The Dissertation Committee for Suriya Gunasekar
certifies that this is the approved version of the following dissertation:

# Mining Structured Matrices in High Dimensions

Committee:

Joydeep Ghosh, Supervisor

Alan C. Bovik

Constantine Caramanis

Pradeep Ravikumar

Sujay Sanghavi

# Mining Structured Matrices in High Dimensions

**by**

**Suriya Gunasekar, B.Tech, M.S.E.**

**DISSERTATION**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**DOCTOR OF PHILOSOPHY**

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2016

# Acknowledgments

I am extremely grateful to have had Joydeep Ghosh as my supervisor. His constant encouragement to pursue independent research ideas, while providing critical guidance to avoid pitfalls has played a pivotal role in culmination of this dissertation. He has been very approachable and supportive throughout my graduate study. He was always available for discussions and advice, be it holidays, or weekends, or late nights before deadlines. He has time and again helped me cope with and accept failed project ideas as essential components of research. I thank him for his support and invaluable guidance.

I have had the privilege of working with some exemplary researchers during the course of my PhD. I have greatly benefited from the skills I have learned from their collaboration and am thankful for their time and effort. Pradeep Ravikumar was very helpful in concertizing one of the early research ideas developed in this dissertation. I have gained several useful insights on research methodology from the discussions with him. Arindam Banerjee's passion for new research directions is highly infectious. I thoroughly enjoyed our numerous skype calls and energetic discussions. Parts of the results in this dissertation were developed during a fun and challenging summer internship at Yahoo under the mentorship of Makoto Yamada. I am also grateful to have briefly worked with Alan Bovik and will endeavor to incorporate his ideals on reproducible and accessible research in the future. Finally, my recent collaboration with David Sontag has been extremely rewarding and inspirational, not to mention enjoyable. I greatly admire his expertise on a broad range of topics and hope to continue learning from him in the future.

The key to good research is in understanding the fundamentals of the field. I thank Sujay Sanghavi, Constantine Caramanis, and Inderjit Dhillon for teaching me to think intuitively about some of the most wonderful concepts in probability, optimization and linear algebra. These courses were foundational to the research presented in this dissertation. I also thank my teachers and mentors from KV, NITW, IITK, and IITB for providing me with a solid educational background prior to UT.

I thank all my labmates in IDEA lab for providing a fun and engaging learning environment (also for putting up with my loud arguments). I would like to express my sincere gratitude to Sanmi and Sreangsu for their guidance through various stages of my graduate school. From suggesting useful research directions to painstakingly overseeing several aspects of my job search, Sanmi's help has been indispensable. Sreangsu has been like a godfather to the entire lab, especially to me. He has been my chalk board for several technical and non-technical ideas and repeatedly pushed me to develop comprehensive researcher qualities from collaborations and internships. I also thank my fellow student collaborators Sindhu, Ayan, Joyce, Sanmi, and Shalmali for being a fun and efficient team.

Graduate school would have been lot less enjoyable and much more tiring but for my wonderful friends. I am specially thankful to Galeej for being an unwavering source of fun, support and encouragement through the best and worst of the times; and Naga for being a wonderful inspiration and for making me look forward to every weekend over the past few years.

Finally, I could not possibly articulate the contributions of my family: mom, dad and Karthik, as none have had greater influence on my life than them. I thank them for their constant and unconditional love and support.

August 2016

# Mining Structured Matrices in High Dimensions

Publication No. _____

Suriya Gunasekar, Ph.D.
The University of Texas at Austin, 2016

Supervisor: Joydeep Ghosh

Structured matrices refer to matrix valued data that are embedded in an inherent lower dimensional manifold with smaller degrees of freedom compared to the ambient or observed dimensions. Such hidden (or latent) structures allow for statistically consistent estimation in high dimensional settings, wherein the number of observations is much smaller than the number of parameters to be estimated. This dissertation makes significant contributions to statistical models, algorithms, and applications of structured matrix estimation in high dimensional settings. The proposed estimators and algorithms are motivated by and evaluated on applications in e–commerce, healthcare, and neuroscience.

In the first line of contributions, substantial generalizations of existing results are derived for a widely studied problem of matrix completion. Tractable estimators with strong statistical guarantees are developed for matrix completion under (a) generalized observation models subsuming heterogeneous data–types, such as count, binary, etc., and heterogeneous noise models beyond additive Gaussian, (b) general structural constraints beyond low rank assumptions, and (c) collective estimation from multiple sources of data.

The second line of contributions focuses on the algorithmic and application specific ideas for generalized structured matrix estimation. Two specific applications of structured matrix estimation are discussed: (a) a constrained latent factor estimation framework that extends the ideas and techniques hitherto discussed, and applies them for the task of learning clinically relevant phenotypes from Electronic Health Records (EHRs), and (b) a novel, efficient, and highly generalized algorithm for collaborative learning to rank (LETOR) applications.

# Table of Contents

# List of Tables

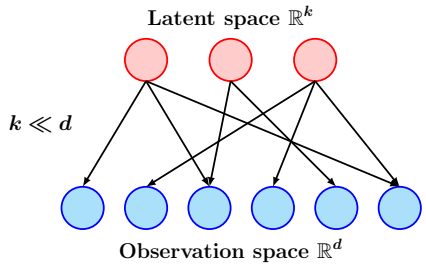
# List of Figures

# Chapter 1

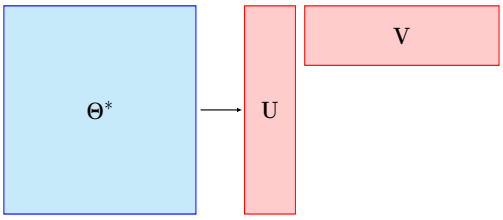
# Introduction

Matrix valued data capturing interactions between a pair of variables — represented along rows and columns — occur naturally in various application settings, e.g., bipartite interactions, network information, spacial interactions in images, covariance matrices, etc. *Structured matrices* refers to matrices that lie in an inherent low dimensional manifold with restricted degrees of freedom compared to the ambient or observed dimensions. A popular example of such a structure is that of low rank, wherein a matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ can be represented as a product of two low rank matrices, say $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{r \times d_2}$ with $r \ll \min\{d_1, d_2\}$. In general, for any matrix that can be represented with smaller number of parameters compared to its ambient dimension, albeit in an unobserved space, its structure can be exploited in various statistical estimation and inference problems. Such hidden low dimensional space of a structured matrix is also commonly referred as its *latent space*. More broadly, learning predictive models by exploiting latent space structures in general vector spaces, not necessarily matrices, has greatly expanded the scope of classical statistical estimation and has led to a surge of research in *high dimensional estimation* problems where the number parameters to be estimated is comparable to (and potentially much larger than) the number of observed samples [27, 29, 46, 33, 114, 10, 150, 13, 153, 23, 126].

The focus of this dissertation is on estimators and algorithms for prediction and inference tasks on structured matrix valued data in high dimensional setting.




(a) Data in observation space $\mathbb{R}^d$ is generated from hidden variables in low dimensional latent space $\mathbb{R}^k$, $k \ll d$.




(b) Low rank matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ factorized as product of two rank $r$ matrices with $(d_1 + d_2 - r)r \ll d_1 d_2$ degrees of freedom

Figure 1.1: Illustration of high dimensional statistics

Mining the latent space representation of structured matrices has been explored in numerous applications including dimensionality reduction techniques such as principal component analysis (PCA) [83, 146, 43], and non-negative matrix factorization (NMF) [98, 101, 148]; topic modeling [72, 19, 110, 124]; collaborative filtering for recommendation systems [56, 94, 167]; subspace clustering [155, 51, 158, 49]; data imputation [39, 66]; covariance matrix estimation [1]; image denoising and other computer vision systems [80, 21]; network coding [65]; distance matrix completion for sensor localization [18, 135]; and more recently for efficient low–rank approximations [17], among several others. A particular problem of interest in high dimensional matrix estimation is that of matrix completion. Matrix completion seeks to recover a low dimensional structured target matrix from noisy measurements of a small fraction of its individual entries. In addition to being a high dimensional estimation problem, the matrix completion task is particularly ill–posed as the observations are not only limited in that the #samples $\ll d_1 d_2$, but each observation is also a highly localized measurement of an individual entry in the matrix. Such localized observations pose additional challenges in analysis of matrix completion estimators in comparison to traditional high dimensional estimation that

assume observations that are global linear combination of all the entries of the target measured using Gaussian or sub–Gaussian operators [53, 52, 128, 33, 13, 153, 126].

This dissertation develops strong statistical models and algorithms for substantial generalizations of high dimensional matrix estimation, focusing on but not limiting to the special case of matrix completion. These models and algorithms are motivated by and evaluated on significant applications in e–commerce, healthcare, and neuroscience. The theoretical and empirical results in this dissertation vastly expand the scope and applicability of structured matrix estimators. Chapters 3–5 address tractable estimators with strong statistical guarantees for matrix completion problems under (a) generalized observation models subsuming heterogeneous data–types, such as count, binary, etc., and heterogeneous noise models beyond additive Gaussian, (b) general structural constraints beyond low rank assumptions, and (c) collective estimation from multiple sources of data, respectively. In Chapter 6, a constrained latent factor estimation framework incorporating ideas developed so far, is discussed for the phenotyping application in healthcare data. Finally, Chapter 7 considers algorithms and applications of structured matrix completion in a collaborative learning to rank (LETOR) formulation.

## 1.1 Generalization of Matrix Completion

A key contribution of this dissertation is the substantial generalization of estimators, statistical analysis, and theoretical guarantees for the high dimensional structured estimation task of matrix completion. As noted earlier, compared to typical high dimensional learning settings, the estimators and analysis of matrix completion are further complicated due to the localized observations. Several novel statistical tools and techniques have been developed in the literature to handle basic formulations of the matrix completion task leading to computationally tractable

estimators with strong statistical guarantees [26, 25, 30, 87, 88, 58, 127, 93, 91, 89, 115, 79, 64]. However, existing literature on matrix completion are specifically well adapted for settings where a subset of entries of a low–rank matrix are observed either deterministically [26], or perturbed by additive noise that is Gaussian [25], or more generally sub–Gaussian [88, 115].

First, let us consider the observation model. While a Gaussian–like noise model for continuous valued data is amenable to the subtle statistical analyses required for the ill–posed problem of matrix completion, it is not always practically suitable for all data settings encountered in matrix completion applications. For instance, a Gaussian error model might not be appropriate in recommender systems based on movie ratings that are either binary (likes or dislikes), or range over the integers one through five. The noise model captures the uncertainty underlying the matrix measurements, and is thus an important component of the problem specification in any application; and it is thus vital for broad applicability of the class of matrix completion estimators to extend to general noise models. Though the generalization of noise models might seem like a narrow technical, although important question, it is related to a broader issue. A Gaussian observation model implicitly assumes the observed matrix values to be continuous (and thin–tail–distributed). But in modern applications, matrix data span the gamut of heterogeneous data–types including skewed–continuous, and categorical–discrete such as binary, count–valued etc., among others. For example, patient electronic health datasets include medication and diagnosis information often recorded as counts, demographics represented as binary or categorical values, and physical measurements as skewed continuous value data. Note that prior to this work there had been some work for the specific case of binary data by [45], but generalizations to other data–types and distributions was largely unexplored. The first problem addressed

in Chapter 3 involves generalization of matrix completion estimator and analysis to observations arising from a rich class of *natural exponential family of distributions* which includes several popular distributions commonly assumed for heterogeneous data types and noise models.

Secondly, while low dimensional structural constraints on the target are understood to be necessary for consistent statistical estimation under high dimensional settings, an (approximate) low rank structure is only one instance of such structures. However, prior to the work discussed in this dissertation, the rich literature on statistical guarantees for consistent matrix completion is exclusively limited to the case of low rank estimation. In the second contribution, a unified statistical analysis of matrix completion under general norm regularization is derived. The framework of general norm regularized estimators proposed in Chapter 4 encompasses a vast variety of low dimensional structures encountered in applications including structured sparseness, superposition structures such as low–rank plus elementwise sparseness, clustered subspace structures, general convex constraint sets, and atomic norms, among others.

Finally, for low rank matrix completion with mild noise assumptions, the known statistical bounds on sample complexity and generalization errors have been shown to be near optimal (upto logarithmic factors) to the information theoretical limits [93, 115]. However, in practice, data commonly arise in the form of multiple matrices sharing correlated information. For example, in e–commerce applications, data containing user preferences in multiple domains such as news, ads, etc., and explicit user/item feature information such as demographics, social network, text description, etc., are made available in the form of a collection of matrices that are coupled through the common set of users/items. The question here is whether such a shared structure among a collection of matrices can be leveraged for accurate pre-

dictions from fewer samples than those required for under single low rank matrix. This setup is analyzed under a convex estimator for collective matrix completion in Chapter 5 and non–trivial sample complexity bounds are derived for the estimate that are optimal for learning from shared information in the entire collection.

## 1.2 Latent Factor Estimation

Mining low dimensional structures in matrices has wider significance beyond the tasks of prediction in high dimensional matrix sensing and completion. The problem class of *latent factor estimation* broadly seeks to reason about the data generation process by identifying the underlying latent structure in the data. While accurate predictions on unseen data for the end task is highly desirable, often black box predictions of the target variable alone is insufficient for informed decision making. In many critical applications, understanding and interpreting the patterns that generate the predictions is crucial for wider deployment in real life systems. Latent factor estimation in matrix valued data is typically studied under low rank assumptions, where additional application specific conditions, such as non–negativity, sparsity, informative priors, etc., are further imposed on the factors. Common examples include PCA, NMF, topic modeling [19, 110, 124] and inference in general graphical models [159, 161].

In Chapter 6, an application of latent factor estimation for high–throughput *electronic health record (EHR)* driven phenotyping is discussed. The increased availability of electronic health records (EHRs) have spearheaded the initiative for precision (personalized) medicine. Essential to this effort is the EHR driven phenotyping task of identifying patients with conditions or characteristics of interest from EHRs. The proposed model incorporates ideas discussed in the earlier chapters towards extracting concise and interpretable phenotypes from heterogeneous EHR

data generated from multiple sources of care givers (e.g., diagnosis, medications, and lab reports).

## 1.3 Collaborative Learning to Rank

A widely popular application of low rank matrix completion is in the collaborative preference completion task of jointly learning missing preferences of set of entities for a shared list of items based on a limited number of observed affinity values, e.g., recommender system [56, 94]. It is commonly assumed that such entity–item preferences are generated from a small number of latent or hidden factors, or equivalently, the underlying preference value matrix is assumed to be low rank. Further, if the observed affinity scores from various explicit and implicit feedback are treated as exact (or mildly perturbed) entries of the unobserved preference value matrix, then the preference completion task naturally fits in the framework of low rank matrix completion.

Recent research in the preference completion literature have noted that using a matrix completion estimator for collaborative preference estimation may be misguided [44, 141, 95] as the observed entity–item affinity scores from implicit/explicit feedback are potentially subject to systematic monotonic transformations arising from limitations in feedback collection, e.g., quantization and inherent biases. In such case, fitting the exact numerical scores in a matrix completion may lead to over-fitting and impair generalization performance. Further, despite the common practice of measuring preferences using numerical scores, predictions are most often deployed or evaluated based on the item ranking e.g. in recommender systems, user recommendations are often presented as a ranked list of items without the underlying scores.

In the final contribution in Chapter 7, a novel, efficient and highly generalized algorithm is developed for the collaborative learning to rank (LETOR) problem, wherein the underlying low rank preferences are learned by fitting the observed order, rather than observed numerical scores. The proposed estimator is also capable of fitting any consistent entity–specific partial ranking over a subset of the items represented as a directed acyclic graph (DAG), further generalizing standard techniques that can only fit preference scores.

# Chapter 2

# Background and Related Work

## 2.1 Notation

Matrices are denoted by capital letters, $X$, $\Theta$, $M$, etc. For a matrix $M$, $M_j$ and $M_{(i)}$ are the $j^{th}$ column and $i^{th}$ row of $M$ respectively, and $M_{ij}$ denotes the $(i,j)^{th}$ entry of $M$. Indexes $i, j$ are typically used to index rows and columns respectively of matrices, and index $s$ is used to index the observations. $e_i$, $e_j$, $e_s$, etc. denote the standard basis in appropriate dimensions*.

Euclidean norm in a vector space is denoted as $\|x\|_2 = \sqrt{\langle x, x\rangle}$. For a matrix $X$ with singular values $\sigma_1 \geq \sigma_2 \geq \ldots$, common norms include the *Frobenius norm* $\|X\|_F = \sqrt{\sum_i \sigma_i^2}$, the *nuclear norm* $\|X\|_* = \sum_i \sigma_i$, the *spectral norm* $\|X\|_{\text{op}} = \sigma_1$, and the *maximum norm* $\|X\|_\infty = \max_{ij} |X_{ij}|$. Also let, $\mathbb{S}^{d_1 d_2 - 1} = \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F = 1\}$ and $\mathbb{B}^{d_1 d_2} = \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F \leq 1\}$.

The *transpose*, *trace*, and *rank* of a matrix $M$ are denoted by $M^\top$, $\text{tr}(M)$, and $\text{rk}(M)$, respectively. The inner product between two matrices is given by $\langle X, Y\rangle = \text{tr}(X^\top Y) = \sum_{(i,j)} X_{ij} Y_{ij}$.

The *Singular Value Decomposition* of a matrix $M \in \mathbb{R}^{d_1 \times d_2}$, of rank $r$ is given by a unique factorization (upto signs) of the form $M = U \Sigma V^\top$, where, $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$ are the left and right singular matrices which have orthonormal columns, and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, such that $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_r$ is

*for brevity the explicit dependence of dimension is omitted unless necessary

the matrix of singular values. For matrix $M$ with singular values $(\sigma_1, \sigma_2, \ldots, \sigma_r)$, common matrix norms include the *nuclear norm* $\|M\|_* = \sum_r \sigma_r$, the *spectral norm* $\|M\|_2 = \sigma_1$, the *Frobenius norm* $\|M\|_F = \sqrt{\sum_r \sigma_r^2}$, and the *maximum norm* $\|M\|_{\max} = \max_{(i,j)} M_{ij}$.

For a linear subspace, $T$, the space orthogonal to $T$ is denoted by $T^\perp$ and the Euclidean projection operator onto a subspace $T$ is denoted by $\mathcal{P}_T$. Given an integer $N$, $[N]$ denotes the set $\{1, 2, \ldots, N\}$. The unit Euclidean sphere and unit Euclidean ball in $\mathbb{R}^{d_1 \times d_2}$ are denoted by $\mathbb{S}^{d_1 d_2 - 1} = \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F = 1\}$ and $\mathbb{B}^{d_1 d_2} = \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F \leq 1\}$, respectively. $\Delta_{d-1} = \{x \in \mathbb{R}^d_+ : \sum_{i=1}^n x_i = 1\}$ denote the $d$ dimensional probability simplex. $\mathbb{P}(.)$ and $\mathbb{E}(.)$ denote the probability of an event and the expectation of a random variable, respectively.

**Definition 2.1.1** (Operator Norm)**.** Let $\mathcal{P} : \mathcal{V} \to \mathcal{W}$ denote a linear operator. The operator norm of $\mathcal{P}$ is given by $\|\mathcal{P}\|_{\text{op}} = \sup_{X \in \mathcal{V} \setminus \{0\}} \frac{\|\mathcal{P}(X)\|_{\mathcal{W}}}{\|X\|_{\mathcal{V}}}$, where $\|.\|_{\mathcal{V}}$ and $\|.\|_{\mathcal{W}}$ are the Euclidean norms in the respective spaces[†].

**Definition 2.1.2** (Dual Norm)**.** Given a norm $\mathcal{R}$ defined on a Banach space $\mathcal{B}$, the dual norm $\mathcal{R}^* : \mathcal{B}^* \to \mathbb{R}_+$ is given by: $\mathcal{R}^*(X) = \sup_{Y : \mathcal{R}(Y) \leq 1} \langle X, Y \rangle$.

**Definition 2.1.3** (Decomposable Norm [114])**.** Norm $\mathcal{R}$ is said to be decomposable over a pair of subspaces $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$ with $\mathcal{M} \subseteq \bar{\mathcal{M}}$, if $\forall\ (X, Y) \in \mathcal{M} \times \bar{\mathcal{M}}^\perp$, $\mathcal{R}(X + Y) = \mathcal{R}(X) + \mathcal{R}(Y)$.

**Definition 2.1.4** (Atomic Norm [33])**.** Let $\mathcal{A}$ denote a set of elementary building blocks called *atoms* such that for a subset $\mathcal{C}$ of interest, $X \in \mathcal{C}$ can be expressed as

[†]Operator norms are in general defined for any pair of norms in the respective spaces, but unless stated otherwise, the notation will be used to refer the operator norm defined on Euclidean norms.

a non–negative affine combination of $\{A_i \in \mathcal{A}\}$ as $X = \sum_i \lambda_i A_i$ for some $\lambda_i \geq 0$. The *atomic norm* with respect to $\mathcal{A}$ is given by the gauge function of conv($\mathcal{A}$):

$$\|X\|_{\mathcal{A}} = \inf\{t > 0 : x \in t.\text{conv}(\mathcal{A})\}.$$

Atomic norm is a norm whenever $\mathcal{A}$ is centrally symmetric, i.e. $A \in \mathcal{A}$ if and only if $-A \in \mathcal{A}$.

**Definition 2.1.5** (Restricted Strong Convexity (RSC))**.** A function $\mathcal{L}$ is said to satisfy *Restricted Strong Convexity (RSC)* at $\Theta$ with respect to a subset $S$, if for some *RSC parameter* $\kappa_{\mathcal{L}} > 0$,

$$\forall \Delta \in S, \mathcal{L}(\Theta + \Delta) - \mathcal{L}(\Theta) - \langle \nabla \mathcal{L}(\Theta), \Delta \rangle \geq \kappa_{\mathcal{L}} \|\Delta\|_F^2. \tag{2.1}$$

**Definition 2.1.6** (Spikiness Ratio [115])**.** Spikiness ratio of $X \in \mathbb{R}^{d_1 \times d_2}$ is given by:

$$\alpha_{\text{sp}}(X) = \frac{\sqrt{d_1 d_2}\|X\|_{\max}}{\|X\|_F}. \tag{2.2}$$

**Definition 2.1.7** (Bregman Divergence)**.** Let $\phi : \text{dom}(\phi) \to \mathbb{R}$ be a strictly convex function differentiable in the relative interior of dom($\phi$). The *Bregman divergence* (associated with $\phi$) between $x \in \text{dom}(\phi)$ and $y \in \text{ri}(\text{dom}(\phi))$ is defined as:

$$B_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle.$$

## 2.2 Related Work

Apart from the following general topics, this dissertation also focuses on special topics of EHR driven phenotyping and collaborative learning to rank. To keep the exposition simple, the background and related literature for these topics are covered in Chapters 6 and 7, respectively.

### 2.2.1 Standard Matrix Completion

Matrix completion and its variants encompass a wide range of applications such as recommendation systems, recovering gene–protein interactions, and modeling text document collections, among others [94, 49, 158]. An broad introduction to classical applications of the problem is covered by Candes et al. [26, 25] and Laurent [96]. As noted earlier, much of the existing literature on matrix completion are specifically well adapted for the special case that assume (a) continuous valued observations with additive thin tailed noise such as Gaussian [25], or more generally sub–Gaussian [88, 115] and (b) low rankness of the target, and are evaluated on exact parameter recovery of a single target matrix. Matrix completion problems under these assumptions will be referred as *Standatd Matrix Completion (SMC)* and is formalized as follows:

Denote the underlying ground truth matrix by $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$. In a matrix completion setting, a subset of the indices $\Omega = \{(i_s, j_s) : i_s \subset [d_1], j_s \subset [d_2], s = 1, 2, \ldots, |\Omega|\} \subset [d_1] \times [d_2]$ of $\Theta^*$ are observed through an additive noise channel:

$$y_s = \Theta^*_{i_s j_s} + \eta_s, \text{ for } s = \{1, 2, \ldots, |\Omega|\},$$

where $\eta_s$ is additive random noise that is assumed to be Gaussian or sub–Gaussian distributed, or bounded.

**Sampling:** $\Omega \subset [d_1] \times [d_2]$ over which $\Theta^*$ is observed is often chosen through a random sampling scheme. The following common sampling assumptions have been shown to be equivalent [30]:

- *Uniform sampling model*: $|\Omega|$ entries of $\Theta^*$ are sampled uniformly and independently at random:

$$\forall\ (i, j) \in \Omega,\ i \sim \text{uniform}([d_1]),\ j \sim \text{uniform}([d_2]). \tag{2.3}$$

- *Bernoulli sampling model*: each element of $[d_1] \times [d_2]$ is independently included in $\Omega$ with a fixed probability of $0 < p < 1$,

$$\forall\ (i,j) \in [d_1] \times [d_2],\ \mathbf{1}_{(i,j)\in\Omega} \sim \text{bernoulli}(p), \tag{2.4}$$

  where $\mathbf{1}_E$ is an indicator variable for an event $E$.

The task in standard matrix completion is to recover $\Theta^*$ from partial and noisy observations, $(\Omega, \{y_s\})$. This task is ill–posed for two reasons:

1. *Limited Sample Size*: Matrix completion is inherently a high dimensional estimation problem and low dimensional structural constraints are necessary for well posed estimation.

2. *Localized Observations*: In a matrix completion, if a small set of entries of the target matrix are overly significant or "spiky" compared to rest of the entries, then a uniform random sampling of observations is likely to miss any information on these significant entries and consistent matrix completion is infeasible [26]. Thus, aside from the low dimensional constraints, further assumptions to eliminate such "spiky" matrices are required for well–posed recovery under localized measurements. Early work analyzing generalization error bounds for various low rank matrix completion algorithms made stringent matrix incoherence assumptions to avoid "spiky" matrices [26, 30, 25, 127, 87, 88, 79]. These assumptions have been made less stringent in more recent results [115, 45] which however guarantees only approximate recovery in low noise settings. In a more recent work [35] also explore leverage score sampling scheme which was shown to be necessary for completing coherent matrix. However, such sampling requires prior knowledge of the elements of the matrix and this line of work is beyond the scope of this dissertation.

Leveraging developments in general high dimensional estimation, numerous models and algorithms have been developed for matrix completion. Theo-

retical results for matrix completion typically quantify bounds on sample complexity and parameter recovery error. Nuclear norm is commonly used as a convex surrogate for low rank constraint in low rank matrix sensing and completion estimators [53, 52, 128]. Early work provide strong statistical analysis for nuclear norm minimization based estimators for matrix completion under observations from thin tailed noise [26, 25, 30, 58, 127]. This line of research generated interest in efficient algorithms for constrained and regularized nuclear norm minimization [81, 22, 147, 109, 77, 108, 12, 75]. More recent work derive approximate recovery guarantees under less restrictive assumptions on incoherence, sampling distributions, and observation model [115, 93, 91, 89]. Apart from nuclear norm minimization, other estimators with theoretical guarantees for consistent matrix completion include the spectral methods [87, 88] and alternating minimization [79, 60, 64]. Besides estimators with theoretical guarantees, a significant line of work for matrix completion includes probabilistic models and other non–convex estimators that have been extensively evaluated on various benchmarked empirical datasets [112, 131, 167, 94]. Extensions of these models to incorporate application–specific additional sources of information such as covariate information, social network, etc. has also been an active area of research [6, 7, 133, 107, 78]

### 2.2.2 High dimensional estimation

High dimensional estimation problems, where the number of parameters to be estimated is much higher than the number of observations are traditionally ill–posed. However, under low dimensional structural constraints, such problems are being extensively studied in the recent literature. Early work focused on the non–asymptotic analysis of estimators for a particular problem of compressed sensing or sparse estimation [46, 27, 29]. More recent work exploit the geometry of general

low dimensional structures in analyzing estimators for generalized linear inverse problems in high dimensions [33, 13, 153, 126]. However, in comparison to matrix completion with localized measurements, such results in general high dimensional estimation assume observations that are global linear combination of all the entries of the target measured using Gaussian or sub–Gaussian operators. In particular, such Gaussian or sub–Gaussian assumption is used to establish some variant of a certain restricted isometry property (RIP) of the measurement ensemble [28]. It has been shown that the localized measurements encountered in matrix completion do not satisfy RIP-like properties [26], and thus novel statistical techniques are generally required to extend the results from general high dimensional estimation to matrix completion settings.

## 2.3 Background

### 2.3.1 Probability

**Lemma 2.3.1** (Bernstein's Inequality (moment version)). *Let $X_i, i = 1, 2, \ldots, N$ be independent zero mean random variables. Further, let $\sigma^2 = \sum_i \mathbb{E}[X_i^2]$, and $M > 0$ be such that the following moment conditions are satisfied for $p \geq 2$,*

$$\mathbb{E}[X_i^p] \leq \frac{p!\sigma^2 M^{p-2}}{2}.$$

*Then the following concentration inequality holds:*

$$\mathbb{P}\Big(\Big|\sum_i X_i\Big| > u\Big) \leq 2\exp\Big(\frac{-u^2}{2\sigma^2 + 2Mu}\Big). \tag{2.5}$$

**Lemma 2.3.2** (Operator Bernstein Inequality [149]). *Let $S_i$, $i = 1, 2, \ldots, m$ be i.i.d self–adjoint operators of dimension $N$. If there exists constants $R$ and $\sigma^2$, such that $\forall i\ \|S_i\|_{op} \leq R$ a.s., and $\sum_i \|E[S_i^2]\|_{op} \leq \sigma^2$,*

$$\textit{then} \quad \forall\, t > 0 \quad Pr\big(\|\textstyle\sum_i S_i\|_{op} > t\big) \leq N\exp\Big(\frac{-t^2/2}{\sigma^2 + \frac{Rt}{3}}\Big).$$

**Lemma 2.3.3** (McDiarmid's Inequality)**.** *Let $X_i, i = 1, 2, \ldots, N$ be independent random variables. Consider a function $f : \mathbb{R}^N \to \mathbb{R}$. If $\forall i$,*

$$\sup_{X_1, X_2, \ldots, X_N, X_i'} |f(X_1, X_2, \ldots, X_N) - f(X_1, X_2, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_N) \leq c_i,$$

*then,*

$$\mathbb{P}(|f(X_1, X_2, \ldots, X_N) - \mathbb{E}f(X_1, X_2, \ldots, X_N)| > u) \leq 2 \exp\left(\frac{-2u^2}{\sum_i c_i^2}\right). \quad (2.6)$$

**Lemma 2.3.4** (Ahlswede-Winter Matrix Bound (Extension))**.** *The Orlicz norm of a random matrix $Z \in \mathbb{R}^{d_1 \times d_2}$ w.r.t to a convex, differentiable and monotonically increasing function, $\phi(x) : \mathbb{R}^+ \to \mathbb{R}$ as follows:*

$$\|Z\|_\phi \triangleq inf\{t \geq 0 : \mathbb{E}\left[\phi\left(|\langle Z, Z' \rangle|/t\right)\right] \leq 1, \\ \forall\ Z' \in \mathbb{R}^{d_1 \times d_2},\ and\ Z'_{ij} \in [0, 1]\}.$$

*Let $Z^{(1)}, Z^{(2)}, \ldots, Z^{(K)}$ be random matrices of dimensions $m \times n$. Let $\|Z^{(i)}\|_\phi \leq M$, $\forall i$. Further, $\sigma_i^2 = \max\{\|\mathbb{E}[Z^{(i)^T} Z^{(i)}]\|_2, \|\mathbb{E}[Z^{(i)} Z^{(i)^T}]\|_2\}$, and $\sigma^2 = \sum_{i=1}^K \sigma_i^2$, then*

$$\mathbb{P}\left(\|\sum_{i=1}^K Z^{(i)}\|_2 \geq t\right) \leq d_1 d_2 \max\left\{e^{-\frac{t^2}{4\sigma^2}}, e^{-\frac{t}{2M}}\right\}.$$

The above lemma is an extension noted by [151] (Theorem 1 and a later remark) for the matrix bounds resulting from [9].

**Lemma 2.3.5** (Symmetrization (Lemma 6.3 in [97]))**.** *Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function, and $X_i, i = 1, 2, \ldots$ be a sequence of mean zero random variables in a Banach space $B$, s.t $\forall i, \mathbb{E}F\|X_i\| < \infty$. Denote a vector of standard Rademacher variables of appropriate dimension as $(\epsilon_i)$, then*

$$\mathbb{E}F\Big(\frac{1}{2}\|\sum_i \epsilon_i X_i\|\Big) \leq \mathbb{E}F\|\sum_i X_i\| \leq \mathbb{E}F\Big(2\|\sum_i \epsilon_i X_i\|\Big). \quad (2.7)$$

*Further, if $X_i$ are not centered, then* $\mathbb{E}F\Big(\|\sum_i X_i - \mathbb{E}[X_i]\|\Big) \leq \mathbb{E}F\Big(2\|\sum_i \epsilon_i X_i\|\Big)$.

**Lemma 2.3.6** (Contraction Principle)**.** *Consider a bounded $T \subset \mathbb{R}^N$, a standard Gaussian and standard Rademacher sequence, $(g_i) \in \mathbb{R}^N$ and $(\epsilon_i) \in \mathbb{R}^N$, respectively. If $\phi_i : \mathbb{R} \to \mathbb{R}$, $i \leq N$ are contractions, i.e. $\forall s, t \in \mathbb{R}$, $|\phi_i(s) - \phi_i(t)| \leq |s-t|$, and with $\phi_i(0) = 0$, then for any convex function $F : \mathbb{R}_+ \to \mathbb{R}_+$, the following results are from Corollary* 3.17*, Theorem* 4.12*, and Lemma* 4.5*, respectively in [97]:*

$$\mathbb{E}F\Big(\frac{1}{2}\sup_{t\in T}\Big|\sum_{i=1}^N g_i\phi_i(t_i)\Big|\Big) \leq \mathbb{E}F\Big(2\sup_{t\in T}\Big|\sum_{i=1}^N g_i t_i\Big|\Big) \tag{2.8}$$

$$\mathbb{E}F\Big(\frac{1}{2}\sup_{t\in T}\Big|\sum_{i=1}^N \epsilon_i\phi_i(t_i)\Big|\Big) \leq \mathbb{E}F\Big(2\sup_{t\in T}\Big|\sum_{i=1}^N \epsilon_i t_i\Big|\Big) \tag{2.9}$$

$$\mathbb{E}F\Big(\|\sum_{i=1}^N \epsilon_i t_i\|\Big) \leq \mathbb{E}F\Big(\sqrt{\frac{\pi}{2}}\|\sum_{i=1}^N g_i t_i\|\Big) \tag{2.10}$$

#### 2.3.1.1 Natural Exponential Family Distributions

**Definition 2.3.1** (Natural Exponential Family)**.** A distribution of a random variable $Y$ in a normed vector space $\mathcal{V}$ is said to belong to the *natural exponential family*, if its probability density function characterized by a natural parameter $\Theta \in \mathcal{V}^*$ can be written as:

$$\mathbb{P}(Y|\Theta) = h(Y)\exp\Big(\langle Y, \Theta\rangle - G(\Theta)\Big),$$

where $h(Y)$ is independent of $\Theta$, and $G(\Theta) = \log\int_Y h(Y)e^{\langle Y,\Theta\rangle}\mathrm{d}Y$, called the log–partition function, is a strictly convex and analytic function,.

The Fenchel conjugate of the log–partition function $G$ is given by: $F(Y) \triangleq \sup_\Theta \langle X, \Theta\rangle - G(\Theta)$. A useful consequence of the exponential family is that the

negative log–likelihood is a strictly convex and analytic function of the natural parameters $\Theta$. Further, Banerjee et al. [14] show that the negative log likelihood as a function of $\Theta$ has a bijection with a large class of divergence functions called *Bregman divergences* (Definition 2.1.7).

#### 2.3.1.2 Sub–Gaussian and Sub–exponential Random Variables

**Definition 2.3.2** (Sub–Gaussian Random Variable [152])**.** The sub–Gaussian norm of a random variable $X$ is given by: $\|X\|_{\Psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$. $X$ is $b$–*sub–Gaussian* if $\|X\|_{\Psi_2} \leq b < \infty$. Equivalently, $X$ is sub–Gaussian if one of the following conditions are satisfied for some constants $k_1$, $k_2$, and $k_3$ [Lemma 5.5 of [152]].

(1) $\forall p \geq 1, (\mathbb{E}|X|^p)^{1/p} \leq b\sqrt{p}$, (2) $\forall t > 0, \mathbb{P}(|X| > t) \leq e^{1-t^2/k_1^2 b^2}$,

(3) $\mathbb{E}[e^{k_2 X^2/b^2}] \leq e$, or (4) if $\mathbb{E}X = 0$, then $\forall s > 0, \mathbb{E}[e^{sX}] \leq e^{k_3 s^2 b^2/2}$.

**Definition 2.3.3** (Sub–Exponential Random Variables)**.** A random variable $X$ is said be *sub-exponential* if it satisfies one of the following equivalent conditions for $k_1$, $k_2$, and $k_3$ differing from one other by constants [Definition 5.13 of [152]].

1. $\mathbb{P}(|X| > t) \leq e^{1-t/k_1}, \forall\, t > 0$,
2. $\forall p \geq 1, (\mathbb{E}[|X|^p])^{1/p} \leq k_2 p$, or
3. $\mathbb{E}[e^{X/k_3}] \leq e$.

The *sub–exponential norm* is given by:

$$\|X\|_{\Psi_1} = \inf\left\{t > 0 : \mathbb{E}\exp\left(\frac{|X|}{t}\right) \leq 2\right\} = \sup_{p\geq 1} p^{-1}(\mathbb{E}[|X|^p])^{1/p}. \tag{2.11}$$

**Lemma 2.3.7** (Hoeffding–type inequality, Proposition **5.10** in [152])**.** *Let $X_1, X_2, \ldots, X_N$ be independent centered sub-Gaussian random variables, and let*

$K = \max_i \|X_i\|_{\Psi_2}$. *Then,* $\forall a \in \mathbb{R}^N$ *and* $t \geq 0$, $\exists$ *constant* $c$ *s.t.,*

$$\mathbb{P}\Big(\big|\sum_{i=1}^{N} a_i X_i\big| \geq t\Big) \leq 2\exp\Big(\frac{-ct^2}{K^2\|a\|_2^2}\Big). \tag{2.12}$$

**Lemma 2.3.8** (Bernstein–type inequality, Proposition **5.16** in [152])**.** *Let* $X_1, X_2, \ldots, X_N$ *be independent centered sub-exponential random variables, and let* $K = \max_i \|X_i\|_{\Psi_1}$. *Then* $\forall a \in \mathbb{R}^N$, *and* $t \geq 0$, *there exists a constant* $c$ *s.t.*

$$\mathbb{P}\Big(\big|\sum_{i=1}^{N} a_i X_i\big| \geq t\Big) \leq 2\exp\Big(-c\min\Big\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\Big\}\Big). \tag{2.13}$$

**Lemma 2.3.9** (Lemma **5.14** in [152])**.** $X$ *is sub–Gaussian if and only if* $X^2$ *is sub–exponential. Further,* $\|X\|_{\Psi_2}^2 \leq \|X^2\|_{\Psi_1} \leq 2\|X\|_{\Psi_2}^2$.

**Lemma 2.3.10** (Remark **5.18** in [152])**.** *If* $X$ *is sub–Gaussian (or sub–exponential), then so is* $X - \mathbb{E}X$. *Further,* $\|X - \mathbb{E}X\|_{\Psi_2} \leq 2\|X\|_{\Psi_2}$; $\|X - \mathbb{E}X\|_{\Psi_1} \leq 2\|X\|_{\Psi_1}$.

### 2.3.2 Gaussian Width

**Definition 2.3.4** (Gaussian Width)**.** Gaussian width of a set $S \subset \mathbb{R}^{d_1 \times d_2}$ is a widely studied measure of complexity of a subset in high dimensional ambient space and is given by:

$$w_G(S) = \mathbb{E}_G \sup_{X \in S} \langle X, G \rangle, \tag{2.14}$$

where $G$ is a matrix of independent standard Gaussian random variables.

Gaussian width plays a key role high dimensional estimation, and plenty of tools have been developed for computing Gaussian widths of compact subsets [48, 97, 145, 33]. The existing work is specially well adapted for computing Gaussian widths for intersection of convex cones with unit norm balls [33], and recent work

of Banerjee et al. [13] propose a mechanism for exploiting these tools for arbitrary compact sets. Some key results that aid in computing Gaussian widths are briefly discussed here. For a cone $\mathcal{C} \in \mathbb{R}^{d_1 \times d_2}$, the polar cone is defined as $\mathcal{C}^\circ = \{X : \langle X, Y \rangle \leq 0, \forall Y \in \mathcal{C}\}$.

#### 2.3.2.1 Direct Estimation

The Gaussian width of a compact set $T$ can be directly estimated as a supremum of Gaussian process over dense countable subset $\bar{T}$ of $T$ as $w_G(T) = \sup_{X \in \bar{T}} \langle X, G \rangle$. The following properties are often used in direct estimation. These properties are consolidated from [145], [33] and [13]. In the following statements, $k$ is a constant not necessarily the same in each occurrence:

- Translation invariant and homogeneous: for any $a \in \mathbb{R}$, $w_G(S+a) = w_G(S)$;
- $w_G(\text{conv}(T)) \leq w_G(T)$
- $w_G(T_1 + T_2) \leq w_G(T_1) + w_G(T_2)$
- If $T_1 \subseteq T_2$, then $w_G(T_1) \leq w_G(T_2)$.
- If $T_1$ and $T_2$ are convex, then $w_G(T_1 \cup T_2) + w_G(T_1 \cap T_2) = w_G(T_1) + w_G(T_2)$

#### 2.3.2.2 Dudley's Inequality and Sudakov Minorization

**Definition 2.3.5** (Covering Number)**.** Consider a metric $d$ defined on $S \subset \mathbb{R}^{d_1 \times d_2}$. Given $\epsilon > 0$, the $\epsilon$–covering number of $S$ with respect to $d$, denoted by $\mathcal{N}(S, \epsilon, d)$, is the minimum number of points $\{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{\mathcal{N}(S,\epsilon,d)}\}$ such that $\forall\, X \in S$, there exists $i \in \{1, 2, \ldots, \mathcal{N}(S, \epsilon, d)\}$ with $d(X, \bar{X}_i) \leq \epsilon$. The set $\{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{\mathcal{N}(S,\epsilon,d)}\}$ is called the $\epsilon$–cover of $S$.

**Lemma 2.3.11** (Dudley's Inequality and Sudakov Minoration)**.** *If $S$ is compact,*

*then for any $\epsilon > 0$, there exists a constant $c$ s. t.*

$$c\epsilon\sqrt{\log N(S, \epsilon, \|.\|_F)} \leq w_G(S) \leq 24 \int_0^\infty \sqrt{N(S, \epsilon, \|.\|_F)} d\epsilon.$$

*The upper bound is the Dudley's inequality and lower bound is by Sudakov minoration.*

#### 2.3.2.3 Geometry of Polar Cone

**Lemma 2.3.12** (Proposition **3.6** and Theorem **3.9** of [33])**.** *If $\mathcal{C} \subset \mathbb{R}^{d_1 \times d_2}$ is a non–empty convex cone and $\mathcal{C}^\circ$ be its polar cone, then:*

$$\text{Distance to polar cone} : w_G(\mathcal{C} \cap \mathbb{S}^{d_1 d_2 - 1}) \leq \mathbb{E}_G[\inf_{X \in \mathcal{C}^\circ} \|G - X\|_F]$$

$$\text{Volume of polar cone} : w_G(\mathcal{C} \cap \mathbb{S}^{d_1 d_2 - 1}) \leq 3\sqrt{\frac{4}{\text{vol}(C^\circ \cap \mathbb{S}^{d_1 d_2 - 1})}}$$

#### 2.3.2.4 Infimum over Translated Cones

**Lemma 2.3.13** (Lemma **3** of [13])**.** *Let $S \subset \mathbb{R}^{d_1 \times d_2}$, and given $X \in S$, define $\rho(X) = \sup_{Y \in S} \|X - Y\|_F$ as the diameter of $S$ measured along $X$. Also define $\mathcal{G}(X) = cone(S - X) \cap \rho(X)\mathbb{B}^{d_1 d_2}$, where $\mathbb{B}^{d_1 d_2}$ is the unit Euclidean ball. Then,*

$$w_G(S) \leq \inf_{X \in S} w_G(\mathcal{G}(X))$$

#### 2.3.2.5 Generic Chaining

Lemma B.1.1 (from [145]) gives the tightest bounds on the Gaussian width of a set. The definition of $\gamma_2$ (B.1) can be used derive tight bounds on the Gaussian width that are optimal upto constants. Further results and examples on using $\gamma$–functionals for Gaussian width computation can be found in the works of Talagrand [143, 144, 145].

# Chapter 3

# Matrix Completion under Generalized Observations

Recent works have proposed computationally tractable estimators with strong statistical guarantees for low rank matrix completion under squared loss minimization over the observed entries. Square loss is implicitly suitable for continuous valued observations perturbed by additive thin–tailed noise like Gaussian or bounded noise. Arguably, common applications of matrix completion require estimators for (a) heterogeneous data–types, such as skewed–continuous, count, binary, etc., and (b) for heterogeneous noise models (beyond Gaussian). This chapter * considers a generalization of matrix completion under the setting where the matrix entries are sampled from a known member of the *exponential family distributions*. A simple convex regularized $M$–estimator is proposed for this generalized framework, and unified and novel statistical analyses for this class of estimators are provided.

