivalent to a standard application of Hotelling's deflation, OHD inherits all of the weaknesses discussed in Section 2.2. However, the same principles may be applied to projection deflation to generate an orthogonalized variant that inherits its desirable properties.

Schur complement deflation is unique in that it preserves orthogonality in all subsequent rounds. That is, if a vector v is orthogonal to A_{t-1} for any t, then $A_t v = A_{t-1}v - \frac{A_{t-1}x_t x_t^T A_{t-1}v}{x_t^T A_{t-1} x_t} = 0$ as $A_{t-1}v = 0$. This further implies the following proposition.

Proposition 2. *Orthogonalized Schur complement deflation is equivalent to Schur complement deflation.*

Proof Consider the t-th round of Schur complement deflation. We may write $x_t = o_t + p_t$, where p_t is in the subspace spanned by all previously extracted pseudo-eigenvectors and o_t is orthogonal to this subspace. Then we know that $A_{t-1}p_t = 0$, as p_t is a linear combination of $x_1, \ldots, x_{t-1}$, and $A_{t-1}x_i = 0, \forall i < t$. Thus, $x_t^T A_t x_t = p_t^T A_t p_t + o_t^T A_t p_t + p_t^T A_t o_t + o_t^T A_t o_t = o_t^T A_t o_t$. Further, $A_{t-1}x_t x_t^T A_{t-1} = A_{t-1}p_t p_t^T A_{t-1} + A_{t-1}p_t o_t^T A_{t-1} + A_{t-1}o_t p_t^T A_{t-1} + A_{t-1}o_t o_t^T A_{t-1} = A_{t-1}o_t o_t^T A_{t-1}$. Hence, $A_t = A_{t-1} - \frac{A_{t-1}o_t o_t^T A_{t-1}}{o_t^T A_{t-1} o_t} = A_{t-1} - \frac{A_{t-1}q_t q_t^T A_{t-1}}{q_t^T A_{t-1} q_t}$ as $q_t = \frac{o_t}{\|o_t\|}$. □

Table 2.1 compares the properties of the various deflation techniques studied in this section.

Method	$x_t^T A_t x_t = 0$	$A_t x_t = 0$	$A_t \in \mathbb{S}_+^p$	$A_s x_t = 0, \forall s > t$
Hotelling's	✓	×	×	×
Projection	✓	✓	✓	×
Schur complement	✓	✓	✓	✓
Orth. Hotelling's	✓	×	×	×
Orth. Projection	✓	✓	✓	✓

Table 2.1: Summary of sparse PCA deflation method properties

2.3 Reformulating sparse PCA

In the previous section, we focused on heuristic deflation techniques that allowed us to reuse the cardinality-constrained optimization problem of Eq. (2.3). In this section, we explore a more principled alternative: reformulating the sparse PCA optimization problem to explicitly reflect our maximization objective on each round.

Recall that the goal of sparse PCA is to find r cardinality-constrained pseudo-eigenvectors which together explain the most variance in the data. If we additionally constrain the sparse loadings to be generated sequentially, as in the PCA setting and the previous section, then a greedy approach of maximizing the *additional* variance of each new vector naturally suggests itself.

On round t, the additional variance of a vector x is given by $\frac{q^T A_0 q}{q^T q}$ where A_0 is the data covariance matrix, $q = (I - \mathcal{P}_{t-1})x$, and $\mathcal{P}_{t-1}$ is the projection onto the space spanned by

previous pseudo-eigenvectors $x_1, \ldots, x_{t-1}$. As $q^T q = x^T(I-\mathcal{P}_{t-1})(I-\mathcal{P}_{t-1})x = x^T(I-\mathcal{P}_{t-1})x$, maximizing additional variance is equivalent to solving a cardinality-constrained maximum generalized eigenvalue problem,

$$\begin{aligned} \max_x \quad & x^T(I-\mathcal{P}_{t-1})A_0(I-\mathcal{P}_{t-1})x \\ \text{subject to } & x^T(I-\mathcal{P}_{t-1})x = 1 \\ & \mathbf{Card}(x) \leq k_t. \end{aligned} \tag{2.7}$$

If we let $q_s = (I-\mathcal{P}_{s-1})x_s, \forall s \leq t-1$, then $q_1, \ldots, q_{t-1}$ form an orthonormal basis for the space spanned by $x_1, \ldots, x_{t-1}$. Writing $I-\mathcal{P}_{t-1} = I - \sum_{s=1}^{t-1} q_s q_s^T = \prod_{s=1}^{t-1}(I - q_s q_s^T)$ suggests a generalized deflation technique that leads to the solution of Eq. (2.7) on each round. We imbed the technique into the following algorithm for sparse PCA:

Algorithm 1 Generalized Deflation Method for Sparse PCA

Given: $A_0 \in S_+^p$, $r \in \mathbb{N}$, $\{k_1, \ldots, k_r\} \subset \mathbb{N}$
Execute:

1. $B_0 \leftarrow I$
2. For $t := 1, \ldots, r$
   - $x_t \leftarrow \underset{x: x^T B_{t-1} x = 1, \mathbf{Card}(x) \leq k_t}{\operatorname{argmax}} x^T A_{t-1} x$
   - $q_t \leftarrow B_{t-1} x_t$
   - $A_t \leftarrow (I - q_t q_t^T) A_{t-1} (I - q_t q_t^T)$
   - $B_t \leftarrow B_{t-1}(I - q_t q_t^T)$
   - $x_t \leftarrow x_t / \|x_t\|$

Return: $\{x_1, \ldots, x_r\}$

Adding a cardinality constraint to a maximum eigenvalue problem renders the optimization problem NP-hard [54], but any of several leading sparse eigenvalue methods, including GSLDA of [54], DCPCA of [73], and DSPCA of [17] (with a modified trace constraint), can be adapted to solve this cardinality-constrained generalized eigenvalue problem.

2.4 Experiments

In this section, we present several experiments on real world datasets to demonstrate the value added by our newly derived deflation techniques. We run our experiments with Matlab implementations of DCPCA [73] (with the continuity correction of [55]) and GSLDA [54], fitted with each of the following deflation techniques: Hotelling's (HD), projection (PD), Schur

complement (SCD), orthogonalized Hotelling's (OHD), orthogonalized projection (OPD), and generalized (GD).

Pit props dataset

The pit props dataset [32] with 13 variables and 180 observations has become a de facto standard for benchmarking sparse PCA methods. To demonstrate the disparate behavior of differing deflation methods, we utilize each sparse PCA algorithm and deflation technique to successively extract six sparse loadings, each constrained to have cardinality less than or equal to $k_t = 4$. We report the additional variances explained by each sparse vector in Table 2.2 and the cumulative percentage variance explained on each iteration in Table 2.3. For reference, the first 6 true principal components of the pit props dataset capture 87% of the variance.

DCPCA						GSLDA					
HD	PD	SCD	OHD	OPD	GD	HD	PD	SCD	OHD	OPD	GD
2.938	2.938	2.938	2.938	2.938	2.938	2.938	2.938	2.938	2.938	2.938	2.938
2.209	2.209	2.076	2.209	2.209	2.209	2.107	2.280	2.065	2.107	2.280	2.280
0.935	1.464	1.926	0.935	1.464	1.477	1.988	2.067	2.243	1.985	2.067	2.072
1.301	1.464	1.164	0.799	1.464	1.464	1.352	1.304	1.120	1.335	1.305	1.360
1.206	1.057	1.477	0.901	1.058	1.178	1.067	1.120	1.164	0.497	1.125	1.127
0.959	0.980	0.725	0.431	0.904	0.988	0.557	0.853	0.841	0.489	0.852	0.908

Table 2.2: Additional variance explained by each of the first 6 sparse loadings extracted from the Pit Props dataset.

On the DCPCA run, Hotelling's deflation explains 73.4% of the variance, while the best performing methods, Schur complement deflation and generalized deflation, explain approximately 79% of the variance each. Projection deflation and its orthogonalized variant also outperform Hotelling's deflation, while orthogonalized Hotelling's shows the worst performance with only 63.2% of the variance explained. Similar results are obtained when the discrete method of GSLDA is used. Generalized deflation and the two projection deflations dominate, with GD achieving the maximum cumulative variance explained on each round. In contrast, the more standard Hotelling's and orthogonalized Hotelling's underperform the remaining techniques.

Gene expression data

The Berkeley Drosophila Transcription Network Project (BDTNP) 3D gene expression data [21] contains gene expression levels measured in each nucleus of developing Drosophila embryos and averaged across many embryos and developmental stages. Here, we analyze 0-3_1160524183713_s10436-29ap05-02.vpc, an aggregate VirtualEmbryo containing 21 genes

DCPCA						GSLDA					
HD	PD	SCD	OHD	OPD	GD	HD	PD	SCD	OHD	OPD	GD
22.6%	22.6%	22.6%	22.6%	22.6%	22.6%	22.6%	22.6%	22.6%	22.6%	22.6%	22.6%
39.6%	39.6%	38.6%	39.6%	39.6%	39.6%	38.8%	40.1%	38.5%	38.8%	40.1%	40.1%
46.8%	50.9%	53.4%	46.8%	50.9%	51.0%	54.1%	56.0%	55.7%	54.1%	56.0%	56.1%
56.8%	62.1%	62.3%	52.9%	62.1%	62.2%	64.5%	66.1%	64.4%	64.3%	66.1%	66.5%
66.1%	70.2%	73.7%	59.9%	70.2%	71.3%	72.7%	74.7%	73.3%	68.2%	74.7%	75.2%
73.4%	77.8%	79.3%	63.2%	77.2%	78.9%	77.0%	81.2%	79.8%	71.9%	81.3%	82.2%

Table 2.3: Cumulative percentage variance explained by the first 6 sparse loadings extracted from the Pit Props dataset.

and 5759 example nuclei. We run GSLDA for eight iterations with cardinality pattern 9,7,6,5,3,2,2,2 and report the results in Table 2.4.

	GSLDA additional variance explained						GSLDA cumulative percentage variance					
	HD	PD	SCD	OHD	OPD	GD	HD	PD	SCD	OHD	OPD	GD
PC 1	1.784	1.784	1.784	1.784	1.784	1.784	21.0%	21.0%	21.0%	21.0%	21.0%	21.0%
PC 2	1.464	1.453	1.453	1.464	1.453	1.466	38.2%	38.1%	38.1%	38.2%	38.1%	38.2%
PC 3	1.178	1.178	1.179	1.176	1.178	1.187	52.1%	51.9%	52.0%	52.0%	51.9%	52.2%
PC 4	0.716	0.736	0.716	0.713	0.721	0.743	60.5%	60.6%	60.4%	60.4%	60.4%	61.0%
PC 5	0.444	0.574	0.571	0.460	0.571	0.616	65.7%	67.4%	67.1%	65.9%	67.1%	68.2%
PC 6	0.303	0.306	0.278	0.354	0.244	0.332	69.3%	71.0%	70.4%	70.0%	70.0%	72.1%
PC 7	0.271	0.256	0.262	0.239	0.313	0.304	72.5%	74.0%	73.4%	72.8%	73.7%	75.7%
PC 8	0.223	0.239	0.299	0.257	0.245	0.329	75.1%	76.8%	77.0%	75.9%	76.6%	79.6%

Table 2.4: Additional variance and cumulative percentage variance explained by the first 8 sparse loadings of GSLDA on the BDTNP VirtualEmbryo.

The results of the gene expression experiment show a clear hierarchy among the deflation methods. The generalized deflation technique performs best, achieving the largest additional variance on every round and a final cumulative variance of 79.6%. Schur complement deflation, projection deflation, and orthogonalized projection deflation all perform comparably, explaining roughly 77% of the total variance after 8 rounds. In last place are the standard Hotelling's and orthogonalized Hotelling's deflations, both of which explain less than 76% of variance after 8 rounds.

2.5 Conclusion

In this chapter, we have exposed the theoretical and empirical shortcomings of Hotelling's deflation in the sparse PCA setting and developed several alternative methods more suitable for non-eigenvector deflation. Notably, the utility of these procedures is not limited to the sparse PCA setting. Indeed, the methods presented can be applied to any of a number of constrained eigendecomposition-based problems, including sparse canonical correlation analysis [78] and linear discriminant analysis [54].

Chapter 3

Mixed Membership Matrix Factorization

3.1 Introduction

This chapter is concerned with unifying discrete mixed membership modeling and continuous latent factor modeling for probabilistic dyadic data prediction. In the dyadic data prediction (DDP) problem [30], we observe labeled *dyads*, i.e., ordered pairs of objects, and form predictions for the labels of unseen dyads. For example, in the collaborative filtering setting, we observe U users, M items, and a training set $\mathcal{T} = \{(u_n, j_n, r_n)\}_{n=1}^{N}$ with real-valued ratings r_n representing the preferences of certain users u_n for certain items j_n. The goal is then to predict unobserved ratings based on users' past preferences. Other concrete examples of DDP include link prediction in social network analysis, binding affinity prediction in bioinformatics, and click prediction in web search.

Matrix factorization methods [68, 18, 71, 70, 75, 41] represent the state of the art for dyadic data prediction tasks. These methods view a dyadic dataset as a sparsely observed ratings matrix, $R \in \mathbb{R}^{U \times M}$, and learn a constrained decomposition of that matrix as a product of two latent factor matrices: $R \approx A^t B$ for $A \in \mathbb{R}^{D \times U}$, $B \in \mathbb{R}^{D \times M}$, and D small. While latent factor methods perform remarkably well on the DDP task, they fail to capture the heterogeneous nature of objects and their interactions. Such models, for instance, do not account for the fact that a user's ratings are influenced by instantaneous mood, that protein interactions are affected by transient functional contexts, or even that users with distinct behaviors may be sharing a single account or web browser.

The fundamental limitation of continuous latent factor methods is a result of the static way in which ratings are assumed to be produced: a user generates all of his item ratings using the same factor vector, without regard for context. Discrete mixed membership models, like Latent Dirichlet Allocation [6], were developed to address a similar limitation of mixture models. Whereas mixture models assume that each generated object is underlyingly a member of a single latent topic, mixed membership models represent objects as distributions

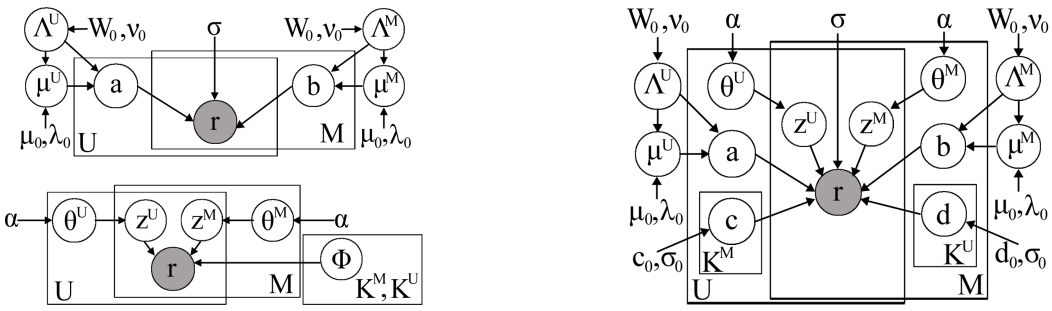

Figure 3.1: Graphical model representations of BPMF (top left), Bi-LDA (bottom left), and M^{3}F-TIB (right).

over topics. Mixed membership dyadic data models such as the Mixed Membership Stochastic Blockmodel [3] for relational prediction and Bi-LDA [66] for rating prediction introduce context dependence by allowing each object to select a new topic for each new interaction. However, the relatively poor predictive performance of Bi-LDA suggests that the blockmodel assumption—that objects only interact via their topics—is too restrictive.

In this chapter we develop a fully Bayesian framework for wedding the strong performance and expressiveness of continuous latent factor models with the context dependence and topic clustering of discrete mixed membership models [46]. In Section 3.2, we provide additional background on matrix factorization and mixed membership modeling. We introduce our Mixed Membership Matrix Factorization (M^{3}F) framework in Section 3.3, and discuss inference and prediction under two M^{3}F models in Section 3.4. Section 3.5 describes experimental evaluation and analysis of our models on a variety of real-world collaborative filtering datasets. The results demonstrate that Mixed-Membership Matrix Factorization methods outperform their context-blind counterparts and simultaneously reveal interesting clustering structure in the data. Finally, we conclude in Section 4.6.

3.2 Background

Latent Factor Models

We begin by considering a prototypical latent factor model, Bayesian Probabilistic Matrix Factorization of Salakhutdinov and Mnih [70] (see Figure 3.1). Like most factor models, BPMF associates with each user u an unknown factor vector $\mathbf{a}_u \in \mathbb{R}^D$ and with each item j an unknown factor vector $\mathbf{b}_j \in \mathbb{R}^D$. A user generates a rating for an item by adding Gaussian noise to the inner product, $r_{uj} = \mathbf{a}_u \cdot \mathbf{b}_j$. We refer to this inner product as the *static rating* for a user-item pair, for, as discussed in the introduction, the latent factor rating mechanism does not model the context in which a rating is given and does not allow a user to don different moods or "hats" in different dyadic interactions. Such contextual flexibility

is desirable for capturing the context-sensitive nature of dyadic interactions, and, as such, we turn our attention to mixed membership models.

Mixed Membership Models

Two recent examples of dyadic mixed membership (DMM) models are the Mixed Membership Stochastic Blockmodel (MMSB) [3] and Bi-LDA [66] (see Figure 3.1). In DMM models, each user u and item j has its own discrete distribution over topics, represented by topic parameters θ_u^U and θ_j^M. When a user desires to rate an item, both the user and the item select interaction-specific topics according to their distributions; the selected topics then determine the distribution over ratings.

One drawback of DMM models is the reliance on purely groupwise interactions: one learns how a user group interacts with an item group but not how a user group interacts directly with a particular item. M^{3}F models address this limitation in two ways—first, by modeling interactions between groups and specific users or items and second, by incorporating the user-item specific static rating of latent factor models.

3.3 Mixed Membership Matrix Factorization

In this section, we present a general Mixed Membership Matrix Factorization framework and two specific models that leverage the predictive power and static specificity of continuous latent factor models while allowing for the clustered context-sensitivity of mixed membership models. In each M^{3}F model, users and items are endowed both with latent factor vectors ($\mathbf{a}_u$ and $\mathbf{b}_j$) and with topic distribution parameters (θ_u^U and θ_j^M). To rate an item, a user first draws a topic z_{uj}^U from his distribution, representing, for example, his mood at the time of rating (in the mood for romance vs. comedy), and the item draws a topic z_{uj}^M from its distribution, representing, for example, the context under which it is being rated (in a theater on opening night vs. in a high-school classroom). The user and item topics, i and k, together with the identity of the user and item, u and j, jointly specify a rating bias, β_{uj}^{ik}, tailored to the user-item pair. Different M^{3}F models will differ principally in the precise form of this *contextual bias*. To generate a complete rating, the user-item-specific static rating $\mathbf{a}_u \cdot \mathbf{b}_j$ is added to the contextual bias β_{uj}^{ik}, along with some noise.

Rather than learn point estimates under our M^{3}F models, we adopt a fully Bayesian methodology and place priors on all parameters of interest. Topic distribution parameters θ_u^U and θ_j^M are given independent exchangeable Dirichlet priors, and the latent factor vectors $\mathbf{a}_u$ and $\mathbf{b}_j$ are drawn independently from $\mathcal{N}\big(\mu^U, (\Lambda^U)^{-1}\big)$ and $\mathcal{N}\big(\mu^M, (\Lambda^M)^{-1}\big)$, respectively. As in Salakhutdinov and Mnih [70], we place normal-Wishart priors on the hyper-parameters (μ^U, Λ^U) and (μ^M, Λ^M). Suppose K^U is the number of user topics and K^M is the number of item topics. Then, given the contextual biases β_{uj}^{ik}, ratings are generated according to the following M^{3}F generative process:

$\Lambda^U \sim \text{Wishart}(\mathbf{W}_0, \nu_0)$, $\Lambda^M \sim \text{Wishart}(\mathbf{W}_0, \nu_0)$

$\mu^U \sim \mathcal{N}\big(\mu_0, (\lambda_0 \Lambda^U)^{-1}\big)$, $\mu^M \sim \mathcal{N}\big(\mu_0, (\lambda_0 \Lambda^M)^{-1}\big)$

For each $u \in \{1, \ldots, U\}$:

$$\mathbf{a}_u \sim \mathcal{N}\big(\mu^U, (\Lambda^U)^{-1}\big)$$
$$\theta_u^U \sim \text{Dir}(\alpha / K^U)$$

For each $j \in \{1, \ldots, M\}$:

$$\mathbf{b}_j \sim \mathcal{N}\big(\mu^M, (\Lambda^M)^{-1}\big)$$
$$\theta_j^M \sim \text{Dir}(\alpha / K^M)$$

For each rating r_{uj}:

$$z_{uj}^U \sim \text{Multi}(1, \theta_u^U),\ z_{uj}^M \sim \text{Multi}(1, \theta_j^M)$$
$$r_{uj} \sim \mathcal{N}\big(\beta_{uj}^{ik} + \mathbf{a}_u \cdot \mathbf{b}_j, \sigma^2\big).$$

For each model discussed below, we let Θ^U denote the collection of all user parameters (e.g., $\mathbf{a}, \theta^U, \Lambda^U, \mu^U$), Θ^M denote all item parameters, and Θ_0 denote all global parameters (e.g., $\mathbf{W}_0, \nu_0, \mu_0, \lambda_0, \alpha, \sigma_0^2, \sigma^2$). We now describe in more detail the specific forms of two M^{3}F models and their contextual biases.

