Convex Optimization Algorithms and Statistical Bounds for Learning Structured Models

Amin Jalali

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Professor Maryam Fazel, Chair

Professor Jeffrey A. Bilmes

Professor Ioana Dumitriu

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

Abstract

Convex Optimization Algorithms and Statistical Bounds for Learning Structured Models

Amin Jalali

Chair of the Supervisory Committee:
Professor Maryam Fazel
Electrical Engineering

Design and analysis of tractable methods for estimation of structured models from massive high-dimensional datasets has been a topic of research in statistics, machine learning and engineering for many years. Regularization, the act of simultaneously optimizing a data fidelity term and a structure-promoting term, is a widely used approach in different machine learning and signal processing tasks. Appropriate regularizers, with efficient optimization techniques, can help in exploiting the prior structural information on the underlying model. This dissertation is focused on exploring new structures, devising efficient convex relaxations for exploiting them, and studying the statistical performance of such estimators. We address three problems under this framework on which we elaborate below.

In many applications, we aim to reconstruct models that are known to have more than one structure at the same time. Having a rich literature on exploiting common structures like sparsity and low rank at hand, one could pose similar questions about *simultaneously structured models* with several low-dimensional structures. Using the respective known convex penalties for the involved structures, we show that multi-objective optimization with these penalties can do no better, order-wise, than exploiting only one of the present structures. This suggests that to fully exploit the multiple structures, we need an entirely new convex relaxation, not one that combines the convex relaxations for each structure. This work, while applicable for general structures, yields interesting results for the case of sparse and low-rank

matrices which arise in applications such as sparse phase retrieval and quadratic compressed sensing.

We then turn our attention to the design and efficient optimization of convex penalties for structured learning. We introduce a general class of semidefinite representable penalties, called *variational Gram functions* (VGF), and provide a list of optimization tools for solving regularized estimation problems involving VGFs. Exploiting the variational structure in VGFs, as well as the variational structure in many common loss functions, enables us to devise efficient optimization techniques as well as to provide guarantees on the solutions of many regularized loss minimization problems.

Finally, we explore the statistical and computational trade-offs in the community detection problem. We study recovery regimes and algorithms for community detection in sparse graphs generated under a *heterogeneous stochastic block model* in its most general form. In this quest, we were able to expand the applicability of semidefinite programs (in exact community detection) to some new and important network configurations, which provides us with a better understanding of the ability of semidefinite programs in reaching statistical identifiability limits.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express my gratitude to my advisor, Maryam Fazel, for her mentorship and support, to Professors Aleksandr Aravkin, Jeffrey Bilmes, James Burke, Ioana Dumitriu, and Marina Meila, for kindly serving on my thesis committee, to Professor Babak Hassibi, Dr. Lin Xiao, and Dr. James Saunderson, with whom I have had the privilege of working, to Professors Mehran Mesbahi and Rekha Thomas for their continuous encouragement, and to my family and friends whose support has been absolutely invaluable over the years.

I also gratefully acknowledge the funding from the National Science Foundation under grant numbers CCF-1409836 and ECCS-0847077.

DEDICATION

To my family.

Chapter 1

INTRODUCTION

Design and analysis of tractable methods for estimation of structured models from massive high-dimensional datasets has been a topic of research in statistics, machine learning and engineering for many years. Regularization, the act of simultaneously optimizing a data fidelity term and a structure-promoting term, is a widely used approach in different machine learning and signal processing tasks. Appropriate regularizers, with efficient optimization techniques, can help in exploiting the prior structural information on the underlying model. This dissertation is focused on exploring new structures, devising efficient convex relaxations for exploiting them, and studying the statistical performance of such estimators. We address three problems under this framework on which we elaborate below.

1.1 Recovery of Simultaneously Structured Models

Recovery of a structured model (signal) given a small number of linear observations has been the focus of many studies in the past decade. Examples include recovering sparse or group-sparse vectors (which gave rise to the area of compressed sensing) [40, 56, 38, 61], low-rank matrices [131, 35], and the sum of sparse and low-rank matrices [41, 32], among others. More generally, the recovery of a signal that can be expressed as the sum of a few atoms out of an appropriate atomic set has been studied in [42]. Canonical questions in this area include: How many linear measurements are enough to recover the model by any means? How many measurements are enough for a tractable approach? In the statistics literature, these questions are often posed in terms of error rates for estimators minimizing the sum of a quadratic loss function and a regularizer that reflects the desired structure [124].

There are many applications where the model of interest is known to have *several* struc-

tures at the same time (see Section 2.1.2 for a list of applications). We then seek a signal that lies in the intersection of several sets defining the individual structures (in a sense that we will later make precise). Often, appropriate regularizers that promote each individual structure are known and allow for recovery using an order-wise optimal number of measurements (e.g., ℓ_1 norm for sparsity, nuclear norm for matrix low rank). Hence, it is reasonable to minimize a combination of such norms. More specifically, one can consider solving

$$\begin{aligned} \underset{x\in\mathcal{C}}{\text{minimize}} \quad & f(x) = h(\|x\|_{(1)}, \ldots, \|x\|_{(\tau)}) \\ \text{subject to} \quad & \mathcal{A}(x) = \mathcal{A}(x_0), \end{aligned} \tag{1.1}$$

where $h : \mathbb{R}_+^\tau \to \mathbb{R}_+$ is convex and non-decreasing in each argument (i.e., non-decreasing and strictly increasing in at least one coordinate), $\mathcal{A}$ is a linear measurement operator, and the norms $\|\cdot\|_{(i)}$, $i = 1, \ldots, \tau$, correspond to the τ simultaneous structures that are present in x_0. However, there has been no general analysis and understanding of how well such regularization performs in terms of the number of observations required for successful recovery of the desired model.

In Chapter 2 we address this ubiquitous yet unexplored problem; i.e., the recovery of *simultaneously structured models*. The setting considered is very broad: any number of structures can be combined, any (convex) combination of the corresponding norms can be analyzed, and the results hold for a variety of popular measurement ensembles. The framework proposed includes special cases that are of interest in their own right, e.g., sparse and low-rank matrix recovery, and low-rank tensor completion [68, 75].

We show that, surprisingly, using multi-objective optimization with these regularizers (which is equivalent to solving an instance of (1.1)) can do no better, order-wise, than exploiting only one of the structures, thus revealing a fundamental limitation in sample complexity. This result suggests that to fully exploit the multiple structures, we need an entirely new convex relaxation. Further, specializing our results to the case of sparse and low-rank matrices, we show that a nonconvex formulation recovers the model from very few measurements (on the order of the degrees of freedom). In contrast, the convex problem

combining the ℓ_1 and nuclear norms requires many more measurements, illustrating a gap between the performance of the convex and nonconvex recovery problems. In nonconvex recovery, we assume we are able to find the global minimum of a nonconvex problem. This is clearly intractable in general, and not a practical recovery method—we consider it as a benchmark for theoretical comparison with the (tractable) convex relaxation in order to determine how powerful the relaxation is. Our framework applies to arbitrary structure-inducing norms as well as to a wide range of measurement ensembles. This allows us to give sample complexity bounds for problems such as sparse phase retrieval and low-rank tensor completion.

A similar bottleneck also appears in a related estimation problem as follows. The problem of identifying an unknown model given noisy observations is commonly referred to as *denoising*. In many denoising tasks where the underlying model admits a low dimensional structure (e.g., a sparse vector or a low-rank matrix), a convex optimization program (proximal mapping) achieves an estimation accuracy that is proportional to the noise level and scales with the number of degrees of freedom in the model; e.g., see [55, 57, 20]. For denoising a noisy observation $y = x_0 + z$, where x_0 is a simultaneously structured model and z is the additive noise, one can use a similar approach as in the aforementioned works and solve the following optimization problem,

$$\underset{x}{\text{argmin}} \quad \tfrac{1}{2}\|y - x\|_2^2 + h(\|x\|_{(1)}, \ldots, \|x\|_{(\tau)}) \,, \tag{1.2}$$

where $h : \mathbb{R}_+^\tau \to \mathbb{R}_+$ is convex and non-decreasing in each argument (i.e., non-decreasing and strictly increasing in at least one coordinate) and the norms $\|\cdot\|_{(i)}$, $i = 1, \ldots, \tau$, correspond to the τ simultaneous structures that are present in x_0 . However, similar to the results presented in Chapter 2, the reconstruction accuracy can be shown [129] to be bounded from below by the reconstruction accuracy of the *best individual* proximal denoiser. More specifically, in the case of denoising a simultaneously sparse and low-rank matrix, fundamental gaps has been shown between the accuracy of denoising via (1.2) and the number of degrees of freedom for such a simultaneously sparse and low-rank matrix [129].

1.1.1 Previously published material

The content of Chapter 2 is based on previous publications:

- Publication [128]: Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C. Eldar, and Babak Hassibi. Simultaneously Structured Models with Application to Sparse and Low-rank Matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, 2015.

- Publication [129]: Samet Oymak, Amin Jalali, Maryam Fazel, and Babak Hassibi. Noisy Estimation of Simultaneously Structured Models: Limitations of Convex Relaxation, *52nd IEEE Conference on Decision and Control (CDC),* pages 6019–6024, 2013.

1.2 Variational Regularization in Structured Learning

Regularization, the act of simultaneously optimizing a data fidelity term (loss) and a structure-promoting term (regularizer), is a widely used approach in different machine learning and signal processing tasks for estimation of structured models from data. For an appropriate loss function $\mathcal{L}(\cdot)$ and regularizer $\Omega(\cdot)$, the *regularized loss minimization problem* is stated as

$$\min_{x} \quad \mathcal{L}(x) + \lambda\Omega(x) \tag{1.3}$$

where λ trades off the importance of the two terms in the original multi-objective optimization problem. Note that when both of the loss function and the regularizer are convex, the regularized loss minimization problem (1.3) is equivalent to (has the same optimal solutions) minimizing the loss subject a constraint of the form $\Omega(x) \leq \eta$ under mild conditions and for an appropriate value of η; e.g., see [24, Sec. 4.3].

In this work, we introduce a general class of semidefinite representable penalties, called variational Gram functions (VGF), defined as follows. Let $x_1, \ldots, x_m$ be vectors in $\mathbb{R}^n$. It is well known that their pairwise inner products $x_i^T x_j$, for $i, j = 1, \ldots, m$, reveal essential

information about their relative orientations, and can serve as a measure for various properties such as orthogonality. Hence, we can consider a class of functions that aggregate the pairwise inner products in a variational form,

$$\Omega_{\mathcal{M}}(x_1, \ldots, x_m) = \max_{M \in \mathcal{M}} \sum_{i,j=1}^{m} M_{ij} x_i^T x_j \,,$$

where $\mathcal{M}$ is a compact subset of the set of m by m symmetric matrices. Let $X = [x_1 \ \cdots \ x_m]$ be an $n \times m$ matrix. Then the pairwise inner products $x_i^T x_j$ are the entries of the Gram matrix $X^T X$ and the function above can be written as

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \langle X^T X, M \rangle = \max_{M \in \mathcal{M}} \operatorname{tr}(XMX^T) \,,$$

where $\langle A, B \rangle = \operatorname{tr}(A^T B)$ denotes the matrix inner product. We call $\Omega_{\mathcal{M}}$ a *variational Gram function* (VGF) of the vectors $x_1, \ldots, x_m$ induced by the set $\mathcal{M}$. As an example, consider the case where $\mathcal{M}$ is given by a box constraint,

$$\mathcal{M} = \left\{ M : \ |M_{ij}| \leq \overline{M}_{ij}, \ i, j = 1, \ldots, m \right\}, \tag{1.4}$$

where $\overline{M}$ is a symmetric nonnegative matrix. In this case, the maximization in the definition of $\Omega_{\mathcal{M}}$ picks either $M_{ij} = \overline{M}_{ij}$ or $M_{ij} = -\overline{M}_{ij}$ depending on the sign of $x_i^T x_j$, for all $i, j = 1, \ldots, m$ (if $x_i^T x_j = 0$, the choice is arbitrary). Therefore,

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \sum_{i,j=1}^{m} M_{ij} x_i^T x_j = \sum_{i,j=1}^{m} \overline{M}_{ij} |x_i^T x_j| \,.$$

In other words, $\Omega_{\mathcal{M}}(X)$ is the weighted sum of the absolute values of the pairwise inner products. This function was proposed in [169] as a regularization function to promote orthogonality between linear classifiers in the context of hierarchical classification.

We focus on problems where $\mathcal{L}(X)$ is smooth or has an explicit variational structure, and show how to exploit the structure of $\mathcal{L}(X)$ and $\Omega(X)$ together and study their interactions

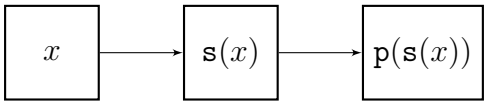


Figure 1.1: Composite penalties. $\mathtt{s}$, often a smooth map, transforms the original structured model x to one with a different structure. $\mathtt{p}$, often a non-smooth real-valued map, penalizes $\mathtt{s}(x)$ according to the latter structure, hence providing a penalization scheme for the original structure.

to provide guarantees on the optimal solutions as well as to derive efficient optimization algorithms. More specifically, we consider a *Fenchel-type representation* for the loss function, as

$$\mathcal{L}(x) = \sup_{g \in \mathcal{G}} \ \langle \mathcal{D}(g), x \rangle - \widehat{\mathcal{L}}(x)\,, \tag{1.5}$$

and *composite penalties* of the form

$$\Omega(x) = \mathtt{p}(\mathtt{s}(x))\,, \tag{1.6}$$

where $\mathcal{D}$ is a linear map encoding the input data, $\mathcal{G}$ is a compact set, $\widehat{\mathcal{L}}$ is a convex function which is inherent to the penalty $\mathcal{L}$, $\mathtt{s}$ is the *structure mapping* (a smooth mapping to change the notion of structure), and, $\mathtt{p}$ is the *outer penalty function*.

Penalties of the form (1.6) arise often in structured learning. Linear composite penalties, corresponding to $\mathtt{s}(x) = Ax$ for some fixed matrix A, include many prominent examples such as the fused lasso [149] (or total variation norm) for smoothing, generalized lasso [150] (including trend filtering [95], wavelet smoothing, etc), OSCAR norm [21] for simultaneous feature selection and grouping, as well as regularizers in [60] for multitask learning, and in [48, 28] for frequency estimation from limited time samples. Variational Gram functions, on the other hand, are derived by using $\mathtt{s}(X) = X^T X$ and are suitable for inducing structures on the relative orientations of columns of X .

For variational Gram functions, with $\mathtt{p}(Y) = \sigma_{\mathcal{M}}(Y) = \sup_{M \in \mathcal{M}} \langle Y, M \rangle$, computing the proximal mapping $\mathrm{prox}_{\mathtt{p}}$ is equivalent to a projection onto the set $\mathcal{M}$ which can be a simple operation for many examples such as the one in (1.4). Hence, in solving (1.3) where Ω is a VGF, instead of using proximal methods through computing the proximal mapping for $\Omega(x) = \mathtt{p}(\mathtt{s}(x))$ which in general amounts to solving a semidefinite program, we consider a reformulation as a convex-concave saddle-point problem

$$\min_x \ \mathcal{L}(x) + \lambda\Omega(x) = \min_x \sup_y \ \mathcal{L}(x) + \lambda\langle y, \mathtt{s}(x)\rangle - \lambda\mathtt{p}^\star(y)\,.$$

This reformulation allows us to rely on the proximal mapping for $\mathtt{p}$ through using primal-dual optimization algorithms. Furthermore, with explicit representations for the loss and the regularizer as in (1.5) and (1.6), the invariance properties of Ω (left unitary invariance for VGFs) can be used to summarize the data term $\mathcal{D}$, hence allowing for possible reductions in the dimensionality of the problem and kernel tricks (see Section 3.5.2).

We motivate the above structural assumptions, on the loss functions and the penalties, by providing examples from the literature as well as new instances. The Fenchel-type representation for a number of common loss functions is provided in Section 3.5. Section 5.2 contains different examples on composite penalties.

1.2.1 Previously published material

The content of Chapter 3 is based on the following work:

- Publication [89]: Amin Jalali, Lin Xiao, and Maryam Fazel. Variational Gram Functions: Convex Analysis and Optimization. *arXiv preprint arXiv:1507.04734,* 2015.

1.3 Community Detection

A fundamental problem in network science and machine learning is to discover structures in large, complex networks (e.g., biological, social, or information networks). Community or cluster detection underlies many decision tasks, as a basic step that uses pairwise relations

between data points in order to understand more global structures in the data. Applications include recommendation systems [163], image segmentation [144, 110], learning gene network structures in bioinformatics, e.g., in protein detection [44] and population genetics [92].

In spite of a long history of heuristic algorithms (see, e.g., [99] for an empirical overview), as well as strong research interest in recent years on the theoretical side as briefly reviewed in the sequel, there are still gaps in understanding the fundamental information theoretic limits of recoverability (i.e., if there is enough information to reveal the communities) and computational tractability (if there are efficient algorithms to recover them). This is particularly true in the case of sparse graphs (that test the limits of recoverability), graphs with heterogeneous communities (communities varying greatly in size and connectivity), graphs with a number of communities that grows with the number of nodes, and partially observed graphs (with various observation models).

In this dissertation, we study recovery regimes and algorithms for community detection in sparse graphs generated under a *heterogeneous stochastic block model* in its most general form. The stochastic block model (SBM), first introduced and studied in mathematical sociology by Holland, Laskey and Leinhardt in 1983 [81], can be described as follows. Consider n vertices partitioned into r *communities* $V_1, V_2, \ldots, V_r$, of sizes $n_1, n_2, \ldots, n_r$. We endow the kth community with an Erdős-Rényi random graph model $\mathcal{G}(n_k, p_k)$ and draw an edge between pairs of nodes in different communities independently with probability q; i.e., the probability of an edge between vertices i and j (denoted by $i \sim j$) is given by

$$\mathbb{P}(i \sim j) = \begin{cases} p_k & \text{if there is a } k \in \{1, 2, \ldots, r\} \text{ such that } i, j \in V_k \\ q & \text{otherwise} \end{cases} \tag{1.7}$$

where we assume $q < \min_k p_k$ in order for the idea of communities to make sense. This defines a distribution over random graphs known as the stochastic block model. In this work, we assume the above model while allowing the number of communities to grow with the number of nodes (similar to [50, 76, 138]). We refer to this model as the *heterogeneous stochastic block model* to contrast our study of this general setting with previous works on

special cases of SBM such as 1) homogenous SBM where the communities are *equivalent* (they are of the same size and the connectivity probabilities are equal,) e.g., as in [47], or, 2) SBM with linear-sized communities, where the number of communities is *fixed* and all community sizes are $\mathrm{O}(n)$; e.g., as in [2].

The community detection problem studied in this work is stated as: given the adjacency matrix of a graph generated by the heterogenous stochastic block model, for what SBM parameters we can recover the labels of *all* vertices, with high probability, using an algorithm that has been proved to do so. We provide guarantees for exact recovery via a semidefinite program in (4.4) as well as upper and (information-theoretic) lower bounds on SBM parameters for exact recoverability. In establishing these performance guarantees, we follow the standard *dual certificate* argument in convex analysis while 1) employing state of the art random matrix theory (e.g., [43, 151, 158, 12]) to strengthen the analysis of standard setups; and 2) tackling the challenge of the general heterogenous SBM and understanding the key descriptors that govern the thresholds. In Section 4.2.3, we extend the above bounds to the case of partial observations, i.e., when each entry of the matrix is observed uniformly with some probability γ and the results are recorded. All of our results only hold with high probability, as this is the best one can hope for; with tiny probability, the model can generate graphs like the complete graph where the partition is unrecoverable.

The results of Chapter 4 provide a clear improvement in the understanding of stochastic block models by exploiting tradeoffs among SBM parameters. We identify a key descriptor (or summary statistic), defined in (4.1) and referred to as *relative density*, which shows up in our results and provides improvements in the statistical assessment and efficient computational approaches for certain configurations of heterogenous SBM; examples are given in in Section 4.3 to illustrate a number of such beneficial tradeoffs such as

- semidefinite programming can successfully recover communities of size $\mathrm{O}(\sqrt{\log n})$ under mild conditions on other communities (see Example 3 of Section 4.3 for details) while $\log n$ has long been believed to be the threshold for the smallest community size.

- The sizes of the communities can be wide spread, or the inter- and intra-community probabilities can be very close, and the model still can be efficiently recoverable, while existing methods (e.g. *peeling strategy* [4]) provide false negatives.

These results are a step towards understanding the information-computational tradeoffs about the heterogenous SBM with a growing number of communities.

1.3.1 Previously published material

The content of Chapter 4 is based on a previous publication:

- Publication [88]: Amin Jalali, Qiyang Han, Ioana Dumitriu, and Maryam Fazel. Exploiting Tradeoffs for Exact Recovery in Heterogeneous Stochastic Block Models. *To appear in the proceedings of the Thirtieth Annual Conference on Neural Information Processing Systems (NIPS).* 2016. (Longer version is available as arXiv preprint arXiv:1512.04937).

Chapter 2

RECOVERY OF SIMULTANEOUSLY STRUCTURED MODELS

Recovering structured models (e.g., sparse or group-sparse vectors, low-rank matrices) given a few linear observations has been well-studied recently. In various applications in signal processing and machine learning, the model of interest exhibits *multiple* structures, for example, a matrix that is simultaneously sparse and low-rank. Often norms that promote the individual structures are known, and allow for recovery using an order-wise optimal number of measurements (e.g., ℓ_1 norm for sparsity, nuclear norm for matrix rank). Hence, it is reasonable to minimize a combination of such norms. We show that, surprisingly, using multi-objective optimization with these norms can do no better, order-wise, than exploiting only one of the structures, thus revealing a fundamental limitation in sample complexity. This result suggests that to fully exploit the multiple structures, we need an entirely new convex relaxation. Further, specializing our results to the case of sparse and low-rank matrices, we show that a nonconvex formulation recovers the model from very few measurements (on the order of the degrees of freedom). In contrast, the convex problem combining the ℓ_1 and nuclear norms requires many more measurements, illustrating a gap between the performance of the convex and nonconvex recovery problems. Our framework applies to arbitrary structure-inducing norms as well as to a wide range of measurement ensembles. This allows us to give sample complexity bounds for problems such as sparse phase retrieval and low-rank tensor completion.

2.1 Introduction

Recovery of a structured model (signal) given a small number of linear observations has been the focus of many studies recently. Examples include recovering sparse or group-

sparse vectors (which gave rise to the area of compressed sensing) [40, 56, 38, 61], low-rank matrices [131, 35], and the sum of sparse and low-rank matrices [41, 32], among others. More generally, the recovery of a signal that can be expressed as the sum of a few atoms out of an appropriate atomic set has been studied in [42]. Canonical questions in this area include: How many linear measurements are enough to recover the model by any means? How many measurements are enough for a tractable approach? In the statistics literature, these questions are often posed in terms of error rates for estimators minimizing the sum of a quadratic loss function and a regularizer that reflects the desired structure [124].

There are many applications where the model of interest is known to have *several* structures at the same time (Section 2.1.2). We then seek a signal that lies in the intersection of several sets defining the individual structures (in a sense that we will later make precise). The most common convex regularizer used to promote all structures together is a linear combination of well-known regularizers for each structure. However, there is currently no general analysis and understanding of how well such regularization performs in terms of the number of observations required for successful recovery of the desired model. This chapter addresses this ubiquitous yet unexplored problem; i.e., the recovery of *simultaneously structured models*.

An example of a simultaneously structured model is a matrix that is *simultaneously sparse and low-rank*. One would like to come up with algorithms that exploit both types of structures to minimize the number of measurements required for recovery. An $n \times n$ matrix with rank $r \ll n$ can be described by $\mathrm{O}(rn)$ parameters, and can be recovered using $\mathrm{O}(rn)$ generic measurements via nuclear norm minimization [131, 34]. On the other hand, a block-sparse matrix with a $k \times k$ nonzero block where $k \ll n$ can be described by k^2 parameters and can be recovered given $\mathrm{O}(k^2 \log \frac{n}{k})$ generic measurements using ℓ_1 minimization. However, a matrix that is *both* rank r and block-sparse can be described by $\mathrm{O}(rk)$ parameters. The question is whether we can exploit this joint structure to efficiently recover such a matrix with $\mathrm{O}(rk)$ measurements.