## 3.1 Introduction

The general problem of matrix completion seeks to recover a structured matrix from noisy and partial measurements. The literature on tractable estimators and statistical guarantees for matrix completion (Section 2.2.1) is specifically well

*The results in this chapter appear in a conference publication [62]. The coauthors contributed equally.

adapted for the setting where a subset of entries of a low rank matrix are observed either deterministically [26], or perturbed by additive noise that is Gaussian [25], or more generally sub–Gaussian [88, 115]. While such a thin–tailed noise model is amenable to the subtle statistical analyses required for the problem of matrix completion, it is not always practically suitable for all data settings encountered in matrix completion applications. For instance, such a Gaussian error model might not be appropriate in recommender systems based on movie ratings that are quantized to either binary values (likes or dislikes), or over a range of integers (one through five, say). The noise model captures the uncertainty in the underlying matrix measurements, and is an important component of the problem specification in any application; and it is thus vital for broad applicability of the class of matrix completion estimators to extend to general noise models.

Though the generalization of noise models might seem like a narrow technical, although important question, it is related to a broader issue. A Gaussian observation model implicitly assumes the observed matrix values to be continuous (and thin–tail–distributed). But in modern applications, matrix data span the gamut of heterogeneous data–types, including skewed–continuous, and categorical–discrete such as binary, count–valued etc., among others.

A key question motivated by these considerations seeks the feasibility of generalization of the standard matrix completion estimators and statistical analyses, suited for continuous values data with additive thin–tailed noise, to (a) a broader family of noise models, and (b) heterogeneous data–types. This chapter considers a generalized matrix completion setting wherein observed matrix entries are sampled from a known member of a rich family of *natural exponential family distributions*. This family of distributions encompass a wide variety of popular distributions including Gaussian, Poisson, binomial, negative–binomial, Bernoulli, etc. The choice

of a particular member of the exponential family can be made depending on the form of the data and assumptions on the noise channel. For instance, thin–tailed continuous data are typically modeled using the Gaussian distribution; count–data are modeled through an appropriate distribution over integers (Poisson, binomial, etc.), binary data through Bernoulli, categorical–discrete through multinomial, etc.

**Contributions:**

- In a key contribution, a simple regularized convex $M$–estimator is proposed for recovering an underlying matrix from generalized observation models described above; and a unified and novel statistical analysis is provided for the proposed estimator.
- Following a standard approach [114], it is (a) first shown that the negative log–likelihood of the subset of observed entries satisfies a form of Restricted Strong Convexity (RSC) (Definition 2.1.5); and (b) under this RSC condition, the proposed $M$–estimator satisfies strong statistical guarantees. The first component showing the RSC condition for generalized class of loss functions is of independent interest.
- Matrix completion under a broader range of decomposable structures beyond low rankness is also briefly discussed in this chapter, although this generalization will be dealt with in greater detail and generality in Chapter 4.

## 3.2 Exponential Family Matrix Completion

Denote the underlying target matrix to be recovered by $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$. In the matrix completion setting considered in this chapter, a subset of individual entries $\{\Theta^*_{ij}\}$ of $\Theta^*$ are observed indirectly via a noisy channel: specifically, as samples $\{Y_{ij}\}$ drawn from some known member of *natural exponential family* (Defini-

tion 2.3.1):

$$\mathbb{P}(Y_{ij}|\Theta^*_{ij}) = h(Y_{ij}) \exp\left\{Y_{ij}\Theta^*_{ij} - G(\Theta^*_{ij})\right\}, \tag{3.1}$$

where $G()$ is a strictly convex, analytic function called the log–partition function.

**Uniformly Sampled Observations:** In this paper, a partially observed setting is considered, where the observations are sampled for a subset of entries of $\Theta^*$ corresponding to indices $\Omega \subset [d_1] \times [d_2]$. A uniform sampling model is assumed:

$$\forall\ (i,j) \in \Omega,\ i \sim \text{uniform}([d_1]),\ j \sim \text{uniform}([d_2]). \tag{3.2}$$

Note that, under the above described sampling scheme, an index $(i,j)$ can be sampled multiple times, in such cases for each instances of $(i,j)$ in $\Omega$ (and not just the unique indices in $\Omega$), and independently sampled $Y_{ij}$ for each occurrence are included in the set of observation $(Y_{ij})_{(i,j)\in\Omega}$.

Given $\Omega$, a linear operator $\mathcal{P}_\Omega : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ is defined as:

$$\mathcal{P}_\Omega(X) = \sum_{(i,j)\in\Omega} X_{ij} e_i e_j^\top.$$

With a slight abuse of notation, $\mathcal{P}_\Omega(Y) = \sum_{(i,j)\in\Omega} Y_{ij} e_i e_j^\top$ is also used for the observation set $(Y_{ij})_{(i,j)\in\Omega}$ sampled from (3.1), although $Y$ need not be a matrix. The matrix completion task now involves estimation of $\Theta^*$ from $(\Omega, (Y_{ij})_{(i,j)\in\Omega})$.

### 3.2.1 Applications

**Gaussian (fixed $\sigma^2$)** is typically used to model continuous data, $x \in \mathbb{R}$, such as measurements with additive errors, affinity datasets. Here, $G(\theta) = \frac{1}{2}\sigma^2\theta^2$.

**Bernoulli** is a popular distribution of choice to model binary data, $x \in \{0,1\}$, with $G(\theta) = \log(1+e^\theta)$. Some examples of data suitable for Bernoulli model include

social networks, gene protein interactions, etc.

**Binomial (fixed $N$)** is used to model number of successes in $N$ trials. Here, $x \in \{0, 1, 2, \ldots, N\}$, and $G(\theta) = N \log(1 + e^{\theta})$. Some applications include predicting success/failure rate, survey outcomes, etc.

**Poisson** is used to model count data $x \in \{0, 1, 2, \ldots\}$, such as arrival times, events per unit time, click–throughs among others. Here, $G(\theta) = e^{\theta}$.

**Exponential** is often used to model positive valued continuous data $x \in \mathbb{R}_+$, specially inter arrival times between events. Here, $G(\theta) = -\log(-\theta)$.

### 3.2.2 Log–likelihood

Denote the gradient map:

$$g(\Theta) \triangleq \nabla G(\Theta) \in \mathbb{R}^{d_1 \times d_2}, \text{ where } g(\Theta)_{ij} = \frac{\partial G(\Theta_{ij})}{\partial \Theta_{ij}}.$$

It can then be verified that the mean and variance of the distribution $\mathbb{P}(Y_{ij}|\Theta^*_{ij})$ are $\mathbb{E}[Y_{ij}] = g(\Theta^*_{ij})$, and $\text{Var}(Y_{ij}) = \nabla^2 G(\Theta^*_{ij})$, respectively. The Fenchel conjugate of the log partition function $G$, is denoted by: $F(X) \triangleq \sup_{\Theta} \langle X, \Theta \rangle - G(\Theta)$.

A useful consequence of the exponential family is that the negative log–likelihood is convex and differentiable in its natural parameters $\Theta^*$, and moreover has a bijection with a large class of *Bregman divergences* (Definition 2.1.7). The following relationship was first noted by Forster et al. [54], and later rigorously established by Banerjee et al. [14]:

$$-\log \mathbb{P}(Y_{ij}|\Theta_{ij}) \propto B_F(Y_{ij}, g(\Theta_{ij})), \ \forall Y_{ij} \in \text{dom(F)}. \tag{3.3}$$

## 3.3 Main Result and Consequences

Matrix completion is in general ill–posed and low dimensional structural constraints on the underlying target matrix $\Theta^*$ are required for well posed estima-

tion. To formalize the notion of such structural constraints, following [114] it is assumed that $\Theta^*$ satisfies $\Theta^* \in \mathcal{M} \subseteq \overline{\mathcal{M}} \subset \mathbb{R}^{d_1 \times d_2}$, for some subspace $\mathcal{M} \subseteq \overline{\mathcal{M}}$, which contains parameter matrices that are structured similar to the target; the setup allows the flexibility of working with a superset $\overline{\mathcal{M}}$ of the model subspace that is potentially easier to analyze.

**Assumption 3.3.1.** (Decomposable Norm Regularizer) There exists a structure inducing matrix norm $\mathcal{R}(.)$ which is decomposable over $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ (Definition 2.1.7).

Although the main result in this work (Theorem 3.3.1) is applicable for general decomposable norms, for the purpose of this chapter, we focus on the special case of low rank structure induced by nuclear norm which has been previously shown to be decomposable under appropriately defined subspaces [115]. Matrix completion under general norm structures will be discussed in greater detail and generality in Chapter 4.

The second assumption restricts the curvature of the log–partition function. This is required to establish a form of RSC (Definition 2.1.5) for the loss function. It can be verified that commonly used members of natural exponential family satisfy this assumption.

**Assumption 3.3.2.** The second derivative of the log–partition function $G : \mathbb{R} \to \mathbb{R}$ has atmost an exponential decay, i.e,

$$\nabla^2 G(u) \geq e^{-\eta|u|}, \ \forall \ u \in \mathbb{R}, \text{ for some } \eta > 0.$$

Finally, for well posed estimation under matrix completion, additional assumptions besides low dimensional structure is required to avoid missing the most informative entries in a localized sampling model. restriction on spikiness ratio is used to preclude "spiky" target matrices in the analysis. Refer Section 2.2.1

for discussion on this assumption. Recall that the *spikiness ratio* is defined as: $\alpha_{\text{sp}}(\Theta) = \frac{\sqrt{d_1 d_2}\|\Theta\|_{\max}}{\|\Theta\|_F}$ (Definition 2.1.6).

**Assumption 3.3.3.** There exists a known $\alpha^* > 0$, such that

$$\|\Theta^*\|_{\max} = \frac{\alpha_{sp}(\Theta^*)}{\sqrt{d_1 d_2}}\|\Theta^*\|_F \leq \frac{\alpha^*}{\sqrt{d_1 d_2}}.$$

### 3.3.1 $M$–estimator for Generalized Matrix Completion

A regularized $M$–estimate as is proposed as our candidate parameter matrix $\widehat{\Theta}$. The norm regularizer $\mathcal{R}(.)$ used is a convex surrogate for the structural constraints, and is assumed to satisfy Assumption 3.3.1. For a suitable $\lambda > 0$,

$$\begin{aligned}\widehat{\Theta} &= \underset{\|\Theta\|_{\max} \leq \frac{\alpha^*}{\sqrt{d_1 d_2}}}{\text{argmin}} \frac{d_1 d_2}{|\Omega|}\Big[\sum_{ij\in\Omega} -\log \mathbb{P}(Y_{ij}|\Theta_{ij})\Big] + \lambda\mathcal{R}(\Theta) \\ &= \underset{\|\Theta\|_{\max} \leq \frac{\alpha^*}{\sqrt{d_1 d_2}}}{\text{argmin}} \frac{d_1 d_2}{|\Omega|}\Big[\sum_{ij\in\Omega} G(\Theta_{ij}) - Y_{ij}\Theta_{ij}\Big] + \lambda\mathcal{R}(\Theta). \qquad (3.4)\end{aligned}$$

In the above estimator, for simplicity it is assumed that the domain of the minimizing function spans all or $\mathbb{R}^{d_1 \times d_2}$. In cases where this is violated, additional constraints to restrict $\Theta$ to the domain could be imposed on the estimator and the results and analysis in the following section still hold. The above optimization problem is a convex program, and can be solved by any off–the shelf convex solvers.

### 3.3.2 Recovery Results

Let $d = \max\{d_1, d_2\}$. Let $\mathcal{R}^*(.) = \sup_{\mathcal{R}(X)\leq 1}\langle X, .\rangle$ be the dual norm of the regularizer $\mathcal{R}(.)$.

- Given a matrix norm $\mathcal{R}(.)$, the maximum and minimum *subspace compatibil-*

*ity* constants of $\mathcal{R}(.)$ w.r.t the subspace $\mathcal{M}$ are defined as follows:

$$\Psi(\mathcal{M}) = \sup_{\Theta \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{R}(\Theta)}{\|\Theta\|_F}, \quad \Psi_{\min} = \inf_{\Theta \neq \{0\}} \frac{\mathcal{R}(\Theta)}{\|\Theta\|_F}.$$

Thus, $\forall \Theta \in \mathcal{M}$, $\Psi_{\min}\|\Theta\|_F \leq \mathcal{R}(\Theta) \leq \Psi(\mathcal{M})\|\Theta\|_F$.

- Finally, the following quantity will later be proved to be the *RSC parameter* (Definition 2.1.5):

$$\kappa_{\mathcal{R}}(d, |\Omega|) := \mathbb{E}\Big[\frac{\sqrt{d_1 d_2}}{|\Omega|}\mathcal{R}^*\Big(\sum_{ij \in \Omega} \epsilon_{ij} e_i e_j^*\Big)\Big], \tag{3.5}$$

where the expectation is over the random sampling index set $\Omega$, and over a Rademacher sequence $\{\epsilon_{ij} : \forall (i,j) \in \Omega\}$; here $\{e_i \in \mathbb{R}^{d_1}\}$, $\{e_j \in \mathbb{R}^{d_2}\}$ are the standard basis. This quantity $\kappa_{\mathcal{R}}(d, |\Omega|)$ captures the interaction between the sampling scheme and the structural constraint as captured by the regularizer (specifically its dual $\mathcal{R}^*$).

**Theorem 3.3.1.** *Let $\widehat{\Theta}$ be the estimate from (3.4) with $\frac{\lambda}{2} \geq \frac{d_1 d_2}{|\Omega|}\mathcal{R}^*(\mathcal{P}_\Omega(Y - g(\Theta^*))$. Under Assumptions 3.3.1–3.3.3, if $|\Omega| \geq c_0 \Psi^2(\overline{\mathcal{M}}) d \log d)$ for large enough $c_0$, then for any given constant $\beta > 0$, $\exists$ a constant $K_\beta > 0$ such that, using $\mu_{\mathcal{L}} := e^{-\frac{2\eta\alpha^*}{\sqrt{d_1 d_2}}}\Big(K_\beta - \frac{64}{c_0}\sqrt{\frac{|\Omega|\kappa_{\mathcal{R}}^2(d,|\Omega|)}{d \log d}}\Big)$, the following holds with probability $> 1 - 4e^{-(1+\beta)\Psi_{\min}^4 \log^2 d}$:*

$$\|\widehat{\Theta} - \Theta^*\|_F^2 \leq \Psi^2(\overline{\mathcal{M}}) \max\left\{\frac{3\lambda^2}{2\mu_{\mathcal{L}}^2}, \frac{c_0^2 \alpha^{*2} d \log d}{|\Omega|}\right\},$$

*provided $\mu_{\mathcal{L}} > 0$.* □

In the above theorem, $\eta$ and $\alpha^* \geq \alpha_{sp}(\Theta^*)\|\Theta^*\|_F$ are constants from Assumptions 3.3.2 and 3.3.3, respectively.

An important special case of the problem is when the parameter matrix $\Theta^*$, is assumed to be of a low rank $r \ll \min\{d_1, d_2\}$ commonly induced using the

decomposable nuclear norm. Let $\{\mathbf{u_k} \in \mathbb{R}^{d_1}\}$ and $\{\mathbf{v_k} \in \mathbb{R}^{d_2}\}$, $k \in [r]$ be the left and right singular vectors, respectively of $\Theta^*$. Let the column and row span of $\Theta^*$ be $U^* \triangleq \text{col}(\Theta^*) = \text{span}\{\mathbf{u_i}\}$ and $V^* \triangleq \text{row}(\Theta^*) = \text{span}\{\mathbf{v_j}\}$, respectively. Define:

$$\begin{aligned} \mathcal{M} &:= \{\Theta : \text{row}(\Theta) \subseteq V^*,\ \text{col}(\Theta) \subseteq U^*\},\ \text{and} \\ \overline{\mathcal{M}}^{\perp} &:= \{\Theta : \text{row}(\Theta) \subseteq V^{*\perp},\ \text{col}(\Theta) \subseteq U^{*\perp}\}. \end{aligned} \tag{3.6}$$

It can be verified that, $\mathcal{M} \neq \overline{\mathcal{M}}$, however, $\mathcal{M} \subset \overline{\mathcal{M}}$.

**Corollary 3.3.2.** *Let $\Theta^*$ be a low rank matrix of rank atmost $r \ll \min\{d_1, d_2\}$. If further, $\forall (i,j)$, $(Y_{ij} - g(\Theta^*_{ij}))$ are sub–Gaussian (Definition 2.3.2) with parameter $b$, and $|\Omega| > c_0 r d \log d$ for large enough constant $c_0$. Given any $\beta > 0$, there exists constants $c_\beta > 0$, $C_\beta > 0$ and $K_\beta > 0$, such that using $\mathcal{R}(.) = \|.\|_*$ and $\frac{\lambda}{2} := c_\beta \sqrt{d_1 d_2} b \sqrt{\frac{d \log d}{|\Omega|}}$ in (3.4), w.p. $> 1 - 4e^{-(1+\beta)\log^2 d} - e^{-(1+\beta)\log(d)}$,*

$$\frac{1}{d_1 d_2} \|\widehat{\Theta} - \Theta^*\|_F^2 \leq C_\beta \frac{\max\{b^2, \alpha^{*2}/d_1 d_2\}}{\mu_{\mathcal{L}}^2} \left(\frac{r d \log d}{|\Omega|}\right),$$

*where $\mu_{\mathcal{L}} = K_\beta e^{-\frac{2\eta\alpha^*}{\sqrt{d_1 d_2}}} > 0$.*

**Remark 1:** Note that the above results hold for the minimizer $\widehat{\Theta}$ of the convex program in (3.4) for any $\alpha^* \geq \alpha_{sp}(\Theta^*)\|\Theta^*\|_F$; in particular it holds with $\alpha^* = \alpha_{sp}(\Theta^*)\|\Theta^*\|_F$, where $1 \leq \alpha_{sp}(\Theta^*) \leq \sqrt{d_1 d_2}$. While in practice $\alpha^*$ is chosen through cross–validation, the theoretical bound in Corollary 3.3.2 can be tightened to the following (if $\|\Theta\|_F \geq 1$):

$$\frac{\|\widehat{\Theta} - \Theta^*\|_F^2}{\|\Theta^*\|_F^2} \leq C_\beta \frac{\alpha_{\text{sp}}^2(\Theta^*) \max\{b^2, 1\}}{\mu_{\mathcal{L}}^2} \left(\frac{r d \log d}{|\Omega|}\right). \tag{3.7}$$

Similar bound can be obtained for Theorem 3.3.1.

**Remark 2:** $b^2$ is a measure of noise per entry; $\forall (i,j), \text{Var}(Y_{ij} - g(\Theta^*_{ij})) \leq b^2$. Note that, in the absence stronger matrix incoherence assumptions, only an approximate recovery is guaranteed even as $b \to 0$.

### 3.3.3 Discussions

The richness of the class of exponential family distributions has been used in other settings to provide general statistical frameworks. Kakade et al. [85] provide a generalization of compressed sensing problem to general exponential family distributions. However, as discussed in Section 2.2.1, the typical analysis from compressed sensing cannot be immediately extended to matrix completion case, since the sampling operator $\mathcal{P}_\Omega$ does not satisfy the restricted isometry like properties. There have been extensions of classical probabilistic PCA [146] from Gaussian noise models to exponential family distributions [43, 113, 57]. There have also been recent extensions of probabilistic graphical model classes, beyond Gaussian and Ising models, to multivariate extensions of exponential family distributions [159, 161]. More complicated probabilistic models have also been proposed in the context of collaborative filtering [112, 131], but these typically involve non–convex optimization, and it is difficult to extend the rigorous statistical analyses of the form in this paper (and in the matrix completion literature) to these models. Finally, prior to this work there had been some work for the specific case of binary data under Bernoulli distribution by [45], but generalizations to other data–types and distributions is largely unexplored.

Proof of the results in Appendix A uses elements from Negahban et al. [115] where authors analyze the case of low rank structure and additive noise, and establish a form of restricted strong convexity (RSC) for squared loss over subset of matrix entries (closely relates to the special case, when the exponential family distribution assumed in (3.1) is Gaussian). However, showing such an RSC condition for structured matrix entries under the negative log–likelihood losses associated with general exponential family distributions involved some non-trivial and novel proof techniques. Further, a much simpler proof of the result is provided that more-

over only required a low–spikiness condition rather than a multiplicative spikiness and structural constraint.

## 3.4 Experiments

Simulated experiments are provided to corroborate the theoretical guarantees, focusing on Corollary 3.3.2 for low rank matrix completion using observations from any member of the general class of exponential family distributions. Three well known members of exponential family are studied which are suitable for different data–types, namely Gaussian, Bernoulli, and binomial — popular choices for modeling continuous valued, binary, and count valued data, respectively.

### 3.4.1 Experimental Setup

Low rank ground truth parameter matrices $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ are created, with sizes $d \in \{50,\ 100,\ 150,\ 200\}$ (for simplicity consider square matrices, $d_1 = d_2 = d$). The rank of $\Theta^*$ are set to $r = 2 \log d$. For each $d$ and various values of $|\Omega|$, subsets $\Omega \subset [d] \times [d]$ are first uniformly sampled, and then observations $(Y_{ij})_{(i,j)\in\Omega}$ are sampled from the different members of exponential family distributions parameterized by $\Theta^*$.

**Evaluation:**

For each member of the exponential family of distributions considered, the performance of the proposed $M$–estimator can be measured either in parameter space as $\frac{\|\widehat{\Theta}-\Theta^*\|_F^2}{\|\Theta^*\|_F^2}$, or in observation space using an appropriate error metric $\text{err}(\widehat{Y}, Y)$, where $\widehat{Y}$ is the maximum likelihood estimate of the recovered distribution, $\widehat{Y} = \text{argmax}_Y \mathbb{P}(Y|\widehat{\Theta})$ (RMSE, MAE are used in the plots). From Corollary 3.3.2, $|\Omega| = \mathcal{O}(rd \log d)$ samples are required for consistent recovery. Thus, the error

metrics are compared against the "normalized" sample size, $\frac{|\Omega|}{rd\log d}$.

**Parameter Recovery Error:** The results are plotted (a) against the proportion of the total entries sampled $\frac{|\Omega|}{d_1 d_2}$ (left), and (b) against the "normalized" sample size $\frac{|\Omega|}{rd\log d}$ (right) for comparison. Figures 3.1–3.3 plot the resultant performance of the proposed estimator for samples from Gaussian, Bernoulli, and binomial distributions, respectively.

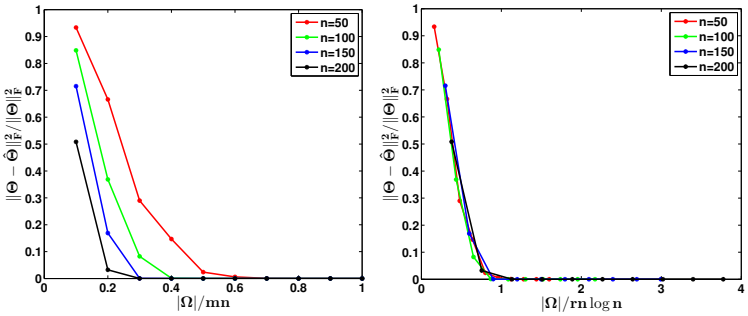



Figure 3.1: Parameter Error when measured $(a)$ against proportion of the sampled values, and $(b)$ against the 'normalized" sample size, when the distribution of the observations $\mathbb{P}(Y|\Theta^*)$, is Gaussian

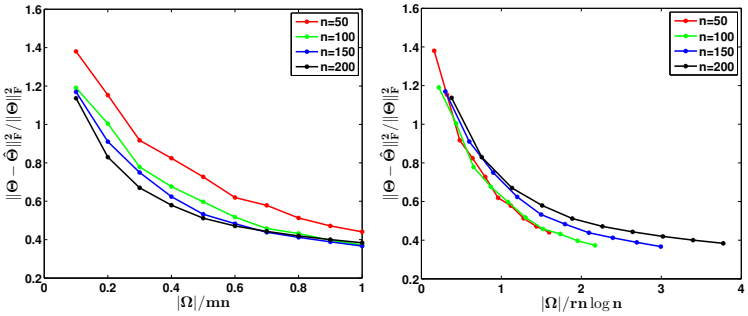



Figure 3.2: Parameter Error when measured $(a)$ against proportion of the sampled values, and $(b)$ against the 'normalized" sample size, when the distribution of the observations $\mathbb{P}(Y|\Theta^*)$, is Bernoulli

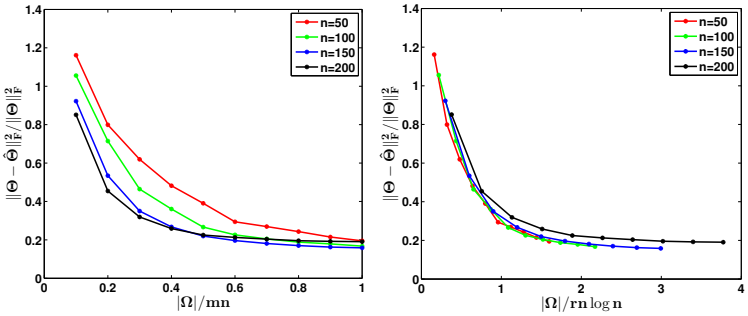



Figure 3.3: Parameter Error when measured $(a)$ against proportion of the sampled values, and $(b)$ against the 'normalized" sample size, when the distribution of the observations $\mathbb{P}(Y|\Theta^*)$, is Binomial

It can be seen from the plots that the error converges to small values proportional to input variance corroborating consistency of estimator; indeed $|\Omega| > 1.5rd\log d$ samples suffice for convergence. Further, aligning of the curves for different $d$ against "normalized" sample size (right) corroborates the convergence rates. Note that the curves do not align against unnormalized sample size (left).

**Sample Recovery Error:** In Figure 3.4 results for sample recovery show trends similar to those under parameter recovery. The curves (for different $d$) plotted against "normalized" sample size, align and converge corroborating our results.

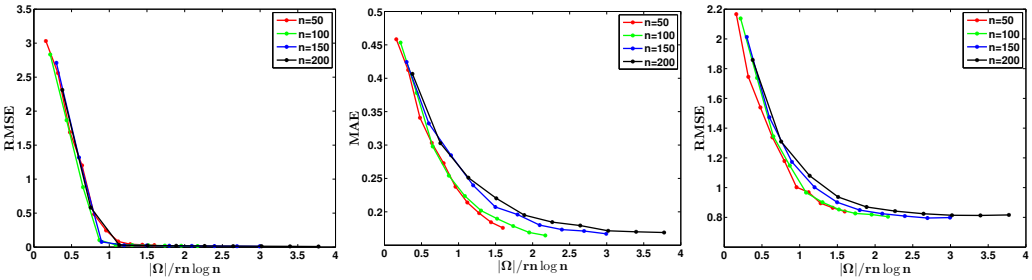



Figure 3.4: Appropriate error metric between observation matrix $Y$, and the MLE estimate from (3.4) $\widehat{Y}$, plotted against "normalized" sample size, when entries of $Y$ are generated from $(a)$ Gaussian, $(b)$ Bernoulli, and $(c)$ binomial distributions

# Chapter 4

# Matrix Completion under General Structures

This chapter [*] presents a unified analysis of matrix completion under general low dimensional structural constraints induced by *any* norm regularization. In a key contribution, two estimators for the general problem of structured matrix completion are proposed, and unified upper bounds on the sample complexity and the recovery error are derived. Further, two intermediate results are derived that are of independent interest: (a) in characterizing the size or complexity of low dimensional subsets in high dimensional ambient space, a certain *partial* complexity measure encountered in the analysis of matrix completion problems is characterized in terms of a well understood complexity measure of Gaussian widths, and (b) it is shown that a useful form of restricted strong convexity (RSC) holds for matrix completion problems under general norm regularization. The proposed framework for general norm regularization is motivated by several non-trivial examples of norm regularized structures, and the special case of the recently proposed spectral $k$-support norm is analysed in detail.

[*]The results in this chapter appear in a conference publication [61]. The coauthors contributed equally.

## 4.1 Introduction

For well–posed estimation in high dimensional problems, including matrix completion, it is imperative that low dimensional structural constraints are imposed on the target (Section 2.2.2). For matrix completion, the special case of low–rank structure has been widely studied and several existing work propose tractable estimators with near–optimal recovery guarantees for (approximate) low–rank matrix completion (see Section 2.2.1 for related work). However, the scope of matrix completion extends for low dimensional structures far beyond simple low–rankness. This chapter presents a unified statistical analysis of matrix completion under general low dimensional structures that are induced by *any* suitable norm regularization. Two norm–regularized matrix completion estimators are studied, the *constrained norm minimizer*, and the *generalized matrix Dantzig selector* (Section 4.2.2). The main results in Theorem 4.3.1a–4.3.1b provide unified upper bounds on the sample complexity and estimation error of these estimators for matrix completion under a general norm regularization. Existing results on matrix completion with low rank or other decomposable structures can be obtained as special cases of Theorem 4.3.1a–4.3.1b.

Such a unified analysis of norm regularized estimators is motivated by recent work on high dimensional estimation using global (sub) Gaussian measurements [33, 10, 150, 13, 153, 23]. A key ingredient in the recovery analysis of high dimensional estimation involves establishing some variation of a certain Restricted Isometry Property (RIP) [28] of the measurement operator. It has been shown that such properties are satisfied by Gaussian and sub–Gaussian measurement operators with high probability. However, as has been noted before by Candes et al. [26], owing to highly localized measurements, such conditions are not satisfied for the matrix completion problem, and the existing results based on global (sub) Gaussian

measurements are not directly applicable. In fact, one of the questions addressed is: given the radically limited measurement model in matrix completion, by how much would the sample complexity of estimation increase beyond the known sample complexity bounds for global (sub) Gaussian measurements? Theorem 4.3.1 provides an upper bound on the sample complexity for matrix completion, which is within a $\log d$ factor over sample complexity bound for estimation under global (sub) Gaussian measurements [33, 13, 23]. While the result was previously known for low rank matrix completion using nuclear norm minimization [115, 89], with a careful use of results from generic chaining [145], it is shown that the $\log d$ factor suffices for structures induced by *any* norm! As a key intermediate result, a useful form of restricted strong convexity (RSC) [116] is derived for the localized measurements encountered in matrix completion over error sets arising from general norm regularization. The result substantially generalizes existing RSC results for matrix completion under the special cases of nuclear norm and decomposable norm regularization [115, 62].

The analysis in this chapter uses tools from generic chaining [145] to characterize the main results (Theorem 4.3.1a–4.3.1b) in terms of the *Gaussian width* (Definition 2.3.4) of certain *error sets*. Gaussian widths provide a powerful geometric characterization for quantifying the complexity of a structured low dimensional subset in a high dimensional ambient space. Numerous tools have been developed in the literature for bounding the Gaussian width of structured sets. A unified characterization of results in terms of Gaussian width has the advantage that this literature can be readily leveraged to derive new recovery guarantees for matrix completion under suitable structural constraints (Section 2.3.2).

In addition to the theoretical elegance of such a unified framework, identifying useful but potentially non–decomposable low dimensional structures is of

significant practical interest. The broad class of structures enforced through symmetric convex bodies and symmetric atomic sets [33] can be analyzed under this paradigm (Section 4.2.1). Such specialized structures can capture the constraints in certain applications better than simple low–rankness. In particular, a non–trivial example of the *spectral $k$–support norm* introduced by McDonald et al. [111] is discussed in detail.

**Contributions**:

- Theorem 4.3.1a–4.3.1b provide unified upper bounds on sample complexity and estimation error for matrix completion estimators using general norm regularization: a substantial generalization of the existing results on matrix completion under structural constraints.
- Theorem 4.3.1a is applied to derive statistical results for the special case of matrix completion under spectral $k$–support norm regularization.
- (a) An intermediate result, Theorem B.3.2 shows that under any norm regularization, a variant of Restricted Strong Convexity (RSC) holds in the matrix completion setting with extremely localized measurements. Further, a certain *partial* measure of complexity of a set is encountered in matrix completion analysis (4.9). (b) Another intermediate result, Theorem 4.3.2 provides bounds on the *partial* complexity measures in terms of a better understood complexity measure of Gaussian width. These intermediate results are of independent interest beyond the scope of this chapter.

## 4.2 Structured Matrix Completion

Denote the ground truth target matrix as $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$; let $d = d_1 + d_2$. In the noisy matrix completion, observations consists of individual entries of $\Theta^*$ observed

through an additive noise channel. In this chapter, notation $G$ and $g$ are reserved to denote a matrix and vector, respectively, with independent standard Gaussian random variables as entries.

**Sub–Gaussian Noise:** Given, a list of independently sampled standard basis $\Omega = \{E_s = e_{i_s} e_{j_s}^\top : i_s \in [d_1], j_s \in [d_2]\}$ with potential duplicates, observations $(y_s)_s \in \mathbb{R}^{|\Omega|}$ are given by:

$$y_s = \langle \Theta^*, E_s \rangle + \xi \eta_s, \text{ for } s = 1, 2, \ldots, |\Omega|, \tag{4.1}$$

where $\eta \in \mathbb{R}^{|\Omega|}$ is the noise vector of independent sub–Gaussian random variables with $\mathbb{E}[\eta_s] = 0$ and $\text{Var}(\eta_s) = 1$, and $\xi^2$ is scaled variance of noise per observation. Also, without loss of generality, assume normalization $\|\Theta^*\|_F = 1$.

**Uniform Sampling:** The entries in $\Omega$ are drawn independently and uniformly:

$$E_s \sim \text{uniform}\{e_i e_j^\top : i \in [d_1], j \in [d_2]\}, \text{ for } E_s \in \Omega. \tag{4.2}$$

Let $\{e_k\}$ be the standard basis of $\mathbb{R}^{|\Omega|}$. Given $\Omega$, define $\mathcal{P}_\Omega : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{|\Omega|}$ as*:

$$\mathcal{P}_\Omega(X) = \textstyle\sum_{s=1}^{|\Omega|} \langle X, E_s \rangle e_s \tag{4.3}$$

**Structural Constraints** For matrix completion with $|\Omega| < d_1 d_2$, low dimensional structural constraints on $\Theta^*$ are necessary for well–posedness. It is assumed that for some low–dimensional *model space* $\mathcal{M}$, $\Theta^* \in \mathcal{M}$ is induced through a surrogate *norm regularizer* $\mathcal{R}(.)$. No further assumptions are made on $\mathcal{R}$ other than it being a norm in $\mathbb{R}^{d_1 \times d_2}$.

**Low Spikiness** As noted earlier for matrix completion under uniform sampling, further restrictions on $\Theta^*$ (beyond low dimensional structure) are required to ensure

*Note that $\mathcal{P}_\Omega$ definition here differs slightly from that in Chapter 3

that the most informative entries of the matrix are observed with high probability (refer Section 2.2.1 for a longer discussion). As in Chapter 3, a restriction on spikiness ratio is used to preclude "spiky" target matrices in the analysis.

**Assumption 4.2.1** (Spikiness Ratio)**.** There exists $\alpha^* > 0$, such that

$$\|\Theta^*\|_\infty = \alpha_{\text{sp}}(\Theta^*)\frac{\|\Theta^*\|_F}{\sqrt{d_1 d_2}} \leq \frac{\alpha^*}{\sqrt{d_1 d_2}}. \qquad \square$$

### 4.2.1 Special Cases and Applications

**Example 1** (Low Rank and Decomposable Norms)**.** *Low–rankness* is the most common structure used in many matrix estimation problems including collaborative filtering, PCA, spectral clustering, etc. Convex estimators for low–rank matrix completion using nuclear norm $\|\Theta\|_*$ regularization has been widely studied statistically [26, 25, 127, 115, 87, 88, 93, 45, 89, 90]. A brief extension of such analysis to general decomposable norms (Definition 2.1.3) — norms $\mathcal{R}$, such that $\forall X, Y \in (\mathcal{M}, \mathcal{M}^\perp), \mathcal{R}(X+Y) = \mathcal{R}(X) + \mathcal{R}(Y)$ — was explored in [62].

**Example 2** (Spectral $k$–support Norm)**.** A non–trivial and significant example of norm regularization that is not decomposable is the *spectral $k$–support* norm recently introduced by McDonald et al. [111]. Spectral $k$–support norm is essentially the vector $k$–support norm (overlapping group lasso penalty over all groups for $k$–sparsity) [11] applied on the singular values $\sigma(\Theta)$ of a matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$.
Without loss of generality, let $\bar{d} = d_1 = d_2$. Let $\mathcal{G}_k = \{g \subseteq [\bar{d}] : |g| \leq k\}$ be the set of all subsets $[\bar{d}]$ of cardinality at most $k$, and let $\mathcal{V}(\mathcal{G}_k) = \{(v_g)_{g \in \mathcal{G}_k} : v_g \in \mathbb{R}^{\bar{d}}, \text{supp}(v_g) \subseteq g\}$. The spectral $k$–support norm is given by:

$$\|\Theta\|_{\text{k-sp}} = \inf_{v \in \mathcal{V}(\mathcal{G}_k)} \Big\{ \sum_{g \in \mathcal{G}_k} \|v_g\|_2 : \sum_{g \in \mathcal{G}_k} v_g = \sigma(\Theta) \Big\}, \tag{4.4}$$

McDonald et al. [111] showed that spectral $k$–support norm is a special case of *cluster norm* [76]. It was further shown that in multi–task learning, wherein the

tasks (columns of $\Theta^*$) are assumed to be clustered into dense groups, the cluster norm provides a trade–off between intra–cluster variance, (inverse) inter–cluster variance, and the norm of the task vectors. These existing work [76, 111] also demonstrate superior empirical performance of cluster norms (and $k$–support norm) over traditional trace norm on bench marked matrix completion and multi–task learning datasets. However, statistical analysis of matrix completion using spectral $k$–support norm regularization has not been previously studied. In Section 4.3.2, the consequence of Theorem 4.3.1 for this non–trivial special case is discussed.

**Example 3** (Additive Decomposition)**.** Elementwise sparsity is a common structure often assumed in high–dimensional estimation problems. However, in matrix completion, elementwise sparsity conflicts with Assumption 4.2.1 (as well as more traditional incoherence assumptions). Indeed, it is easy to see that with high probability most of the $|\Omega| \ll d_1 d_2$ uniformly sampled observations will be zero, and an informed prediction is infeasible. However, elementwise sparse structures can often be modelled within an *additive decomposition* framework, where $\Theta^* = \sum_k \Theta^{(k)}$ and each component matrix $\Theta^{(k)}$ is in turn structured (e.g. low rank+sparse used for robust PCA [24]). In such structures, there is no scope for recovering sparse components outside the observed indices, and it is assumed that: $\Theta^{(k)}$ is sparse $\Rightarrow$ supp$(\Theta^{(k)}) \subseteq \Omega$. These cases can be studied within the proposed framework under additional regularity assumptions that enforces non–spikiness on the superposed matrix. A candidate norm regularizer for such structures is the weighted infimum convolution of individual structure inducing norms [24, 160],

$$\mathcal{R}_w(\Theta) = \inf \Big\{ \sum_k w_k \mathcal{R}_k(\Theta^{(k)}) : \sum_k \Theta^{(k)} = \Theta \Big\}.$$

**Example 4** (Other Applications)**.** Other potential applications including *cut matrices* [140, 33], structures induced by *compact convex sets*, norms inducing *struc-*

*tured sparsity assumptions on the spectrum of* $\Theta^*$, etc. can also be handled under the paradigm of this chapter.

### 4.2.2 Structured Matrix Estimator

Let $\mathcal{R}$ be the norm surrogate for the structural constraints on $\Theta^*$, and $\mathcal{R}^*$ denote its dual norm.

**Constrained Norm Minimizer**

$$\widehat{\Theta}_{\text{cn}} = \underset{\|\Theta\|_\infty \le \frac{\alpha^*}{\sqrt{d_1 d_2}}}{\operatorname{argmin}} \mathcal{R}(\Theta) \qquad \text{s.t. } \|\mathcal{P}_\Omega(\Theta) - y\|_2 \le \lambda_{\text{cn}}. \tag{4.5}$$

**Generalized Matrix Dantzig Selector**

$$\widehat{\Theta}_{\text{ds}} = \underset{\|\Theta\|_\infty \le \frac{\alpha^*}{\sqrt{d_1 d_2}}}{\operatorname{argmin}} \mathcal{R}(\Theta) \qquad \text{s.t. } \frac{\sqrt{d_1 d_2}}{|\Omega|} \mathcal{R}^* \mathcal{P}_\Omega^*(\mathcal{P}_\Omega(\Theta) - y) \le \lambda_{\text{ds}}, \tag{4.6}$$

where $\mathcal{P}_\Omega^* : \mathbb{R}^\Omega \to \mathbb{R}^{d_1 \times d_2}$ is the linear adjoint of $\mathcal{P}_\Omega$, i.e. $\langle \mathcal{P}_\Omega(X), y \rangle = \langle X, \mathcal{P}_\Omega^*(y) \rangle$.

Theorem 4.3.1a–4.3.1b give consistency results for (4.5) and (4.6), respectively, under certain conditions on the parameters $\lambda_{\text{cn}} > 0$, $\lambda_{\text{ds}} > 0$, and $\alpha^* > 1$. In particular, these conditions assume knowledge of tight bounds on noise variance $\xi^2$ and spikiness ratio $\alpha_{\text{sp}}(\Theta^*)$. In practice, typically $\xi$ and $\alpha_{\text{sp}}(\Theta^*)$ are unknown and the parameters are tuned by validating on held out data.

## 4.3 Main Results

Define the following "restricted" *error cone* and its subset:

$$\mathcal{T}_\mathcal{R} = \mathcal{T}_\mathcal{R}(\Theta^*) = \text{cone}\{\Delta : \mathcal{R}(\Theta^* + \Delta) \le \mathcal{R}(\Theta^*)\}, \text{and } \mathcal{E}_\mathcal{R} = \mathcal{T}_\mathcal{R} \cap \mathbb{S}^{d_1 d_2 - 1}, \tag{4.7}$$

where recall $\mathbb{S}^{d_1 d_2 - 1} = \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F = 1\}$.

Let $\widehat{\Theta}_{\text{cn}}$ and $\widehat{\Theta}_{\text{ds}}$ be the estimates from (4.5) and (4.6), respectively. If $\lambda_{\text{cn}}$ and $\lambda_{\text{ds}}$ are chosen such that $\Theta^*$ belongs to the feasible sets in (4.5) and (4.6), respectively, then the error matrices $\widehat{\Delta}_{\text{cn}} = \widehat{\Theta}_{\text{cn}} - \Theta^*$ and $\widehat{\Delta}_{\text{ds}} = \widehat{\Theta}_{\text{ds}} - \Theta^*$ are contained in $\mathcal{T}_{\mathcal{R}}$.

Recall definition of Gaussian width $w_G$ from (2.14). Further, define the following norm compatibility constant.