The M^{3}F Topic-Indexed Bias Model

The M^{3}F Topic-Indexed Bias (TIB) model assumes that the contextual bias decomposes into a latent user bias and a latent item bias. The user bias is influenced by the interaction-specific topic selected by the item. Similarly, the item bias is influenced by the user's selected topic. We denote the latent rating bias of user u under item topic k as c_u^k and denote the bias for item j under user topic i as d_j^i. The contextual bias for a given user-item interaction is then found by summing the two latent biases and a fixed global bias, χ_0[1]:

$$\beta_{uj}^{ik} = \chi_0 + c_u^k + d_j^i.$$

Topic-indexed biases c_u^k and d_j^i are drawn independently from Gaussian priors with variance σ_0^2 and means c_0 and d_0 respectively. Figure 3.1 compares the graphical model representations of M^{3}F-TIB, BPMF, and Bi-LDA. Note that M^{3}F-TIB reduces to BPMF when K^U and K^M are both zero.

Intuitively, the topic-indexed bias model captures the "*Napoleon Dynamite* effect," [76] whereby certain movies provoke strongly differing reactions from otherwise similar users. Each user-topic-indexed bias d_j^i represents one of K^U possible predispositions towards liking or disliking each item in the database, irrespective of the static latent factor parameterization. Thus, in the movie-recommendation problem, we expect the variance in user reactions to

[1]The global bias, χ_0, is suppressed in the remainder of the chapter for clarity.

Algorithm 1 Gibbs Sampling for M^{3}F-TIB.

Input: $(\mathbf{a}^{(0)}, \mathbf{b}^{(0)}, \mathbf{c}^{(0)}, \mathbf{d}^{(0)}, \theta^{U(0)}, \theta^{M(0)}, \mathbf{z}^{M(0)})$
  for $t = 1$ **to** T **do**
    // Sample Hyperparameters
    for $(u, j) \in \mathcal{T}$ **do**
      $(\mu^U, \Lambda^U)^t \sim \mu^U, \Lambda^U \mid \mathbf{a}^{t-1}, \Theta_0$
      $(\mu^M, \Lambda^M)^t \sim \mu^M, \Lambda^M \mid \mathbf{b}^{t-1}, \Theta_0$
    end for
    // Sample Topics
    for $(u, j) \in \mathcal{T}$ **do**
      $z_{uj}^{U(t)} \sim z_{uj}^U | (z_{uj}^M, \theta_u^U, \mathbf{a}_u, \mathbf{b}_j, \mathbf{c}_u, \mathbf{d}_j)^{t-1}, \mathbf{r}^{(\mathrm{v})}, \Theta_0$
      $z_{uj}^{M(t)} \sim z_{uj}^M | (\theta_j^M, \mathbf{a}_u, \mathbf{b}_j, \mathbf{c}_u, \mathbf{d}_j)^{t-1}, z_{uj}^{U(t)}, \mathbf{r}^{(\mathrm{v})}, \Theta_0$
    end for
    // Sample User Parameters
    for $u = 1$ **to** U **do**
      $\theta_u^{U(t)} \sim \theta_u^U \mid \mathbf{z}^{U(t)}, \Theta_0$
      $\mathbf{a}_u^t \sim \mathbf{a}_u \mid (\Lambda^U, \mu^U, \mathbf{z}_u^U, \mathbf{z}^M)^t, (\mathbf{b}, \mathbf{c}_u, \mathbf{d})^{t-1}, \Theta_0$
      for $i = 1$ **to** K^M **do**
        $c_u^{i(t)} \sim c_u^i \mid (\mathbf{z}^U, \mathbf{z}^M, \mathbf{a}_u)^t, (\mathbf{b}, \mathbf{d})^{t-1}, \mathbf{r}^{(\mathrm{v})}, \Theta_0$
      end for
    end for
    // Sample Item Parameters
    for $j = 1$ **to** M **do**
      $\theta_j^{M(t)} \sim \theta_j^M \mid \mathbf{z}^{M(t)}, \Theta_0$
      $\mathbf{b}_j^t \sim \mathbf{b}_j \mid (\Lambda^U, \mu^U, \mathbf{z}_u^U, \mathbf{z}^M, \mathbf{a}, \mathbf{c}_u)^t, \mathbf{d}^{t-1}, \Theta_0$
      for $k = 1$ **to** K^U **do**
        $d_j^{k(t)} \sim d_j^k \mid (\mathbf{z}^U, \mathbf{z}^M, \mathbf{a}, \mathbf{b}_j, \mathbf{c})^t, \mathbf{r}^{(\mathrm{v})}, \Theta_0$
      end for
    end for
  end for

movies such as *Napoleon Dynamite* to be captured in part by a corresponding variance in the bias parameters d_j^i (see Section 3.5). Moreover, because the model is symmetric, each rating is also influenced by the item-topic-indexed bias c_u^k. This can be interpreted as the predisposition of each perceived item class towards being liked or disliked by each user in the database. Finally, because M^{3}F-TIB is a mixed-membership model, each user and item can choose a different topic and hence a different bias for each rating (e.g., when multiple users share a single account).

The M^{3}F Topic-Indexed Factor Model

The M^{3}F Topic-Indexed Factor (TIF) model assumes that the joint contextual bias is an inner product of topic-indexed factor vectors, rather than the sum of topic-indexed biases as in the TIB model. Each item topic k maintains a latent factor vector $\mathbf{c}_u^k \in \mathbb{R}^{\tilde{D}}$ for each user, and each user topic i maintains a latent factor vector $\mathbf{d}_j^i \in \mathbb{R}^{\tilde{D}}$ for each item. Each user and each item additionally maintains a single static rating bias, ξ_u and χ_j respectively. The joint contextual bias is formed by summing the user bias, the item bias, and the inner product between the topic-indexed factor vectors:

$$\beta_{uj}^{ik} = \xi_u + \chi_j + \mathbf{c}_u^k \cdot \mathbf{d}_j^i.$$

The topic-indexed factors $\mathbf{c}_u^k$ and $\mathbf{d}_j^i$ are drawn independently from $\mathcal{N}\left(\tilde{\mu}^U, (\tilde{\Lambda}^U)^{-1}\right)$ and $\mathcal{N}\left(\tilde{\mu}^M, (\tilde{\Lambda}^M)^{-1}\right)$ priors, and conjugate normal-Wishart priors are placed on the hyper-parameters $(\tilde{\mu}^U, \tilde{\Lambda}^U)$ and $(\tilde{\mu}^M, \tilde{\Lambda}^M)$. The static user and item biases, ξ_u and χ_j, are drawn independently from Gaussian priors with variance σ_0^2 and means ξ_0 and χ_0 respectively.[2]

Intuitively, the topic-indexed factor model can be interpreted as an extended matrix factorization with both *global* and *local* low-dimensional representations. Each user u has a single global factor $\mathbf{a}_u$ but K^U local factors $\mathbf{c}_u^k$; similarly, each item j has both a global factor $\mathbf{b}_j$ and multiple local factors $\mathbf{d}_j^i$. A strength of latent factor methods is their ability to discover globally predictive intrinsic properties of users and items. The topic-indexed factor model extends this representation to allow for intrinsic properties that are predictive in some but perhaps not all contexts. For example, in the movie-recommendation setting, is *Lost In Translation* a dark comedy or a romance film? The answer may vary from user to user and thus may be captured by different vectors $\mathbf{d}_j^i$ for each user-indexed topic.

3.4 Inference and Prediction

The goal in dyadic data prediction is to predict unobserved ratings $\mathbf{r}^{(\mathrm{h})}$ given observed ratings $\mathbf{r}^{(\mathrm{v})}$. As in Salakhutdinov and Mnih [71, 70] and Takács et al. [75], we adopt root mean squared error (RMSE)[3] as our primary error metric and note that the Bayes optimal prediction under RMSE loss is the posterior mean of the predictive distribution $p(\mathbf{r}^{(\mathrm{h})}|\mathbf{r}^{(\mathrm{v})}, \Theta_0)$.

In our M^{3}F models, the predictive distribution over unobserved ratings is found by integrating out all topics and parameters. The posterior distribution $p(\mathbf{z}^U, \mathbf{z}^M, \Theta^U, \Theta^M|\mathbf{r}^{(\mathrm{v})}, \Theta_0)$ is thus our main inferential quantity of interest. Unfortunately, as in both LDA and BPMF, analytical computation of this posterior is intractable, due to complex coupling in the marginal distribution $p(\mathbf{r}^{(\mathrm{v})}|\Theta_0)$ [6, 70].

[2]Static biases ξ and χ are suppressed in the remainder of the chapter for clarity.

[3]For work linking improved RMSE with better top-K recommendation rankings, see Koren [37].

Table 3.1: 1M MovieLens and EachMovie RMSE scores for varying static factor dimensionalities and topic counts for both M^{3}F models. All scores are averaged across 3 standardized cross-validation splits. Parentheses indicate topic counts (K^U, K^M). For M^{3}F-TIF, $\tilde{D} = 2$ throughout. L&U (2009) refers to [41]. Best results for each D are boldened. Asterisks indicate significant improvement over BPMF under a one-tailed paired t-test with level 0.05.

	1M MovieLens				EachMovie			
Method	**D=10**	**D=20**	**D=30**	**D=40**	**D=10**	**D=20**	**D=30**	**D=40**
BPMF	0.8695	0.8622	0.8621	0.8609	1.1229	1.1212	1.1203	1.1163
M^{3}F-TIB (1,1)	0.8671	0.8614	0.8616	0.8605	1.1205	1.1188	1.1183	1.1168
M^{3}F-TIF (1,2)	0.8664	0.8629	0.8622	0.8616	1.1351	1.1179	1.1095	1.1072
M^{3}F-TIF (2,1)	0.8674	0.8605	0.8605	0.8595	1.1366	1.1161	1.1088	1.1058
M^{3}F-TIF (2,2)	**0.8642**	**0.8584***	0.8584	0.8592	1.1211	1.1043	1.1035	1.1020
M^{3}F-TIB (1,2)	0.8669	0.8611	0.8604	0.8603	1.1217	1.1081	1.1016	1.0978
M^{3}F-TIB (2,1)	0.8649	0.8593	**0.8581***	**0.8577***	1.1186	1.1004	1.0952	1.0936
M^{3}F-TIB (2,2)	0.8658	0.8609	0.8605	0.8599	**1.1101***	**1.0961***	**1.0918***	**1.0905***
L&U (2009)	0.8801 (RBF)		0.8791 (Linear)		1.1111 (RBF)		1.0981 (Linear)	

Inference via Gibbs Sampling

In this chapter, we use a Gibbs sampling MCMC procedure [23] to draw samples of topic and parameter variables $\{(\mathbf{z}^{U(t)}, \mathbf{z}^{M(t)}, \Theta^{U(t)}, \Theta^{M(t)})\}_{t=1}^{T}$ from their joint posterior. Our use of conjugate priors ensures that each Gibbs conditional has a simple closed form (see Section 3.7 for the exact conditional distributions).

Alg. 1 displays the Gibbs sampling algorithm for the M^{3}F-TIB model; the M^{3}F-TIF Gibbs sampler is similar. Note that we choose to sample the topic parameters θ^U and θ^M rather than integrate them out as in a collapsed Gibbs sampler (see, e.g., [66]). This decision allows us to sample the interaction-specific topic variables in parallel. Indeed, each loop in Alg. 1 corresponds to a block of parameters that can be sampled in parallel. In practice, such parallel computation yields substantial savings in sampling time for large-scale dyadic datasets.

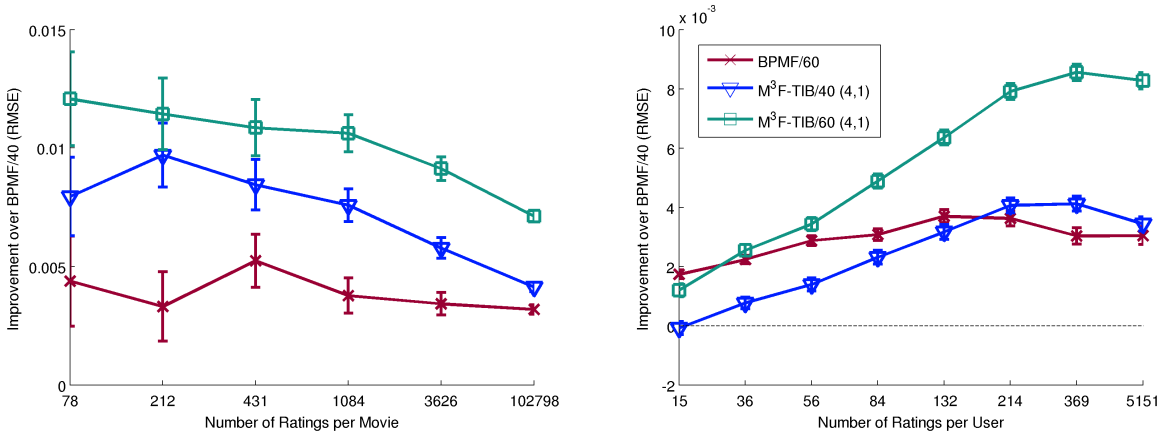

Figure 3.2: RMSE improvements over BPMF/40 on the Netflix Prize as a function of movie or user rating count. Left: Improvement as a function of movie rating count. Each x-axis label represents the average rating count of 1/6 of the movie base. Right: Improvement over BPMF as a function of user rating count. Each bin represents 1/8 of the user base.

Prediction

Given posterior samples of parameters, we can approximate the true predictive distribution by the Monte Carlo expectation

$$\hat{p}(\mathbf{r}^{(\mathrm{h})}|\mathbf{r}^{(\mathrm{v})},\Theta_0) = \frac{1}{T}\sum_{t=1}^{T}\sum_{\mathbf{z}^U,\mathbf{z}^M} p(\mathbf{z}^U,\mathbf{z}^M|\Theta^{U(t)},\Theta^{M(t)}) \\ p(\mathbf{r}^{(\mathrm{h})}|\mathbf{z}^U,\mathbf{z}^M,\Theta^{U(t)},\Theta^{M(t)},\Theta_0), \tag{3.1}$$

where we have integrated over the unknown topic variables. Eq. 3.1 yields the following posterior mean prediction for each user-item pair under the M^{3}F-TIB model:

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{a}_u^{(t)}\cdot\mathbf{b}_j^{(t)} + \sum_{k=1}^{K^M} c_u^{k(t)}\theta_{jk}^{M(t)} + \sum_{i=1}^{K^U} d_j^{i(t)}\theta_{ui}^{U(t)}\right).$$

Under the M^{3}F-TIF model, posterior mean prediction takes the form

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{a}_u^{(t)}\cdot\mathbf{b}_j^{(t)} + \sum_{i=1}^{K^U}\sum_{k=1}^{K^M}\theta_{ui}^{U(t)}\theta_{jk}^{M(t)}\mathbf{c}_u^{k(t)}\cdot\mathbf{d}_j^{i(t)}\right).$$

3.5 Experimental Evaluation

We evaluate our models on several movie rating collaborative filtering datasets including the Netflix Prize dataset[4], the EachMovie dataset, and the 1M and 10M MovieLens datasets[5]. The Netflix Prize dataset contains 100 million ratings in $\{1, \ldots, 5\}$ distributed across 17,770 movies and 480,189 users. The EachMovie dataset contains 2.8 million ratings in $\{1, \ldots, 6\}$ distributed across 1,648 movies and 74,424 users. The 1M MovieLens dataset has 6,040 users, 3,952 movies, and 1 million ratings in $\{1, \ldots, 5\}$. The 10M MovieLens dataset has 10,681 movies, 71,567 users, and 10 million ratings on a .5 to 5 scale with half-star increments. In all experiments, we set W_0 equal to the identity matrix, ν_0 equal to the number of static matrix factors, μ_0 equal to the all-zeros vector, χ_0 equal to the mean rating in the data set, and $(\lambda_0, \sigma^2, \sigma_0^2) = (10, .5, .1)$. For M^{3}F-TIB experiments, we set $(c_0, d_0, \alpha) = (0, 0, 10000)$, and for M^{3}F-TIF, we set $\tilde{W}_0$ equal to the identity matrix, $\tilde{\nu}_0$ equal to the number of topic-indexed factors, $\tilde{\mu}_0$ equal to the all-zeros vector, and $(\tilde{D}, \xi_0, \alpha, \tilde{\lambda}_0) = (2, 0, 10, 10000)$. Free parameters were selected by grid search on an EachMovie hold-out set, disjoint from the test sets used for evaluation. Throughout, reported error intervals are of plus or minus one standard error from the mean.

1M MovieLens and EachMovie Datasets

We first evaluated our models on the smaller datasets, 1M MovieLens and EachMovie. We conducted the "weak generalization" ratings prediction experiment of Marlin [50], where, for each user in the training set, a single rating is withheld for the test set. All reported results are averaged over the same 3 random train-test splits used in [51, 50, 68, 18, 61, 41]. Our Gibbs samplers were initialized with draws from the prior and run for 3000 samples for M^{3}F-TIB and 512 samples for M^{3}F-TIF. No samples were discarded for "burn-in."

Table 3.1 reports the predictive performance of our models for a variety of static factor dimensionalities (D) and topic counts (K^U, K^M). We compared all models against BPMF as a baseline by running the M^{3}F-TIB model with K^U and K^M set to zero. For comparison with previous results that report the normalized mean average error (NMAE) of Marlin [50], we additionally ran M^{3}F-TIB with $(D, K^U, K^M) = (300, 2, 1)$ on EachMovie and achieved a weak RMSE of $(\mathbf{1.0878} \pm 0.0025)$ and a weak NMAE of $(\mathbf{0.4293} \pm 0.0013)$.

On both the EachMovie and the 1M MovieLens datasets, both M^{3}F models systematically outperformed the BPMF baseline for almost every setting of latent dimensionality and topic counts. For $D = 20$, increasing K^U to 2 provided a boost in accuracy for both M^{3}F models equivalent to doubling the number of BPMF static factor parameters ($D = 40$). We also found that the M^{3}F-TIB model outperformed the more recent Gaussian process matrix factorization model of Lawrence and Urtasun [41].

The results indicate that the mixed-membership component of M^{3}F offers greater predictive power than simply increasing the dimensionality of a pure latent factor model. While

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

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

the M^{3}F-TIF model sometimes failed to outperform the BPMF baseline due to overfitting, the M^{3}F-TIB model always outperformed BPMF regardless of the setting of K^U, K^M, or D. Note that the increase in the number of parameters from the BPMF model to the M^{3}F models is independent of D (M^{3}F-TIB requires $(U+M)(K^U+K^M)$ more parameters than BPMF with equal D), and therefore the ratio of the number of parameters of BPMF and M^{3}F approaches 1 if D increases while K^U, K^M, and $\tilde{D}$ are held fixed. Nonetheless, the modeling of joint contextual bias in the M^{3}F-TIB model continues to improve predictive performance even as D increases, suggesting that the M^{3}F-TIB model is capturing aspects of the data that are not captured by a pure latent factor model.

Finally, because the M^{3}F-TIB model offered superior performance to the M^{3}F-TIF model in most experiments, we focus on the M^{3}F-TIB model in the remainder of this section.

10M MovieLens Dataset

For the larger datasets, we initialized the Gibbs samplers with MAP estimates of **a** and **b** under simple Gaussian priors, which we trained with stochastic gradient descent. This is similar to the PMF initialization scheme of Salakhutdinov and Mnih [70]. All other parameters were initialized to their model means.

For the 10M MovieLens dataset, we averaged our results across the r_a and r_b train-test splits provided with the dataset after removing those test set ratings with no corresponding item in the training set. For comparison with the Gaussian process matrix factorization model of Lawrence and Urtasun [41], we adopted a static factor dimensionality of $D = 10$. Our M^{3}F-TIB model with $(K^U, K^M) = (4, 1)$ achieved an RMSE of (**0.8447** $\pm$ 0.0095), representing a significant improvement ($p = 0.034$) over BPMF with RMSE (**0.8472** $\pm$ 0.0093) and a substantial increase in accuracy over the Gaussian process model with RMSE (**0.8740** $\pm$ 0.0197).

Netflix Prize Dataset

The unobserved ratings for the 100 million dyad Netflix Prize dataset are partitioned into two standard sets, known as the Quiz Set and the Test Set. Prior to September of 2009, public evaluation was only available on the Quiz Set, and, as a result, most prior published "test set" results were evaluated on the Quiz Set. In Table 3.2, we compare the performance of BPMF and M^{3}F-TIB with $(K^U, K^M) = (4, 1)$ on the Quiz Set, the Test Set, and on their union (the Qualifying Set), across a wide range of static dimensionalities. We also report running times of our Matlab/MEX implementation on dual quad-core 2.67GHz Intel Xeon CPUs. We used the initialization scheme described in Section 3.5 and ran the Gibbs samplers for 500 iterations.