In this chapter we give a negative answer to this question in the following sense: if we use

multi-objective optimization with the ℓ_1 and nuclear norms (used for sparse signals and low rank matrices, respectively), then the number of measurements required is lower bounded by $\mathrm{O}(\min\{k^2, rn\})$. In other words, we need at least this number of observations for the desired signal to be recoverable by a combination of the ℓ_1 norm and the nuclear norm. This means we can do *no better than an algorithm that exploits only one of the two structures.*

We introduce a framework to express general simultaneously structured models, and as our main result, we prove that the same phenomenon happens for a general set of structures. We analyze a wide range of measurement ensembles, including the subsampled standard basis (i.e. matrix completion), Gaussian and subgaussian measurements, and quadratic measurements. Table 2.1 summarizes known results on recovery of some common structured models, along with a result of this manuscript specialized to the problem of low-rank and sparse matrix recovery. The first column gives the number of parameters needed to describe the model (often referred to as its 'degrees of freedom'), while the second and third columns show how many generic measurements are needed for successful recovery. In 'nonconvex recovery', we assume we are able to find the global minimum of a nonconvex problem. This is clearly intractable in general, and not a practical recovery method—we consider it as a benchmark for theoretical comparison with the (tractable) convex relaxation in order to determine how powerful the relaxation is.

The first and second rows are the results on k sparse vectors in $\mathbb{R}^n$ and rank r matrices in $\mathbb{R}^{n\times n}$ respectively, [37, 34]. The third row considers the recovery of "low-rank plus sparse" matrices. Consider a matrix $X \in \mathbb{R}^{n\times n}$ that can be decomposed as $X = X_L + X_S$ where X_L is a rank r matrix and X_S is a matrix with only k nonzero entries. The degrees of freedom in X are $\mathrm{O}(rn + k)$. Minimizing the combination of the ℓ_1 norm and nuclear norm, i.e., $f(X) = \min_Y \|Y\|_\star + \lambda\|X - Y\|_1$ subject to random Gaussian measurements on X, gives a convex approach for recovering X. It has been shown that under reasonable incoherence assumptions, X can be recovered from $\mathrm{O}((rn+k)\log^2 n)$ measurements which is suboptimal only by a logarithmic factor [162]. Finally, the last row in Table 2.1 shows one of the results in this manuscript. Let $X \in \mathbb{R}^{n\times n}$ be a rank r matrix whose entries are zero outside a

$k_1 \times k_2$ submatrix. The degrees of freedom of X are $\mathrm{O}((k_1 + k_2)r)$. We consider both convex and non-convex programs for the recovery of this type of matrix. The nonconvex method involves minimizing the number of nonzero rows, columns and rank of the matrix jointly, as discussed in Section 2.3.2. As shown later, $\mathrm{O}((k_1 + k_2)r \log n)$ measurements suffice for this program to successfully recover the original matrix. The convex method minimizes any convex combination of the individual structure-inducing norms, namely the nuclear norm and the $\ell_{1,2}$ norm of the matrix, which encourage low-rank and column/row-sparse solutions respectively. We show that with high probability this program cannot recover the original matrix with fewer than $\Omega(rn)$ measurements. In summary, while the nonconvex method is only slightly suboptimal, the convex method performs poorly as the number of measurements scales with n rather than $k_1 + k_2$.

Model	Degrees of Freedom	Nonconvex Recovery	Convex Recovery
Sparse vectors	k	$\mathrm{O}(k)$	$\mathrm{O}(k \log \frac{n}{k})$
Low rank matrices	$r(2n - r)$	$\mathrm{O}(rn)$	$\mathrm{O}(rn)$
Low rank plus sparse	$\mathrm{O}(rn + k)$	not analyzed	$\mathrm{O}((rn + k) \log^2 n)$
Low rank and sparse	$\mathrm{O}(r(k_1 + k_2))$	$\mathrm{O}(r(k_1 + k_2) \log n)$	$\Omega(rn)$

Table 2.1: Summary of results in recovery of structured signals. This work shows a gap between the performance of convex and nonconvex recovery programs for simultaneously structured matrices (last row).

2.1.1 Contributions

This work presents an analysis for the recovery of models with more than one structure, by combining penalties corresponding to each structure. The setting considered is very broad: any number of structures can be combined, any (convex) combination of the corresponding norms can be analyzed, and the results hold for a variety of popular measurement ensembles.

The framework proposed includes special cases that are of interest in their own right, e.g., sparse and low-rank matrix recovery, and low-rank tensor completion [68, 75].

More specifically, our contributions can be summarized as follows.

Poor performance of convex relaxations We consider a model with several structures and associated structure-inducing norms. For recovery, we consider a multi-objective optimization problem to minimize the individual norms simultaneously. Given the convexity of the problem, we know that minimizing a weighted sum of the norms and varying the weights traces out all points of the Pareto-optimal front (Section 2.2). We obtain a lower bound on the number of measurements for any convex function combining the individual norms. A sketch of our main result is as follows.

> *Given a model x_0 with τ simultaneous structures, the number of measurements required for recovery with high probability using any linear combination of the individual norms satisfies the lower bound*
>
> $$m \geq c \min_{i=1,\ldots,\tau} m_i$$
>
> *where m_i is an intrinsic lower bound on the required number of measurements when minimizing the ith norm only. The term c depends on the measurement ensemble.*

For the norms of interest, m_i is proportional to the degrees of freedom of the ith structure. With $\min_i m_i$ as the bottleneck, this result indicates that the combination of norms performs no better than using only one (the best) of the norms, even though the target model has small degrees of freedom.

Different measurement ensembles Our characterization of recovery failure is easy to interpret and deterministic in nature. We show that it can be used to obtain probabilistic failure results for various random measurement ensembles. In particular, our results hold

for measurement matrices with i.i.d. subgaussian rows, quadratic measurements and matrix completion type measurements.

Understanding the effect of weighting We characterize the sample complexity of the multi-objective function as a function of the weights associated with the individual norms. Our upper and lower bounds reveal that the sample complexity of the multi-objective function is related to a certain convex combination of the sample complexities associated with the individual norms. We provide formulas for this combination as a function of the weights.

Incorporating general cone constraints In addition, we can incorporate side information on x_0, expressed as convex cone constraints. This additional information helps in recovery; however, quantifying how much the cone constraints help is not trivial. Our analysis explicitly determines the role of the cone constraint: Geometric properties of the cone such as its Gaussian width determines the constant factors in the bound on the number of measurements.

Illustrating a gap for the recovery and denoising of sparse and low-rank matrices As a special case, we consider the recovery of simultaneously sparse and low-rank matrices and prove that there is a significant gap between the performance of convex and non-convex recovery programs. This gap is surprising when one considers similar results on low-dimensional model recovery as shown in Table 2.1. As mentioned earlier in Section 1.1, we also show the existence of a similar gap when *denoising* a simultaneously sparse and low-rank matrix and the interested reader is referred to [129] for more details.

2.1.2 Applications

We survey several applications where simultaneous structures arise, as well as existing results specific to these applications.

Sparse signal recovery from quadratic measurements Sparsity has long been exploited in signal processing, applied mathematics, statistics and computer science for tasks such as compression, denoising, model selection, image processing and more. Despite the great interest in exploiting sparsity in various applications, most of the work to date has focused on recovering sparse or low rank data from linear measurements. Recently, the basic sparse recovery problem has been generalized to the case in which the measurements are given by nonlinear transforms of the unknown input, [14]. A special case of this more general setting is quadratic compressed sensing [143] in which the goal is to recover a sparse vector x from quadratic measurements $b_i = x^T A_i x$. This problem can be linearized by *lifting*, where we wish to recover a "low rank and sparse" matrix $X = xx^T$ subject to measurements $b_i = \langle A_i, X \rangle$.

Sparse recovery problems from quadratic measurements arise in a variety of problems in optics. One example is sub-wavelength optical imaging [148, 143] in which the goal is to recover a sparse image from its far-field measurements, where due to the laws of physics the relationship between the (clean) measurement and the unknown image is quadratic. In [143] the quadratic relationship is a result of using partially-incoherent light. The quadratic behavior of the measurements in [148] arises from coherent diffractive imaging in which the image is recovered from its intensity pattern. Under an appropriate experimental setup, this problem amounts to reconstruction of a sparse signal from the magnitude of its Fourier transform.

A related and notable problem involving sparse and low-rank matrices is Sparse Principal Component Analysis (SPCA), mentioned in Section 5.1.

Sparse phase retrieval Quadratic measurements appear in phase retrieval problems, in which a signal is to be recovered from the magnitude of its measurements $b_i = |a_i^T x|$, where each measurement is a linear transform of the input $x \in \mathbb{R}^n$ and a_i's are arbitrary, possibly complex-valued measurement vectors. An important case is when $a_i^T x$ is the Fourier transform and b_i^2 is the signal's power. Phase retrieval is of great interest in many applications

such as optical imaging [159, 112], crystallography [78], and more [83, 70, 66].

The problem becomes linear when x is *lifted* and we consider the recovery of $X = xx^T$ where each measurement takes the form $b_i^2 = \langle a_i a_i^T, X \rangle$. In [143], an algorithm was developed to treat phase retrieval problems with sparse x based on a semidefinite relaxation, and low-rank matrix recovery combined with a row-sparsity constraint on the resulting matrix. More recent works also proposed the use of semidefinite relaxation together with sparsity constraints for phase retrieval [127, 102, 86, 104]. An alternative algorithm was recently designed in [142] based on a greedy search. In [86], the authors consider sparse signal recovery based on combinatorial and probabilistic approaches and provide uniqueness results under certain conditions. Stable uniqueness in phase retrieval problems is studied in [62]. The results of [31, 39] applies to general (non-sparse) signals where in some cases *masked* versions of the signal are required.

Fused lasso Suppose the signal of interest x_0 is sparse and its entries vary slowly, i.e., the signal can be approximated by a piecewise constant function. To encourage sparsity, one can use the ℓ_1 norm. To promote the piecewise constant structure, discrete total variation can be used, defined as

$$\|x\|_{\mathrm{TV}} = \sum_{i=1}^{n-1} |x_{i+1} - x_i| \,,$$

where $\|\cdot\|_{\mathrm{TV}}$ is basically the ℓ_1 norm of the gradient and is approximately sparse. The resulting optimization problem that estimates x_0 from its samples Ax_0 is known as fused-lasso [149], and is given as

$$\min_x \quad \|x\|_1 + \lambda \|x\|_{\mathrm{TV}} \quad \text{s.t.} \quad Ax = Ax_0. \tag{2.1}$$

To the best of our knowledge, the sample complexity of fused lasso has not been analyzed from a compressed sensing point of view. However, there is a series of recent works [123, 122] on total variation minimization; which may lead to analysis of (2.1).

We remark that TV regularization is also used together with nuclear norm to encourage a low-rank and smooth (i.e., slowly varying entries) solution. This regularization finds applications in imaging and physics [141, 72].

Low-rank tensors Tensors with small Tucker rank can be seen as a generalization of low-rank matrices [154]. In this setup, the signal of interest is a tensor $X_0 \in \mathbb{R}^{n_1 \times \cdots \times n_\tau}$, where X_0 is low-rank along its unfoldings which are obtained by reshaping X_0 as a matrix with size $n_i \times \frac{n}{n_i}$, with $n = \prod_{i=1}^{\tau} n_i$. Denoting the ith unfolding by $\mathcal{U}_i(X_0)$, a standard approach to estimate X_0 from $y = \mathcal{A}(X_0)$ is minimizing the weighted nuclear norms of the unfoldings,

$$\min_{X} \ \sum_{i=1}^{\tau} \lambda_i \| \mathcal{U}_i(X) \|_\star \quad \text{subject to} \quad y = \mathcal{A}(X_0). \tag{2.2}$$

Low-rank tensors have applications in machine learning, physics, computational finance and high dimensional PDE's [75]. The problem (2.2) has been investigated in several papers [103, 68]. Closer to us, [120] recently showed that the convex relaxation (2.2) performs poorly compared to information theoretically optimal bounds for Gaussian measurements. Our results can extend those to the more applicable tensor completion setup, where we observe the entries of the tensor.

Other applications of simultaneously structured signals include Collaborative Hierarchical Sparse Modeling [145] where sparsity is considered within the non-zero blocks in a block-sparse vector, and the recovery of hyperspectral images where we aim to recover a simultaneously block sparse and low rank matrix from compressed observations [71].

2.1.3 Outline of the Chapter

This chapter is structured as follows. Background and definitions are given in Section 2.2. An overview of the main results is provided in Section 2.3. Section 2.4 discusses some measurement ensembles for which our results apply. Section 2.5 derives upper bounds for the convex relaxations assuming a Gaussian measurement ensemble. Proofs of the general results are presented in Section 2.6. The proofs for the special case of simultaneously sparse

and low-rank matrices are given in Section A.1, where we compare corollaries of the general results with the results on non-convex recovery approaches, and illustrate a gap. Numerical simulations in Section 2.7 empirically support the results on sparse and low-rank matrices. A discussion on the results is presented in Section 5.1.

2.2 Problem Setup

We begin by reviewing some basic definitions. Our results will be on structure-inducing norms; examples include the ℓ_1 norm, the $\ell_{1,2}$ norm, and the nuclear norm.

2.2.1 Definitions and Signal Model

The nuclear norm of a matrix is denoted by $\|\cdot\|_\star$ and is the sum of the singular values of the matrix. The $\ell_{1,2}$ norm is the sum of the ℓ_2 norms of the columns of a matrix. Minimizing the ℓ_1 norm encourages sparse solutions, while the $\ell_{1,2}$ norm and nuclear norm encourage column-sparse and low-rank solutions respectively, [131, 35, 146, 164, 63]; see Section 2.6.4 for more detailed discussion of these norms and their subdifferentials. The Euclidean norm is denoted by $\|\cdot\|_2$, i.e., the ℓ_2 norm for vectors and the Frobenius norm $\|\cdot\|_F$ for matrices.

Overlines denote normalization, i.e., for a vector x and a matrix X, $\bar{x} = \frac{x}{\|x\|_2}$ and $\bar{X} = \frac{X}{\|X\|_F}$. The minimum and maximum singular values of a matrix A are denoted by $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$. The set of $n \times n$ positive semidefinite (PSD) and symmetric matrices are denoted by $\mathbb{S}^n_+$ and $\mathbb{S}^n$ respectively. cone(S) denotes the conic hull of a given set S. $\mathcal{A}(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ is a linear measurement operator if $\mathcal{A}(x)$ is equivalent to the matrix multiplication Ax where $A \in \mathbb{R}^{m\times n}$. If x is a matrix, $\mathcal{A}(x)$ will be a matrix multiplication with a suitably vectorized x. In some of our results, we consider Gaussian measurements, in which case A has independent $\mathcal{N}(0,1)$ entries.

For a vector $x \in \mathbb{R}^n$, $\|x\|$ denotes a general norm and $\|x\|^\star = \sup_{\|z\|\leq 1}\langle x, z\rangle$ is the corresponding dual norm. A subgradient of the norm $\|\cdot\|$ at x is a vector g for which $\|z\| \geq \|x\| + \langle g, z - x\rangle$ holds for any z. The set of all subgradients is called the subdifferential

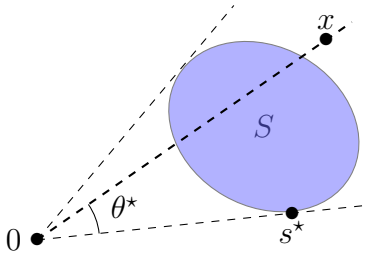


Figure 2.1: Depiction of the correlation between a vector x and a set S. $s^\star$ achieves the largest angle with x, hence $s^\star$ has the minimum correlation with x; i.e., $\rho(x,S) = |\bar{x}^T \bar{s}^\star|$.

and is denoted by $\partial\|x\|$. The Lipschitz constant of the norm is defined as

$$L = \sup_{z_1 \neq z_2 \in \mathbb{R}^n} \frac{\|z_1\| - \|z_2\|}{\|z_1 - z_2\|_2}.$$

Definition 1 (Correlation) *Given a nonzero vector x and a set S, $\rho(x,S)$ is defined as*

$$\rho(x,S) := \inf_{0 \neq s \in S} \frac{|x^T s|}{\|x\|_2 \|s\|_2}.$$

$\rho(x,S)$ corresponds to the minimum absolute-valued correlation between the vector x and elements of S. Let $\bar{x} = \frac{x}{\|x\|_2}$. The correlation between x and the associated subdifferential has a simple form:

$$\rho(x, \partial\|x\|) = \inf_{g \in \partial\|x\|} \frac{\bar{x}^T g}{\|g\|_2} = \frac{\|\bar{x}\|}{\sup_{g \in \partial\|x\|} \|g\|_2}.$$

Here, we used the fact that, for norms, subgradients $g \in \partial\|x\|$ satisfy $x^T g = \|x\|$, [160]. The denominator of the right hand side is the local Lipschitz constant of $\|\cdot\|$ at x and is upper bounded by L. Consequently, $\rho(x, \partial\|x\|) \geq \frac{\|\bar{x}\|}{L}$. We will denote $\frac{\|\bar{x}\|}{L}$ by κ. Recently, this quantity has been studied by Mu et al. to analyze the simultaneously structured signals in a similar spirit to us for Gaussian measurements [120][1]. Similar calculations as above gives an alternative interpretation for κ which is illustrated in Figure 2.2.

[1]The work [120] was submitted after the initial version of [128]; in which we projected the subdifferential onto a carefully chosen subspace to obtain bounds on the sample complexity (see Proposition 27). Inspired by [120], projection onto x_0 and the use of κ led to the simplification of the notation and improvement of the results in the current manuscript, in particular, Section 2.4.

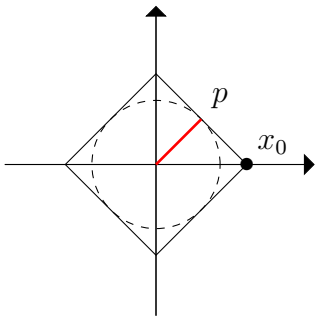


Figure 2.2: For a scaled norm ball passing through x_0, $\kappa = \frac{\|p\|_2}{\|x_0\|_2}$, where p is any of the closest points on the scaled norm ball to the origin.

κ is a measure of alignment between the vector x and the subdifferential. For the norms of interest, it is associated with the model complexity. For instance, for a k-sparse vector x, $\|\bar{x}\|_1$ lies between 1 and $\sqrt{k}$ depending on how spiky nonzero entries are. Also the Lipschitz constant for the ℓ_1 norm in $\mathbb{R}^n$ is $L = \sqrt{n}$. Hence, when the nonzero entries are ± 1, we find $\kappa^2 = \frac{k}{n}$. Similarly, given a $d \times d$, rank r matrix X, $\|\bar{X}\|_\star$ lies between 1 and $\sqrt{r}$. If the singular values are spread (i.e. ± 1), we find $\kappa^2 = \frac{r}{d} = \frac{rd}{d^2}$. In these cases, κ^2 is proportional to the model complexity normalized by the ambient dimension.

Simultaneously structured models We consider a signal x_0 which has several low-dimensional structures S_1, S_2, ..., S_τ (e.g., sparsity, group sparsity, low-rank). Suppose each structure i corresponds to a norm denoted by $\|\cdot\|_{(i)}$ which promotes that structure (e.g., ℓ_1, $\ell_{1,2}$, nuclear norm). We refer to such an x_0 as a *simultaneously structured model*.

2.2.2 Convex Recovery Program

We investigate the recovery of the simultaneously structured x_0 from its linear measurements $\mathcal{A}(x_0)$. To recover x_0, we would like to simultaneously minimize the norms $\|\cdot\|_{(i)}$, $i = 1, \ldots, \tau$, which leads to a multi-objective (vector-valued) optimization problem. For all feasible points x satisfying $\mathcal{A}(x) = \mathcal{A}(x_0)$ and side information $x \in \mathcal{C}$, consider the set of achievable norms

$\big(\|x\|_{(1)}, \ldots, \|x\|_{(\tau)}\big)$ denoted as points in $\mathbb{R}^\tau$. The minimal points of this set with respect to the positive orthant $\mathbb{R}^\tau_+$ form the *Pareto-optimal* front, as illustrated in Figure 2.3. Since the problem is convex, one can alternatively consider the set

$$\{\mathbf{v} \in \mathbb{R}^\tau : \exists x \in \mathbb{R}^n \text{ such that } x \in \mathcal{C},\, \mathcal{A}(x) = \mathcal{A}(x_0),\, v_i \geq \|x\|_{(i)},\, \text{for } i = 1, \ldots, \tau\},$$

which is convex and has the same Pareto optimal points as the original set (see, e.g., [25, Chapter 4]).

Definition 2 (Recoverability) *We call x_0 recoverable if it is a Pareto optimal point; i.e., there does not exist a feasible $x' \neq x$ satisfying $\mathcal{A}(x') = \mathcal{A}(x_0)$ and $x' \in \mathcal{C}$, with $\|x'\|_{(i)} \leq \|x_0\|_{(i)}$ for $i = 1, \ldots, \tau$.*

The vector-valued convex recovery program can be turned into a scalar optimization problem as

$$\begin{array}{ll} \underset{x \in \mathcal{C}}{\text{minimize}} & f(x) = h(\|x\|_{(1)}, \ldots, \|x\|_{(\tau)}) \\ \text{subject to} & \mathcal{A}(x) = \mathcal{A}(x_0), \end{array} \tag{2.3}$$

where $h : \mathbb{R}^\tau_+ \to \mathbb{R}_+$ is convex and non-decreasing in each argument (i.e., non-decreasing and strictly increasing in at least one coordinate). For convex problems with strong duality, it is known that we can recover all of the Pareto optimal points by optimizing weighted sums $f(x) = \sum_{i=1}^\tau \lambda_i \|x\|_{(i)}$, with positive weights λ_i , among all possible functions $f(x) = h(\|x\|_{(1)}, \ldots, \|x\|_{(\tau)})$. For each x_0 on the Pareto, the coefficients of such a recovering function are given by the hyperplane supporting the Pareto at x_0 [25, Chapter 4].

In Figure 2.3, consider the smallest m that makes x_0 recoverable. Then one can choose a function h and recover x_0 by (2.3) using the m measurements. If the number of measurements is any less, then *no* function can recover x_0. Our goal is to provide lower bounds on m.

In [42], Chandrasekaran et al. propose a general theory for constructing a suitable penalty, called an *atomic norm*, given a single set of atoms that describes the structure of the target object. In the case of simultaneous structures, this construction requires defining new atoms,

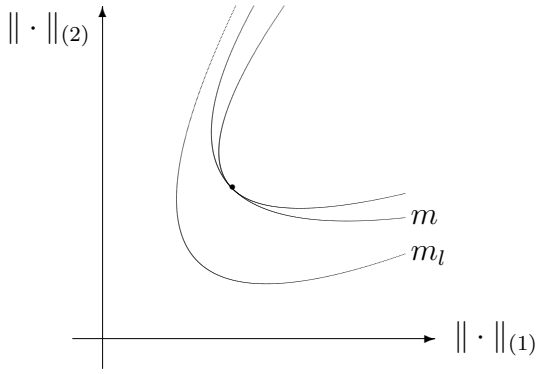


Figure 2.3: Consider a point x_0 represented in the figure by the dot. We need at least m measurements for x_0 to be recoverable since for any $m_l < m$ this point is not on the Pareto optimal front.

and then ensuring the resulting atomic norm can be minimized in a computationally tractable way, which is nontrivial and often intractable. We briefly discuss such constructions as a future research direction in Section 5.1.

2.3 Main Results

In this section, we state our main theorems that aim to characterize the number of measurements needed to recover a simultaneously structured signal by convex or nonconvex programs. We first present our general results, followed by results for simultaneously sparse and low-rank matrices as a specific but important instance of the general case. The proofs are given in Sections 2.6 and A.1. All of our statements will implicitly assume $x_0 \neq 0$. This will ensure that x_0 is not a trivial minimizer and 0 is not in the subdifferentials.

2.3.1 General Simultaneously Structured Models

Consider the recovery of a signal x_0 that is simultaneously structured with $S_1, S_2, \ldots, S_\tau$ as described in Section 2.2.1. We provide a lower bound on the required number of measurements, using the geometric properties of the individual norms.