**Definition 4.3.1** (Norm Compatibility Constant [116])**.** The compatibility constant of a norm $\mathcal{R} : \mathcal{V} \to \mathbb{R}$ under a closed convex cone $\mathcal{C} \subset \mathcal{V}$ is defined as follows:

$$\Psi_{\mathcal{R}}(\mathcal{C}) = \sup_{X \in \mathcal{C} \setminus \{0\}} \frac{\mathcal{R}(X)}{\|X\|_F}. \tag{4.8}$$

**Theorem 4.3.1a** (Constrained Norm Minimizer)**.** *Under the problem setup in Section 4.2, let $\widehat{\Theta}_{cn} = \Theta^* + \widehat{\Delta}_{cn}$ be the estimate from (4.5) with $\lambda_{cn} = 2\xi\sqrt{|\Omega|}$. For large enough $c_0$, if $|\Omega| > c_0^2 w_G^2(\mathcal{E}_{\mathcal{R}}) \log d$, then there exists an RSC parameter $\kappa_{c_0} > 0$ with $\kappa_{c_0} \approx 1 - o\left(\frac{1}{\sqrt{\log d}}\right)$, and constants $c_1$ and $c_2$ such that, with probability greater than $1 - \exp\left(-c_1 w_G^2(\mathcal{E}_{\mathcal{R}})\right) - 2\exp\left(-c_2 w_G^2(\mathcal{E}_{\mathcal{R}}) \log d\right)$,*

$$\frac{1}{d_1 d_2}\|\widehat{\Delta}_{cn}\|_F^2 \leq 4 \max\left\{\frac{\xi^2}{\kappa_{c_0}}, \frac{\alpha^{*2}}{d_1 d_2}\sqrt{\frac{c_0^2 w_G^2(\mathcal{E}_{\mathcal{R}}) \log d}{|\Omega|}}\right\}.$$

**Theorem 4.3.1b** (Matrix Dantzig Selector)**.** *Under the problem setup in Section 4.2, let $\widehat{\Theta}_{ds} = \Theta^* + \widehat{\Delta}_{ds}$ be the estimate from (4.6) with $\lambda_{ds} \geq 2\xi\frac{\sqrt{d_1 d_2}}{|\Omega|}\mathcal{R}^*\mathcal{P}_{\Omega}^*(\eta)$. For large enough $c_0$, if $|\Omega| > c_0^2 w_G^2(\mathcal{E}_{\mathcal{R}}) \log d$, then there exists an RSC parameter $\kappa_{c_0} > 0$ with $\kappa_{c_0} \approx 1 - o\left(\frac{1}{\sqrt{\log d}}\right)$, and a constant $c_1$ such that, with probability greater than $1 - \exp(-c_1 w_G^2(\mathcal{E}_{\mathcal{R}}))$,*

$$\frac{1}{d_1 d_2}\|\widehat{\Delta}_{ds}\|_F^2 \leq 16 \max\left\{\frac{\lambda_{ds}^2 \Psi_{\mathcal{R}}^2(\mathcal{T}_{\mathcal{R}})}{\kappa_{c_0}^2}, \frac{\alpha^{*2}}{d_1 d_2}\sqrt{\frac{c_0^2 w_G^2(\mathcal{E}_{\mathcal{R}}) \log d}{|\Omega|}}\right\}.$$

**Remarks:**

1. If $\mathcal{R}(\Theta) = \|\Theta\|_*$ and $\text{rank}(\Theta^*) = r$, then $w_G^2(\mathcal{E}_\mathcal{R}) \leq 3dr$, $\Psi_\mathcal{R}(\mathcal{T}_\mathcal{R}) \leq 8\sqrt{r}$ and $\frac{\sqrt{d_1 d_2}}{|\Omega|}\|\mathcal{P}^*_\Omega(\eta)\|_2 \leq 2\sqrt{\frac{d \log d}{|\Omega|}}$ w.h.p [33, 53, 115]. Using these bounds in Theorem 4.3.1b recovers near–optimal results for low rank matrix completion under spikiness assumption [115].
2. For both estimators, upper bound on sample complexity is dominated by the square of Gaussian width which is often considered the *effective dimension* of a subset in high dimensional space and plays a key role in high dimensional estimation under Gaussian measurement ensembles. The results show that, independent of $\mathcal{R}(.)$, the upper bound on sample complexity for consistent matrix completion with highly localized measurements is within a $\log d$ factor of the known sample complexity of $\sim w_G^2(\mathcal{E}_\mathcal{R})$ for estimation from Gaussian measurements [13, 33, 153, 23].
3. First term in estimation error bounds in Theorem 4.3.1a–4.3.1b scales with $\xi^2$ which is the per observation noise variance. The second term is an upper bound on error that arises due to unidentifiability of $\Theta^*$ within a certain radius under the spikiness constraints [115]; in contrast [25] show exact recovery when $\xi = 0$ using more stringent matrix incoherence conditions.
4. Bound on $\widehat{\Delta}_{\text{cn}}$ from Theorem 4.3.1a is comparable to the result by Candés et al. [25] for low rank matrix completion under non–low–noise regime, where the first term dominates, and those of [33, 150] for high dimensional estimation under Gaussian measurements. With a bound on $w_G^2(\mathcal{E}_\mathcal{R})$, it is easy to specialize this result for new structural constraints. However, this bound is potentially loose and asymptotically converges to a constant error proportional to the noise variance $\xi^2$.
5. The estimation error bound in Theorem 4.3.1b is typically sharper than that in

Theorem 4.3.1a. However, for specific structures, using application of Theorem 4.3.1b requires additional bounds on $\mathcal{R}^*\mathcal{P}^*_\Omega(\eta)$ and $\Psi_\mathcal{R}(\mathcal{T}_\mathcal{R})$ besides $w^2_G(\mathcal{E}_\mathcal{R})$.

### 4.3.1 Partial Complexity Measures

Recall that $G \in \mathbb{R}^{d_1 \times d_2}$, and $g \in \mathbb{R}^{|\Omega|}$ denotes a random matrix and vector respectively with each entry sampled independently from standard normal distribution, and $w_G(S) = \mathbb{E}\sup_{X\in S}\langle X, G\rangle$ (Definition 2.3.4).

**Definition 4.3.2** (Partial Complexity Measures)**.** Given a randomly sampled $\Omega = \{E_s \in \mathbb{R}^{d_1\times d_2}\}$, and a centered random vector $\eta \in \mathbb{R}^{|\Omega|}$, the *partial $\eta$–complexity measure* of $S$ is given by:

$$w_{\Omega,\eta}(S) = \mathbb{E}_{\Omega,\eta} \sup_{X\in S-S} \langle X, \mathcal{P}^*_\Omega(\eta)\rangle. \tag{4.9}$$

Special cases of $\eta$ being a vector of standard Gaussian $g$, or standard Rademacher $\epsilon$ (i.e. $\epsilon_s \in \{-1, 1\}$ w.p. $1/2$) variables, are of particular interest.

**Note:** For symmetric $\eta$, like $g$ and $\epsilon$, $w_{\Omega,\eta}(S) = 2\mathbb{E}_{\Omega,\eta}\sup_{X\in S}\langle X, \mathcal{P}^*_\Omega(\eta)\rangle$, and the later expression will be used interchangeably ignoring the constant term. □

**Theorem 4.3.2** (Partial Gaussian Complexity)**.** *Let $S \subseteq \mathbb{B}^{d_1 d_2}$ with non–empty interior, and let $\Omega$ be sampled according to* (4.2)*. $\exists$ universal constants $k_1$, $k_2$, $K_1$ and $K_2$ such that:*

$$\begin{aligned} w_{\Omega,g}(S) &\le k_1\sqrt{\frac{|\Omega|}{d_1 d_2}}w_G(S) + k_2\sqrt{\mathbb{E}_\Omega \sup_{X,Y\in S}\|\mathcal{P}_\Omega(X-Y)\|_2^2} \\ w_{\Omega,g}(S) &\le K_1\sqrt{\frac{|\Omega|}{d_1 d_2}}w_G(S) + K_2 \sup_{X,Y\in S}\|X-Y\|_\infty. \end{aligned} \tag{4.10}$$

*Also, for centered i.i.d. sub–Gaussian vector $\eta \in \mathbb{R}^{|\Omega|}$, $\exists$ constant $K_3$ s.t. $w_{\Omega,\eta}(S) \le K_3 w_{\Omega,g}(S)$.*

**Note:** For $\Omega \subsetneq [d_1] \times [d_2]$, the second term in (4.10) is a consequence of the localized measurements.

### 4.3.2 Spectral $k$–Support Norm

Spectral $k$–support norm was introduced in Section 4.2.1. The estimators from (4.5) and (4.6) for spectral $k$–support norm can be efficiently solved via proximal methods [111]. The analysis for upper bounding the Gaussian width of the descent cone for the vector $k$–support norm by [129] is extended to the case of spectral $k$–support norm. WLOG let $d_1 = d_2 = \bar{d}$. Let $\sigma^* \in \mathbb{R}^{\bar{d}}$ be the vector of singular values of $\Theta^*$ sorted in non–ascending order. Let $r \in \{0, 1, 2, \ldots, k-1\}$ be the unique integer satisfying: $\sigma^*_{k-r-1} > \frac{1}{r+1}\sum_{i=k-r}^{p} \sigma^*_i \geq \sigma^*_{k-r}$. Denote $I_2 = \{1, 2, \ldots, k-r-1\}$ and $I_1 = \{k-r, k-r+1, \ldots, s\}$. Finally, for $I \subseteq [\bar{d}]$, $(\sigma^*_I)_i = 0 \; \forall i \in I^c$, and $(\sigma^*_I)_i = \sigma^*_i \; \forall i \in I$.

**Lemma 4.3.3.** *If rank of $\Theta^*$ is $s$ and $\mathcal{E}_{\mathcal{R}}$ is the error set for $\mathcal{R}(\Theta) = \|\Theta\|_{k\text{–}sp}$, then*

$$w_G^2(\mathcal{E}_{\mathcal{R}}) \leq s(2\bar{d} - s) + \Big(\frac{(r+1)^2\|\sigma^*_{I_2}\|_2^2}{\|\sigma^*_{I_1}\|_1^2} + |I_1|\Big)(2\bar{d} - s).$$

Proof of the above lemma is provided in the appendix. Lemma 4.3.3 can be combined with Theorem 4.3.1a to obtain recovery guarantees for matrix completion under spectral $k$–support norm.

## 4.4 Discussions and Comparisons to Related Work

**Sample Complexity:** For consistent recovery in high dimensional convex estimation, it is desirable that the descent cone at the target parameter $\Theta^*$ is "small" relative to the feasible set (enforced by the observations) of the estimator. Thus, it is not surprising that the sample complexity and estimation error bounds of an estimator depends on a measure of complexity/size of the error cone at $\Theta^*$. Results in this chapter are largely characterized in terms of a widely used complexity measure of Gaussian width $w_G(.)$, and can be compared with the literature on estimation from Gaussian measurements.

**Error Bounds:** Theorem 4.3.1a provides estimation error bounds that depends only on the Gaussian width of the descent cone. In non–low–noise regime, this result is comparable to analogous results of constrained norm minimization [24, 33, 150]. However, this bound is potentially loose owing to data–fit term using squared loss rather than a matching dual norm, and asymptotically converges to a constant error proportional to the noise variance $\xi^2$.

A tighter analysis on the estimation error can be obtained for the matrix Dantzig selector (4.6) from Theorem 4.3.1b. However, application of Theorem 4.3.1b requires computing high probability upper bound on $\mathcal{R}^*\mathcal{P}^*_\Omega(\eta)$. The literature on norms of random matrices [50, 103, 152, 149] can be exploited in computing such bounds. Beside, in special cases: if $\mathcal{R}(.) \geq K\|.\|_*$, then $K\mathcal{R}^*(.) \leq \|.\|_{\text{op}}$ can be used to obtain asymptotically consistent results.

Finally, under near zero–noise, the second term in the results of Theorem 4.3.1 dominates. In this low noise setting, the bounds are weaker than that of [24, 88] owing to the relaxation of stronger incoherence assumption. The closest related work is the result on consistency of matrix completion under decomposable norm regularization briefly discussed in Section 3.3.2 (refer [62]). Results in this chapter are a strict generalization to general norm regularized (not necessarily decomposable) matrix completion. Non–trivial examples of applications where structures enforced by such non–decomposable norms are of interest are discussed in Section 4.2.1. Further, in contrast to the results derived in this chapter that are based on Gaussian width, the RSC parameter in [62] depends on a modified complexity measure $\kappa_{\mathcal{R}}(d, |\Omega|)$ (3.5). An advantage of results based on Gaussian width is that, application of Theorem 4.3.1 for special cases can greatly benefit from the numerous tools in the literature for the computation of $w_G(.)$.

# Chapter 5

# Collective Matrix Completion

In this chapter* , the collective matrix completion problem of jointly recovering a collection of matrices with shared structure from partial (and potentially noisy) observations is addressed. The problem is studied under a joint low–rank structure, wherein each component matrix is low–rank and the latent space of the low rank factors corresponding to each entity is shared across the entire collection. A rigorous algebra for the collective–matrix structure is developed, and a convex estimate for solving the collective matrix completion problem is proposed. The main result in this chapter provides the first non–trivial theoretical guarantees for consistency of collective matrix completion. It is shown that, for a subset of entity–relationship structures defining a collection of matrices (see Assumption 5.3.3), with high probability, the proposed estimator exactly recovers the true matrices whenever certain sample complexity requirements (dictated by Theorem 5.4.1) are met. A scalable approximate algorithm is proposed to solve the proposed convex program, and the results are corroborated using simulated and real data experiments.

*The results in this chapter appear in a conference publication [63]. The coauthors contributed equally.

## 5.1 Introduction

In practical applications, data commonly arise in the form of multiple matrices sharing correlated information. For example, in e–commerce applications, data containing user preferences in multiple domains such as news, ads, etc., and explicit user/item feature information such as demographics, social network, text description, etc., are made available in the form of a collection of matrices that are coupled through the common set of users/items. In such scenarios, the shared structure among the matrices can be leveraged for better predictions.

In collective matrix completion problem setup, there are $K \geq 2$ types of *entities*, and data consists of a collection of $V \geq 1$ interaction matrices called *views*. Each view (component matrix) is an affinity relation between a pair of entity types, e.g. user–movie rating matrix, item–features matrix, etc. The task in collective matrix completion is to simultaneously complete one or more partially observed views by potentially leveraging data from the entire collection. To this end, a joint low–rank structure is commonly assumed, wherein each view is individually a low–rank matrix, and the low dimensional factors for each entity type are shared across all the views involving that entity type, i.e., for all entity types $k$, there is a low dimensional factor representation $U_k$, and each view representing the interaction between entity types $k_1$ and $k_2$ is a low rank matrix given by $U_{k_1} U_{k_2}^\top$. One could trivially address collective matrix completion through separate low–rank matrix completions; however, estimators that leverage shared structure are more attractive as they can potentially alleviate two major problems that arise in standard matrix completion:

1. *Data Sparsity*: The algorithms and estimators proposed for traditional matrix completion setting, fail under extremely sparse data. In a collective matrix setting, this data sparsity issue can be mitigated by transferring information from one or more related views. For example, in a multiple recommendation

system where the data consists of ratings of a set of users for a subset of items in multiple domains, say movies, books and TV shows. It is reasonable to assume that user interests are related across the domains, and leveraging this shared information can help mitigate data sparsity.

2. *Cold Start*: The existing estimators for matrix completion cannot handle an entire missing row or column in a matrix. This problem, often referred as *cold start*, can be overcome in a collective matrix setting if the entity corresponding to the missing row/column in a particular view has data in other views sharing the entity. For example in recommendation systems with access to user's explicit features, recommendation for new user with no known rating can be provided by jointly factorizing the user–feature and user–item ratings matrices.

In this chapter a convex estimator is proposed for jointly estimating the a collection of matrices under joint low rank structure and provide first non–trivial theoretical guarantees for a large subset of collective matrix structures. Further, a vanilla adaption of the Singular Value Thresholding (SVT) algorithm for the proposed estimate [20, 22, 147] requires computing the complete SVD of a very large sized blockwise concatenated matrix (5.2) within each iteration and thus is not scalable to large datasets. To address sclability, an approximate algorithm is proposed by adapting Hazan's algorithm [67] for the proposed convex program.

A closely related work is the paper by Bouchard et al. [20], where the authors propose the first convex estimator for collective matrix completion without addressing the recovery guarantees of the estimator. Besides the convex estimator, related work for collective matrix completion includes various non–convex estimators and probabilistic models. A seminal paper on low rank collective matrix factorization is the work by Singh et al. [138], in which the views are parameterized

by the shared latent factor representation. The latent factors are learnt by minimizing a regularized loss function over the estimates. A Bayesian model for collective matrix factorization was also proposed by the same authors [136, 137]. Various algorithms and models for learning collective matrices are summarized in the thesis of Singh [136]. Collective matrix factorization is also related to applications involving multi–task learning and tensor factorization [106, 102, 8, 164, 166]. However, this line of work involves complex non–convex optimizations and is difficult to provide rigorous statistical analysis for.

**Contributions**:

- A convex program is proposed for the task of collective–matrix completion building on a rigorous algebra developed for representation of collective matrix structures (Section 5.2 and 5.3).
- The main result quantified first non–trivial sample complexity bounds for consistent collective matrix completion. It is shown that for a subset of entity–relationship structures of the collective matrix, with high probability, the proposed estimator exactly recovers the true matrices whenever the sample complexity satisfies $\forall k$, $|\Omega_k| \sim O(n_k R \text{polylog} N)$ (Section 5.4.1).
- A scalable approximate algorithm is proposed for the optimization problem (Section 5.4.3) and the results are corroborated through experiments on simulated and real life datasets (Section 5.5).

## 5.2 Collective–Matrix Structure

A collective–matrix structure denoted using script letters, $\mathcal{X}$, $\mathcal{M}$, etc, is used to represent a collection of pairwise affinity relations among a set of $K$ types of *entities*. A collective–matrix $\mathcal{X}$ consists of a list of $V$ matrices $\mathcal{X} = [X_v]_{v=1}^{V} = [X_v :$

$v = 1, 2, \ldots, V]$, wherein each component matrix $X_v$, called a *view*, is the affinity matrix between a pair of entity types $r_v$ (along rows) and $c_v$ (along columns). In this chapter, the collective matrices are restricted to static undirected affinity relations, under which for a given pair of entity types $k_1, k_2 \in \{1, 2, \ldots K\}$, there is at most one affinity relation, $X_v$, defined between $k_1$ and $k_2$.

The entity–relationship structure defining a collective–matrix, is represented by an undirected graph $\mathcal{G}$, with nodes denoting the $K$ entity types, and an edge between nodes $k_1$ and $k_2$ implying that a view $X_v$ with either $(r_v = k_1, c_v = k_2)$ or $(r_v = k_2, c_v = k_1)$ exists in the collective matrix. Without loss of generality, let the graph $\mathcal{G}$ form a single connected component, if not, each connected component could be handled separately. An illustration of a collective matrix structure $\mathcal{X}$ and its entity–relationship graph $\mathcal{G}$ is given in Figure 5.1 $(a)$–$(b)$.

Let $n_k$ for $k = 1, 2, \ldots, K$ denote the number of instances of the $k^{\text{th}}$ entity type, and let $N = \sum_k n_k$. Then $\forall v$, $X_v \in \mathbb{R}^{n_{r_v} \times n_{c_v}}$ and collective–matrices defined by common entity–relationship graph $\mathcal{G}$, belong to the following vector space:

$$\mathfrak{X} = \mathbb{R}^{n_{r_1} \times n_{c_1}} \times \mathbb{R}^{n_{r_2} \times n_{c_2}} \times \ldots \times \mathbb{R}^{n_{r_V} \times n_{c_V}}$$

Finally, for $v \in \{1, 2, \ldots, V\}$, $\mathcal{I}(v) = \{(i, j) : i \in [n_{r_v}], j \in [n_{c_v}]\} = [n_{r_v}] \times [n_{c_v}]$ denotes the set of indices representing the elements in view $v$.

### 5.2.1 Equivalent Representations

For mathematical convenience, two alternate (equivalent) representations are introduced for the collective-matrix structure. These representations will be used interchangeably.

1. **Entity Matrix Set Representation:** A collective–matrix $\mathcal{X}$, can be equivalently represented as a set of $K$ matrices $\mathbb{X} = [\mathbb{X}_k]_{k=1}^K$, such that $\mathbb{X}_k$ is a

matrix formed by concatenating views involving the entity type $k$, i.e. the views with either $r_v = k$ or $c_v = k$. Let $m_k = \sum_{v=1}^{V} n_{c_v} \mathbb{1}_{(r_v=k)} + n_{r_v} \mathbb{1}_{(c_v=k)}$, where $\mathbb{1}_E$ is an indicator variable for event $E$, then:

$$\mathbb{X}_k := \text{hcat}\{[X_v \mathbb{1}_{(r_v=k)}, X_v^\top \mathbb{1}_{(c_v=k)}]_{v=1}^{V}\} \in \mathbb{R}^{n_k \times m_k}. \tag{5.1}$$

2. **Block Matrix Representation:** Collective–matrices can also be represented as blocks in a symmetric matrix of size $N \times N$, where $N = \sum_k n_k$ [20]. Consider a symmetric matrix $Z \in \mathbb{S}^N$ consisting of $K \times K$ blocks, wherein the $(k_1, k_2)$ block, denoted as $Z[k_1, k_2]$, is of dimension $n_{k_1} \times n_{k_2}$. The block matrix representation of $\mathcal{X}$ is denoted as $\mathcal{B}(\mathcal{X}) \in \mathbb{S}^N$, such that:

$$\mathcal{B}(\mathcal{X})[k_1, k_2] = \begin{cases} X_v & \text{if } \exists v, \text{ s.t. } r_v = k_1, c_v = k_2 \\ X_v^\top & \text{if } \exists v, \text{ s.t. } r_v = k_2, c_v = k_1 \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

Define a projection operator $P_v : S^N \to \mathbb{R}^{n_{r_v} \times n_{c_v}}$, such that $P_v(Z) = Z[r_v, c_v]$. Alternatively, for any $Z \in \mathbb{S}^N$, $\mathcal{Z} = [P_v(Z)]_{v=1}^{V} \in \mathcal{X}$

These alternate representations for collective–matrix structure are illustrated in Figure 5.1 $(c)$ and $(d)$, respectively.

### 5.2.2 Collective–Matrix Algebra

**Collective–Matrix Inner Product** $\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{v=1}^{V} \langle X_v, Y_v \rangle = \frac{1}{2} \sum_{k=1}^{K} \langle \mathbb{X}_k, \mathbb{Y}_k \rangle = \frac{1}{2} \langle \mathcal{B}(\mathcal{X}), \mathcal{B}(\mathcal{Y}) \rangle$.

**Standard Orthonormal Basis** The *standard orthonormal basis* for $\mathfrak{X}$ is given by $\{\mathcal{E}^{(v,i_v,j_v)} : v \in [V], (i_v, j_v) \in \mathcal{I}(v)\}$, where $\mathcal{E}^{(v,i_v,j_v)} \in \mathfrak{X}$ has a value of 1 in the $(i_v, j_v)^{\text{th}}$ element of view $v$, and 0 everywhere else. Recall that $\mathcal{I}(v) = [n_{r_v}] \times [n_{c_v}]$.

**Collective–Matrix Frobenius Norm** $\|\mathcal{X}\|_F = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$.

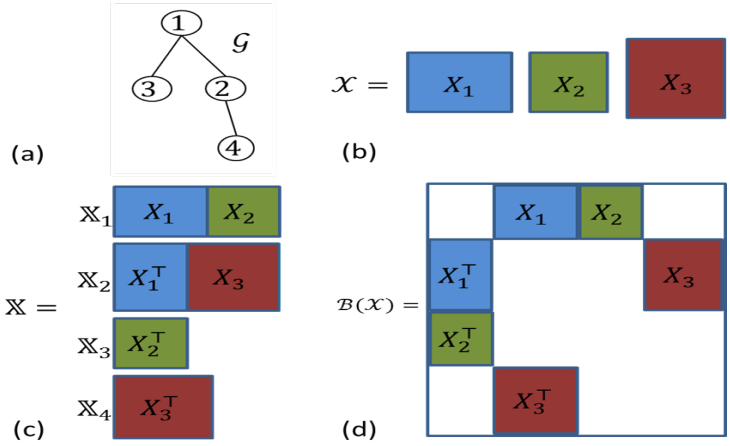



Figure 5.1: An illustration of the various collective–matrix representations described in Section 5.2

**Joint Factorization and Collective–Matrix Rank** A collective-matrix $\mathcal{X} \in \mathfrak{X}$ is said to possess an $R$–*dimensional joint factorization* structure, if there exists a set of factors $\{U_k \in \mathbb{R}^{n_k \times R}\}_{k=1}^{K}$, such that $\forall v,\ X_v = U_{r_v} U_{c_v}^\top$. The set of collective–matrices in $\mathfrak{X}$ that have a joint factorization structure of dimension $R < \infty$ is denoted by $\bar{\mathfrak{X}} \subseteq \mathfrak{X}$. For $\mathcal{X} \in \bar{\mathfrak{X}}$, the *collective–matrix rank* is defined as the minimum value of $R$ such that an $R$-dimensional joint factorization exists for $\mathcal{X}$.

### 5.2.3 Atomic Decomposition of Collective–Matrices

Consider the following atomic set of rank–1 collective–matrices.

$$\mathscr{A} = \text{ext}\big(\text{conv}\{[P_v(uu^\top)]_{v=1}^{V} : u \in \mathbb{R}^N, \|u\|_2 = 1\}\big), \tag{5.3}$$

where conv(), and ext() return the convex hull, and extreme points of a set, respectively. Recall that $N = \sum_k n_k$, and $P_v : \mathbb{S}^N \to \mathbb{R}^{n_{r_v} \times n_{c_v}}$ extracts the block corresponding to the view $v$ in an $N \times N$ symmetric matrix. From the block matrix representation of collective matrices (5.2), it can be noted that $\mathfrak{X} = \text{aff}(\mathscr{A})$.

**Proposition 5.2.1.** *A collective–matrix has a joint factorization structure if and only if it belongs to the conic hull of $\mathscr{A}$, i.e. $\bar{\mathfrak{X}} = cone(\mathscr{A})$.*

*Proof:* $\mathcal{X}$ has a joint factorization $\iff$ $\mathcal{X} = P_v(UU^\top)$ for $U \in \mathbb{R}^{N\times k}$ $\iff$ $\mathcal{X} = \sum_k \sigma_k P_v(u_k u_k^\top)$, where $\|u_k\|_2 = 1, \sigma_k \geq 0$ $\iff$ $\mathcal{X} \in \text{cone}(\mathscr{A})$. $\square$

The following functions are defined on $\mathscr{A}$:

1. The gauge function of $\mathscr{A}$ will henceforth be referred as the **Collective–Matrix Atomic Norm**:

$$\|\mathcal{X}\|_{\mathscr{A}} := \inf\{t > 0 : \mathcal{X} \in t \cdot \text{conv}(\mathscr{A})\}. \tag{5.4}$$

   By convention, $\|\mathcal{X}\|_{\mathscr{A}} = \infty$ if $\mathcal{X} \in \mathfrak{X} \setminus \bar{\mathfrak{X}}$.
2. The support function of $\mathscr{A}$:

$$\|\mathcal{X}\|^*_{\mathscr{A}} := \sup\{\langle \mathcal{X}, \mathcal{A}\rangle : \mathcal{A} \in \mathscr{A}\}. \tag{5.5}$$

**Remarks**

1. $\|\mathcal{X}\|_{\mathscr{A}}$ is not always a norm. It is a norm if $\mathscr{A}$ is centrally symmetric, i.e. if $\mathcal{A} \in \mathscr{A} \Leftrightarrow -\mathcal{A} \in \mathscr{A}$.
2. However, $\|\mathcal{X}\|_{\mathscr{A}}$ is always a convex function and exhibits many norm–like properties. $\forall \mathcal{X} \in \mathfrak{X}$, $\|\mathcal{X}\|_{\mathscr{A}} \geq 0$ and $\|\mathcal{X}\|_{\mathscr{A}} = 0$ iff $\mathcal{X} = 0$ (positivity); $\forall a \geq 0$, $\|a\mathcal{X}\|_{\mathscr{A}} = a\|\mathcal{X}\|_{\mathscr{A}}$ (positive homogeneity); and $\|\mathcal{X} + \mathcal{Y}\|_{\mathscr{A}} \leq \|\mathcal{X}\|_{\mathscr{A}} + \|\mathcal{Y}\|_{\mathscr{A}}$ (triangle inequality). The only property of norm $\|.\|_{\mathscr{A}}$ does not satisfy for general $\mathscr{A}$ is that of absolute homogeneity, specifically, in general it is possible that $\|\mathcal{X}\|_{\mathscr{A}} < \infty$ and $\|-\mathcal{X}\|_{\mathscr{A}} = \infty \neq \|\mathcal{X}\|_{\mathscr{A}}$.
3. If $\|\mathcal{X}\|_{\mathscr{A}}$ is a norm, then $\|\mathcal{X}\|^*_{\mathscr{A}}$ is its dual norm.

#### 5.2.3.1 Primal Dual representation

For all $\mathcal{X} \in \bar{\mathfrak{X}}$, $\|\mathcal{X}\|_{\mathscr{A}} < \infty$, and the atomic norm defined in (5.4), can be equivalently defined using the following primal and dual optimization problems.

$$(P) \quad \|\mathcal{X}\|_{\mathscr{A}} = \min_{\{\lambda_r \geq 0\}} \textstyle\sum_r \lambda_r \;\; \text{s.t.} \;\; \textstyle\sum_r \lambda_r \mathcal{A}_r = \mathcal{X} \tag{5.6}$$

$$(D) \quad \|\mathcal{X}\|_{\mathscr{A}} = \max_{\mathcal{Y} \in \mathfrak{X}} \langle \mathcal{X}, \mathcal{Y} \rangle \;\; \text{s.t.} \;\; \|\mathcal{Y}\|_{\mathscr{A}}^* \leq 1 \tag{5.7}$$

**Proposition 5.2.2.** *$\forall \mathcal{X} \in \bar{\mathfrak{X}}$, convex programs $(P)$ and $(D)$ defined above are equivalent to:*

$$(P) \quad \|\mathcal{X}\|_{\mathscr{A}} = \min_{Z \in \mathbb{S}^N} tr(Z) \quad s.t. \; P_v(Z) = X_v \forall \; v,$$

$$(D) \quad \|\mathcal{X}\|_{\mathscr{A}} = \max_{\mathcal{Y} \in \mathfrak{X}} \langle \mathcal{X}, \mathcal{Y} \rangle \quad s.t. \; \frac{1}{2}\mathcal{B}(\mathcal{Y}) \preccurlyeq \mathbb{I},$$

*where recall $\mathcal{B}(.)$ and $P_v$ from (5.2).* □

Finally, define the following set of **"sign" collective–matrices**:

$$\mathscr{E}(\mathcal{X}) = \{\mathcal{E} \in \mathfrak{X} : \|\mathcal{X}\|_{\mathscr{A}} = \langle \mathcal{E}, \mathcal{X} \rangle, \|\mathcal{E}\|_{\mathscr{A}}^* = 1\} \tag{5.8}$$

## 5.3 Convex Collective–Matrix Completion

Denote the ground truth collective–matrix as $\mathcal{M} \in \bar{\mathfrak{X}}$. A partially observed setting is considered in which only a subset of the entries of $\mathcal{M}$ are observed under a random sampling model. Denote the set of observed entries as $\Omega = \{(v_s, i_s, j_s) : (i_s, j_s) \in \mathcal{I}(v_s), s = 1, 2, \ldots, |\Omega|\}$. For conciseness, denote the standard basis corresponding to the entries in $\Omega$ as $\mathcal{E}^{(s)} = \mathcal{E}^{(v_s, i_s, j_s)}$, for $s = 1, 2, \ldots, |\Omega|$. Consider two observation models:

1. Noise–free Model: $\mathcal{M}$ is observed on $\Omega$ without any noise, $\forall s, y_s = \langle \mathcal{M}, \mathcal{E}^{(s)} \rangle$.

2. Additive Noise Model: The values of $\mathcal{M}$ on $\Omega$ are observed with additive random noise, i.e. $\forall s, y_s = \langle \mathcal{M}, \mathcal{E}^{(s)} \rangle + \eta_s$.

The task in collective–matrix completion is to recover $\mathcal{M}$ from $\{y_s\}_{s=1}^{|\Omega|}$. Given $\Omega$ and $\mathcal{X} \in \mathfrak{X}$, define the following projection:

$$\mathcal{P}_\Omega(\mathcal{X}) = \sum_{v=1}^{V} \sum_{(i,j)\in\mathcal{J}(v)} \mathbb{1}_{[(v,i,j)\in\Omega]} \langle \mathcal{X}, \mathcal{E}^{(v,i,j)} \rangle \mathcal{E}^{(v,i,j)}. \tag{5.9}$$

### 5.3.1 Assumptions

**Assumption 5.3.1** ($R$–dimensional joint factorization)**.** The ground truth collective–matrix $\mathcal{M}$ is assumed to have a collective–matrix rank of $R \ll N$, i.e. $\exists \{U_k \in \mathbb{R}^{n_k \times R}\}$, such that $\forall v$, $M_v = U_{r_v} U_{c_v}^\top$. □

Analogous to matrices, $\forall \mathcal{X} \in \bar{\mathfrak{X}}$, define the following:

$$\begin{aligned} T(\mathcal{X}) =& \text{aff}\{\mathcal{Y} \in \bar{\mathfrak{X}} : \forall\, v, \text{rowSpan}(\mathbb{Y}_{r_v}) \subseteq \text{rowSpan}(\mathbb{X}_{r_v}) \\ & \textit{or } \text{rowSpan}(\mathbb{Y}_{c_v}) \subseteq \text{rowSpan}(\mathbb{X}_{c_v})\} \qquad (5.10) \\ T^\perp(\mathcal{X}) =& \{\mathcal{Y} \in \bar{\mathfrak{X}} : \forall\, v, \text{rowSpan}(Y_v) \perp \text{rowSpan}(X_v) \\ & \textit{and } \text{colSpan}(Y_v) \perp \text{colSpan}(X_v)\} \qquad (5.11) \end{aligned}$$

Note: the entity matrix set representation (5.1) is used in (5.10).

For conciseness, $T(\mathcal{M})$ and $T^\perp(\mathcal{M})$ will henceforth be denoted simply as $T$ and $T^\perp$, respectively. Let $\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$ be projection operators onto $T$ (or $T(\mathcal{M})$), and $T^\perp$ (or $T^\perp(\mathcal{M})$), respectively.

**Lemma 5.3.1.** $\forall \mathcal{X} \in \bar{\mathfrak{X}}$, $\mathcal{X} \in T^\perp$ *iff* $\langle \mathcal{X}, \mathcal{Y} \rangle = 0$, $\forall \mathcal{Y} \in T$.

The lemma is proved in Appendix C.1.

As with matrix completion, in a localized observation setting, consistent recovery is infeasible if any entry in $\mathcal{M}$ is overly significant (Refer Section 2.2.1). Thus, it is required that every element in $\mathcal{M}$ have some significant information about the model subspace $T$. This is enforced through the following analogue of incoherence conditions for matrix completion [26, 58].

**Assumption 5.3.2** (Incoherence)**.** $\exists\ (\mu_0, \mu_1)$ such that the following incoherence conditions with respect to standard basis are satisfied for all $\mathcal{E}^{(v,i,j)}$:

$$\|\mathcal{P}_T(\mathcal{E}^{(v,i,j)})\|_F^2 \leq \frac{\mu_0 R}{m_{r_v}} + \frac{\mu_0 R}{m_{c_v}} \tag{5.12}$$

$$\exists \mathcal{E}_{\mathcal{M}} \in \mathscr{E}(\mathcal{M}) \cap T, \text{ s.t. } \langle \mathcal{E}^{(v,i,j)}, \mathcal{E}_{\mathcal{M}} \rangle^2 \leq \frac{\mu_1 R}{N^2} \tag{5.13}$$

Recall $\mathscr{E}(\mathcal{M})$ from (5.8), and $m_k = \sum_{v=1}^{V} n_{c_v} \mathbb{1}_{(r_v=k)} + n_{r_v} \mathbb{1}_{(c_v=k)}$

Note that $\|\mathcal{P}_T(\mathcal{E}^{(v,i,j)})\|_F^2$ is upper bounded by a sum of norms of projections of $m_{r_v}$ and $m_{c_v}$ dimensional standard basis (in $\mathbb{R}^{m_{r_v}}$ and $\mathbb{R}^{m_{c_v}}$) onto the $R$ dimensional latent factor space. Equation (5.12) ensures that no single latent dimension is overly dominant. □

Further, in Section 5.2.2 it was noted that in general $\bar{\mathfrak{X}} \subseteq \mathfrak{X}$, and the set of atoms spanning $\bar{\mathfrak{X}}$ defined in (5.3) need not be centrally symmetric. This poses subtle challenges in analyzing the consistency of collective–matrix completion. To mitigate these difficulties, a restricted set of collective–matrix structures is considered, under which $\mathfrak{X} = \bar{\mathfrak{X}}$.

**Assumption 5.3.3** (Bipartite $\mathcal{G}$)**.** Recall from Section 5.2 that the entity–relationship structure of $\mathfrak{X}$ is represented through an undirected graph $\mathcal{G}$. Assume that $\mathcal{G}$ is bipartite, or equivalently $\mathcal{G}$ does not contain any odd length cycles.

Using induction, it can be easily verified that Assumption 5.3.3 is equivalent to the condition, $\mathfrak{X} = \bar{\mathfrak{X}}$. Under this assumption, $\|.\|_{\mathscr{A}}$ and $\|.\|_{\mathscr{A}}^*$ are norms, and $\|\mathcal{X}\|_{\mathscr{A}}^* = \frac{1}{2}\lambda_{\max}(\mathcal{B}(\mathcal{X})) \leq \frac{1}{2}\|\mathcal{B}(\mathcal{X})\|_2$.

Note: for the well–posedness of collective–matrix completion, some variation of Assumptions 5.3.1, and 5.3.2 is necessary. However, it is not clear if Assumption 5.3.3 is necessary. □

**Assumption 5.3.4** (Sampling Scheme)**.** For all $k$, define $\Omega_k = \{(v_s, i_s, j_s) \in \Omega : r_{v_s} = k \text{ or } c_{v_s}\}$. Let $|\Omega_k|$ be the expected number of observations in $\Omega_k$.

For $s = 1, 2, \ldots, |\Omega|$, independently (a) sample $k_s : k_s = k$ w.p. $\frac{|\Omega_k|}{2|\Omega|}$; (b) sample $i_{k_s} \sim$ uniform$([n_k])$; and (c) sample $j_{k_s} \sim$ uniform$([m_k])$. For $s = 1, 2, \ldots, |\Omega|$, $(v_s, i_s, j_s)$ is the element corresponding to the $(i_{k_s}, j_{k_s})$ in $\mathbb{M}_{k_s}$.

For a given $v \in [V]$ and $(i, j) \in \mathcal{I}(v)$, and $s = 1, 2, \ldots, |\Omega|$:

$$\mathbb{P}\big((v, i, j) = \Omega_s\big) = \frac{|\Omega_{r_v}|}{2|\Omega| n_{r_v} m_{r_v}} + \frac{|\Omega_{c_v}|}{2|\Omega| n_{c_v} m_{c_v}} \tag{5.14}$$

**Remarks:**

1. *Why* $|\Omega_k|$*?*: For consistent recovery of $\mathcal{M}$, the low dimensional factors of $\mathcal{M}$, $\{U_k \in \mathbb{R}^{n_k \times R}\}$ need to be learnt. For a given $k$, information about $U_k$ is entirely contained in $\mathbb{M}_k$. Intuitively for consistent recovery, sample complexity bounds are needed on individual $|\Omega_k|$. Thus, the sampling scheme is chosen in terms of the expected number of observations within each entity type.
2. Hoeffdings's inequality can be used to show that the cardinality of $\Omega_k$ concentrates sharply around $|\Omega_k|$.
3. Note that the notation for cardinality of the set is overloaded: $|\Omega_k|$ is the expected cardinality of the set $\Omega_k$, while $|\Omega|$ is the number of samples observed.

### 5.3.2 Atomic Norm Minimization

Collective–matrix rank of $\mathcal{M} \in \bar{\mathfrak{X}}$ is given by:

$$\text{rank}(\mathcal{M}) = \min_{\{\lambda_r \geq 0\}} \sum_r \mathbb{1}_{\lambda_r \neq 0} \quad s.t. \sum_r \lambda_r \mathcal{A}_r = \mathcal{M},$$

where $\mathcal{A}_r \in \mathscr{A}$. However, minimizing the rank of a collective matrix is intractable, thus the atomic norm (5.4) is proposed as a convex surrogate for the rank function leading to the following convex estimator under noise–free model:

$$\widehat{\mathcal{M}} = \underset{\mathcal{X} \in \bar{\mathfrak{X}}}{\text{argmin}} \ \|\mathcal{X}\|_{\mathscr{A}} \quad \text{s.t. } \mathcal{P}_\Omega(\mathcal{X}) = \mathcal{P}_\Omega(\mathcal{M}). \tag{5.15}$$

The above convex formulation can be suitably modified in the presence of additive noise. For the additive-noise model, the following estimators are equivalent.

$$\widehat{\mathcal{M}} = \underset{\mathcal{X} \in \bar{\mathfrak{X}}}{\text{argmin}} \ \|\mathcal{X}\|_{\mathscr{A}} \ \text{ s.t. } \|\mathcal{P}_\Omega(\mathcal{X} - \mathcal{M})\|_F^2 \leq \omega^2, \tag{5.16}$$

$$\widehat{\mathcal{M}} = \underset{\mathcal{X} \in \bar{\mathfrak{X}}}{\text{argmin}} \ \|\mathcal{P}_\Omega(\mathcal{X} - \mathcal{M})\|_F^2 \ \text{ s.t. } \|\mathcal{X}\|_{\mathscr{A}} \leq \eta, \tag{5.17}$$

$$\widehat{\mathcal{M}} = \underset{\mathcal{X} \in \bar{\mathfrak{X}}}{\text{argmin}} \ \|\mathcal{P}_\Omega(\mathcal{X} - \mathcal{M})\|_F^2 + \gamma \|\mathcal{X}\|_{\mathscr{A}}. \tag{5.18}$$

The estimators are theoretically equivalent in the sense that for some combination of $\omega$, $t$, and $\gamma$, the estimate from the three convex programs are identical. In practice, the parameters are set through cross validation, and the choice of a convex program for noisy collective–matrix completion is often made from the algorithmic considerations.

## 5.4 Main Results

In this section, the theoretical and algorithmic aspects of collective–matrix completion using the proposed estimator are discussed. Thorem 5.4.1 provides the first non–trivial sample complexity bounds under which the convex program

in (5.15) exactly recovers the ground truth matrix with high probability. In Section 5.4.3, a scalable greedy algorithm is proposed for solving noisy collective–matrix completion using the convex program in (5.17). In comparison to the algorithms proposed by Bouchard et. al. [20], the proposed algorithm has a better computational complexity, and strong convergence guarantees.

### 5.4.1 Consistency under Noise–Free Model

In the proof section the following quantity is used which scales atmost as $N^4$ for general $\Omega$ and as $N^2$ under the sample complexity conditions of Theorem 5.4.1:

$$\kappa_{\Omega}(N) = \frac{3|\Omega|\sqrt{\max_k \frac{|\Omega_k|}{n_k m_k}}}{\min_k \frac{|\Omega_k|}{n_k m_k}}$$

Recall that $|\Omega_k|$ is the expected cardinality of $\Omega_k = \{(v,i,j) \in \Omega : r_v = k \text{ or } c_v = k\}$, $|\Omega|$ is the cardinality of $\Omega$, $n_k$ is the number of instances of type $k$, $R$ is the collective–matrix rank of $\mathcal{M}$, and $\mu_0$ and $\mu_1$ are the incoherence parameters (Assumption 5.3.2).

**Theorem 5.4.1.** *Under Assumption 5.3.1–5.3.4, if*

*(i)* $\forall k,\ |\Omega_k| > C_0 \mu_0 n_k R \beta \log N \log\left(N\kappa_{\Omega}(N)\right)$ *and* $\frac{|\Omega_k|}{n_k m_k} \geq c\frac{|\Omega|}{N^2}$, *and*

*(ii)* $|\Omega| > C_1 \max\{\mu_0, \mu_1\} N R \beta \log N \log\left(N\kappa_{\Omega}(N)\right)$,

*for large enough constants* $c$, $C_0$, *and* $C_1$, *then for the noise–free observation model, the convex program in* (5.15) *exactly recovers the true collective–matrix* $\mathcal{M}$ *with probability greater than* $1 - N^{-\beta} - C_2 N^{-\beta} \log\left(N\kappa_{\Omega}(N)\right)$ *for a constant* $C_2$.

### 5.4.2 Discussion and Directions for Future Work

As noted earlier, for consistent recovery of $\mathcal{M}$, the low dimensional factors of $\mathcal{M}$, $\{U_k \in \mathbb{R}^{n_k \times R}\}$ need to be learnt. For a given $k$, information about $U_k$

is entirely contained in $P_{\Omega_k}(\mathbb{M}_k)$. Thus, an obvious lower bound on the sample complexity for well-posedness is given by $|\Omega_k| \sim O(n_k R)$. The results presented are optimal upto a poly–logarithmic factor.

A trivial way to address the collective–matrix completion task is to perform matrix completion on the component matrices independently. Since a joint low–rank structure also imposes low rank structure on the component matrices, this is feasible if each component matrix satisfies the sample complexity requirements of standard matrix completion, i.e. $|\Omega_v| > C\mu_0 R(n_{r_v} + n_{c_v}) \log(n_{r_v} + n_{c_v})$. However, the proposed collective matrix completion setting leverages the shared structure introduced by the jointly factorizability of collective–matrices to obtain a better sample complexity.

The collective–matrix completion problem can also be cast as standard matrix completion problem of completing an incomplete $N \times N$ symmetric matrix, in which blocks corresponding to the collective–matrix are partially observed. However, the existing theoretical results on the consistency of matrix completion algorithms require either uniform random sampling [26, 87, 79], or coherent sampling [35] of the entries of the matrix; and these results cannot be directly applied for blockwise random sampled matrix. Thus, the results in this chapter provide a strict generalization to existing matrix completion results for the task of collective–matrix completion.

### 5.4.3 Algorithm

Recently, Jaggi et. al. [77] proposed a scalable approximate algorithm for solving nuclear norm regularized matrix estimation by casting nuclear norm minimization as a semi definite program (SDP), and then using the approximate SDP solver of Hazan [67]. The robust estimate for collective atomic norm proposed in

Proposition 5.2.2, is of similar flavor. Using Proposition 5.2.2, the optimization problem in (5.17) for solving noisy collective–matrix completion can be cast as the following SDP:

$$\min_{Z \succcurlyeq 0,} \sum_{v=1}^{V} \|P_{\Omega_v}(M_v - P_v(Z))\|_F^2 \quad \text{s.t. } \mathrm{tr}(Z) \leq \eta, \tag{5.19}$$

where $\Omega_v = \{(v_s, i_s, j_s) \in \Omega : v_s = v\}$. Hazan's algorithm to solve (5.19) is given in Algorithm 1.