In addition to outperforming the BPMF baselines of comparable dimensionality, the M^{3}F-TIB models routinely proved to be more accurate than higher dimensional BPMF models with longer running times and many more learned parameters. This major advantage of

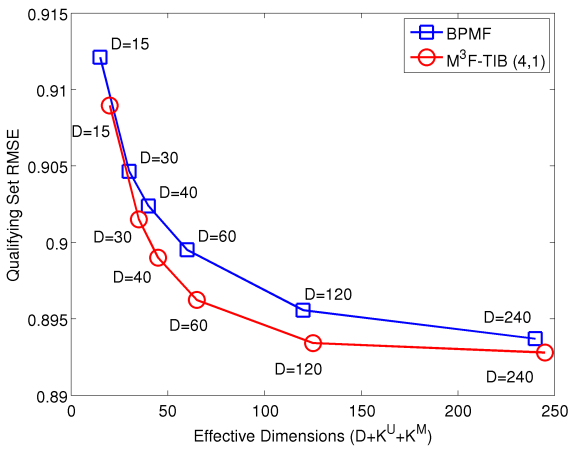

Figure 3.3: RMSE performance of BPMF and M^{3}F-TIB with $(K^U, K^M) = (4, 1)$ on the Netflix Prize Qualifying set as a function of the number of parameters modeled per user or item.

M^{3}F modeling is highlighted in Figure 3.3, which plots error as a function of the number of parameters modeled per user or item $(D + K^U + K^M)$.

To determine where our models were providing the most improvement over BPMF, we divided the Qualifying Set into bins based on the number of ratings associated with each user and movie in the database. Figure 3.2 displays the improvements of BPMF/60, M^{3}F-TIB/40, and M^{3}F-TIB/60 over BPMF/40 as a function of the number of user or movie ratings. Consistent with our expectations, we found that adopting an M^{3}F model yielded improved accuracy for movies of small rating counts, with the greatest improvement over BPMF occurring for those high-variance movies with relatively few ratings. Moreover, the improvements realized by either M^{3}F-TIB model uniformly dominated the improvements realized by BPMF/60 across movie rating counts. At the same time, we found that the improvements of the M^{3}F-TIB models were skewed toward users with larger rating counts.

M^{3}F & The *Napoleon Dynamite* Effect

In our introduction to the M^{3}F-TIB model we discussed the joint contextual bias as a potential solution to the problem of making predictions for movies that have high variance. To investigate whether or not M^{3}F-TIB achieved progress towards this goal, we analyzed the correlation between the improvement in RMSE over the BPMF baseline and the variance of ratings for the 1000 most popular movies in the database. While the improvements for BPMF/60 were not significantly correlated with movie variance ($\rho = -0.016$), the improvements of the M^{3}F-TIB models were strongly correlated with $\rho = 0.117 (p < 0.001)$ and $\rho = 0.15$ $(p < 10^{-7})$ for the $(40, 4, 1)$ and $(60, 4, 1)$ models, respectively. These results indicate that a strength of the M^{3}F-TIB model lies in the ability of the topic-indexed biases

Table 3.2: Netflix Prize results for BPMF and M^{3}F-TIB with $(K^U, K^M) = (4, 1)$. Hidden ratings are partitioned into Quiz and Test sets; the Qualifying set is their union. Best results in each block are boldened. Reported times are average running times per sample.

Method	**Test**	**Quiz**	**Qual**	**Time**
BPMF/15	0.9125	0.9117	0.9121	27.8s
TIB/15	**0.9093**	**0.9086**	**0.9090**	46.3s
BPMF/30	0.9049	0.9044	0.9047	38.6s
TIB/30	**0.9018**	**0.9012**	**0.9015**	56.9s
BPMF/40	0.9029	0.9026	0.9027	48.3s
TIB/40	**0.8992**	**0.8988**	**0.8990**	70.5s
BPMF/60	0.9004	0.9001	0.9002	94.3s
TIB/60	**0.8965**	**0.8960**	**0.8962**	97.0s
BPMF/120	0.8958	0.8953	0.8956	273.7s
TIB/120	**0.8937**	**0.8931**	**0.8934**	285.2s
BPMF/240	0.8939	0.8936	0.8938	1152.0s
TIB/240	**0.8931**	**0.8927**	**0.8929**	1158.2s

to model variance in user biases toward specific items.

To further illuminate this property of the model, we computed the posterior expectation of the movie bias parameters, $\mathbb{E}\mathbf{d}_j|\mathbf{r}^{(\mathrm{v})}$, for the 200 most popular movies in the database. For these movies, the variance of $\mathbb{E}d^i_j|\mathbf{r}^{(\mathrm{v})}$ across topics and the variance of the ratings of these movies were very strongly correlated ($\rho = 0.682, p < 10^{-10}$). The five movies with the highest and lowest variance in $\mathbb{E}d^i_j|\mathbf{r}^{(\mathrm{v})}$ across topics are shown in Table 3.3. The results are easily interpretable, with high-variance movies such as *Napoleon Dynamite* dominating the high-variance positions and universally acclaimed blockbusters dominating the low-variance positions.

3.6 Conclusion

In this chapter, we developed a fully Bayesian dyadic data prediction framework for integrating the complementary approaches of discrete mixed membership modeling and continuous latent factor modeling. We introduced two Mixed Membership Matrix Factorization models, developed MCMC inference procedures, and evaluated our methods on the EachMovie, MovieLens, and Netflix Prize datasets. On each dataset, we found that M^{3}F-TIB significantly outperformed BPMF and other state-of-the-art baselines, even when fitting fewer parameters. We further discovered that the greatest performance improvements occurred for the high-variance, sparsely-rated items, for which accurate DDP is typically the hardest.

Table 3.3: Top 200 Movies from the Netflix Prize dataset with the highest and lowest cross-topic variance in $\mathbb{E}d_j^i|\mathbf{r}^{(\mathrm{v})}$. Reported intervals are of the mean value of $\mathbb{E}d_j^i|\mathbf{r}^{(\mathrm{v})}$ plus or minus one standard deviation.

Movie Title	$\mathbb{E}d_j^i\|\mathbf{r}^{(\mathrm{v})}$
Napoleon Dynamite	-0.11 ± 0.93
Fahrenheit 9/11	-0.06 ± 0.90
Chicago	-0.12 ± 0.78
The Village	-0.14 ± 0.71
Lost in Translation	-0.02 ± 0.70
LotR: The Fellowship of the Ring	0.15 ± 0.00
LotR: The Two Towers	0.18 ± 0.00
LotR: The Return of the King	0.24 ± 0.00
Star Wars: Episode V	0.35 ± 0.00
Raiders of the Lost Ark	0.29 ± 0.00

3.7 Gibbs Sampling Conditionals for M^{3}F Models

The M^{3}F-TIB Model

In this section, we specify the conditional distributions used by the Gibbs sampler for the M^{3}F-TIB model.

Normal-Wishart Parameters

$$\Lambda^U|rest\backslash\{\mu^U\} \sim \text{Wishart}((\mathbf{W}_0^{-1} + \sum_{u=1}^{U}(\mathbf{a}_u - \bar{\mathbf{a}})(\mathbf{a}_u - \bar{\mathbf{a}})^t + \frac{\lambda_0 U}{\lambda_0 + U}(\mu_0 - \bar{\mathbf{a}})(\mu_0 - \bar{\mathbf{a}})^t)^{-1},$$
$$\nu_0 + U) \text{ where } \bar{\mathbf{a}} = \tfrac{1}{U}\textstyle\sum_{u=1}^{U}\mathbf{a}_u.$$

$$\Lambda^M|rest\backslash\{\mu^M\} \sim \text{Wishart}((\mathbf{W}_0^{-1} + \sum_{j=1}^{M}(\mathbf{b}_j - \bar{\mathbf{b}})(\mathbf{b}_j - \bar{\mathbf{b}})^t + \frac{\lambda_0 M}{\lambda_0 + M}(\mu_0 - \bar{\mathbf{b}})(\mu_0 - \bar{\mathbf{b}})^t)^{-1},$$
$$\nu_0 + M) \text{ where } \bar{\mathbf{b}} = \tfrac{1}{M}\textstyle\sum_{j=1}^{M}\mathbf{b}_j.$$

$$\mu^U|rest \sim \mathcal{N}\left(\frac{\lambda_0\mu_0 + \sum_{u=1}^{U}\mathbf{a}_u}{\lambda_0 + U}, (\Lambda^U(\lambda_0 + U))^{-1}\right).$$

$$\mu^M|rest \sim \mathcal{N}\left(\frac{\lambda_0\mu_0 + \sum_{j=1}^{M}\mathbf{b}_j}{\lambda_0 + M}, (\Lambda^M(\lambda_0 + M))^{-1}\right).$$

Bias Parameters

For each u and $i \in \{1, \ldots, K^M\}$,

$$c_u^i | rest \sim \mathcal{N}\left(\frac{\frac{c_0}{\sigma_0^2} + \sum_{j \in V_u} \frac{1}{\sigma^2} z_{uji}^M (r_{uj} - \chi_0 - d_j^{z_{uj}^U} - \mathbf{a}_u \cdot \mathbf{b}_j)}{\frac{1}{\sigma_0^2} + \sum_{j \in V_u} \frac{1}{\sigma^2} z_{uji}^M}, \frac{1}{\frac{1}{\sigma_0^2} + \sum_{j \in V_u} \frac{1}{\sigma^2} z_{uji}^M} \right).$$

For each j and $i \in \{1, \ldots, K^U\}$,

$$d_j^i | rest \sim \mathcal{N}\left(\frac{\frac{d_0}{\sigma_0^2} + \sum_{u : j \in V_u} \frac{1}{\sigma^2} z_{uji}^U (r_{uj} - \chi_0 - c_u^{z_{uj}^M} - \mathbf{a}_u \cdot \mathbf{b}_j)}{\frac{1}{\sigma_0^2} + \sum_{u : j \in V_u} \frac{1}{\sigma^2} z_{uji}^U}, \frac{1}{\frac{1}{\sigma_0^2} + \sum_{u : j \in V_u} \frac{1}{\sigma^2} z_{uji}^U} \right).$$

Static Factors

For each u,

$$\mathbf{a}_u | rest \sim \mathcal{N}\left((\Lambda_u^{U*})^{-1} (\Lambda^U \mu^U + \sum_{j \in V_u} \frac{1}{\sigma^2} \mathbf{b}_j (r_{uj} - \chi_0 - c_u^{z_{uj}^M} - d_j^{z_{uj}^U})), (\Lambda_u^{U*})^{-1} \right)$$

where $\Lambda_u^{U*} = (\Lambda^U + \sum_{j \in V_u} \frac{1}{\sigma^2} \mathbf{b}_j (\mathbf{b}_j)^t)$.

For each j,

$$\mathbf{b}_j | rest \sim \mathcal{N}\left((\Lambda_j^{M*})^{-1} (\Lambda^M \mu^M + \sum_{u : j \in V_u} \frac{1}{\sigma^2} \mathbf{a}_u (r_{uj} - \chi_0 - c_u^{z_{uj}^M} - d_j^{z_{uj}^U})), (\Lambda_j^{M*})^{-1} \right)$$

where $\Lambda_j^{M*} = (\Lambda^M + \sum_{u : j \in V_u} \frac{1}{\sigma^2} \mathbf{a}_u (\mathbf{a}_u)^t)$.

Dirichlet Parameters

For each u, $\theta_u^U | rest \sim Dir(\alpha / K^U + \sum_{j \in V_u} z_{uj}^U)$.

For each j, $\theta_j^M | rest \sim Dir(\alpha / K^M + \sum_{u : j \in V_u} z_{uj}^M)$.

Topic Variables

For each u and $j \in V_u$, $z_{uj}^U | rest \sim \text{Multi}(1, \theta_{uj}^{U*})$ where

$$\theta_{uji}^{U*} \propto \theta_{ui}^U \exp\left(-\frac{(r_{uj} - \chi_0 - c_u^{z_{uj}^M} - d_j^i - \mathbf{a}_u \cdot \mathbf{b}_j)^2}{2\sigma^2} \right).$$

For each j and $u : j \in V_u$, $z_{uj}^M | rest \sim \text{Multi}(1, \theta_{uj}^{M*})$ where

$$\theta_{uji}^{M*} \propto \theta_{ji}^M \exp\left(-\frac{(r_{uj} - \chi_0 - c_u^i - d_j^{z_{uj}^U} - \mathbf{a}_u \cdot \mathbf{b}_j)^2}{2\sigma^2} \right).$$

The M^{3}F-TIF Model

In this section, we specify the conditional distributions used by the Gibbs sampler for the M^{3}F-TIF model.

Normal-Wishart Parameters

$\Lambda^U|rest\backslash\{\mu^U\} \sim \text{Wishart}((\mathbf{W}_0^{-1} + \sum_{u=1}^{U}(\mathbf{a}_u - \bar{\mathbf{a}})(\mathbf{a}_u - \bar{\mathbf{a}})^t + \frac{\lambda_0 U}{\lambda_0 + U}(\mu_0 - \bar{\mathbf{a}})(\mu_0 - \bar{\mathbf{a}})^t)^{-1}, \nu_0 + U)$ where $\bar{\mathbf{a}} = \frac{1}{U}\sum_{u=1}^{U}\mathbf{a}_u$.

$\Lambda^M|rest\backslash\{\mu^M\} \sim \text{Wishart}((\mathbf{W}_0^{-1} + \sum_{j=1}^{M}(\mathbf{b}_j - \bar{\mathbf{b}})(\mathbf{b}_j - \bar{\mathbf{b}})^t + \frac{\lambda_0 M}{\lambda_0 + M}(\mu_0 - \bar{\mathbf{b}})(\mu_0 - \bar{\mathbf{b}})^t)^{-1}, \nu_0 + M)$ where $\bar{\mathbf{b}} = \frac{1}{M}\sum_{j=1}^{M}\mathbf{b}_j$.

$$\mu^U|rest \sim \mathcal{N}\left(\frac{\lambda_0\mu_0 + \sum_{u=1}^{U}\mathbf{a}_u}{\lambda_0 + U}, (\Lambda^U(\lambda_0 + U))^{-1}\right).$$

$$\mu^M|rest \sim \mathcal{N}\left(\frac{\lambda_0\mu_0 + \sum_{j=1}^{M}\mathbf{b}_j}{\lambda_0 + M}, (\Lambda^M(\lambda_0 + M))^{-1}\right).$$

$\tilde{\Lambda}^U|rest\backslash\{\tilde{\mu}^U\} \sim \text{Wishart}((\tilde{\mathbf{W}}_0^{-1} + \sum_{u=1}^{U}\sum_{i=1}^{K^M}(\mathbf{c}_u^i - \bar{\mathbf{c}})(\mathbf{c}_u^i - \bar{\mathbf{c}})^t + \frac{\tilde{\lambda}_0 U K^M}{\tilde{\lambda}_0 + U K^M}(\tilde{\mu}_0 - \bar{\mathbf{c}})(\tilde{\mu}_0 - \bar{\mathbf{c}})^t)^{-1}, \tilde{\nu}_0 + UK^M)$ where $\bar{\mathbf{c}} = \frac{1}{UK^M}\sum_{u=1}^{U}\sum_{i=1}^{K^M}\mathbf{c}_u^i$.

$\tilde{\Lambda}^M|rest\backslash\{\tilde{\mu}^M\} \sim \text{Wishart}((\tilde{\mathbf{W}}_0^{-1} + \sum_{j=1}^{M}\sum_{i=1}^{K^U}(\mathbf{d}_j^i - \bar{\mathbf{d}})(\mathbf{d}_j^i - \bar{\mathbf{d}})^t + \frac{\tilde{\lambda}_0 M K^U}{\tilde{\lambda}_0 + M K^U}(\tilde{\mu}_0 - \bar{\mathbf{d}})(\tilde{\mu}_0 - \bar{\mathbf{d}})^t)^{-1}, \tilde{\nu}_0 + MK^U)$ where $\bar{\mathbf{d}} = \frac{1}{MK^U}\sum_{j=1}^{M}\sum_{i=1}^{K^U}\mathbf{d}_j^i$.

$$\tilde{\mu}^U|rest \sim \mathcal{N}\left(\frac{\tilde{\lambda}_0\tilde{\mu}_0 + \sum_{u=1}^{U}\sum_{i=1}^{K^M}\mathbf{c}_u^i}{\tilde{\lambda}_0 + UK^M}, (\tilde{\Lambda}^U(\tilde{\lambda}_0 + UK^M))^{-1}\right).$$

$$\tilde{\mu}^M|rest \sim \mathcal{N}\left(\frac{\tilde{\lambda}_0\tilde{\mu}_0 + \sum_{j=1}^{M}\sum_{i=1}^{K^U}\mathbf{d}_j^i}{\tilde{\lambda}_0 + MK^U}, (\tilde{\Lambda}^M(\tilde{\lambda}_0 + MK^U))^{-1}\right).$$

Bias Parameters

For each u,

$$\xi_u|rest \sim \mathcal{N}\left(\frac{\frac{\xi_0}{\sigma_0^2} + \sum_{j\in V_u}\frac{1}{\sigma^2}(r_{uj} - \chi_j - \mathbf{a}_u\cdot\mathbf{b}_j - \mathbf{c}_u^{z_{uj}^M}\cdot\mathbf{d}_j^{z_{uj}^U})}{\frac{1}{\sigma_0^2} + \sum_{j\in V_u}\frac{1}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2} + \sum_{j\in V_u}\frac{1}{\sigma^2}}\right).$$

For each j,

$$\chi_j|rest \sim \mathcal{N}\left(\frac{\frac{\chi_0}{\sigma_0^2} + \sum_{u:j\in V_u}\frac{1}{\sigma^2}(r_{uj} - \xi_u - \mathbf{a}_u\cdot\mathbf{b}_j - \mathbf{c}_u^{z_{uj}^M}\cdot\mathbf{d}_j^{z_{uj}^U})}{\frac{1}{\sigma_0^2} + \sum_{u:j\in V_u}\frac{1}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2} + \sum_{u:j\in V_u}\frac{1}{\sigma^2}}\right).$$

Static Factors

For each u,

$$\mathbf{a}_u|rest \sim \mathcal{N}\left((\Lambda_u^{U*})^{-1}(\Lambda^U\mu^U + \sum_{j\in V_u}\frac{1}{\sigma^2}\mathbf{b}_j(r_{uj} - \xi_u - \chi_j - \mathbf{c}_u^{z_{uj}^M}\cdot\mathbf{d}_j^{z_{uj}^U})), (\Lambda_u^{U*})^{-1}\right)$$

where $\Lambda_u^{U*} = (\Lambda^U + \sum_{j\in V_u}\frac{1}{\sigma^2}\mathbf{b}_j(\mathbf{b}_j)^t)$.

For each j,

$$\mathbf{b}_j|rest \sim \mathcal{N}\left((\Lambda_j^{M*})^{-1}(\Lambda^M\mu^M + \sum_{u:j\in V_u}\frac{1}{\sigma^2}\mathbf{a}_u(r_{uj} - \xi_u - \chi_j - \mathbf{c}_u^{z_{uj}^M}\cdot\mathbf{d}_j^{z_{uj}^U})), (\Lambda_j^{M*})^{-1}\right)$$

where $\Lambda_j^{M*} = (\Lambda^M + \sum_{u:j\in V_u}\frac{1}{\sigma^2}\mathbf{a}_u(\mathbf{a}_u)^t)$.

Topic-indexed Factors

For each u and each $i \in 1,\ldots,K^M$,

$$\mathbf{c}_u^i|rest \sim \mathcal{N}\left((\tilde{\Lambda}_{ui}^{U*})^{-1}(\tilde{\Lambda}^U\tilde{\mu}^U + \sum_{j\in V_u}\frac{1}{\sigma^2}z_{uji}^M\mathbf{d}_j^{z_{uj}^U}(r_{uj} - \xi_u - \chi_j - \mathbf{a}_u\cdot\mathbf{b}_j)), (\tilde{\Lambda}_{ui}^{U*})^{-1}\right)$$

where $\tilde{\Lambda}_{ui}^{U*} = (\tilde{\Lambda}^U + \sum_{j\in V_u}\frac{1}{\sigma^2}z_{uji}^M\mathbf{d}_j^{z_{uj}^U}(\mathbf{d}_j^{z_{uj}^U})^t)$.

For each j and each $i \in 1,\ldots,K^U$,

$$\mathbf{d}_j^i|rest \sim \mathcal{N}\left((\tilde{\Lambda}_{ji}^{M*})^{-1}(\tilde{\Lambda}^M\tilde{\mu}^M + \sum_{u:j\in V_u}\frac{1}{\sigma^2}z_{uji}^U\mathbf{c}_u^{z_{uj}^M}(r_{uj} - \xi_u - \chi_j - \mathbf{a}_u\cdot\mathbf{b}_j)), (\tilde{\Lambda}_{ji}^{M*})^{-1}\right)$$

where $\tilde{\Lambda}_{ji}^{M*} = (\tilde{\Lambda}^M + \sum_{u:j\in V_u}\frac{1}{\sigma^2}z_{uji}^U\mathbf{c}_u^{z_{uj}^M}(\mathbf{c}_u^{z_{uj}^M})^t)$.

Dirichlet Parameters

For each u, $\theta_u^U | rest \sim Dir(\alpha/K^U + \sum_{j \in V_u} z_{uj}^U)$.

For each j, $\theta_j^M | rest \sim Dir(\alpha/K^M + \sum_{u:j \in V_u} z_{uj}^M)$.