Theorem 3 (Deterministic failure) *Suppose $\mathcal{C} = \mathbb{R}^n$ and,*

$$\rho(x_0, \partial f(x_0)) := \inf_{g \in \partial f(x_0)} |\bar{g}^T \bar{x}_0| > \frac{\|A\bar{x}_0\|_2}{\sigma_{\min}(A^T)} \,. \tag{2.4}$$

Then, x_0 is not a minimizer of (2.3).

Theorem 3 is deterministic in nature. However, it can be easily specialized to specific random measurement ensembles. The left hand side of (2.4) depends only on the vector x_0 and the subdifferential $\partial f(x_0)$, hence it is independent of the measurement matrix A. For simultaneously structured models, we will argue that, the left hand side cannot be made too small, as the subgradients are *aligned* with the signal. On the other hand, the right hand side depends only on A and x_0 and is independent of the subdifferential. In linear inverse problems, A is often assumed to be random. For large class of random matrices, we will argue that, the right hand side is approximately $\sim \sqrt{\frac{m}{n}}$ which will yield a lower bound on the number of required measurements.

Typical measurement ensembles include the following:

- **Sampling entries:** In low-rank matrix and tensor completion problems, we observe the entries of x_0 uniformly at random. In this case, rows of A are chosen from the standard basis in $\mathbb{R}^n$. We should remark that, instead of the standard basis, one can consider other orthonormal bases such as the Fourier basis.

- **Matrices with i.i.d. rows:** A has independent and identically distributed rows with certain moment conditions. This is a widely used setup in compressed sensing as each measurement we make is associated with the corresponding row of A [33].

- **Quadratic measurements:** Arises in the phase retrieval problem as discussed in Section 2.1.2.

In Section 2.4, we find upper bounds on the right hand side of (2.4) for these ensembles. As discussed in Section 2.4, we can modify the rows of A to get better bounds as long as this does not affect its null space. For instance, one can discard the identical rows to improve conditioning. However, as m increases and A has more linearly independent rows, $\sigma_{\min}(A^T)$ will naturally decrease and (2.4) will no longer hold after a certain point. In particular, (2.4) cannot hold beyond $m \geq n$ as $\sigma_{\min}(A^T) = 0$. This is indeed natural as the system becomes overdetermined.

The following proposition lower bounds the left hand side of (2.4) in an interpretable manner. In particular, the correlation $\rho(x_0, \partial f(x_0))$ can be lower bounded by the smallest individual correlation.

Proposition 4 *Let L_i be the Lipschitz constant of the ith norm and $\kappa_i = \frac{\|\bar{x}_0\|_{(i)}}{L_i}$ for $1 \leq i \leq \tau$. Set $\kappa_{\min} = \min\{\kappa_i : \ i = 1, \ldots, \tau\}$. Then:*

- *All functions $f(\cdot)$ in (2.3) satisfy, $\rho(x_0, \partial f(x_0)) \geq {\kappa_{\min}}^2$.*

- *Suppose $f(\cdot)$ is a weighted linear combination $f(x) = \sum_{i=1}^{\tau} \lambda_i \|x\|_{(i)}$ for nonnegative $\{\lambda_i\}_{i=1}^{\tau}$. Let $\bar{\lambda}_i = \frac{\lambda_i L_i}{\sum_{i=1}^{\tau} \lambda_i L_i}$ for $1 \leq i \leq \tau$. Then, $\rho(x_0, \partial f(x_0)) \geq \sum_{i=1}^{\tau} \bar{\lambda}_i \kappa_i$.*

Proof. From Lemma 26, any subgradient of $f(\cdot)$ can be written as, $g = \sum_{i=1}^{\tau} w_i g_i$ for some nonnegative w_i's. On the other hand, from [160], $\langle \bar{x}_0, g_i \rangle = \|\bar{x}_0\|_{(i)}$. Combining these results,

$$g^T \bar{x}_0 = \sum_{i=1}^{\tau} w_i \|\bar{x}_0\|_{(i)}.$$

From the triangle inequality, $\|g\|_2 \leq \sum_{i=1}^{\tau} w_i L_i$. Therefore,

$$\frac{\sum_{i=1}^{\tau} w_i \|\bar{x}_0\|_{(i)}}{\sum_{i=1}^{\tau} w_i L_i} \ \geq \ \min_{1 \leq i \leq \tau} \frac{w_i \|\bar{x}_0\|_{(i)}}{w_i L_i} = \kappa_{\min} \, . \tag{2.5}$$

[2]The lower bound $\kappa_{\min}$ is directly comparable to Theorem 5 of [120]. Indeed, our lower bounds on the sample complexity will have the form $\mathrm{O}(\kappa_{\min}^2 n)$.

Model	$f(\cdot)$	L	$\|\bar{x}_0\| \leq$	$n\kappa^2 \leq$
k sparse vector	$\|\cdot\|_1$	$\sqrt{n}$	$\sqrt{k}$	k
k column-sparse matrix	$\|\cdot\|_{1,2}$	$\sqrt{d}$	$\sqrt{k}$	kd
Rank r matrix	$\|\cdot\|_\star$	$\sqrt{d}$	$\sqrt{r}$	rd
S&L (k,k,r) matrix	$h(\|\cdot\|_\star, \|\cdot\|_1)$	$-$	$-$	$\min\{k^2, rd\}$

Table 2.2: Summary of the parameters that are discussed in this section. The last three lines are for a $d \times d$ S&L (k,k,r) matrix where $n = d^2$. In the fourth column, the corresponding entry for S&L is $\kappa_{\min} = \min\{\kappa_{\ell_1}, \kappa_\star\}$.

To prove the second part, we use the fact that for the weighted sums of norms, $w_i = \lambda_i$ and the subgradients have the form $g = \sum_{i=1}^{\tau} \lambda_i g_i$, [25]. Then, substitute $\bar{\lambda}_i$ for λ_i in the left hand side of (2.5). ■

Before stating the next result, let us give a relevant definition regarding the average distance between a set and a random vector.

Definition 5 (Gaussian distance) *Let $\mathcal{M}$ be a closed convex set in $\mathbb{R}^n$ and let $\mathbf{h} \in \mathbb{R}^n$ be a vector with independent standard normal entries. Then, the Gaussian distance of $\mathcal{M}$ is defined as*

$$\mathbf{D}(\mathcal{M}) = \mathbb{E} \inf_{\mathbf{v} \in \mathcal{M}} \|\mathbf{h} - \mathbf{v}\|_2$$

When $\mathcal{M}$ is a cone, we have $0 \leq \mathbf{D}(\mathcal{M}) \leq \sqrt{n}$. Similar definitions have been used extensively in the literature, such as Gaussian width [42], statistical dimension [6] and mean width [156]. For notational simplicity, let the normalized distance be $\bar{\mathbf{D}}(\mathcal{M}) = \frac{\mathbf{D}(\mathcal{M})}{\sqrt{n}}$.

We will now state our result for Gaussian measurements; which can additionally include cone constraints for the lower bound. Other ensembles are considered in Section 2.4.

Theorem 6 (Gaussian lower bound) *Suppose A has independent $\mathcal{N}(0,1)$ entries. Whenever $m \leq m_{\text{low}}$, x_0 will not be a minimizer of any of the recovery programs in (2.3) with probability at least $1 - 10\exp\big(-\frac{1}{16}\min\{m_{\text{low}}, (1-\bar{\mathbf{D}}(\mathcal{C}))^2 n\}\big)$, where*

$$m_{\text{low}} \;:=\; \frac{(1-\bar{\mathbf{D}}(\mathcal{C}))\, n\, \kappa_{\min}^2}{100}\,.$$

Remark 7 *When $\mathcal{C} = \mathbb{R}^n$, $\bar{\mathbf{D}}(\mathcal{C}) = 0$ hence, the lower bound simplifies to $m_{\text{low}} = \frac{n\kappa_{\min}^2}{100}$.*

Note that $\bar{\mathbf{D}}(\mathcal{C})$ depends only on $\mathcal{C}$ and can be viewed as a constant. For instance, for the positive semidefinite cone, we show that $\bar{\mathbf{D}}(\mathbb{S}_+^n) < \frac{\sqrt{3}}{2}$. Observe that for a smaller cone $\mathcal{C}$, it is reasonable to expect a smaller lower bound on the required number of measurements. Indeed, as $\mathcal{C}$ gets smaller, $\mathbf{D}(\mathcal{C})$ increases.

As discussed before, there are various options for the scalarizing function in (2.3), with one choice being the weighted sum of norms. In fact, for a recoverable point x_0 there always exists a weighted sum of norms which recovers it. This function is also often the choice in applications, where the space of positive weights is searched for a good combination. Thus, we can state the following corollary as a general result.

Corollary 8 (Weighted lower bound) *Suppose A has i.i.d. $\mathcal{N}(0,1)$ entries and $f(x) = \sum_{i=1}^{\tau} \lambda_i \|x\|_{(i)}$ for nonnegative weights $\{\lambda_i\}_{i=1}^{\tau}$. Whenever $m \leq m'_{\text{low}}$, x_0 will not be a minimizer of the recovery program (2.3) with probability at least $1 - 10\exp\big(-\frac{1}{16}\min\{m'_{\text{low}}, (1-\bar{\mathbf{D}}(\mathcal{C}))^2 n\}\big)$, where*

$$m'_{\text{low}} \;:=\; \frac{n(1-\bar{\mathbf{D}}(\mathcal{C}))(\sum_{i=1}^{\tau} \bar{\lambda}_i \kappa_i)^2}{100}\,,$$

and $\bar{\lambda}_i = \frac{\lambda_i L_i}{\sum_{i=1}^{\tau} \lambda_i L_i}$.

Observe that Theorem 6 is stronger than stating "a particular function $h(\|x\|_{(1)}, \ldots, \|x\|_{(\tau)})$ will not work". Instead, our result states that with high probability none of the programs in the class (2.3) can return x_0 as optimal unless the number of measurements is sufficiently large.

To understand the result better, note that the required number of measurements is proportional to $\kappa_{min}^2 n$ which is often proportional to the sample complexity of the best individual norm. As we have argued in Section 2.2.1, $\kappa_i^2 n$ corresponds to how structured the signal is. For sparse signals it is equal to the sparsity, and for a rank r matrix, it is equal to the degrees of freedom of the set of rank r matrices. Consequently, Theorem 6 suggests that even if the signal satisfies multiple structures, the required number of measurements is effectively determined by only one dominant structure.

Intuitively, the degrees of freedom of a simultaneously structured signal can be much lower, which is provable for simultaneously sparse and low-rank (S&L) matrices. Hence, there is a considerable gap between the expected measurements based on model complexity and the number of measurements needed for recovery via (2.3) ($\kappa_{\min}^2 n$).

2.3.2 Simultaneously Sparse and Low-rank Matrices

We now focus on a special case, namely simultaneously sparse and low-rank (S&L) matrices. We consider matrices with nonzero entries contained in a small submatrix where the submatrix itself is low rank. Here, norms of interest are $\|\cdot\|_{1,2}$, $\|\cdot\|_1$ and $\|\cdot\|_\star$ and the cone of interest is the PSD cone. We also consider nonconvex approaches and contrast the results with convex approaches. For the nonconvex problem, we replace the norms $\|\cdot\|_1, \|\cdot\|_{1,2}, \|\cdot\|_\star$ with the functions $\|\cdot\|_0, \|\cdot\|_{0,2}, \text{rank}(\cdot)$ which give the number of nonzero entries, the number of nonzero columns and rank of a matrix respectively and use the same cone constraint as the convex method. We show that convex methods perform poorly as predicted by the general result in Theorem 6, while nonconvex methods require optimal number of measurements (up to a logarithmic factor). Proofs are given in Section A.1.

Definition 9 *We say $X_0 \in \mathbb{R}^{d_1 \times d_2}$ is an S&L matrix with (k_1, k_2, r) if the smallest submatrix that contains nonzero entries of X_0 has size $k_1 \times k_2$ and $\text{rank}(X_0) = r$. When X_0 is symmetric, let $d = d_1 = d_2$ and $k = k_1 = k_2$. We consider the following cases.*

(a) General: *$X_0 \in \mathbb{R}^{d_1 \times d_2}$ is S&L with (k_1, k_2, r).*

(b) PSD model: *$X_0 \in \mathbb{R}^{n\times n}$ is PSD and S&L with (k,k,r).*

We are interested in S&L matrices with $k_1 \ll d_1, k_2 \ll d_2$ so that the matrix is sparse, and $r \ll \min\{k_1,k_2\}$ so that the submatrix containing the nonzero entries is low rank. Recall from Section 2.2.2 that our goal is to recover X_0 from linear observations $\mathcal{A}(X_0)$ via convex or nonconvex optimization. The measurements can be equivalently written as $A\text{vec}(X_0)$, where $A \in \mathbb{R}^{m\times d_1 d_2}$ and $\text{vec}(X_0) \in \mathbb{R}^{d_1 d_2}$ denotes the vector obtained by stacking the columns of X_0.

Based on the results in Section 2.3.1, we obtain lower bounds on the number of measurements for convex recovery. We additionally show that significantly fewer measurements are sufficient for non-convex programs to uniquely recover X_0; thus proving a performance gap between convex and nonconvex approaches. The following theorem summarizes the results.

Theorem 10 (Performance of S&L matrix recovery) *Suppose $\mathcal{A}(\cdot)$ is an i.i.d. Gaussian map and consider recovering $X_0 \in \mathbb{R}^{d_1\times d_2}$ via*

$$\underset{X\in\mathcal{C}}{\text{minimize}}\ f(X) \quad \text{subject to} \quad \mathcal{A}(X) = \mathcal{A}(X_0). \tag{2.6}$$

For the cases given in Definition 9, the following convex and nonconvex recovery results hold for some positive constants c_1, c_2.

(a) General model:

(a1) Let $f(X) = \|X\|_{1,2} + \lambda_1\|X^T\|_{1,2} + \lambda_2\|X\|_\star$ where $\lambda_1,\lambda_2 \geq 0$ and $\mathcal{C} = \mathbb{R}^{d_1\times d_2}$. Then, (2.6) will fail to recover X_0 with probability $1-\exp(-c_1 m_0)$ whenever $m \leq c_2 m_0$ where $m_0 = \min\{d_1k_2, d_2k_1, (d_1+d_2)r\}$.

(a2) Let $f(X) = \frac{1}{k_2}\|X\|_{0,2} + \frac{1}{k_1}\|X^T\|_{0,2} + \frac{1}{r}\text{rank}(X)$ and $\mathcal{C} = \mathbb{R}^{d_1\times d_2}$. Then, (2.6) will uniquely recover X_0 with probability $1-\exp(-c_1 m)$ whenever $m \geq c_2 \max\{(k_1+k_2)r, k_1\log\frac{d_1}{k_1}, k_2\log\frac{d_2}{k_2}\}$.

(b) PSD with $\ell_{1,2}$:

Setting	Nonconvex Sufficient m	Convex Required m
General model	$\mathrm{O}(\max\{rk, k\log\frac{d}{k}\})$	$\Omega(rd)$
PSD with $\ell_{1,2}$	$\mathrm{O}(\max\{rk, k\log\frac{d}{k}\})$	$\Omega(rd)$
PSD with ℓ_1	$\mathrm{O}(k\log\frac{d}{k})$	$\Omega(\min\{k^2, rd\})$

Table 2.3: Summary of recovery results for models in Definition 9, assuming $d_1 = d_2 = d$ and $k_1 = k_2 = k$. For the 'PSD with ℓ_1' case, we assume $\frac{1}{k}\|\bar{X}_0\|_1$ and $\frac{1}{\sqrt{r}}\|\bar{X}_0\|_\star$ to be approximately constant for the sake of simplicity. Nonconvex approaches are optimal up to a logarithmic factor, while convex approaches perform poorly.

(b1) Let $f(X) = \|X\|_{1,2} + \lambda\|X\|_\star$ *where* $\lambda \geq 0$ *and* $\mathcal{C} = \mathbb{S}^d_+$. *Then,* (2.6) *will fail to recover* X_0 *with probability* $1 - \exp(-c_1 rd)$ *whenever* $m \leq c_2 rd$.

(b2) Let $f(X) = \frac{2}{k}\|X\|_{0,2} + \frac{1}{r}\mathrm{rank}(X)$ *and* $\mathcal{C} = \mathbb{S}^d$. *Then,* (2.6) *will uniquely recover* X_0 *with probability* $1 - \exp(-c_1 m)$ *whenever* $m \geq c_2 \max\{rk, k\log\frac{d}{k}\}$.

(c) PSD with ℓ_1*:*

(c1) Let $f(X) = \|X\|_1 + \lambda\|X\|_\star$ *and* $\mathcal{C} = \mathbb{S}^d_+$. *Then,* (2.6) *will fail to recover* X_0 *with probability* $1 - \exp(-c_1 m_0)$ *for all possible* $\lambda \geq 0$ *whenever* $m \leq c_2 m_0$ *where* $m_0 = \min\{\|\bar{X}_0\|_1^2, \|\bar{X}_0\|_\star^2 d\}$.

(c2) Suppose $\mathrm{rank}(X_0) = 1$. *Let* $f(X) = \frac{1}{k^2}\|X\|_0 + \mathrm{rank}(X)$ *and* $\mathcal{C} = \mathbb{S}^d$. *Then,* (2.6) *will uniquely recover* X_0 *with probability* $1 - \exp(-c_1 m)$ *whenever* $m \geq c_2 k\log\frac{d}{k}$.

Remark 11 (case of PSD with ℓ_1) *In the special case,* $X_0 = \mathbf{a}\mathbf{a}^T$ *for a* k*-sparse vector* $\mathbf{a}$*, we have* $m_0 = \min\{\|\bar{\mathbf{a}}\|_1^4, d\}$. *When nonzero entries of* $\mathbf{a}$ *are* ± 1*, we have* $m_0 = \min\{k^2, d\}$.

The nonconvex programs require almost the same number of measurements as the degrees of freedom (or number of parameters) of the underlying model. For instance, it is known

that the degrees of freedom of a rank r matrix of size $k_1 \times k_2$ is simply $r(k_1 + k_2 - r)$ which is $\mathrm{O}((k_1 + k_2)r)$. Hence, the nonconvex results are optimal up to a logarithmic factor. On the other hand, our results on the convex programs that follow from Theorem 6 show that the required number of measurements are significantly larger. Table 2.3 provides a quick comparison of the results on S&L.

For the S&L (k, k, r) model, from standard results one can easily deduce that [131, 146, 38],

- ℓ_1 penalty only: requires at least k^2 measurements,
- $\ell_{1,2}$ penalty only: requires at least kd measurements,
- Nuclear norm penalty only: requires at least rd measurements.

These follow from the model complexity of the sparse, column-sparse and low-rank matrices. Theorem 6 shows that, combination of norms require at least as much as the best individual norm. For instance, combination of ℓ_1 and the nuclear norm penalization yields the lower bound $\mathrm{O}(\min\{k^2, rd\})$ for S&L matrices whose singular values and nonzero entries are spread. This is indeed what we would expect from the interpretation that $\kappa^2 n$ is often proportional to the sample complexity of the corresponding norm and, the lower bound $\kappa^2_{\min} n$ is proportional to that of the best individual norm.

As we saw in Section 2.3.1, adding a cone constraint to the recovery program does not help in reducing the lower bound by more than a constant factor. In particular, we discuss the positive semidefiniteness assumption that is beneficial in the sparse phase retrieval problem and show that the number of measurements remain high even when we include this extra information. On the other hand, the nonconvex recovery programs performs well even without the PSD constraint.

We remark that, we could have stated Theorem 10 for more general measurements given in Section 2.4 without the cone constraint. For instance, the following result holds for the weighted linear combination of individual norms and for the subgaussian ensemble.

Corollary 12 *Suppose $X_0 \in \mathbb{R}^{d\times d}$ obeys the general model with $k_1 = k_2 = k$ and $\mathcal{A}$ is a linear subgaussian map as described in Proposition 16. Choose $f(X) = \lambda_{\ell_1}\|X\|_1 + \lambda_\star\|X\|_\star$, where $\lambda_{\ell_1} = \beta$, $\lambda_\star = (1-\beta)\sqrt{d}$ and $0 \le \beta \le 1$. Then, whenever, $m \le \min\{m_{\text{low}}, c_1 n\}$, where,*

$$m_{\text{low}} = \frac{1}{2}\left(\beta\|\bar{X}_0\|_1 + (1-\beta)\|\bar{X}_0\|_\star\sqrt{d}\right)^2,$$

(2.6) fails with probability $1 - 4\exp(-c_2 m_{\text{low}})$. Here $c_1, c_2 > 0$ are constants as described in Proposition 16.

Remark 13 *Choosing $X_0 = \mathbf{a}\mathbf{a}^T$ where nonzero entries of $\mathbf{a}$ are ± 1 yields $\frac{1}{2}(\beta k + (1-\beta)\sqrt{d})^2$ on the right hand side. An explicit construction of an S&L matrix with maximal $\|\bar{X}\|_1, \|\bar{X}\|_\star$ is provided in Section A.1.3.*

This corollary compares well with the upper bound obtained in Corollary 23 of Section 2.5. In particular, both the bounds and the penalty parameters match up to logarithmic factors. Hence, together, they sandwich the sample complexity of the combined cost $f(X)$.

2.4 Measurement Ensembles

This section will make use of standard results on sub-gaussian random variables and random matrix theory to obtain probabilistic statements. We will explain how one can analyze the right hand side of (2.4) for,

- Matrices with sub-gaussian rows,
- Subsampled standard basis (in matrix completion),
- Quadratic measurements arising in phase retrieval.

2.4.1 Sub-gaussian Measurements

We first consider the measurement maps with sub-gaussian entries. The following definitions are borrowed from [155].

Definition 14 (Sub-gaussian random variable) *A random variable x is sub-gaussian if there exists a constant $K > 0$ such that for all $p \geq 1$,*

$$(\mathbb{E}|x|^p)^{1/p} \leq K\sqrt{p}.$$

The smallest such K is called the sub-gaussian norm of x and is denoted by $\|x\|_{\Psi_2}$. A sub-exponential random variable y is one for which there exists a constant K' such that, $\mathbb{P}(|y| > t) \leq \exp(1 - \frac{t}{K'})$. The variable x is sub-gaussian if and only if x^2 is sub-exponential.

Definition 15 (Isotropic sub-gaussian vector) *A random vector $x \in \mathbb{R}^n$ is sub-gaussian if the one dimensional marginals $x^T\mathbf{v}$ are sub-gaussian random variables for all $\mathbf{v} \in \mathbb{R}^n$. The sub-gaussian norm of x is defined as,*

$$\|x\|_{\Psi_2} = \sup_{\|\mathbf{v}\|=1} \|x^T\mathbf{v}\|_{\Psi_2}.$$

The vector x is also isotropic if its covariance is equal to the identity, i.e. $\mathbb{E}xx^T = \mathbf{I}_n$.

Proposition 16 (Sub-gaussian measurements) *Suppose A has i.i.d. rows in either of the following forms,*

- *a copy of a zero-mean isotropic sub-gaussian vector $\mathbf{a} \in \mathbb{R}^n$, where $\|\mathbf{a}\|_2 = \sqrt{n}$ almost surely.*

- *the rows consist of i.i.d. zero-mean and unit-variance sub-gaussian entries.*

Then, there exists constants c_1, c_2 depending only on the sub-gaussian norm of the rows, such that, whenever $m \leq c_1 n$, with probability $1 - 4\exp(-c_2 m)$, we have,

$$\frac{\|A\bar{x}_0\|_2^2}{\sigma_{\min}^2(A^T)} \leq \frac{2m}{n}.$$

Proof. Using Theorem 5.58 of [155], there exists constants c, C depending only on the sub-gaussian norm of $\mathbf{a}$ such that for any $t \geq 0$, with probability $1 - 2\exp(-ct^2)$

$$\sigma_{\min}(A^T) \geq \sqrt{n} - C\sqrt{m} - t\,.$$

Choosing $t = C\sqrt{m}$ and $m \leq \frac{n}{100C^2}$ ensures that $\sigma_{\min}(A^T) \geq \frac{4\sqrt{n}}{5}$.