**Algorithm 1** Hazan's Algorithm for Convex Collective–Matrix Completion (5.19)

Rescale loss function as $\widehat{f}_\eta(Z) = \sum_v \|P_{\Omega_v}(M_v - P_v(\eta Z))\|_F^2$
Initialize $Z^{(1)}$
**for all** $t = 1, 2 \ldots, T = \frac{4}{\epsilon}$ **do**
  Compute $u^{(t)} = \text{approxEV}\big(-\nabla \widehat{f}_\eta(Z^{(t)}), \frac{1}{t^2}\big)$
  $\alpha_t := \frac{2}{2+t}$
  $Z^{(t+1)} = Z^{(t)} + \alpha_t u^{(t)} u^{(t)\top}$
**return** $[P_v(Z^{(T)})]_{v=1}^V$

**Lemma 5.4.2.** *Algorithm 1 solves (5.17) upto $\epsilon$ error in time $O\big(\frac{|\Omega|}{\epsilon^2}\big)$.*

*Proof:* From Theorem 2 of Hazan's work [67], the proposed algorithm returns an estimate with primal–dual error of atmost $\epsilon$ in $\frac{4C_f}{\epsilon}$ iterations, where $C_f$ is a curvature constant. For squared loss, $C_f \leq 1$ (Lemma 4 in [77]). Iteration $t$ in the algorithm involves computing an $\frac{1}{t^2}$–approximate largest eigen value of a sparse matrix with $|\Omega|$ non–zero elements. Using Lanczos algorithm, each iteration requires $O(\frac{|\Omega|}{t})$ computation. □

In comparison to the proposed algorithm, the SVT algorithm proposed by Bouchard et. al. [20] converges in $O(\frac{1}{\sqrt{\epsilon}})$ iterations, however, each iteration requires computing all the non–zero eigenvectors of a $N \times N$ matrix, which has significantly

higher computational cost. For the task of matrix completion, Jaggi et. al. [77] observe upto $\sim 5$x speedup on Hazan's algorithm over SVT algorithm.

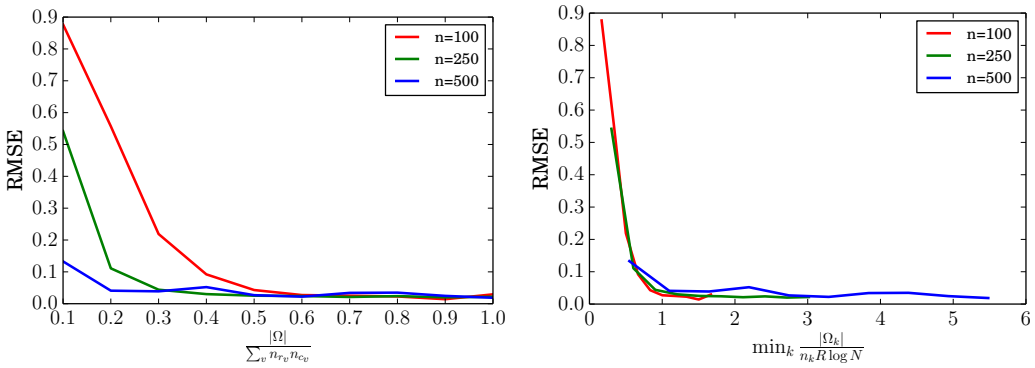
## 5.5 Experiments

### 5.5.1 Simulated Experiments

Low–rank ground truth collective–matrices with $K = 4$, $V = 3$ were simulated, where view 1 is a relation between entity types 1 and 2, view 2 is a relation between entity types 1 and 3, and view 3 is a relation between entity types 2 and 4 respectively. For simplicity assume a common $n_k = n$. Collective matrices with $n \in \{100,\ 250,\ 500\}$ with rank to $R = 2 \log n$ were generated. The matrices were partially observed with the fraction of observed entries, $\frac{|\Omega|}{\sum_v n_{r_v} n_{c_v}}$ varying in $[0.1, 0.2, \ldots, 1]$ and the errors were plotted against the unnormalized fraction of observations, $\frac{|\Omega|}{\sum_v n_{r_v} n_{c_v}}$ in Figure 5.2a, and against the normalized sample complexity provided by the theoretical analysis, $\min_k \frac{|\Omega_k|}{n_k R \log N}$ in Figure 5.2b. It can be seen from the plots that the error decays with increasing sample size, indeed $|\Omega_k| > 1.5 n_k R \log N$ samples suffice for the errors to decay to a very small value. The aligning of the curves (for different $n$) given the normalized sample size corroborates the theoretical sample complexity requirements.

### 5.5.2 Experiments with Commercial News Recommendation Dataset

The proposed approach was evaluated on two datasets from a commercial news recommendation engine. The entities include users, news articles, and news–categories. The datasets consists of two views (a) user–article click information in a 3hr time window, (b) an aggregation of the categories clicked by users was used to train a classifier that gives a dense and complete user–category preference.




(a) RMSE vs unnormalized sample size (b) RMSE vs normalized sample size

Figure 5.2: Error convergence against normalized and unnormalized sample size

The first dataset "News–Cold–Start", consists of $\sim 180K$ users, $\sim 750$ articles, and $34$ categories. In this dataset, $\sim 25000$ users have only one click. Randomly chosen negative samples were added to give dataset of $\sim 1.25$ million user–article ratings, and $\sim 1.4$ million user–category annotations. The dataset was split in $80 : 10 : 20$ proportion as training, validation, and test. The $20\%$ of the test dataset contains cold start users with no rating information. In the second dataset "News–No–Cold–Start", all the cold start users in the test dataset were removed leading to a much smaller datasets consisting of $\sim 6500$ users, $\sim 750$ articles and $34$ categories, with $\sim 150K$ user–article ratings and $\sim 50K$ user–category ratings (including the randomly chosen negatives). The negatives in each dataset were sampled independently in each cross–validation iteration to remove bias.

Mean absolute error (MAE) on the test dataset obtained from the proposed Hazans algorithm for Collective–Matrix Completion (CMF–Hazans) and Standard Matrix Factorization (SMF) are reported in Table 5.1.

| Method | News-Cold-Start | News-No-Cold-Start |
|---|---|---|
| CMF–Hazans | $0.27408 \pm 0.00016$ | $0.21559 \pm 0.00143$ |
| SMF | $0.29051 \pm 0.00074$ | $0.21488 \pm 0.00076$ |

Table 5.1: MAE of the predictors on the two news recommendation datasets

It is observed that collective matrix factorization does not add much value for warm–start cases as the ratings give accurate predictor. On the other hand, for test dataset consisting on both warm–start and cold–start test cases, the proposed joint estimation potentially leverages the information in the user–category affinities and shows significant improvement.

# Chapter 6

# Phenotyping using Structured Estimation

The increased availability of electronic health records (EHRs) have spearheaded the initiative for precision medicine using data driven approaches. Essential to this effort is the ability to identify patients with certain medical conditions of interest from simple queries on EHRs, or EHR-based phenotypes. Existing rule–based phenotyping approaches are extremely labor intensive. Instead, dimensionality reduction and latent factor estimation techniques from machine learning can be adapted for phenotype extraction with no (or minimal) human supervision.

Building on the results of Chapter 3–5, this chapter proposes to identify an easily interpretable latent space shared across various sources of EHR data as potential candidates for phenotypes. By incorporating multiple EHR data sources (e.g., diagnosis, medications, and lab reports) available in heterogeneous datatypes in a generalized *Collective Matrix Factorization (CMF)*, the proposed methods can generate rich phenotypes. Further, easy interpretability in phenotyping application requires sparse representations of the candidate phenotypes, for example each phenotype derived from patients' medication and diagnosis data should preferably be represented by handful of diagnosis and medications, (5–10 active components). The CMF framework is extended for learning and interpretable phenotypes from multiple sources of EHR data. Non–negativity and sparsity inducing constraints are imposed to enhance the interpretability of the candidate phenotypes. The proposed model is applied on EHR data from Vanderbilt University Medical Center.

## 6.1 Introduction

EHR-driven phenotyping refers to the task of identification of a set of clinical features or characteristic indicative of a medical condition from EHR data and has been a major focus of EHR data analyses [118]. Phenotypes are important for targeting patients for screening tests and interventions, improving multisite clinical trials, and to support surveillance of infectious diseases and rare disease complications. While existing efforts (e.g., eMerge Network, Phenotype KnowedgeBase, and the SHARPn program) have illustrated the promise of EHR–driven phenotypes, state of the art phenotype development generally requires an iterative and collaborative effort between clinicians and IT professionals to compose a series of rules for reproducible queries of EHR databases [74, 117]. A single phenotype takes substantial time, effort, and expert knowledge to develop. Data mining tools such as support vector machines [32], active learning approaches [36] and inductive logic programming [123], have been recently used to partially automate the phenotyping process. Yet, these work require annotated samples to obtain good performance. As such annotations are expensive and time consuming to obtain, it is of interest to investigate unsupervised learning tools for automated phenotyping.

Phenotyping can be viewed as a form of dimensionality reduction of EHR data, where each phenotype or medical condition of interest represents a latent space [74] and the rich literature in the field of machine learning for latent space estimation can be suitably adapted to automate and speed up the phenotype extraction process. Several factors contribute to the quality of phenotypes extracted from EHR data, and it is advantageous to consider these factors in choosing the appropriate dimensionality reduction tools for phenotyping. A review of the top 10 phenotypes across different studies showed that several data sources are typically used to define a phenotype [134]. Additionally, EHR data is commonly available

in heterogeneous datatypes. For example, laboratory test results are often in the form of a real–valued number, patient demographic information can be encoded as a binary value, and procedure codes contain the number of times, a non-negative integer, the procedure is performed. Thus, an automated phenotyping process that can incorporate data from heterogeneous datatypes and diverse sources can help identify rich existing as well as novel medical concepts.

Recent work has illustrated the promise of tensor factorization to generate phenotypes with minimal human supervision [71, 154, 68]. Latent space shared by various modes of higher order tensors are easier to interpret; and also more accurately capture the multi–source nature of phenotypes. However, rich multi–way interactions required to form tensors is often not available in existing EHR data, for example, in a simple 3rd order patient-diagnosis-medication tensor, the $(i, j, k)^{\text{th}}$ entry of the observation requires detailed information on the number of times patient $i$ was prescribed medication $k$ in response to diagnosis $j$. In practice, much of the EHR data is available in flat formats that are more readily represented as matrices rather than tensors, e.g., a patient-diagnosis and a patient-medication matrix. Moreover, maintaining infrastructure to record and store higher order multi–way interactions is resource–intensive as the number of such possible interactions exponentially increase with each additional source. Alternatively, tensors constructed by approximating higher order interactions from flat format data could lead to noisy correlations and biased results. These motivate the exploration of tools that directly work with multiple sources of matrix valued data.

In this chapter, unsupervised models are proposed for learning phenotypes from EHR data that are available as a collection of matrices. *Collective Matrix Factorization (CMF)* [138] (also see Chapter 5) is an effective tool for identifying a latent space shared across multiple sources of data. In CMF, a collection of re-

lated matrices are jointly factorized into low–rank factors that are shared across the entire collection. For the phenotyping application, various structural and methodological modifications are introduced to the basic CMF model towards enhancing interpretability of candidate phenotypes.

- *Heterogeneous datatypes:* Each source of EHR data can contain diverse datatype representations, such as numeric, count, or integer elements. Thus, it is desirable to use loss functions that are appropriate for the data in each source. The class of *Bregman divergences* are chosen to be appropriate for the phenotyping application as this class includes divergence functions appropriate for various datatypes, including continuous real–valued, binary and count data (Chapter 3).

- *Collective Factorization:* The challenge in effectively combining heterogeneous divergences in a collective matrix factorization is that such divergences often span different numerical scales and simple unweighted combinations tend to overfit datatypes or source matrices whose divergences are in the higher numerical range. An effective heuristic approach to estimate appropriate weights for individual source matrices.

- *Non–negativity and Sparsity:* Physically interpretable latent factors are necessary to extract clinically meaningful phenotypes from EHR data. Non–negative matrix factorization (NMF) [120, 98] in comparison to the more traditional *principal component analysis (PCA)* provides better interpretability of the low–rank factors as sum–of–parts representation. Such non–negativity constraints can be readily extended into the CMF framework. Further, sparsity of latent factors representing the phenotypes plays a crucial role in the usefulness of the phenotypes as human experts need to analyze the factors and conduct further investigation to validate its clinical relevance. Thus, each phenotype should be ideally be represented by very few active components ($\leq$ 10 non–zero loading of entities) from

each source. In one of the proposed variations of the generalized collective NMF formulation, convex sparsity inducing constraints are introduced to enhance the interpretability of extracted latent factors.

The proposed models were empirically evaluated on real EHR data from Vanderbilt University. The clinical relevance of the extracted phenotypes were evaluated by domain experts.

## 6.2 Related Work

Inferring low–dimensional representation of matrix data is a fundamental problem in machine learning. PCA [83], the most popular and widely used tool dimensionality reduction, learns latent factors as low rank matrices whose values are unconstrained and can contain both positive and negative entries. However, in many applications it is desirable to interpret the low rank factors as physical concepts and negative entries often contradict physical reality. This motivated a related line of dimensionality reduction techniques called the Non–Negative Matrix factorization (NMF) [120, 98]. Several existing work extend matrix factorization tools to analyze data from multiple matrices. Collective matrix factorization (CMF) and its non–negative variants [138] incorporate information from multiple sources of matrix data using shared latent variables/factors. Alternatively, regularized NMF variants have been proposed combining data from multiple sources [165, 104]. The tools for matrix valued data have also been generalized to higher order tensors, or multi–way arrays (see [92] for a review). Variants of non-negative tensor factorization (NTF) based on CANDECOMP–PARAFAC, one of the most popular tensor decomposition models have been applied to extract interpretable latent/hidden factors, e.g. [42, 99, 41, 2, 164, 3] and references therein. However, most of these methods primarily utilize the least square loss and may not be appropriate for all

data types. This work builds on these tools to propose techniques for efficiently extracting structured latent factors from multiple sources of heterogeneous EHR data. The primary focus is on the interpretability of the low–dimensional factors as meaningful phenotypes.

Although existing phenotyping methods rely on a labor–intensive process, unsupervised models have been proposed to leverage the vast amount of EHR data for automatic phenotype discovery. These models include the use of probabilistic graphical models to cluster patient's longitudinal trajectories [132], deep learning to detect characteristic patterns in clinical time series data [34], and generative models on static data [37]. Yet these methods are not scalable and are ill-suited for incorporating data from patients over a prolonged period of time (6+ months). Recent work has illustrated the promise of NTF to generate phenotypes with minimal human supervision using data over several years [70, 71, 154, 68]. However, as noted earlier, a tensor representation is not always available in EHR data, at least not without introducing assumptions and potentially biasing the results.

## 6.3 Phenotyping from EHR Data

The notations used in the rest this chapter are summarized in Table 6.1. The patient EHR data from $V$ sources, such as medications, diagnosis, laboratory measurements, etc. are represented as matrix valued data whose rows correspond to a common set of patients, and columns represent entities from the respective sources (medications, diagnosis, laboratory measurements, etc.). Let $d_0$ denote the number of patients, and for each source $v \in \{1, 2, \ldots, V\}$, let $d_v$ denote the number of unique entities within the source $v$. The collection of $V$ matrices containing EHR data from multiple sources is denoted by $\mathcal{X} = [X_v]_{v=1}^{V}$, where $X_v \in \mathbb{R}^{d_0 \times d_v}$ denotes the matrix data from source $v$.

| Notation | Description |
|---|---|
| **Input** | |
| $v = 1, 2, \ldots, V$ | Index over $V$ sources of EHRs, e.g. medication, diagnosis, etc. |
| $d_0$ | Number of patients |
| $d_v$ | Number of entities in source type $v$ |
| $X_v \in \mathbb{R}^{d_0 \times d_v}$ | EHR data matrix from source $v$ |
| $\mathcal{X} = [X_v]_{v=1}^{V}$ | Collection of $V$ EHR data matrices |
| $D_v$ | Bregman divergence appropriate for approximating $X_v$ |
| **Estimates** | |
| $\widehat{\mathcal{X}} = [\widehat{X}_v]_{v=1}^{V}$ | Estimate of $\mathcal{X}$ from models |
| $W \in \mathbb{R}^{d_0 \times r}$ | Patients' loading along the $R$ dimensional latent space |
| $H_v \in \mathbb{R}^{d_v \times r}$ | Latent factor representation for features in source $v$ |
| $b_v \in \mathbb{R}^{d_v}$ | Bias factors associated with columns of the data matrix $X_v$ |

Table 6.1: Additional notations for phenotyping using structured estimation

### 6.3.1 Dataset Overview

The proposed models are evaluated on an EHR data set from Vanderbilt University Medical Center. This section contains a brief exploration of the data and the empirical results.

The dataset consists of de-identified electronic medical records corresponding to the first ∼10,000 patients in BioVU*, the Vanderbilt DNA databank, spanning over 20 years. The details of the inclusion and exclusion criteria for the databank are described in [130]. For evaluation purposes, a subset of data containing the case and control patients for type–2 diabetes and resistant hypertension is used. These patients and their labels were selected by using the respective rule–based phenotype algorithms defined in the Phenotype KnowledgeBase†. However, the labels

*https://victr.vanderbilt.edu/pub/biovu/

†https://phekb.org

from these rule-based algorithms were *not* used in the phenotyping models which are learned in a completely unsupervised setting.

Although the proposed model is general enough to be applied to multiple data types, the empirical study work with counts of diagnoses and medications for evaluation purposes. The diagnosis codes, in the form of International Classification of Disease, 9th edition (ICD-9) codes, were grouped using the PheWAS code groups[‡], a custom-developed hierarchy which currently contains $\sim 1600$ groups. Medications were aggregated based on Medical Subject Headings (MeSH) pharmacological actions provided by the RxClass REST API, a product of the US National Library of Medicine. Note that a medication may belong to multiple categories. Figure 6.1 provides example aggregations performed on the original table for the purpose of the study.

Finally, BioVU dataset assigns an index (reference) date to each patient, which corresponds either to the date where the criteria was met (case patients) or the last encounter date (control patients). The EHR records of patients falling in the date range of one year prior to their index date up until the index date were used in the experiments. Any patient without at least one diagnosis and medication during the relevant time period was not included in the study. The resulting data set contains $2039$ patients, $936$ diagnosis groups, and $161$ medication classes.

The dataset is summarized in Table 6.2 and the top five diagnosis and medication categories that appear in the data are shown in Table 6.3.

[‡]http://phewas.mc.vanderbilt.edu/

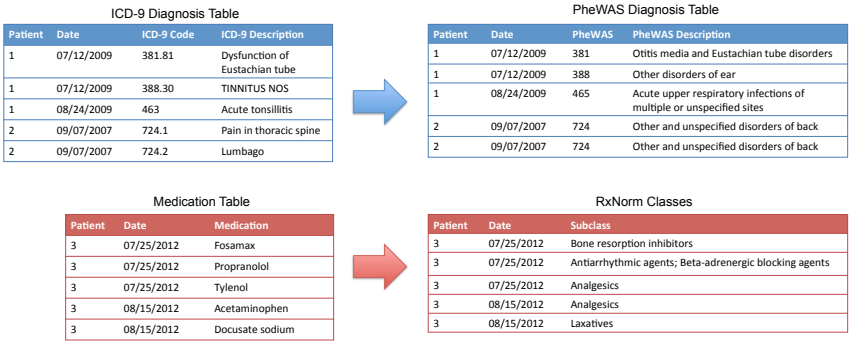



Figure 6.1: Examples of the aggregation from ICD-9 diagnosis codes to PheWAS code groups and original medications to the MeSH pharmacological actions classes.

| $v$ | Source Matrix $X_v$ | $d_0 \times n_v$ | Datatype |
|---|---|---|---|
| 1 | Patient–Diagnosis | $2039 \times 936$ | Count |
| 2 | Patient–Medication | $2039 \times 161$ | Count |

Table 6.2: Dataset summary of BioVU dataset used for phenotyping.

## 6.4 Structured Collective Matrix Factorization for Phenotyping

For each source $v \in [V]$, $X_v$ is approximated by structured estimates $\widehat{X}_v$ which incorporates model constraints appropriate for effective phenotyping.

### 6.4.1 Heterogeneous Datatypes

In EHR data from multiple sources, each source matrix $X_v$ may contain data represented in diverse datatypes (e.g., binary values for demographics, count values for medications, or continuous values for laboratory measurements). In the proposed phenotyping models, the data fidelity of $\widehat{X}_v$ is quantified using an appropriately chosen source–specific divergence $D_v(X_v, \widehat{X}_v)$. The divergence functions are selected from a class of *Bregman divergence* 2.1.7. The motivation for using Bregman divergences are two fold (cf. Chapter 3). Bregman divergences include

| Source | Top five entities |
|---|---|
| Diagnosis | Hypertension; Incision, excision, and division of other bones; Ischemic Heart Disease; Secondary malignant neoplasm of respiratory and digestive systems; Disorders of lipoid metabolism. |
| Medication | Analgesics; Vitamins; Anticonvulsants; Anxiolytics, sedatives, and hypnotics; Antihyperlipidemic agents. |

Table 6.3: The top five diagnosis and medications of the patients in the study.

rich classes of loss functions that are appropriate for a variety of datatypes including (weighted) squared loss for continuous valued data, logistic loss for binary valued data, and generalized KL divergence for count valued data among others [54, 14]. These loss functions are also equivalent to the negative log–likelihood of members of exponential family distributions including Gaussian, Bernoulli, Poisson, exponential among others [54, 14]. Thus, the domain knowledge of data distribution can be potentially incorporated in choosing the appropriate divergence. Secondly, Bregman divergences are strictly convex and differentiable in the first parameter, and accurate and tractable estimators for $\widehat{X}_v$ can be developed using gradient descent and alternating minimization algorithms.

In the dataset described in Section 6.3.1, as both the matrices described in Section 6.3.1 have count valued data, the generalized KL divergence given by the following equation is used as the divergence for both sources:

$$D(X, \widehat{X}) = \sum_{ij} \widehat{X}_{ij} - X_{ij} + X_{ij} \log \frac{X_{ij}}{\widehat{X}_{ij}}. \tag{6.1}$$

### 6.4.2 Generalized Collective NMF (CNMF)

As noted earlier, non–negativity constraints on the patient loading and latent factor matrices allow for better interpretability as sum–of–parts representation. A generalized collective NMF (CNMF) is proposed as a basic model for extracting phenotypes from multiple sources of patient data available in heterogeneous datatypes. In Section 6.4.3, additional structures are introduced to enhance interpretability.

Each source of EHR data $v$ is associated with a structured latent factor matrix $H_v \in \mathbb{R}^{d_v \times R}$, and these factors jointly span a shared latent space. The columns of $H_v$ concatenated across the $V$ sources are potential candidates for phenotypes. The loading of the patients along these latent dimensions are given by the matrix $W \in \mathbb{R}^{d_0 \times R}$. Additionally, the raw EHR data often contains generic features that are not necessarily indicative of any medical condition of interest. For example, medications like pain reliever, laboratory measurements like body temperature, etc. are frequently encountered in patient data, but are not discriminative of patient conditions. EHR data from such frequent and non–discriminative features are captured through an explicit (and potentially dense) column or feature bias factor $b_v \in \mathbb{R}^{n_v}$ for each source $v$.

For $v \in [V]$, the source $X_v$ is approximated as $WH_v^\top + \mathbf{1}b_v^\top$, where $\mathbf{1}$ is a vector of all ones in appropriate dimensions. A Bergman divergence $D_v$ appropriate for each source is used to measure the data fidelity of the estimate to the observed data. Finally, as the heterogeneous divergences are in different scales, the divergences are weighted using parameters $\alpha_v, v = 1, 2, \ldots, V$. The basic CNMF

estimator is given by the following optimization problem.

$$\begin{aligned}\widehat{\mathfrak{X}} = \underset{\{\widehat{X}_v\}_{v\in[V]}}{\text{argmin}} & \sum_{v=1}^{V} \alpha_v D_v(X_v, \widehat{X}_v), \\ \text{s.t. } & \widehat{X}_v = WH_v^\top + \mathbf{1}b_v^\top \text{ for } v=1,2,\ldots,V, \\ & W \in \mathbb{R}_+^{d_0\times R},\ H_v \in \mathbb{R}_+^{d_v\times R},\ b_v \in \mathbb{R}_+^{d_v}.\end{aligned} \tag{6.2}$$

#### 6.4.2.1 Computing $\{\alpha_v : v = 1, 2, \ldots, V\}$

As noted earlier, since the divergences associated with difference datatypes span different numerical scales, unweighted objective in (6.2) will tend to overfit the matrices whose divergences are in the higher numerical range. An effective heuristic approach is proposed to estimate contribution of each source matrix $X_v$ in the joint estimation. To motivate the idea, consider a source matrix $X_v$. If a joint factorization is not required, i.e. $W$ need not be shared, then the optimization problem in (6.2) can be solved as $V$ independent structured factorization $\widetilde{X}_v^{\text{ind}} = W_v H_v^\top + \mathbf{1}b_v^\top$ without the weights $\alpha_v$. In a preprocessing step, for each source and independent factorization of the form $\widetilde{X}_v^{\text{ind}}$ is learned by minimizing $D_v(X_v, \widetilde{X}_v^{\text{ind}})$ assuming the sources to be independent of each other. The resultant divergence from independent factorization is treated as the effective scale of divergence for each source. In order to assign equal importance to all source matrices, the choice of $\forall v,\ \alpha_v = \frac{1}{D_v(X_v, \widetilde{X}_v^{\text{ind}})}$ is proposed.

### 6.4.3 Sparsity–inducing CNMF (SiCNMF)

As phenotypes learned form data analysis tools are further investigated by human experts, it is desirable that candidate phenotypes learned from EHRs are sparse combinations of the source entities, i.e., columns $H_v$ are sparse.

To illustrate sparsity–inducing constraints for enhanced interpretability, first consider a single source of EHR data matrix $X \in \mathbb{R}^{d_0 \times n}$ and an appropriate divergence function $D(.)$. Explicit sparsity constraints on the factor matrix $H$ lead to intractable combinatorial optimization problems. A commonly used convex surrogate for sparsity involves restricting the $\ell_1$ norm of the columns of $H$, i.e., constraints of the form $\{\|H^{(k)}\|_1 \leq s : k \in [R]\}$, for some parameter $s$. However, in (6.2) if the scaling of $W$ is unrestricted, then due to multiplicative nature of the factorization, restrictions on norm of $H$ tend to be ineffective as any scaling of $H$ can be easily absorbed by $W$. Thus, additionally the scale of $W$ is constrained using a Frobenius norm constraint of the form $\|W\|_F \leq \eta$, for another parameter $\eta$. Note that, $s$ and $\eta$ effectively work as single parameter due to the multiplicative update. Thus, WLOG,fix $s = 1$ and use $\eta$ as a tunable parameter to control the sparsity level.

The following generalized SiCNMF model is proposed as an extension of vanilla CMF which incorporates (a) sparsity–inducing and non–negativity constraints for enhanced interpretability, (b) feature specific bias factors $\{b_v : v \in [V]\}$ to capture data specific offsets, and (c) appropriately weighted heterogeneous divergences to handle varied datatypes.

$$\begin{aligned} \widehat{\mathcal{X}} = \underset{\{\widehat{X}_v\}_{v\in[V]}}{\text{argmin}} & \sum_{v=1}^{V} \alpha_v D_v(X_v, \widehat{X}_v), \\ \text{s.t. } & \widehat{X}_v = W H_v^\top + \mathbf{1} b_v^\top \text{ for } v = 1, 2, \ldots, V, \\ & W \in \mathbb{R}_+^{d_0 \times R},\ H_v \in \mathbb{R}_+^{d_v \times R},\ b_v \in \mathbb{R}_+^{d_v}, \\ & \|W\|_F \leq \eta, \|H_v^{(k)}\|_1 = 1\ \forall k \in [R], \end{aligned} \tag{6.3}$$

where recall that $H_v^{(k)}$ is the $k^{\text{th}}$ column of $H_v$, and $\alpha_v$ are either (a) all ones (unweighted SiCNMF), or (b) computed using the methodology described in Section 6.4.2.1 (weighted SiCNMF). Note that the higher the value of $\eta$, the weaker the sparsity constraint. In the limiting case of $\eta = \infty$, the model is equivalent to

the heterogeneous collective non–negative matrix factorization (CNMF) as scaling constraints of $H_v$ are captured by $W$.

## 6.5 SICNMF: Algorithm Details

For any set of Bregman divergences $\{D_v : v = 1, 2, \ldots V\}$ and positive parameters $\eta, \{\alpha_v\} > 0$ , the optimization problem (6.3) is convex in $[(H_v, b_v)\ \forall v]$ when $W$ is fixed and vice versa. The proposed algorithm uses alternating minimization to solve (6.3) where each iteration alternatively minimizes $[(H_v, b_v)\ \forall v]$ and $W$, while keeping the other fixed. Each such component update involves minimizing a smooth convex objective subject to convex constraint set and is solved using projected gradient decent algorithm with backtracking line search to determine step size [101].

Recent work has shown that projected gradient methods are computationally competitive and have better convergence properties than standard multiplicative update approaches [101]. Moreover, compared to multiplicative updates, projected gradient descent based algorithms can be easily extended for convex constraints beyond simple non–negativity. Although [101] ignore the KL divergence problem as ill-defined, a more recent work [40] provide convergence for related tensor factorization task by showing that the convex hull of the level sets of the KL divergence problem is compact. To project onto the simplex, the simple and fast algorithm proposed by Chen and Ye is used [38].

The algorithm for solving (6.3) is summarized in Algorithm 2.

**Algorithm 2** Alternating minimization for (6.3) using projected gradient descent

**Input:** EHR data $X_v$, $D_v(.)$ for $v = 1, 2, \ldots, V$

**Parameters:** Divergence weights $\{\alpha_v\}$ and tunable sparsity inducing parameter $\eta \in (0, \infty)$

**while** not converged **do**

$$\widehat{W} = \underset{W \geq 0}{\text{argmin}} \sum_{v=1}^{V} \alpha_v D_v(X_v, W\widehat{H}_v^\top + \mathbf{1}\widehat{b}_v^\top)$$
$$\text{s.t.} \quad |W\|_F \leq \eta, \; W \geq 0.$$

**for** $v \in [V]$ **do**

$$\widehat{H}_v, \widehat{b}_v = \underset{H_v \geq 0, b_v \geq 0}{\text{argmin}} \; D_v(X_v, \widehat{W}H_v^\top + \mathbf{1}b_v^\top)$$
$$\text{s.t.} \quad \forall k, \|H_v^{(k)}\|_1 = 1.$$

**return** Patient loadings $\widehat{W}$, factors/phenotypes $\{\widehat{H}_v\}$ and feature biases $\{\widehat{b}_v\}$

## 6.6 Experiments

The generalized KL–divergence (6.1) is used as loss function for both matrices (patient by diagnosis and patient by medication) in the collective matrix factorization models as well as the baselines described in the following subsection.

### 6.6.1 Baseline Models

The primary focus of this work is the clinical relevance of candidate phenotypes obtained from unsupervised dimensionality reduction techniques. Since the Vanderbilt data contains flat files associated with the diagnosis codes and medications, construction of the patient–medication and patient–diagnosis matrices for the collective matrix factorization models were straightforward. The proposed models of CNMF (6.2) and SiCNMF (6.3) are compared with two baseline models described below:

- **Non–negative matrix factorization (NMF) [98]**: In order to evaluate traditional NMF in identifying a shared latent space, the patient information is

aggregated into a third matrix, diagnosis by medication, wherein each element represents the number of patients who have at least one occurrence of both the diagnosis and the medication during the one year time window of the dataset. It is important to note that under this construction, a patient with two encounters almost one year apart, one with the diagnosis A and one with medication B would be counted in the $(A, B)^{\text{th}}$ entry of the matrix. A non–negative matrix factorization model with the generalized KL divergence as the objective and no sparsity constraints is performed on the diagnosis by medication matrix.

- **Marble [71]**: Marble is a sparse non–negative tensor factorization model that has been used to obtain highly effective and interpretable phenotypes provided a multiway tensor EHR data is available. However, the BioVU dataset does not have rich multi–way interactions to easily construct tensors. For example, in a patient–diagnosis–medication tensor, a entry $x_{ijk}$ denotes the number of times a patient $i$ was prescribed medication $k$ in order to treat a diagnosis $j$. To construct tensors from available flat files, these interactions were approximated by assuming that a medication was used to treat a specific diagnosis if both diagnosis and medication occur within a one week time interval, that is the counter for $x_{ijk}$ is incremented if patient $i$ was prescribed medication $k$ within one week of an encounter with diagnosis $j$. Marble applied to this approximated tensor is the second baseline used in experiments.

The baselines described above are compared to three CMF based models described in this chapter: (a) CNMF (6.2) which does not incorporate the sparsity inducing constraints, (b) unweighted SiCNMF which incorporates sparsity–inducing constraints proposed in Section 6.4.3, but uses a simple aggregation of various source divergences, i.e., solves (6.3) with $\alpha_v = 1$ for all $v \in [V]$, and finally

(c) weighted SiCNMF which incorporates both the sparsity–inducing structure and the weights $\alpha_v$ computed using the heuristic described in Section 6.4.2.1.

All the models described above involve non–convex optimization and the estimates from the algorithm are sensitive to initialization. To mitigate this issue from local minima, each algorithm was run independently multiple times and pick the run with best fit to the objective. All the competing models learn a $R = 20$ rank factorization.

### 6.6.2 Sparsity–accuracy trade off: Data fit

The sparsity of the candidate phenotypes plays a crucial role in the interpretability and wider applicability of the estimates. Concise representations allow domain experts to more easily reason about a particular group of patients queried using the phenotypes. As noted earlier, while non–negativity constraint in matrix and tensor factorization inherently induce sparsity as a by–product, there is no explicit control over the sparsity levels. Thus in order to deriving extremely sparse phenotypes involve, sparsity inducing regularization was introduced, whose sparsity levels can be controlled by an tunable knob of $\eta$ in (6.3).

The expected sparsity–accuracy trade-off in the data fit can be observed in Figure 6.2. Note that higher values of $\eta$ in (6.3) correspond to a weaker sparsity constraints as the $W$ factor can more easily absorb the scaling constraint on $H_v$.

### 6.6.3 Type-2 diabetes and Resistant hypertension prediction

With relaxed sparsity constraints, while a monotonic decay of objective function on training data fit as observed in Figure 6.2 is expected, such a monotonic accuracy trade-off does extend for predictions on held out test datasets. Besides improving interpretability, the sparsity constraints further function as regularization to

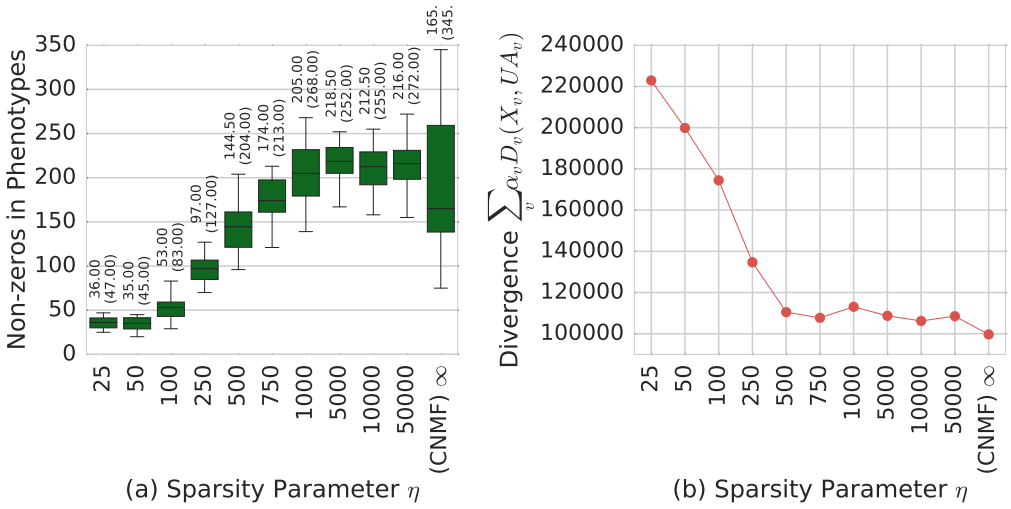



Figure 6.2: Sparsity–accuracy trade-off in data fit of weighted SiCNMF. Sparsity is measured as the median number of non-zero entries in columns of the phenotype matrices concatenated from all sources $\{\widehat{H}_v : v = 1, 2, \ldots, V\}$. (a) Each box plot represents the spread of the number of non–zeros in $R = 20$ candidate phenotypes learned from weighted SiCNMF using $\eta$ represented along the x–axis in (6.3). (b) Plot of decay of divergence between the fitted estimate and the observed data as the sparsity constraint is relaxed using higher $\eta$. Note that the values of $\eta$ along x–axis are not in linear scale and higher values correspond to weaker sparsity–inducing regularization.

prevent overfitting.

To quantitatively evaluate the effectiveness of the extracted phenotypes, consider the classification problem of predicting two chronic conditions prevalent in the patient population of the dataset (Section 6.3.1): (a) type-2 diabetes, and (b) resistant hypertension. As described in Section 6.3.1, for each patient in the dataset, the class labels for these chronic conditions were estimated from rule-based phenotyping algorithm from PheKb.

The full dataset of $\sim 2000$ patients is divided into $5$ stratified cross valida-

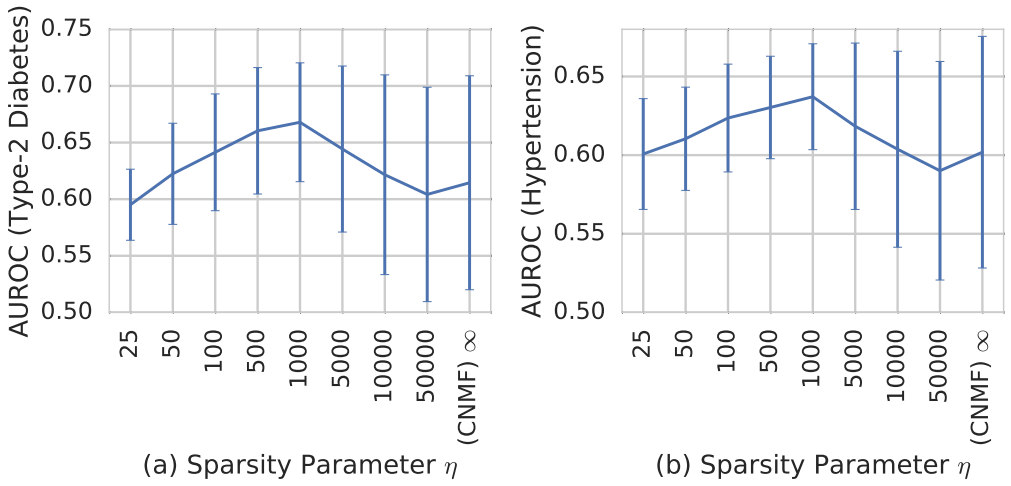



Figure 6.3: Sparsity–accuracy tradeoff in prediction of (a) Type–2 diabetes and (b) resistant hypertension. The results are for weighted SiCNMF, but similar tradeoff was also observed for unweighted SiCNMF. Note that x–axis is not linear and higher $\eta$ leads to lower sparsity (more number of non-zeros in phenotype representations)

tion folds of $80\%$ training and $20\%$ test patients. For each cross-validation fold, the models described in Section 6.6.1 were applied on training EHR dataset to extract the phenotype matrices $\{\widehat{H}_v^{\text{train}} : v \in [V]\}$. It is clarified that, for all the competing models, the phenotypes (latent factors) were extracted (a) only from EHR data of patients in the training set, and (b) the estimates were learned in a completely unsupervised setting. In particular, the test EHR data and the labels were *not* used in the phenotype extraction phase. For each patient, the $R$ dimensional loading along the phenotype/latent space spanned by $\{\widehat{H}_v^{\text{train}} : v \in [V]\}$ is used as features for learning the classifiers. Such representations are computed by projecting the EHR matrix into the fixed phenotype factors. For CMF variants, the features for a patient

with EHR $[X_v^{\text{patient}}]$ is given by:

$$W^{\text{patient}} = \underset{W \geq 0}{\text{argmin}} \sum_v \alpha_v D_v(X_v^{\text{patient}}, W H_v^{\text{train}} + \mathbf{1} b_v^{\text{train}\top}).$$

The sparsity–accuracy trade-off in prediction performance on held out dataset is plotted in Figure 6.3. Although, the predictive performance at various $\eta$ levels are comparable, the mild regularization effect of sparsity constraints can be observed the plots.

### 6.6.4 Sparsity and Prediction Comparison to Baseline Models

In this subsection, the performance CMF based estimators is compared to strong baselines models.

#### 6.6.4.1 Sparsity

The sparsity patterns obtained by the competing phenotyping algorithms described in Section 6.6.1 are compared in Figure 6.4. As expected, the sparsity of SiCNMF models are better than those of non–sparsity–inducing CNMF and NMF models. NMF [98] on dense aggregated data which does not incorporate explicit sparsity constraints learns dense factor matrices. Note that CNMF models multiple sparse matrices jointly learns much sparser factors compared to NMF on single aggregated matrix. Marble [71] induces sparsity by truncation and achieves the best sparsity performance.

#### 6.6.4.2 Prediction

The classification performance of baseline models for predicting type-2 diabetes and resistant hypertension are compared in Figure 6.5. As NMF uses aggregated data of patients and there is no effective approach to learn individual patient

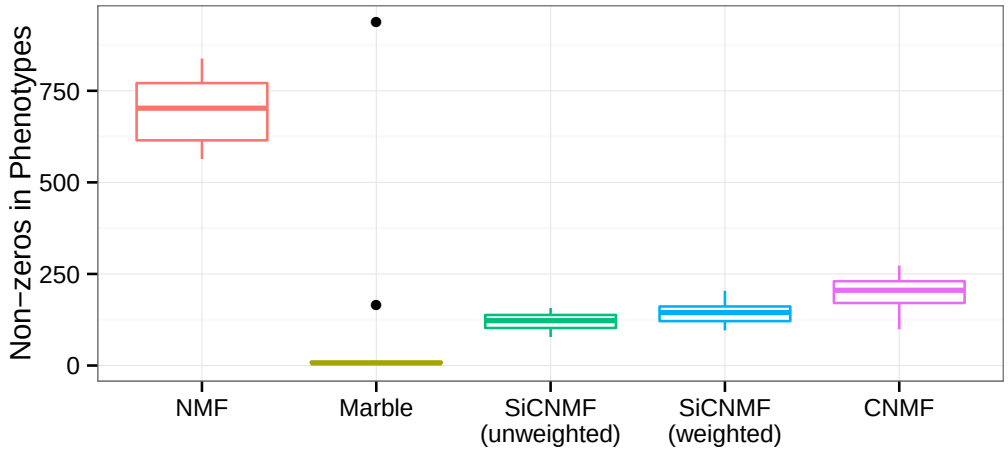



Figure 6.4: Box plots showing the inherent sparsity induced by the models.

representations in the phenotype space. Thus, NMF is excluded from this set of experiments. Instead, the classifiers learned on *full* concatenated EHR matrix is used as an additional baseline for prediction performance. Note that the concatenated EHR matrix has $> 1000$ features compared to the $20$ dimensional representation of the rest of the models. It is observed that the phenotype based models with $20$ dimensional feature representation have comparable performance. However, the classifiers from full EHR matrix with $> 1000$ features outperforms the phenotype–

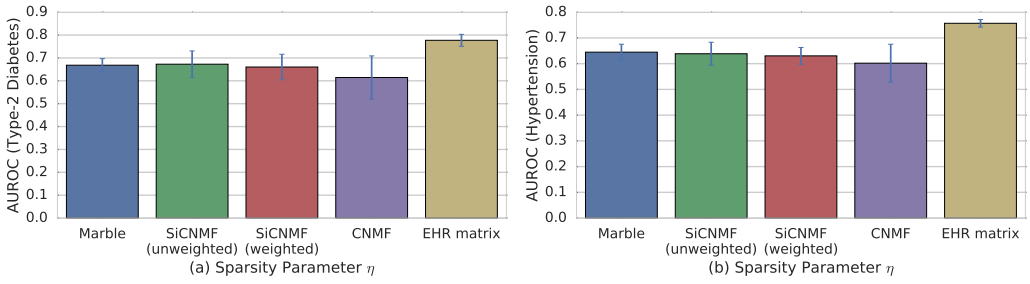



Figure 6.5: Accuracysparsity tradeoff in prediction

based models. While the EHR matrix provides a richer set of features for prediction performance, the high dimensional EHR data are not not useful for phenotyping applications and interpretability.

### 6.6.5 Clinical Relevance of Phenotypes

The phenotypes extracted from the models described above are evaluated by a human expert for clinical relevance. For reasonable evaluation of the phenotypes by humans, it is desirable that each phenotype be represented by a very small number of diagnoses and medications groups. Based on a round of feedback from a clinical expert, in post processing, just the top $5$ medications and top $5$ diagnosis from phenotypes learned from *all* the models were retained for evaluations in this section.