Topic Variables

For each u and $j \in V_u$, $z_{uj}^U | rest \sim \text{Multi}(1, \theta_{uj}^{U*})$ where

$$\theta_{uji}^{U*} \propto \theta_{ui}^U \exp\left(-\frac{(r_{uj} - \xi_u - \chi_j - \mathbf{a}_u \cdot \mathbf{b}_j - \mathbf{c}_u^{z_{uj}^M} \cdot \mathbf{d}_j^i)^2}{2\sigma^2} \right).$$

For each j and $u : j \in V_u$, $z_{uj}^M | rest \sim \text{Multi}(1, \theta_{uj}^{M*})$ where

$$\theta_{uji}^{M*} \propto \theta_{ji}^M \exp\left(-\frac{(r_{uj} - \xi_u - \chi_j - \mathbf{a}_u \cdot \mathbf{b}_j - \mathbf{c}_u^i \cdot \mathbf{d}_j^{z_{uj}^U})^2}{2\sigma^2} \right).$$

Chapter 4

Divide-and-Conquer Matrix Factorization

4.1 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 [10, 11, 12], 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 [10], graphical model selection [12], document modeling [53], and image alignment [63].

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, 10, 11, 58], and a variety of algorithms (e.g., [43, 45, 77]) 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 [49]. 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 speed-ups in practice, while our theory provides high-probability recovery guarantees for DFC comparable to those enjoyed by its base algorithm.

The remainder of the chapter is organized as follows. In Section 4.2, 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 4.3. Our theoretical analysis follows in Section 4.4. 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.

4.2 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.

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

$$(\mathcal{P}_\Omega(\mathbf{M}))_{ij} = \mathbf{M}_{ij}, \text{ if } (i,j) \in \Omega \quad \text{and} \quad (\mathcal{P}_\Omega(\mathbf{M}))_{ij} = 0 \text{ otherwise.}^1$$

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.

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

- **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 $\mathcal{P}_\Omega(\mathbf{M}) = \mathbf{M}$.

Divide-Factor-Combine

Algorithms 2 and 3 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, $\{\mathcal{P}_\Omega(\mathbf{C}_1), \ldots, \mathcal{P}_\Omega(\mathbf{C}_t)\}$[2], 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 4.2.

Algorithm 2 DFC-Proj

Input: $\mathcal{P}_\Omega(\mathbf{M})$, t
$\{\mathcal{P}_\Omega(\mathbf{C}_i)\}_{1 \le i \le t} = \textsc{SampCol}(\mathcal{P}_\Omega(\mathbf{M}), t)$
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 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
$\hat{\mathbf{L}}^{nys} = \textsc{GenNyström}(\hat{\mathbf{C}}, \hat{\mathbf{R}})$

[2]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.

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 $\mathbf{M}$.

Column Projection This approximation, introduced by Frieze, Kannan, and Vempala [22], 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 [40] 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 [81] and has been generalized for arbitrary real-valued matrices [25]. 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 [40] 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 $\mathbf{U}_W^{\top}\mathbf{R}$.

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 $\mathrm{O}(mlk_{C_i})$ time for $\mathbf{C}_i$ (or $\mathbf{C}$) and $\mathrm{O}(ndk_R)$ 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 $\mathrm{O}(mk^2+lk^2)$ time for $k \triangleq \max_i k_{C_i}$: $\mathrm{O}(mk^2+lk^2)$ time for the pseudoinversion of $\hat{\mathbf{C}}_1$ and $\mathrm{O}(mk^2+lk^2)$ time for matrix multiplication with each $\hat{\mathbf{C}}_i$ in parallel. Similarly, the generalized Nyström step of DFC-Nys requires only $\mathrm{O}(l\bar{k}^2 + d\bar{k}^2 + \min(m,n)\bar{k}^2)$ 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.

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 [39]. 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 4.3 highlights the empirical effectiveness of ensembling.

4.3 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 [77] as our base noisy MC algorithm and the APG algorithm of [43] as our base noisy RMF algorithm. In all experiments, we use the default parameter settings suggested by [77] and [43], 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.

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., [10, 35, 84]. 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.

We first explored the recovery error of DFC as a function of s, using ($m = 10$K, $r = 10$) with varying observation sparsity for MC and ($m = 1$K, $r = 10$) with a varying percentage

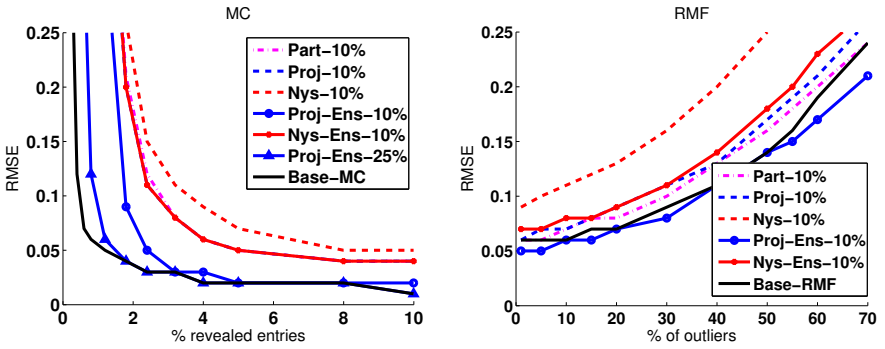

Figure 4.1: Recovery error of DFC relative to base algorithms.

of outliers for RMF. The results are summarized in Figure 4.1.[3] 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 4.1 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 4.2, 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.

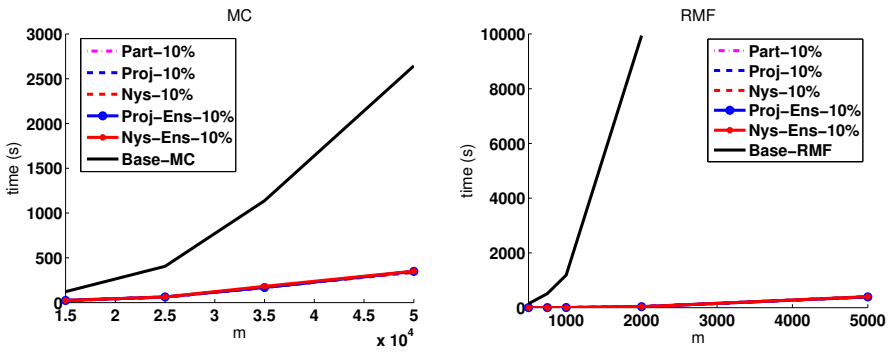

Figure 4.2: Speed-up of DFC relative to base algorithms.

[3]In the left-hand plot of Figure 4.1, the lines for Proj-10% and Proj-Ens-10% overlap.

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[4] ($m = 4$K, $n = 6$K, $s > 10$M) and the Netflix Prize dataset[5] ($m = 18$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 4.3. 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 4.1: 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

[4] http://www.grouplens.org/

[5] http://www.netflixprize.com/

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 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 [10], while 'Lobby' (1546 frames of size 168×120) includes many changes in illumination (a smaller video with 250 frames was studied by [10]). 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 4.3. 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 4.3: Sample 'Hall' recovery by APG, DFC-Proj-Ens-5%, and DFC-Proj-Ens-.5%.

4.4 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.

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 [11, 35, 58] and RMF [1, 84]. 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 [67]:

Definition 3 (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:*

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

Definition 4 (μ_1-Coherence)**.** *Let $\mathbf{L} \in \mathbb{R}^{m\times n}$ have rank r. Then, the μ_1-coherence of $\mathbf{L}$ is:*

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

For any $\mu > 0$, we will call a matrix $\mathbf{L}$ (μ, r)-*coherent* if $\text{rank}(\mathbf{L}) = r$, $\max(\mu_0(\mathbf{U}_L), \mu_0(\mathbf{V}_L)) \leq \mu$, 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.

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 $\left\|\mathcal{P}_\Omega(\mathbf{Z}_0)\right\|_F \leq \Delta$. 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. 5 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 Section 4.5.

Theorem 5. *Choose $t = n/l$ and $l \geq cr\mu\log(n)\log(2/\delta)/\epsilon^2$, where c is a fixed positive constant, and fix any $c_e \geq 0$. Under the notation of Algorithm 2, if a base MF algorithm yields $\mathbf{P}\Big(\left\|\mathbf{C}_{0,i} - \hat{\mathbf{C}}_i\right\|_F > c_e\sqrt{ml}\Delta \mid A(\mathbf{C}_{0,i})\Big) \leq \delta_C$ 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*

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

Under Algorithm 3, if a base MF algorithm yields $\mathbf{P}\Big(\left\|\mathbf{C}_0 - \hat{\mathbf{C}}\right\|_F > c_e\sqrt{ml}\Delta \mid A(\mathbf{C})\Big) \leq \delta_C$ and $\mathbf{P}\Big(\left\|\mathbf{R}_0 - \hat{\mathbf{R}}\right\|_F > c_e\sqrt{dn}\Delta \mid A(\mathbf{R})\Big) \leq \delta_R$ for $d \geq cl\mu_0(\hat{\mathbf{C}})\log(m)\log(1/\delta)/\epsilon^2$, then, with probability at least $(1-\delta)(1-\delta-0.2)(1-\delta_C-\delta_R)$, DFC-Nys *guarantees*

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

To understand the conclusions of Thm. 5, consider a typical base algorithm which, when applied to $\mathcal{P}_\Omega(\mathbf{M})$, recovers an estimate $\hat{\mathbf{L}}$ satisfying $\left\|\mathbf{L}_0 - \hat{\mathbf{L}}\right\|_F \leq c_e\sqrt{mn}\Delta$ with high probability. Thm. 5 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 $2+3\epsilon$.[6] 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. 5 can be applied to derive specific DFC recovery guarantees for noisy MC and noisy RMF. In these sections, we let $\bar{n} \triangleq \max(m, n)$.

Consequences for Noisy MC

Our first corollary of Thm. 5 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 [11]:

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

Then, Cor. 6 follows from a novel guarantee for noisy convex MC, proved in the Section 4.14.

Corollary 6. *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 $1 < \beta \leq \beta_s$. Then, if $\|\mathcal{P}_\Omega(\mathbf{M}) - \mathcal{P}_\Omega(\mathbf{L}_0)\|_F \leq \Delta$ a.s., it suffices to choose $t = n/l$ and

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

to achieve

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

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

with probability at least

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

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

respectively, with c as in Thm. 5 and c_e' a positive constant.

[6] 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. 6 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. 6 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. 6 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. 6.

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 [84]:

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

with $\lambda = 1/\sqrt{\bar{n}}$. Then, Cor. 7 follows from Thm. 5 and the noisy RMF guarantee of [84, Thm. 2].

Corollary 7. *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 $4\beta_s - 3/\rho_s \leq \beta' \leq \beta_s$. Then, if $\|\mathbf{M}-\mathbf{L}_0-\mathbf{S}_0\|_F \leq \Delta$ a.s., it suffices to choose $t = n/l$ and

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

to have

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

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

with probability at least

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

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

respectively, with c as in Thm. 5 and $\rho_r, c_e'', $ and c_p positive constants.

Note that Cor. 7 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.

4.5 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.

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. 4.7.

Lemma 8. *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. 9, 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 $1 - \delta/n$.

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 [19] and the matrix concentration analysis of [31] and immediately gives rise to estimation guarantees for column projection and the generalized Nyström method. The proof of Thm. 9 will be given in Sec. 4.8.

Theorem 9. *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. 9 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 [19] by providing guarantees relative to an arbitrary low-rank approximation. The proof is given in Sec. 4.9.

Corollary 10. *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. 9 and Cor. 10 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. 4.10.

Corollary 11. *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. 10, 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 $(1-\delta)(1-\delta'-0.2)$*.*

4.6 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 [57] 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 [57] 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.

4.7 Proof of Lemma 8

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 13 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 [56] 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 13, 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

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

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 13 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 13, 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 13 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 $\|\mathbf{B}^\top\mathbf{B}\|_2 \leq \frac{n}{l(1-\epsilon/2)}$ and $\|\mathbf{B}\|_2 \leq \sqrt{\frac{n}{l(1-\epsilon/2)}}$. Thus, we conclude that $\mu_1(\mathbf{L}_C) \leq \sqrt{r\mu_0(\mathbf{U}_L)\mu_0(\mathbf{V}_L)}/\sqrt{1-\epsilon/2}$.

4.8 Proof of Theorem 9

We now give a proof of Thm. 9. 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 [29, Sec. 6] implies the same conclusions when columns and rows are sampled without replacement.

Our proof of Thm. 9 will require a strengthened version of the randomized ℓ_2 regression work of [19, Thm. 5]. The proof of Thm. 5 of [19] 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 [31], 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 12 (Sec. 3.4.3 of [31])**.** *Given matrices $\mathbf{A} \in \mathbb{R}^{m\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$ with $r \geq \max(\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{B}))$, an error tolerance $\epsilon \in (0,1]$, and a failure probability $\delta \in (0,1]$, define probabilities p_j satisfying*

$$p_j \geq \frac{\beta}{Z}\|\mathbf{A}_{(j)}\|\|\mathbf{B}_{(j)}\|, \quad Z = \sum_j \|\mathbf{A}_{(j)}\|\|\mathbf{B}_{(j)}\|, \quad \text{and} \quad \textstyle\sum_{j=1}^k p_j = 1 \tag{4.1}$$

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 $l \geq 48r\log(4r/(\beta\delta))/(\beta\epsilon^2)$ columns are selected in i.i.d. trials in which the j-th column is chosen with probability p_j, 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 $1-\delta$,

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

Using Lemma 12, we now establish a stronger version of Lemma 1 of [19]. 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

$$p_j \geq \frac{\beta}{r}\|(\mathbf{V}_L)_{(j)}\|^2 \quad \text{and} \quad \textstyle\sum_{j=1}^n p_j = 1. \tag{4.2}$$

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 $t \in \{1,\ldots,l\}$. 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 13. *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 $1-\delta$:*

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

Proof By Lemma 12, 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 [19]. □

Lemma 13 immediately yields improved sampling complexity for the randomized ℓ_2 regression of [19]:

Proposition 14. *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 [19] once Lemma 13 is substituted for Lemma 1 of [19]. □

A typical application of Prop. 14 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. (4.2). Here, we will take advantage of the slack term, β, allowed in the sampling probabilities of Eq. (4.2) 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. 9, 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. 14 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.

4.9 Proof of Corollary 10

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

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

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

we may apply Prop. 14 independently for each i to yield

$$\|\mathbf{M} - \mathbf{C}_i\mathbf{L}_{C_i}^+\mathbf{L}\|_F \leq (1+\epsilon)\|\mathbf{M} - \mathbf{M}\mathbf{L}^+\mathbf{L}\|_F \leq (1+\epsilon)\|\mathbf{M} - \mathbf{L}\|_F \tag{4.3}$$

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

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

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

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. (4.3) also fails to hold. The desired conclusion therefore must hold with probability at least $1 - 0.45^b = 1 - \delta$.

4.10 Proof of Corollary 11

With $c = 48000/\log(1/0.45)$ as in Cor. 10, we notice that for $m > 1$,

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

Therefore,

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

for all $m > 1$ and $\delta' \leq 0.8$. Hence, we may apply Thm. 9 and Cor. 10 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.

4.11 Proof of Theorem 5

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 8, while H holds with probability at least $1 - \delta/2$ by Thm. 9. Hence, by the union bound,

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

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

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

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

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

imply that $A(\mathbf{C})$ and $A(\mathbf{R})$ hold with probability at least $1 - \delta/(2n)$ and $1 - \delta/n$ respectively by Lemma 8, while K holds with probability at least $(1-\delta/2)(1-\delta)$ by Cor. 11. Hence, by the union bound,

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

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

$$\begin{aligned} \mathbf{P}(J) &\geq \mathbf{P}(B_C \cap B_R \mid K \cap A(\mathbf{C}) \cap A(\mathbf{R}))\mathbf{P}(K \cap A(\mathbf{C}) \cap A(\mathbf{R})) \\ &\geq (1 - \delta_C - \delta_R)(1-\delta)^2. \end{aligned}$$

4.12 Proof of Corollary 6

Cor. 6 is based on a new noisy MC theorem, which we prove in Sec. 4.14. A similar recovery guarantee is obtained by [11] under stronger assumptions.

Theorem 15. *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

$$\text{minimize}_{\mathbf{L}} \quad \|\mathbf{L}\|_* \quad \text{subject to} \quad \|\mathcal{P}_\Omega(\mathbf{M} - \mathbf{L})\|_F \leq \Delta \tag{4.4}$$

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. 5, 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. 15 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 $\mathbb{E}s_i = \frac{sl}{n}$ and hence satisfies Hoeffding's inequality for the hypergeometric distribution [29, Sec. 6]:

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

It therefore follows that

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

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. 5.

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. 5.

4.13 Proof of Corollary 7

Cor. 7 is based on the following theorem of Zhou et al. [84], reformulated for a generic rate parameter β, as described in [10, Section 3.1].

Theorem 16 (Thm. 2 of [84]). *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*

$$\text{minimize}_{\mathbf{L},\mathbf{S}} \quad \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 \quad \text{subject to} \quad \|\mathbf{M} - \mathbf{L} - \mathbf{S}\|_F \leq \Delta \tag{4.5}$$

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

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

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

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. 5, 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. 16 and the definitions of β' and β'',

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

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 $\mathbb{E}s_i = \frac{sl}{n}$ and hence satisfies Bernstein's inequality for the hypergeometric [29, Sec. 6]:

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

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

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

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 from Thm. 5.

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 $\mathbf{P}(B_C \mid A(\mathbf{C})) \leq (c_p+1)\bar{n}^{-\beta}$ and $\mathbf{P}(B_R \mid A(\mathbf{R})) \leq (c_p+1)\bar{n}^{-\beta}$. Thus, the DFC-Nys bound also follows from Thm. 5.

4.14 Proof of Theorem 15

In the spirit of [11], our proof will extend the noiseless analysis of [67] to the noisy matrix completion setting. As suggested in [26], 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 [67].

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 recovery guarantee.

Theorem 17. *Under the assumptions of Thm. 15, suppose that*

$$\frac{mn}{s}\left\|\mathcal{P}_T\mathcal{P}_\Omega\mathcal{P}_T - \frac{s}{mn}\mathcal{P}_T\right\|_2 \leq \frac{1}{2} \tag{4.6}$$

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

$$\|\mathcal{P}_T(\mathbf{Y}) - \mathbf{U}_{L_0}\mathbf{V}_{L_0}^\top\|_F \leq \sqrt{\frac{s}{32mn}} \quad \textit{and} \quad \|\mathcal{P}_{T^\perp}(\mathbf{Y})\|_2 < \frac{1}{2}. \tag{4.7}$$

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. (4.6),

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

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

$$\sqrt{\frac{s}{2mn}}\|\mathcal{P}_T(\mathbf{H})\|_F \leq \|\mathcal{P}_\Omega\mathcal{P}_T(\mathbf{H})\|_F \leq \|\mathcal{P}_\Omega\mathcal{P}_{T^\perp}(\mathbf{H})\|_F \leq \|\mathcal{P}_{T^\perp}(\mathbf{H})\|_F \leq \|\mathcal{P}_{T^\perp}(\mathbf{H})\|_*, \quad (4.8)$$

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. (4.7), and the final inequality follows from Eq. (4.8).

Since $\mathbf{L}_0$ is feasible for Eq. (4.4), $\|\mathbf{L}_0\|_* \geq \|\hat{\mathbf{L}}\|_*$, 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

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

Hence

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

for some constant c_e, by our assumption on s. □

To show that the sufficient conditions of Thm. 17 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 18. *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 19. *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 20. *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 [67, 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 [67, Thm. 3.2] also holds under sampling without replacement, as shown in [26].

Lemma 18 guarantees that Eq. (4.6) holds with high probability. To construct a matrix $\mathbf{Y} = \mathcal{P}_\Omega(\mathbf{Y})$ satisfying Eq. (4.7), we consider a sampling with batch replacement scheme recommended in [26] and developed in [14]. 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. (4.7) under this batch replacement scheme, as shown in the next lemma.

Lemma 21. *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. (4.7). 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 $\mathbf{P}(A(\tilde{\Omega}(s))) \leq \mathbf{P}(A(\Omega(s)))$.*

Proof As sketched in [26]

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

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 [67] to obtain $\mathbf{Y} = \mathcal{P}_{\tilde{\Omega}}(\mathbf{Y})$ satisfying Eq. (4.7). 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

$$\frac{mn}{q}\left\|\mathcal{P}_T\mathcal{P}_{\tilde{\Omega}_k}\mathcal{P}_T - \frac{q}{mn}\mathcal{P}_T\right\|_2 \leq \frac{1}{2} \tag{4.9}$$

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. (4.7).

The second condition of Eq. (4.7) follows from the assumptions

$$\left\|\mathbf{W}_{k-1} - \frac{mn}{q}\mathcal{P}_T\mathcal{P}_{\tilde{\Omega}_k}(\mathbf{W}_{k-1})\right\|_\infty \leq \frac{1}{2}\|\mathbf{W}_{k-1}\|_\infty \tag{4.10}$$

$$\left\|\left(\frac{mn}{q}\mathcal{P}_{\tilde{\Omega}_k} - \mathcal{I}\right)(\mathbf{W}_{k-1})\right\|_2 \leq \sqrt{\frac{8mn^2\beta\log(m+n)}{3q}}\|\mathbf{W}_{k-1}\|_\infty \tag{4.11}$$

for all k, since Eq. (4.10) 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 18 implies that Eq. (4.6) fails to hold with probability at most $2n^{2-2\beta}$. For each k, Eq. (4.9) fails to hold with probability at most $2n^{2-2\beta}$ by Lemma 18, Eq. (4.10) fails to hold with probability at most $2n^{2-2\beta}$ by Lemma 20, and Eq. (4.11) fails to hold with probability at most $(m+n)^{1-2\beta}$

by Lemma 19. Hence, by the union bound, the conclusion of Thm. 17 holds with probability at least

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

Chapter 5

Matrix Concentration Inequalities via the Method of Exchangeable Pairs

5.1 Introduction

In this chapter, we derive concentration inequalities for random matrices using Stein's method of exchangeable pairs [74]. Such inequalities are fundamental to the analysis of randomized procedures like matrix recovery from sparse random measurements [27, 67, 49], randomized matrix multiplication and factorization [19, 31], and convex relaxation of robust or chance-constrained optimization [59, 72, 15].