Next, we shall estimate $\|A\bar{x}_0\|_2^2$, which is a sum of i.i.d. sub-exponential random variables identical to $|\mathbf{a}^T\bar{x}_0|^2$. Note that $\mathbb{E}|\mathbf{a}^T\bar{x}_0|^2 = 1$. Hence, Proposition 5.16 of [155] gives,

$$\mathbb{P}(\|A\bar{x}_0\|_2^2 \geq m + t) \leq 2\exp(-c' \min\{\frac{t^2}{m}, t\}).$$

Choosing $t = \frac{7m}{25}$, we find that $\mathbb{P}(\|A\bar{x}_0\|_2^2 \geq \frac{32m}{25}) \leq 2\exp(-c''m)$. Combining the two, we obtain,

$$\mathbb{P}\left(\frac{\|A\bar{x}_0\|_2^2}{\sigma_{\min}^2(A^T)} \leq \frac{2m}{n}\right) \geq 1 - 4\exp(-c'''m).$$

The second statement can be proved in the exact same manner by using Theorem 5.39 of [155] instead of Theorem 5.58. ∎

Remark 17 *While Proposition 16 assumes* $\mathbf{a}$ *has fixed* ℓ_2 *norm, this can be ensured by properly normalizing rows of* A *(assuming they stay sub-gaussian). For instance, if the* ℓ_2 *norm of the rows are larger than* $c\sqrt{n}$ *for a positive constant* c*, normalization will not affect sub-gaussianity. Note that, scaling rows of a matrix does not change its null space.*

2.4.2 Randomly Sampling Entries

We now consider the scenario where each row of A is chosen from the standard basis uniformly at random. Note that, when m is comparable to n, there is a nonnegligible probability that A will have duplicate rows. Theorem 3 does not take this situation into account which would make $\sigma_{\min}(A^T) = 0$. In this case, one can discard the copies as they don't affect the recoverability of x_0. This would get rid of the ill-conditioning, since the new matrix is well-conditioned with the exact same null space as the original. This corresponds to a "sampling without replacement" scheme where we ensure each row is different.

Similar to achievability results in matrix completion [35], the following failure result requires true signal to be incoherent with the standard basis, where incoherence is characterized by $\|\bar{x}_0\|_\infty$, which lies between $\frac{1}{\sqrt{n}}$ and 1.

Proposition 18 (Sampling entries) *Let $\{\mathbf{e}_i\}_{i=1}^n$ be the standard basis in $\mathbb{R}^n$ and suppose each row of A is chosen from $\{\mathbf{e}_i\}_{i=1}^n$ uniformly at random. Let $\widehat{A}$ be the matrix obtained by removing duplicate rows in A. Then, with probability $1-\exp(-\frac{m}{4n\|\bar{x}_0\|_\infty^2})$, we have,*

$$\frac{\|\widehat{A}\bar{x}_0\|_2^2}{\sigma_{\min}^2(\widehat{A})} \leq \frac{2m}{n}.$$

Proof. Let $\widehat{A}$ be the matrix obtained by discarding the rows of A that occur multiple times except one of them. Clearly $\text{Null}(\widehat{A}) = \text{Null}(A)$ hence they are equivalent for the purpose of recovering x_0. Furthermore, $\sigma_{\min}(\widehat{A}) = 1$. Therefore, we are interested in upper bounding $\|\widehat{A}\bar{x}_0\|_2$.

Clearly $\|\widehat{A}\bar{x}_0\|_2 \leq \|A\bar{x}_0\|_2$. Hence, we will bound $\|A\bar{x}_0\|_2^2$ probabilistically. Let $\mathbf{a}$ be the first row of A. $|\mathbf{a}^T\bar{x}_0|^2$ is a random variable, with mean $\frac{1}{n}$ and is upper bounded by $\|\bar{x}_0\|_\infty^2$. Applying the Chernoff Bound yields

$$\mathbb{P}(\|A\bar{x}_0\|_2^2 \geq \frac{m}{n}(1+\delta)) \leq \exp(-\frac{m\delta^2}{2(1+\delta)n\|\bar{x}_0\|_\infty^2}).$$

Setting $\delta = 1$, we find that, with probability $1-\exp(-\frac{m}{4n\|\bar{x}_0\|_\infty^2})$, we have,

$$\frac{\|\widehat{A}\bar{x}_0\|_2^2}{\sigma_{\min}(\widehat{A})^2} \leq \frac{\|A\bar{x}_0\|_2^2}{\sigma_{\min}(\widehat{A})^2} \leq \frac{2m}{n},$$

completing the proof. ∎

A significant application of this result would be for the low-rank tensor completion problem, where we randomly observe some entries of a low-rank tensor and try to reconstruct it. A promising approach for this problem is using the weighted linear combinations of nuclear norms of the unfoldings of the tensor to induce the low-rank tensor structure described in (2.2), [68, 75]. Related work [120] shows the poor performance of (2.2) for the special case

of Gaussian measurements. Combination of Theorem 3 and Proposition 18 will immediately extend the results of [120] to the more applicable tensor completion setup (under proper incoherence conditions that bound $\|\bar{x}_0\|_\infty$).

Remark 19 *In Propositions 16 and 18, we can make the upper bound for the ratio* $\frac{\|A\bar{x}_0\|_2^2}{\sigma_{\min}(A)^2}$ *arbitrarily close to* $\frac{m}{n}$ *by changing the proof parameters. Combined with Proposition 4, this would suggest that, failure happens, when* $m < n\kappa_{\min}$.

2.4.3 Quadratic Measurements

As mentioned in the phase retrieval problem, quadratic measurements $|\mathbf{v}^T\mathbf{a}|^2$ of the vector $\mathbf{a} \in \mathbb{R}^d$ can be linearized by the change of variable $\mathbf{a} \to X_0 = \mathbf{a}\mathbf{a}^T$ and using $\mathbf{V} = \mathbf{v}\mathbf{v}^T$. The following proposition can be used to obtain a lower bound for such ensembles when combined with Theorem 3.

Proposition 20 *Suppose we observe quadratic measurements* $\mathcal{A}(X_0) \in \mathbb{R}^m$ *of a matrix* $X_0 = \mathbf{a}\mathbf{a}^T \in \mathbb{R}^{d\times d}$. *Here, assume that ith entry of* $\mathcal{A}(X_0)$ *is equal to* $|\mathbf{v}_i^T\mathbf{a}|^2$ *where* $\{\mathbf{v}_i\}_{i=1}^m$ *are independent vectors, either with* $\mathcal{N}(0,1)$ *entries or are uniformly distributed over the sphere with radius* $\sqrt{d}$. *Then, there exists absolute constants* $c_1, c_2 > 0$ *such that whenever* $m < \frac{c_1 d}{\log d}$, *with probability* $1 - 2ed^{-2}$,

$$\frac{\|\mathcal{A}(\bar{X}_0)\|_2}{\sigma_{\min}(A^T)} \leq \frac{c_2\sqrt{m}\log d}{d}.$$

Proof. Let $\mathbf{V}_i = \mathbf{v}_i\mathbf{v}_i^T$. Without loss of generality, assume $\mathbf{v}_i$'s are uniformly distributed over the sphere with radius $\sqrt{d}$. To lower bound $\sigma_{\min}(A^T)$, we estimate the coherence of its columns, defined by,

$$\mu(A^T) = \max_{i\neq j} \frac{|\langle \mathbf{V}_i, \mathbf{V}_j\rangle|}{\|\mathbf{V}_i\|_F\|\mathbf{V}_j\|_F} = \frac{(\mathbf{v}_i^T\mathbf{v}_j)^2}{d^2}.$$

Section 5.2.5 of [155] states that the sub-gaussian norm of $\mathbf{v}_i$ is bounded by an absolute constant. Hence, conditioned on $\mathbf{v}_j$ (which satisfies $\|\mathbf{v}_j\|_2 = \sqrt{d}$), $\frac{(\mathbf{v}_i^T\mathbf{v}_j)^2}{d}$ is a subexponential

random variable with mean 1. Using Definition 14, there exists a constant $c > 0$ such that,

$$\mathbb{P}\left(\frac{(\mathbf{v}_i^T \mathbf{v}_j)^2}{d} > c \log d\right) \leq e d^{-4}.$$

Union bounding over all i, j pairs ensures that with probability ed^{-2} we have $\mu(A^T) \leq c\frac{\log d}{d}$. Next, we use the standard result that for a matrix with columns of equal length, $\sigma_{\min}(A^T) \geq d(1-(m-1)\mu)$. The reader is referred to Proposition 1 of [152]. Hence, $m \leq \frac{d}{2c \log d}$, gives $\sigma_{\min}(A^T) \geq \frac{d}{2}$.

It remains to upper bound $\|\mathcal{A}(\bar{X}_0)\|_2$. The ith entry of $\mathcal{A}(\bar{X}_0)$ is equal to $|\mathbf{v}_i^T \bar{\mathbf{a}}|^2$, hence it is subexponential. Consequently, there exists a constant c' so that each entry is upper bounded by $\frac{c'}{2} \log d$ with probability $1 - ed^{-3}$. Union bounding, and using $m \leq d$, we find that $\|\mathcal{A}(\bar{X}_0)\|_2 \leq \frac{c'}{2}\sqrt{m} \log d$ with probability $1 - ed^{-2}$. Combining with the estimate of $\sigma_{\min}(A^T)$ completes the proof. ∎

Comparison to existing literature Proposition 20 is useful to estimate the performance of the sparse phase retrieval problem, in which $\mathbf{a}$ is a k sparse vector, and we minimize a combination of the ℓ_1 norm and the nuclear norm to recover X_0. Combined with Theorem 3, Proposition 20 gives that, whenever $m \leq \frac{c_1 d}{\log d}$ and $\frac{c_2\sqrt{m}\log d}{d} \leq \min\{\frac{\|\bar{X}_0\|_1}{d}, \frac{\|\bar{X}_0\|_\star}{\sqrt{d}}\}$, recovery fails with high probability. Since $\|\bar{X}_0\|_\star = 1$ and $\|\bar{X}_0\|_1 = \|\bar{\mathbf{a}}\|_1^2$, the failure condition reduces to,

$$m \leq \frac{c}{\log^2 d} \min\{\|\bar{\mathbf{a}}\|_1^4, d\}.$$

When $\bar{\mathbf{a}}$ is a k-sparse vector with ± 1 entries, in a similar flavor to Theorem 10, the right hand side has the form $\frac{c}{\log^2 d} \min\{k^2, d\}$.

We emphasize that the lower bound provided in [102] is directly comparable to our results. The authors in [102] consider the same problem and give two results: first, if $m \geq \mathrm{O}(\|\bar{\mathbf{a}}\|_1^2 k \log d)$ then minimizing $\|X\|_1 + \lambda \mathrm{tr}(X)$ for suitable value of λ over the set of PSD matrices will exactly recover X_0 with high probability. Secondly, their Theorem 1.3 provides a necessary condition (lower bound) on the number of measurements, under which

the recovery program fails to recover X_0 with high probability. In particular, their failure condition is $m \leq \min\{m_0, \frac{d}{40\log d}\}$ where $m_0 = \frac{\max(\|\bar{\mathbf{a}}\|_1^2 - k/2, 0)^2}{500\log^2 d}$.

Observe that both results require $m \leq \mathrm{O}(\frac{d}{\log d})$. Focusing on the sparsity requirements, when the nonzero entries are sufficiently diffused (i.e. $\|\mathbf{a}\|_1^2 \approx k$) both results yield $\mathcal{O}(\frac{\|\bar{\mathbf{a}}\|^4}{\log^2 d})$ as a lower bound. On the other hand, if $\|\bar{\mathbf{a}}\|_1 \leq \sqrt{k/2}$, their lower bound becomes trivial while our lower bound still requires $\mathcal{O}(\frac{\|\bar{\mathbf{a}}\|^4}{\log^2 d})$ measurements. $\|\bar{\mathbf{a}}\|_1 \leq \sqrt{k/2}$ can happen as soon as the nonzero entries are spiky, i.e. some of the entries are much larger than the rest. In this sense, our bounds are tighter. On the other hand, their lower bound includes the PSD constraint unlike ours.

2.4.4 Asymptotic Regime

While we discussed two cases in the nonasymptotic setup, we believe significantly more general results can be stated asymptotically ($m, n \to \infty$). For instance, under finite fourth moment constraint, thanks to the Bai-Yin law [11], asymptotically, the smallest singular value of a matrix with i.i.d. unit variance entries concentrates around $\sqrt{n} - \sqrt{m}$. Similarly, $\|A\bar{x}_0\|_2^2$ is the sum of independent variables; hence thanks to the law of large numbers, we will have $\frac{\|A\bar{x}_0\|_2^2}{m} \to 1$. Together, these yield $\frac{\|A\bar{x}_0\|_2}{\sigma_{\min}(A^T)} \to \frac{\sqrt{m}}{\sqrt{n}-\sqrt{m}}$.

2.5 Upper bounds

We now state an upper bound on the simultaneous optimization for Gaussian measurement ensemble. Our upper bound will be in terms of distance to the dilated subdifferentials.

To accomplish this, we make use of the recent theory on the sample complexity of linear inverse problems. It has been recently shown that (2.3) exhibits a phase transition from failure with high probability to success with high probability when the number of Gaussian measurements are around the quantity $m_{PT} = \mathbf{D}(\mathrm{cone}(\partial f(x_0)))^2$ [6, 42]. This phenomenon was first observed by Donoho and Tanner, who calculated the phase transitions for ℓ_1 minimization and showed that $\mathbf{D}(\mathrm{cone}(\partial\|x_0\|_1))^2 \leq 2k\log\frac{en}{k}$ for a k-sparse vector in $\mathbb{R}^n$ [58]. All of these works focus on signals with a single structure and do not study properties of a

penalty that is a combination of norms. The next theorem relates the phase transition point of the joint optimization (2.3) to the individual subdifferentials.

Theorem 21 *Suppose A has i.i.d. $\mathcal{N}(0,1)$ entries and let $f(x) = \sum_{i=1}^{\tau} \lambda_i \|x\|_{(i)}$. For positive scalars $\{\alpha_i\}_{i=1}^{\tau}$, let $\bar{\lambda}_i = \frac{\lambda_i \alpha_i^{-1}}{\sum_{i=1}^{\tau} \lambda_i \alpha_i^{-1}}$ and define,*

$$m_{\text{up}}(\{\alpha_i\}_{i=1}^{\tau}) \coloneqq \left(\sum_i \bar{\lambda}_i \mathbf{D}(\alpha_i \partial \|x_0\|_{(i)}) \right)^2$$

If $m \geq (\sqrt{m_{\text{up}}} + t)^2 + 1$, then program (2.3) will succeed with probability $1 - 2\exp(-\frac{t^2}{2})$.

Proof. Fix $\mathbf{h}$ as an i.i.d. standard normal vector. Let g_i be such that $\alpha_i g_i$ is closest to $\mathbf{h}$ over $\alpha_i \partial \|x_0\|_{(i)}$. Let $\gamma = (\sum_i \frac{\lambda_i}{\alpha_i})^{-1}$. Then, we may write,

$$\begin{aligned}
\inf_{g' \in \text{cone}(\partial f(x_0))} \|\mathbf{h} - g'\|_2 &\leq \inf_{g \in \partial f(x_0)} \|\mathbf{h} - \gamma g\|_2 \\
&\leq \|\mathbf{h} - \gamma \sum_i \lambda_i g_i\|_2 \\
&= \|\mathbf{h} - \gamma \sum_i \frac{\lambda_i}{\alpha_i} \alpha_i g_i\|_2 \\
&= \|\mathbf{h} - \sum_i \bar{\lambda}_i \alpha_i g_i\|_2 \\
&\leq \sum_i \bar{\lambda}_i \|\mathbf{h} - \alpha_i g_i\|_2 \\
&= \sum_i \bar{\lambda}_i \inf_{g_i' \in \partial \|x_0\|_{(i)}} \|\mathbf{h} - \alpha_i g_i'\|_2.
\end{aligned}$$

Taking the expectations of both sides and using the definition of $\mathbf{D}(\cdot)$, we find that

$$\mathbf{D}(\text{cone}(\partial f(x_0))) \leq \sum_i \bar{\lambda}_i \mathbf{D}(\alpha_i \partial \|x_0\|_{(i)}),$$

and $m_{\text{up}} \geq \mathbf{D}(\text{cone}(\partial f(x_0)))^2$. The result then follows from the fact that, when $m \geq (\mathbf{D}(\text{cone}(\partial f(x_0))) + t)^2 + 1$, recovery succeeds with probability $1 - 2\exp(-\frac{t^2}{2})$. To see this, first, as discussed in Proposition 3.6 of [42], $\mathbf{D}(\text{cone}(\partial f(x_0)))$ is equal to the Gaussian width of the "tangent cone intersected with the unit ball" (see Theorem 58 for a definition of Gaussian width). Then, Corollary 3.3 of [42] yields the probabilistic statement. ∎

For Theorem 21 to be useful, choices of α_i should be made wisely. An obvious choice is letting,

$$\alpha_i^* = \arg\min_{\alpha_i \geq 0} \mathbf{D}(\alpha_i \partial \|x_0\|_{(i)}). \tag{2.7}$$

With this choice, our upper bounds can be related to the individual sample complexities, which is equal to $\mathbf{D}(\text{cone}(\partial \|x_0\|_{(i)}))^2$. Proposition 1 of [67] shows that, if $\|\cdot\|_{(i)}$ is a *decomposable* norm, then,

$$0 \leq \mathbf{D}(\alpha_i^* \partial \|x_0\|_{(i)}) - \mathbf{D}(\text{cone}(\partial \|x_0\|_{(i)})) \leq 6.$$

Decomposability is defined and discussed in detail in Section 2.6.4. In particular, $\ell_1, \ell_{1,2}$ and the nuclear norm are decomposable. With this assumption, our upper bound suggests that the sample complexity of the simultaneous optimization is smaller than a certain convex combination of individual sample complexities.

Corollary 22 *Suppose A has i.i.d. $\mathcal{N}(0,1)$ entries and let $f(x) = \sum_{i=1}^{\tau} \lambda_i \|x\|_{(i)}$ for decomposable norms $\{\|\cdot\|_{(i)}\}_{i=1}^{\tau}$. Let $\{\alpha_i^*\}_{i=1}^{\tau}$ be as in (2.7) and assume they are strictly positive. Let $\bar{\lambda}_i^* = \frac{\lambda_i (\alpha_i^*)^{-1}}{\sum_{i=1}^{\tau} \lambda_i (\alpha_i^*)^{-1}}$ and define*

$$\sqrt{m_{\text{up}}(\{\alpha_i^*\}_{i=1}^{\tau})} \coloneqq \sum_i \bar{\lambda}_i^* \mathbf{D}(\text{cone}(\partial \|x_0\|_{(i)})) + 6.$$

If $m \geq (\sqrt{m_{\text{up}}} + t)^2 + 1$, then program (2.3) will succeed with probability $1 - 2\exp(-\frac{t^2}{2})$.

Here, we used the fact that $\sum_i \bar{\lambda}_i^* = 1$ to take 6 out of the sum over i. We note that Corollaries 8 and 22 can be related in the case of sparse and low-rank matrices. For norms of interest, roughly speaking,

- $n\kappa_i^2$ is proportional to the sample complexity $\mathbf{D}(\text{cone}(\partial \|x_0\|_{(i)}))^2$.
- L_i is proportional to $\frac{\sqrt{n}}{\alpha_i^*}$.

Consequently, the sample complexity of (2.3) will be upper and lower bounded by similar convex combinations.

2.5.1 Upper Bounds for the S&L Model

We now apply the bound of Theorem 21 to S&L matrices. To obtain simple and closed form bounds, we make use of existing results in the literature.

- Table II of [67]: If $x_0 \in \mathbb{R}^n$ is a k sparse vector, choosing $\alpha_{\ell_1} = \sqrt{2\log\frac{n}{k}}$, $\mathbf{D}(\alpha_{\ell_1}\partial\|x_0\|_1)^2 \leq 2k\log\frac{en}{k}$.

- Table 3 of [130]: If $X_0 \in \mathbb{R}^{d\times d}$ is a rank r matrix, choosing $\alpha_\star = 2\sqrt{d}$, $\mathbf{D}(\alpha_\star\partial\|X_0\|_\star)^2 \leq 6dr + 2d$.

Proposition 23 *Suppose A has i.i.d. $\mathcal{N}(0,1)$ entries and $X_0 \in \mathbb{R}^{d\times d}$ is a rank r matrix whose nonzero entries lie on a $k\times k$ submatrix. For $0 \leq \beta \leq 1$, let $f(X) = \lambda_{\ell_1}\|X\|_1 + \lambda_\star\|X\|_\star$ where $\lambda_{\ell_1} = \beta\sqrt{\log\frac{d}{k}}$ and $\lambda_\star = (1-\beta)\sqrt{d}$. Then, whenever,*

$$m \geq \left(2\beta k\sqrt{\log\frac{ed}{k}} + (1-\beta)\sqrt{6dr+2d} + t\right)^2 + 1,$$

X_0 can be recovered via (2.3) with probability $1 - 2\exp(-\frac{t^2}{2})$.

Proof. To apply Theorem 21, we choose $\alpha_{\ell_1} = \sqrt{4\log\frac{d}{k}}$ and $\alpha_\star = 2\sqrt{d}$. X_0 is effectively an (at most) k^2 sparse vector of size d^2. Hence, $\alpha_{\ell_1} = \sqrt{2\log\frac{d^2}{k^2}}$ and $\mathbf{D}(\alpha_{\ell_1}\|X_0\|_1)^2 \leq 4k^2\log\frac{ed}{k}$.

Now, for the choice of $\alpha_\star$, we have, $\mathbf{D}(\alpha_\star\|X_0\|_\star)^2 \leq 6dr + 2d$. Observe that $\alpha_{\ell_1}^{-1}\lambda_{\ell_1} = \frac{\beta}{2}$, $\alpha_\star^{-1}\lambda_\star = \frac{1-\beta}{2}$ and apply Theorem 21 to conclude. ∎

2.6 General Simultaneously Structured Model Recovery

Recall the setup from Section 2.2 where we consider a vector $x_0 \in \mathbb{R}^n$ whose structures are associated with a family of norms $\{\|\cdot\|_{(i)}\}_{i=1}^\tau$ and x_0 satisfies the cone constraint $x_0 \in \mathcal{C}$. This section is dedicated to the proofs of theorems in Section 2.3.1 and additional side results where the goal is to find lower bounds on the required number of measurements to recover x_0.

The following definitions will be helpful for the rest of our discussion. For a subspace M, denote its orthogonal complement by $M^{\perp}$. For a convex set M and a point x, we define the projection operator as

$$\mathcal{P}_M(x) = \arg\min_{\mathbf{u}\in M} \|x - \mathbf{u}\|_2 \,.$$

Given a cone $\mathcal{C}$, denote its dual cone by $\mathcal{C}^*$ and polar cone by $\mathcal{C}^\circ = -\mathcal{C}^*$, where $\mathcal{C}^*$ is defined as

$$\mathcal{C}^* = \{z \big| \, \langle z, \mathbf{v}\rangle \geq 0 \text{ for all } \mathbf{v} \in \mathcal{C}\} \,.$$

2.6.1 Preliminary Lemmas

We first show that the objective function $\max_{1\leq i\leq \tau} \frac{\|x\|_{(i)}}{\|x_0\|_{(i)}}$ can be viewed as the 'best' among the functions mentioned in (2.3) for recovery of x_0.

Lemma 24 *Consider the class of recovery programs in* (2.3)*. If the program*

$$\begin{aligned} \underset{x\in\mathcal{C}}{\textit{minimize}} \quad & f_{\text{best}}(x) \coloneqq \max_{i=1,\ldots,\tau} \frac{\|x\|_{(i)}}{\|x_0\|_{(i)}} \\ \textit{subject to} \quad & \mathcal{A}(x) = \mathcal{A}(x_0) \end{aligned} \tag{2.8}$$

fails to recover x_0, then any member of this class will also fail to recover x_0.