The clinical relevance of the resulting phenotypes were evaluated from the phenotyping models by conducting a survey with a domain expert. The domain expert was given $20$ phenotypes from each model to assess and were not informed apriori the correspondence between the models and the results. For each of the individual phenotypes, the experts assigned one of three values: **(1) yes – it was clinically meaningful, (2) possible – the phenotype has some clinical meaningfulness, and (3) no – it was not meaningful at all**.

The annotated results for the models are compared in Figure 6.6. The results show that weighted CMF based algorithms perform significantly better in producing potentially clinical meaningful groupings. In an earlier work, Ho et. al. [71] show that for tensor valued data, Marble is very effective for phenotyping. The improved performance of CMF based algorithms compared to Marble signifies the shortcomings of approximating the tensor from flat files, besides the additional computational cost of factorizing higher order tensors. Moreover, the improved

performance of weighted SiCNMF compared to unweighted SiCNMF corroborates the efficacy of the weighing scheme described in Section 6.4.2.1.

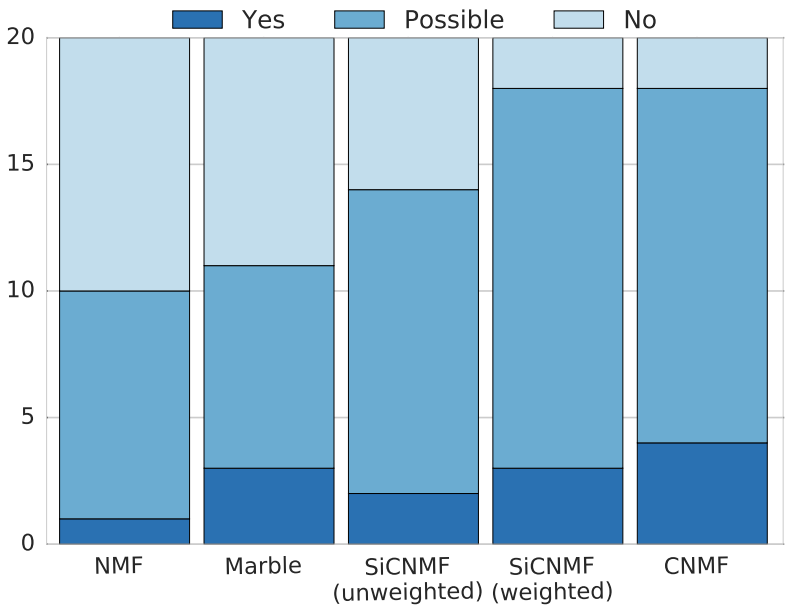



Figure 6.6: Distribution of the clinical relevance scores across the various models.

Although SiCNMF on data that contains only the case patients could potentially yield more clinically relevant phenotypes, the experiments were intended to demonstrate the unsupervised nature of the algorithm on a heterogeneous patient population.

Finally, Tables 6.4 and 6.5 show examples of phenotypes derived from weighted SiCNMF and CNMF, that were rated to be clinically meaningful by the domain experts.

| Diagnosis | Medication |
|---|---|
| ischemic heart disease; hypertension; disorders of lipoid metabolism; late effects of cerebrovascular disease; occlusion of cerebral arteries; | antihyperlipidemic agents; cholinesterase inhibitors; antianginal agents; analgesics; antiplatelet agents; |
| chronic airway obstruction, not elsewhere classified; other diseases of lung; dyspnea and respiratory abnormalities; pneumonia, organism unspecified; hypertension; | bronchodilators; antiarrhythmic agents; calcium channel blocking agents; antiviral agents; medical gas; |
| malignant neoplasm of colon; rheumatoid arthritis and other inflammatory polyarthropathies; malignant neoplasm of rectum, rectosigmoid junction, and anus; secondary malignant neoplasm of respiratory and digestive systems; disorders involving the immune mechanism; | immunosuppressive agents; antirheumatics; antimetabolites; antipsoriatics; adrenal cortical steroid; |

Table 6.4: Phenotypes from weighted–SiCNMF ($\eta = 500$) that were evaluated as "clinically meaningful" by a domain expert.

| Diagnosis | Medication |
|---|---|
| ischemic heart disease; hypertension; disorders of lipoid metabolism; unspecified chest pain; myocardial infarction; | antianginal agents; antihyperlipidemic agents; vasodilators; antiplatelet agents; angiotensin converting enzyme inhibitors |
| heart failure; atrial fibrillation and flutter; hypertension; pulmonary heart disease; dyspnea and respiratory abnormalities; | diuretics; antiarrhythmic agents; calcium channel blocking agents bronchodilators;aldosterone receptor antagonists; |
| malignant neoplasm of colon; rheumatoid arthritis and other inflammatory polyarthropathies; regional enteritis; malignant neoplasm of rectum, rectosigmoid junction, and anus; ulcerative colitis; | immunosuppressive agents; antirheumatics; analgesics; vitamins; antimetabolites; |
| chronic kidney disease (CKD); diabetes mellitus, type 2 Complications peculiar to certain specified procedures; other and unspecified anemias; diabetes mellitus, Type 1; | antidiabetic agents; miscellaneous antibiotics; sulfonamides; recombinant human erythropoietins; glucose elevating agents; |

Table 6.5: Phenotypes from CNMF (no sparsity constraints) that were evaluated as clinically meaningful by a domain expert.

# Chapter 7

# Collaborative Preference Completion from Partial Rankings

In this chapter, a novel and efficient algorithm for low rank matrix estimation in a collaborative learning to rank (LETOR) framework is developed, which involves jointly estimating individualized rankings for a set of entities over a shared set of items, based on a limited number of observed affinity values. The approach exploits the observation that while preferences are often recorded as numerical scores, the predictive quantity of interest is the underlying rankings. Thus, attempts to closely match the recorded scores may lead to overfitting and impair generalization performance. Instead, an estimator is proposed that directly fits the underlying rank order, combined with nuclear norm constraints to encourage low rank parameters. Besides (approximate) correctness of the ranking order, the proposed estimator makes no generative assumption on the numerical scores of the observations. One consequence is that the proposed estimator can fit any consistent entity–specific partial ranking over a subset of the items represented as a directed acyclic graph (DAG), generalizing standard techniques that can only fit preference scores. Despite this generality, for supervision representing total or blockwise total orders, the computational complexity of the proposed algorithm is within a $\log$ factor of the standard algorithms for nuclear norm regularization based estimates for matrix completion.

## 7.1 Introduction

Collaborative preference completion is the task of jointly learning bipartite (or dyadic) preferences of set of entities for a shared list of items, e.g., user–item interactions in a recommender system [56, 94]. It is commonly assumed that such entity–item preferences are generated from a small number of latent or hidden factors, or equivalently, the underlying preference value matrix is assumed to be low rank. Further, if the observed affinity scores from various explicit and implicit feedback are treated as exact (or mildly perturbed) entries of the unobserved preference value matrix, then the preference completion task naturally fits in the framework of low rank matrix completion [94, 167].

Recent research in the preference completion literature have noted that using a matrix completion estimator for collaborative preference estimation may be misguided [44, 141, 95] as the observed entity–item affinity scores from implicit/explicit feedback are potentially subject to systematic monotonic transformations arising from limitations in feedback collection, e.g., quantization and inherent biases. Such monotonic transformations can significantly increase the rank of the observed preference score matrix, thus adversely affecting recovery using low rank matrix completion methods [55]. Further, despite the common practice of measuring preferences using numerical scores, predictions are most often deployed or evaluated based on the item ranking e.g. in recommender systems, user recommendations are often presented as a ranked list of items without the underlying scores. Indeed several authors have shown that favorable empirical/theoretical performance in mean square error for the preference matrix often does not translate to better performance when performance is measured using ranking metrics [44, 141, 95]. Thus, collaborative preference estimation may be better posed as a collection of coupled *learning to rank (LETOR)* problems [105] that seek to jointly learn the preference

rankings of a set of entities, particularly exploiting the low dimensional latent structure of the underlying preference values.

This chapter considers preference completion in a general collaborative LETOR setting. Importantly, while the observations are assumed to be reliable indicators for relative preference ranking, their numerical scores may be quite deviant from the ground truth low rank preference matrix. Therefore, the aim in this chapter is to address preference completion under the following generalizations:

1. In a simple setting, for each entity, a score vector representing the its observed affinity interactions is assumed to be generated from an *arbitrary monotonic transformation* of the corresponding entries of the ground truth preference matrix. *No* further generative assumptions is made on the observed scores beyond monotonicity with respect to the underlying low rank preference matrix.
2. A more general setting is also considered, where observed preferences of each entity represent specifications of a *partial ranking* in the form of a directed acyclic graph (DAG) – the nodes represent a subset of items, and each edge represents a strict ordering between a pair of nodes. Such rankings may be encountered when the preference scores are consolidated from multiple sources of feedback, e.g., comparative feedback (pairwise or listwise) solicited for independent subsets of items. This generalized setting cannot be handled by standard matrix completion without some way of transforming the DAG orderings into a score vector.

This work is in part motivated by an application to neuroimaging meta-analysis as outlined in the following. Cognitive neuroscience aims to quantify the link between brain function with behavior. This interaction is most often measured in humans using Functional Magnetic Resonance Imaging (fMRI) experiments that measure brain activity in response to behavioral tasks. After analysis, the conclu-

sions are often summarized in neuroscience publications which include a table of brain locations that are most actively activated in response to an experimental stimulus. These results can then be synthesized using meta-analysis techniques to derive accurate predictions of brain activity associated with cognitive terms (also known as forward inference) and prediction of cognitive terms associated with brain regions (also known as reverse inference).

**Contributions:**

- A convex estimator for low rank preference completion is proposed using limited supervision, addressing: (a) arbitrary monotonic transformations of preference scores; and (b) partial rankings over items and simple generalization error bounds for a surrogate ranking loss that quantifies the trade–off between data–fit and regularization (Section 7.5).
- Efficient algorithms for the estimate are proposed under total and partially ordered observations. In the case of total orders, in spite of increased generality, the computational complexity of the proposed algorithm is within a $\log$ factor of the standard convex algorithms for matrix completion (Section 7.4).
- The proposed algorithm is evaluated for a novel application of identifying associations between brain–regions and cognitive terms from the neurosynth dataset [162] (Section 7.6). Such a large scale meta-analysis synthesizing information from the literature and related tasks has the potential to lead to novel insights into the role of brain regions in cognition and behavior.

## 7.2 Related Work

**Matrix Completion:** The bulk of the matrix completion works discussed in the precvious chapters including those in the context of ranking/recommendation applications focus on (a) fitting the observed numerical scores using squared loss,

and (b) evaluating the results on parameter/rating recovery metrics such as root mean squared error (RMSE). The shortcomings of such estimators and results using squared loss in ranking applications have been studied in some recent research [47, 44]. Motivated by collaborative ranking applications, there has been growing interest in addressing matrix completion within an explicit LETOR framework. [157] and [95] propose estimators that involve non–convex optimization problems and their algorithmic convergence and generalization behavior are not well understood. Some recent works provide parameter recovery guarantees for pairwise/listwise ranking observations under specific probabilistic distributional assumptions on the observed rankings [122, 119]. In comparison, the estimators and algorithms in this paper are agnostic to the generative distribution, and hence have much wider applicability.

**Learning to rank (LETOR):** LETOR is a structured prediction task of rank ordering relevance of a list of items as a function of pre–selected features [105]. Currently, leading algorithms for LETOR are listwise methods [31], which fully exploit the ranking structure of ordered observations, and offer better modeling flexibility compared to the pointwise [100] and pairwise methods [69, 82]. A recent listwise LETOR algorithm proposed the idea of monotone retargeting (MR) [5], which elegantly addresses listwise learning to rank (LETOR) task while maintaining the relative simplicity and scalability of pointwise estimation. MR was further extended to incorporate margins in the *margin equipped monotonic retargeting (MEMR)* formulation [4] to preclude trivial solutions that arise from scale invariance of the initial MR estimate in [5]. The estimator proposed in the paper is inspired from the idea of MR and will be revisited later in the paper. In collaborative preference completion, rather than learning a functional mapping from features to ranking, we seek to exploit the low rank structure in jointly modeling

the preferences of a collection of entities without access to preference indicative features.

**Single Index Models (SIMs)** Finally, literature on monotonic single index models (SIMs) also considers estimation under unknown monotonic transformations [73, 86]. However, algorithms for SIMs are designed to solve a harder problem of exactly estimating the non–parametric monotonic transformation and are evaluated for parameter recovery rather than the ranking performance. In general, with no further assumptions, sample complexity of SIM estimators lends them unsuitable for high dimensional estimation. The existing high dimensional estimators for learning SIMs typically assume Lipschitz continuity of the monotonic transformation which explicitly uses the observed score values in bounding the Lipsciptz constant of the monotonic transformation [84, 55]. In comparison, the proposed model is completely agnostic to the numerical values preference scores.

## 7.3 Preference Completion from Partial Rankings

Let the unobserved true preference scores of $d_2$ entities for $d_1$ items be denoted by a rank $r \ll \min\{d_1, d_2\}$ matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$. For each entity $j \in [d_2]$, a partial or total ordering of preferences for a subset of items denoted by $\mathcal{I}_j \subset [d_1]$ is observed. Let $n_j = |\mathcal{I}_j|$ denote the length of the ranked list of observed preferences associated with entity $j$, so that $\Omega_j = \{(i, j) : i \in \mathcal{I}_j\}$ denotes the entity-item index set for $j$, and $\Omega = \bigcup_j \Omega_j$ denotes the index set collected across entities. Let $\mathbb{P}_\Omega$ denote the sampling distribution for $\Omega$. The observed preferences of entity $j$ are typically represented by a listwise preference score vector $y^{(j)} \in \mathbb{R}^{n_j}$.

$$\forall j \in [d_2],\ y^{(j)} = g_j(\mathcal{P}_{\Omega_j}(\Theta^* + W)), \tag{7.1}$$

where each $(g_j)$ are an *arbitrary and unknown monotonic transformations*, and $W \in \mathbb{R}^{d_1 \times d_2}$ is some non–adversarial noise matrix sampled from the distribution $\mathbb{P}_W$. The *preference completion task* is to estimate a unseen rankings within each column of $\Theta^*$ from a subset of orderings $(\Omega_j, y^{(j)})_{j \in [d_2]}$.

As $(g_j)$ are arbitrary, the exact values of $(y^{(j)})$ are inconsequential, and the observed preference order can be specified by a constraint set parameterized by a margin parameter $\epsilon$ as follows:

**Definition 7.3.1** ($\epsilon$–margin Isotonic Set)**.** The following set of vectors are isotonic to $y \in \mathbb{R}^n$ with an $\epsilon > 0$ margin parameter:

$$\mathcal{R}^n_{\downarrow \epsilon}(y) = \{x \in \mathbb{R}^n : \forall\ i, k \in [n],\ y_i < y_k \Rightarrow x_i \leq x_k - \epsilon\}.$$

In addition to score vectors, isotonic sets of the form $\mathcal{R}^n_{\downarrow \epsilon}(y)$ are equivalently defined for any DAG $y = \mathcal{G}([n], E)$ which denotes a partial ranking among the vertices, with the convention that $(i, k) \in E \Rightarrow \forall x \in \mathcal{R}^n_{\downarrow \epsilon}(y),\ x_i \leq x_k - \epsilon$. Note that in Definition 7.3.1 that ties are *not* broken at random, e.g., if $y_{i_1} = y_{i_2} < y_k$, then $\forall x \in \mathcal{R}^n_{\downarrow \epsilon}(y), x_{i_1} \leq x_k - \epsilon, x_{i_2} \leq x_k - \epsilon$, but no particular ordering between $x_{i_1}$ and $x_{i_2}$ is specified.

Let $y_{(k)}$ denote the $k^{\text{th}}$ smallest entry of $y \in \mathbb{R}^n$. Three special cases of an observation $y$ representing a partial ranking over $[n]$ are distinguished.

(A) *Strict Total Order*: $y_{(1)} < y_{(2)} < \ldots < y_{(n)}$.

(B) *Blockwise Total Order*: $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$, with $K \leq n$ unique values.

(C) *Arbitrary DAG*: Partial order induced by a DAG $y = \mathcal{G}([n], E)$.

### 7.3.1 Monotone Retargeted Low Rank Estimator

Consider any scalable pointwise learning algorithm that fits a model to exact preferences scores. Since no generative model (besides monotonicity) is assumed

for the raw numerical scores in the observations, in principle, the scores $y^{(j)}$ for entity $j$ can be replaced or *retargeted* to any ranking-preserving scores, i.e., by any vector in $\mathcal{R}^{n_j}_{\downarrow\epsilon}(y^{(j)})$. Monotone Retargeting (MR) [5] exploits this observation to address the combinatorial listwise ranking problem [105] while maintaining the relative simplicity and scalability of pointwise estimates (regression). The key idea in MR is to alternately fit a pointwise algorithm to current relevance scores, and *retarget* the scores by searching over the space of all monotonic transformations of the scores. The proposed approach extends and generalizes monotone retargeting for the preference prediction task.

The algorithm is first motivated for the noise free setting, where it is clear that $\Theta^*_{\Omega_j} \in \mathcal{R}^{n_j}_{\downarrow\epsilon}(y^{(j)})$, so a candidate preference matrix $X$ lies in the intersection of (a) the data constraints from the observed preference rankings $\{X_{\Omega_j} \in \mathcal{R}^{n_j}_{\downarrow\epsilon}(y^{(j)})\}$, and (b) the model constraints – in this case low rankness induced by constraining the nuclear norm $\|X\|_*$. For robust estimation in the presence of noise, the noise free approach is extended by incorporating a soft penalty on constraint violations. Let $z \in \mathbb{R}^{|\Omega|}$, and with slight abuse of notation, let $z_{\Omega_j} \in \mathbb{R}^{n_j}$ denote vector of the entries of $z \in \mathbb{R}^{|\Omega|}$ corresponding to $\Omega_j \subset \Omega$. Upon incorporating the soft penalties, the monotone retargeted low rank estimator is given by:

$$\begin{aligned} \widehat{\mathcal{X}} = \underset{X}{\text{Argmin}} \ \min_{x \in \mathbb{R}^{|\Omega|}} \quad & \lambda \|X\|_* + \tfrac{1}{2}\|z - \mathcal{P}_\Omega(X)\|_2^2 \\ \text{s.t.} \quad & \forall j, \ z_{\Omega_j} \in \mathcal{R}^{n_j}_{\downarrow\epsilon}(y^{(j)}), \end{aligned} \tag{7.2}$$

where the parameter $\lambda$ controls the trade–off between nuclear norm regularization and data fit, and $\widehat{\mathcal{X}}$ is the set of minimizers of (7.2). The proposed estimate can handle arbitrary monotonic transformation of the preference scores and provides higher flexibility compared to the standard matrix completion.

Note that $\mathcal{R}^n_{\downarrow\epsilon}(y)$ is convex, and $\forall \lambda \geq 1$, the scaling $\lambda\mathcal{R}^n_{\downarrow\epsilon}(y) = \{\lambda x \ \forall \ x \in \mathcal{R}^n_{\downarrow\epsilon}(y)\} \subseteq \mathcal{R}^n_{\downarrow\epsilon}(y)$. Although (7.2) is specified in terms of two parameters, due to

the geometry of the problem, it turns out that $\lambda$ and $\epsilon$ are not jointly identifiable, as discussed in the following proposition.

**Proposition 7.3.1.** *The optimization in* (7.2) *is jointly convex in* $(X, z)$. *Further,* $\forall \gamma > 0$, $(\lambda, \gamma\epsilon)$ *and* $(\gamma^{-1}\lambda, \epsilon)$ *lead to equivalent estimators, specifically* $\widehat{\mathfrak{X}}(\lambda, \gamma\epsilon) = \gamma^{-1}\widehat{\mathfrak{X}}(\gamma^{-1}\lambda, \epsilon)$.

Since, positive scaling of $\widehat{\mathfrak{X}}$ preserves the resultant preference order, using Proposition 7.3.1 without loss of generality, only one of $\epsilon$ or $\lambda$ requires tuning with the other remaining fixed. In the experiments $\epsilon$ is chosen arbitrarily, and tune $\lambda$ using validation.

## 7.4 Optimization Algorithm

The optimization problem in (7.2) is jointly convex in $(X, z)$. Further, it is later shown that the proximal operator of the non–differential component of the estimate $\lambda\|X\|_* + \sum_j \mathbb{I}(z_{\Omega_j} \in \mathcal{R}^{n_j}_{\downarrow\epsilon}(y^{(j)}))$ is efficiently computable. This motivates using the proximal gradient descent algorithm [121] to jointly update $(X, z)$. A fixed step size of $\alpha = 1/2$ is used and the resulting updates are as follows:

- ***X* Update: Singular Value Thresholding** The proximal operator for $\tau\|.\|_*$ is the singular value thresholding operator $\mathcal{S}_\tau$. For $X$ with singular value decomposition $X = U\Sigma V$ and $\tau \geq 0$, $\mathcal{S}_\tau(X) = U s_\tau(\Sigma)V$, where $s_\tau$ is the soft thresholding operator given by $s_\tau(x)_i = \max\{x_i - \tau, 0\}$.
- ***z* Update: Parallel Projections** For hard constraints on $z$, the proximal operator at $v$ is the Euclidean projection on the constraints given by $z \leftarrow \text{argmin}_z \|z - v\|_2^2$, s.t. $z_{\Omega_j} \in \mathcal{R}^{n_j}_{\downarrow\epsilon}(y^{(j)})\ \forall j \in [d_2]$. These updates decouple along each entity (column) $z_{\Omega_j}$ and can be trivially parallelized. Efficient projections onto $\mathcal{R}^{n_j}_{\downarrow\epsilon}(y^{(j)})$ are discussed Section 7.4.1.

**Algorithm 3** Proximal Gradient Descent for (7.2) with input $\Omega, \{y_j^{(j)}\}, \epsilon$ and paramter $\lambda$

**for** $k = 0, 1, 2, \ldots,$ **Until** (stopping criterion)

$$X^{(k+1)} = \mathcal{S}_{\lambda/2}\Big(X^{(k)} + \frac{1}{2}(\mathcal{P}^*_\Omega(z^{(k)} - X^{(k)}_\Omega)\Big), \tag{7.3}$$

$$\forall j, z^{(k+1)}_{\Omega_j} = \text{Proj}_{\mathcal{R}^{n_j}_{\downarrow\epsilon}(y_j)}\Big(\frac{z^{(k)}_{\Omega_j} + X^{(k)}_{\Omega_j}}{2}\Big). \tag{7.4}$$

### 7.4.1 Projection onto $\mathcal{R}^n_{\downarrow\epsilon}(y)$

The following definitions that are used in characterizing $\mathcal{R}^n_{\downarrow\epsilon}(y)$.

**Definition 7.4.1** (Adjacent difference operator)**.** The adjacent difference operator in $\mathbb{R}^n$, denoted by $\boldsymbol{D_n} : \mathbb{R}^n \to \mathbb{R}^{n-1}$ is defined as $(\boldsymbol{D_n}x)_i = x_i - x_{i+1}$, for $i \in [n-1]$.

**Definition 7.4.2** (Incidence Matrix)**.** For a directed graph $\mathcal{G}(V, E)$, the incidence matrix $A_\mathcal{G} \in \mathbb{R}^{|V|\times|E|}$ is such that: if the $j^{\text{th}}$ directed edge $e_j \in E$ is from $i^{\text{th}}$ node to $k^{\text{th}}$ node, then $(A_\mathcal{G})_{ij} = 1$, $(A_\mathcal{G})_{kj} = -1$, and $(A_\mathcal{G})_{lj} = 0$, $\forall l \neq i$ or $k$.

Projection onto $\mathcal{R}^n_{\downarrow\epsilon}(y)$ is closely related to the *isotonic regression* problem of finding a univariate least squares fit under consistent order constraints (without margins). This isotonic regression problem in $\mathbb{R}^n$ can be solved exactly in $\mathcal{O}(n)$ complexity using the classical Pool of Adjacent Violators (PAV) algorithm [59, 16]

$$\text{PAV}(v) = \underset{z'\in\mathbb{R}^n}{\text{argmin}} ||z' - v||^2 \text{ s.t. } z'_i - z'_{i+1} \leq 0. \tag{7.5}$$

Simple adaptations of isotonic regression can be used for projection onto $\epsilon$-margin isotonic sets for the three special cases of interest as summarized in Table 7.1.

**(A) Strict Total Order:** $y_{(1)} < y_{(2)} < \ldots y_{(n)}$

In this setting, the constraint set can be characterized as $\mathcal{R}^n_{\downarrow\epsilon}(y) = \{x : \boldsymbol{D_n}x \leq$

$-\epsilon\mathbb{1}\}$, where $\mathbb{1}$ is a vector of ones. For this case projection onto $\mathcal{R}^n_{\downarrow\epsilon}(y)$ differs from (7.5) only in requiring an $\epsilon$–separation and a straight forward extension of the PAV algorithm [16] can be used. Let $\boldsymbol{d}^{\text{sl}} \in \mathbb{R}^n$ be any vector such that $\mathbb{1} = -\boldsymbol{D_n}\boldsymbol{d}^{\text{sl}}$, then by simple substitutions, $\text{Proj}_{\mathcal{R}^n_{\downarrow\epsilon}(y)}(x) = \text{PAV}(x - \epsilon\boldsymbol{d}^{\text{sl}}) + \epsilon\boldsymbol{d}^{\text{sl}}$.

**(B) Blockwise Total Order:** $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$

This is a common setting for supervision in many preference completion applications, where the listwise ranking preferences obtained from ratings over discrete quantized levels $1, 2, \ldots, K$, with $K \ll n$ are prevalent. Let $y$ be partitioned into $K \leq n$ blocks $P = \{P_1, P_2, \ldots P_K\}$, such that the entries of $y$ within each partition are equal, and the blocks themselves are strictly ordered,

$$\text{i.e., } \forall k \in [K],\ \sup y(P_{k-1}) < \inf y(P_k) = \sup y(P_k) < \inf y(P_{k+1}),$$

where $P_0 = P_{K+1} = \phi$, and $y(P) = \{y_i : i \in P\}$.

Let $\boldsymbol{d}^{\text{bl}} \in \mathbb{R}^n$ be such that $\boldsymbol{d}^{\text{bl}}_{\boldsymbol{i}} = \sum_{k=1}^{K} k\, \mathbb{I}_{i \in P_k}$ is a vector of block indices $\boldsymbol{d}^{\text{bl}} = [1, 1, ..|2, 2, ..|K, K, .., K]^\top$. Let $\Pi_P$ be a set of valid permutations that permute entries only within blocks $\{P_k \in P\}$, then $\mathcal{R}^n_{\downarrow\epsilon}(y) = \{x : \exists \pi \in \Pi_P, \boldsymbol{D_n}\pi(x) \leq -\epsilon\boldsymbol{D_n}\boldsymbol{d}^{\text{bl}}\}$. The following steps are proposed to compute $\widehat{z} = \text{Proj}_{\mathcal{R}^n_{\downarrow\epsilon}(y)}(x)$ in this case:

$$\begin{aligned} &\text{Step 1. } \pi^*(x) \text{ s.t. } \forall k \in [K],\ \pi^*(x)_{P_k} = \text{sort}(x_{P_k}) \\ &\text{Step 2. } \widehat{z} = PAV(\pi^*(x) - \epsilon\boldsymbol{d}^{\text{bl}}) + \epsilon\boldsymbol{d}^{\text{bl}}. \end{aligned} \tag{7.6}$$

The correctness of (7.6) is summarized by the following Lemma.

**Lemma 7.4.1.** *Estimate $\widehat{z}$ from (7.6) is the unique minimizer for*

$$\underset{z}{\textit{argmin}} \|z - x\|_2^2 \textit{ s.t. } \exists \pi \in \Pi_P : \boldsymbol{D_n}\pi(z) \leq \epsilon\boldsymbol{D_n}\boldsymbol{d}^{bl}.$$

**(C) Arbitrary DAG:** $y = \mathcal{G}([n], E)$

An arbitrary DAG (not necessarily connected) can be used to represent *any* consistent order constraints over its vertices, e.g., partial rankings consolidated from multiple listwise/pairwise scores. In this case, the $\epsilon$–margin isotonic set is given by $\mathcal{R}^n_{\downarrow_\epsilon}(y) = \{x : A_{\mathcal{G}}^\top x \leq -\epsilon \mathbb{1}\}$ (c.f. Definition 7.4.2). Consider $\boldsymbol{d}^{\text{DAG}} \in \mathbb{R}^n$ such that $i^{\text{th}}$ entry $\boldsymbol{d}_{\boldsymbol{i}}^{\text{DAG}}$ is the length of the longest directed chain connecting the topological descendants of the node $i$. It can be easily verified that, the isotonic regression algorithm for arbitrary DAGs applied on $x - \epsilon \boldsymbol{d}^{\text{DAG}}$ gives the projection onto $\mathcal{R}^n_{\downarrow_\epsilon}(y)$. In this most general setting, the best isotonic regression algorithm for exact solution requires $\mathcal{O}(nm^2 + n^3 \log n^2)$ computation [142], where $m$ is the number of edges in $\mathcal{G}$. While even in the best case of $m = o(n)$, the computation can be prohibitive, this case is included for completeness. Also note that this case of partial DAG ordering cannot be handled in the standard matrix completion setting without consolidating the partial ranks to total order.

| | $\mathcal{R}^n_{\downarrow_\epsilon}(y)$ | $\text{Proj}_{\mathcal{R}^n_{\downarrow_\epsilon}(y)}(x)$ | Computation |
|---|---|---|---|
| (A) | $\{x : \boldsymbol{D_n} x \leq -\epsilon \mathbb{1}\}$ | $\text{PAV}(x - \epsilon \boldsymbol{d}^{\text{sl}}) + \epsilon \boldsymbol{d}^{\text{sl}}$ | $\mathcal{O}(n)$ |
| (B) | $\left\{x : \begin{smallmatrix}\exists \pi \in \Pi_P, \\ \boldsymbol{D_n} \pi(x) \leq -\epsilon \mathbb{1}\end{smallmatrix}\right\}$ | $\pi_P^{*-1}\left(\text{PAV}(\pi_P^*(x) - \epsilon \boldsymbol{d}^{\text{bl}}) + \epsilon \boldsymbol{d}^{\text{bl}}\right)$ | $\mathcal{O}(n \log n)$ |
| (C) | $\{x : A_{\mathcal{G}}^\top x \leq -\epsilon \mathbb{1}\}$ | $\text{IsoReg}(x - \epsilon \boldsymbol{d}^{\text{DAG}}, \mathcal{G}) + \epsilon \boldsymbol{d}^{\text{DAG}}$ | $\mathcal{O}(n^2 m + n^3 \log n)$ |

Table 7.1: Summary of algorithms for $\text{Proj}_{\mathcal{R}^n_{\downarrow_\epsilon}(y)}(x)$

### 7.4.2 Computational Complexity

Let $H(X, z) := \frac{1}{2}\|\mathcal{P}_\Omega(X) - z\|_2^2$, and

$$f_\lambda(X) = \min_z \lambda \|X\|_* + H(X, z) \text{ s.t. } z_{\Omega_j} \in \mathcal{R}^{n_j}_{\downarrow_\epsilon}(y^{(j)}).$$

It can be derived that $\nabla_{X,z} H$ is 2–Lipscitz continuous. The following proposition is a standard result on convergence of proximal gradient descent [121].

**Proposition 7.4.2.** *If* $(X^{(k)}, z^{(k)})$ *are the sequence of updates from Algorithm 3, them for some* $\bar{X} \in \widehat{\mathcal{X}}_\lambda$, *(a)* $f_\lambda(X^{(k)}) - f_\lambda(\bar{X}) \leq \frac{1}{k}\|X^{(0)} - \bar{X}\|_F$, *and (b)* $\lim_{k\to\infty} X^{(k)} = \bar{X}$.

Compared to proximal algorithms for standard matrix completion [22, 108], the additional complexity in Algorithm 3 arises in the $z$ update (7.4), which is a simple substitution $z^{(k)} = X_\Omega^{(k)}$ in standard matrix completion. For total orders, the $z$ update of (7.4) is highly efficient and is asymptotically within an additional $\log|\Omega|$ factor of the computational costs for standard matrix completion.

## 7.5 Generalization Error

The estimator and the algorithms described so far are independent of the sampling distribution generating $(\Omega, \{y_j\})$. In this section a generalization error bound for (7.2) is derived under simple assumptions on the observations.

Recall that $y_j$ are (noisy) partial rankings of subset of items for each user, obtained from $g_j(\Theta_j^* + W_j)$ where $W$ is a noise matrix and $g_j$ are unknown and arbitrary transformations that only preserve that ranking order within each column. For simplicity, provide generalization error bounds for observations with linear order which can be analogously extended to partial order from DAGs.

### 7.5.1 Sampling

For a fixed noise matrix $W$ and ground truth $\Theta^*$, assume the following sampling distribution:

**Assumption 7.5.1** (Sampling $(\mathbb{P}_\Omega)$)**.** Let be $c_0$ a fixed constant and $R$ be pre–specified parameter denoting the length of single listwise observation. For $s =$

$1, 2, \ldots, |S| = c_0 d_2 \log d_2$,

$$\begin{aligned} j(s) \sim \text{uniform}[d_2], \quad \mathcal{I}(s) \sim \text{randsample}([d_1], R), \\ \Omega(s) = \{(i, j(s)) : i \in \mathcal{I}(s)\}, \quad y(s) = g_{j(s)}(\mathcal{P}_{\Omega(s)}(\Theta^* + W)). \end{aligned} \tag{7.7}$$

Further, the following notation is defined:

$$\forall j,\ \mathcal{I}_j = \bigcup_{s:j(s)=j} \mathcal{I}(s), \quad \Omega_j = \bigcup_{s:j(s)=j} \Omega(s), \ \text{ and } \ n_j = |\Omega_j|. \tag{7.8}$$

For each column $j$, the listwise scores $\{y(s) : j(s) = j\}$ jointly define a consistent partial ranking of $\mathcal{I}_j$ as the scores are subsets of a monotonically transformed preference vector $g_j(\Theta^*_j + W_j)$. This consistent ordering is represented by a DAG $y^{(j)} = \text{PartialOrder}(\{y(s) : j(s) = j\})$. Also note that $\mathcal{O}(d_2 \log d_2)$ samples ensures that each column is included in the sampling with high probability, This follows from standard concentration results on uniform sampling.

**Definition 7.5.1** (Projection Loss)**.** Let $y = \mathcal{G}([n], E)$ or $y \in \mathbb{R}^n$ define a partial ordering or total order in $\mathbb{R}^n$, respectively. The following convex surrogate loss is defined over the partial rankings.

$$\Phi(x, y) = \min_{z \in \mathcal{R}^n_{\downarrow_\epsilon}(y)} \|x - z\|_2$$

**Theorem 7.5.1** (Generalization Bound)**.** *Let $\widehat{X}$ be an estimate from (7.2). With appropriate scaling let $\|\widehat{X}\|_F = 1$ , then for constants $K_1$ $K_2$, the following holds with probability greater than $1 - \delta$ over all observed rankings $\{y^{(j)}, \Omega_j : j \in [d_2]\}$ drawn from (7.7) with $|S| \geq c_0 d_2 \log d_2$:*

$$\begin{aligned} \mathbb{E}_{y(s),\Omega(s)} \Phi(\widehat{X}_{\Omega(s)}, y(s)) \leq & \frac{1}{|S|} \sum_{s=1}^{|S|} \Phi(\widehat{X}_{\Omega(s)}, y(s)) \\ & + K_1 \frac{\|\widehat{X}\|_* \log^{1/4} d}{\sqrt{d_1 d_2}} \sqrt{\frac{d \log d}{R|S|}} + K_2 \sqrt{\frac{\log 2/\delta}{|S|}}. \end{aligned}$$

Theorem 7.5.1 quantifies the test projection loss over a random $R$ length items $\mathcal{I}(s)$ drawn for a random entity/user $j(s)$. The bound provides a trade–off between observable training error and complexity defined by nuclear norm of the estimate. Finally, in contrast to sample complexity results often seen for exact matrix completion (like [26, 127]), although, Theorem 7.5.1 does not provide a–priori guarantee on the performance the estimate, such generalization error bounds are useful to quantify test errors in terms of observable quantities of training error and estimate complexity [15]. Moreover, ground truth recovery guarantees in the style of [127, 26, 87] typically require additional assumptions on the generative model to uniquely identify a point estimate.

## 7.6 Experiments

**Movielens Dataset:** Movielens* is a movie recommendation website administered by GroupLens Research. The competitive benchmarked movielens 100K dataset † is used, which consists of 943 users and 1682 items, and Ratings are partially ordered – taking one of 5 values in the set $\{1.0, 2.0, \ldots, 5.0\}$. The two train/test splits provided with the dataset are used and the results are compared to the baselines of standard matrix completion (SMC) and ranking based collaborative filtering algorithm called Cofi-Rank [157]. In each split, a different list of 10 items are held out for each user in the test set. All algorithms rank the set of items for each user, and the score is averaged across users. Table 7.2 presents the results of evaluation on the Movielens dataset.

The models are evaluated on three ranking metrics: Kendall Tau, Normal-

*movielens.umn.edu
†www.grouplens.org/node/73

| | Kendall Tau | NDCG@5 | Precision@5 |
|---|---|---|---|
| **MR Preference Completion** | **0.3858 (0.0175)** | **0.8010 (0.0021)** | **0.7005 (0.0130)** |
| Standard Matrix Completion | 0.3584 (0.0012) | 0.7875 (0.0055) | 0.6897 (0.0022) |
| COFI-Rank | 0.3101 (0.0057) | 0.7711 (0.0008) | 0.6735 (0.0004) |

Table 7.2: Comparison of ranking performance on Movielens 100K dataset. Higher values are better.

ized Discounted Cummulative Gain or NDCG at top 5 and Precision at top 5 (See [105] for evaluation metrics). Note that test set consists of a list of 10 items per user. It can be observed that the proposed retargeted matrix completion (RMC) outperforms SMC and Cofi-Rank [157] in terms of Kendall Tau, NDCG@5 and Precision@5 [105].

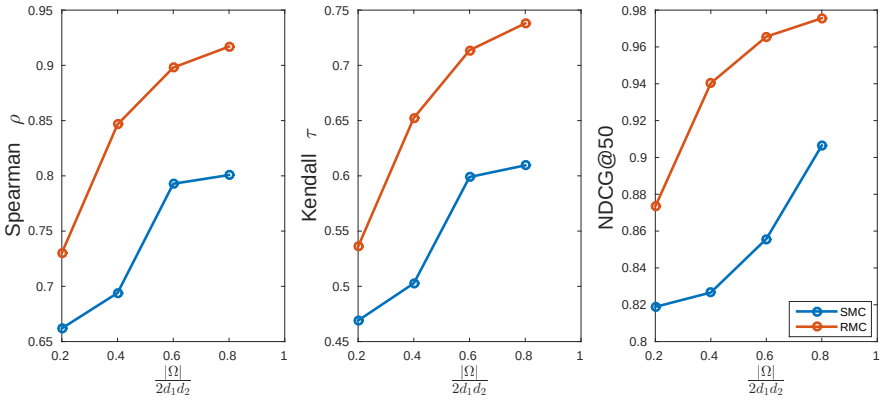



Figure 7.1: Comparison of ranking performance of proposed method, RMC for Retargeted Matrix Completion, with Standard Matrix Completion (SMC) using nuclear norm minimization. For all the three popular ranking metrics shown, higher values are better[5].

**Neurosynth Dataset:** As discussed in the introduction, from Neurosynth an association matrix between $\sim$ 3000 terms by 1000 consolidated brain regions is ex-

tracted, where the entries indicate the number of times a term/task and a brain region are mentioned in the same manuscript. Using this data, the reverse inference task is considered – the ranking of cognitive concepts for each brain region. As acknowledged by the developers of Neurosynth[163], the literature scraping approach is inherently noisy, and hence the numerical values may not be precise. Thus, ranking based approaches may be preferable to a parametric modeling approach. Further, the given brain regions $\times$ term matrix is inherently sparse, as many associations are never observed. The proposed estimator in (7.2) is compared against standard low rank matrix completion using nuclear norm minimization. The results in Fig. 7.1show that the proposed estimate from (7.2) significantly outperforms standard matrix completion in term of popular ranking metrics, namely Spearman $\tau$, Kendall $\rho$ and NDCG [5].

# Chapter 8

# Conclusions and Furture Work

This dissertation makes several contributions towards extending the statistical and empirical results on structured matrix estimation. In Chapters 3–5, tractable estimators with strong statistical guarantees are developed for matrix completion problems. The results in these chapters substantially extend the scope of provable matrix completion. The existing analysis for provable matrix completion are restricted to completion of a low rank structured matrix from observations under noiseless of additive thin–tailed noise distributions. In Chapter 3, the recovery results for matrix completion is extended to observations arising from a rich class of exponential family of distributions. This class of distributions are often the distributions of choice to model a variety of common datatypes and noise models. In Chapter 4, a unified statistical analysis is derived for matrix completion under general low dimensional structural constraints that can be enforced using any norm regularization. Several significant structures in high dimensional learning beyond low–rankness, including those arising from superposition structure, atomic norms and convex constraints can be analyzed under this framework. In Chapter 5, collective estimation from multiple sources of data is addressed. First non–trivial sample complexity result is derived for the collective matrix completion problem of learning completion of multiple matrices sharing a joint low dimensional structure. Intermediate results arising in the proofs of these chapters are of independent interest beyond the scope of this dissertation.

In Chapter 6, unsupervised, structured collective matrix factorization tools that incorporates various application specific constraints into a joint low rank factorization framework is proposed. Unsupervised learning approaches for automated phenotyping has the potential to enable improved clinical trials, properly target patients for screening tests and interventions, and support surveillance of infectious diseases. This framework is used for phenotype extraction from multi–source EHR data from Vanderbilt University. The clinical relevance of extracted candidate phenotypes is evaluated by domain experts and the results show improved performance over naive baselines. The utility of the phenotype descriptions is further quantitatively evaluated on the classification of case and control patients with Type-2 diabetes and hypertension.

Finally, in Chapter 7, algorithms and applications for the problem of collaboratively ranking multiple preference lists is discussed. A general setting is considered, wherein the observed preferences are either (a) numerical scores that are arbitrary monotonic transformations of the underlying low rank matrix values, or (b) DAG's representing partial orders. In both these cases, using matrix completion for estimating missing preferences is misguided or not applicable. A novel convex estimator efficient algorithm for the collaborative LETOR task is proposed, wherein a missing low rank preferences of the entities are learned by fitting the total or partial ordering of the observed preferences rather than the numerical scores. Remarkably, in the case of complete order, the complexity of our algorithm is within a $\log$ factor of the state–of–the–art algorithms for standard matrix completion. The efficacy proposed estimator is validated by experiments on real data applications.

## 8.1 Future Work

In Chapters 3–5, the matrix completion problems analyzed assume that the observations $(X_{ij})$ are sampled independently from the other entries. Further, the probability of observing a specific entry $X_{ij}$, under uniform sampling is independent of the noise channel or the distribution $\mathbb{P}(X_{ij}|\Theta^*_{ij})$. However, in some applications, it might be beneficial to have a sampling scheme involving dependencies among the observed entries as well among the sampled entries and the noise channel. In future work, it would be interesting to extend the analysis of this dissertation to such a dependent sampling settings.

The phenotyping applications discussed in Chapter 6, can also be extend along several directions. EHR data is often subject to noise and missing data. Further, in latent factor estimation using non-convex optimization algorithms, the estimated latent factors are typically interchangeable and lack identifiability. It is of interest to investigate algorithms under domain specific constraints for learning identifiable phenotypes that are robust to (a) missing data, (b) noise in data, and (c) varying patient populations. Moreover, in certain datasets a fully shared latent space is overly restrictive and models that allow for partial sharing of latent space could be explored in future work.

Finally, the collaborative learning to rank framework discussed in Chapter 7 could be extended along both theoretical and application domains. Stronger theoretical guarantees for the estimator and the algorithms under common cases of ranking observation are a potential line of exploration. Collaborative ranking problems that incorporate knowledge of features associated with entities and items could also be investigated within the monotone retargeting framework. Moreover, the preliminary results motivate future collaboration with neuroscience practitioners in extending the results towards developing systems for significant applications.

# Appendices

# Appendix A

# Proof of Results in Chapter 3

## A.1 Proof of Theorem 3.3.1

Proof of Theorem 3.3.1 involves two key steps:

- Show that, under assumptions Assumption 3.3.1–3.3.3, *RSC* of the form in Definition 2.1.5 holds for the loss function in (3.4) over a *large subset* of the solution space.
- When the RSC condition holds, the result follows from a few simple calculations; we handle the case where RSC does not hold separately.