A primary difficulty in establishing matrix concentration is the lack of multiplicative commutativity: many classical proof techniques for scalar concentration rely on commuting elements and hence break down in the non-commutative matrix setting. In recent years, authors have begun to surmount this difficulty [2, 60, 79] by appealing to deep results from matrix analysis like the Golden-Thompson inequality [5, Section IX.3] or Lieb's concave trace inequality [42, Theorem 6]. Here we take a fundamentally different approach, building upon the work of Chatterjee [13], who demonstrated how the method of exchangeable pairs could be used to derive concentration inequalities for scalar random variables. Our analysis will extend to both independent and dependent sums of random matrices and to more general matrix functions satisfying a self-bounding property.

In the sequel, we describe the main results of our exchangeable pairs analysis. We present exponential tail inequalities for Hermitian matrices in Section 5.2, showing application to sums of random matrices and to more general self-bounding matrix functions. In Section 5.2, we present a complementary set of Hermitian moment inequalities and demonstrate their use in deriving tail inequalities. We extend our results to non-Hermitian matrices in Section 5.2 and conclude with proofs of all results in Section 5.3.

Notation Throughout, $\mathbb{H}_d$ denotes the set of Hermitian matrices in $\mathbb{C}^{d\times d}$. That is,

$$\mathbb{H}_d \triangleq \{\boldsymbol{A} \in \mathbb{C}^{d\times d} : \boldsymbol{A} = \boldsymbol{A}^*\}$$

where $\boldsymbol{A}^*$ is the conjugate transpose of $\boldsymbol{A}$. The Hermitian component of a generic square matrix $\boldsymbol{B} \in \mathbb{C}^{d\times d}$ is given by $\mathrm{Re}[\boldsymbol{B}] \triangleq \frac{1}{2}(\boldsymbol{B} + \boldsymbol{B}^*)$. Further, $\boldsymbol{I}$ denotes an identity matrix, $\boldsymbol{0}$ denotes a matrix of all zeros, $\mathrm{Tr}[\cdot]$ denotes the trace of a given matrix, and $\|\cdot\|$ denotes the spectral norm, i.e., the largest singular value of a given matrix. For $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_d$, $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ are the maximum and minimum eigenvalues of $\boldsymbol{A}$ respectively, and $\boldsymbol{A} \preceq \boldsymbol{H}$ or $\boldsymbol{H} \succeq \boldsymbol{A}$ signifies that $\boldsymbol{H} - \boldsymbol{A}$ is positive semidefinite. Given any function $h : \mathbb{R} \to \mathbb{R}$, we define a lifted function on Hermitian matrices via the eigenvalue decomposition:

$$h(\boldsymbol{A}) \triangleq \boldsymbol{Q}\begin{bmatrix} h(\lambda_1) & & \\ & \ddots & \\ & & h(\lambda_d) \end{bmatrix}\boldsymbol{Q}^* \quad \text{where} \quad \boldsymbol{A} = \boldsymbol{Q}\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}\boldsymbol{Q}^*$$

for $(\lambda_1, \ldots, \lambda_d)$ the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{Q}$ the matrix of associated eigenvectors.

5.2 Matrix concentration inequalities

Exponential tail inequalities

Our first result bounds the trace of the moment-generating function of a random matrix using Stein's method of exchangeable pairs. Combined with a matrix analogue of the Laplace transform method [2, 60],

$$\mathbb{P}(\lambda_{\max}(\boldsymbol{Y}) \geq t) \leq \inf_{\theta>0}\big\{e^{-\theta t}\,\mathrm{Tr}\big[\mathbb{E}\big[e^{\theta \boldsymbol{Y}}\big]\big]\big\},$$

this yields an exponential tail inequality for the maximum eigenvalue of a matrix.

Theorem 22. *Let $\mathcal{X}$ be a separable metric space, and suppose (X, X') is an exchangeable pair of $\mathcal{X}$-valued random variables. Suppose $\boldsymbol{f} : \mathcal{X} \to \mathbb{H}_d$ and $\boldsymbol{F} : \mathcal{X} \times \mathcal{X} \to \mathbb{H}_d$ are square-integrable functions such that for some non-decreasing $g : \mathbb{R} \to \mathbb{R}$,*

$$\boldsymbol{F}(X, X') = g(\boldsymbol{f}(X)) - g(\boldsymbol{f}(X')) \text{ a.s.}, \quad \mathbb{E}[\boldsymbol{F}(X, X') \mid X] = \boldsymbol{f}(X) \text{ a.s.},$$

and $\mathbb{E}\big[\|e^{\theta \boldsymbol{f}(X)}\boldsymbol{F}(X, X')\|\big] < \infty$. Let

$$\boldsymbol{\Delta}(X) \triangleq \frac{1}{2}\,\mathrm{Re}\,\mathbb{E}[(\boldsymbol{f}(X) - \boldsymbol{f}(X'))\boldsymbol{F}(X, X') \mid X].$$

If there exist real constants $b \geq 0$ and $c > 0$ such that $\boldsymbol{\Delta}(X) \preceq b\boldsymbol{f}(X) + c\boldsymbol{I}$ almost surely, then for any $0 \leq \theta < 1/b$,

$$\begin{aligned}\mathrm{Tr}\big[\mathbb{E}\big[e^{\theta \boldsymbol{f}(X)}\big]\big] &\leq d \cdot \exp\big(-\tfrac{c}{b^2}(b\theta + \log(1 - b\theta))\big)\\ &\leq d \cdot \exp\big(c\theta^2/(2 - 2b\theta)\big),\end{aligned}$$

and for any $t \geq 0$

$$\begin{aligned}\mathbb{P}(\lambda_{\max}(\boldsymbol{f}(X)) \geq t) &\leq d \cdot \exp\big(-\tfrac{t}{b} + \tfrac{c}{b^2}\log(1 + \tfrac{bt}{c})\big)\\ &\leq d \cdot \exp\big(-t^2/(2c + 2bt)\big).\end{aligned}$$

Remark Theorem 22 also yields a tail inequality for the minimum eigenvalue of a matrix, due to the identity

$$\lambda_{\min}(\boldsymbol{f}(X)) = -\lambda_{\max}(-\boldsymbol{f}(X)).$$

Comparable inequalities are obtained for intermediate eigenvalues when Theorem 22 is combined with the minimax Laplace transform method of Gittens and Tropp [24].

When applied to sums of independent matrices, Theorem 22 delivers tail bounds reminiscent of the classical inequalities due to Bernstein [4].

Theorem 23 (Hermitian Bernstein)**.** *Let $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n \in \mathbb{H}_d$ be independent random matrices satisfying*

$$\mathbb{E}[\boldsymbol{Y}_k] = \mathbf{0} \quad \text{and} \quad \boldsymbol{Y}_k^2 \preceq r\boldsymbol{Y}_k + \boldsymbol{A}_k^2 \quad \text{a.s.}, \quad \forall k \in \{1, \dots, n\},$$

for fixed $\boldsymbol{A}_k \in \mathbb{H}_d$ and $r \geq 0$, and define $\sigma^2 \triangleq \|\sum_{k=1}^n \boldsymbol{A}_k^2 + \mathbb{E}[\boldsymbol{Y}_k^2]\|$. Then, for all $t \geq 0$,

$$\begin{aligned}\mathbb{P}(\lambda_{\max}(\textstyle\sum_{k=1}^n \boldsymbol{Y}_k) \geq t) &\leq d \cdot \exp\left(\frac{-t^2}{\sigma^2 + rt}\right) \\ &\leq \begin{cases} d \cdot \exp(-t^2/\sigma^2) & \text{for } r = 0 \\ d \cdot \exp(-t^2/2\sigma^2) & \text{for } r > 0,\ t \leq \sigma^2/r \\ d \cdot \exp(-t/2r) & \text{for } r > 0,\ t \geq \sigma^2/r. \end{cases}\end{aligned}$$

An immediate consequence of Theorem 23 is a natural generalization of Hoeffding's inequality [29] to sums of bounded, independent random matrices. The following bound recovers the classical, scalar Hoeffding inequality when $d = 1$ and improves upon the recent Hoeffding generalization of Tropp [79, Theorem 1.3] by a factor of 4 in the exponent.

Corollary 24 (Hermitian Hoeffding)**.** *Let $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n \in \mathbb{H}_d$ be independent random matrices satisfying*

$$\mathbb{E}[\boldsymbol{Y}_k] = \mathbf{0} \quad \text{and} \quad \boldsymbol{Y}_k^2 \preceq \boldsymbol{A}_k^2 \quad \text{a.s.}, \quad \forall k \in \{1, \dots, n\},$$

and let $\sigma^2 \triangleq \|\sum_{k=1}^n \boldsymbol{A}_k^2\|$. Then, for all $t \geq 0$,

$$\mathbb{P}(\lambda_{\max}(\textstyle\sum_{k=1}^n \boldsymbol{Y}_k) \geq t) \leq d e^{-t^2/2\sigma^2}.$$

Remark Theorem 23 and Corollary 24 hold more generally for sums of dependent matrices satisfying a martingale difference-type property:

$$\mathbb{E}[\boldsymbol{Y}_k \mid \boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{k-1}, \boldsymbol{Y}_{k+1}, \dots, \boldsymbol{Y}_n] = \mathbf{0} \quad \text{a.s.}, \quad \forall k \in \{1, \dots, n\}.$$

The utility of Theorem 22 is by no means limited to sums of independent random matrices. Indeed, comparable concentration inequalities are available for all matrix functions satisfying

a certain self-bounding property, even when the underlying random elements are dependent. Self-bounding functions were introduced in [7] to establish concentration for scalar functions of independent random variables. Our next theorem extends these concentration results to the dependent, matrix-variate setting.

Theorem 25 (Self-bounding Hermitian Functions)**.** *For a separable metric space $\mathcal{X}$, let $X = (X_1, \dots, X_n)$ be a vector of $\mathcal{X}$-valued random variables. For each $x \in \mathcal{X}^n$, define $x_{\backslash k} \triangleq (x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)$ for each $k \in \{1, \dots, n\}$, and let $\boldsymbol{H} : \mathcal{X}^n \to \mathbb{H}_d$ be a square-integrable function satisfying*

$$\sum_{k=1}^n \mathbb{E}\big[\boldsymbol{H}(x_1, \dots, X_k, \dots, x_n) \mid x_{\backslash k}\big] = s\boldsymbol{H}(x) + (n-s)\mathbb{E}[\boldsymbol{H}(X)] \quad \text{and}$$
$$\frac{1}{n-s}\sum_{k=1}^n \mathbb{E}\big[(\boldsymbol{H}(x) - \boldsymbol{H}(x_1, \dots, X_k, \dots, x_n))^2 \mid x_{\backslash k}\big] \preceq r\boldsymbol{H}(x) + \boldsymbol{A}^2,$$

for fixed $\boldsymbol{A} \in \mathbb{H}_d$, real $s \neq n$, $r \geq 0$, and all $x \in \mathcal{X}^n$. If $\sigma^2 \triangleq \lambda_{\max}\big(\boldsymbol{A}^2 + r\mathbb{E}[\boldsymbol{H}(X)]\big)$, then, for all $t \geq 0$,

$$\mathbb{P}(\lambda_{\max}(\boldsymbol{H}(X) - \mathbb{E}[\boldsymbol{H}(X)]) \geq t) \leq d \cdot \exp\left(\frac{-t^2}{\sigma^2 + rt}\right)$$
$$\leq \begin{cases} d \cdot \exp(-t^2/\sigma^2) & \text{for } r = 0 \\ d \cdot \exp(-t^2/2\sigma^2) & \text{for } r > 0,\ t \leq \sigma^2/r \\ d \cdot \exp(-t/2r) & \text{for } r > 0,\ t \geq \sigma^2/r. \end{cases}$$

Notably, when $r = 0$, Theorem 25 delivers a dependent, Hermitian version of the bounded differences inequality due to McDiarmid [52].

To give a more exotic example of dependence treated by Theorem 22, we next develop a Bernstein-type inequality for a Hermitian analogue of Hoeffding's combinatorial statistics [28].

Theorem 26 (Combinatorial Hermitian Bernstein)**.** *Let $(\boldsymbol{A}_{ij})_{1 \leq i,j \leq n}$ be a fixed collection of matrices satisfying*

$$\boldsymbol{A}_{ij} \in \mathbb{H}_d \quad \text{and} \quad \boldsymbol{0} \preceq \boldsymbol{A}_{ij} \preceq \boldsymbol{I}, \quad \forall i, j \in \{1, \dots, n\},$$

and define

$$\mu \triangleq \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \boldsymbol{A}_{ij}\right).$$

If π is drawn uniformly from the set of all permutations over $\{1, \dots, n\}$, then, for all $t \geq 0$,

$$\mathbb{P}\big(\lambda_{\max}\big(\textstyle\sum_{i=1}^n \boldsymbol{A}_{i\pi(i)} - \mathbb{E}\big[\boldsymbol{A}_{i\pi(i)}\big]\big) \geq t\big) \leq d \cdot \exp\left(\frac{-t^2}{8\mu + 4t}\right)$$
$$\leq \begin{cases} d \cdot \exp(-t^2/16\mu) & \text{for } t \leq 2\mu \\ d \cdot \exp(-t/8) & \text{for } t \geq 2\mu. \end{cases}$$

Non-commutative moment inequalities

In addition to providing exponential tail inequalities for random matrices, Stein's method can be used to develop non-commutative moment inequalities, in the tradition of Lust-Piquard [44] and Pisier and Xu [65]:

Theorem 27. *Let $\mathcal{X}$ be a separable metric space, and suppose (X, X') is an exchangeable pair of $\mathcal{X}$-valued random variables. Suppose $\boldsymbol{f} : \mathcal{X} \to \mathbb{H}_d$ and $\boldsymbol{F} : \mathcal{X} \times \mathcal{X} \to \mathbb{H}_d$ are square-integrable functions such that*

$$\boldsymbol{F}(X, X') = g(\boldsymbol{f}(X)) - g(\boldsymbol{f}(X')) \quad \text{and} \quad \mathbb{E}[\boldsymbol{F}(X, X') \mid X] = \boldsymbol{f}(X) \text{ a.s.}$$

for some non-decreasing $g : \mathbb{R} \to \mathbb{R}$. Let

$$\boldsymbol{\Delta}(X) \triangleq \frac{1}{2} \operatorname{Re} \mathbb{E}[(\boldsymbol{f}(X) - \boldsymbol{f}(X'))\boldsymbol{F}(X, X') \mid X].$$

Then, for any positive integer p, we have

$$\mathbb{E}\big[\operatorname{Tr}\big[\boldsymbol{f}(X)^{2p}\big]\big] \leq (2p-1)^p \mathbb{E}[\operatorname{Tr}[\boldsymbol{\Delta}(X)^p]].$$

When combined with Markov's inequality, the moment inequalities of Theorem 27 give rise to polynomial tail probabilities for the maximum eigenvalue of $\boldsymbol{f}(X)$. That is, for all $t > 0$ and integers $p > 0$,

$$\begin{aligned}
\mathbb{P}(\lambda_{\max}(\boldsymbol{f}(X)) \geq t) &\leq \mathbb{E}\big[\lambda_{\max}(\boldsymbol{f}(X))^{2p}\big]/t^{2p} \\
&\leq \mathbb{E}\big[\lambda_{\max}\big(\boldsymbol{f}(X)^{2p}\big)\big]/t^{2p} \\
&\leq \mathbb{E}\big[\operatorname{Tr}\big[\boldsymbol{f}(X)^{2p}\big]\big]/t^{2p} \\
&\leq \frac{(2p-1)^p}{t^{2p}} \mathbb{E}[\operatorname{Tr}[\boldsymbol{\Delta}(X)^p]].
\end{aligned}$$

Moreover, control over all even moments lets us bound the trace of the moment generating function of $\boldsymbol{f}(X)$. To see this, note that $e^{\boldsymbol{A}} \prec e^{\boldsymbol{A}} + e^{-\boldsymbol{A}} = 2\sum_{p=0}^{\infty} \boldsymbol{A}^{2p}/(2p)!$ for all $\boldsymbol{A} \in \mathbb{H}_d$.[1] Thus,

$$\begin{aligned}
\operatorname{Tr}\big[\mathbb{E}\big[e^{\theta \boldsymbol{f}(X)}\big]\big] &< 2\textstyle\sum_{p=0}^{\infty} \theta^{2p} \mathbb{E}\big[\operatorname{Tr}\big[\boldsymbol{f}(X)^{2p}\big]\big]/(2p)! \\
&\leq 2\textstyle\sum_{p=0}^{\infty} \theta^{2p} (2p-1)^p \mathbb{E}[\operatorname{Tr}[\boldsymbol{\Delta}(X)^p]]/(2p)! \\
&\leq 2\textstyle\sum_{p=0}^{\infty} \theta^{2p} e^p \mathbb{E}[\operatorname{Tr}[\boldsymbol{\Delta}(X)^p]]/(p! 2^p) \\
&= 2 \operatorname{Tr}\Big[\mathbb{E}\Big[e^{\theta^2 \boldsymbol{\Delta}(X) e/2}\Big]\Big], \qquad (5.1)
\end{aligned}$$

where we have used the fact that $(2p-1)^p/(2p)! \leq e^p/(p! 2^p)$ for all $p > 0$. Combined with appropriate assumptions on the growth of $\boldsymbol{\Delta}(X)$, Eq. 5.1 gives rise to exponential tail probabilities, like those of Section 5.2, albeit with worse constants.

An example application of Theorem 27 is to sums of independent random matrices. In this case, we obtain a matrix version of the Burkholder-Davis-Gundy moment inequalities [8],

[1] The additional factor of two can be avoided when $\mathbb{E}[\operatorname{Tr}[\boldsymbol{f}(X)^p]] \leq 0$ for all odd positive integers p.

Theorem 28 (Hermitian Burkholder-Davis-Gundy)**.** *Let $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n \in \mathbb{H}_d$ be independent random matrices satisfying*

$$\mathbb{E}[\boldsymbol{Y}_k] = \mathbf{0}, \quad \forall k \in \{1, \dots, n\}.$$

Then, for any positive integer p, we have

$$\mathbb{E}\Big[\operatorname{Tr}\Big[\big(\textstyle\sum_{k=1}^n \boldsymbol{Y}_k\big)^{2p}\Big]\Big] \leq (2p-1)^p \mathbb{E}\big[\operatorname{Tr}\big[\big(\textstyle\sum_{k=1}^n \boldsymbol{Y}_k^2\big)^p\big]\big].$$

Theorem 28 may in turn be used to generalize the classical Khintchine inequalities [36] to sums of fixed matrices with random scalings.

Corollary 29 (Hermitian Khintchine)**.** *Fix $\boldsymbol{A}_1, \dots, \boldsymbol{A}_n \in \mathbb{H}_d$, and let $\xi_1, \dots, \xi_n \in \mathbb{R}$ be independent random variables satisfying*

$$\mathbb{E}[\xi_k] = 0 \quad \textit{and} \quad \xi_k \in [-1, 1], \quad \forall k \in \{1, \dots, n\}.$$

Then, for any positive integer p, we have

$$\mathbb{E}\Big[\operatorname{Tr}\Big[\big(\textstyle\sum_{k=1}^n \xi_k \boldsymbol{A}_k\big)^{2p}\Big]\Big] \leq (2p-1)^p \operatorname{Tr}\big[\big(\textstyle\sum_{k=1}^n \boldsymbol{A}_k^2\big)^p\big].$$

Recently, such non-commutative Khintchine inequalities have been used to analyze convex relaxations of robust and chance-constrained optimization problems [72].

The conclusions of Theorem 27 apply equally to matrices constructed from dependent sequences. As an example, we give a Burkholder-Davis-Gundy-type bound for the moments of the Hermitian combinatorial sums introduced in Theorem 26.