Proof. Suppose (2.8) does not have x_0 as an optimal solution and there exists x' such that $f_{\text{best}}(x') \leq f_{\text{best}}(x_0)$. Then

$$\frac{1}{\|x_0\|_{(i)}} \|x'\|_{(i)} \leq f_{\text{best}}(x') \leq f_{\text{best}}(x_0) = 1,$$

for $i = 1, \ldots, \tau$. This implies that

$$\|x'\|_{(i)} \leq \|x_0\|_{(i)}, \quad \text{for all } i = 1, \ldots, \tau. \tag{2.9}$$

Conversely, given (2.9), we have $f_{\text{best}}(x') \leq f_{\text{best}}(x_0)$ from the definition of f_{best}.

Furthermore, since we assume $h(\cdot)$ in (2.3) is non-decreasing in its arguments and increasing in at least one of them, (2.9) implies $f(x') \leq f(x_0)$ for any such function $f(\cdot)$. Thus, failure of $f_{\text{best}}(\cdot)$ in recovery of x_0 implies failure of any other function in (2.3). ∎

The following lemma provides necessary conditions for x_0 to be a minimizer of the problem (2.3).

Lemma 25 *If x_0 is a minimizer of the program (2.3), then there exist $\mathbf{v} \in \mathcal{C}^*$, z, and $g \in \partial f(x_0)$ such that*

$$g - \mathbf{v} - A^T z = 0 \quad \textit{and} \quad \langle x_0, \mathbf{v} \rangle = 0.$$

The proof of Lemma 25 follows from the KKT conditions for (2.3) to have x_0 as an optimal solution [19, Section 4.7].

The next lemma describes the subdifferential of any general function $f(x) = h(\|x\|_{(1)}, \ldots, \|x\|_{(\tau)})$ as discussed in Section 2.2.2.

Lemma 26 *For any subgradient of the function $f(x) = h(\|x\|_{(1)}, \ldots, \|x\|_{(\tau)})$ at $x \neq 0$ defined by convex function $h(\cdot)$, there exists non-negative constants w_i, $i = 1, \ldots, \tau$ such that*

$$g = \sum_{i=1}^{\tau} w_i g_i$$

where $g_i \in \partial \|x_0\|_{(i)}$.

Proof. Consider the function $N(x) = \begin{bmatrix} \|x\|_{(1)}, & \ldots, & \|x\|_{(\tau)} \end{bmatrix}^T$ by which we have $f(x) = h(N(x))$. By Theorem 10.49 in [137] we have

$$\partial f(x) = \bigcup \left\{ \partial (y^T N(x)) : \; y \in \partial h(N(x)) \right\}$$

where we used the convexity of f and h. Now notice that any $y \in \partial h(N(x))$ is a non-negative vector because of the monotonicity assumption on $h(\cdot)$. This implies that any subgradient $g \in \partial f(x)$ is in the form of $\partial (w^T N(x))$ for some nonnegative vector w. The desired result simply follows because subgradients of conic combinations of norms are conic combinations of their subgradients (see e.g. [136]). ∎

Using Lemmas 25 and 26, we now provide the proofs of Theorems 3 and 6.

2.6.2 Proof of Theorem 3

We prove the more general version of Theorem 3, which can take care of the cone constraint and alignment of subgradients over arbitrary subspaces. This will require us to extend the definition of correlation to handle subspaces. For a linear subspace $\mathcal{R} \in \mathbb{R}^n$ and a set $S \in \mathbb{R}^n$, we define,

$$\rho(\mathcal{R}, S) = \inf_{0 \neq s \in S} \frac{\|\mathcal{P}_{\mathcal{R}}(s)\|_2}{\|s\|_2}.$$

Proposition 27 *Let*

$$\sigma_{\mathcal{C}}(A^T) = \inf_{\|z\|_2=1} \frac{\|\mathcal{P}_{\mathcal{C}}(A^T z)\|_2}{\|A^T z\|_2}.$$

Let $\mathcal{R}$ be an arbitrary linear subspace orthogonal to the following cone,

$$\left\{ y \in \mathbb{R}^n \middle| x_0^T y = 0,\ y \in \mathcal{C}^* \right\}. \tag{2.10}$$

Suppose,

$$\rho(\mathcal{R}, \partial f(x_0)) \coloneqq \inf_{g \in \partial f(x_0)} \frac{\|\mathcal{P}_{\mathcal{R}}(g)\|_2}{\|g\|_2} > \frac{\sigma_{\max}(\mathcal{P}_{\mathcal{R}}(A^T))}{\sigma_{\mathcal{C}}(A^T)\sigma_{\min}(A^T)}.$$

Then, x_0 is not a minimizer of (2.3).

Proof. Suppose x_0 is a minimizer of (2.3). From Lemma 25, there exist a $g \in \partial f(x_0)$, $z \in \mathbb{R}^m$ and $\mathbf{v} \in \mathcal{C}^*$ such that

$$g = A^T z + \mathbf{v} \tag{2.11}$$

and $\langle x_0, \mathbf{v} \rangle = 0$. We first eliminate the contribution of $\mathbf{v}$ in equation (2.11). Projecting both sides of (2.11) onto the subspace $\mathcal{R}$ gives,

$$\mathcal{P}_{\mathcal{R}}(g) = \mathcal{P}_{\mathcal{R}}(A^T z) = \mathcal{P}_{\mathcal{R}}(A^T) z. \tag{2.12}$$

Taking the ℓ_2 norms,

$$\|\mathcal{P}_{\mathcal{R}}(g)\|_2 = \|\mathcal{P}_{\mathcal{R}}(A^T) z\|_2 \leq \sigma_{\max}(\mathcal{P}_{\mathcal{R}}(A^T)) \|z\|_2. \tag{2.13}$$

Since $\mathbf{v} \in \mathcal{C}^*$, from Lemma 55 we have $\mathcal{P}_{\mathcal{C}}(-\mathbf{v}) = \mathcal{P}_{\mathcal{C}}(A^T z - g) = 0$. Using Corollary 57,

$$\|g\|_2 \geq \|\mathcal{P}_{\mathcal{C}}(A^T z)\|_2. \tag{2.14}$$

From the initial assumption, for any $z \in \mathbb{R}^m$, we have,

$$\sigma_{\mathcal{C}}(A^T)\|A^T z\|_2 \leq \|\mathcal{P}_{\mathcal{C}}(A^T z)\|_2 \tag{2.15}$$

Combining (2.14) and (2.15) yields $\|g\|_2 \geq \sigma_{\mathcal{C}}(A^T)\|A^T z\|_2$. Further incorporating (2.13), we find,

$$\frac{\|\mathcal{P}_{\mathcal{R}}(g)\|_2}{\sigma_{\max}(\mathcal{P}_{\mathcal{R}}(A^T))} \leq \|z\|_2 \leq \frac{\|A^T z\|_2}{\sigma_{\min}(A^T)} \leq \frac{\|g\|_2}{\sigma_{\mathcal{C}}(A^T)\sigma_{\min}(A^T)}\,. \tag{2.16}$$

Hence, if x_0 is recoverable, then there exists $g \in \partial f(x_0)$ satisfying

$$\frac{\|\mathcal{P}_{\mathcal{R}}(g)\|_2}{\|g\|_2} \leq \frac{\sigma_{\max}(\mathcal{P}_{\mathcal{R}}(A^T))}{\sigma_{\mathcal{C}}(A^T)\sigma_{\min}(A^T)}\,,$$

completing the proof. ■

To obtain Theorem 3, choose $\mathcal{R} = \text{span}(\{x_0\})$ and $\mathcal{C} = \mathbb{R}^n$. This choice of $\mathcal{R}$ yields $\sigma_{\max}(\mathcal{P}_{\mathcal{R}}(A^T)) = \|\bar{x}_0\bar{x}_0^T A^T\|_2 = \|A\bar{x}_0\|_2$ and $\|\mathcal{P}_{\mathcal{R}}(g)\|_2 = |\bar{x}_0^T g|$. Choosing $\mathcal{C} = \mathbb{R}^n$ yields $\sigma_{\mathcal{C}}(A) = 1$. Also note that, for any choice of $\mathcal{C}$, x_0 is orthogonal to (2.10) by definition.

2.6.3 Proof of Theorem 6

Rotational invariance of Gaussian measurements allows us to make full use of Proposition 27. The following is a generalization of Theorem 6.

Proposition 28 *Consider the setup in Proposition 27 where A has i.i.d. $\mathcal{N}(0,1)$ entries. Let,*

$$m_{\text{low}} = \frac{n(1-\bar{\mathbf{D}}(\mathcal{C}))\rho(\mathcal{R}, \partial f(x_0))^2}{100},$$

and suppose that $\dim(\mathcal{R}) \leq m_{\text{low}}$. Then, whenever $m \leq m_{\text{low}}$, with probability

$$1 - 10\exp(-\frac{1}{16}\min\{m_{\text{low}}, (1-\bar{\mathbf{D}}(\mathcal{C}))^2 n\}),$$

(2.3) will fail for all functions $f(\cdot)$.

Proof. More measurements can only increase the chance of success. Hence, without losing generality, assume $m = m_{\text{low}}$ and $\dim(\mathcal{R}) \leq m$. The result will follow from Proposition 27. Recall that $m \leq \frac{(1-\bar{\mathbf{D}}(\mathcal{C}))n}{100}$.

- $\mathcal{P}_{\mathcal{R}}(A^T)$ is statistically identical to a $\dim(\mathcal{R}) \times m$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries under proper unitary rotation. Hence, using Corollary 5.35 of [155], with probability $1-2\exp(-\frac{m}{8})$, $\sigma_{\max}(\mathcal{P}_{\mathcal{R}}(A^T)) \leq 1.5\sqrt{m}+\sqrt{\dim(\mathcal{R})} \leq 2.5\sqrt{m}$. With the same probability, $\sigma_{\min}(A^T) \geq \sqrt{n} - 1.5\sqrt{m}$.

- From Theorem 59, using $m \leq \frac{(1-\bar{\mathbf{D}}(\mathcal{C}))n}{100}$, with probability $1 - 6\exp(-\frac{(1-\bar{\mathbf{D}}(\mathcal{C}))^2 n}{16})$, $\sigma^2_{\mathcal{C}}(A^T) \geq \frac{1-\bar{\mathbf{D}}(\mathcal{C})}{4(1+\bar{\mathbf{D}}(\mathcal{C}))} \geq \frac{1-\bar{\mathbf{D}}(\mathcal{C})}{8}$.

Since $\frac{m}{n} \leq \frac{1}{30}$, combining these, with the desired probability,

$$\frac{\sigma_{\max}(\mathcal{P}_{\mathcal{R}}(A^T))}{\sigma_{\mathcal{C}}(A^T)\sigma_{\min}(A^T)} \leq \sqrt{\frac{8}{1-\bar{\mathbf{D}}(\mathcal{C})}}\frac{2.5\sqrt{m}}{\sqrt{n}-1.5\sqrt{m}} < \frac{10\sqrt{m}}{\sqrt{(1-\bar{\mathbf{D}}(\mathcal{C}))n}}.$$

Finally, using Proposition 27 and the fact that $m \leq \frac{n(1-\bar{\mathbf{D}}(\mathcal{C}))}{100}\rho(\mathcal{R}, \partial f(x_0))^2$, with the same probability (2.3) fails. ∎

To prove Theorem 6, choose $\mathcal{R} = \text{span}(\{x_0\})$ and use the first statement of Proposition 4. To show Corollary 8, choose $\mathcal{R} = \text{span}(\{x_0\})$ and use the second statement of Proposition 4.

2.6.4 Enhanced Lower Bounds

From our initial results, it may look like our lower bounds are suboptimal. For instance, considering only the ℓ_1 norm, $\kappa = \frac{\|\bar{x}_0\|_1}{\sqrt{n}}$ lies between $\frac{1}{\sqrt{n}}$ and $\sqrt{\frac{k}{n}}$ for a k sparse signal. Combined with Theorem 6, this gives a lower bound of $\|\bar{x}_0\|_1^2$ measurements. On the other hand, clearly, we need at least $\mathrm{O}(k)$ measurements to estimate a k sparse vector.

Indeed, Proposition 27 gives such a bound with a better choice of $\mathcal{R}$. In particular, let us choose $\mathcal{R} = \text{span}(\{\text{sign}(x_0)\})$. For any $g \in \partial\|x_0\|_1$, we have that,

$$\frac{\left\langle g, \frac{\text{sign}(x_0)}{\sqrt{k}} \right\rangle}{L} = \sqrt{\frac{k}{n}} \implies \rho(\text{sign}(x_0), \partial\|x_0\|_1) = \sqrt{\frac{k}{n}}.$$

Hence, we immediately have $m \geq \text{O}(k)$ as a lower bound. The idea of choosing such sign vectors can be generalized to the so-called decomposable norms.

Definition 29 (Decomposable Norm) *A norm $\|\cdot\|$ is decomposable at $x \in \mathbb{R}^n$ if there exist a subspace $T \subset \mathbb{R}^n$ and a vector $\mathbf{e} \in T$ such that the subdifferential at x has the form*

$$\partial\|x\| = \{z \in \mathbb{R}^n \ : \ \mathcal{P}_T(z) = \mathbf{e} \ , \ \|\mathcal{P}_{T^\perp}(z)\|^* \leq 1\}.$$

We refer to T as the support *and $\mathbf{e}$ as the* sign vector *of x with respect to $\|\cdot\|$.*

Similar definitions are used in [36] and [162]. Our definition is simpler and less strict compared to these works. Note that L is a global property of the norm while $\mathbf{e}$ and T depend on both the norm and the point under consideration (decomposability is a local property in this sense).

To give some intuition for Definition 29, we review examples of norms that arise when considering simultaneously sparse and low rank matrices. For a matrix $X \in \mathbb{R}^{d_1 \times d_2}$, let $X_{i,j}$, $X_{i,\cdot}$ and $X_{\cdot,j}$ denote its (i,j) entry, ith row and jth column respectively.

Lemma 30 (see [36]) *The ℓ_1 norm, the $\ell_{1,2}$ norm and the nuclear norm are decomposable as follows.*

- **ℓ_1 norm** *is decomposable at every $x \in \mathbb{R}^n$, with sign $\mathbf{e} = \text{sgn}(x)$, and support*

$$T = \text{supp}(x) = \{y \in \mathbb{R}^n : x_i = 0 \Rightarrow y_i = 0 \ \ \forall \ i \leq n\}.$$

- **$\ell_{1,2}$ norm** *is decomposable at every $X \in \mathbb{R}^{d_1 \times d_2}$. The support is*

$$T = \left\{Y \in \mathbb{R}^{d_1 \times d_2} : \ \ X_{\cdot,i} = \mathbf{0} \Rightarrow Y_{\cdot,i} = \mathbf{0} \ \ \forall \ i \leq d_2\right\},$$

and the sign vector $\mathbf{e} \in \mathbb{R}^{d_1 \times d_2}$ *is obtained by normalizing the columns of* X *present in the support,* $\mathbf{e}_{\cdot,j} = \frac{X_{\cdot,j}}{\|X_{\cdot,j}\|_2}$ *if* $\|X_{\cdot,j}\|_2 \neq 0$*, and setting the rest of the columns to zero.*

- **Nuclear norm** *is decomposable at every* $X \in \mathbb{R}^{d_1 \times d_2}$*. For a matrix* X *with rank* r *and compact singular value decomposition* $X = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ *where* $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$*, we have* $\mathbf{e} = \mathbf{U}\mathbf{V}^T$ *and*

$$\begin{aligned} T &= \left\{ Y \in \mathbb{R}^{d_1 \times d_2} : \ (\mathbf{I} - \mathbf{U}\mathbf{U}^T) Y (\mathbf{I} - \mathbf{V}\mathbf{V}^T) = \mathbf{0} \right\} \\ &= \left\{ \mathbf{Z}_1 \mathbf{V}^T + \mathbf{U}\mathbf{Z}_2^T \mid \mathbf{Z}_1 \in \mathbb{R}^{d_1 \times r}, \mathbf{Z}_2 \in \mathbb{R}^{d_2 \times r} \right\}. \end{aligned}$$

The next lemma shows that the sign vector $\mathbf{e}$ will yield the largest correlation with the subdifferential and the best lower bound for such norms.

Lemma 31 *Let* $\|\cdot\|$ *be a decomposable norm with support* T *and sign vector* $\mathbf{e}$*. For any* $\mathbf{v} \neq 0$*, we have that,*

$$\rho(\mathbf{v}, \partial\|x_0\|) \leq \rho(\mathbf{e}, \partial\|x_0\|) \tag{2.17}$$

Also $\rho(\mathbf{e}, \partial\|x_0\|) \geq \frac{\|\mathbf{e}\|_2}{L}$.

Proof. Let $\mathbf{v}$ be a unit vector. Without losing generality, assume $\mathbf{v}^T\mathbf{e} \geq 0$. Pick a vector $z \in T^\perp$ with $\|z\|^* = 1$ such that $z^T\mathbf{v} \leq 0$ (otherwise pick $-z$). Now, consider the class of subgradients $g(\alpha) = \mathbf{e} + \alpha z$ for $1 \geq \alpha \geq -1$. Then,

$$\inf_{-1 \leq \alpha \leq 1} \frac{|\mathbf{v}^T g(\alpha)|}{\|g(\alpha)\|_2} = \inf_{0 \leq \alpha \leq 1} \frac{|\mathbf{v}^T g(\alpha)|}{\|g(\alpha)\|_2} = \inf_{0 \leq \alpha \leq 1} \frac{|\mathbf{e}^T\mathbf{v} - \alpha |z^T\mathbf{v}||}{(\|\mathbf{e}\|_2^2 + \alpha^2 \|z\|_2^2)^{1/2}}.$$

If $|z^T\mathbf{v}| \geq \mathbf{e}^T\mathbf{v}$, then, the numerator can be made 0 and $\rho(\mathbf{v}, \partial\|x_0\|) = 0$. Otherwise, the right hand side is decreasing function of α, hence the minimum is achieved at $\alpha = 1$, which

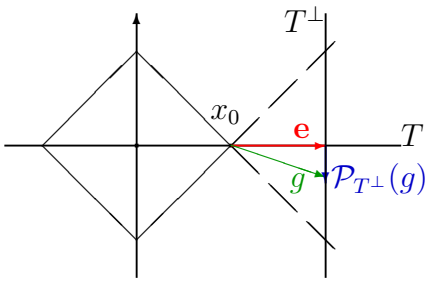


Figure 2.4: An example of a decomposable norm: ℓ_1 norm is decomposable at $x_0 = (1, 0)$. The sign vector $\mathbf{e}$, the support T, and shifted subspace $T^\perp$ are illustrated. A subgradient g at x_0 and its projection onto $T^\perp$ are also shown.

gives,

$$\begin{aligned}
\inf_{-1\leq\alpha\leq 1} \frac{|\mathbf{v}^T g(\alpha)|}{\|g(\alpha)\|_2} &= \frac{|\mathbf{e}^T\mathbf{v} - |z^T\mathbf{v}||}{(\|\mathbf{e}\|_2^2 + \|z\|_2^2)^{1/2}} \\
&\leq \frac{|\mathbf{e}^T\mathbf{v}|}{(\|\mathbf{e}\|_2^2 + \|z\|_2^2)^{1/2}} \\
&\leq \frac{\|\mathbf{e}\|_2}{(\|\mathbf{e}\|_2^2 + \|z\|_2^2)^{1/2}} \\
&= \inf_{-1\leq\alpha\leq 1} \frac{|\bar{\mathbf{e}}^T g(\alpha)|}{\|g(\alpha)\|_2}
\end{aligned}$$

where we used $\mathbf{e}^T g(\alpha) = \mathbf{e}^T\mathbf{e} = \|\mathbf{e}\|_2^2$. Hence, along any direction z, $\mathbf{e}$ yields a higher minimum correlation than $\mathbf{v}$. To obtain (2.17), further take the infimum over all $z \in T^\perp, \|z\|^* \leq 1$ which yields the infimum over $\partial\|x_0\|$. Finally, use $\|g(\alpha)\|_2 \leq L$ to lower bound $\rho(\mathbf{e}, \partial\|x_0\|)$.

■

Based on Lemma 31, the individual lower bounds are $\mathrm{O}(\frac{\|\mathbf{e}\|_2^2}{L^2})n$. Calculating $\frac{\|\mathbf{e}\|_2^2}{L^2}n$ for the norms in Lemma 30, reveals that, this quantity is k for a k sparse vector, cd_1 for a c-column sparse matrix and $r\max\{d_1, d_2\}$ for a rank r matrix. Compared to bounds obtained by using $\bar{x}_0$, these new quantities are directly proportional to the true model complexities. Finally, we remark that, these new bounds correspond to choosing x_0 that maximizes the value of $\|\bar{x}_0\|_1, \|\bar{x}_0\|_\star$ or $\|\bar{x}_0\|_{1,2}$ while keeping sparsity, rank or column sparsity fixed. In particular,

in these examples, $\mathbf{e}$ has the same sparsity, rank, column sparsity as x_0.

The next proposition derives a correlation bound for the combination of decomposable norms as well as a simple lower bound on the sample complexity.

Proposition 32 *Given decomposable norms $\|\cdot\|_{(i)}$ with supports T_i and sign vectors $\mathbf{e}_i$. Let $T_\cap = \bigcap_{1\le i\le\tau} T_i$. Choose the subspace $\mathcal{R}$ to be a subset of $T_\cap$.*

- *Assume $\langle \mathcal{P}_\mathcal{R}(\mathbf{e}_i), \mathcal{P}_\mathcal{R}(\mathbf{e}_j)\rangle \ge 0$ for all i, j and $\min_{1\le i\le\tau} \frac{\|\mathcal{P}_\mathcal{R}(\mathbf{e}_i)\|_2}{\|\mathbf{e}_i\|_2} \ge \upsilon$. Then,*

$$\rho(\mathcal{R}, \partial f(x_0)) \ge \frac{\upsilon}{\sqrt{\tau}} \min_{1\le i\le\tau} \rho(\mathbf{e}_i, \partial\|x_0\|_{(i)}).$$

- *Consider Proposition 27 with Gaussian measurements and suppose $\mathcal{R}$ is orthogonal to the set (2.10). Let $f(x) = \sum_{i=1}^{\tau} \lambda_i \|x\|_{(i)}$ for some nonnegative $\{\lambda_i\}$'s. Then, if $m < \dim(\mathcal{R})$, (2.3) fails with probability 1.*

Proof. Let $g = \sum_{i=1}^{\tau} w_i g_i$ for some $g_i \in \partial\|x_0\|_{(i)}$. First, $\|g\|_2 \le \sum_{i=1}^{\tau} w_i \|g_i\|_2$. Next,

$$\begin{aligned}\|\mathcal{P}_\mathcal{R}(g)\|_2^2 = \|\sum_{i=1}^{\tau} w_i \mathcal{P}_\mathcal{R}(\mathbf{e}_i)\|_2^2 &\ge \sum_{i=1}^{\tau} w_i^2 \|\mathcal{P}_\mathcal{R}(\mathbf{e}_i)\|_2^2 \\ &\ge \upsilon^2 \sum_{i=1}^{\tau} w_i^2 \|\mathbf{e}_i\|_2^2 \ge \frac{\upsilon^2}{\tau}(\sum_{i=1}^{\tau} w_i \|\mathbf{e}_i\|_2)^2.\end{aligned}$$

To see the second statement, consider the line (2.12) from the proof of Proposition 27. $\mathcal{P}_\mathcal{R}(g) = \sum_{i=1}^{\tau} \lambda_i \mathcal{P}_\mathcal{R}(\mathbf{e}_i)$. On the other hand, the column space of $\mathcal{P}_\mathcal{R}(A^T)$ is an m-dimensional random subspace of $\mathcal{R}$. If $m < \dim(\mathcal{R})$, $\mathcal{P}_\mathcal{R}(g)$ is linearly independent of $\mathcal{P}_\mathcal{R}(A^T)$ with probability 1 and (2.12) will not hold. ∎

In the next section, we show how better choices of $\mathcal{R}$ (based on the decomposability assumption) can improve the lower bounds for S&L recovery.