Let $\widehat{\Delta} = \widehat{\Theta} - \Theta^*$. Recall the notation $\alpha_{\text{sp}}(\Delta) = \frac{\sqrt{d_1 d_2}\|\Delta\|_{\max}}{\|\Delta\|_F}$. Consider two cases, depending on whether the following condition holds for the constant $c_0 > 0$ in Theorem 3.3.1:

$$\alpha_{\text{sp}}(\widehat{\Delta}) \leq \frac{1}{c_0 \Psi(\overline{\mathcal{M}})} \sqrt{\frac{|\Omega|}{d \log d}}. \tag{A.1}$$

**Case 1:** Suppose condition in (A.1) does not hold; so that $\alpha_{\text{sp}}(\widehat{\Delta}) > \frac{1}{c_0 \Psi(\overline{\mathcal{M}})} \sqrt{\frac{|\Omega|}{d \log d}}$. From the constraints of the optimization problem (3.4), we have that $\|\widehat{\Delta}\|_{\max} \leq \|\widehat{\Theta}\|_{\max} + \|\Theta^*\|_{\max} \leq (2\alpha^*/\sqrt{d_1 d_2})$. Thus,

$$\|\widehat{\Delta}\|_F = \frac{\sqrt{d_1 d_2}\|\widehat{\Delta}\|_{\max}}{\alpha_{\text{sp}}(\widehat{\Delta})} \leq 2c_0 \alpha^* \sqrt{\frac{\Psi^2(\overline{\mathcal{M}}) d \log d}{|\Omega|}}. \tag{A.2}$$

**Case 2:** Suppose condition in (A.1) does hold. Then, the following theorem shows that an RSC condition of the form in Definition 2.1.5 holds.

**Theorem A.1.1** (Restricted Strong Convexity). *For $c_0$ given by Theorem 3.3.1, let $\alpha_{sp}(\widehat{\Delta}) \leq \frac{1}{c_0 \Psi(\overline{\mathcal{M}})} \sqrt{\frac{|\Omega|}{d \log d}}$. For large enough $c_0$, given any constant $\beta > 0$, there exists constant $K_\beta > 0$ such that, under the assumptions in Theorem 3.3.1, w.p. $> 1 - 4e^{-(1+\beta)\Psi_{\min}^4 \log^2 d}$:*

$$\frac{d_1 d_2}{|\Omega|} \sum_{ij \in \Omega} B_G(\widehat{\Theta}_{ij}, \Theta^*_{ij}) \geq \mu_{\mathcal{L}} \|\widehat{\Delta}\|_F^2,$$

*where* $\mu_{\mathcal{L}} = e^{-\frac{2\eta\alpha^*}{\sqrt{d_1 d_2}}} \left( K_\beta - \frac{64}{c_0} \sqrt{\frac{|\Omega| \kappa_{\mathfrak{R}}^2(d, |\Omega|)}{d \log d}} \right)$.

As noted earlier, such an RSC result for the special case of squared loss under low–rank constraints was shown in [115]. The theorem in Section A.3.

**Lemma A.1.2.** *Let $\widehat{\Theta}$ be the minimizer of (3.4). If $\frac{\lambda}{2} \geq \frac{d_1 d_2}{|\Omega|} \mathfrak{R}^*(\mathcal{P}_\Omega(Y - g(\Theta^*))$, then:*

$$\frac{d_1 d_2}{|\Omega|} \sum_{(i,j) \in \Omega} B_G(\widehat{\Theta}_{ij}, \Theta^*_{ij}) \leq \frac{3\lambda \Psi(\overline{\mathcal{M}})}{2} \|\Theta^* - \widehat{\Theta}\|_F$$

The proof is provided in Appendix A.4.1. □

**Remaining steps of the proof of Theorem 3.3.1:** Thus, if $\alpha_{\text{sp}}(\widehat{\Delta}) \leq \frac{1}{c_0 \Psi(\overline{\mathcal{M}})} \sqrt{\frac{|\Omega|}{d \log d}}$, and $\mu_{\mathcal{L}} > 0$, from Theorem A.1.1 and Lemma A.1.2, w.h.p.:

$$\mu_{\mathcal{L}} \|\widehat{\Delta}\|_F^2 \leq \frac{d_1 d_2}{|\Omega|} \sum_{ij \in \Omega} B_G(\widehat{\Theta}_{ij}, \Theta^*_{ij}) \leq \frac{3\lambda \Psi(\overline{\mathcal{M}})}{2} \|\widehat{\Delta}\|_F \tag{A.3}$$

From (A.2) and (A.3), under assumptions of Theorem 3.3.1, w.p. $> 1 - 4e^{-(1+\beta)\Psi_{\min}^4 \log^2 d}$,

$$\|\widehat{\Delta}\|_F^2 \leq \Psi^2(\overline{\mathcal{M}}) \max \left\{ \frac{3\lambda^2}{2\mu_{\mathcal{L}}^2}, \frac{\alpha^{*2} c_0^2 d \log d}{|\Omega|} \right\}.$$

## A.2 Proof of Corollary 3.3.2

Using the definition of $\overline{\mathcal{M}}^\perp$ in (3.6), $\overline{\mathcal{M}} = \text{span}\{\mathbf{u_i} x^\top, y \mathbf{v_j}^\top : x \in \mathbb{R}^n,\ y \in \mathbb{R}^m\}$. Let $P_{U^*} \in \mathbb{R}^{m \times m}$ and $P_{V^*} \in \mathbb{R}^{d \times d}$, be the projection matrices onto the column and row spaces ($U^*$, $V^*$) of $\Theta^*$, respectively. Then, $\forall X \in \mathbb{R}^{d_1 \times d_2}$, $X_{\overline{\mathcal{M}}} =$

$P_{U^*}X + XP_{V^*} - P_{U^*}XP_{V^*}$. Also, $\text{rk}(P_{U^*}) = \text{rk}(P_{V^*}) = \text{rk}(\Theta^*) = r$. Thus, $\forall \Phi \in \overline{\mathcal{M}}$, $\text{rk}(\Phi) \leq 2r$; and hence,

$$\Psi(\overline{\mathcal{M}}) = \sup_{\Phi \in \overline{\mathcal{M}} \setminus \{0\}} \frac{\|\Phi\|_*}{\|\Phi\|_F} \leq \sqrt{2r}. \text{ Further, } \Psi_{\min} = 1.$$

The following proposition by [115] is used to bound $\kappa_{\mathcal{R}}(d, |\Omega|)$ in Theorem 3.3.1.

**Lemma A.2.1.** *If $\Omega \subset [d_1] \times [d_2]$ is sampled using uniform sampling and $|\Omega| > d \log d$, then for a Rademacher sequence $\{\epsilon_{ij}, \forall (i,j) \in \Omega\}$,*

$$\mathbb{E}\Big[\frac{1}{|\Omega|}\|\sum_{ij\in\Omega} \sqrt{d_1 d_2}\epsilon_{ij} e_i e_j^*\|_2\Big] \leq 10\sqrt{\frac{d \log d}{|\Omega|}}.$$

This follows from Lemma 6 of [115], using $|\Omega| > d \log d$. □

Thus, for large enough $c_0 > 640$, using $\kappa_{\mathcal{R}}(d, |\Omega|) = 10\sqrt{\frac{d \log d}{|\Omega|}}$ in Theorem A.1.1, for some $K'_\beta > 0$,

$$\mu_{\mathcal{L}} = e^{-\frac{2\eta\alpha^*}{\sqrt{d_1 d_2}}}\left(K_\beta - \frac{640}{c_0}\right) = K'_\beta e^{-\frac{2\eta\alpha^*}{\sqrt{d_1 d_2}}}. \tag{A.4}$$

Finally, to prove the corollary, we derive a bound on $\|\mathcal{P}_\Omega(Y - g(\Theta^*))\|_2$ using the Ahlswede–Winter Matrix bound (Lemma 2.3.4). Let $\phi(x) = \psi_2(x) = e^{x^2} - 1$; and let $Z^{(ij)} \triangleq \sqrt{d_1 d_2}(Y_{ij} - g(\Theta^*_{ij}))e_i e_j^\top$, such that, $\frac{\sqrt{d_1 d_2}}{|\Omega|}\|\mathcal{P}_\Omega(Y - g(\Theta^*))\|_2 = \|\frac{1}{|\Omega|}\sum_{ij\in\Omega} Z^{(ij)}\|_2$.

From the equivalence of sub-Gaussian definitions Definition 2.3.2, there exists a constant $c_1$ such that $\|Y_{ij} - g(\Theta^*_{ij})\|_\phi \leq c_1 b$, $\forall (i,j)$. Since, $Z^{(ij)}$ has a single sub–Gaussian element $\sqrt{d_1 d_2}(Y_{ij} - g(\Theta^*_{ij}))$, $\|Z^{(ij)}\|_{\psi_2} \leq c_1\sqrt{d_1 d_2} b$. Further,

$$\mathbb{E}[Z^{(ij)^T} Z^{(ij)}] = \mathbb{E}[d_1 d_2 (Y_{ij} - g(\Theta^*_{ij}))^2 e_j e_j^*] \overset{(a)}{=} d_1 d_2 \mathbb{E}_{(ij\in\Omega)}\mathbb{E}_Y (Y_{ij} - g(\Theta^*_{ij}))^2 e_j e_j^*$$
$$\overset{(b)}{\leq} d_1 d_2 b^2 \mathbb{E}_{(ij\in\Omega)}[e_j e_j^*] \overset{(c)}{=} d_1 d_2 b^2 \frac{1}{d_2} I_{d_2 \times d_2}, \tag{A.5}$$

where $(a)$ follows from Fubini's Theorem, $(b)$ follows as $(Y_{ij} - g(\Theta^*_{ij}))$ is $b$-sub–Gaussian, and $(c)$ follows from the uniform sampling

model. Similarly, $\mathbb{E}[Z^{(ij)}Z^{(ij)^T}] = d_1 d_2 b^2 I_{d_1 \times d_1}$. Define $\sigma_{ij}^2 := \max\{\mathbb{E}[Z^{(ij)^T}Z^{(ij)}], \mathbb{E}[Z^{(ij)}Z^{(ij)^T}]\} \leq db^2$

In Lemma 2.3.4, using $\sigma^2 := \sum_{ij\in\Omega} \sigma_{ij}^2 = d|\Omega|b^2$, $M = c_1\sqrt{d_1 d_2}b \leq c_1 db$, and $t = |\Omega|\delta$,

$$\mathbb{P}\Big(\big\|\frac{1}{|\Omega|}\sum_{ij\in\Omega} Z^{(ij)}\big\|_2 \geq \delta\Big) \leq d^2 \max\Big\{e^{-\frac{\delta^2|\Omega|}{4db^2}}, e^{-\frac{\delta|\Omega|}{2c_1 db}}\Big\}.$$

If $|\Omega| > c_0 d\log d$ for large enough $c_0 > 0$, then for any constant $C$, using $\delta = Cb\sqrt{\frac{d\log d}{|\Omega|}}$,

$$\mathbb{P}\left(\frac{\sqrt{d_1 d_2}}{|\Omega|}\|\mathcal{P}_\Omega(Y - g(\Theta^*))\|_2 \geq Cb\sqrt{\frac{d\log d}{|\Omega|}}\right) \leq d^2 e^{-\frac{C^2}{4}\log d}. \tag{A.6}$$

Re-parameterizing the constants: for $\beta > 0$, $\exists C_\beta > 0$ such that with probability greater than $1 - e^{(1+\beta)\log d}$, $\frac{\sqrt{d_1 d_2}}{|\Omega|}\|\mathcal{P}_\Omega(Y - g(\Theta^*))\|_2 \leq C_\beta b\sqrt{\frac{d\log d}{|\Omega|}}$. Thus, using $\Psi_{\min} \geq 1$, $\mu_{\mathcal{L}} = K'_\beta e^{-\frac{2\eta\alpha^*}{\sqrt{d_1 d_2}}}$ (from (A.4)), and $\frac{\lambda}{2} := C_\beta\sqrt{d_1 d_2}b\sqrt{\frac{d\log d}{|\Omega|}}$ in Theorem 3.3.1 leads to the corollary.

## A.3 Proof of Theorem A.1.1

**Lemma A.3.1** (Lemma 1 of [114]). *Define the following subset:*

$$\mathcal{V} = \{\Theta \in \mathbb{R}^{d_1\times d_2} : \mathcal{R}(\Theta_{\overline{\mathcal{M}}^\perp}) \leq 3\mathcal{R}(\Theta_{\overline{\mathcal{M}}})\},$$

*where recall $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ from Assumption 3.3.1, and $\Theta_{\overline{\mathcal{M}}}$ is the projection of $\Theta$ onto the subspace $\overline{\mathcal{M}}$. If $\widehat{\Theta}$ is the minimizer of (3.4), and $\frac{\lambda}{2} \geq \frac{d_1 d_2}{|\Omega|}\mathcal{R}^*(\mathcal{P}_\Omega(Y - g(\Theta^*))$, then $\widehat{\Delta} = \widehat{\Theta} - \Theta^* \in \mathcal{V}$.* □

**Lemma A.3.2.** *Under Theorem A.1.1, consider the subset*

$$\mathcal{E} = \Big\{\Delta \in \mathcal{V} : \alpha_{sp}(\Delta) \leq \frac{1}{c_0\Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|}{d\log d}}, \|\Delta\|_F = 1\Big\}.$$

*Given any constant $\beta > 0$, there exists a constant $k_\beta > 0$, such that w.p. $> 1 - 4e^{-(1+\beta)\Psi^4_{\min}\log^2 d}$, $\forall\ \Delta \in \mathcal{E}$:*

$$\Big|\frac{d_1 d_2}{|\Omega|}\sum_{ij\in\Omega}\Delta_{ij}^2 - 1\Big| \leq \frac{16\mathcal{R}(\Delta)}{c_0\Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|\kappa_{\mathcal{R}}^2(d,|\Omega|)}{d\log d}} + \frac{k_\beta \mathcal{R}(\Delta)}{c_0^2\Psi(\overline{\mathcal{M}})}.$$

Proof is provided in Appendix A.4.2. □

From the assumptions in Theorem A.1.1 and Proposition A.3.1, $\frac{\widehat{\Delta}}{\|\widehat{\Delta}\|_F} \in \mathcal{E}$. Also, $\widehat{\Delta} \in \mathcal{V} \Rightarrow \mathcal{R}(\widehat{\Delta}) \leq \mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}}) + \mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}^\perp}) \leq 4\mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}}) \leq 4\Psi(\overline{\mathcal{M}})\|\widehat{\Delta}\|_F$. Further, $\forall\ (i,j) \in \Omega$, $\exists v_{ij} \in [0,1]$, s.t.

$$\begin{aligned} B_G(\widehat{\Theta}_{ij}, \Theta^*_{ij}) &= G(\widehat{\Theta}_{ij}) - G(\Theta^*_{ij}) - g(\Theta^*_{ij})(\widehat{\Theta}_{ij} - \Theta^*_{ij}) \\ &= \nabla^2 G((1-v_{ij})\Theta^*_{ij} + v_{ij}\widehat{\Theta}_{ij})\widehat{\Delta}_{ij}^2 \overset{(a)}{\geq} e^{-\frac{2\eta\alpha^*}{\sqrt{d_1 d_2}}}\widehat{\Delta}_{ij}^2. \end{aligned} \tag{A.7}$$

where $(a)$ holds as $|(1-v_{ij})\Theta^*_{ij} + v_{ij}\widehat{\Theta}_{ij}| \leq \|\Theta^*\|_{\max} + \|\widehat{\Theta}\|_{\max} \leq \frac{2\alpha^*}{\sqrt{d_1 d_2}}$, and $\nabla^2 G(u) \geq e^{-\eta|u|}$ (Assumption 3.3.2).

Using Lemma A.3.2 and (A.7), for large enough $c_0$, if $|\Omega| > c_0\Psi^2(\overline{\mathcal{M}})d\log d$, then $K_\beta := 1 - \frac{4k_\beta}{c_0^2} > 0$. Finally, let $\mu_{\mathcal{L}} := e^{-\frac{2\eta\alpha^*}{\sqrt{d_1 d_2}}}\Big(K_\beta - \frac{64}{c_0}\sqrt{\frac{|\Omega|\kappa_{\mathcal{R}}^2(d,|\Omega|)}{d\log d}}\Big)$; if $\mu_{\mathcal{L}} > 0$, then w.h.p., $\frac{d_1 d_2}{|\Omega|}\sum_{ij\in\Omega} B_G(\widehat{\Theta}_{ij}, \Theta^*_{ij}) \geq \mu_{\mathcal{L}}\|\widehat{\Delta}\|_F^2$.

## A.4 Proofs of Lemma

### A.4.1 Proof of Lemma A.1.2

Let $\widehat{\Delta} = \widehat{\Theta} - \Theta^*$.

$$\begin{aligned} \mathcal{R}(\widehat{\Theta}) &= \mathcal{R}\big(\Theta^* + \widehat{\Delta}_{\overline{\mathcal{M}}} + \widehat{\Delta}_{\overline{\mathcal{M}}^\perp}\big) \geq \mathcal{R}\big(\Theta^* + \widehat{\Delta}_{\overline{\mathcal{M}}^\perp}\big) - \mathcal{R}\big(\widehat{\Delta}_{\overline{\mathcal{M}}}\big) \\ &= \mathcal{R}\big(\Theta^*\big) + \mathcal{R}\big(\widehat{\Delta}_{\overline{\mathcal{M}}^\perp}\big) - \mathcal{R}\big(\widehat{\Delta}_{\overline{\mathcal{M}}}\big) \end{aligned} \tag{A.8}$$

The above inequalities hold due to triangle inequality, and decomposability of $\mathcal{R}$ over $\Theta^* \in \mathcal{M}$ and $\Delta_{\overline{\mathcal{M}}^\perp} \in \overline{\mathcal{M}}^\perp$.

$$
\begin{aligned}
&\frac{d_1 d_2}{|\Omega|} \sum_{(i,j)\in\Omega} B_G(\widehat{\Theta}_{ij}, \Theta^*_{ij}) \\
&= \frac{d_1 d_2}{|\Omega|} \big[ \sum_{(i,j)\in\Omega} G(\widehat{\Theta}_{ij}) - Y_{ij}\widehat{\Theta}_{ij} - G(\Theta^*_{ij}) + Y_{ij}\Theta^*_{ij} + \langle \mathcal{P}_\Omega(Y - g(\Theta^*), \widehat{\Delta} \rangle \big] \\
&\overset{(a)}{\leq} \lambda\mathcal{R}(\Theta^*) - \lambda\mathcal{R}(\widehat{\Theta}) + \frac{d_1 d_2}{|\Omega|}\mathcal{R}^*(\mathcal{P}_\Omega(Y - g(\Theta^*)))\mathcal{R}(\widehat{\Delta}) \\
&\overset{(b)}{\leq} \lambda\mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}}) - \lambda\mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}^\perp}) + \frac{\lambda}{2}\mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}} + \widehat{\Delta}_{\overline{\mathcal{M}}^\perp}) \overset{(c)}{\leq} \frac{3\lambda}{2}\mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}}) - \frac{\lambda}{2}\mathcal{R}(\widehat{\Delta}_{\overline{\mathcal{M}}^\perp}) \\
&\leq \frac{3\lambda\Psi(\overline{\mathcal{M}})}{2}\|\Theta^* - \widehat{\Theta}_{\overline{\mathcal{M}}}\|_F \leq \frac{3\lambda\Psi(\overline{\mathcal{M}})}{2}\|\Theta^* - \widehat{\Theta}\|_F
\end{aligned} \tag{A.9}
$$

where $(a)$ follows as $\widehat{\Theta}$ is the minimizer of (3.4) and using Cauchy Schwartz, $(b)$ follows from (A.8) and using $\frac{d_1 d_2}{|\Omega|}\mathcal{R}^*(\mathcal{P}_\Omega(Y - g(\Theta^*)) \leq \frac{\lambda}{2}$, and $(c)$ follows from triangle inequality. □

### A.4.2 Proof of Lemma A.3.2

Recall that $\mathcal{V} = \{\Delta : \mathcal{R}(\Delta_{\overline{\mathcal{M}}^\perp}) \leq 3\mathcal{R}(\Delta_{\overline{\mathcal{M}}})\}$. To prove Lemma 4, consider the nuclear norm ball $S_{\mathcal{R}}(t) = \{\Delta : \mathcal{R}(\Delta) \leq t\}$.

1. Show that, $\mathbb{P}\left( \sup_{\Delta\in\mathcal{E}\cap S_{\mathcal{R}}(t)} \left| \frac{d_1 d_2}{|\Omega|} \sum_{ij\in\Omega} \Delta_{ij}^2 - 1 \right| > \frac{8t}{c_0\Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|\kappa^2(d,|\Omega|)}{d\log d}} + \frac{k_\beta t}{2c_0^2\Psi(\overline{\mathcal{M}})} \right)$ is small; where $\kappa(d, |\Omega|)$ is a quantity that depends only on the dimensions $d$ and $|\Omega|$. This is done by:
   (a) Bounding the expectation, $\mathbb{E}\Big[ \sup_{\Delta\in\mathcal{E}\cap S_{\mathcal{R}}(t)} \left| \frac{d_1 d_2}{|\Omega|} \sum_{ij\in\Omega} \Delta_{ij}^2 - 1 \right| \Big]$
   (b) Showing an exponential decay of the tail.
2. Then use a peeling argument [125] to derive at the result in Lemma A.3.2.

#### A.4.2.1 Bounding Expectation

Note that $\forall\ \Delta \in \mathcal{E},\ \mathbb{E}[\frac{d_1 d_2}{|\Omega|}\sum_{ij\in\Omega}\Delta_{ij}^2] = \|\Delta\|_F^2 = 1$. Thus, by using standard symmetrization argument (Lemma 6.3 of [97], with a Rademacher sequence, $\{\epsilon_{ij}, \forall\ ij \in \Omega\}$, we have:

$$\mathbb{E}\Big[\sup_{\Delta\in\mathcal{E}\cap S_{\mathfrak{R}}(t)}\Big|\frac{d_1 d_2}{|\Omega|}\sum_{ij\in\Omega}\Delta_{ij}^2 - 1\Big|\Big] \le \frac{2d_1 d_2}{|\Omega|}\mathbb{E}\Big[\sup_{\Delta\in\mathcal{E}\cap S_{\mathfrak{R}}(t)}\Big|\sum_{ij\in\Omega}\epsilon_{ij}\Delta_{ij}^2\Big|\Big] \tag{A.10}$$

Also, $\forall\Delta \in \mathcal{E}$, $\phi_{ij}(\Delta) \triangleq \frac{\Delta_{ij}^2}{2\sup\limits_{\Delta\in\mathcal{E}}\|\Delta\|_{\max}}$ is a contraction, and $\forall\Delta \in \mathcal{E}$, $\|\Delta\|_{\max} = \frac{\alpha_{\text{sp}}(\Delta)}{\sqrt{d_1 d_2}} \le \frac{1}{c_0\Psi(\overline{\mathcal{M}})\sqrt{d_1 d_2}}\sqrt{\frac{|\Omega|}{d\log d}}$.

Thus, using Theorem 4.12 of [97] in (A.10),

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\Delta\in\mathcal{E}\cap S_{\mathfrak{R}}(t)}\Big|\frac{d_1 d_2}{|\Omega|}\sum_{ij\in\Omega}\Delta_{ij}^2 - 1\Big|\Big] \\
&\le \frac{8}{c_0\Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|}{d\log d}}\mathbb{E}\left[\sup_{\Delta\in\mathcal{E}\cap S_{\mathfrak{R}}(t)}\Big|\frac{\sqrt{d_1 d_2}}{|\Omega|}\Big\langle\sum_{ij\in\Omega}\epsilon_{ij}e_i e_j^*, \Delta\Big\rangle\Big|\right] \\
&\overset{(a)}{\le} \frac{8t}{c_0\Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|}{d\log d}}\mathbb{E}\left[\frac{\sqrt{d_1 d_2}}{|\Omega|}\mathcal{R}^*\Big(\sum_{ij\in\Omega}\epsilon_{ij}e_i e_j^*\Big)\right]
\end{aligned} \tag{A.11}$$

where $(a)$ follows from Cauchy–Schwartz and as $\mathcal{R}(\Delta) \le t$. Note that $\mathcal{R}^*\Big(\sum_{ij\in\Omega}\epsilon_{ij}e_i e_j^*\Big)$ is independent of $\Delta$ and depends only on $d$ and $|\Omega|$. Let $\kappa(d, |\Omega|) \ge \mathbb{E}\left[\frac{\sqrt{d_1 d_2}}{|\Omega|}\mathcal{R}^*\Big(\sum_{ij\in\Omega}\epsilon_{ij}e_i e_j^*\Big)\right]$ be a suitable upper bound.

$$\mathbb{E}\Big[\sup_{\Delta\in\mathcal{E}\cap S_{\mathfrak{R}}(t)}\Big|\frac{d_1 d_2}{|\Omega|}\sum_{ij\in\Omega}\Delta_{ij}^2 - 1\Big|\Big] \le \frac{8t}{c_0\Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|\kappa^2(d, |\Omega|)}{d\log d}} \tag{A.12}$$

#### A.4.2.2 Tail Behavior

Let $G_t(\Omega) \triangleq \sup_{\Delta \in \mathcal{E} \cap S_{\mathcal{R}}(t)} \left| \frac{d_1 d_2}{|\Omega|} \sum_{ij \in \Omega} \Delta_{ij}^2 - 1 \right|$. Let $\Omega' \subset [d_1] \times [d_2]$ be another set of indices that differ from $\Omega$ in exactly one element. We then have:

$$\begin{aligned}
G_t(\Omega) - G_t(\Omega') &= \sup_{\Delta \in \mathcal{E} \cap S_{\mathcal{R}}(t)} \left| \frac{d_1 d_2}{|\Omega|} \sum_{ij \in \Omega} \Delta_{ij}^2 - 1 \right| - \sup_{\Delta \in \mathcal{E} \cap S_{\mathcal{R}}(t)} \left| \frac{d_1 d_2}{|\Omega|} \sum_{kl \in \Omega'} \Delta_{kl}^2 - 1 \right| \\
&\leq \frac{d_1 d_2}{|\Omega|} \sup_{\Delta \in \mathcal{E} \cap S_{\mathcal{R}}(t)} \left( \left| \sum_{ij \in \Omega} \Delta_{ij}^2 - 1 \right| - \left| \sum_{kl \in \Omega'} \Delta_{kl}^2 - 1 \right| \right) \\
&\leq \frac{d_1 d_2}{|\Omega|} \sup_{\Delta \in \mathcal{E} \cap S_{\mathcal{R}}(t)} \left( \left| \sum_{ij \in \Omega} \Delta_{ij}^2 - \sum_{kl \in \Omega'} \Delta_{kl}^2 \right| \right) \\
&\leq \frac{2 d_1 d_2}{|\Omega|} \sup_{\Delta \in \mathcal{E} \cap S_{\mathcal{R}}(t)} \|\Delta\|_{\max}^2 \leq \frac{2}{c_0^2 \Psi^2(\overline{\mathcal{M}}) d \log d}
\end{aligned}$$

By similar arguments on $G_t(\Omega') - G_t(\Omega)$, we conclude that $|G_t(\Omega) - G_t(\Omega')| \leq \frac{2}{c_0^2 \Psi^2(\overline{\mathcal{M}}) d \log d}$. Therefore, using Mc Diarmid's inequality, we have $\mathbb{P}(|G_t(\Omega) - \mathbb{E}[G_t(\Omega)]| > \delta) \leq 2 \exp\left( - \frac{c_0^4 \delta^2 \Psi^4(\overline{\mathcal{M}}) d^2 \log^2 d}{2|\Omega|} \right)$. Fix $\delta = \frac{2 k_1 t}{c_0^2 \Psi(\overline{\mathcal{M}})}$ for appropriate constant $k_1$. Recall that $\Psi_{\min} = \inf_{X \setminus \{0\}} \frac{\mathcal{R}(X)}{\|X\|_F} \leq \Psi(\overline{\mathcal{M}})$. Using $|\Omega| \leq d^2$,

$$\mathbb{P}\Big( G_t(\Omega) > \frac{8t}{c_0 \Psi(\overline{\mathcal{M}})} \sqrt{\frac{|\Omega| \kappa^2(d, |\Omega|)}{d \log d}} + \frac{2 k_1 t}{c_0^2 \Psi(\overline{\mathcal{M}})} \Big) \leq 2 \exp\left( -2 k_1^2 t^2 \Psi_{\min}^2 \log^2 d \right)$$

#### A.4.2.3 Peeling Argument

Consider the following sets, $S_\ell = \{\Delta \in \mathcal{E} : 2^{\ell-1} \Psi_{\min} \leq \mathcal{R}(\Delta) \leq 2^\ell \Psi_{\min}\}$, for all (integers) $\ell \geq 1$. Since, $\forall \Delta \in \mathcal{E}$, $\mathcal{R}(\Delta) \geq \Psi_{\min} \|\Delta\|_F = \Psi_{\min}$, for each $\Delta \in \mathcal{E}$, $\Delta \in S_\ell$ for some $\ell \geq 1$. Further, if for some $\Delta \in \mathcal{E}$, $\left| \frac{d_1 d_2}{|\Omega|} \sum_{ij \in \Omega} \Delta_{ij}^2 - 1 \right| > \frac{16 \mathcal{R}(\Delta)}{c_0 \Psi(\overline{\mathcal{M}})} \sqrt{\frac{|\Omega| \kappa^2(d, |\Omega|)}{d \log d}} + \frac{4 k_1 \mathcal{R}(\Delta)}{c_0^2 \Psi(\overline{\mathcal{M}})}$, then for some $\ell$:

$$\begin{aligned}
\left| \frac{d_1 d_2}{|\Omega|} \sum_{ij \in \Omega} \Delta_{ij}^2 - 1 \right| &> \frac{16 (2^{\ell-1}) \Psi_{\min}}{c_0 \Psi(\overline{\mathcal{M}})} \sqrt{\frac{|\Omega| \kappa^2(d, |\Omega|)}{d \log d}} + \frac{4 k_1 2^{\ell-1} \Psi_{\min}}{c_0^2 \Psi(\overline{\mathcal{M}})} \\
&= \frac{8 (2^\ell \Psi_{\min})}{c_0 \Psi(\overline{\mathcal{M}})} \sqrt{\frac{|\Omega| \kappa^2(d, |\Omega|)}{d \log d}} + \frac{2 k_1 (2^\ell \Psi_{\min})}{c_0^2 \Psi(\overline{\mathcal{M}})}
\end{aligned} \tag{A.13}$$

Thus,

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{\Delta\in\mathcal{E}}\bigg|\frac{d_1d_2}{|\Omega|}\sum_{ij\in\Omega}\Delta_{ij}^2-1\bigg|>\frac{16\mathcal{R}(\Delta)}{c_0\Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|\kappa^2(d,|\Omega|)}{d\log d}}+\frac{4k_1\mathcal{R}(\Delta)}{c_0^2\Psi(\overline{\mathcal{M}})}\bigg)\\
&\quad\leq\sum_{\ell=1}^{\infty}\mathbb{P}\bigg(G_{2^\ell}(\Omega)>\frac{8(2^\ell\Psi_{\min})}{c_0\Psi(\overline{\mathcal{M}})}\sqrt{\frac{|\Omega|\kappa^2(d,|\Omega|)}{d\log d}}+\frac{2k_1(2^\ell\Psi_{\min})}{c_0^2\Psi(\overline{\mathcal{M}})}\bigg)\\
&\quad\leq\sum_{\ell=1}^{\infty}2\exp\left(-2k_1^2 2^{2l}\Psi_{\min}^4\log^2 d\right)\overset{(a)}{\leq}\sum_{\ell=1}^{\infty}2\exp(-4\log 2\;k_1^2\ell\Psi_{\min}^4\log^2 d)\\
&\quad\leq\frac{2e^{-4k_1^2\Psi_{\min}^4\log^2 d}}{1-e^{-4k_1^2\Psi_{\min}^4\log^2 d}}\leq 4e^{-4k_1^2\Psi_{\min}^4\log^2 d}
\end{aligned}\tag{A.14}$$

where $(a)$ follows as $x\geq\log x$ for $x>1$, and the last step holds for $d>1$. The lemma follows by re-parametrization of constants in terms of $\beta$.

# Appendix B

# Proof of Results in Chapter 4

## B.1 Results from Generic Chaining

In this section, $K$ denotes a universal constant, not necessarily the same at each occurrence.

**Definition B.1.1** (Gamma Functional (Definition 2.2.19 in [145]))**.** Given a complete pseudometric space $(T, d)$, an *admissible sequence* is an increasing sequence $(\mathcal{A}_n)$ of partitions of $T$ such that $|\mathcal{A}_0| = 1$ and $|\mathcal{A}_n| \leq 2^{2^n}$ for $n \geq 1$. For $\alpha > 0$, define the Gamma functional $\gamma_\alpha(T, d)$ as follows:

$$\gamma_\alpha(T, d) = \inf_{(\mathcal{A}_n)_{n\geq 0}} \sup_{t\in T} \sum_{n\geq 0} 2^{n/\alpha} \Delta_d(A_n(t)), \tag{B.1}$$

where $\inf$ is over all admissible sequences $(\mathcal{A}_n)$, $A_n(t)$ is the unique element of $\mathcal{A}_n$ that contains $t$, and $\Delta_d(A)$ is the diameter of the set $A$ measured in metric $d$.

**Lemma B.1.1** (Majorizing Measures Theorem (Theorem 2.4.1 in [145]))**.** *Given a closed set $T$ in a metric space, let $(X_t)_{t\in T}$ be a centered Gaussian process indexed by $t \in T$, i.e. $(X_t)$ are jointly Gaussian. For $s, t \in T$, let $d_X(s, t) := \sqrt{\mathbb{E}(X_s - X_t)^2}$ denote the canonical pseudometric associated with $(X_t)$. Then,*

$$\frac{1}{K}\gamma_2(T, d_X) \leq \mathbb{E} \sup_{t\in T} X_t \leq K\gamma_2(T, d_X).$$

*In particular, for the canonical Gaussian process $(\sum_i t_i g_i)_{t\in T}$,*

$$\frac{1}{K}\gamma_2(T, \|.\|_F) \leq w_G(T) \leq K\gamma_2(T, \|.\|_F).$$

**Lemma B.1.2** (Theorem 2.4.12 in [145]). *Let $(X_t)_{t\in T}$ be a centered Gaussian process with canonical distance $d_X = \sqrt{\mathbb{E}(X_s - X_t)^2}$. Let $(Y_t)_{t\in T}$ be another centered process indexed by the same set $T$, such that*

$$\forall s,t \in T, u > 0, \quad \mathbb{P}(|Y_s - Y_t| > u) \leq 2\exp\Big(-\frac{u^2}{2d_X^2(s,t)}\Big),$$

*then, $\mathbb{E}\sup_{s,t\in T}|Y_s - Y_t| \leq K\mathbb{E}\sup_{t\in T} X_t$. If further, $(Y_t)_{t\in T}$ is symmetric, then $\mathbb{E}\sup_t |Y_t| \leq \mathbb{E}\sup_{s,t\in T}|Y_s - Y_t| = 2\mathbb{E}\sup_{t\in T} Y_t$.*

**Note:** From the definition of sub–Gaussian random variables (Section 2.3.1.2), using the above lemma, sub–Gaussian complexity measures can be directly bounded by Gaussian complexities.

**Lemma B.1.3** (Theorem 3.1.4 in [145]). *Let $T$ be a compact set with non–empty interior. Consider a translation invariant random distance $d_\omega$ on $T$, that depends on a random parameter $\omega$; and let $d(s,t) = \sqrt{\mathbb{E}d_\omega^2(s,t)}$, then :*

$$\big(\mathbb{E}\gamma_2^2(T,d_\omega)\big)^{1/2} \leq K\gamma_2(T,d) + K\big(\mathbb{E}\sup_{s,t\in T} d_\omega^2(s,t)\big)^{1/2}$$

## B.2 Proof of Theorem 4.3.2

Let the entries of $\Omega = \{E_s = e_{i_s}e_{j_s}^\top : s = 1,2,\ldots,|\Omega|\}$ be sampled from (4.2). Define the following random process over $S$:

$$(\mathcal{X}_{\Omega,g}(X))_{X\in S}, \text{where } \mathcal{X}_{\Omega,g}(X) = \langle X, \mathcal{P}_\Omega^*(g)\rangle = \textstyle\sum_s \langle X, E_s\rangle g_s. \tag{B.2}$$

The following lemmata are proved in Appendix B.5.

**Lemma B.2.1.** *For a compact subset $S \subseteq \mathbb{R}^{d_1\times d_2}$ with non–empty interior, $\exists$ constants $k_1$, $k_2$ such that:*

$$w_{\Omega,g}(S) = \mathbb{E}\sup_{X\in S}\mathcal{X}_{\Omega,g}(X) \leq k_1\sqrt{\frac{|\Omega|}{d_1d_2}}w_G(S) + k_2\sqrt{\mathbb{E}\sup_{X,Y\in S}\|\mathcal{P}_\Omega(X-Y)\|_2^2}.$$

**Lemma B.2.2.** *There exists constants $k_3$, $k_4$, such that for compact $S \subseteq \mathbb{B}^{d_1 d_2}$ with non–empty interior*

$$\mathbb{E} \sup_{X,Y \in S} \|\mathcal{P}_\Omega(X-Y)\|_2^2 \leq k_3 \frac{|\Omega|}{d_1 d_2} w_G^2(S) + k_4 (\sup_{X,Y \in S} \|X-Y\|_\infty) w_{\Omega,g}(S)$$

From Lemma B.2.2, the following holds:

$$\begin{aligned}\sqrt{\mathbb{E} \sup_{X,Y \in S} \|\mathcal{P}_\Omega(X-Y)\|_2^2} &\overset{(a)}{\leq} K_3 \sqrt{\frac{|\Omega|}{d_1 d_2}} w_G(S) + \sqrt{k_4 w_{\Omega,g}(S) \sup_{X,Y \in S} \|X-Y\|_\infty} \\ &\overset{(b)}{\leq} K_3 \sqrt{\frac{|\Omega|}{d_1 d_2}} w_G(S) + K_4 (\sup_{X,Y \in S} \|X-Y\|_\infty) + \frac{1}{2} w_{\Omega,g}(S), \end{aligned} \tag{B.3}$$

where $(a)$ follows from triangle inequality, $(b)$ using $\sqrt{ab} \leq a/2 + b/2$. Bound on $w_{\Omega,g}(S)$ in Theorem 4.3.2 follows by using (B.3) in Lemma B.2.1.

The statement in Theorem 4.3.2 about partial sub–Gaussian complexity follows from a standard result in empirical process given in Lemma B.1.2.

## B.3 Proof of Theorem 4.3.1

Define the following set of $\beta$–*non–spiky* matrices in $\mathbb{R}^{d_1 \times d_2}$ for constant $c_0$ from Theorem 4.3.1:

$$\mathbb{A}(\beta) = \left\{ X : \alpha_{\text{sp}}(X) = \frac{\sqrt{d_1 d_2} \|X\|_\infty}{\|X\|_F} < \beta \right\}. \tag{B.4}$$

Define, $$\beta_{c_0}^2 = \sqrt{\frac{|\Omega|}{c_0^2 w_G^2(\mathcal{E}_\Re) \log d}} \tag{B.5}$$

**Case 1: Spiky Error Matrix** When the error matrix from (4.5) or (4.6) has large spikiness ratio, following bound on error is immediate using $\|\widehat{\Delta}\|_\infty \leq \|\widehat{\Theta}\|_\infty + \|\Theta^*\|_\infty \leq 2\alpha^*/\sqrt{d_1 d_2}$ in (2.2).

**Proposition B.3.1** (Spiky Error Matrix). *For the constant $c_0$ in Theorem 4.3.1a, if $\alpha_{sp}(\widehat{\Delta}_{cn}) \notin \mathbb{A}(\beta_{c_0})$, then $\|\widehat{\Delta}_{cn}\|_F^2 \leq \frac{4\alpha^{*2}}{\beta_{c_0}^2} = 4\alpha^{*2}\sqrt{\frac{c_0^2 w_G^2(\mathcal{E}_{\mathcal{R}})\log d}{|\Omega|}}$. An analogous result also holds for $\widehat{\Delta}_{ds}$.* □

**Case 2: Non–Spiky Error Matrix** Let $\widehat{\Delta}_{\text{ds}}, \widehat{\Delta}_{\text{cn}} \in \mathbb{A}(\beta_{c_0})$. Recall from (4.1), that $y - \mathcal{P}_\Omega(\Theta^*) = \xi\eta$, where $\eta \in \mathbb{R}^{|\Omega|}$ consists of independent sub–Gaussian random variables with $\mathbb{E}[\eta_s] = 0$, $\text{Var}(\eta_s) = 1$. Further, as $\eta$ is sub–Gaussian, let $\|\eta_s\|_{\Psi_2} \leq b$ for a constant $b$.

### B.3.1 Restricted Strong Convexity (RSC)

Recall $\mathcal{T}_{\mathcal{R}}$ and $\mathcal{E}_{\mathcal{R}}$ from (4.7). An important step in the proof of Theorem 4.3.1 involves showing that over a subset of $\mathcal{T}_{\mathcal{R}}$, a form of RSC (2.1) is satisfied by a squared loss penalty.

**Theorem B.3.2** (Restricted Strong Convexity). *Let $|\Omega| > c_0^2 w_G^2(\mathcal{E}_{\mathcal{R}})\log d$, for large enough constant $c_0$. There exists a RSC parameter $\kappa_{c_0} > 0$ with $\kappa_{c_0} \approx 1 - o\Big(\frac{1}{\sqrt{\log d}}\Big)$, and a constant $c_1$ such that, the following holds w.p. greater that $1 - \exp(-c_1 w_G^2(\mathcal{E}_{\mathcal{R}}))$,*

$$\forall X \in \mathcal{T}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0}), \quad \frac{d_1 d_2}{|\Omega|}\|\mathcal{P}_\Omega(X)\|_2^2 \geq \kappa_{c_0}\|X\|_F^2.$$

Proof in Appendix B.4 combines tools from empirical process along with Theorem 4.3.2. □

### B.3.2 Constrained Norm Minimizer

**Lemma B.3.3.** *Under the conditions of Theorem 4.3.1, let $b$ be a constant such that $\forall s$, $\|\eta_s\|_{\Psi_2} \leq b$. There exists a universal constant $c_2$ such that, if $\lambda_{cn} \geq 2\xi\sqrt{|\Omega|}$, then w.p. greater than $1 - 2\exp(-c_2|\Omega|)$, (a) $\widehat{\Delta}_{ds} \in \mathcal{T}_{\mathcal{R}}$, and (b) $\|\mathcal{P}_\Omega(\widehat{\Delta}_{cn})\|_2 \leq 2\lambda_{cn}$.* □

Using $\lambda_{\text{cn}} = 2\xi\sqrt{|\Omega|}$ in (4.5), if $\widehat{\Delta}_{\text{cn}} \in \mathbb{A}(\beta_{c_0})$, then using Theorem B.3.2 and Lemma B.3.3, w.h.p.

$$\frac{\|\widehat{\Delta}_{\text{cn}}\|_F^2}{d_1 d_2} \leq \frac{1}{\kappa_{c_0}} \frac{\|\mathcal{P}_\Omega(\widehat{\Delta}_{\text{cn}})\|_2^2}{|\Omega|} \leq \frac{4\xi^2}{\kappa_{c_0}}. \tag{B.6}$$

### B.3.3 Matrix Dantzig Selector

**Proposition B.3.4.** $\lambda_{ds} \geq \xi \frac{\sqrt{d_1 d_2}}{|\Omega|} \mathcal{R}^* \mathcal{P}_\Omega^*(\eta) \Rightarrow$ *(a)* $\widehat{\Delta}_{ds} \in \mathcal{T}_\mathcal{R}$; *(b)* $\frac{\sqrt{d_1 d_2}}{|\Omega|} \mathcal{R}^* \mathcal{P}_\Omega^*(\mathcal{P}_\Omega(\widehat{\Delta}_{ds})) \leq 2\lambda_{ds}$.

Above result follows from optimality of $\widehat{\Theta}_{\text{ds}}$ and triangle inequality. Also,

$$\frac{\sqrt{d_1 d_2}}{|\Omega|} \|\mathcal{P}_\Omega(\widehat{\Delta}_{\text{ds}})\|_2^2 \leq \frac{\sqrt{d_1 d_2}}{|\Omega|} \mathcal{R}^* \mathcal{P}_\Omega^*(\mathcal{P}_\Omega(\widehat{\Delta}_{\text{ds}})) \mathcal{R}(\widehat{\Delta}_{\text{ds}}) \leq 2\lambda_{\text{ds}} \Psi_\mathcal{R}(\mathcal{T}_\mathcal{R}) \|\widehat{\Delta}_{\text{ds}}\|_F,$$

where recall norm compatibility constant $\Psi_\mathcal{R}(\mathcal{T}_\mathcal{R})$ from (4.8). Finally, using Theorem B.3.2, w.h.p.

$$\frac{\|\widehat{\Delta}_{\text{ds}}\|_F^2}{d_1 d_2} \leq \frac{1}{|\Omega|} \frac{\|\mathcal{P}_\Omega(\widehat{\Delta}_{\text{ds}})\|_2^2}{\kappa_{c_0}} \leq \frac{4\lambda_{\text{ds}} \Psi_\mathcal{R}(\mathcal{T}_\mathcal{R})}{\kappa_{c_0}} \frac{\|\widehat{\Delta}_{\text{ds}}\|_F}{\sqrt{d_1 d_2}}. \tag{B.7}$$

Theorem 4.3.1 follows from Proposition B.3.1, (B.6) and (B.7).