Theorem 30 (Combinatorial Hermitian Burkholder-Davis-Gundy)**.** *Let $(\boldsymbol{A}_{ij})_{1 \leq i,j \leq n}$ be a fixed collection of matrices satisfying*

$$\boldsymbol{A}_{ij} \in \mathbb{H}_d \quad \textit{and} \quad \mathbf{0} \preceq \boldsymbol{A}_{ij} \preceq \boldsymbol{I}, \quad \forall i, j \in \{1, \dots, n\}.$$

If π is drawn uniformly from the set of all permutations over $\{1, \dots, n\}$, and

$$\boldsymbol{\Delta} \triangleq \frac{1}{4n} \textstyle\sum_{i=1}^n \sum_{j=1}^n \boldsymbol{A}_{i\pi(i)}^2 + \boldsymbol{A}_{j\pi(j)}^2 - \boldsymbol{A}_{i\pi(j)}^2 - \boldsymbol{A}_{j\pi(i)}^2,$$

then, for any positive integer p, we have

$$\mathbb{E}\Big[\operatorname{Tr}\Big[\big(\textstyle\sum_{i=1}^n \boldsymbol{A}_{i\pi(i)} - \mathbb{E}\big[\boldsymbol{A}_{i\pi(i)}\big]\big)^{2p}\Big]\Big] \leq (2p-1)^p \mathbb{E}[\operatorname{Tr}[\boldsymbol{\Delta}^p]].$$

Extension to non-Hermitian matrices

We extend our results to a generic non-Hermitian matrix $\boldsymbol{B} \in \mathbb{C}^{d_1 \times d_2}$ by drawing upon a technique from operator theory known as self-adjoint dilation [62]:

$$\mathscr{D}(\boldsymbol{B}) \triangleq \begin{bmatrix} \mathbf{0} & \boldsymbol{B} \\ \boldsymbol{B}^* & \mathbf{0} \end{bmatrix}.$$

By construction, $\mathscr{D}(\boldsymbol{B})$ is Hermitian, and, moreover, $\lambda_{\max}(\mathscr{D}(\boldsymbol{B})) = \|\boldsymbol{B}\|$. Hence the following non-Hermitian variants of Theorem 22 and Theorem 27 also apply.

Corollary 31. *Under the conditions of Theorem 22, if $\boldsymbol{f}(X) = \mathscr{D}(\boldsymbol{h}(X))$ a.s. for $\boldsymbol{h} : \mathcal{X} \to \mathbb{C}^{d_1 \times d_2}$, then for all $t \geq 0$*

$$\begin{aligned} \mathbb{P}(\|\boldsymbol{h}(X)\| \geq t) &\leq (d_1 + d_2) \exp\big(-\tfrac{t}{b} + \tfrac{c}{b^2} \log(1 + \tfrac{bt}{c})\big) \\ &\leq (d_1 + d_2) \exp\big(-t^2/(2c + 2bt)\big). \end{aligned}$$

Corollary 32. *Under the conditions of Theorem 27, if $\boldsymbol{f}(X) = \mathscr{D}(\boldsymbol{h}(X))$ a.s. for $\boldsymbol{h} : \mathcal{X} \to \mathbb{C}^{d_1 \times d_2}$, then, for any positive integer p, we have*

$$\mathbb{E}[\mathrm{Tr}[(\boldsymbol{h}(X)\boldsymbol{h}(X)^*)^p]] \leq \frac{(2p-1)^p}{2} \mathbb{E}[\mathrm{Tr}[\boldsymbol{\Delta}(X)^p]].$$

5.3 Proofs via Stein's Method

Proof of Theorem 22

Proof Our proof extends that of [13, Theorem 1.5], which establishes analogous results for real-valued f. We begin with a lemma:

Lemma 33. *Under the conditions of Theorem 22, suppose that $\boldsymbol{h} : \mathcal{X} \to \mathbb{H}_d$ is a measurable map satisfying $\mathbb{E}[\|\boldsymbol{h}(X)\boldsymbol{F}(X, X')\|] < \infty$. Then*

$$\mathbb{E}[\boldsymbol{h}(X)\boldsymbol{f}(X)] = \frac{1}{2} \mathbb{E}[(\boldsymbol{h}(X) - \boldsymbol{h}(X'))\boldsymbol{F}(X, X')]. \tag{5.2}$$

Proof First note that $\boldsymbol{F}$ is antisymmetric:

$$\boldsymbol{F}(X, X') = g(\boldsymbol{f}(X)) - g(\boldsymbol{f}(X')) = -g(\boldsymbol{f}(X')) - g(\boldsymbol{f}(X)) = -\boldsymbol{F}(X', X).$$

Further, $\mathbb{E}[\boldsymbol{h}(X)\boldsymbol{f}(X)] = \mathbb{E}[\boldsymbol{h}(X)\mathbb{E}[\boldsymbol{F}(X, X') \mid X]] = \mathbb{E}[\boldsymbol{h}(X)\boldsymbol{F}(X, X')]$. Since X and X' are exchangeable and $\boldsymbol{F}$ is antisymmetric, it follows that

$$\mathbb{E}[\boldsymbol{h}(X)\boldsymbol{F}(X, X')] = \mathbb{E}[\boldsymbol{h}(X')\boldsymbol{F}(X', X)] = -\mathbb{E}[\boldsymbol{h}(X')\boldsymbol{F}(X, X')].$$

Hence,

$$\mathbb{E}[\boldsymbol{h}(X)\boldsymbol{f}(X)] = \mathbb{E}[\boldsymbol{h}(X)\boldsymbol{F}(X,X')] = \frac{1}{2}\mathbb{E}[(\boldsymbol{h}(X) - \boldsymbol{h}(X'))\boldsymbol{F}(X,X')].$$

□

We next let $\boldsymbol{m}(\theta) \triangleq \mathbb{E}\big[e^{\theta \boldsymbol{f}(X)}\big]$, the moment generating function of $\boldsymbol{f}(X)$, for all $\theta \in \mathbb{R}$ and consider its derivative, $\boldsymbol{m}'$. We are free to take the derivative inside of the expectation, due to our assumption that $\mathbb{E}\big[\|e^{\theta \boldsymbol{f}(X)}\boldsymbol{F}(X,X')\|\big] < \infty$ for all θ. Hence, Lemma 33 implies that

$$\begin{aligned}\boldsymbol{m}'(\theta) &= \mathbb{E}\big[e^{\theta \boldsymbol{f}(X)}\boldsymbol{f}(X)\big] = \frac{1}{2}\mathbb{E}\Big[(e^{\theta \boldsymbol{f}(X)} - e^{\theta \boldsymbol{f}(X')})\boldsymbol{F}(X,X')\Big] \\ &= \frac{1}{2}\mathbb{E}\Big[(e^{\theta \boldsymbol{f}(X)} - e^{\theta \boldsymbol{f}(X')})(g(\boldsymbol{f}(X)) - g(\boldsymbol{f}(X')))\Big].\end{aligned}$$

We will bound the trace of $\boldsymbol{m}'(\theta)$ using the following lemma:

Lemma 34. *If $g : \mathbb{R} \to \mathbb{R}$ is non-decreasing, $h : \mathbb{R} \to \mathbb{R}$ is differentiable, and $x \mapsto |h'(x)|$ is convex, then*

$$\begin{gathered}\operatorname{Tr}[(h(\boldsymbol{A}) - h(\boldsymbol{H}))(g(\boldsymbol{A}) - g(\boldsymbol{H}))] \leq \\ \frac{1}{2}\operatorname{Tr}[(|h'(\boldsymbol{A})| + |h'(\boldsymbol{H})|)\operatorname{Re}[(\boldsymbol{A} - \boldsymbol{H})(g(\boldsymbol{A}) - g(\boldsymbol{H}))]]\end{gathered}$$

for all $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_d$.

Proof Since g is non-decreasing, $(x-y)(g(x)-g(y)) \geq 0$ for all $x, y \in \mathbb{R}$. The fundamental theorem of calculus and the convexity of h' moreover imply that

$$\begin{aligned}(h(x) - h(y))(g(x) - g(y)) &= (x-y)(g(x) - g(y))\int_0^1 h'(tx + (1-t)y)dt \\ &\leq (x-y)(g(x) - g(y))\int_0^1 |h'(tx + (1-t)y)|dt \\ &\leq (x-y)(g(x) - g(y))\int_0^1 (t|h'(x)| + (1-t)|h'(y)|)dt \\ &= \frac{1}{2}(|h'(x)| + |h'(y)|)(x-y)(g(x) - g(y)) \qquad (5.3)\end{aligned}$$

for all $x, y \in \mathbb{R}$. The following proposition (see [64, Proposition 3] for a concise proof) allows us to establish a Hermitian analogue of Eq. 5.3:

Proposition 35. *If f_k and g_k are functions $\mathbb{R} \to \mathbb{R}$ such that for some $c_k \in \mathbb{R}$,*

$$\textstyle\sum_k c_k f_k(x) g_k(y) \geq 0$$

for every $x, y \in S \subseteq \mathbb{R}$, then for all $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_d$ having all eigenvalues in S

$$\textstyle\sum_k c_k \operatorname{Tr}[f_k(\boldsymbol{A})g_k(\boldsymbol{H})] \geq 0.$$

The inequality of Eq. 5.3 can be manipulated into the form $\sum_k c_k f_k(x) g_k(y) \geq 0$ for all $x, y \in \mathbb{R}$ as

$$\begin{aligned} 0 \leq \frac{1}{2}(&|h'(x)|xg(x) - g(x)|h'(x)|y - |h'(x)|xg(y) + |h'(x)|yg(y) \\ &+ xg(x)|h'(y)| - g(x)|h'(y)|y - xg(y)|h'(y)| + |h'(y)|yg(y)) \\ &- h(x)g(x) + h(x)g(y) + g(x)h(y) - h(y)g(y). \end{aligned}$$

Hence, for all $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_d$, Proposition 35 implies that

$$\begin{aligned} 0 \leq &\frac{1}{2}\operatorname{Tr}[|h'(\boldsymbol{A})|\boldsymbol{A}g(\boldsymbol{A}) - g(\boldsymbol{A})|h'(\boldsymbol{A})|\boldsymbol{H} - |h'(\boldsymbol{A})|\boldsymbol{A}g(\boldsymbol{H}) + |h'(\boldsymbol{A})|\boldsymbol{H}g(\boldsymbol{H}) \\ &+ \boldsymbol{A}g(\boldsymbol{A})|h'(\boldsymbol{H})| - g(\boldsymbol{A})|h'(\boldsymbol{H})|\boldsymbol{H} - \boldsymbol{A}g(\boldsymbol{H})|h'(\boldsymbol{H})| + |h'(\boldsymbol{H})|\boldsymbol{H}g(\boldsymbol{H})] \\ &- \operatorname{Tr}[h(\boldsymbol{A})g(\boldsymbol{A}) - h(\boldsymbol{A})g(\boldsymbol{H}) - g(\boldsymbol{A})h(\boldsymbol{H}) + h(\boldsymbol{H})g(\boldsymbol{H})] \\ = &\frac{1}{2}\operatorname{Tr}[|h'(\boldsymbol{A})|\boldsymbol{A}g(\boldsymbol{A}) - |h'(\boldsymbol{A})|\boldsymbol{H}g(\boldsymbol{A}) - |h'(\boldsymbol{A})|\boldsymbol{A}g(\boldsymbol{H}) + |h'(\boldsymbol{A})|\boldsymbol{H}g(\boldsymbol{H}) \\ &+ |h'(\boldsymbol{H})|\boldsymbol{A}g(\boldsymbol{A}) - |h'(\boldsymbol{H})|\boldsymbol{H}g(\boldsymbol{A}) - |h'(\boldsymbol{H})|\boldsymbol{A}g(\boldsymbol{H}) + |h'(\boldsymbol{H})|\boldsymbol{H}g(\boldsymbol{H})] \\ &- \operatorname{Tr}[h(\boldsymbol{A})g(\boldsymbol{A}) - h(\boldsymbol{A})g(\boldsymbol{H}) - g(\boldsymbol{A})h(\boldsymbol{H}) + h(\boldsymbol{H})g(\boldsymbol{H})] \\ = &\frac{1}{2}\operatorname{Tr}[(|h'(\boldsymbol{A})| + |h'(\boldsymbol{H})|)(\boldsymbol{A} - \boldsymbol{H})(g(\boldsymbol{A}) - g(\boldsymbol{H}))] \\ &- \operatorname{Tr}[(h(\boldsymbol{A}) - h(\boldsymbol{H}))(g(\boldsymbol{A}) - g(\boldsymbol{H}))]. \end{aligned}$$

where the first equality follows from the cyclic property of the trace. An identical argument yields

$$\begin{aligned} \operatorname{Tr}[(h(\boldsymbol{A}) - h(\boldsymbol{H}))(g(\boldsymbol{A}) - g(\boldsymbol{H}))] \leq \\ \frac{1}{2}\operatorname{Tr}[(|h'(\boldsymbol{A})| + |h'(\boldsymbol{H})|)(g(\boldsymbol{A}) - g(\boldsymbol{H}))(\boldsymbol{A} - \boldsymbol{H})]. \end{aligned}$$

Since $\boldsymbol{A}$ and $\boldsymbol{H}$ are Hermitian,

$$\operatorname{Re}[(\boldsymbol{A} - \boldsymbol{H})(g(\boldsymbol{A}) - g(\boldsymbol{H}))] = \frac{1}{2}((\boldsymbol{A} - \boldsymbol{H})(g(\boldsymbol{A}) - g(\boldsymbol{H})) + (g(\boldsymbol{A}) - g(\boldsymbol{H}))(\boldsymbol{A} - \boldsymbol{H})),$$

and the desired result follows from the two preceding inequalities. □

For each $\theta \in \mathbb{R}$, $x \mapsto e^{\theta x}$ has derivative $x \mapsto \theta e^{\theta x}$, and $x \mapsto |\theta e^{\theta x}|$ is convex on $\mathbb{R}$, so Lemma 34 implies

$$\begin{aligned} \operatorname{Tr}\big[(e^{\theta \boldsymbol{A}} - e^{\theta \boldsymbol{H}})(g(\boldsymbol{A}) - g(\boldsymbol{H}))\big] \leq \\ \frac{|\theta|}{2}\operatorname{Tr}\big[(e^{\theta \boldsymbol{A}} + e^{\theta \boldsymbol{H}})\operatorname{Re}[(\boldsymbol{A} - \boldsymbol{H})(g(\boldsymbol{A}) - g(\boldsymbol{H}))]\big] \end{aligned}$$

for all $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_d$. Combining this result with the exchangeability of X and X', we obtain

$$\begin{aligned}
\operatorname{Tr}[\boldsymbol{m}'(\theta)] &= \frac{1}{2}\mathbb{E}\Big[\operatorname{Tr}\Big[(e^{\theta \boldsymbol{f}(X)} - e^{\theta \boldsymbol{f}(X')})\boldsymbol{F}(X, X')\Big]\Big] \\
&\le \frac{1}{2}\mathbb{E}\bigg[\frac{|\theta|}{2}\operatorname{Tr}\Big[(e^{\theta \boldsymbol{f}(X)} + e^{\theta \boldsymbol{f}(X')})\operatorname{Re}[(\boldsymbol{f}(X) - \boldsymbol{f}(X'))\boldsymbol{F}(X, X')]\Big]\bigg] \\
&= \frac{|\theta|}{2}\operatorname{Tr}\bigg[\mathbb{E}\bigg[e^{\theta \boldsymbol{f}(X)}\frac{1}{2}\operatorname{Re}\mathbb{E}[(\boldsymbol{f}(X) - \boldsymbol{f}(X'))\boldsymbol{F}(X, X') \mid X] + \\
&\qquad\qquad e^{\theta \boldsymbol{f}(X')}\frac{1}{2}\operatorname{Re}\mathbb{E}[(\boldsymbol{f}(X') - \boldsymbol{f}(X))\boldsymbol{F}(X', X) \mid X']\bigg]\bigg] \\
&= \frac{|\theta|}{2}\operatorname{Tr}\Big[\mathbb{E}\Big[e^{\theta \boldsymbol{f}(X)}\boldsymbol{\Delta}(X) + e^{\theta \boldsymbol{f}(X')}\boldsymbol{\Delta}(X')\Big]\Big] \\
&= |\theta|\mathbb{E}\big[\operatorname{Tr}\big[e^{\theta \boldsymbol{f}(X)}\boldsymbol{\Delta}(X)\big]\big].
\end{aligned}$$

Introducing our bound on $\boldsymbol{\Delta}(X)$ requires the following proposition.

Proposition 36. *If* $\mathbf{0} \preceq \boldsymbol{A}$ *and* $\boldsymbol{H} \preceq \boldsymbol{W}$*, then* $\operatorname{Tr}[\boldsymbol{A}\boldsymbol{H}] \le \operatorname{Tr}[\boldsymbol{A}\boldsymbol{W}]$.

Proof Since $\mathbf{0} \preceq \boldsymbol{W} - \boldsymbol{H}$ and $xy \ge 0$ for all $x, y \ge 0$, Proposition 35 implies that $\operatorname{Tr}[\boldsymbol{A}(\boldsymbol{W} - \boldsymbol{H})] \ge 0$. □

Since $0 \preceq e^{\theta \boldsymbol{f}(X)}$, Proposition 36 and our assumed bound on $\boldsymbol{\Delta}(X)$ now give

$$\begin{aligned}
\operatorname{Tr}[\boldsymbol{m}'(\theta)] &\le |\theta|\mathbb{E}\big[\operatorname{Tr}\big[e^{\theta \boldsymbol{f}(X)}(b\boldsymbol{f}(X) + cI)\big]\big] \\
&= b|\theta|\operatorname{Tr}[\boldsymbol{m}'(\theta)] + c|\theta|\operatorname{Tr}[\boldsymbol{m}(\theta)]
\end{aligned}$$

which, for all $0 \le \theta < 1/b$, may be rewritten as

$$\frac{d}{d\theta}\log\operatorname{Tr}[\boldsymbol{m}(\theta)] \le \frac{c\theta}{1 - b\theta}.$$

Integrating and noting that $\operatorname{Tr}[\boldsymbol{m}(0)] = d$, we obtain

$$\log\operatorname{Tr}[\boldsymbol{m}(\theta)] - \log d \le \int_0^\theta \frac{cu}{1 - bu}du = -\frac{c}{b^2}(b\theta + \log(1 - b\theta)),$$

which evaluates to $c\theta^2/2$ when $b = 0$. A second fruitful bound is obtained by observing that

$$\int_0^\theta \frac{cu}{1 - bu}du \le \int_0^\theta \frac{cu}{1 - b\theta}du \le \frac{c\theta^2}{2 - 2b\theta}.$$

To derive the desired concentration inequalities, note that for any $0 \le \theta < 1/b$ and all $t \ge 0$

$$\begin{aligned}
\mathbb{P}(\lambda_{\max}(\boldsymbol{f}(X)) \ge t) &\le \exp(-\theta t + \log \mathbb{E}[\exp(\theta \lambda_{\max}(\boldsymbol{f}(X)))]) \\
&\le \exp(-\theta t + \log \mathbb{E}[\lambda_{\max}(\exp(\theta \boldsymbol{f}(X)))]) \\
&\le \exp(-\theta t + \log \mathbb{E}[\operatorname{Tr}[\exp(\theta \boldsymbol{f}(X))]]) \\
&= \exp(-\theta t + \log \operatorname{Tr}[\boldsymbol{m}(\theta)]) \\
&\le d \cdot \exp\bigl(-\theta t - \tfrac{c}{b^2}(b\theta + \log(1 - b\theta))\bigr) \qquad (5.4) \\
&\le d \cdot \exp\bigl(-\theta t + c\theta^2/(2 - 2b\theta)\bigr), \qquad (5.5)
\end{aligned}$$

since $0 \preceq e^{\theta \boldsymbol{f}(X)}$ and $\lambda_{\max}(\boldsymbol{A}) \le \operatorname{Tr}[\boldsymbol{A}]$ for any $\boldsymbol{A} \succeq 0$. The advertised inequalities follow by letting $\theta = t/(c + bt) < 1/b$ in Eq. 5.4 and Eq. 5.5. □

Proof of Theorem 23

Proof We will prove a generalization of Theorem 23 for dependent $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n \in \mathbb{H}_d$ satisfying

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{Y}_k \mid \boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{k-1}, \boldsymbol{Y}_{k+1}, \dots, \boldsymbol{Y}_n] = \boldsymbol{0} \\
&\mathbb{E}\bigl[\boldsymbol{Y}_k^2 \mid \boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{k-1}, \boldsymbol{Y}_{k+1}, \dots, \boldsymbol{Y}_n\bigr] \preceq \boldsymbol{H}_k^2 \quad \text{a.s.}, \quad \forall k \in \{1, \dots, n\}
\end{aligned}$$

for deterministic $\boldsymbol{H}_k \in \mathbb{H}_d$ and $\sigma^2 \triangleq \|\sum_{k=1}^n \boldsymbol{A}_k^2 + \boldsymbol{H}_k^2\|$. The original statement for independent matrices will follow as a special case.

Let $\boldsymbol{X} \triangleq \sum_{k=1}^n \boldsymbol{Y}_k$ and $\boldsymbol{f}(\boldsymbol{X}) \triangleq \boldsymbol{X} - \mathbb{E}[\boldsymbol{X}] = \boldsymbol{X}$. For each k, define

$$\boldsymbol{Y}_{\setminus k} \triangleq (\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{k-1}, \boldsymbol{Y}_{k+1}, \dots, \boldsymbol{Y}_n),$$

and let $\boldsymbol{Y}_k'$ be drawn, independently of $\boldsymbol{Y}_k$, from the conditional distribution of $\boldsymbol{Y}_k$ given $\boldsymbol{Y}_{\setminus k}$. To create an exchangeable pair, we define

$$\boldsymbol{X}' \triangleq \boldsymbol{Y}_K' + \textstyle\sum_{k \ne K} \boldsymbol{Y}_k$$

where K is independent of $(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n, \boldsymbol{Y}_1', \dots, \boldsymbol{Y}_n')$ and distributed uniformly on $\{1, \dots, n\}$. Since $\boldsymbol{Y}_k'$ and $\boldsymbol{Y}_k$ are conditionally i.i.d. given $\boldsymbol{Y}_{\setminus k}$ for all k, it follows that $\boldsymbol{X}$ and $\boldsymbol{X}'$ are conditionally i.i.d. given K and $\boldsymbol{Y}_{\setminus K}$. Hence, $\boldsymbol{X}$ and $\boldsymbol{X}'$ are exchangeable.