2.7 Numerical Experiments

In this section, we numerically verify our theoretical bounds on the number of measurements for the sparse and low-rank recovery problem. We demonstrate the empirical performance

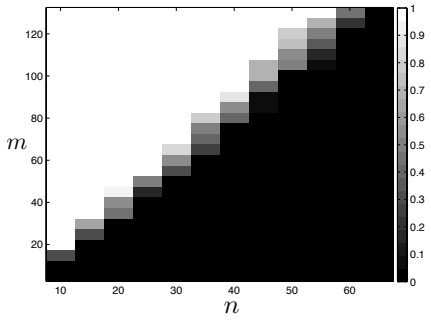


Figure 2.5: Performance of the recovery program minimizing $\max\{\frac{\mathrm{tr}(X)}{\mathrm{tr}(X_0)}, \frac{\|X\|_{1,2}}{\|X_0\|_{1,2}}\}$ with a PSD constraint. The dark region corresponds to the experimental region of failure due to insufficient measurements. As predicted by Theorem 10, the number of required measurements increases linearly with rd.

of the weighted maximum of the norms f_{best} (see Lemma 24), as well as the weighted sum of norms.

The experimental setup is as follows. Our goal is to explore how the number of required measurements m scales with the size of the matrix d. We consider a grid of (m, d) values, and generate at least 100 test instances for each grid point (in the boundary areas, we increase the number of instances to at least 200).

We generate the target matrix X_0 by generating a $k \times r$ i.i.d. Gaussian matrix $\mathbf{G}$, and inserting the $k \times k$ matrix $\mathbf{G}\mathbf{G}^T$ in an $d \times d$ matrix of zeros. We take $r = 1$ and $k = 8$ in all of the following experiments; even with these small values, we can observe the scaling predicted by our bounds. In each test, we measure the normalized recovery error $\frac{\|X - X_0\|_F}{\|X_0\|_F}$ and declare successful recovery when this error is less than 10^{-4}. The optimization programs are solved using the CVX package [74], which calls the SDP solver SeDuMi [147].

We first test our bound in part (b) of Theorem 10, $\Omega(rd)$, on the number of measurements for recovery in the case of minimizing $\max\{\frac{\mathrm{tr}(X)}{\mathrm{tr}(X_0)}, \frac{\|X\|_{1,2}}{\|X_0\|_{1,2}}\}$ over the set of PSD matrices. Figure 2.5 shows the results, which demonstrates m scaling linearly with d (note that $r = 1$).

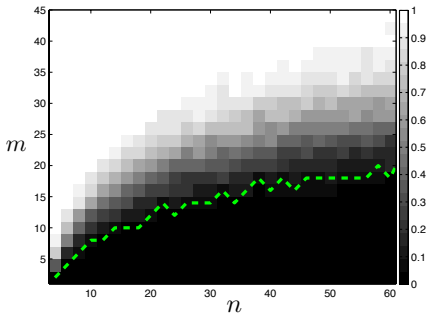


Figure 2.6: Performance of the recovery program minimizing $\max\{\frac{\mathrm{tr}(X)}{\mathrm{tr}(X_0)}, \frac{\|X\|_1}{\|X_0\|_1}\}$ with a PSD constraint. $r = 1, k = 8$ and d is allowed to vary. The plot shows m versus d to illustrate the lower bound $\Omega(\min\{k^2, dr\})$ predicted by Theorem 10.

Next, we replace the $\ell_{1,2}$ norm with the ℓ_1 norm and consider a recovery program that emphasizes entry-wise sparsity rather than block sparsity. Figure 2.6 demonstrates the lower bound $\Omega(\min\{k^2, d\})$ in Part (c) of Theorem 10 where we attempt to recover a rank-1 PSD matrix X_0 by minimizing $\max\{\frac{\mathrm{tr}(X)}{\mathrm{tr}(X_0)}, \frac{\|X\|_1}{\|X_0\|_1}\}$ subject to the measurements and a PSD constraint. The dashed green curve in the figure shows the empirical 95% failure boundary, depicting the region of failure with high probability that our results have predicted. It starts off growing linearly with d, when the term rd dominates the term k^2, and then saturates as d grows and the k^2 term (which is a constant in our experiments) becomes dominant.

The penalty function $\max\{\frac{\mathrm{tr}(X)}{\mathrm{tr}(X_0)}, \frac{\|X\|_1}{\|X_0\|_1}\}$ depends on the norm of X_0. In practice the norm of the solution is not known beforehand; a weighted sum of norms is used instead. In Figure 2.7 we examine the performance of the weighted sum of norms penalty in recovery of a rank-1 PSD matrix, for different weights. We pick $\lambda = 0.20$ and $\lambda = 0.35$ for a randomly generated matrix X_0. It can be seen that we get a reasonable result which is comparable to the performance of $\max\{\frac{\mathrm{tr}(X)}{\mathrm{tr}(X_0)}, \frac{\|X\|_1}{\|X_0\|_1}\}$.

In addition, we consider the *amount of error* in the recovery when the program fails. Figure 2.8 shows two curves below which we get a 90% failure, where for the upper (green)

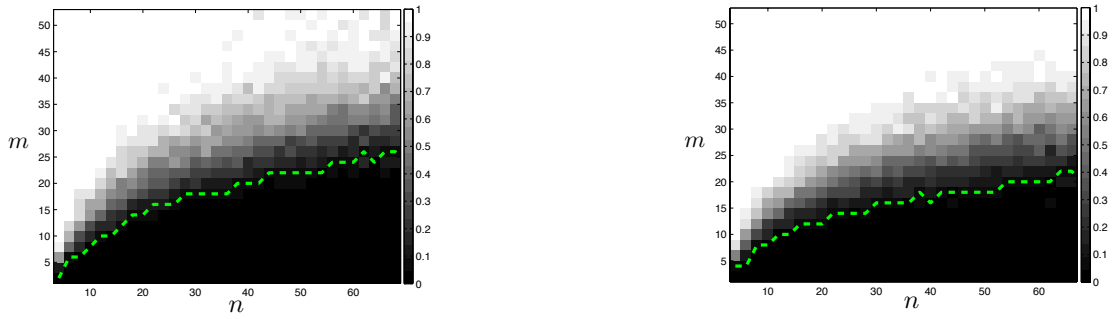


Figure 2.7: Performance of the recovery program minimizing $\mathrm{tr}(X) + \lambda\|X\|_1$ with a PSD constraint, for $\lambda = 0.2$ (left) and $\lambda = 0.35$ (right).

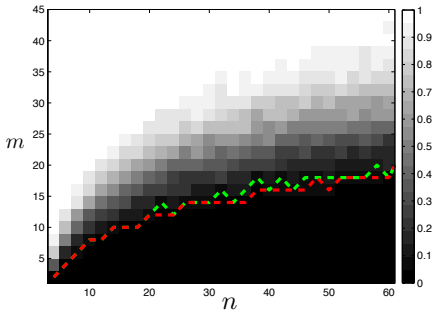


Figure 2.8: 90% frequency of failure where the threshold of recovery is 10^{-4} for the green (upper) and 0.05 for the red (lower) curve. $\max\{\frac{\mathrm{tr}(X)}{\mathrm{tr}(X_0)}, \frac{\|X\|_1}{\|X_0\|_1}\}$ is minimized subject to the PSD constraint and the measurements.

curve the normalized error threshold for declaring failure is 10^{-4}, and for the lower (red) curve it is a larger value of 0.05. We minimize $\max\{\frac{\mathrm{tr}(X)}{\mathrm{tr}(X_0)}, \frac{\|X\|_1}{\|X_0\|_1}\}$ as the objective. We observe that when the recovery program has an error, it is very likely that this error is large, as the curves for 10^{-4} and 0.05 almost overlap. Thus, when the program fails, it fails badly. This observation agrees with intuition from similar problems in compressed sensing where sharp phase transition is observed.

As a final comment, observe that, in Figures 2.6, 2.7 and 2.8 the required amount of

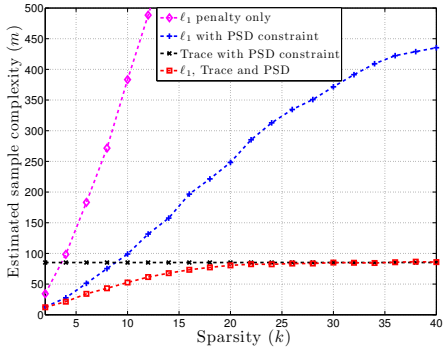


Figure 2.9: We compare sample complexities of different approaches for a rank 1, 40×40 matrix as function of sparsity. The sample complexities were estimated by a search over m, where we chose the m with success rate closest to 50% (over 100 iterations).

measurements slowly increases even when d is large and $k^2 = 64$ is the dominant constant term. While this is consistent with our lower bound of $\Omega(k^2, d)$, the slow increase for constant k, can be explained by the fact that, as d gets larger, sparsity becomes the dominant structure and ℓ_1 minimization by itself requires $\mathrm{O}(k^2 \log \frac{d}{k})$ measurements rather than $\mathrm{O}(k^2)$. Hence for large d, the number of measurements can be expected to grow logarithmically in d.

In Figure 2.9, we compare the estimated phase transition points for different approaches for varying sparsity levels. The algorithms we compare are,

- Minimize ℓ_1 norm,
- Minimize ℓ_1 norm subject to the positive-semidefinite constraint,
- Minimize trace norm subject to the positive-semidefinite constraint,
- Minimize $\max\{\frac{\mathrm{tr}(X)}{\mathrm{tr}(X_0)}, \frac{\|X\|_1}{\|X_0\|_1}\}$ subject to the positive-semidefinite constraint.

Not surprisingly, the last option outperforms the rest in all cases. On the other hand, its performance is highly comparable to the minimum of the second and third approaches.

For all regimes of sparsity, we observe that, measurements required by the last method is at least half as much as the minimum of the second and third methods.

2.8 Discussion

We have considered the problem of recovery of a simultaneously structured object from limited measurements. It is common in practice to combine known norm penalties corresponding to the individual structures (also known as regularizers in statistics and machine learning applications), and minimize this combined objective in order to recover the object of interest. The common use of this approach motivated us to analyze its performance, in terms of the smallest number of generic measurements needed for correct recovery. We showed that, under a certain assumption on the norms involved, the combined penalty requires more generic measurements than one would expect based on the degrees of freedom of the desired object. Our lower bounds on the required number of measurements implies that the combined norm penalty cannot perform significantly better than the best individual norm.

These results raise several interesting questions, and lead to directions for future work. We briefly outline some of these directions, as well as connections to some related problems in Chapter 5.

Chapter 3

VARIATIONAL GRAM FUNCTIONS

In this chapter, we propose a new class of convex penalty functions, called *variational Gram functions* (VGFs), that can promote pairwise relations, such as orthogonality, among a set of vectors in a vector space. These functions can serve as regularizers in convex optimization problems arising from hierarchical classification, multitask learning, and estimating vectors with disjoint supports, among other applications. We study necessary and sufficient conditions under which a VGF is convex, and give a characterization of its subdifferential. We show how to compute its proximal operator, and discuss efficient optimization algorithms for regularized loss minimization problems where the loss admits a simple variational representation and the regularizer is a VGF. We also establish a general representer theorem for such learning problems. Lastly, numerical experiments on a hierarchical classification problem are presented to demonstrate the effectiveness of VGFs and the associated optimization algorithms.

3.1 Introduction

Let $x_1, \ldots, x_m$ be vectors in $\mathbb{R}^n$. It is well known that their pairwise inner products $x_i^T x_j$, for $i, j = 1, \ldots, m$, reveal essential information about their relative orientations, and can serve as a measure for various properties such as orthogonality. In this chapter, we consider a class of functions that aggregate the pairwise inner products in a variational form,

$$\Omega_{\mathcal{M}}(x_1, \ldots, x_m) = \max_{M \in \mathcal{M}} \sum_{i,j=1}^{m} M_{ij} x_i^T x_j \,, \tag{3.1}$$

where $\mathcal{M}$ is a compact subset of the set of m by m symmetric matrices. Let $X = [x_1 \ \cdots \ x_m]$ be an $n \times m$ matrix. Then the pairwise inner products $x_i^T x_j$ are the entries of the Gram

matrix X^TX and the function above can be written as

$$\Omega_{\mathcal{M}}(X) = \max_{M\in\mathcal{M}} \langle X^TX, M\rangle = \max_{M\in\mathcal{M}} \operatorname{tr}(XMX^T)\,, \tag{3.2}$$

where $\langle A, B\rangle = \operatorname{tr}(A^TB)$ denotes the matrix inner product. We call $\Omega_{\mathcal{M}}$ a *variational Gram function* (VGF) of the vectors $x_1, \ldots, x_m$ induced by the set $\mathcal{M}$. If the set $\mathcal{M}$ is clear from the context, we may write $\Omega(X)$ to simplify notation.

As an example, consider the case where $\mathcal{M}$ is given by a box constraint,

$$\mathcal{M} = \left\{M : \ |M_{ij}| \le \overline{M}_{ij},\ i,j = 1,\ldots,m\right\}, \tag{3.3}$$

where $\overline{M}$ is a symmetric nonnegative matrix. In this case, the maximization in the definition of $\Omega_{\mathcal{M}}$ picks either $M_{ij} = \overline{M}_{ij}$ or $M_{ij} = -\overline{M}_{ij}$ depending on the sign of $x_i^Tx_j$, for all $i,j = 1,\ldots,m$ (if $x_i^Tx_j = 0$, the choice is arbitrary). Therefore,

$$\Omega_{\mathcal{M}}(X) = \max_{M\in\mathcal{M}} \sum_{i,j=1}^{m} M_{ij}x_i^Tx_j = \sum_{i,j=1}^{m} \overline{M}_{ij}|x_i^Tx_j|\,. \tag{3.4}$$

In other words, $\Omega_{\mathcal{M}}(X)$ is the weighted sum of the absolute values of the pairwise inner products. This function was proposed in [169] as a regularization function to promote orthogonality between linear classifiers in the context of hierarchical classification.

An important question from both the theoretical and algorithmic points of view is: what are the conditions on $\mathcal{M}$ so that a VGF is convex? Observe that the function $\operatorname{tr}(XMX^T)$ is a convex quadratic function of X if M is positive semidefinite. As a result, the variational form $\Omega_{\mathcal{M}}(X)$ is convex if $\mathcal{M}$ is a subset of the positive semidefinite cone $\mathbb{S}_+^m$, because then it is the pointwise maximum of a family of convex functions indexed by $M \in \mathcal{M}$ (see, e.g., [134, Theorem 5.5]). However, this is not a necessary condition. For example, the set $\mathcal{M}$ in (3.3) is not a subset of $\mathbb{S}_+^m$ unless $\overline{M} = 0$, but the VGF in (3.4) is convex provided that the *comparison matrix* of $\overline{M}$ (derived by negating the off-diagonal entries) is positive semidefinite [169]. In this chapter, we give more careful analysis of the conditions for a VGF to be convex, and characterize its subdifferential and associated proximal operator.

Given a convex VGF, we can define a semi-norm[1] by taking its square root as

$$\|X\|_{\mathcal{M}} := \sqrt{\Omega_{\mathcal{M}}(X)} = \max_{M \in \mathcal{M}} \left(\sum_{i,j=1}^{m} M_{ij} x_i^T x_j \right)^{1/2}. \tag{3.5}$$

If $\mathcal{M} \subset \mathbb{S}_+^m$, then $\|X\|_{\mathcal{M}}$ is the pointwise maximum of the semi-norms $\|XM^{1/2}\|_F$ over all $M \in \mathcal{M}$.

VGFs and the associated norms can serve as penalties or regularization functions in optimization problems to promote certain pairwise properties among a set of vector variables (such as orthogonality in the above example). In this chapter, we consider optimization problems of the form

$$\underset{X \in \mathbb{R}^{n \times m}}{\text{minimize}} \quad \mathcal{L}(X) + \lambda\, \Omega_{\mathcal{M}}(X)\,, \tag{3.6}$$

where $\mathcal{L}(X)$ is a convex loss function of the variable $X = [x_1 \ \cdots \ x_m]$, $\Omega(X)$ is a convex VGF, and $\lambda > 0$ is a parameter to trade off the relative importance of these two functions. We will focus on problems where $\mathcal{L}(X)$ is smooth or has an explicit variational structure, and show how to exploit the structure of $\mathcal{L}(X)$ and $\Omega(X)$ together to derive efficient optimization algorithms.

Organization In Section 3.2, we give more examples of VGFs and explain the connections with functions of Euclidean distance matrices and robust optimization. Section 3.3 studies the convexity of VGFs and their conjugates, semidefinite representability, corresponding norms and their subdifferentials. Their proximal operators are derived in Section 3.4. In Section 3.5, we study a class of structured loss minimization problems with VGF penalities, and show how to exploit their structure using the mirror-prox algorithm. Finally, in Section 3.6, we present a numerical experiment on hierarchical classification to illustrate the application of VGFs.

---

[1] a semi-norm satisfies all the properties of a norm except definiteness; i.e. it can have a zero value for a nonzero input.

Notation We use $\mathbb{S}^m$ to denote the set of symmetric matrices in $\mathbb{R}^{m\times m}$, and $\mathbb{S}^m_+ \subset \mathbb{S}^m$ is the cone of positive semidefinite (PSD) matrices. The symbol $\preceq$ represents the Loewner partial order and $\langle\cdot,\cdot\rangle$ denotes the inner product. We use capital letters for matrices and bold lower case letters for vectors. We use $X \in \mathbb{R}^{n\times m}$ and $x = \mathrm{vec}(X) \in \mathbb{R}^{nm}$ interchangeably, with x_i denoting the ith column of X; i.e., $X = [x_1 \ \cdots \ x_m]$. We use $\mathbf{1}$ and $\mathbf{0}$ to denote matrices or vectors of all ones and all zeros respectively, whose sizes would be clear from the context. The entry-wise absolute value of X is denoted by $|X|$. We use $\|\cdot\|_p$ to denote the ℓ_p norm of the input vector or matrix, and $\|\cdot\|_F$ as the Frobenius norm (similar to ℓ_2 vector norm). The convex conjugate of a function f is defined as $f^\star(y) = \sup_x \langle x, y\rangle - f(x)$, and the dual norm of $\|\cdot\|$ is defined as $\|y\|^\star = \sup\{\langle x, y\rangle : \ \|x\| \le 1\}$. Finally argmin (argmax) returns an optimal point to a minimization (maximization) program while Arg min (or Arg max) is the set of all optimal points. The operator diag$(\cdot)$ is used to put a vector on the diagonal of a zero matrix of corresponding size, extract the diagonal entries of a matrix as a vector, or zeroing out the off-diagonal entries of a matrix. We use $f \equiv g$ to denote $f(x) = g(x)$ for all $x \in \mathrm{dom}(f) = \mathrm{dom}(g)$.

3.2 Examples and Connections

In this section, we present examples of VGFs corresponding to different choices of the set $\mathcal{M}$. The list includes some well known functions that can be expressed in the variational form of (3.1), as well as some new examples.

Vector norms Any vector norm $\|\cdot\|$ on $\mathbb{R}^m$ is the square root of a VGF defined by $\mathcal{M} = \{uu^T : \ \|u\|^\star \le 1\}$. For a column vector $x \in \mathbb{R}^m$, the VGF is given by

$$\Omega_{\mathcal{M}}(x^T) = \max_u \{\mathrm{tr}(x^T uu^T x) : \ \|u\|^\star \le 1\} = \max_u \{(x^T u)^2 : \ \|u\|^\star \le 1\} = \|x\|^2.$$

As another example for when $n = 1$, consider the case where $\mathcal{M}$ is a compact convex set of diagonal matrices with positive diagonal entries. The corresponding VGF (and norm) is

defined as

$$\Omega_{\mathcal{M}}(x^T) = \max_{\theta \in \mathrm{diag}(\mathcal{M})} \sum_{i=1}^{m} \theta_i x_i^2 = \|x\|^2, \tag{3.7}$$

and the dual norm can be expressed as $(\|x\|^\star)^2 = \inf_{\theta \in \mathrm{diag}(\mathcal{M})} \sum_{i=1}^m \frac{1}{\theta_i} x_i^2$. This norm and its dual were first introduced in [111], in the context of regularization for structured sparsity, and later discussed in [9]. The k-support norm [7], which is a norm used to encourage vectors to have k or fewer nonzero entries, is a special case of the dual norm given above, corresponding to $\mathcal{M} = \{\mathrm{diag}(\theta) : \ 0 \leq \theta_i \leq 1 \,, \ \mathbf{1}^T\theta = k\}$.

Norms of the Gram matrix Given a symmetric nonnegative matrix $\overline{M}$, we can define a class of VGFs based on any norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$. Consider

$$\mathcal{M} = \{K \circ \overline{M} : \ \|K\|^\star \leq 1, \ K^T = K\}, \tag{3.8}$$

where $\circ$ represents the matrix Hadamard product, i.e., $(K \circ \overline{M})_{ij} = K_{ij}\overline{M}_{ij}$ for all i, j. Then we have

$$\begin{aligned}\Omega_{\mathcal{M}}(X) &= \max_{M \in \mathcal{M}} \langle M, X^T X\rangle = \max_{\|K\|^\star \leq 1} \langle K \circ \overline{M}, X^T X\rangle \\ &= \max_{\|K\|^\star \leq 1} \langle K, \overline{M} \circ (X^T X)\rangle = \|\overline{M} \circ (X^T X)\| \,.\end{aligned}$$

The following are several concrete examples.

1. If we let $\|\cdot\|^*$ in (3.8) be the ℓ_∞ norm, then $\mathcal{M} = \{M \ : \ |M_{ij}/\overline{M}_{ij}| \leq 1, \ i, j = 1, \ldots, m\}$, which is the same as in (3.3). Here we use the convention $0/0 = 0$, thus $M_{ij} = 0$ whenever $\overline{M}_{ij} = 0$. In this case, we obtain the VGF in (3.4):

$$\Omega_{\mathcal{M}}(X) = \|\overline{M} \circ (X^T X)\|_1 = \sum_{i,j=1}^{m} \overline{M}_{ij} |x_i^T x_j|$$

2. If we use the ℓ_2 norm in (3.8), then $\mathcal{M} = \big\{M \ : \ \sum_{i,j}^m (M_{ij}/\overline{M}_{ij})^2 \leq 1\big\}$. In this case, we have

$$\Omega_{\mathcal{M}}(X) = \|\overline{M} \circ (X^T X)\|_F = \big(\sum_{i,j=1}^{m} (\overline{M}_{ij} x_i^T x_j)^2\big)^{1/2}. \tag{3.9}$$

This function has been considered in multi-task learning [139], and also in the context of super-saturated designs [22, 49].

3. We can use the ℓ_1 norm in (3.8) to define $\mathcal{M} = \left\{M \ : \ \sum_{i,j}^{m} |M_{ij}/\overline{M}_{ij}| \leq 1\right\}$, which results in

$$\Omega_{\mathcal{M}}(X) = \|\overline{M} \circ (X^T X)\|_\infty = \max_{i,j=1,\ldots,m} \overline{M}_{ij} |x_i^T x_j| \,. \tag{3.10}$$

This case can also be traced back to [22] in the statistics literature, where the maximum of $|x_i^T x_j|$ for $i \neq j$ is used as the measure to choose among supersaturated designs.

Many other interesting examples can be constructed this way. For example, one can model *sharing* vs *competition* using group-ℓ_1 norm of the Gram matrix which was considered in vision tasks [90]. We will revisit the above examples to discuss their convexity conditions in Section 3.3.

Spectral functions From the definition, the value of a VGF is invariant under left-multiplication of X by an orthogonal matrix, but this is not true for right multiplication. Hence, VGFs are *not* functions of singular values (e.g. see [100]) in general, and are functions of the row space of X as well. This also implies that in general $\Omega(X) \not\equiv \Omega(X^T)$. However, if the set $\mathcal{M}$ is closed under left and right multiplication by orthogonal matrices, then $\Omega_{\mathcal{M}}(X)$ becomes a function of squared singular values of X. For any matrix $M \in \mathbb{S}^m$, denote the sorted vector of its singular values by $\sigma(M)$ and let $\Theta = \{\sigma(M) : \ M \in \mathcal{M}\}$. Then we have

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \ \mathrm{tr}(XMX^T) = \max_{\theta \in \Theta} \ \sum_{i=1}^{\min(n,m)} \theta_i \, \sigma_i(X)^2 \,, \tag{3.11}$$

as a result of Von Neumann's trace inequality [113]. Note the similarity of the above to the VGF in (3.7). As an example, consider

$$\mathcal{M} = \{M : \ \alpha_1 I \preceq M \preceq \alpha_2 I, \ \ \mathrm{tr}(M) = \alpha_3\}, \tag{3.12}$$

where $0 < \alpha_1 < \alpha_2$ and $\alpha_3 \in [m\alpha_1,\ m\alpha_2]$ are given constants. The so called *spectral box-norm* [108] is the dual to the norm of the form (3.5) defined via this $\mathcal{M}$. Note that in this case, $\mathcal{M} \subset \mathbb{S}^m_+$, so $\Omega_{\mathcal{M}}$ is convex. The square of this norm has been considered in [85] for clustered multitask learning where it is presented as a convex relaxation for k-means.