## B.4 Proof of Theorem B.3.2

**Statement of Theorem B.3.2**:

Let $|\Omega| > c_0^2 w_G^2(\mathcal{E}_\mathcal{R}) \log d$, for large enough constant $c_0$. There exists a RSC parameter $\kappa_{c_0} > 0$ with $\kappa_{c_0} \approx 1 - o\left(\frac{1}{\sqrt{\log d}}\right)$, and a constant $c_1$ such that, the following holds w.p. greater that $1 - \exp(-c_1 w_G^2(\mathcal{E}_\mathcal{R}))$,

$$\forall X \in \mathcal{T}_\mathcal{R} \cap \mathbb{A}(\beta_{c_0}), \quad \frac{d_1 d_2}{|\Omega|} \|\mathcal{P}_\Omega(X)\|_2^2 \geq \kappa_{c_0} \|X\|_F^2.$$

**Proof:** Recall that $\mathcal{T}_{\mathcal{R}} = \text{cone}\{\Delta : \mathcal{R}(\Theta^* + \Delta) \leq \mathcal{R}(\Theta^*)$ and $\mathcal{E}_{\mathcal{R}} = \mathcal{T}_{\mathcal{R}} \cap \mathbb{S}^{d_1 d_2 - 1}$. Using the properties of norms, it can be easily verified that for the non–trivial case of $\Theta^* \neq 0$, $\mathcal{T}_{\mathcal{R}}$ is a cone with non–empty interior. Theorem 4.3.2 is used as a key result in this proof.

Define $\bar{\mathcal{E}}_{\mathcal{R}} = \mathcal{T}_{\mathcal{R}} \cap \mathbb{B}^{d_1 d_2}$. $\bar{\mathcal{E}}_{\mathcal{R}} \supset \mathcal{E}_{\mathcal{R}}$ is a compact subset of $\mathcal{T}_{\mathcal{R}}$ with non–empty interior, which satisfies the conditions of Theorem 4.3.2. Also, since $\mathcal{T}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})$ is a cone, the following can be easily verified:

$$\begin{aligned} w_{\Omega,g}(\bar{\mathcal{E}}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})) &= w_{\Omega,g}(\mathcal{E}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})) \\ w_G(\bar{\mathcal{E}}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})) &= w_G(\mathcal{E}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})) \leq w_G(\mathcal{E}_{\mathcal{R}}) \end{aligned} \tag{B.8}$$

Define a random variable $V(\Omega) = \sup_{X \in \mathcal{E}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})} \left| \frac{d_1 d_2}{|\Omega|} \|\mathcal{P}_\Omega(X)\|_2^2 - 1 \right|$.
Note that: for $X \in \mathcal{E}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})$, $\mathbb{E}\frac{d_1 d_2}{|\Omega|}\|\mathcal{P}_\Omega(X)\|^2 = 1$; and
for $X \in \bar{\mathcal{E}}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})$, $\|X\|_\infty \leq \frac{\beta_{c_0}}{\sqrt{d_2 d_2}}\|X\|_F^2 \leq \frac{\beta_{c_0}}{\sqrt{d_2 d_2}}$.

### B.4.1 Expectation of $V(\Omega)$

Recall that $\Omega = \{E_s : s = 1, 2, \ldots |\Omega|\}$ are sampled uniformly form standard basis for $\mathbb{R}^{d_1 \times d_2}$, $(\epsilon_s)$ are a sequence of independent Rademacher variables, and $w_G(.)$ denotes the Gaussian width. For constant $k_1, k_2, k_3$ not necessarily same in each occurrence:

$$\begin{aligned} \mathbb{E}V(\Omega) &\overset{(a)}{\leq} \frac{2d_1 d_2}{|\Omega|}\mathbb{E}\sup_{X \in \mathcal{E}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})} \left| \sum_{s=1}^{|\Omega|} \langle X, E_s \rangle^2 \epsilon_s \right| \\ &\overset{(b)}{\leq} k_1 \beta_{c_0} \frac{\sqrt{d_1 d_2}}{|\Omega|}\,\mathbb{E}\sup_{X \in \mathcal{E}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})} \left| \sum_{s=1}^{|\Omega|} \langle X, E_s \rangle \epsilon_s \right| \\ &= k_1 \beta_{c_0} \frac{\sqrt{d_1 d_2}}{|\Omega|} w_{\Omega,\epsilon}(\mathcal{E}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})) = k_1 \beta_{c_0} \frac{\sqrt{d_1 d_2}}{|\Omega|} w_{\Omega,\epsilon}(\bar{\mathcal{E}}_{\mathcal{R}} \cap \mathbb{A}(\beta_{c_0})) \\ &\overset{(c)}{\leq} k_1 \sqrt{\frac{\beta_{c_0}^2 w_G^2(\mathcal{E}_{\mathcal{R}})}{|\Omega|}} + k_2 \frac{\beta_{c_0}^2}{|\Omega|} \overset{(d)}{\leq} \frac{k_3}{c_0 \sqrt{\log d}}, \end{aligned} \tag{B.9}$$

where $(a)$ follows from symmetrization (Lemma 2.3.5), $(b)$ from contraction principle as $\phi_k(\langle X, E_s\rangle) = \frac{\langle X,E_s\rangle^2}{2\sup_{X\in\mathcal{E}_{\mathcal{R}}\cap\mathbb{A}(\beta_{c_0})}\|X\|_\infty}$ is a contraction (Lemma 2.3.6), $(c)$ follows from Theorem 4.3.2, and $(d)$ using $|\Omega| > c_0^2 w_G^2(\mathcal{E}_{\mathcal{R}})\log d$.

### B.4.2 Concentration about $\mathbb{E}V(\Omega)$

Given $\Omega$, let $\Omega' \subset [m]\times[n]$ be another set of indices that differ from $\Omega$ in exactly one element. Then:

$$\begin{aligned} V(\Omega) - V(\Omega') &= \sup_{X\in\mathcal{E}_{\mathcal{R}}\cap\mathbb{A}(\beta_{c_0})} \left|\frac{d_1d_2}{|\Omega|}\sum_{ij\in\Omega} X_{ij}^2 - 1\right| - \sup_{X\in\mathcal{E}_{\mathcal{R}}\cap\mathbb{A}(\beta_{c_0})} \left|\frac{d_1d_2}{|\Omega|}\sum_{kl\in\Omega'} X_{kl}^2 - 1\right| \\ &\le \frac{d_1d_2}{|\Omega|}\sup_{X\in\mathcal{E}_{\mathcal{R}}\cap\mathbb{A}(\beta_{c_0})}\left(\left|\sum_{ij\in\Omega} X_{ij}^2 - \sum_{kl\in\Omega'} X_{kl}^2\right|\right) \\ &\le \frac{2d_1d_2}{|\Omega|}\sup_{X\in\mathcal{E}_{\mathcal{R}}\cap\mathbb{A}(\beta_{c_0})}\|X\|_\infty^2 \le \frac{2\beta_{c_0}^2}{|\Omega|}. \end{aligned} \tag{B.10}$$

By similar arguments on $V(\Omega') - V(\Omega)$, $|V(\Omega) - V(\Omega')| \le \frac{2\beta_{c_0}^2}{|\Omega|}$. Therefore, using Mc Diarmid's inequality (2.6), $\mathbb{P}(V(\Omega) > \mathbb{E}V(\Omega) + \delta) \le \exp\left(-c_1'\frac{\delta^2|\Omega|}{\beta_{c_0}^4}\right)$. Using $\delta = \frac{1}{c_0\sqrt{\log d}}$,

$$\mathbb{P}\left(V(\Omega) > \frac{k_3'}{c_0\sqrt{\log d}}\right) \le \exp\Big(-c_1 w_G^2(\mathcal{E}_{\mathcal{R}})\Big),$$

where $c_0$ is a constant that can be chosen independent of $k_3$. Choosing $c_0$ large enough, set $\kappa_{c_0} := 1 - \delta_{c_0} = 1 - \frac{k_3'}{c_0\sqrt{\log d}}$ close to 1. □

## B.5 Lemmata in Proof of Theorem 4.3.1 and Theorem 4.3.2

### B.5.1 Proof of Lemma B.2.1

Recall definition of $(\mathcal{X}_{\Omega,g}(X))_{X\in S}$ from (B.2): $\mathcal{X}_{\Omega,g}(X) = \sum_s \langle X, E_s\rangle g_s$. By Fubini's theorem $\mathbb{E}_{\Omega,g}\sup_{X\in S}\mathcal{X}_{\Omega,g}(X) = \mathbb{E}_\Omega\mathbb{E}_g\sup_{X\in S}\mathcal{X}_{\Omega,g}(X)$. Further,

- Given random variable $\Omega$, $(\mathcal{X}_{\Omega,g}(X))$ is a Gaussian process with a translation invariant canonical distance given by $d_\Omega(X,Y) = \|\mathcal{P}_\Omega(X-Y)\|_2^2$.
- $d(X,Y) := \sqrt{\mathbb{E}_\Omega d_\Omega^2(X,Y)} = \sqrt{\frac{|\Omega|}{d_1 d_2}}\|X-Y\|_F$

Using Lemma B.1.1, $\mathbb{E}_g \sup_{X\in S} \mathcal{X}_{\Omega,g}(X) \leq K\gamma_2(S, d_\Omega)$, and the following holds:

$$w_{\Omega,g}(S) = \mathbb{E}_\Omega \mathbb{E}_g \sup_{X\in S} \mathcal{X}_{\Omega,g}(X) \leq K\mathbb{E}_\Omega \gamma_2(S,d_\Omega) \overset{(a)}{\leq} \sqrt{\mathbb{E}_\Omega \gamma_2^2(S,d_\Omega)}$$
$$\overset{(b)}{\leq} K\sqrt{\frac{|\Omega|}{d_1 d_2}}\gamma_2(S,\|.\|_F) + K\sqrt{\mathbb{E}\sup_{X,Y\in S}\|\mathcal{P}_\Omega(X-Y)\|_2^2}, \tag{B.11}$$

where $(a)$ follows from Jensen's inequality, $(b)$ from Lemma B.1.3 and noting that by Definition B.1.1 $\forall M > 0$, $\gamma_2(T, M\dot{d}) = M\gamma_2(T,d)$. Lemma B.2.1 now follows from (B.11) and Lemma B.1.1. □

### B.5.2 Proof of Lemma B.2.2

Using triangle inequality, :

$$\mathbb{E}\sup_{X,Y\in S}\|\mathcal{P}_\Omega(X-Y)\|_2^2 \leq \mathbb{E}\sup_{X,Y\in S}|\|\mathcal{P}_\Omega(X-Y)\|_2^2 - \mathbb{E}\|\mathcal{P}_\Omega(X-Y)\|_2^2| + \sup_{X,Y\in S}\mathbb{E}\|\mathcal{P}_\Omega(X-Y)\|_2^2. \tag{B.12}$$

Further,

$$\sup_{X,Y\in S}\mathbb{E}\|\mathcal{P}_\Omega(X-Y)\|_2^2 = \frac{|\Omega|}{d_1 d_2}\sup_{X,Y\in S}\|X-Y\|_F^2 \leq \frac{|\Omega|}{d_1 d_2}\gamma_2^2(S,\|.\|_F), \tag{B.13}$$

where the last inequality follows from the definition of $\gamma_\alpha$. Finally, the following set of equations hold:

$$
\begin{aligned}
&\mathbb{E} \sup_{X,Y\in S} \left| \|\mathcal{P}_\Omega(X-Y)\|_2^2 - \mathbb{E}\|\mathcal{P}_\Omega(X-Y)\|_2^2 \right| \\
&= \mathbb{E} \sup_{X,Y\in S} \Big| \sum_{s=1}^{|\Omega|} \langle X-Y, E_s\rangle^2 - \mathbb{E}\langle X-Y, E_s\rangle^2 \Big| \\
&\overset{(a)}{\leq} 2\mathbb{E}_{\Omega,(\epsilon_s)} \sup_{X,Y\in S} |\sum_{s=1}^{|\Omega|} \langle X-Y, E_s\rangle^2 \epsilon_s| \\
&\overset{(b)}{\leq} k_4' \sup_{X,Y\in S} \|X-Y\|_\infty \mathbb{E}_{\Omega,g} \sup_{X,Y\in S} |\sum_{s=1}^{|\Omega|} \langle X-Y, E_s\rangle g_s| \\
&\overset{(c)}{\leq} 2k_4' \sup_{X,Y\in S} \|X-Y\|_\infty \mathbb{E}_{\Omega,g} \sup_{X\in S} |\sum_{s=1}^{|\Omega|} \langle X, E_s\rangle g_s| \\
&\overset{(d)}{\leq} 4k_4' \sup_{X,Y\in S} \|X-Y\|_\infty w_{\Omega,g}(S),
\end{aligned} \tag{B.14}
$$

where $(\epsilon_s)$ are standard Rademacher variables, i.e. $\epsilon_s \in \{-1, 1\}$ with equal probability, $(a)$ follows from symmetrization argument (Lemma 2.3.5), $(b)$ follows from contraction principles (Lemma 2.3.6) and using $\phi(\langle X, E_s\rangle) = \frac{\langle X,E_s\rangle^2}{2\sup_{X\in S}\|X\|_\infty}$ as a contraction, $(c)$ follows from triangle inequality, and $(d)$ follows from $g_s$ being symmetric (Lemma 2.2.1 in [145]). The lemma follows by combining (B.12), (B.13), and (B.14), along with Lemma B.1.1. □

### B.5.3 Proof of Lemma B.3.3

Recall that $\eta \in \mathbb{R}^{|\Omega|}$ is a vector of centered, unit variance sub-Gaussian random variables. Further, let $\|\eta_s\|_{\Psi_2} \leq b$, for some constant $b$ (Definition 2.3.2). Combining Lemma 2.3.9 and Lemma 2.3.10: $\eta_s^2$ and $\eta_s^2 - 1$ are sub–exponential with $\|\eta_s^2 - 1\|_{\Psi_1} \leq 2\|\eta_s^2\|_{\Psi_1} \leq 4\|\eta_s\|_{\Psi_2} \leq 4b^2$. Thus, using Lemma 2.3.8, for a

constant $c_2'$,

$$\mathbb{P}\Big(\Big|\frac{1}{|\Omega|}\sum_{s=1}^{|\Omega|}\eta_s^2 - 1\Big| > \tau\Big) \leq 2\exp\Big(-c_2'|\Omega|\min\big\{\frac{\tau^2}{16b^4}, \frac{\tau}{4b^2}\big\}\Big). \tag{B.15}$$

Choosing $\tau$ to be an appropriate constant, $\|\mathcal{P}_\Omega(\Theta^*) - y)\|_2 \leq 2\xi\sqrt{|\Omega|} \leq \lambda_{\text{cn}}$ w.p. greater than $1 - \exp(-c_2\tau|\Omega|)$, and the lemma follows from the optimality of $\widehat{\Theta}_{\text{cn}}$ and triangle inequality.

## B.6 Spectral k–Support Norm

Recall the following definition of spectral $k$–support norm $\|\Theta\|_{\text{k–sp}}$ from (4.4):

$$\|\Theta\|_{\text{k–sp}} = \inf_{v\in\mathcal{V}(\mathcal{G}_k)}\Big\{\sum_{g\in\mathcal{G}_k}\|v_g\|_2 : \sum_{g\in\mathcal{G}_k} v_g = \sigma(\Theta)\Big\}, \tag{B.16}$$

where $\mathcal{G}_k = \{g \subseteq [\bar{d}] : |g| \leq k\}$ is the set of all subsets $[\bar{d}]$ of cardinality at most $k$, and $\mathcal{V}(\mathcal{G}_k) = \{(v_g)_{g\in\mathcal{G}_k} : v_g \in \mathbb{R}^{d_1}, \text{supp}(v_g) \subseteq g\}$.

**Proposition B.6.1** (Proposition 2.1 in [11]). *For $\Theta \in \mathbb{R}^{\bar{d}\times\bar{d}}$ with singular values $\sigma(\Theta) = \{\sigma_1, \sigma_2, \ldots, \sigma_{\bar{d}}\}$, such that $\sigma_1 \geq \sigma_2 \geq \ldots, \geq \sigma_{\bar{d}}$. Then,*

$$\|\Theta\|_{k\text{–}sp} = \left(\sum_{i=1}^{k-r-1}\sigma_i^2 + \frac{1}{r+1}\left(\sum_{i=k-r}^{\bar{d}}\sigma_i\right)^2\right)^{\frac{1}{2}}, \tag{B.17}$$

*where $r \in \{0, 1, 2, \ldots, k-1\}$ is the unique integer satisfying $\sigma_{k-r-1} > \frac{1}{r+1}\sum_{i=k-r}^{d_1}\sigma_i \geq \sigma_{k-r}$.* □

### B.6.1 Proof of Lemma 4.3.3

#### Statement of Lemma 4.3.3

If rank of $\Theta^*$ is $s$ and $\mathcal{E}_\mathcal{R}$ is the error set from $\mathcal{R}(\Theta) = \|\Theta\|_{\text{k–sp}}$, then

$$w_G^2(\mathcal{E}_{\mathcal{R}}) \leq s(2\bar{d}-s) + \Big(\frac{(r+1)^2\|\sigma^*_{I_2}\|_2^2}{\|\sigma^*_{I_1}\|_1^2} + |I_1|\Big)(2\bar{d}-s).$$

□

**Proof** The following lemmas are stated from existing work.

**Lemma B.6.2** (Equation 60 in [129]). *Let $z$ be an $s \geq k$ sparse vector in $\mathbb{R}^p$, and let $\tilde{z}$ is the vector $z$ sorted in non increasing order of $|z_i|$. Denote $r \in \{0, 1, 2, \ldots, k-1\}$ to be the unique integer satisfying*

$$|\tilde{z}_{k-r-1}| > \frac{1}{r+1}\sum_{i=k-r}^{p}|\tilde{z}_i| \geq |\tilde{z}_{k-r}|.$$

*Define $I_2 = \{1, 2, \ldots, k-r-1\}$, $I_1 = \{k-r, k-r+1, \ldots, s\}$, and $I_0 = \{s+1, s+2, \ldots, p\}$; and let $\tilde{z}_I$ denote the vector $\tilde{z}$ restricted to indices in $I$. Then the sub–differential of the vector $k$–support norm denoted by $\|.\|_{vk\text{-}sp}$ at $w$ is given by:*

$$\partial\|z\|_{vk\text{-}sp} = \frac{1}{\|z\|_{vk\text{-}sp}}\Big\{\tilde{z}_{I_2} + \frac{1}{r+1}\|\tilde{z}_{I_1}\|_1(sign(\tilde{z}_{I_1}) + h_{I_0}) : \|h\|_\infty \leq 1\Big\},$$

**Lemma B.6.3** (Theorem 2 in [156]). *Let $\mathcal{R} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}_+$ be an orthogonally invariant norm; i.e. $\mathcal{R}(X) = \phi(\sigma(X))$ such that $\phi : \mathbb{R}^{d_1} \to \mathbb{R}_+$ is a symmetric gauge function satisfying: (a) $\phi(x) > 0\ \forall x \neq 0$, (b) $\phi(\alpha x) = |\alpha|\phi(x)$, (c) $\phi(x+y) \leq \phi(x) + \phi(y)$, and (d) $\phi(x) = \phi(|x|)$.*

*Further let $\partial\phi(x)$ denote the sub–differential of $\phi$ at $x$. Then for $X \in \mathbb{R}^{\bar{d}\times\bar{d}}$ with singular value decomposition (SVD) $X = U_X\Sigma_X V_X^\top$ and $\sigma_X = diag(\Sigma_X)$, the sub–differential of $\mathcal{R}(X)$ is given by:*

$$\partial\mathcal{R}(X) = \{U_X D V_X^\top : D = diag(d),\ and\ d \in \partial\Phi(\sigma_X)\}.$$

Since spectral $k$–support norm of a matrix $X = U_X \Sigma_X V_X^\top$ is the vector $k$–support norm applied to the singular values $\sigma_X = \text{diag}(\Sigma_X)$, Lemma B.6.2 and B.6.3 can be used to infer the following:

$$\partial \|X\|_{\text{k–sp}} = \Big\{ U_X D V_X^\top : \text{diag}(D) \in \frac{1}{\|\sigma_X\|_{\text{vk-sp}}} \Big\{ \sigma_{X_{I_2}} + \frac{\|\sigma_{X_{I_1}}\|_1}{r+1} (\mathbf{1}_{I_1} + h_{I_0}) : \|h\|_\infty \leq 1 \Big\} \Big\}. \tag{B.18}$$

where $\mathbf{1} \in \mathbb{R}^{\bar{d}}$ denotes a vector of all ones.

The error cone for $\mathcal{R}(.) = \|.\|_{\text{k–sp}}$ is given by the tangent cone:

$$\mathcal{T}_{\mathcal{R}} = \text{cone}\{\Delta : \|\Theta^* + \Delta\|_{\text{k–sp}} \leq \|\Theta^*\|_{\text{k–sp}}\},$$

and the polar of the tangent cone – the *normal cone* is given by

$$\mathcal{T}_{\mathcal{R}}^* = \mathcal{N}_{\mathcal{R}}(\Theta^*) = \{Y : \langle Y, X \rangle \leq 0 \; \forall X \in \mathcal{T}_{\mathcal{R}}\} = \text{cone}(\partial \mathcal{R}(\Theta^*))$$

Let $\Theta^* = U^* \Sigma^* V^{*\top}$ be the full SVD of $\Theta^*$, such that $\sigma^* = \text{diag}(\Sigma^*) \in \mathbb{R}^{\bar{d}}$ and $\sigma_1^* \geq \sigma_2^* \ldots \geq \sigma_{\bar{d}}^*$. Let $u_i^*$ and $v_i^*$ for $i \in [\bar{d}]$ denote the $i^{\text{th}}$ column of $U^*$ and $V^*$, respectively. Further, let the rank of $\Theta^*$ be $\text{rk}(\Theta^*) = \|\sigma^*\|_0 = s$.

Like for the vector case, denote $r \in \{0, 1, 2, \ldots, k-1\}$ to be the unique integer satisfying $\sigma_{k-r-1}^* > \frac{1}{r+1} \sum_{i=k-r}^{p} \sigma_i^* \geq \sigma_{k-r}^*$. Define $I_2 = \{1, 2, \ldots, k-r-1\}$, $I_1 = \{k-r, k-r+1, \ldots, s\}$, and $I_0 = \{s+1, s+2, \ldots, p\}$; Also define the subspace:

$$T = \text{span}\{u_i^* x^\top : i \in I_2 \cup I_1, x \in \mathbb{R}^{\bar{d}}\} \cup \text{span}\{y v_i^{*\top} : i \in I_2 \cup I_1, y \in \mathbb{R}^{\bar{d}}\}$$

Let $T^\perp$ be the subspace orthogonal to $T$ and let $P_T$ and $P_{T^\perp}$ be the projection operators onto $T$ and $T^\top$ respectively. From (B.18),

$$\mathcal{N}_{\mathcal{R}}(\Theta^*) = \left\{ Y = U^* D V^{*\top} : D = \text{diag}\Big( t \frac{r+1}{\|\sigma_{I_1}^*\|_1} \sigma_{I_2}^* + t \mathbf{1}_{I_1} + t h_{I_0} \Big) : t \geq 0, \|h\|_\infty \leq 1 \right\},$$

Finally, from Lemma 2.3.12,

$$w_G^2(\mathcal{T}_{\mathcal{R}} \cap \mathbb{S}^{\bar{d}\bar{d}-1}) \leq \mathbb{E}_G \inf_{X \in \mathcal{N}_{\mathcal{R}}(\Theta^*)} \|G - X\|_F^2$$
$$\leq \mathbb{E}_G \inf_{\substack{t>0 \\ \|h\|_\infty \leq 1}} \left\| P_T(G) - t\frac{r+1}{\|\sigma^*_{I_1}\|_1} \sum_{i \in I_2} \sigma_i^* u_i^* v_i^{*\top} + t \sum_{i \in I_1} u_i^* v_i^{*\top} + P_{T^\perp}(G) - t \sum_{i \in I_0} h_i u_i^* v_i^{*\top} \right\|_F^2$$

Let $P_{T^\perp}(G) = \sum_{i \in I_0} \sigma_i(P_{T^\perp} G) u_i^* v_i^{*\top}$ be the decomposition of $P_T^\perp(G)$ in the basis of of $\{u_i^* v_i^{*\top}\}_{i \in I_0}$. Taking $t = \|P_{T^\perp}(G)\|_{\text{op}} = \max_{i \in I_0} \sigma_i(P_{T^\perp}(G))$, and $h_i = \sigma_i(P_{T^\perp}(G))/\|P_{T^\perp}(G)\|_{\text{op}} \leq 1$,

$$w_G^2(\mathcal{T}_{\mathcal{R}} \cap \mathbb{S}^{\bar{d}\bar{d}-1}) \leq \mathbb{E}_G \|P_T(G)\|_F^2 + \left( \frac{(r+1)^2 \|\sigma^*_{I_2}\|_2^2}{\|\sigma^*_{I_1}\|_1^2} + |I_1| \right) \mathbb{E}_G \|P_T(G)\|_2^2. \quad \text{(B.19)}$$

Lemma 4.3.3 follows by using $\mathbb{E}_G \|P_T(G)\|_F^2 = s(2\bar{d} - s)$ and $\mathbb{E}_G \|P_T(G)\|_{\text{op}}^2 \leq 2(2\bar{d} - s)$ from [33].

## B.7 Extension to GLMs

This section provides directions for extending the work to matrix completion under generalized linear models. This section has not been rigorously formalized. An accurate version will be included in a longer version of the paper.

Consider an observation model wherein the observation matrix $Y$ is drawn from a member of *natural exponential family* parametrized by a structured ground truth matrix $\Theta^*$, such that:

$$\mathbb{P}(Y|\Theta^*) = \prod_{ij} p(Y_{ij})\, e^{Y_{ij}\Theta^*_{ij} - A(\Theta^*_{ij})}, \quad \text{(B.20)}$$

where $A : \text{dom}(\Theta_{ij}) \to \mathbb{R}$ is called the *log–partition function* and is strictly convex and analytic, and $p(.)$ is called the *base measure*. This family of distributions encompass a wide range of common distributions including Gaussian, Bernoulli,

binomial, Poisson, and exponential among others. In a generalized linear matrix completion setting [62], the task is to estimate $\Theta^*$ from a subset of entries $\Omega$ of $Y$, i.e. $(\Omega, \mathcal{P}_\Omega(Y))$.

A useful consequence of exponential family distribution assumption for observation matrix is that the negative log–likelihood loss over the observed entries is convex with respect to the natural parameter $\Theta^*$, and have a one-to-one correspondence with a rich class of divergence functions called the Bregman Divergence [54, 14]. The negative log likelihood is proportional to:

$$\mathcal{L}_\Omega(\Theta) = \sum_{(i,j)\in\Omega} A(\Theta_{ij}) - Y_{ij}\Theta_{ij}$$

The following *regularized matrix estimator* is proposed for generalized matrix completion:

$$\widehat{\Theta}_{\text{re}} = \underset{\|\Theta\|_\infty \le \frac{\alpha^*}{\sqrt{d_1 d_2}}}{\operatorname{argmin}} \frac{d_1 d_2}{|\Omega|}\mathcal{L}_\Omega(\Theta) + \lambda_{\text{re}}\mathcal{R}(\Theta). \tag{B.21}$$

**Hypothesis 1.** *Let $\widehat{\Theta}_{re} = \Theta^* + \widehat{\Delta}_{re}$. In addition to the assumptions in Section 4.2, assume that for some $\eta \ge 0$, $\nabla^2 A(u) \ge e^{-\eta|u|} \forall\ u \in \mathbb{R}$. The following result holds for any fixed $\gamma > 1$. Define:*

$$\widetilde{\mathcal{T}}_{\mathcal{R},\gamma} = \text{cone}\{\Delta : \mathcal{R}(\Theta^*+\Delta) \le \mathcal{R}(\Theta^*)+\frac{1}{\gamma}\mathcal{R}(\Theta^*)\}, \quad \text{and} \quad \widetilde{\mathcal{E}}_{\mathcal{R},\gamma} = \widetilde{\mathcal{T}}_{\mathcal{R},\gamma}\cap\mathbb{S}^{d_1 d_2 - 1}. \tag{B.22}$$

*Let $\lambda_{re} \ge \gamma\frac{d_1 d_2}{|\Omega|}\mathcal{R}^*(\nabla\mathcal{L}_\Omega(\Theta^*))$, and for some $c_0$, $|\Omega| > \left(\frac{\gamma+1}{\gamma-1}\right)^2 c_0^2 w_G^2(\widetilde{\mathcal{E}}_{\mathcal{R},\gamma})\log d$. There exists a constant $k_1$ such that for large enough $c_0$, there exists $\kappa_{c_0} > 0$, such that with high probability,*

$$\|\widehat{\Delta}_{re}\|_F^2 \le 4\alpha^{*2}\left(\frac{\gamma+1}{\gamma-1}\right)^2 \max\left\{\frac{\lambda_{re}^2\Psi_{\mathcal{R}}^2(\widetilde{\mathcal{T}}_{\mathcal{R},\gamma})}{\zeta(\eta,\alpha^*)\kappa_{c_0}^2}, \frac{c_0^2 w_G^2(\widetilde{\mathcal{E}}_{\mathcal{R},\gamma})\log d}{|\Omega|}\right\},$$

*where $\zeta(\eta,\alpha^*) = e^{\frac{-4\eta\alpha^*}{\sqrt{d_1 d_2}}}$, and $\alpha^*$, $w_G(.)$, and $\Psi_{\mathcal{R}}(.)$ are notations from Section 5.4.*

The conjectures follows by combining the results in this paper along with the results from [13], and [62]. This result is beyond the scope of this paper and will be dealt with more rigorously in a longer version of the paper.

# Appendix C

# Proof of Results in Chapter 5

## C.1 Proof of Lemma 5.3.1

Recall that:

$$\begin{aligned} T = T(\mathcal{M}) =& \text{aff}\{\mathcal{Y} \in \bar{\mathfrak{X}} : \forall\, v, \text{rowSpan}(\mathbb{Y}_{r_v}) \subseteq \text{rowSpan}(\mathbb{M}_{r_v}) \\ & \textit{or } \text{rowSpan}(\mathbb{Y}_{c_v}) \subseteq \text{rowSpan}(\mathbb{M}_{c_v})\} \\ T^{\perp} = T^{\perp}(\mathcal{M}) =& \{\mathcal{Y} \in \bar{\mathfrak{X}} : \forall\, v, \text{rowSpan}(Y_v) \perp \text{rowSpan}(_v) \\ & \textit{and } \text{colSpan}(Y_v) \perp \text{colSpan}(M_v)\} \end{aligned}$$

Need to show that $\forall \mathcal{X} \in \bar{\mathfrak{X}}$, $\mathcal{X} \in T^{\perp}$ iff $\langle \mathcal{X}, \mathcal{Y} \rangle = 0, \forall \mathcal{Y} \in T$.

$\Longrightarrow$ Let $\mathcal{X} \in \{\mathcal{X} \in \bar{\mathfrak{X}} : \langle \mathcal{X}, \mathcal{Y} \rangle = 0, \forall \mathcal{Y} \in T\}$, if $\mathcal{X} \notin T^{\perp}$, then $\exists v$ such that atleast one of the statements below hold true:

(a) $\text{rowSpan}(X_v) \not\perp \text{rowSpan}(M_v)$, or

(b) $\text{colSpan}(X_v) \not\perp \text{colSpan}(M_v)$

WLOG let us assume that $(a)$ is true, the proof for the other case is analogous. Consider the decompositon $X_v = X_v^{(1)} + X_v^{(2)}$ such that $\text{rowSpan}(X_v^{(1)}) \perp \text{rowSpan}(M_v)$ and $\text{rowSpan}(X_v^{(2)}) \subseteq \text{rowSpan}(M_v)$. Consider the collective matrix $\mathcal{Y}$ such that $Y_{v'} = X_v^{(2)}$ if $v' = v$, and $Y_{v'} = 0$ otherwise. Clearly, $\mathcal{Y} \in T$ and $\langle \mathcal{X}, \mathcal{Y} \rangle \neq 0$, a contradiction.

$\Longleftarrow$ If $\mathcal{X} \in T^{\perp}$, then by the definitions, $\forall \mathcal{Y} \in T$, $\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_v \langle X_v, Y_v \rangle = 0$.

## C.2 Proof of Theorem 5.4.1

The proof uses ideas of *dual certificate* from existing matrix completion literature [26, 30, 127], and further adapts the *golfing scheme* introduced by Gross et al. [58], for constructing the dual certificate.

Let $\widehat{\mathcal{M}} = \mathcal{M} + \Delta$ be the output of the convex program in (5.15). Key steps in the proof are:

1. Show that under the sample complexity requirements of Theorem 5.4.1, $\|\mathcal{P}_T(\Delta)\|_F$ can be upper bounded by a finite multiple of $\|\mathcal{P}_{T^\perp}(\Delta)\|_F$, where $T$ and $T^\perp$ are defined in 5.10 and 5.11, respectively.
2. Under the above condition, show optimality of $\mathcal{M}$ for (5.15) if a *dual certificate* $\mathcal{Y}$ satisfying certain conditions exists.
3. Adapt the *golfing scheme* to construct $\mathcal{Y}$.

### C.2.1 Bound on $\|\mathcal{P}_T(\Delta)\|_F$

Let $p(v,i,j) = \frac{|\Omega_{r_v}|}{2n_{r_v}m_{r_v}} + \frac{|\Omega_{c_v}|}{2n_{c_v}m_{c_v}} = |\Omega|\mathbb{P}\big((v,i,j) = \Omega_s\big)$. Define the following operators for $s = 1, 2, \ldots, |\Omega|$:

$$\mathcal{R}_s : \mathcal{X} \to \tfrac{1}{p(v_s,i_s,j_s)}\langle \mathcal{X}, \mathcal{E}^{(s)}\rangle\ \mathcal{E}^{(s)}, \text{ and} \tag{C.1}$$

$$\mathcal{R}_\Omega : \mathcal{X} \to \textstyle\sum_{s=1}^{|\Omega|} \mathcal{R}_s(\mathcal{X}) \text{ with } \mathbb{E}[\mathcal{R}_\Omega] = \mathcal{I}, \tag{C.2}$$

where $\mathcal{I}$ is the identity operator, and recall that $\mathcal{E}^{(s)} = \mathcal{E}^{(v_s,i_s,j_s)}$.

**Lemma C.2.1.** *Let $\forall\ k$, $|\Omega_k| \geq c_0\mu_0 n_k R\beta \log N$ for a sufficiently large constant $c_0$. Then, under the assumptions in Section 5.3.1, the following holds w.p. greater than $1 - N^{-\beta}$,*

$$\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \mathcal{P}_T\|_{op} \leq \tfrac{1}{2}.$$

Proof in Appendix C.3. □

Let $M_\Omega(v,i,j)$ denote the multiplicity of $(v,i,j)$ in $\Omega$, i.e. $M_\Omega(v,i,j) = \sum_s \mathbb{1}_{(v,i,j)=(v_s,i_s,j_s)}$. Note that $M_\Omega(v,i,j) \le |\Omega|$, and $\min_k \frac{|\Omega_k|}{n_k m_k} \le p(v,i,j) \le \max_k \frac{|\Omega_k|}{n_k m_k}$. Further, for all $\mathcal{X}$,

$$\|\mathcal{R}_\Omega(\mathcal{X})\|_F = \Big\| \sum_{v=1}^{V} \sum_{(i,j)\in\mathcal{J}(v)} \frac{M_\Omega(v,i,j)}{p(v,i,j)} \langle \mathcal{X}, \mathcal{E}^{(v,i,j)} \rangle \mathcal{E}^{(v,i,j)} \Big\|_F \le \frac{|\Omega|}{\min_k \frac{|\Omega_k|}{n_k m_k}} \|\mathcal{X}\|_F, \tag{C.3}$$

and w.h.p,

$$\begin{aligned} \|\mathcal{R}_\Omega \mathcal{P}_T(\Delta)\|_F^2 &\ge \frac{1}{\max_k \frac{|\Omega_k|}{n_k m_k}} \langle \mathcal{R}_\Omega \mathcal{P}_T(\Delta), \mathcal{P}_T(\Delta) \rangle \\ &= \frac{1}{\max_k \frac{|\Omega_k|}{n_k m_k}} \langle \mathcal{P}_T \mathcal{R}_\Omega \mathcal{P}_T(\Delta), \mathcal{P}_T(\Delta) \rangle \ge \frac{1}{2 \max_k \frac{|\Omega_k|}{n_k m_k}} \|\mathcal{P}_T(\Delta)\|_F^2, \end{aligned} \tag{C.4}$$

where the last inequality follows from Lemma C.2.1.

Combining (C.3) and (C.4), along with $0 = \|\mathcal{R}_\Omega(\Delta)\|_F \ge \|\mathcal{R}_\Omega \mathcal{P}_T(\Delta)\|_F - \|\mathcal{R}_\Omega \mathcal{P}_{T^\perp}(\Delta)\|_F$,

$$\|\mathcal{P}_T(\Delta)\|_F \le \frac{1}{2} \kappa_\Omega(N) \|\mathcal{P}_{T^\perp}(\Delta)\|_F, \tag{C.5}$$

where $\kappa_\Omega(N) = \frac{3|\Omega| \sqrt{\max_k |\Omega_k|/n_k m_k}}{\min_k |\Omega_k|/n_k m_k}$.

### C.2.2 Optimality of $\mathcal{M}$

**Lemma C.2.2.** *Under assumptions in Section 5.3.1, for a sufficiently large constant $c_0$, let $|\Omega_k| \ge c_0 \mu_0 n_k R \beta \log N \ \forall k$. If there exists a dual certificate $\mathcal{Y} = \mathcal{P}_\Omega(\mathcal{Y})$ satisfying the following conditions, then $\mathcal{M}$ is the unique minimizer to (5.15) w.p. greater than $1 - N^{-\beta}$:*

1. $\|\mathcal{P}_T(\mathcal{Y}) - \mathcal{E}_\mathcal{M}\|_F \le \frac{1}{\kappa_\Omega(N)}$, *and*
2. $\|\mathcal{P}_{T^\perp}(\mathcal{Y})\|_\mathscr{A}^* \le 1/2$

*where recall $\mathcal{E}_{\mathcal{M}}$ from Assumption 5.3.2.*

Proof is provided in the Appendix C.3.

### C.2.3 Constructing Dual Certificate

The proof is completed by constructing a dual certificate $\mathcal{Y}$ satisfying the conditions in Lemma C.2.2. Partition $\Omega$ into $p \geq c_1 \log(N\kappa_\Omega(N))$ partitions denoted by $\Omega^{(j)}$, for $j = 1, 2, \ldots, p$. The partitioning is done such that for all $j$: (a) $|\Omega_k^{(j)}| > c_0\mu_0\beta R n_k \log N$ and $\frac{|\Omega_k^{(j)}|}{n_k m_k} \leq c\frac{|\Omega^{(j)}|}{N^2}$ for all $k$, and (b) $|\Omega^{(j)}| > c_2 \max\{\mu_0, \mu_1\}\beta RN \log N$ , where $\Omega_k^{(j)} = \{(v, i, j) \in \Omega^{(j)} : r_v = k \text{ or } c_v = k\}$.

Define $\mathcal{W}_0 = \mathcal{E}_{\mathcal{M}}$ where $\mathcal{E}_{\mathcal{M}}$ is the sign matrix from Assumption 5.3.2. Define the following processes for $j = 1, 2, \ldots$ s.t. :

$$\mathcal{Y}_j = \textstyle\sum_{j'=1}^{j} \mathcal{R}_{\Omega^{(j')}} \mathcal{W}_{j'-1} = \mathcal{R}_{\Omega^{(j)}} \mathcal{W}_{j-1} + \mathcal{Y}_{j-1}, \text{ and } \mathcal{W}_j = \mathcal{E}_{\mathcal{M}} - \mathcal{P}_T(\mathcal{Y}_j). \quad \text{(C.6)}$$

Note that $\forall\ j$, $\mathcal{P}_\Omega(\mathcal{Y}_j) = \mathcal{Y}_j$, and $\mathcal{P}_T(\mathcal{W}_j) = \mathcal{W}_j$.

1. *Claim:* $\mathcal{Y}_p$ for $p \geq c_1 \log(N\kappa_\Omega(N))$ satisfies the first condition in Lemma C.2.2:

   *Proof:* It is easy to verify that $\frac{1}{2}\mathcal{E}^{(v,i,j)} \in \mathscr{A}$ for all basis vectors $\mathcal{E}^{(v,i,j)}$; and from Assumption 5.3.3, $-\frac{1}{2}\mathcal{E}^{(v,i,j)} \in \mathscr{A}$. Thus,

   $$\forall \mathcal{X} \in \bar{\mathfrak{X}}, \|\mathcal{X}\|_{\mathscr{A}}^* = \sum_{\mathcal{A} \in \mathscr{A}} \langle \mathcal{X}, \mathcal{A} \rangle \geq \frac{1}{2} \max_{v \in [V], (i,j) \in \mathcal{J}(v)} |\langle \mathcal{X}, \mathcal{E}^{(v,i,j)} \rangle| \geq \frac{1}{2N} \|\mathcal{X}\|_F.$$

   Also, $1 = \|\mathcal{E}_{\mathcal{M}}\|_{\mathscr{A}}^* \geq \frac{1}{2N}\|\mathcal{E}_{\mathcal{M}}\|_F$, and $\mathcal{P}_T(\mathcal{Y}_p) - \mathcal{E}_{\mathcal{M}} = \mathcal{W}_p$. Using the above inequalities,

   $$\begin{aligned} \|\mathcal{P}_T(\mathcal{Y}_p) - \mathcal{E}_{\mathcal{M}}\|_F = \|\mathcal{W}_{p-1} - \mathcal{P}_T \mathcal{R}_{\Omega^{(p)}} \mathcal{W}_{p-1}\|_F &\overset{(a)}{\leq} \frac{1}{2}\|\mathcal{W}_{p-1}\|_F \\ \leq \frac{1}{2^p}\|\mathcal{E}_{\mathcal{M}}\|_F &\overset{(b)}{<} \frac{1}{\kappa_\Omega(N)} \end{aligned} \quad \text{(C.7)}$$

where $(a)$ follows from Lemma C.2.1, and $(b)$ follows for large enough $c_1$ s.t. $p > c_1 \log(N\kappa_\Omega(N))$.

2. The proof for second condition follows directly from the analogous proof for standard matrix completion by Recht [127]. It is derived for completeness in Appendix C.3.3.

## C.3 Proof of Lemmata in Appendix C.2

### C.3.1 Proof of Lemma C.2.1

Recall $\mathcal{R}_s$ and $\mathcal{R}_\Omega$ from (C.1) and (C.2), and $\mathcal{X} = \sum_{v=1}^{V} \sum_{(i,j)\in\mathcal{I}(v)} \langle \mathcal{X}, \mathcal{E}^{(v,i,j)} \rangle \mathcal{E}^{(v,i,j)}$.

Hence,

$$\mathcal{P}_T(\mathcal{X}) = \sum_{v=1}^{V} \sum_{(i,j)\in\mathcal{I}(v)} \langle \mathcal{P}_T(\mathcal{X}), \mathcal{E}^{(v,i,j)} \rangle \mathcal{E}^{(v,i,j)} = \sum_{v=1}^{V} \sum_{(i,j)\in\mathcal{I}(v)} \langle \mathcal{X}, \mathcal{P}_T(\mathcal{E}^{(v,i,j)}) \rangle \mathcal{E}^{(v,i,j)}$$

Define $\mathcal{V}_s := \mathcal{P}_T \mathcal{R}_s \mathcal{P}_T : \mathcal{X} \to \frac{1}{p(v_s,i_s,j_s)} \langle \mathcal{X}, \mathcal{P}_T(\mathcal{E}^{(s)}) \rangle \mathcal{P}_T(\mathcal{E}^{(s)})$, where $p(v,i,j) = \frac{|\Omega_{r_v}|}{2n_{r_v}m_{r_v}} + \frac{|\Omega_{c_v}|}{2n_{c_v}m_{c_v}}$.

Thus, $\mathbb{E}[\mathcal{V}_s] = \frac{1}{|\Omega|}\mathcal{P}_T$, and

$$\begin{aligned} \|\mathcal{V}_s\|_{\text{op}} &= \sup_{\|\mathcal{X}\|_F=1} \frac{1}{p(v_s,i_s,j_s)} \langle \mathcal{X}, \mathcal{P}_T(\mathcal{E}^{(s)}) \rangle \|\mathcal{P}_T(\mathcal{E}^{(s)})\|_F = \frac{1}{p(v_s,i_s,j_s)} \|\mathcal{P}_T(\mathcal{E}^{(s)})\|_F^2 \\ &\overset{(a)}{\leq} \frac{1}{p(v_s,i_s,j_s)} \left( \frac{\mu_0 R}{m_{r_{v_s}}} + \frac{\mu_0 R}{m_{c_{v_s}}} \right) \overset{(b)}{\leq} \frac{1}{c_0 \beta \log N}, \end{aligned} \tag{C.8}$$

where $(a)$ follows from Assumption 5.3.2, and $(b)$ follows as $\forall k$, $|\Omega_k| > c_0 \mu_0 n_k R \beta \log N$.