Let $\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \triangleq n(\boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{f}(\boldsymbol{X}'))$, and note that

$$\begin{aligned}
\mathbb{E}[\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{X}] &= n\mathbb{E}[\boldsymbol{Y}_K - \boldsymbol{Y}_K' \mid \boldsymbol{X}] \\
&= \frac{n}{n} \textstyle\sum_{k=1}^n \boldsymbol{Y}_k - \mathbb{E}\bigl[\mathbb{E}\bigl[\boldsymbol{Y}_k' \mid \boldsymbol{Y}_{\setminus k}\bigr] \mid \boldsymbol{X}\bigr] = \boldsymbol{X}
\end{aligned}$$

as $\mathbb{E}\bigl[\boldsymbol{Y}_k' \mid \boldsymbol{Y}_{\setminus k}\bigr] = 0$. So, $\mathbb{E}[\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{X}] = \mathbb{E}[\mathbb{E}_n[\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}')] \mid \boldsymbol{X}] = \boldsymbol{f}(\boldsymbol{X})$, as desired.

Furthermore, our assumptions imply that

$$\begin{aligned}
\boldsymbol{\Delta}(\boldsymbol{X}) &= \frac{n}{2}\mathbb{E}\big[(\boldsymbol{X}-\boldsymbol{X}')^2 \mid \boldsymbol{X}\big] = \frac{1}{2}\textstyle\sum_{k=1}^n \mathbb{E}\big[(\boldsymbol{Y}_k - \boldsymbol{Y}'_k)^2 \mid \boldsymbol{X}\big] \\
&= \frac{1}{2}\textstyle\sum_{k=1}^n \mathbb{E}\big[\boldsymbol{Y}_k^2 \mid \boldsymbol{X}\big] + \mathbb{E}\Big[{\boldsymbol{Y}'_k}^2 \mid \boldsymbol{X}\Big] - \mathbb{E}[\boldsymbol{Y}_k\boldsymbol{Y}'_k \mid \boldsymbol{X}] - \mathbb{E}[\boldsymbol{Y}'_k\boldsymbol{Y}_k \mid \boldsymbol{X}] \\
&\preceq \frac{1}{2}\textstyle\sum_{k=1}^n \mathbb{E}\big[\boldsymbol{Y}_k^2 \mid \boldsymbol{X}\big] + \mathbb{E}\Big[{\boldsymbol{Y}'_k}^2 \mid \boldsymbol{X}\Big] - \mathbb{E}\big[\boldsymbol{Y}_k\mathbb{E}\big[\boldsymbol{Y}'_k \mid \boldsymbol{Y}_{\backslash k}\big] \mid \boldsymbol{X}\big] \\
&\qquad - \mathbb{E}\big[\mathbb{E}\big[\boldsymbol{Y}'_k \mid \boldsymbol{Y}_{\backslash k}\big]\boldsymbol{Y}_k \mid \boldsymbol{X}\big] \\
&\preceq \frac{1}{2}\textstyle\sum_{k=1}^n r\mathbb{E}[\boldsymbol{Y}_k \mid \boldsymbol{X}] + \frac{1}{2}\sum_{k=1}^n (\boldsymbol{A}_k^2 + \mathbb{E}\big[{\boldsymbol{Y}_k}^2 \mid \boldsymbol{Y}_{\backslash k}\big]) \\
&\preceq \frac{r}{2}\boldsymbol{f}(\boldsymbol{X}) + \frac{\sigma^2}{2}\boldsymbol{I}.
\end{aligned}$$

since $\boldsymbol{Y}_k$ is conditionally independent of $\boldsymbol{Y}'_k$ given $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{k-1}$. Hence, Theorem 22 applies with $b = r/2$ and $c = \sigma^2/2$, and we obtain

$$\mathbb{P}(\lambda_{\max}(\textstyle\sum_{k=1}^n \boldsymbol{Y}_k) \geq t) \leq d \cdot \exp\big(-t^2/(\sigma^2 + rt)\big).$$

□

Proof of Corollary 24

Proof By the triangle inequality and our boundedness assumption,

$$\|\textstyle\sum_{k=1}^n \boldsymbol{A}_k^2 + \mathbb{E}\big[\boldsymbol{Y}_k^2\big]\| \leq \|\sum_{k=1}^n \boldsymbol{A}_k^2\| + \|\sum_{k=1}^n \mathbb{E}\big[\boldsymbol{Y}_k^2\big]\| \leq 2\|\textstyle\sum_{k=1}^n \boldsymbol{A}_k^2\| = 2\sigma^2.$$

Thus, Theorem 23 implies

$$\mathbb{P}(\lambda_{\max}(\textstyle\sum_{k=1}^n \boldsymbol{Y}_k) \geq t) \leq d \cdot \exp\left(\frac{-t^2}{\|\sum_{k=1}^n \boldsymbol{A}_k^2 + \mathbb{E}\big[\boldsymbol{Y}_k^2\big]\|}\right) \leq de^{-t^2/2\sigma^2}.$$

□

Proof of Theorem 25

Proof Let $\boldsymbol{f}(X) \triangleq \boldsymbol{H}(X) - \mathbb{E}[\boldsymbol{H}(X)]$. To create an exchangeable pair, we independently choose a random coordinate K uniformly from $\{1, \dots, n\}$ and define

$$X' \triangleq (X_1, \dots, X_{K-1}, X'_K, X_{K+1}, \dots, X_n)$$

where X'_k is drawn, independently of X_k, from the conditional distribution of X_k given $X_{\backslash k}$. Since X'_k and X_k are conditionally i.i.d. given $X_{\backslash k}$ for all k, it follows that X and X' are conditionally i.i.d. given K and $X_{\backslash K}$. Hence, X and X' are exchangeable.

Now let $\boldsymbol{F}(X, X') \triangleq \frac{n}{n-s}(\boldsymbol{f}(X) - \boldsymbol{f}(X'))$, and note that, by our assumptions,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{F}(X, X') \mid X] &= \frac{n}{n-s}\mathbb{E}[\boldsymbol{H}(X) - \boldsymbol{H}(X_1, \dots, X'_K, \dots, X_n) \mid X] \\
&= \frac{n}{n-s}\boldsymbol{H}(X) - \frac{1}{n-s}\textstyle\sum_{k=1}^n \mathbb{E}[\boldsymbol{H}(X_1, \dots, X'_k, \dots, X_n) \mid X] \\
&= \frac{n}{n-s}\boldsymbol{H}(X) - \frac{1}{n-s}\textstyle\sum_{k=1}^n \mathbb{E}\big[\boldsymbol{H}(X) \mid X_{\setminus k}\big] \\
&= \frac{n}{n-s}\boldsymbol{H}(X) - \frac{s}{n-s}\boldsymbol{H}(X) - \mathbb{E}[\boldsymbol{H}(X)] \\
&= \boldsymbol{H}(X) - \mathbb{E}[\boldsymbol{H}(X)]
\end{aligned}$$

as desired.

Furthermore,

$$\begin{aligned}
\boldsymbol{\Delta}(X) &= \frac{n}{2(n-s)}\mathbb{E}\big[(\boldsymbol{H}(X_1, \dots, X_K, \dots, X_n) - \boldsymbol{H}(X_1, \dots, X'_K, \dots, X_n))^2 \mid X\big] \\
&= \frac{1}{2(n-s)}\textstyle\sum_{k=1}^n \mathbb{E}\big[(\boldsymbol{H}(X_1, \dots, X_k, \dots, X_n) - \boldsymbol{H}(X_1, \dots, X'_k, \dots, X_n))^2 \mid X\big] \\
&\preceq \frac{1}{2}(r\boldsymbol{H}(X) + \boldsymbol{A}^2) = \frac{1}{2}(r\boldsymbol{f}(X) + \boldsymbol{A}^2 + r\mathbb{E}[\boldsymbol{H}(X)]) \preceq \frac{1}{2}(r\boldsymbol{f}(X) + \sigma^2\boldsymbol{I})
\end{aligned}$$

Thus, we may apply Theorem 22 with $b = r/2$ and $c = \sigma^2/2$ to obtain

$$\mathbb{P}(\lambda_{\max}(\textstyle\sum_{k=1}^n \boldsymbol{Y}_k) \geq t) \leq d e^{-t^2/(\sigma^2 + rt)}.$$

□

Proof of Theorem 26

Proof Our argument extends that of [13, Proposition 1.1], which establishes a related result for scalar random variables. Let $\boldsymbol{X} \triangleq \sum_{i=1}^n \boldsymbol{A}_{i\pi(i)}$ and

$$\boldsymbol{f}(\boldsymbol{X}) \triangleq \boldsymbol{X} - \mathbb{E}[\boldsymbol{X}] = \boldsymbol{X} - \frac{1}{n}\textstyle\sum_{i=1}^n\sum_{j=1}^n \boldsymbol{A}_{ij}.$$

To create an exchangeable pair, we independently choose a pair of indices (I, J) uniformly from $\{1, \dots, n\}^2$ and define a new permutation π' as the composition of π with the transposition (I, J), i.e. $\pi' \triangleq \pi \circ (I, J)$. Since π and π' are exchangeable, so too are $\boldsymbol{X}$ and $\boldsymbol{X}'$ when

$$\boldsymbol{X}' \triangleq \textstyle\sum_{i=1}^n \boldsymbol{A}_{i\pi'(i)}.$$

Now let $\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \triangleq (n/2)(\boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{f}(\boldsymbol{X}'))$ and note that

$$\begin{aligned}
\mathbb{E}[\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \mid \pi] &= \frac{n}{2}\mathbb{E}\big[\boldsymbol{A}_{I\pi(I)} + \boldsymbol{A}_{J\pi(J)} - \boldsymbol{A}_{J\pi(I)} - \boldsymbol{A}_{I\pi(J)} \mid \pi\big] \\
&= \textstyle\sum_{i=1}^n \boldsymbol{A}_{i\pi(i)} - \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \boldsymbol{A}_{i\pi(j)} = \boldsymbol{f}(\boldsymbol{X}).
\end{aligned}$$

So, $\mathbb{E}[\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{X}] = \mathbb{E}[\mathbb{E}[\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \mid \pi] \mid \boldsymbol{X}] = \boldsymbol{f}(\boldsymbol{X})$, as desired.

Furthermore, our assumptions imply that

$$\begin{aligned}
&\frac{1}{2}\operatorname{Re}\mathbb{E}[(\boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{f}(\boldsymbol{X}'))\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \mid \pi] \\
&\quad = \frac{n}{4}\mathbb{E}\big[(\boldsymbol{X} - \boldsymbol{X}')^2 \mid \pi\big] \\
&\quad = \frac{n}{4}\mathbb{E}\big[(\boldsymbol{A}_{I\pi(I)} + \boldsymbol{A}_{J\pi(J)} - \boldsymbol{A}_{J\pi(I)} - \boldsymbol{A}_{I\pi(J)})^2 \mid \pi\big] \\
&\quad = \frac{1}{4n}\textstyle\sum_{i=1}^n\sum_{j=1}^n(\boldsymbol{A}_{i\pi(i)} + \boldsymbol{A}_{j\pi(j)} - \boldsymbol{A}_{j\pi(i)} - \boldsymbol{A}_{i\pi(j)})^2 \\
&\quad \preceq \frac{1}{2n}\textstyle\sum_{i=1}^n\sum_{j=1}^n(\boldsymbol{A}_{i\pi(i)} + \boldsymbol{A}_{j\pi(j)})^2 + (\boldsymbol{A}_{j\pi(i)} + \boldsymbol{A}_{i\pi(j)})^2 \\
&\quad \preceq \frac{1}{n}\textstyle\sum_{i=1}^n\sum_{j=1}^n(\boldsymbol{A}_{i\pi(i)} + \boldsymbol{A}_{j\pi(j)} + \boldsymbol{A}_{j\pi(i)} + \boldsymbol{A}_{i\pi(j)}) \\
&\quad = 2\boldsymbol{X} + 2\mathbb{E}[\boldsymbol{X}] = 2\boldsymbol{f}(\boldsymbol{X}) + 4\mathbb{E}[\boldsymbol{X}],
\end{aligned}$$

where the first inequality follows from the operator convexity of the matrix square:

$$\left(\frac{\boldsymbol{H} + \boldsymbol{W}}{2}\right)^2 \preceq \frac{\boldsymbol{H}^2}{2} + \frac{\boldsymbol{W}^2}{2} \quad \text{for all} \quad \boldsymbol{H}, \boldsymbol{W} \in \mathbb{H}_d \quad \text{since} \quad \boldsymbol{0} \preceq \left(\frac{\boldsymbol{H}}{2} - \frac{\boldsymbol{W}}{2}\right)^2,$$

and the second inequality follows from $\boldsymbol{0} \preceq \boldsymbol{A}_{i\pi(i)} + \boldsymbol{A}_{j\pi(j)} \preceq 2\boldsymbol{I}$ and $\boldsymbol{0} \preceq \boldsymbol{A}_{j\pi(i)} + \boldsymbol{A}_{i\pi(j)} \preceq 2\boldsymbol{I}$. Therefore,

$$\begin{aligned}
\boldsymbol{\Delta}(\boldsymbol{X}) &= \mathbb{E}\left[\frac{1}{2}\operatorname{Re}\mathbb{E}[(\boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{f}(\boldsymbol{X}'))\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \mid \pi] \mid \boldsymbol{X}\right] \\
&\preceq 2\boldsymbol{f}(\boldsymbol{X}) + 4\lambda_{\max}(\mathbb{E}[\boldsymbol{X}])\boldsymbol{I},
\end{aligned}$$

and thus Theorem 22 applies with $b = 2$ and $c = 4\lambda_{\max}(\mathbb{E}[\boldsymbol{X}])$, and we obtain

$$\mathbb{P}\big(\lambda_{\max}\big(\textstyle\sum_{i=1}^n \boldsymbol{A}_{i\pi(i)} - \mathbb{E}\big[\boldsymbol{A}_{i\pi(i)}\big]\big) \geq t\big) \leq d \cdot \exp\big(-t^2/(8\lambda_{\max}(\mathbb{E}[\boldsymbol{X}]) + 4t)\big).$$

□

Proof of Theorem 27

Proof Our argument extends that of [13, Theorem 1.5], which establishes a related result for scalar random variables. Fix any integer $p > 0$ and notice that Lemma 33 implies

$$\mathbb{E}\big[\boldsymbol{f}(X)^{2p}\big] = \frac{1}{2}\mathbb{E}\big[(\boldsymbol{f}(X)^{2p-1} - \boldsymbol{f}(X')^{2p-1})\boldsymbol{F}(X, X')\big].$$

Further, $x \mapsto x^{2p-1}$ has nonnegative convex derivative $x \mapsto (2p-1)x^{2p-2}$ on $\mathbb{R}$, so Lemma 34 implies

$$\operatorname{Tr}\big[(\boldsymbol{A}^{2p-1} - \boldsymbol{H}^{2p-1})(g(\boldsymbol{A}) - g(\boldsymbol{H}))\big] \le \frac{2p-1}{2} \operatorname{Tr}\big[(\boldsymbol{A}^{2p-2} + \boldsymbol{H}^{2p-2}) \operatorname{Re}[(\boldsymbol{A} - \boldsymbol{H})(g(\boldsymbol{A}) - g(\boldsymbol{H}))]\big]$$

for all $\boldsymbol{A}, \boldsymbol{H} \in \mathbb{H}_d$.

Combining this result with the exchangeability of X and X', we obtain

$$\begin{aligned}
&\mathbb{E}\big[\operatorname{Tr}\big[\boldsymbol{f}(X)^{2p}\big]\big] \\
&= \frac{1}{2}\mathbb{E}\big[\operatorname{Tr}\big[(\boldsymbol{f}(X)^{2p-1} - \boldsymbol{f}(X')^{2p-1})\boldsymbol{F}(X, X')\big]\big]. \\
&\le \frac{1}{2}\mathbb{E}\left[\operatorname{Tr}\left[\frac{2p-1}{2}(\boldsymbol{f}(X)^{2p-2} + \boldsymbol{f}(X')^{2p-2}) \operatorname{Re}[(\boldsymbol{f}(X) - \boldsymbol{f}(X'))\boldsymbol{F}(X, X')]\right]\right] \\
&= \frac{2p-1}{2} \operatorname{Tr}\left[\mathbb{E}\left[\boldsymbol{f}(X)^{2p-2}\frac{1}{2} \operatorname{Re}\mathbb{E}[(\boldsymbol{f}(X) - \boldsymbol{f}(X'))\boldsymbol{F}(X, X') \mid X] + \right.\right. \\
&\qquad\qquad \left.\left. \boldsymbol{f}(X')^{2p-2}\frac{1}{2} \operatorname{Re}\mathbb{E}[(\boldsymbol{f}(X') - \boldsymbol{f}(X))\boldsymbol{F}(X')) \mid X']\right]\right] \\
&= \frac{2p-1}{2} \operatorname{Tr}\big[\mathbb{E}\big[\boldsymbol{f}(X)^{2p-2}\boldsymbol{\Delta}(X) + \boldsymbol{f}(X')^{2p-2}\boldsymbol{\Delta}(X')\big]\big] \\
&= (2p-1)\mathbb{E}\big[\operatorname{Tr}\big[\boldsymbol{f}(X)^{2p-2}\boldsymbol{\Delta}(X)\big]\big] \\
&\le (2p-1)\mathbb{E}\Big[\big(\operatorname{Tr}\big[(\boldsymbol{f}(X)^{2p-2})^{p/(p-1)}\big]\big)^{(p-1)/p}(\operatorname{Tr}[\boldsymbol{\Delta}(X)^p])^{1/p}\Big] \\
&= (2p-1)\mathbb{E}\big[(\operatorname{Tr}\big[\boldsymbol{f}(X)^{2p}\big])^{(p-1)/p}(\operatorname{Tr}[\boldsymbol{\Delta}(X)^p])^{1/p}\big] \\
&\le (2p-1)(\mathbb{E}\big[\operatorname{Tr}\big[\boldsymbol{f}(X)^{2p}\big]\big])^{(p-1)/p}(\mathbb{E}[\operatorname{Tr}[\boldsymbol{\Delta}(X)^p]])^{1/p},
\end{aligned}$$

where the penultimate inequality follows from Hölder's inequality for Schätten p-norms, and the final inequality is Hölder's inequality for real random variables. Hence

$$(\mathbb{E}\big[\operatorname{Tr}\big[\boldsymbol{f}(X)^{2p}\big]\big])^{1/p} \le (2p-1)(\mathbb{E}[\operatorname{Tr}[\boldsymbol{\Delta}(X)^p]])^{1/p}$$

and thus

$$\mathbb{E}\big[\operatorname{Tr}\big[\boldsymbol{f}(X)^{2p}\big]\big] \le (2p-1)^p\mathbb{E}[\operatorname{Tr}[\boldsymbol{\Delta}(X)^p]]$$

as desired. □

Proof of Theorem 28

Proof Fix any positive integer p, and, as in the proof of Theorem 23, let

$$\boldsymbol{X} \triangleq \sum_{k=1}^n \boldsymbol{Y}_k, \quad \boldsymbol{f}(\boldsymbol{X}) \triangleq \boldsymbol{X}, \quad \boldsymbol{X}' \triangleq \boldsymbol{Y}'_K + \sum_{k \ne K} \boldsymbol{Y}_k,$$

and

$$\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \triangleq n(\boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{f}(\boldsymbol{X}'))$$

where K is chosen independently and uniformly from $\{1, \dots, n\}$ and $\boldsymbol{Y}'_1, \dots, \boldsymbol{Y}'_n$ is an independent copy of $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$. Then,

$$\begin{aligned}\boldsymbol{\Delta}(\boldsymbol{X}) &= \frac{n}{2}\mathbb{E}\big[(\boldsymbol{X} - \boldsymbol{X}')^2 \mid \boldsymbol{X}\big] = \frac{1}{2}\textstyle\sum_{k=1}^n \mathbb{E}\big[(\boldsymbol{Y}_k - \boldsymbol{Y}'_k)^2 \mid \boldsymbol{X}\big] \\ &= \frac{1}{2}\textstyle\sum_{k=1}^n \mathbb{E}\big[\boldsymbol{Y}_k^2 \mid \boldsymbol{X}\big] + \mathbb{E}\Big[{\boldsymbol{Y}'_k}^2\Big] - \mathbb{E}[\boldsymbol{Y}_k \mid \boldsymbol{X}]\mathbb{E}[\boldsymbol{Y}'_k] - \mathbb{E}[\boldsymbol{Y}'_k]\mathbb{E}[\boldsymbol{Y}_k \mid \boldsymbol{X}] \\ &= \frac{1}{2}\textstyle\sum_{k=1}^n \mathbb{E}\big[\boldsymbol{Y}_k^2 \mid \boldsymbol{X}\big] + \mathbb{E}\Big[{\boldsymbol{Y}'_k}^2\Big].\end{aligned}$$

To proceed, consider the following proposition concerning the convexity of trace functions (see [64, Proposition 2] for a short proof):

Proposition 37. *If $g : [\alpha, \beta] \to \mathbb{R}$ is convex for $[\alpha, \beta] \subseteq \mathbb{R}$, then $\boldsymbol{A} \mapsto \mathrm{Tr}[g(\boldsymbol{A})]$ is convex on $\{\boldsymbol{A} \in \mathbb{H}_d : \alpha\boldsymbol{I} \preceq \boldsymbol{A} \preceq \beta\boldsymbol{I}\}$.*