Finite set $\mathcal{M}$ For a finite set $\mathcal{M} = \{M_1, \ldots, M_p\} \subset \mathbb{S}^m_+$, the VGF is given by

$$\Omega_{\mathcal{M}}(X) = \max_{i=1,\ldots,p} \|XM_i^{1/2}\|_F^2,$$

which is the pointwise maximum of a finite number of squared weighted Frobenius norms.

In the following subsections, we explore other interpretations of a VGF that show its effect in promoting diversity, its connection to Euclidean distance matrices, and give a robust optimization interpretation.

3.2.1 Diversification

VGFs can be used for *diversifying* the columns of the input matrix; e.g., minimizing (3.4) pushes to zero the inner products $x_i^T x_j$ corresponding to the nonzero entries in $\overline{M}$ as much as possible. As another example, observe that two non-negative vectors have disjoint supports if and only if they are orthogonal to each other. Hence, using a VGF as (3.4), $\Omega_{\mathcal{M}}(X) = \sum_{i,j=1}^m \overline{M}_{ij}|x_i^T x_j|$, that promotes orthogonality, we can define

$$\Psi(X) = \Omega_{\mathcal{M}}(|X|) \tag{3.13}$$

to promote disjoint supports among the columns of X; hence diversifying the supports of columns of X. Convexity of (3.13) is discussed in Section 3.3.6. Different approaches has been used in machine learning applications for promoting diversity; e.g., see[105, 97, 84] and references therein.

3.2.2 Functions of Euclidean distance matrix

Consider a set $\mathcal{M} \subset \mathbb{S}^m$ with the property that $M\mathbf{1} = \mathbf{0}$ for all $M \in \mathcal{M}$. For every $M \in \mathcal{M}$, let $A = \mathrm{diag}(M) - M$ and observe that

$$\mathrm{tr}(XMX^T) = \sum_{i,j=1}^{m} M_{ij} x_i^T x_j = \tfrac{1}{2} \sum_{i,j=1}^{m} A_{ij} \|x_i - x_j\|_2^2 \,.$$

This allows us to express the associated VGF as a function of the *Euclidean distance matrix* D, which is defined by $D_{ij} = \frac{1}{2}\|x_i - x_j\|_2^2$ for $i, j = 1, \ldots, m$ (see, e.g., [25, Section 8.3]). Let $\mathcal{A} = \{\mathrm{diag}(M) - M : \ M \in \mathcal{M}\}$. Then we have

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \ \mathrm{tr}(XMX^T) = \max_{A \in \mathcal{A}} \ \langle A, D \rangle \,.$$

A sufficient condition for the above function to be convex in X is that each $A \in \mathcal{A}$ is entrywise nonnegative, which implies that the corresponding $M = \mathrm{diag}(A\mathbf{1}) - A$ is diagonally dominant with nonnegative diagonal elements, hence positive semidefinite. However, this is not a necessary condition and the function can be convex without all A's being entrywise nonegative. In Section 3.3 we will discuss more general conditions for convexity of VGFs. See [64] and references therein for applications of this VGF.

3.2.3 Connection with robust optimization

The VGF-regularized loss minimization problem has the following connection to robust optimization (see, e.g., [18]): the optimization program

$$\underset{X}{\text{minimize}} \ \max_{M \in \mathcal{M}} \ \left\{ \mathcal{L}(X) + \mathrm{tr}(XMX^T) \right\}$$

can be interpreted as seeking an X with minimal worst-case value over an uncertainty set $\mathcal{M}$. Alternatively, when $\mathcal{M} \subset \mathbb{S}_+^m$, this can be viewed as a problem with Tikhonov regularization $\|XM^{1/2}\|_F^2$ where the weight matrix $M^{1/2}$ is subject to errors characterized by the set $\mathcal{M}$.

3.3 Convex Analysis of VGF

In this section, we study the convexity of VGFs, their conjugate functions and subdifferentials.

First, we review some basic properties. Notice that $\Omega_{\mathcal{M}}$ is the *support function* of the set $\mathcal{M}$ at the Gram matrix X^TX; i.e.,

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \operatorname{tr}(XMX^T) = \sigma_{\mathcal{M}}(X^TX)$$

where the support function of a set $\mathcal{M}$ is defined as $\sigma_{\mathcal{M}}(Y) = \sup_{M \in \mathcal{M}} \langle M, Y \rangle$; see, e.g., [134, Section 13]. By properties of the support function (see [134, Section 15]), we have

$$\Omega_{\mathcal{M}} \equiv \Omega_{\operatorname{conv}(\mathcal{M})},$$

where $\operatorname{conv}(\mathcal{M})$ denotes the convex hull of $\mathcal{M}$. It is clear that the representation of a VGF (i.e., the associated set $\mathcal{M}$) is not unique. Henceforth, without loss of generality we assume $\mathcal{M}$ is convex unless explicitly noted otherwise. Also, for simplicity we assume $\mathcal{M}$ is a compact set, while all we need is that the maximum in (3.1) is attained. For example, a non-compact $\mathcal{M}$ that is unbounded along any negative semidefinite direction is allowed.

Moreover, VGFs are left unitarily invariant; for any $Y \in \mathbb{R}^{n \times m}$ and any orthogonal matrix $U \in \mathbb{R}^{n \times n}$, where $UU^T = U^TU = I$, we have $\Omega(Y) = \Omega(UY)$ and $\Omega^*(Y) = \Omega^*(UY)$; use the definitions in (3.2) and (3.17). We use this property in simplifying computations involving VGFs (such as proximal mapping calculations in Section 3.4) as well as in establishing a general kernel trick and representer theorem in Section 3.5.2.

As we mentioned in the introduction, a sufficient condition for the convexity of a VGF is that $\mathcal{M} \subset \mathbb{S}^m_+$. The following theorem gives a necessary and sufficient condition for a VGF to be convex. Basically, it only requires the VGF to admit a representation as in (3.2) with a set of PSD matrices.

Theorem 33 *Suppose that $\mathcal{M}$ is compact. Then $\Omega_{\mathcal{M}}$ is convex if and only if for every X there exists an $M \in \mathcal{M} \cap \mathbb{S}_+$ that achieves the maximum value in the definition of $\Omega_{\mathcal{M}}(X)$. In other words, $\Omega_{\mathcal{M}}$ is convex if and only if $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M} \cap \mathbb{S}_+}$.*

The above theorem means that a convex VGF is essentially the point-wise maximum of a family of squared weighted Frobenius norms. We postpone the proof (which uses conjugate functions) until after Lemma 38.

While Theorem 33 gives a necessary and sufficient condition for the convexity of VGFs, it does not provide an effective procedure to check whether or not such a condition holds. In Section 3.3.1, we discuss more concrete conditions for determining convexity when the set $\mathcal{M}$ is a polytope. In Section 3.3.2, we describe a more tangible sufficient condition for general sets.

3.3.1 Convexity with polytope $\mathcal{M}$

Consider the case where $\mathcal{M}$ is a polytope with p vertices, i.e., $\mathcal{M} = \text{conv}\{M_1, \dots, M_p\}$. The support function of this set is given as $\sigma_{\mathcal{M}}(Y) = \max_{i=1,\dots,p} \langle Y, M_i \rangle$ and is piecewise linear [137, Section 8.E]. We define $\mathcal{M}_{\text{eff}}$ as a subset of $\{M_1, \dots, M_p\}$ with smallest possible size satisfying $\sigma_{\mathcal{M}}(X^T X) = \sigma_{\mathcal{M}_{\text{eff}}}(X^T X)$ for all $X \in \mathbb{R}^{n \times m}$.

As an example, for $\mathcal{M} = \{M : \ |M_{ij}| \leq \overline{M}_{ij}, \ i, j = 1, \dots, m\}$ which gives the function defined in (3.4), we have

$$\mathcal{M}_{\text{eff}} \subseteq \{M : \ M_{ii} = \overline{M}_{ii} \ , \ M_{ij} = \pm \overline{M}_{ij} \text{ for } i \neq j \} . \tag{3.14}$$

Theorem 34 *For a polytope* $\mathcal{M} \subset \mathbb{R}^{m \times m}$*, the associated VGF is convex if and only if* $\mathcal{M}_{\text{eff}} \subset \mathbb{S}^m_+$.

Proof. Obviously, $\mathcal{M}_{\text{eff}} \subset \mathbb{S}^m_+$ ensures convexity of $\max_{M \in \mathcal{M}_{\text{eff}}} \text{tr}(XMX^T) = \Omega_{\mathcal{M}}(X)$. Next, we prove necessity for any $\mathcal{M}_{\text{eff}}$. Take any $M_i \in \mathcal{M}_{\text{eff}}$. If for every $X \in \mathbb{R}^{n \times m}$ with $\Omega(X) = \text{tr}(XM_iX^T)$ there exists another $M_j \in \mathcal{M}_{\text{eff}}$ with $\Omega(X) = \text{tr}(XM_jX^T)$, then $\mathcal{M}_{\text{eff}} \backslash \{M_i\}$ is an effective subset of $\mathcal{M}$ which contradicts the minimality of $\mathcal{M}_{\text{eff}}$. Hence, there exists X_i such that $\Omega(X_i) = \text{tr}(X_i M_i X_i^T) > \text{tr}(X_i M_j X_i^T)$ for all $j \neq i$. Hence, Ω is twice continuously differentiable in a small neighborhood of X_i with Hessian $\nabla^2 \Omega(\text{vec}(X_i)) = M_i \otimes I_n$, where $\otimes$ denotes the matrix Kronecker product. Since Ω is assumed to be convex, the Hessian has to be PSD which gives $M_i \succeq \mathbf{0}$. ∎

The definition of $\mathcal{M}_{\text{eff}}$ requires $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}_{\text{eff}}}$, and the condition in Theorem 34 is $\mathcal{M}_{\text{eff}} \subset \mathbb{S}^m_+$. Comparing with Theorem 33, here we have $\mathcal{M}_{\text{eff}} \subset \mathcal{M} \cap \mathbb{S}^m_+$, which can be a strict inclusion

(even $\text{conv}(\mathcal{M}_{\text{eff}})$ can be a strict subset of $\mathcal{M} \cap \mathbb{S}_+^m$). Next we give a few examples to illustrate the use of Theorem 34.

1. We begin with the example defined in (3.4). Authors in [169] provided the necessary (when $n \geq m-1$) and sufficient condition for convexity using results from M-matrix theory: First, define the comparison matrix $\widetilde{M}$ associated to the nonnegative matrix $\overline{M}$ as $\widetilde{M}_{ii} = \overline{M}_{ii}$ and $\widetilde{M}_{ij} = -\overline{M}_{ij}$ for $i \neq j$. Then $\Omega_{\mathcal{M}}$ is convex if $\widetilde{M}$ is positive semidefinite, and this condition is also necessary when $n \geq m-1$ [169]. Theorem 34 provides an alternative and more general proof. Let $\lambda_{\min}(M)$ be the minimum eigenvalue of a symmetric matrix M. From the characterization of $\mathcal{M}_{\text{eff}}$ in (3.14), we have

$$\begin{aligned} \min_{M \in \mathcal{M}_{\text{eff}}} \lambda_{\min}(M) &= \min_{\substack{M \in \mathcal{M}_{\text{eff}} \\ \|z\|_2 = 1}} z^T M z \;\geq\; \min_{\|z\|_2=1} \sum_i \overline{M}_{ii} z_i^2 - \sum_{i \neq j} \overline{M}_{ij} |z_i z_j| \\ &= \min_{\|z\|_2=1} |z|^T \widetilde{M} |z| \;\geq\; \lambda_{\min}(\widetilde{M}). \end{aligned} \tag{3.15}$$

   When $n \geq m-1$, one can construct $X \in \mathbb{R}^{n \times m}$ such that all off-diagonal entries of $X^T X$ are negative (see the example in Appendix A.2 of [169]). On the other hand, Lemma 2.1(2) of [30] states that the existence of such a matrix implies $n \geq m-1$. Hence, $\widetilde{M} \in \mathcal{M}_{\text{eff}}$ if and only if $n \geq m-1$. Therefore, both inequalities in (3.15) should hold with equality, which means that $\mathcal{M}_{\text{eff}} \subset \mathbb{S}_+^m$ if and only if $\widetilde{M} \succeq 0$. By Theorem 34, this is equivalent to the VGF in (3.4) being convex. If $n < m-1$, then $\widetilde{M}$ may not belong to $\mathcal{M}_{\text{eff}}$, thus $\widetilde{M} \succeq 0$ is only a "sufficient" condition for convexity for general n.

2. Similar to the set $\mathcal{M}$ above, consider a box that is not necessarily symmetric around the origin. More specifically, let $\mathcal{M} = \{M \in \mathbb{S}^m : \; M_{ii} = C_{ii}\,, \; |M - C| \leq D\}$ where C (denoting the center) is a symmetric matrix with zero diagonal, and D is a symmetric nonnegative matrix. In this case, we have $\mathcal{M}_{\text{eff}} \subseteq \{M : \; M_{ii} = D_{ii}\,, \; M_{ij} = C_{ij} \pm D_{ij} \text{ for } i \neq j\}$. When used as a penalty function in applications, this can capture the prior information that when $x_i^T x_j$ is not zero, a particular range of acute or obtuse angles (depending on the sign of C_{ij}) between the vectors is preferred. Similar to (3.15),

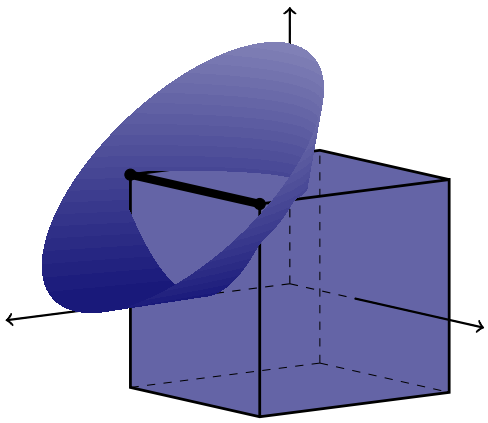


Figure 3.1: The set in (3.3) (defined by $\overline{M} = [1,\ 0.8;\ 0.8,\ 1]$) and the cone of positive semidefinite matrices where 2×2 symmetric matrices are embedded into $\mathbb{R}^3$. The thick edge of the cube is the set of all points with the same diagonal elements as $\overline{M}$ (see (3.40)), and the two endpoints constitute $\mathcal{M}_{\text{eff}}$. Positive semidefiniteness of $\widetilde{M}$ is a necessary and sufficient condition for the convexity of $\Omega_{\mathcal{M}} : \mathbb{R}^{n \times 2} \to \mathbb{R}$ for all $n \geq m - 1 = 1$.

we have

$$\min_{M\in\mathcal{M}_{\text{eff}}} \lambda_{\min}(M) \geq \min_{\|z\|_2=1} |z|^T \widetilde{D}|z| + z^T C z \geq \lambda_{\min}(\widetilde{D}) + \lambda_{\min}(C),$$

where $\widetilde{D}$ is the comparison matrix associated to D. Note that C has zero diagonals and cannot be PSD. Hence, a sufficient condition for convexity of $\Omega_{\mathcal{M}}$ defined by an asymmetric box is that $\lambda_{\min}(\widetilde{D}) + \lambda_{\min}(C) \geq 0$.

3. Consider the VGF defined in (3.10), whose associated variational set is

$$\mathcal{M} = \{M \in \mathbb{S}^m : \sum_{(i,j):\, \overline{M}_{ij}\neq 0} |M_{ij}/\overline{M}_{ij}| \leq 1\,,\ M_{ij} = 0 \text{ if } \overline{M}_{ij} = 0\}, \tag{3.16}$$

where $\overline{M}$ is a symmetric nonnegative matrix. Vertices of $\mathcal{M}$ are matrices with either only one nonzero value $\overline{M}_{ii}$ on the diagonal, or two nonzero off-diagonal entries at (i,j) and (j,i) equal to $\frac{1}{2}\overline{M}_{ij}$ or $-\frac{1}{2}\overline{M}_{ij}$. The second type of matrices cannot be PSD as their diagonal is zero, and according to Theorem 34, convexity of $\Omega_{\mathcal{M}}$ requires these vertices do not belong to $\mathcal{M}_{\text{eff}}$. Therefore, the matrices in $\mathcal{M}_{\text{eff}}$ should be diagonal. Hence, a convex VGF corresponding to the set (3.16) has the form $\Omega(X) = \max_{i=1,\ldots,m} \overline{M}_{ii}\|x_i\|_2^2$. To ensure such a description for $\mathcal{M}_{\text{eff}}$ we need $\max\{\overline{M}_{ii}\|x_i\|_2^2, \overline{M}_{jj}\|x_j\|_2^2\} \geq \overline{M}_{ij}|x_i^T x_j|$ for all i, j and any $X \in \mathbb{R}^{n\times m}$, which is equivalent to $\overline{M}_{ii}\overline{M}_{jj} \geq \overline{M}_{ij}^2$ for all i, j. This is satisfied if $\overline{M} \succeq \mathbf{0}$.

3.3.2 A spectral sufficient condition

As mentioned before, it is generally not clear how to provide easy-to-check necessary and sufficient convexity guarantees for the case of non-polytope sets $\mathcal{M}$. However, simple sufficient conditions can be easily checked for certain classes of sets $\mathcal{M}$, for example spectral sets (Lemma 35). We first provide an example and consider a specialized approach to establish convexity, which illustrates the advantage of a simple guarantee as the one in Lemma 35.

1. Consider the VGF defined in (3.9) and its associated set given in (3.8) when we plug

in the Frobenius norm; i.e.,

$$\mathcal{M} = \{K \circ \overline{M} : \ \|K\|_F \leq 1, \ K^T = K\}.$$

In this case, $\mathcal{M}$ is not a polytope, but we can proceed with a similar analysis as in the previous subsection. In particular, given any $X \in \mathbb{R}^{n\times m}$, the value of $\Omega_{\mathcal{M}}(X)$ is achieved by an optimal matrix $K = (\overline{M} \circ X^T X)/\|\overline{M} \circ X^T X\|_F$. By Theorem 33, $\Omega_{\mathcal{M}}$ is convex provided that $K \circ \overline{M} \succeq 0$, which is equivalent to $\overline{M} \circ \overline{M} \circ X^T X \succeq 0$. Since this should hold for every X, we need $\overline{M} \circ \overline{M} \succeq 0$. The Schur Product Theorem [82, Theorem 7.5.1] states that $\overline{M} \succeq 0$ is sufficient for this requirement to hold, hence it is also a sufficient condition for convexity of $\Omega_{\mathcal{M}}$.

Denote by M_+ the orthogonal projection of a symmetric matrix M onto the PSD cone, which is given by the matrix formed by only positive eigenvalues and their associated eigenvectors of M.

Lemma 35 (a sufficient condition) *$\Omega_{\mathcal{M}}$ is convex provided that for any $M \in \mathcal{M}$ there exists $M' \in \mathcal{M}$ such that $M_+ \preceq M'$.*

Proof. It is easy to see that for any X we have $\mathrm{tr}(XMX^T) \leq \mathrm{tr}(XM_+X^T)$. Therefore,

$$\Omega_{\mathcal{M}}(X) = \max_{M\in\mathcal{M}} \ \mathrm{tr}(XMX^T) \leq \max_{M\in\mathcal{M}} \ \mathrm{tr}(XM_+X^T).$$

On the other hand, the assumption of the lemma gives

$$\max_{M\in\mathcal{M}} \ \mathrm{tr}(XM_+X^T) \leq \max_{M'\in\mathcal{M}} \ \mathrm{tr}(XM'X^T) = \Omega_{\mathcal{M}}(X)$$

which implies that the inequalities have to hold with equality. Now, $\Omega_{\mathcal{M}}(X)$ is convex by Theorem 33. Note that the assumption of the lemma can hold while $\mathcal{M}_+ \not\subseteq \mathcal{M}$. ∎

On the other hand, it is easy to see that the condition in Lemma 35 is not necessary. Consider $\mathcal{M} = \{M \in \mathbb{S}^2 : \ |M_{ij}| \leq 1\}$. Although the associated VGF is convex (because the comparison matrix is PSD), there is no matrix in $M' \in \mathcal{M}$ for which $M' \succeq M$ where

$$M = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \in \mathcal{M}, \quad \text{but} \quad M_+ \simeq \begin{bmatrix} 0.44 & .72 \\ .72 & 1.17 \end{bmatrix}$$

as for any $M' \in \mathcal{M}$ we have $(M' - M_+)_{22} < 0$.

Similar to the proof of Lemma 35, one can check that another sufficient condition for convexity of a VGF is that all of the maximal points of $\mathcal{M}$ with respect to $\mathbb{S}_+$ are PSD.

Lemma 36 *For any $Z \succeq 0$, consider $P = \Pi_{\mathcal{M}}(Z)$ and its Moreau decomposition with respect to the positive semidefinite cone as $P = P_+ - P_-$ where $P_+, P_- \succeq 0$ and $\langle P_+, P_- \rangle = 0$. Then, $P_+ \in \mathcal{M}$ implies $P_- = 0$.*

Proof. Recall the firm nonexpansive property of the projection operator onto a convex set [137] applied to $P = \Pi_{\mathcal{M}}(Z)$ and $P_+ = \Pi_{\mathcal{M}}(P_+)$; where the latter is from the assumption. We have,

$$\|P - P_+\|_F^2 \leq \langle P - P_+, Z - P_+ \rangle \implies \langle P_-, Z \rangle + \|P_-\|_F^2 \leq 0$$

which gives $P_- = 0$ as $\langle Z, P_- \rangle \succeq 0$. ∎

Now, we can state the following corollary.

Corollary 37 *Provided that for any $M \in \mathcal{M}$ we have $M_+ \in \mathcal{M}$, then $\Omega_{\mathcal{M}}$ is convex and $\Pi_{\mathcal{M}}(Z) \succeq 0$ for any $Z \succeq 0$.*

Corollary 37 establishes an interesting property about the iterates of the mirror-prox algorithm in Section 3.5.1, with ℓ_2 norm as the mirror map, as follows. If we initialize M to be a positive semidefinite matrix, all of the iterations stay PSD as we add a PSD matrix to the previous iteration and project it to the PSD cone. Notice that such condition is required for applying the mirror-prox algorithm: the objective has to be convex-concave and the positive semidefiniteness of all iterations guarantees this property.

3.3.3 Conjugate function and proof of Theorem 33

For any function Ω, the conjugate function is defined as $\Omega^\star(Y) = \sup_X \langle X, Y \rangle - \Omega(X)$ and the transformation that maps Ω to $\Omega^\star$ is called the Legendre-Fenchel transform; e.g., see [134, Section 12].

Lemma 38 (conjugate VGF) *Consider a convex VGF associated to a compact convex set* $\mathcal{M}$. *The conjugate function is*

$$\Omega^\star_{\mathcal{M}}(Y) = \tfrac{1}{4} \inf_M \left\{ \mathrm{tr}(YM^\dagger Y^T) \ : \ \mathrm{range}(Y^T) \subseteq \mathrm{range}(M)\,,\ M \in \mathcal{M} \cap \mathbb{S}^m_+ \right\}, \tag{3.17}$$

where $M^\dagger$ *is the Moore-Penrose pseudoinverse of* M.