(i) Bound on $\|\mathcal{V}_s - \mathbb{E}[\mathcal{V}_s]\|_{\text{op}}$

$$\begin{aligned} \|\mathcal{V}_s - \mathbb{E}[\mathcal{V}_s]\|_{\text{op}} &\overset{(a)}{\leq} \max\left( \|\mathcal{V}_s\|_{\text{op}}, \|\mathbb{E}[\mathcal{V}_s]\|_{\text{op}} \right) \\ &\leq \max\left( \frac{1}{c_0 \beta \log N}, \frac{1}{\Omega} \right) = \frac{1}{c_0 \beta \log N} \end{aligned} \tag{C.9}$$

where $(a)$ follows as both $\mathcal{V}_s$ and $\mathbb{E}[\mathcal{V}_s]$ are positive semidefinite.

(ii) Bound on $\sum_{s=1}^{|\Omega|} \|\mathbb{E}[(\mathcal{V}_s - \mathbb{E}[\mathcal{V}_s])^2]\|_{\text{op}}$.

$$\begin{aligned}\mathbb{E}[(\mathcal{V}_s)^2(X)] &= \mathbb{E}\left[\frac{1}{p(v_s,i_s,j_s)^2}\langle \mathcal{X}, \mathcal{P}_T(\mathcal{E}^{(s)})\rangle \|\mathcal{P}_T(\mathcal{E}^{(s)})\|_F^2 \mathcal{P}_T(\mathcal{E}^{(s)})\right] \\ &\preccurlyeq \frac{1}{c_0\beta\log N}\mathbb{E}\left[\frac{1}{p(v_s,i_s,j_s)}\langle \mathcal{X}, \mathcal{P}_T(\mathcal{E}^{(s)})\rangle \mathcal{P}_T(\mathcal{E}^{(s)})\right] \\ &= \frac{1}{|\Omega| c_0\beta\log N}\mathcal{P}_T(\mathcal{X}). \end{aligned} \tag{C.10}$$

$$\begin{aligned}\|\mathbb{E}[(\mathcal{V}_s - \mathbb{E}[\mathcal{V}_s])^2]\|_{\text{op}} &= \|\mathbb{E}[\mathcal{V}_s^2] - (\mathbb{E}[\mathcal{V}_s])^2]\|_{\text{op}} \\ &\leq \max\left(\|\mathbb{E}[\mathcal{V}_s^2]\|_{\text{op}}, \|(\mathbb{E}[\mathcal{V}_s])^2\|_{\text{op}}\right) \overset{(a)}{\leq} \frac{1}{|\Omega| c_0\beta\log N}, \end{aligned} \tag{C.11}$$

where $(a)$ follows as $\|\mathcal{P}_T\|_{\text{op}} \leq 1$.

Thus, $\sigma^2 := \sum_{s=1}^{|\Omega|} \|\mathbb{E}[(\mathcal{V}_s - \mathbb{E}[\mathcal{V}_s])^2]\|_{\text{op}} \leq \frac{1}{c_0\beta\log N}$

(iii) Lemma follows from applying $(i)$ and $(ii)$ above in operator Bernstein inequality (Lemma 2.3.2).

### C.3.2 Proof of Lemma C.2.2

Recall that under the assumptions in Section 5.3.1, $\|\cdot\|_{\mathscr{A}}$ is norm, and by the sub–differential characterization of norms the following holds:

$$\begin{aligned}\partial\|\mathcal{M}\|_{\mathscr{A}} =& \text{conv}\{\mathcal{Y} : \langle \mathcal{M}, \mathcal{Y}\rangle = \|\mathcal{M}\|_{\mathscr{A}}, \|\mathcal{Y}\|_{\mathscr{A}}^* \leq 1\} \\ =& \text{conv}\{\mathcal{E} + \mathcal{W} : \mathcal{E} \in \mathscr{E}(\mathcal{M}), \mathcal{W} \in T^{\perp}, \|\mathcal{W}\|_{\mathcal{A}}^* \leq 1\}\end{aligned} \tag{C.12}$$

Recall $\mathscr{E}(\mathcal{M})$ from (5.8). In particular the set $\{\mathcal{E}_{\mathcal{M}} + \mathcal{W} : \mathcal{W} \in T^{\perp}, \|\mathcal{W}\|_{\mathcal{A}}^* \leq 1\} \subset \partial\|\mathcal{M}\|_{\mathscr{A}}$, where $\mathcal{E}_{\mathcal{M}}$ is the sign vector from Assumption 5.3.2.

Given any $\Delta$, with $\mathcal{P}_{\Omega}(\Delta) = 0$, consider any $\mathcal{W} \in T^{\perp}$, such that $\|\mathcal{P}_{T^{\perp}}(\Delta)\|_{\mathscr{A}} = \langle \mathcal{W}, \mathcal{P}_{T^{\perp}}(\Delta)\rangle$ and $\mathcal{E}_{\mathcal{M}} + \mathcal{W} \in \partial\|\mathcal{M}\|_{\mathscr{A}}$. Let $\mathcal{Y} = \mathcal{P}_{\Omega}(\mathcal{Y})$ be a

dual certificate satisfying the conditions stated in the Lemma.

$$\begin{aligned}\|\mathcal{M}+\Delta\|_{\mathscr{A}} &\overset{(a)}{\geq} \|\mathcal{M}\|_{\mathscr{A}} + \langle \mathcal{E}_{\mathcal{M}} + \mathcal{W} - \mathcal{Y}, \Delta\rangle \\ &= \|\mathcal{M}\|_{\mathscr{A}} + \langle \mathcal{E}_{\mathcal{M}} - \mathcal{P}_T(\mathcal{Y}), \mathcal{P}_T(\Delta)\rangle + \langle \mathcal{W} - \mathcal{P}_{T^\perp}(\mathcal{Y}), \mathcal{P}_{T^\perp}(\Delta)\rangle \\ &\overset{(b)}{\geq} \|\mathcal{M}\|_{\mathscr{A}} - \|\mathcal{E}_{\mathcal{M}} - \mathcal{P}_T(\mathcal{Y})\|_F \|\mathcal{P}_T(\Delta)\|_F + \|\mathcal{P}_{T^\perp}(\Delta)\|_{\mathscr{A}}(1 - \|\mathcal{P}_{T^\perp}(\mathcal{Y})\|^*_{\mathscr{A}}) \\ &\overset{(c)}{\geq} \|\mathcal{M}\|_{\mathscr{A}} - \frac{1}{2}\kappa_\Omega(N)\|\mathcal{E}_{\mathcal{M}} - \mathcal{P}_T(\mathcal{Y})\|_F \|\mathcal{P}_{T^\perp}(\Delta)\|_F + \frac{1}{2}\|\mathcal{P}_{T^\perp}(\Delta)\|_{\mathscr{A}} \\ &\overset{(d)}{>} \|\mathcal{M}\|_{\mathscr{A}}, \end{aligned} \tag{C.13}$$

where $(a)$ follows as $\langle \Delta, \mathcal{Y}\rangle = 0$, $(b)$ follows from triangle inequality, $(c)$ follows as $\|\mathcal{P}_{T^\perp}(\mathcal{Y})\|^*_{\mathscr{A}} \leq \frac{1}{2}$ and $\frac{1}{2}\kappa_\Omega(N)\|\mathcal{P}_{T^\perp}(\Delta)\|_F \geq \|\mathcal{P}_T(\Delta)\|_F$ w.h.p. (from (C.5)), and $(d)$ follows as $\|\mathcal{E}_{\mathcal{M}} - \mathcal{P}_T(\mathcal{Y})\|_F < \frac{1}{\kappa_\Omega(N)}$ and using $\|\mathcal{X}\|_{\mathscr{A}} = \min_{Z\succcurlyeq 0} tr(Z)\ \ s.t. P_v[Z] = X_v\ \forall v \geq \min_{Z\succcurlyeq 0}\|Z\|_F\ \ s.t. P_v[Z] = X_v\ \forall v \geq \|\mathcal{X}\|_F$.

### C.3.3 Dual Certificate–Bound on $\|\mathcal{P}_{T^\perp}\mathcal{Y}_p\|^*_{\mathscr{A}}$

Recall that $\mathcal{Y}_p$ from Appendix C.2.3 following a golfing scheme introduced by Gross et al. [58]. The proof for the second property of the dual certificate, extends directly from the analogous proof for matrix completion by Recht [127].

$$\begin{aligned}\|\mathcal{P}_{T^\perp}\mathcal{Y}_p\|^*_{\mathscr{A}} &\leq \sum_{j=1}^{p} \|\mathcal{P}_{T^\perp}\mathcal{R}_{\Omega^{(j)}}\mathcal{W}_{j-1}\|^*_{\mathscr{A}} = \sum_{j=1}^{p} \|\mathcal{P}_{T^\perp}(\mathcal{R}_{\Omega^{(j)}} - \mathcal{I})\mathcal{W}_{j-1}\|^*_{\mathscr{A}} \\ &\leq \sum_{j=1}^{p} \|(\mathcal{R}_{\Omega^{(j)}} - \mathcal{I})\mathcal{W}_{j-1}\|^*_{\mathscr{A}} \end{aligned} \tag{C.14}$$

Denote $\max_{(v,i,j)} |\langle \mathcal{X}, \mathcal{E}^{(v,i,j)}\rangle| = \|\mathcal{X}\|_{\max}$. The following lemmas are directly adapted from Theorem 3.5 and Lemma 3.6 in [127]:

**Lemma C.3.1.** *Let $\Omega$ be any subset of entries of size $|\Omega|$ sampled independently such that $\mathbb{E}[\mathcal{R}_\Omega(\mathcal{W})] = \mathcal{W}$, then for all $\beta > 1$ and $N \geq 2$, the following holds with*

*probability greater than* $1 - N^{-\beta}$ *provided* $|\Omega| > 6N\beta \log N$, *and* $\frac{|\Omega_k|}{n_k m_k} \geq \frac{|\Omega|}{N^2}; \forall k$:

$$\|(\mathcal{R}_\Omega - \mathcal{I})\mathcal{W}\|_{\mathscr{A}}^* \leq \|\mathcal{B}(\mathcal{R}_\Omega \mathcal{W}) - \mathcal{B}(\mathcal{W})\|_{op} \leq \sqrt{\frac{8\beta N^3 \log N}{3|\Omega|}} \|\mathcal{W}\|_{max} \tag{C.15}$$

*Proof.* The proof is obtained by applying the steps described for the analogous proof in [127] on $\|\mathcal{B}(\mathcal{R}_\Omega \mathcal{W}) - \mathcal{B}(\mathcal{W})\|_{\text{op}}$. For $s = 1, 2, \ldots, |\Omega|$, let $\mathcal{V}_s = \mathcal{B}(\mathcal{R}_s(\mathcal{W}))$, then $\mathcal{B}(\mathcal{R}_\Omega \mathcal{W}) - \mathcal{B}(\mathcal{W}) = \sum_{s=1}^{|\Omega|} (\mathcal{V}_s - \mathbb{E}[\mathcal{V}_s])$ is a sum of independent zero mean random variables. From the proof of Theorem 3.5 in the work by Recht [128], for any $N \times N$ matrix $Z$, $\|Z\|_{\text{op}} \leq N \|Z\|_{\max}$.

(i) for $n \geq 2$,

$$\begin{aligned}
\|\mathcal{V}_s - \mathbb{E}[\mathcal{V}_s]\|_{\text{op}} &\leq \|\mathcal{V}_s\|_{\text{op}} + \|\mathbb{E}[\mathcal{V}_s]\|_{\text{op}} \\
&\overset{(a)}{\leq} \frac{N^2}{|\Omega|} \|\mathcal{W}\|_{\max} + \frac{N}{|\Omega|} \|\mathcal{W}\|_{\max} \leq \frac{3N^2}{2|\Omega|} \|\mathcal{W}\|_{\max}
\end{aligned}$$

where $(a)$ follows as $\frac{1}{p(v,i,j)} \leq \frac{1}{\min_k \frac{|\Omega_k|}{n_k m_k}} \leq \frac{N^2}{|\Omega|}$ if $\frac{|\Omega_k|}{n_k m_k} \geq \frac{|\Omega|}{N^2}; \forall k$.

(ii) The following holds:

$$\begin{aligned}
&\|\mathbb{E}[(\mathcal{V}_s - \mathbb{E}[\mathcal{V}_s])^2]\|_2 \leq \max\{\|\mathbb{E}[\mathcal{V}_s^2]\|_2, \|(\mathbb{E}[\mathcal{V}_s])^2\|_2\}, \\
&\|(\mathbb{E}[\mathcal{V}_s])^2\|_2 = \frac{1}{|\Omega|^2} \|\mathcal{B}(\mathcal{W})\|_2^2 \leq \frac{N^2}{|\Omega|^2} \|\mathcal{W}\|_{\max}^2, \text{ and} \\
&\|\mathbb{E}[\mathcal{V}_s^2]\|_2 = \frac{1}{|\Omega|} \Big\| \sum_{v=1}^{V} \sum_{(i,j) \in \mathcal{I}(v)} \frac{1}{p(v,i,j)} \langle \mathcal{W}, \mathcal{E}^{(v,i,j)} \rangle^2 \mathcal{B}(\mathcal{E}^{(v,i,j)}) \Big\|_2 \\
&\qquad \leq \frac{N^2}{|\Omega|^2} \|\mathcal{W}\|_{\max}^2 \Big\| \sum_{v=1}^{V} \sum_{(i,j) \in \mathcal{I}(v)} \mathcal{B}(\mathcal{E}^{(v,i,j)}) \Big\|_2 \leq \frac{N^3}{|\Omega|^2} \|\mathcal{W}\|_{\max}^2.
\end{aligned}$$

The proof follows by using Lemma 2.3.2 with $t = \sqrt{\frac{8\beta N^3 \log N}{3|\Omega|}} \|\mathcal{W}\|_{\max}$. □

**Lemma C.3.2.** *If* $\forall k$, $|\Omega_k| \geq c_0 \beta n_k R \log N$, *and the assumptions in Section 5.3.1 are satisfied, then for sufficiently large* $c_0$, *the following holds with probability*

*greater that* $1 - N^{-\beta}$*:*

$$\forall\ \mathcal{W} \in T,\ \|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{W} - \mathcal{W}\|_{max} \leq \frac{1}{2}\|\mathcal{W}\|_{max} \tag{C.16}$$

*Proof:* Using union bound and noting that $\sum_v n_{r_v} n_{c_v} \leq N^2$:

$$\mathbb{P}(\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{W} - \mathcal{W}\|_{\max} > \frac{1}{2}) \leq N^2\mathbb{P}(\langle \mathcal{P}_T\mathcal{R}_\Omega\mathcal{W} - \mathcal{W}, \mathcal{E}^{(v,i,j)}\rangle > 1/2 \text{ for any} (v,i,j))$$

For each $(v,i,j)$, sample $\mathcal{E}^{(s')} = \mathcal{E}^{(v_{s'},i_{s'},j_{s'})}$ according to the sampling distribution in Assumption 5.3.4. Define $\Psi_{(v,i,j)} = \langle \mathcal{E}^{(v,i,j)}, \mathcal{P}_T\mathcal{R}_{s'}\mathcal{W} - \frac{1}{|\Omega|}\mathcal{W}\rangle$. Recall the definition of $\mathcal{R}_s$. Now each entry of $\mathcal{P}_T\mathcal{R}_\Omega\mathcal{W} - \mathcal{W}$ is distributed as $\sum_{s=1}^{|\Omega|} \Psi^{(s)}_{(v,i,j)}$, where $\Psi^{(s)}_{(v,i,j)}$ are iid samples of $\Psi_{(v,i,j)}$.

Further: $|\Psi_{(v,i,j)}| \leq \frac{1}{p(v,i,j)}\|\mathcal{P}_T(\mathcal{E}^{(v,i,j)})\|_F^2 |\langle \mathcal{E}^{(v,i,j)}, \mathcal{W}\rangle| \leq \frac{1}{c'\beta \log N}\|\mathcal{W}\|_{\max}$.

Also, $\mathbb{E}[\Psi^2_{(v,i,j)}] = \mathbb{E}[\frac{1}{p(v,i,j)^2}\langle \mathcal{E}^{(v,i,j)}, \mathcal{W}\rangle^2 \langle \mathcal{E}^{(v,i,j)}, \mathcal{E}^{(s')}\rangle^2] \leq \frac{1}{|\Omega|c'\beta \log N}$, where the expectation is over $s'$. Standard Bernstein inequality (2.5) can be used with the above bounds to prove the lemma. □

**Remaining steps in the proof:** Using the above lemmas in the (C.14):

$$\begin{aligned}
\|\mathcal{P}_{T^\perp}\mathcal{Y}_p\|^*_{\mathscr{A}} &\leq \sum_{j=1}^{p} \|(\mathcal{R}_{\Omega^{(j)}} - \mathcal{I})\mathcal{W}_{j-1}\|^*_{\mathscr{A}} \overset{(a)}{\leq} \sum_{j=1}^{p} \sqrt{\frac{8\beta N^3 \log N}{3|\Omega^{(j)}|}}\|\mathcal{W}_{j-1}\|_{\max} \\
&\overset{(b)}{\leq} 2\sum_{j=1}^{p} 2^{-j}\sqrt{\frac{8\beta N^3 \log N}{3|\Omega^{(j)}|}}\|\mathcal{E}_{\mathcal{M}}\|_{\max} \\
&\overset{(c)}{\leq} 2\sum_{j=1}^{p} 2^{-j}\sqrt{\frac{8\beta\mu_1 RN \log N}{3|\Omega^{(j)}|}} \overset{(d)}{\leq} \frac{1}{2},
\end{aligned} \tag{C.17}$$

where $(a)$ follows from Lemma C.3.1, $(b)$ from Lemma C.3.1 as $\mathcal{W}_j = \mathcal{W}_{j-1} - \mathcal{P}_T\mathcal{R}_\Omega\mathcal{W}_{j-1}$, $(c)$ from the second incoherence condition stated in Assumption 5.3.2, and finally $(d)$ if for large enough $c_1$, $|\Omega^{(j)}| > c_1\mu_1\beta RN \log N$.

Finally, the probability that the proposed dual certificate $\mathcal{Y}_p$ fails the conditions of Lemma C.2.2 is given by a union bound of the failure probabilities of (C.7), and Lemma C.3.1 and C.3.1 for each partition $\Omega^{(j)}$: $3c_1 \log(N\kappa_\Omega(N))N^{-\beta}$; thus proving Theorem 5.4.1.

# Appendix D

# Appendix for Preference Completion from Partial Rankings

## D.1 Estimator and Algorithm

### D.1.1 Proof of Proposition 7.3.1

**Statement of the Proposition:** The optimization in (7.2) is jointly convex in $(X, z)$. Further, $\forall \gamma > 0$, $(\lambda, \gamma\epsilon)$ and $(\gamma^{-1}\lambda, \epsilon)$ lead to equivalent estimators, specifically $\widehat{\mathfrak{X}}(\lambda, \gamma\epsilon) = \gamma^{-1}\widehat{\mathfrak{X}}(\gamma^{-1}\lambda, \epsilon)$.

**Proof:** Let $f_{\lambda,\epsilon}(X) = \min\limits_{z \in \mathbb{R}^{|\Omega|}} \lambda\|X\|_* + \frac{1}{2}\|z - \mathcal{P}_\Omega(X)\|_2^2$ . s.t. $\forall j,\ z_{\Omega_j} \in \mathcal{R}^{n_j}_{\downarrow\epsilon}(y^{(j)})$,

We have,

$$
\begin{aligned}
f_{\lambda,\gamma\epsilon}(X) &= \min_z \lambda\|X\|_* + \frac{1}{2}\|z - \mathcal{P}_\Omega(X)\|_2^2 \ \text{ s.t. } \ z_{\Omega_j} \in \mathcal{R}^{n_j}_{\downarrow\gamma\epsilon}(y^{(j)}), \\
&\overset{(a)}{=} \min_{\bar{z}} \lambda\|X\|_* + \frac{1}{2}\|\gamma\bar{z} - \mathcal{P}_\Omega(X)\|_2^2 \ \text{ s.t. } \ \bar{z}_{\Omega_j} \in \mathcal{R}^{n_j}_{\downarrow\epsilon}(y^{(j)}), \\
&= \gamma^2 \min_{\bar{z}} \frac{\lambda}{\gamma}\|X/\gamma\|_* + \frac{1}{2}\|\bar{z} - \mathcal{P}_\Omega(X/\gamma)\|_2^2 \ \text{ s.t. } \ \bar{z}_{\Omega_j} \in \mathcal{R}^{n_j}_{\downarrow\epsilon}(y^{(j)}), \\
&= \gamma^2 f_{\gamma^{-1}\lambda,\epsilon}(X/\gamma),
\end{aligned}
\tag{D.1}
$$

where $(a)$ follows from reparameterizing the optimization using $\bar{z} = z/\gamma$ as the geometry of $\mathcal{R}^{n_j}_{\downarrow\gamma\epsilon}(y^{(j)})$ which is set of linear constraints of the form $z_i - z_k \leq \gamma\epsilon$. From above set of equations, if $X \in \operatorname{Argmin}_X f_{\lambda,\gamma\epsilon}(X)$, then $\gamma^{-1}X \in \operatorname{Argmin}_X f_{\gamma^{-1}\lambda,\epsilon}(X)$.

### D.1.2 Proof of Lemma 7.4.1

**Statement of the Lemma:** Consider the following steps,

$$\begin{aligned} &\text{Step 1. } \pi^*(x) \text{ s.t. } \forall k \in [K],\ \pi^*(x)_{P_k} = \text{sort}(x_{P_k}) \\ &\text{Step 2. } \widehat{z} = PAV(\pi^*(x) - \epsilon \boldsymbol{d}^{\text{bl}}) + \epsilon \boldsymbol{d}^{\text{bl}}. \end{aligned} \tag{D.2}$$

Estimate $\widehat{z}$ is the unique minimizer for

$$\underset{z}{\text{argmin}} \|z - x\|_2^2 \text{ s.t. } \exists \pi \in \Pi_P : \boldsymbol{D_n} \pi(z) \leq \epsilon \boldsymbol{D_n} \boldsymbol{d}^{\text{bl}}.$$

**Proof:** A version of the lemma for linear orders was proved in [5]. In general,

$$\begin{aligned} &\min_z \|z - x\|_2^2 \text{ s.t. } \exists \pi \in \Pi_P : \boldsymbol{D_n} \pi(z) \leq \epsilon \boldsymbol{D_n} \boldsymbol{d}^{\text{bl}} \\ &= \min_{z, \pi \in \Pi_P} \|z - x\|_2^2 \text{ s.t. } \boldsymbol{D_n} \pi(z) \leq \epsilon \boldsymbol{D_n} \boldsymbol{d}^{\text{bl}} \\ &\overset{(a)}{=} \min_w \min_{\pi \in \Pi_P} \|\pi^{-1}(w + \epsilon \boldsymbol{d}^{\text{bl}}) - x\|_2^2 \text{ s.t. } \boldsymbol{D_n} w \\ &\leq 0 \overset{(b)}{=} \min_{w: \boldsymbol{D_n} w \leq 0} \min_{\pi \in \Pi_P} \|w + \epsilon \boldsymbol{d}^{\text{bl}} - \pi(x)\|_2^2 \\ &\overset{(c)}{=} \min_{w: \boldsymbol{D_n} w \leq 0} \|w + \epsilon \boldsymbol{d}^{\text{bl}} - \pi^*(x)\|_2^2, \end{aligned} \tag{D.3}$$

where $\pi^*(x)$ is the update from Step 1 stated above, $(a)$ follows reparametrizing $w := \pi(z) - \epsilon \boldsymbol{d}^{\text{bl}}$, $(b)$ follows as for all permutations $\pi$ using $\|x\|_2^2 = \|\pi(x)\|_2^2$, and $(c)$ follows form Proposition D.1.1 as $\boldsymbol{D_n} w \leq 0$ from constraints and $\epsilon \boldsymbol{D_n} \boldsymbol{d}^{\text{bl}} \leq 0$ by construction. The final minimization is solved using Step 2. □

**Proposition D.1.1.** *For any sorted* $z \in \mathbb{R}^n$ *such* $\boldsymbol{D_n} z \leq 0$, $\pi^* = \underset{\pi \in \Pi_P}{\text{argmin}} \|z - \pi(x)\|_2^2$, *where* $\pi^*$ *is the permutation from Step* 1.

$\Pi_P$ allows for all possible permutations within each partition $P_k$. Proposition follows from optimality of sorting within each block. □

## D.2 Generalization Error

### D.2.1 Background

**Definition D.2.1** (Rademacher Complexity)**.** Let $X_1, X_2, \ldots, X_n \in \mathcal{X}$ be drawn iid from a distribution $\mathbb{P}_X$. For a function class $\mathcal{F} : \mathcal{X} \to \mathcal{A}$, the empirical Rademacher complexity is defined as,

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \Big( \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \Big),$$

where $\sigma_1, \sigma_2, \ldots, \sigma_n$ are iid Rademacher variables, i.e., $\pm 1$ with probability $1/2$.

The Rademacher complexity with respect tp $\mathbb{P}_X$ is then defined as $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{\mathbb{P}_X} \widehat{\mathfrak{R}}_n(\mathcal{F})$.

**Theorem D.2.1** (Generalization Error Bound (Corollary 15 in [15]))**.** *Consider a loss function $\ell : \mathcal{Y} \times \mathbb{R}^m \to [0, 1]$ and a bounded function class $\mathcal{F} : \mathcal{X} \to \mathbb{R}^m$ such that $\mathcal{F}$ is a direct sum of $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_m$. Further, if $\ell$ is $L$–Lipschitz continuous with respect to Euclidean distance on $\mathbb{R}^m$ and is uniformly bounded. Let $\{(X_i, Y_i), i = 1, 2, \ldots, n\}$ be sampled form a distribution $\mathbb{P}_{X,Y}$. Then there exists a constant $c$ such that, for any integer $n$ and any $\delta \in (0, 1)$, with probability atleast $1 - \delta$, over all sample of length $n$, the following holds for every $f \in \mathcal{F}$:*

$$\mathbb{E}_{X,Y} \ell(Y, f(X)) \leq \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) + cL \sum_{i=1}^{m} \widehat{\mathfrak{R}}_n(\mathcal{F}_m) + \sqrt{\frac{8 \log(2/\delta)}{n}}$$

### D.2.2 Proof of Theorem 7.5.1

**Lemma D.2.2.** *$\phi(., y)$ is convex and $2$–Lipschitz continuous with respect to $\ell_2$ norm.*

**Proof:** Convexity follows form $\Phi$ being a marginal of a convex function. For a any convex set $C$ and its projection operator $P_C$, we have the following for all

$x, x'$:

$$\begin{aligned}|\|x - P_C(x)\|_2 - \|x' - P_C(x')\|_2| &\leq \|x - P_C(x) - x' + P_C(x')\|_2 \\ &\leq \|x - x'\|_2 + \|P_C(x) - P_C(x')\|_2 \leq 2\|x - x'\|_2\end{aligned}$$

Consider a vector class of functions in $\mathbb{R}^R$, $\mathcal{F}_R = \{\Omega(s) \to X_{\Omega(s)} \in \mathbb{R}^R : \|X\|_* \leq M\}$, where $\Omega(s)$ are sampled as in the main paper. Also, consider another function classes $\mathcal{F}_{ij} = \{(i,j) \to X_{ij} : \|X\|_* \leq M\}$. It can be seen that $\mathcal{F}_R$ is an $R$ way direct sum of $\mathcal{F}_{ij}$. In order to use Theorem D.2.1, we need to estimate the Rademacher complexity of $\mathcal{F}_{ij}$.

**Lemma D.2.3.** *Let $\Omega = \cup_j \Omega_j$ obtained from combining samples form Assumption 7.5.1. The distribution of $\Omega$ is equivalent to uniformly sampling with replacement $|\Omega| = c_0 d_2 R \log d_2$ entries from $[d_1] \times [d_2]$.*

*Proof*: For $k = 1, 2 \ldots |\Omega|$, $\forall (i,j) \in [d_1] \times [d_2]$,
$\mathbb{P}((i,j) = \Omega_k) = \frac{1}{d_1 d_2}$.

Thus, given $(i,j) \in [d_1] \times [d_2]$, $\mathbb{P}((i,j) \in \Omega) = \frac{|\Omega|}{d_1 d_2}$. □

**Lemma D.2.4** (Theorem 29 in [139])**.** *For a universal constant $K$, the Rademacher complexity of matrices in $\mathbb{R}^{d_1 \times d_2}$ of trace norm $M$, over uniform sampling of index pairs $\Omega$ is bounded by the following whenever $|\Omega| > d \log d$*

$$\mathfrak{R}(\{\|X\|_* \leq M\}) \leq K \frac{M \log^{1/4} d}{\sqrt{d_1 d_2}} \sqrt{\frac{d \log d}{|\Omega|}} \tag{D.4}$$

From Lemma D.2.3, it can be seen that Lemma D.2.4 applies to samples drawn according to Assumption 7.5.1.

For the function class $\mathcal{F}_R = \{\Omega(s) \to X_{\Omega(s)} : \|X\|_* \leq M\}$, for some $M$. The theorem now follows by using the Rademacher complexity bound in Lemma D.2.4 and Lipschitz continuity of $\Phi(., y)$ from D.2.2 in Theorem D.2.1.

# Bibliography


[1] Y. Abramovich, N. Spencer, and A. Gorokhov. Positive-definite toeplitz completion in doa estimation for nonuniform linear antenna arrays. ii. partially augmentable arrays. *IEEE Transactions on Signal Processing*, 1999.

[2] E. Acar, T. G. Kolda, and D. M. Dunlavy. All-at-once optimization for coupled matrix and tensor factorizations. In *International Workshop on Mining and Learning with Graphs*, 2011.

[3] E. Acar, E. E. Papalexakis, G. Gurdeniz, M. A. Rasmussen, A. J. Lawaetz, M. Nilsson, and R. Bro. Structure-revealing data fusion. *BMC Bioinformatics*, 2014.

[4] S. Acharyya and J. Ghosh. MEMR: A margin equipped monotone retargeting framework for ranking. In *Conference on Uncertainty in Artificial Intelligence*, 2013.

[5] S. Acharyya, O. Koyejo, and J. Ghosh. Learning to rank with bregman divergences and monotone retargeting. In *Conference on Uncertainty in Artificial Intelligence*, 2012.

[6] D. Agarwal and B. C. Chen. Regression-based latent factor models. In *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009.

[7] D. Agarwal and B. C. Chen. flda: matrix factorization through latent dirichlet allocation. In *ACM International Conference on Web Search and Data Mining*, 2010.

[8] D. Agarwal, B. C. Chen, and B. Long. Localized factor models for multi-context recommendation. In *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2011.

[9] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 2002.

[10] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *IMA Information and Inference*, 2014.

[11] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the $k$-support norm. In *Advances in Neural Information Processing Systems*, 2012.

[12] H. Avron, S. Kale, S. Kasiviswanathan, and V. Sindhwani. Efficient and practical stochastic subgradient descent for nuclear norm regularization. *International Conference on Machine Learning*, 2012.

[13] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, 2014.

[14] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 2005.

[15] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2003.

[16] M. J. Best and N. Chakravarti. Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 1990.

[17] S. Bhojanapalli, P. Jain, and S. Sanghavi. Tighter low-rank approximation via sampling the leveraged element. *arXiv preprint*, 2014.

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

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[20] G. Bouchard, S. Guo, and D. Yin. Convex collective matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, 2013.

[21] R. S. Cabral, F. Torre, J. P. Costeira, and A. Bernardino. Matrix completion for multi-label image classification. In *Advances in Neural Information Processing Systems*, 2011.

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

[23] T. Cai, T. Liang, and A. Rakhlin. Geometrizing local rates of convergence for linear inverse problems. *The Annals of Statistics*, 2016.

[24] E. J. Candés, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 2011.

[25] E. J. Candés and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 2010.

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

[27] E. J. Candés, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 2006.

[28] E. J. Candés and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 2005.

[29] E. J. Candés and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies. *IEEE Transactions on Information Theory*, 2006.

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

[31] Z. Cao, T. Qin, T. Y. Liu, M. F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *International Conference on Machine Learning*, 2007.

[32] R. J. Carroll, A. E. Eyler, and J. C. Denny. Naïve electronic health record phenotype identification for rheumatoid arthritis. In *AMIA Annual Symposium*, 2011.

[33] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 2012.

[34] Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu. Deep Computational Phenotyping. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[35] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Completing any low-rank matrix, provably. *Journal of Machine Learning Research*, 2015.

[36] Y. Chen, R. J. Carroll, E. Hinz, A. Shah, A. E. Eyler, J. C. Denny, and H. Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 2013.

[37] Y. Chen, N. M. Lorenzi, W. S. Sandberg, K. Wolgast, and B. A. Malin. Identifying collaborative care teams through electronic medical record utilization patterns. *Journal of the American Medical Informatics Association*, page ocw124, 2016.

[38] Y. Chen and X Ye. Projection Onto A Simplex. *arXiv preprint*, January 2011.

[39] E. Chi, H. Zhou, G. Chen, D. O. Del Vecchyo, and K. Lange. Genotype imputation via matrix completion. *Genome research*, 2013.

[40] E. C. Chi and T. G. Kolda. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 2012.

[41] A. Cichocki, A. H. Phan, and C. Caiafa. Flexible HALS algorithms for sparse non-negative matrix/tensor factorization. In *IEEE International Workshop on Machine Learning for Signal Processing*, 2008.

[42] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[43] M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, 2001.

[44] P. Cremonesi, Y. Koren, and R. Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *ACM Recommender Systems Conference*, 2010.

[45] M. A. Davenport, Y. Plan, E. Berg, and M. Wootters. 1-bit matrix completion. *IMA Information and Inference*, 2014.

[46] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 2006.

[47] J. C. Duchi, L. W Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *International Conference on Machine Learning*, 2010.

[48] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1967.

[49] D. Dueck, Q. D. Morris, and B. J. Frey. Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics*, 2005.

[50] A. Edelman. Eigenvalues and condition numbers of random matrices. *Journal on Matrix Analysis and Applications*, 1988.

[51] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *International Conference on Computer Vision and Pattern Recognition*, 2012.

[52] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

[53] M. Fazel, H Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference*, 2001.

[54] J. Forster and M. Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 2002.

[55] R. Ganti, L. Balzano, and R. Willett. Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, 2015.

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

[57] G. J. Gordon. Generalizedˆ 2 linearˆ 2 models. In *Advances in Neural Information Processing Systems*, 2002.

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

[59] S. J. Grotzinger and C. Witzgall. Projections onto order simplexes. *Applied Mathematics and Optimization*, 1984.

[60] S. Gunasekar, A. Acharya, N. Gaur, and J. Ghosh. Noisy matrix completion using alternating minimization. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery*, 2013.

[61] S. Gunasekar, A. Banerjee, and J. Ghosh. Unified view of matrix completion under general structural constraints. In *Advances in Neural Information Processing Systems*, 2015.

[62] S. Gunasekar, P. Ravikumar, and J. Ghosh. Exponential family matrix completion under structural constraints. In *International Conference on Machine Learning*, 2014.

[63] S. Gunasekar, M. Yamada, D. Yin, and Y. Chang. Consistent collective matrix completion under joint low rank structure. In *International Conference on Artificial Intelligence and Statistics*, 2015.

[64] M. Hardt. On the provable convergence of alternating minimization for matrix completion. *arXiv preprint*, 2013.

[65] N. Harvey, David R Karger, and Kazuo Murota. Deterministic network coding by matrix completion. In *ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2005.

[66] S. Hasan, G. T. Duncan, D. B. Neill, and R. Padman. Automatic detection of omissions in medication lists. *Journal of the American Medical Informatics Association*, 2011.

[67] E. Hazan. Sparse approximate solutions to semidefinite programs. In *Latin American Theoretical Informatics Symposium*. Springer, 2008.

[68] R. Henao, J. T. Lu, J. E. Lucas, J. Ferranti, and L. Carin. Electronic health record analysis via deep poisson factor models. *Journal of Machine Learning Research*, 2015.

[69] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Neural Information Processing Systems*, 1999.

[70] J. C Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 2014.

[71] J. C. Ho, J. Ghosh, and J. Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

[72] T. Hofmann. Probabilistic latent semantic indexing. In *International ACM SIGIR conference on Research and Development in Information Retrieval*, 1999.

[73] J. L. Horowitz. *Semiparametric and nonparametric methods in econometrics*. Springer, 2009.

[74] G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 2013.

[75] C. Hsieh and P. Olsen. Nuclear norm minimization via active subspace selection. In *International Conference on Machine Learning*, 2014.

[76] L. Jacob, J. P. Vert, and F. R. Bach. Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems*, 2009.

[77] M. Jaggi and M. Sulovsk. A simple algorithm for nuclear norm regularized problems. In *International Conference on Machine Learning*, 2010.

[78] P. Jain and I. S. Dhillon. Provable inductive matrix completion. *arXiv preprint*, 2013.

[79] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Annual Symposium on the Theory of Computing*, 2013.

[80] H. Ji, C. Liu, Z. Shen, and Y. Xu. Robust video denoising using low rank matrix completion. In *International Conference on Computer Vision and Pattern Recognition*, 2010.

[81] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *International Conference on Machine Learning*, 2009.

[82] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[83] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[84] S. M. Kakade, V. Kanade, O. Shamir, and A. Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, 2011.

[85] S. M. Kakade, O. Shamir, K. Sridharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *International Conference on Artificial Intelligence and Statistics*, 2010.

[86] A. T. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *International Conference on Computational Learning Theory*, 2009.

[87] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 2010.

[88] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 2010.

[89] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 2014.

[90] O. Klopp. Matrix completion by singular value thresholding: sharp bounds. *Electronic Journal of Statistics*, 2015.

[91] Olga Klopp. High dimensional matrix estimation with unknown variance of the noise. *arXiv preprint*, 2011.

[92] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 2009.

[93] V. Koltchinskii, K. Lounici, A. B. Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 2011.

[94] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 2009.

[95] O. Koyejo, S. Acharyya, and J. Ghosh. Retargeted matrix factorization for collaborative filtering. In *ACM Recommender Systems Conference*, 2013.

[96] M. Laurent. Matrix completion problems. *Encyclopedia of Optimization*, 2009.

[97] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

[98] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999.

[99] H. Lee, Y. Kim, A. Cichocki, and S. Choi. Nonnegative tensor factorization for continuous EEG classification. *International Journal of Neural Systems*, 2007.

[100] P. Li, Q. Wu, and C. J. Burges. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in Neural Information Processing Systems*, 2007.

[101] C. J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 2007.

[102] C. Lippert, S. H. Weber, Y. Huang, V. Tresp, M. Schubert, and H. P. Kriegel. Relation prediction in multi-relational domains using matrix factorization. In *NIPS Workshop: Structured Input-Structured Output*, 2008.

[103] A. E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Advances in Mathematics*, 2005.

[104] J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *SIAM International Conference on Data Mining*, 2013.

[105] T. Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 2009.

[106] B. Long, Z. M. Zhang, X. Wu, and P. S. Yu. Spectral clustering for multi-type relational data. In *International Conference on Machine Learning*, 2006.

[107] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *ACM International Conference on Web Search and Data Mining*, 2011.

[108] S. Ma, D. Goldfarb, and L. Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 2011.

[109] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 2010.

[110] J. D. Mcauliffe and D. M. Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, 2008.

[111] A. M. McDonald, M. Pontil, and D. Stamos. New perspectives on k-support and cluster norms. *arXiv preprint*, 2014.

[112] A. Mnih and R. Salakhutdinov. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, 2007.

[113] S. Mohamed, Z. Ghahramani, and K. A. Heller. Bayesian exponential family PCA. In *Advances in Neural Information Processing Systems*, 2008.

[114] S. Negahban. *Structured Estimation in High-Dimensions*. PhD thesis, EECS, University of California, Berkeley, 2012.

[115] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 2012.

[116] S. Negahban, B. Yu, M. J. Wainwright, and P. Ravikumar. A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, 2009.

[117] K. M. Newton, P. L. Peissig, A. N. Kho, S. J. Bielinski, R. L. Berg, V. Choudhary, M. Basford, C. G. Chute, I. J. Kullo, R. Li, J. A. Pacheco, L. V. Rasmussen, L. Spangler, and J. C. Denny. Validation of electronic medical record-based phenotyping algorithms: Results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association*, 2013.

[118] NIH Health Care Systems Research Collaboratory. Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials, 2014.

[119] S. Oh, K. K. Thekumparampil, and J. Xu. Collaboratively learning preferences from ordinal data. In *Advances in Neural Information Processing Systems*, 2015.

[120] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 1994.

[121] N. Parikh and S. P. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 2014.

[122] D. Park, J. Neeman, J. Zhang, S. Sanghavi, and I. Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *International Conference on Machine Learning*, 2015.

[123] P. L. Peissig, V. S. Costa, M. D. Caldwell, C. Rottscheit, R. L. Berg, E. A. Mendonca, and D. Page. Relational machine learning for electronic health record-driven phenotyping. *Journal of Biomedical Informatics*, 2014.

[124] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Conference on Empirical Methods in Natural Language Processing*, 2009.

[125] G. Raskutti, M. J Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 2010.

[126] G. Raskutti and M. Yuan. Convex regularization for high-dimensional tensor regression. *arXiv preprint*, 2015.

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

[128] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 2010.

[129] E. Richard, G. Obozinski, and J.-P. Vert. Tight convex relaxations for sparse matrix factorization. *arXiv preprints*, 2014.

[130] M. D. Ritchie, J. C. Denny, D. C. Crawford, A. H. Ramirez, J. B. Weiner, J. M. Pulley, M. A. Basford, K. Brown-Gentry, J. R. Balser, D. R. Masys,

J. L. Haines, and D. M. Roden. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *The American Journal of Human Genetics*, 2010.

[131] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *International Conference on Machine Learning*, 2008.

[132] P. Schulam, F. Wigley, and S. Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *AAAI Conference on Artificial Intelligence*, 2015.

[133] H. Shan and A. Banerjee. Generalized probabilistic matrix factorizations for collaborative filtering. In *IEEE International Conference on Data Mining*, 2010.

[134] C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 2014.

[135] A. Singer and M. Cucuringu. Uniqueness of low-rank matrix completion by rigidity theory. *SIAM Journal on Matrix Analysis and Applications*, 2010.

[136] A. P. Singh. Efficient matrix models for relational learning. Technical report, Carnegie Mellon University, 2009.

[137] A. P. Singh and G. Gordon. A Bayesian matrix factorization model for relational data. *Conference on Uncertainty in Artificial Intelligence*, 2010.

[138] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2008.

[139] N. Srebro. *Learning with matrix factorizations*. PhD thesis, Massachusetts Institute of Technology, 2004.

[140] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, 2005.

[141] H. Steck. Training and testing of recommender systems on data missing not at random. In *SIGKDD international conference on Knowledge discovery and data mining*, 2010.

[142] Q. F. Stout. Isotonic regression via partitioning. *Algorithmica*, 2013.

[143] M. Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 1996.

[144] M. Talagrand. Majorizing measures without measures. *The Annals of Probability*, 2001.

[145] M. Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014.

[146] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1999.

[147] K. C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 2010.

[148] J. A. Tropp. Literature survey: Nonnegative matrix factorization. Technical report, The University of Texas at Asutin, 2003.

[149] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 2012.

[150] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*. Springer, 2015.

[151] R. Vershynin. A note on sums of independent random matrices after ahlswede-winter. *Lecture notes*, 2009.

[152] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed sensing*, 2012.

[153] R. Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, a Renaissance*. Springer, 2015.

[154] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun. Rubik: Knowledge Guided Tensor Factorization and Completion for Health Data Analytics. In *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015.

[155] Y. X. Wang, H. Xu, and C. Leng. Provable subspace clustering: When lrr meets ssc. In *Advances in Neural Information Processing Systems*, 2013.

[156] A. G. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra Applications*, 1992.

[157] M. Weimer, A. Karatzoglou, Q. V. Le, and A. J. Smola. COFIRANK - maximum margin matrix factorization for collaborative ranking. In *Advances in Neural Information Processing Systems*, 2008.

[158] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR conference on Research and development in informaion retrieval*, 2003.

[159] E. Yang, G. Allen, Z. Liu, and P. Ravikumar. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, 2012.

[160] E. Yang and P. Ravikumar. Dirty statistical models. In *Advances in Neural Information Processing Systems*, 2013.

[161] E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Conditional random fields via univariate exponential families. In *Advances in Neural Information Processing Systems*, 2013.

[162] T. Yarkoni. http://neurosynth.org/. http://neurosynth.org/, 2011.

[163] T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, and T. D. Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 2011.

[164] K. Y. Yilmaz, A. T. Cemgil, and U. Simsekli. Generalised coupled tensor factorisation. In *Advances in Neural Information Processing Systems*, 2011.

[165] S. Zhang, Q. Li, J. Liu, and X. J. Zhou. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics*, 2011.

[166] Y. Zhang, B. Cao, and D. Y. Yeung. Multi-domain collaborative filtering. In *Conference on Uncertainty in Artificial Intelligence*, 2010.

[167] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the netflix prize. In *International Conference on Algorithmic Aspects in Information and Management*, 2008.