Proposition 37 implies that $\boldsymbol{A} \mapsto \mathrm{Tr}[\boldsymbol{A}^p]$ is convex for $\boldsymbol{A} \succeq \boldsymbol{0}$, since $x \mapsto x^p$ is convex for $x \geq 0$. Thus, we may apply Jensen's inequality twice to obtain

$$\begin{aligned}\mathbb{E}[\mathrm{Tr}[\boldsymbol{\Delta}(\boldsymbol{X})^p]] &= \mathbb{E}\left[\mathrm{Tr}\left[\left(\frac{1}{2}\textstyle\sum_{k=1}^n \mathbb{E}\big[\boldsymbol{Y}_k^2 \mid \boldsymbol{X}\big] + \mathbb{E}\Big[{\boldsymbol{Y}'_k}^2\Big]\right)^p\right]\right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\mathrm{Tr}\left[\left(\frac{1}{2}\textstyle\sum_{k=1}^n \boldsymbol{Y}_k^2 + {\boldsymbol{Y}'_k}^2\right)^p\right] \mid \boldsymbol{X}\right]\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\left[\left(\frac{1}{2}\textstyle\sum_{k=1}^n \boldsymbol{Y}_k^2 + {\boldsymbol{Y}'_k}^2\right)^p\right]\right] \\ &\leq \mathbb{E}\left[\frac{1}{2}\mathrm{Tr}\big[\big(\textstyle\sum_{k=1}^n \boldsymbol{Y}_k^2\big)^p\big] + \frac{1}{2}\mathrm{Tr}\Big[\Big(\textstyle\sum_{k=1}^n {\boldsymbol{Y}'_k}^2\Big)^p\Big]\right] \\ &= \mathbb{E}\big[\mathrm{Tr}\big[\big(\textstyle\sum_{k=1}^n \boldsymbol{Y}_k^2\big)^p\big]\big].\end{aligned}$$

The proof of Theorem 23 established that $\boldsymbol{X}$ and $\boldsymbol{X}'$ are exchangeable and that

$$\mathbb{E}[\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{X}] = \boldsymbol{f}(\boldsymbol{X}),$$

so Theorem 27 now implies

$$\mathbb{E}\big[\mathrm{Tr}\big[\boldsymbol{f}(\boldsymbol{X})^{2p}\big]\big] \leq (2p-1)^p \mathbb{E}[\mathrm{Tr}[\boldsymbol{\Delta}(\boldsymbol{X})^p]] \leq (2p-1)^p \mathbb{E}\big[\mathrm{Tr}\big[\big(\textstyle\sum_{k=1}^n \boldsymbol{Y}_k^2\big)^p\big]\big].$$

□

Proof of Corollary 29

Proof Let $\boldsymbol{Y}_k \triangleq \xi_k \boldsymbol{A}_k$. To establish

$$\operatorname{Tr}\big[\big(\sum_{k=1}^n \boldsymbol{Y}_k^2\big)^p\big] \le \operatorname{Tr}\big[\big(\sum_{k=1}^n \boldsymbol{A}_k^2\big)^p\big] \quad \text{a.s.}$$

when $\sum_{k=1}^n \boldsymbol{Y}_k^2 = \sum_{k=1}^n \xi_k^2 \boldsymbol{A}_k^2 \preceq \sum_{k=1}^n \boldsymbol{A}_k^2$ a.s., we appeal to the monotonicity of trace functions (see [64, Proposition 1] for a concise proof):

Proposition 38. *If $g : [\alpha, \beta] \to \mathbb{R}$ is nondecreasing for $[\alpha, \beta] \subseteq \mathbb{R}$, and $\alpha \boldsymbol{I} \preceq \boldsymbol{A}, \boldsymbol{H} \preceq \beta \boldsymbol{I}$, then $\boldsymbol{A} \preceq \boldsymbol{H}$ implies $\operatorname{Tr}[g(\boldsymbol{A})] \le \operatorname{Tr}[g(\boldsymbol{H})]$.*

Applying Theorem 28 to $\sum_{k=1}^n \boldsymbol{Y}_k$ now yields the result. □

Proof of Theorem 30

Proof Fix any positive integer p, and, as in the proof of Theorem 26, let

$$\boldsymbol{X} \triangleq \textstyle\sum_{i=1}^n \boldsymbol{A}_{i\pi(i)}, \quad \boldsymbol{f}(\boldsymbol{X}) \triangleq \boldsymbol{X} - \mathbb{E}[\boldsymbol{X}], \quad \boldsymbol{X}' \triangleq \textstyle\sum_{i=1}^n \boldsymbol{A}_{i\pi'(i)},$$

and

$$\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \triangleq \frac{n}{2}(\boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{f}(\boldsymbol{X}'))$$

where $\pi' \triangleq \pi \circ (I, J)$ for indices (I, J) drawn independently and uniformly from $\{1, \dots, n\}^2$. Then,

$$\begin{aligned}
&\frac{1}{2} \operatorname{Re} \mathbb{E}[(\boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{f}(\boldsymbol{X}'))\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \mid \pi] \\
&\quad = \frac{n}{4} \mathbb{E}\big[(\boldsymbol{X} - \boldsymbol{X}')^2 \mid \pi\big] \\
&\quad = \frac{n}{4} \mathbb{E}\big[(\boldsymbol{A}_{I\pi(I)} + \boldsymbol{A}_{J\pi(J)} - \boldsymbol{A}_{J\pi(I)} - \boldsymbol{A}_{I\pi(J)})^2 \mid \pi\big] \\
&\quad = \frac{1}{4n} \textstyle\sum_{i=1}^n \sum_{j=1}^n (\boldsymbol{A}_{i\pi(i)} + \boldsymbol{A}_{j\pi(j)} - \boldsymbol{A}_{j\pi(i)} - \boldsymbol{A}_{i\pi(j)})^2 \\
&\quad = \boldsymbol{\Delta}.
\end{aligned}$$

The proof of Theorem 26 established that $\boldsymbol{X}$ and $\boldsymbol{X}'$ are exchangeable and that

$$\mathbb{E}[\boldsymbol{F}(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{X}] = \boldsymbol{f}(\boldsymbol{X}),$$

so Theorem 27 now implies

$$\mathbb{E}\big[\operatorname{Tr}\big[\boldsymbol{f}(\boldsymbol{X})^{2p}\big]\big] \le (2p-1)^p \mathbb{E}[\operatorname{Tr}[\mathbb{E}[\boldsymbol{\Delta} \mid \boldsymbol{X}]^p]] \le (2p-1)^p \mathbb{E}[\operatorname{Tr}[\boldsymbol{\Delta}^p]]$$

by Jensen's inequality since $\boldsymbol{H} \mapsto \operatorname{Tr}[\boldsymbol{H}^p]$ is convex for $\boldsymbol{H} \succeq \boldsymbol{0}$ by Proposition 37. □

Proof of Corollary 31

Proof Since $\|\boldsymbol{h}(X)\| = \lambda_{\max}(\mathscr{D}(\boldsymbol{h}(X)))$ the result follows from Theorem 22. □

Proof of Corollary 32

Proof We apply Theorem 27 to obtain

$$\begin{aligned}(2p-1)^p \mathbb{E}[\operatorname{Tr}[\boldsymbol{\Delta}(X)^p]] &\geq \mathbb{E}\big[\operatorname{Tr}\big[\boldsymbol{f}(X)^{2p}\big]\big] \\ &= \mathbb{E}\left[\operatorname{Tr}\left[\begin{bmatrix}\boldsymbol{h}(X)\boldsymbol{h}(X)^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{h}(X)^*\boldsymbol{h}(X)\end{bmatrix}^p\right]\right] \\ &= 2\mathbb{E}[\operatorname{Tr}[(\boldsymbol{h}(X)\boldsymbol{h}(X)^*)^p]]\end{aligned}$$

since $\operatorname{Tr}[(\boldsymbol{h}(X)\boldsymbol{h}(X)^*)^p] = \operatorname{Tr}[(\boldsymbol{h}(X)^*\boldsymbol{h}(X))^p]$. □

Bibliography

[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] R. Ahlswede and A. Winter. "Strong converse for identification via quantum channels". In: *IEEE Transactions on Information Theory* 48.3 (2002), pp. 569–579.

[3] E. Airoldi et al. "Mixed Membership Stochastic Blockmodels". In: *Journal of Machine Learning Research* 9 (2008), pp. 1981–2014.

[4] S. Bernstein. "The Theory of Probabilities". In: *Gastehizdat Publishing House* (1946).

[5] R. Bhatia. *Matrix Analysis*. New York: Springer-Verlag, 1997, pp. xii+347. ISBN: 0-387-94846-5.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[7] S. Boucheron, G. Lugosi, and P. Massart. "A sharp concentration inequality with applications". In: *Random Struct. Algorithms* 16.3 (2000), pp. 277–292.

[8] D. Burkholder. "Distribution function inequalities for martingales." English. In: *Ann. Probab.* 1 (1973), pp. 19–42. DOI: 10.1214/aop/1176997023.

[9] J. Cadima and I. Jolliffe. "Loadings and correlations in the interpretation of principal components". In: *Applied Statistics* 22 (1995), p. 203.214.

[10] E. J. Candès et al. "Robust Principal Component Analysis?" In: *Journal of the ACM* 58.3 (2011), pp. 1–37.

[11] E. Candès and Y. Plan. "Matrix Completion With Noise". In: *Proceedings of the IEEE* 98.6 (2010), pp. 925 –936.

[12] V. Chandrasekaran et al. "Sparse and low-rank matrix decompositions". In: *Allerton Conference on Communication, Control, and Computing*. Monticello, Illinois, USA, 2009.

[13] S. Chatterjee. "Stein's method for concentration inequalities". In: *Probability Theory and Related Fields* 138 (2007), pp. 305–321.

[14] Y. Chen et al. "Robust Matrix Completion and Corrupted Columns". In: *International Conference on Machine Learning*. 2011.

[15] S.-S. Cheung, A. M.-C. So, and K. Wang. "Chance-Constrained Linear Matrix Inequalities with Dependent Perturbations: A Safe Tractable Approximation Approach." Preprint. 2011.

[16] A. d'Aspremont, F. R. Bach, and L. E. Ghaoui. "Full regularization path for sparse principal component analysis". In: *Proceedings of the 24th international Conference on Machine Learning*. Ed. by Z. Ghahramani. vol. 227. ACM, New York, NY: ICML '07, 2007, pp. 177–184.

[17] A. d'Aspremont et al. "A Direct Formulation for Sparse PCA using Semidefinite Programming". In: *Advances in Neural Information Processing Systems (NIPS)*. Vancouver, BC, 2004.

[18] D. DeCoste. "Collaborative prediction using ensembles of Maximum Margin Matrix Factorizations". In: *Proceedings of the Twenty-Third International Conference on Machine Learning*. 2006.

[19] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. "Relative-Error CUR Matrix Decompositions". In: *SIAM Journal on Matrix Analysis and Applications* 30 (2 2008), pp. 844–881.

[20] F. Z. (Ed.) *The Schur Complement and Its Applications*. Kluwer, Dordrecht, Springer, 2005.

[21] C. C. Fowlkes et al. "A Quantitative Spatio-temporal Atlas of Gene Expression in the Drosophila Blastoderm". In: *Cell* 133 (2008), pp. 364–374.

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

[23] S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Pattern Analysis and Machine Intelligence* 6 (1984), pp. 721–741.

[24] A. Gittens and J. A. Tropp. "Tail bounds for all eigenvalues of a sum of random matrices". In: *ArXiv e-prints* (Apr. 2011). eprint: `1104.4513`.

[25] S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin. "A theory of pseudoskeleton approximations". In: *Linear Algebra and its Applications* 261.1-3 (1997), pp. 1 –21.

[26] D. Gross and V. Nesme. "Note on sampling without replacing from a finite collection of matrices". In: *CoRR* abs/1001.2738 (2010).

[27] D. Gross. "Recovering Low-Rank Matrices From Few Coefficients in Any Basis". In: *IEEE Transactions on Information Theory* 57.3 (2011), pp. 1548–1566.

[28] W. Hoeffding. "A combinatorial central limit theorem." English. In: *Ann. Math. Stat.* 22 (1951), pp. 558–566.

[29] W Hoeffding. "Probability inequalities for sums of bounded random variables". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30.

[30] T. Hofmann, J. Puzicha, and M. I. Jordan. "Learning from dyadic data". In: *Neural Information Processing Systems*. 1999.

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

[32] J. Jeffers. "Two case studies in the application of principal components". In: *Applied Statistics* 16 (1967), pp. 225–236.

[33] I. T. Jolliffe. "Rotation of principal components: choice of normalization constraints". In: *Journal of Applied Statistics* 22 (1995), pp. 29–35.

[34] I. T. Jolliffe and M. Uddin. "A Modified Principal Component Technique based on the Lasso". In: *Journal of Computational and Graphical Statistics* 12 (2003), p. 531.547.

[35] R. H. Keshavan, A. Montanari, and S. Oh. "Matrix Completion from Noisy Entries". In: *Journal of Machine Learning Research* 99 (2010), pp. 2057–2078.

[36] A. Khintchine. "Über dyadische Brüche". In: *Mathematische Zeitschrift* 18 (1 1923). 10.1007/BF01192399, pp. 109–116. ISSN: 0025-5874.

[37] Y. Koren. "Factorization meets the neighborhood: a multifaceted collaborative filtering model". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2008.

[38] Y. Koren, R. M. Bell, and C. Volinsky. "Matrix Factorization Techniques for Recommender Systems". In: *IEEE Computer* 42.8 (2009), pp. 30–37.

[39] S. Kumar, M. Mohri, and A. Talwalkar. "Ensemble Nyström Method". In: *NIPS*. 2009.

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

[41] N. Lawrence and R. Urtasun. "Non-linear matrix factorization with Gaussian processes". In: *Proceedings of the Twenty-Sixth International Conference on Machine Learning*. 2009.

[42] E. H. Lieb. "Convex trace functions and the Wigner-Yanase-Dyson conjecture". In: *Advances in Mathematics* 11.3 (1973), pp. 267–288.

[43] Z. Lin et al. *Fast Convex Optimization Algorithms for Exact Recovery of a Corrupted Low-Rank Matrix*. UIUC Technical Report UILU-ENG-09-2214. 2009.

[44] F. Lust-Piquard. "Inégalités de Khintchine dans C_p $(1 < p < \infty)$". In: *C. R. Acad. Sci. Paris Sér. I Math.* 303.7 (1986), pp. 289–292. ISSN: 0249-6291.

[45] S. Ma, D. Goldfarb, and L. Chen. "Fixed point and Bregman iterative methods for matrix rank minimization". In: *Mathematical Programming* 128.1-2 (2011), pp. 321–353.

[46] L. Mackey, D. Weiss, and M. I. Jordan. "Mixed Membership Matrix Factorization". In: *Proceedings of the 27th International Conference on Machine Learning*. 2010.

[47] L. Mackey et al. "Matrix Concentration Inequalities via the Method of Exchangeable Pairs". In: *ArXiv e-prints* (Jan. 2012). eprint: 1201.6002.

[48] L. Mackey. "Deflation Methods for Sparse PCA". In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller et al. 2009, pp. 1017–1024.

[49] L. Mackey, A. Talwalkar, and M. I. Jordan. "Divide-and-Conquer Matrix Factorization". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor et al. 2011, pp. 1134–1142.

[50] B. Marlin. "Collaborative Filtering: A Machine Learning Perspective". en. MA thesis. University of Toronto, 2004.

[51] B. Marlin. "Modeling User Rating Profiles For Collaborative Filtering". In: *Neural Information Processing Systems*. 2003.

[52] C. McDiarmid. "On the method of bounded differences". In: *Surveys in Combinatorics 1989*. London Mathematical Society Lecture Notes, 1989, pp. 148–188.

[53] K. Min et al. "Decomposing background topics from keywords by principal component pursuit". In: *Conference on Information and Knowledge Management*. 2010.

[54] B. Moghaddam, Y. Weiss, and S. Avidan. *Generalized spectral bounds for sparse LDA*. In Proc. ICML, 2006.

[55] B. Moghaddam, Y. Weiss, and S. Avidan. *Spectral bounds for sparse PCA: Exact and greedy algorithms*. 18: Advances in Neural Information Processing Systems, 2006.

[56] M. Mohri and A. Talwalkar. "Can Matrix Coherence be Efficiently and Accurately Estimated?" In: *Conference on Artificial Intelligence and Statistics*. 2011.

[57] Y. Mu et al. "Accelerated Low-Rank Visual Recovery by Random Projection". In: *Conference on Computer Vision and Pattern Recognition*. 2011.

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

[59] A. Nemirovski. "Sums of random symmetric matrices and quadratic optimization under orthogonality constraints". In: *Math. Program.* 109 (2 2007), pp. 283–317. ISSN: 0025-5610. DOI: 10.1007/s10107-006-0033-0. URL: http://dl.acm.org/citation.cfm?id=1229716.1229726.

[60] R. I. Oliveira. "Sums of random Hermitian matrices and an inequality by Rudelson". In: *ArXiv e-prints* (Apr. 2010). eprint: 1004.3821.

[61] S.-T. Park and D. M. Pennock. "Applying collaborative filtering techniques to movie search for better ranking and browsing". In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2007.

[62] V. Paulsen. *Completely bounded maps and operator algebras*. Cambridge studies in advanced mathematics. Cambridge University Press, 2002. ISBN: 9780521816694.

[63] Y. Peng et al. "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images". In: *Conference on Computer Vision and Pattern Recognition*. 2010.

[64] D. Petz. "A survey of certain trace inequalities". In: *Functional analysis and operator theory* 30 (1994), 287298.

[65] G. Pisier and Q. Xu. "Non-commutative martingale inequalities". In: *Comm. Math. Phys.* 189.3 (1997), pp. 667–698. ISSN: 0010-3616. DOI: 10.1007/s002200050224. URL: http://dx.doi.org/10.1007/s002200050224.

[66] I. Porteous, E. Bart, and M. Welling. "Multi-HDP: A Non Parametric Bayesian Model for Tensor Factorization". In: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*. 2008.

[67] B. Recht. *A Simpler Approach to Matrix Completion*. arXiv:0910.0651v2[cs.IT]. 2009.

[68] J. Rennie and N. Srebro. "Fast maximum margin matrix factorization for collaborative prediction". In: *Proceedings of the Twenty-Second International Conference on Machine Learning*. 2005.

[69] Y. Saad. "Projection and deflation methods for partial pole assignment in linear state feedback". In: *IEEE Trans. Automat. Contr.* 33 (1998), pp. 290–297.

[70] R. Salakhutdinov and A. Mnih. "Bayesian probabilistic matrix factorization using Markov chain Monte Marlo". In: *Proceedings of the Twenty-Fifth International Conference on Machine Learning*. 2008.

[71] R. Salakhutdinov and A. Mnih. "Probabilistic Matrix Factorization". In: *Advances in Neural Information Processing Systems 20*. 2007.

[72] A. M.-C. So. "Moment inequalities for sums of random matrices and their applications in optimization". In: *Math. Program.* 130.1 (2011), pp. 125–151.

[73] B. K. Sriperumbudur, D. A. Torres, and G. R. G. Lanckriet. "Sparse eigen methods by DC programming". In: *Proceedings of the 24th International Conference on Machine learning* (2007), pp. 831–838.

[74] C. Stein. *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables.* English. Proc. 6th Berkeley Sympos. math. Statist. Probab., Univ. Calif. 1970, 2, 583-602. 1972.

[75] G. Takács et al. "Scalable collaborative filtering approaches for large recommender systems". In: *Journal of Machine Learning Research* 10 (2009), pp. 623–656.

[76] C. Thompson. "If You Liked This, You're Sure to Love That". In: *New York Times Magazine* (2008).

[77] K. Toh and S. Yun. "An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems". In: *Pacific Journal of Optimization* 6.3 (2010), pp. 615–640.

[78] D. Torres, B. K. Sriperumbudur, and G. Lanckriet. *Finding Musically Meaningful Words by Sparse CCA*. Neural Information Processing Systems (NIPS) Workshop on Music, the Brain and Cognition, 2007.

[79] J. A. Tropp. "User-friendly tail bounds for sums of random matrices". In: *Found. Comput. Math.* (2011).

[80] P. White. "The Computation of Eigenvalues and Eigenvectors of a Matrix". In: *Journal of the Society for Industrial and Applied Mathematics, Vol* 6.4 (1958), pp. 393–437.

[81] C. Williams and M. Seeger. "Using the Nyström Method to Speed Up Kernel Machines". In: *NIPS*. 2000.

[82] Z. Zhang, H. Zha, and H. Simon. "Low-rank approximations with sparse factors I: Basic algorithms and error analysis". In: *SIAM J. Matrix Anal. Appl.* 23 (2002), pp. 706–727.

[83] Z. Zhang, H. Zha, and H. Simon. "Low-rank approximations with sparse factors II: Penalized methods with discrete Newton-like iterations". In: *SIAM J. Matrix Anal. Appl.* 25 (2004), pp. 901–920.

[84] Z. Zhou et al. *Stable Principal Component Pursuit*. `arXiv:1001.2363v1[cs.IT]`. 2010.

[85] H. Zou, T. Hastie, and R. Tibshirani. "Sparse Principal Component Analysis". Technical Report, Statistics Department, Stanford University. 2004.