Note that $\Omega^\star(Y)$ is $+\infty$ if the optimization problem in (3.17) is infeasible; i.e., if $\mathrm{range}(Y^T) \not\subseteq \mathrm{range}(M)$ for all $M \in \mathcal{M} \cap \mathbb{S}^m_+$. This is equivalent to: $Y(I - MM^\dagger)$ is nonzero for all $M \in \mathcal{M} \cap \mathbb{S}^m_+$, where $MM^\dagger$ is the orthogonal projection onto the range of M. This can be seen using generalized Schur complements; e.g., see Appendix A.5.5 in [25] or [27]. *Proof.* Applying the definition of conjugate function to a VGF gives

$$\Omega^\star_{\mathcal{M}}(Y) \ = \ \sup_X \left[\langle X, Y \rangle - \sup_{M \in \mathcal{M}} \mathrm{tr}(XMX^T) \right] \ = \ \sup_X \inf_{M \in \mathcal{M}} \ \langle X, Y \rangle - \mathrm{tr}(XMX^T)\,.$$

Since $\mathcal{M}$ is compact and convex, we can change the order of sup and inf. Note that for any non-PSD $M \in \mathcal{M}$, the maximization with respect to X is unbounded above. With this, and after another change of order for sup and inf, we get

$$\Omega^\star_{\mathcal{M}}(Y) \ = \ \inf_{M \in \mathcal{M} \cap \mathbb{S}^m_+} \ \sup_X \ \langle X, Y \rangle - \mathrm{tr}(XMX^T)\,.$$

Therefore, $\Omega^\star_{\mathcal{M}} \equiv \Omega^\star_{\mathcal{M} \cap \mathbb{S}_+}$. Next we define

$$f_{\mathcal{M}}(Y) \ = \ \tfrac{1}{4} \inf_{M,\,C} \left\{ \mathrm{tr}(C) \ : \ \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq \mathbf{0}\,,\ M \in \mathcal{M} \right\}. \tag{3.18}$$

The positive semidefiniteness constraint implies $M \succeq \mathbf{0}$, therefore $f_{\mathcal{M}} \equiv f_{\mathcal{M} \cap \mathbb{S}_+}$. Its conjugate function is

$$\begin{aligned} f^\star_{\mathcal{M}}(X) \ &= \ \sup_{Y,M,C} \left\{ \langle X, Y \rangle - \tfrac{1}{4}\,\mathrm{tr}(C) \ : \ \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq \mathbf{0}\,,\ M \in \mathcal{M} \right\} \\ &= \ \sup_{Y,M,C} \left\{ \langle X, Y \rangle - \tfrac{1}{4}\,\mathrm{tr}(C) \ : \ \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq \mathbf{0}\,,\ M \in \mathcal{M} \cap \mathbb{S}_+ \right\}. \end{aligned} \tag{3.19}$$

Consider the dual of the optimization program with respect to $Y$and C. Let $W \succeq \mathbf{0}$ be the dual variable with corresponding blocks, and write the Lagrangian as

$$L(Y, C, W) = \langle X, Y \rangle - \tfrac{1}{4}\operatorname{tr}(C) + \langle W_{11}, M \rangle + 2\langle W_{21}, Y \rangle + \langle W_{22}, C \rangle ,$$

whose maximum value is finite only if $W_{21} = -\frac{1}{2}X$ and $W_{22} = \frac{1}{4}I$. Therefore, the dual problem is

$$\min_{W_{11}} \left\{ \langle W_{11}, M \rangle : \begin{bmatrix} W_{11} & -\frac{1}{2}X^T \\ -\frac{1}{2}X & \frac{1}{4}I \end{bmatrix} \succeq \mathbf{0} \right\} = \min_{W_{11}} \left\{ \langle W_{11}, M \rangle : \; W_{11} \succeq X^T X \right\},$$

which is equal to $\langle M, X^T X \rangle$. Plugging in (3.19), we conclude $f^\star_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$.

Next, convexity and lower semi-continuity of $f_{\mathcal{M}}$ imply $f^{\star\star}_{\mathcal{M}} = f_{\mathcal{M}}$ (e.g. [137, Theorem 11.1]). Therefore, $f_{\mathcal{M}}$ is equal to $\Omega^\star_{\mathcal{M}\cap\mathbb{S}_+}$ which we showed to be equal to $\Omega^\star_{\mathcal{M}}$. Taking the generalized Schur complement of the semidefinite constraint in (3.18) gives the desired representation in (3.17). ∎

Following Lemma 38, we are ready to prove Theorem 33.

Proof. [of Theorem 33] We know that $\Omega_{\mathcal{M}\cap\mathbb{S}_+}$ is convex because it is the pointwise maximum of convex quadratic functions parametrized by $M \in \mathcal{M}\cap\mathbb{S}_+$. Therefore, if $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$ then $\Omega_{\mathcal{M}}$ is convex. To prove the other direction, notice that the proof of Lemma 38 shows that $\Omega^\star_{\mathcal{M}} \equiv \Omega^\star_{\mathcal{M}\cap\mathbb{S}_+}$, which in turn gives $\Omega^{\star\star}_{\mathcal{M}} \equiv \Omega^{\star\star}_{\mathcal{M}\cap\mathbb{S}_+}$. Since both $\Omega_{\mathcal{M}}$ and $\Omega_{\mathcal{M}\cap\mathbb{S}_+}$ are proper, lower semi-continuous convex functions, they are equal to their biconjugates [137, Theorem 11.1]. Therefore, $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$. ∎

Recall from convex analysis (e.g., [134, Cor. 12.1.1]) that the convex envelope (also referred to as the convex hull) of a non-convex function (the greatest convex function that is majorized by this function; see [134, pp. 36] for definition) is given by the biconjugate function. As it is apparent from the representation in Lemma 38, biconjugation of a possibly non-convex VGF $\Omega_{\mathcal{M}}$ is equivalent to intersecting $\mathcal{M}$ with the positive semidefinite cone; i.e., $\Omega^{**}_{\mathcal{M}} = \Omega_{\mathcal{M}\cap\mathbb{S}_+}$. Hence, a VGF representation of a function provides a natural way for finding its convex envelope.

3.3.4 Related norms

Given a convex VGF $\Omega_{\mathcal{M}}$, Theorem 33 states that $\Omega_{\mathcal{M}} \equiv \Omega_{\mathcal{M}\cap\mathbb{S}_+}$, which implies

$$\Omega_{\mathcal{M}}(X) = \sup_{M\in\mathcal{M}\cap\mathbb{S}_+} \operatorname{tr}(XMX^T) = \sup_{M\in\mathcal{M}\cap\mathbb{S}_+} \|XM^{1/2}\|_F^2 .$$

This representation shows that $\sqrt{\Omega_{\mathcal{M}}}$ is a semi-norm: absolute homogeneity holds, and it is easy to prove the triangle inequality for the maximum of semi-norms. The next lemma, which can be seen from Corollary 15.3.2 of [134], generalizes this assertion; we provide another proof in the Appendix.

Lemma 39 *Suppose a function $\Omega : \mathbb{R}^{n\times m} \to \mathbb{R}$ is homogeneous of order 2, i.e., $\Omega(\theta X) = \theta^2\Omega(X)$. Then its square root $\|X\| = \sqrt{\Omega(X)}$ is a semi-norm if and only if Ω is convex. If Ω is strictly convex then $\sqrt{\Omega}$ is a norm.*

Dual Norm Considering $\|\cdot\|_{\mathcal{M}} \equiv \sqrt{\Omega_{\mathcal{M}}}$, we have $\frac{1}{2}\Omega_{\mathcal{M}} \equiv \frac{1}{2}\|\cdot\|_{\mathcal{M}}^2$. Taking the conjugate function of both sides yields $2\Omega_{\mathcal{M}}^\star \equiv \frac{1}{2}(\|\cdot\|_{\mathcal{M}}^\star)^2$ where we used the order-2 homogeneity of $\Omega_{\mathcal{M}}$. Therefore, $\|\cdot\|_{\mathcal{M}}^\star \equiv 2\sqrt{\Omega_{\mathcal{M}}^\star}$. Given the representation of $\Omega_{\mathcal{M}}^\star$ in Lemma 38, one can derive a similar representation for $\sqrt{\Omega_{\mathcal{M}}^\star}$ as follows.

Theorem 40 *Consider a convex VGF $\Omega_{\mathcal{M}}$, where $\mathcal{M}$ is a compact convex set. We have*

$$\|Y\|_{\mathcal{M}}^\star = 2\sqrt{\Omega_{\mathcal{M}}^\star(Y)} = \tfrac{1}{2}\inf_{M,C}\left\{\operatorname{tr}(C) + \gamma_{\mathcal{M}}(M) : \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq \mathbf{0}\right\} . \tag{3.20}$$

where $\gamma_{\mathcal{M}}(M) = \inf\{\lambda \geq 0 : \ M \in \lambda\mathcal{M}\}$ is the gauge function associated to the nonempty convex set $\mathcal{M}$.

Proof. The square root function, over positive numbers, can be represented in a variational form as $\sqrt{y} = \min\{\alpha + \frac{y}{4\alpha} : \ \alpha > 0\}$. Without loss of generality, suppose $\mathcal{M}$ is a compact convex set containing the origin. Provided that $\Omega_{\mathcal{M}}^\star(Y) > 0$, from the variational

representation of a conjugate VGF function we have

$$\sqrt{\Omega^\star_{\mathcal{M}}(Y)} = \tfrac{1}{4} \inf_{M,\alpha\geq 0} \left\{\alpha + \tfrac{1}{\alpha}\operatorname{tr}(YM^\dagger Y^T) \;:\; \operatorname{range}(Y^T) \subseteq \operatorname{range}(M)\,,\; M \in \mathcal{M} \cap \mathbb{S}^m_+\right\}$$
$$= \tfrac{1}{4} \inf_{M,\alpha\geq 0} \left\{\alpha + \operatorname{tr}(YM^\dagger Y^T) \;:\; \operatorname{range}(Y^T) \subseteq \operatorname{range}(M)\,,\; M \in \alpha(\mathcal{M} \cap \mathbb{S}^m_+)\right\}$$

where we used $(\alpha M)^\dagger = M^\dagger/\alpha$ and performed a change of variable. The last representation is the same as the one given in the statement of the lemma. On the other hand, when $\Omega^\star_{\mathcal{M}}(Y) = 0$, the claimed representation returns 0 as well because $\mathcal{M}$ contains the origin. ■

As an example, $\mathcal{M} = \{M \succeq \mathbf{0} :\ \operatorname{tr}(M) \leq 1\}$ corresponds to $\gamma_{\mathcal{M}}(M) = \operatorname{tr}(M)$ which if plugged in (3.20) gives the well-known semidefinite representation for nuclear norm.

3.3.5 Subdifferentials

In this section, we characterize the subdifferential of VGFs and their conjugate functions, as well as that of their corresponding norms. Due to the variational definition of a VGF where the objective function is linear in M, and the fact that $\mathcal{M}$ is assumed to be compact, it is straightforward to obtain the subdifferential of $\Omega_{\mathcal{M}}$ (e.g., see [80, Theorem 4.4.2]).

Proposition 41 *The subdifferential of a convex VGF $\Omega_{\mathcal{M}}$ at a matrix X is given by*

$$\partial\,\Omega_{\mathcal{M}}(X) = \operatorname{conv}\left\{2XM :\ \operatorname{tr}(XMX^T) = \Omega(X),\ M \in \mathcal{M} \cap \mathbb{S}_+\right\}.$$

For the norm $\|X\|_{\mathcal{M}} \equiv \sqrt{\Omega_{\mathcal{M}}}$, we have $\partial\|X\|_{\mathcal{M}} = \frac{1}{2\|X\|_{\mathcal{M}}}\partial\,\Omega_{\mathcal{M}}(X)$ if $\Omega_{\mathcal{M}}(X) \neq 0$.

As an example, the subdifferential of $\Omega(X) = \sum_{i,j=1}^m \overline{M}_{ij}|x_i^T x_j|$, defined in (3.4), is given by

$$\partial\,\Omega(X) = \{2XM :\ M_{ij} = \overline{M}_{ij}\operatorname{sign}(x_i^T x_j) \ \text{ if } \ \langle x_i, x_j\rangle \neq 0\,, \qquad (3.21)$$
$$M_{ii} = \overline{M}_{ii}\,,\, |M_{ij}| \leq \overline{M}_{ij} \ \text{ otherwise}\}.$$

Proposition 42 *For a convex VGF $\Omega_{\mathcal{M}}$, the subdifferential of its conjugate function is given by*

$$\partial\,\Omega^\star_{\mathcal{M}}(Y) = \left\{\tfrac{1}{2}(YM^\dagger + W) :\ \Omega(YM^\dagger + W) = 4\Omega^\star(Y) = \operatorname{tr}(YM^\dagger Y^T)\,,\right.$$
$$\left.\operatorname{range}(W^T) \subseteq \ker(M) \subseteq \ker(Y)\,,\ M \in \mathcal{M} \cap \mathbb{S}_+\right\}. \quad (3.22)$$

When $\Omega^\star_{\mathcal{M}}(Y) \neq 0$ *we have* $\partial\|Y\|^\star_{\mathcal{M}} = \frac{2}{\|Y\|^\star_{\mathcal{M}}}\partial\,\Omega^\star_{\mathcal{M}}(Y)$.

The proof of Proposition 42 is given in the Appendix.

Since $\partial\Omega^\star(Y)$ is non-empty, for any choice of M_0 , there exists a W such that $\frac{1}{2}(YM_0^\dagger + W) \in \partial\Omega^\star(Y)$. However, finding such W is not trivial. The following lemma characterizes the subdifferential as the solution set of a convex optimization problem involving Ω and affine constraints.

Lemma 43 *Given* Y *and an optimal* M_0 *, which by optimality satisfies* $\ker(M_0) \subseteq \ker(Y)$ *, we have*

$$\begin{aligned}\partial\Omega^\star(Y) &= \operatorname{Arg\,min}_Z \left\{\Omega(Z) :\ Z = \tfrac{1}{2}(YM_0^\dagger + W)\,,\ \operatorname{range}(W^T) \subseteq \ker(M_0) \subseteq \ker(Y)\right\} \\ &= \operatorname{Arg\,min}_Z \left\{\Omega(Z) :\ Z = \tfrac{1}{2}(YM_0^\dagger + W)\,,\ WM_0M_0^\dagger = \mathbf{0}\right\} .\end{aligned}$$

This is because for all feasible Z we have $\Omega(Z) \geq \operatorname{tr}(ZM_0Z^T) = \Omega^\star(Y)$.

The characterization of the whole subdifferential is helpful for understanding optimality conditions, but algorithms only need to compute a single subgradient, which is easier than computing the whole subdifferential.

3.3.6 Composition of VGF and absolute values

The characterization of the subdifferential allows us to establish conditions for convexity of $\Psi(X) = \Omega(|X|)$ defined in (3.13). Our result is based on the following Lemma.

Lemma 44 *Given a function* $f : \mathbb{R}^n \to \mathbb{R}$ *, consider* $g(x) = \min_{y \geq |x|} f(y)$ *, and* $h(x) = f(|x|)$ *, where the absolute values and inequalities are all entry-wise. Then,*

(a) $h^{\star\star} \leq g \leq h$.

(b) If f *is convex then* g *is convex and* $g = h^{\star\star}$.

Proof. (a) In $h^\star(y) = \sup_x \{\langle x, y\rangle - f(|x|)\}$, the optimal x has the same sign pattern as y; hence $h^\star(y) = \sup_{x\geq \mathbf{0}} \{\langle x, |y|\rangle - f(x)\}$. Next, we have

$$\begin{aligned} h^{\star\star}(z) &= \sup_y \left\{\langle y, z\rangle - \sup_{x\geq \mathbf{0}} \{\langle x, |y|\rangle - f(x)\}\right\} = \sup_{y\geq \mathbf{0}} \inf_{x\geq \mathbf{0}} \left\{\langle y, |z|\rangle - \langle x, y\rangle + f(x)\right\} \\ &\leq \inf_{x\geq \mathbf{0}} \sup_{y\geq \mathbf{0}} \left\{\langle y, |z|\rangle - \langle x, y\rangle + f(x)\right\} = \inf_{x\geq \mathbf{0}} \sup_{y\geq \mathbf{0}} \left\{\langle y, |z| - x\rangle + f(x)\right\} \\ &= \inf_{x\geq |z|} f(x) = g(z). \end{aligned}$$

This shows the first inequality in part (a). The second inequality follows directly from the definition of g and h.

(b) Consider $x_1, x_2 \in \mathbb{R}^n$ and $\theta \in [0,1]$. Suppose $g(x_i) = f(y_i)$ for some $y_i \geq |x_i|$, for $i = 1, 2$. In other words, y_i is the minimizer in the definition of $g(x_i)$, for $i = 1, 2$. Then,

$$\theta y_1 + (1-\theta) y_2 \geq \theta |x_1| + (1-\theta)|x_2| \geq |\theta x_1 + (1-\theta) x_2|.$$

By definition of g and convexity of f

$$g(\theta x_1 + (1-\theta)x_2) \leq f(\theta y_1 + (1-\theta) y_2) \leq \theta f(y_1) + (1-\theta) f(y_2) = \theta g(x_1) + (1-\theta) g(x_2),$$

which implies that g is convex. It is a classical result that the epigraph of the biconjugate h^{**} is the closed convex hull of the epigraph of h; in other words, h^{**} is the largest lower semi-continuous convex function that is no larger than h (e.g., [134, Theorem 12.2]). Since g is convex and $h^{\star\star} \leq g \leq h$, we must have $h^{\star\star} = g$. ∎

Corollary 45 *Let $\Omega_{\mathcal{M}}$ be a convex VGF. Then, $\Omega_{\mathcal{M}}(|X|)$ is a convex function of X if and only if $\Omega_{\mathcal{M}}(|X|) = \min_{Y\geq |X|} \Omega_{\mathcal{M}}(Y)$.*

Proof. Let $\Omega_{\mathcal{M}}$ be the function f in Lemma 44. Then we have $g(X) = \min_{Y\geq |X|} \Omega_{\mathcal{M}}(Y)$ and $h(X) = \Omega_{\mathcal{M}}(|X|)$. Since here h is a closed convex function, we have $h = h^{\star\star}$ [134, Theorem 12.2], thus part (a) of Lemma 44 implies $h = g$. On the other hand, given a convex function f, part (b) of Lemma 44 states that $g = h^{**}$ is also convex. Hence, $h = g$ implies convexity of h. ∎

Another proof of Corollary 45, in the case where $\sqrt{\Omega_{\mathcal{M}}}$ is a norm and not a semi-norm, is given as Lemma 66 in the Appendix.

Lemma 46 *Let $\Omega_{\mathcal{M}}$ be a convex VGF. If $\partial\Omega_{\mathcal{M}}(X) \cap \mathbb{R}^{n\times m}_{+} \neq \emptyset$ holds for any $X \geq \mathbf{0}$, then $\Psi(X) = \Omega_{\mathcal{M}}(|X|)$ is convex.*

Proof. Using the definition of subgradients for Ω at $|X|$ we have

$$\Omega(|X| + \Delta) \geq \Omega(|X|) + \sup\{\langle G, |X| + \Delta\rangle : \ G \in \partial\Omega \text{ at } |X|\},$$

where the right-most term is the directional derivative of Ω at $|X|$ in the direction Δ. From the assumption, we get $\Omega(Y) \geq \Omega(|X|)$ for all $Y \geq |X|$. Therefore, $\Psi(X) = \Omega_{\mathcal{M}}(|X|) = \min_{Y\geq|X|} \ \Omega_{\mathcal{M}}(Y)$. Corollary 45 establishes the convexity of Ψ. ∎

For example, consider the VGF $\Omega_{\mathcal{M}}$ defined in (3.4), and assume that it is convex. Its subdifferential $\partial\Omega_{\mathcal{M}}$ given in (3.21). For each $X \geq 0$, the matrix product $X\overline{M} \geq \mathbf{0}$ since $\overline{M}$ is also a nonnegative matrix, hence it belongs to $\partial\Omega_{\mathcal{M}}(X)$. Therefore the condition in the above lemma is satisfied, and the function $\Psi(X) = \Omega_{\mathcal{M}}(|X|)$ is convex and has an alternative representation $\Psi(X) = \min_{Y\geq|X|} \ \Omega_{\mathcal{M}}(Y)$. This specific function Ψ has been used in [157] for learning matrices with disjoint supports.

3.4 Proximal Operators

The proximal operator of a closed convex function $h(\cdot)$ is defined as

$$\text{prox}_h(x) = \text{argmin}_u \left\{ h(u) + \tfrac{1}{2}\|u - x\|_2^2 \right\},$$

which always exists and is unique; e.g., see [134, Section 31]. Computing the proximal operator is the essential step in the proximal point algorithm ([106, 135]) and the proximal gradient methods (e.g., [126]). In each iteration of such algorithms, we need to compute $\text{prox}_{\tau h}$ where $\tau > 0$ is a step size parameter. For a convex VGF Ω (for which, without loss of generality, we assume $\mathcal{M} \subseteq \mathbb{S}^m_+$), we have

$$\text{prox}_{\tau\Omega}(X) = \operatorname*{argmin}_{Y} \max_{M\in\mathcal{M}} \ \left\{ \tfrac{1}{2}\|Y - X\|_F^2 + \tau \,\text{tr}(YMY^T) \right\}. \tag{3.23}$$

If $\mathcal{M}$ is a compact convex set, one can change the order of min and max and first solve for Y in terms of any given X and M, which gives $Y = X(I + 2\tau M)^{-1}$. Then we can find the optimal $M_0 \in \mathcal{M}$ given X as

$$M_0 = \text{argmin}_{M \in \mathcal{M}} \ \text{tr}\left(X(I + 2\tau M)^{-1} X^T\right).$$

which gives $\text{prox}_{\tau\Omega}(X) = X(I + 2\tau M_0)^{-1}$. To compute the proximal operator for the conjugate function Ω^*, one can use Moreau's formula (see, e.g., [134, Theorem 31.5]):

$$\text{prox}_{\tau\Omega}(X) + \tau^{-1}\text{prox}_{\tau^{-1}\Omega^\star}(X) = X. \tag{3.24}$$

Next we discuss proximal operators of norms induced by VGFs (section 3.3.4). Since computing the proximal operator of a norm is equivalent to projection onto the dual norm ball, we can express the proximal operator of the norm $\|\cdot\| \equiv \sqrt{\Omega_{\mathcal{M}}(\cdot)}$ as

$$\begin{aligned}\text{prox}_{\tau\|\cdot\|}(X) &= X - \Pi_{\|\cdot\|^\star \le \tau}(X) \\ &= X - \operatorname*{argmin}_{Y} \min_{M,C} \left\{ \|Y - X\|_F^2 : \ \text{tr}(C) \le \tau^2, \ \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq 0, \ M \in \mathcal{M} \right\},\end{aligned}$$

where we used the representation of the conjugate VGF in (3.18) and the dual norm in (3.20). On the other hand, using the definition of proximal operator for the dual norm computed via (3.20) we have

$$\text{prox}_{\tau\|\cdot\|_{\mathcal{M}}^\star}(X) = \operatorname*{argmin}_{Y} \min_{M,C} \left\{ \|Y - X\|_F^2 + \tau(\text{tr}(C) + \gamma_{\mathcal{M}}(M)) : \ \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq \mathbf{0} \right\},$$

where $\gamma_{\mathcal{M}}(M) = \inf\{\lambda \ge 0 : \ M \in \lambda\mathcal{M}\}$ is the gauge function associated to the nonempty convex set $\mathcal{M}$. The computational cost for computing proximal operators can be high in general (involving solving semidefinite programs); however, they may be simplified for special cases of $\mathcal{M}$. For example, a fast algorithm for computing the proximal operator of the VGF associated with the set $\mathcal{M}$ defined in (3.12) is presented in [108]. For general problems, due to the convex-concave saddle point structure in (3.23), we may use the mirror-prox algorithm [125] to obtain an inexact solution.

Left unitarily invariance and QR factorization As mentioned before, VGFs and their conjugates are left unitarily invariant. We can use this fact to simplify the computation of corresponding proximal operators when $n \geq m$. Consider the QR decomposition of a matrix $Y = QR$ where Q is an orthogonal matrix with $Q^TQ = QQ^T = I$ and $R = [R_Y^T \ \mathbf{0}]^T$ is an upper triangular matrix with $R_Y \in \mathbb{R}^{m\times m}$. From the definition, we have $\Omega(Y) = \Omega(R_Y)$ and $\Omega^\star(Y) = \Omega^\star(R_Y)$. For the proximal operators, we can simply plug in R_X from the QR decomposition $X = Q[R_X^T \ \mathbf{0}]^T$ to get

$$\begin{aligned}\operatorname{prox}_{\tau\Omega^\star}(X) &= \operatorname*{argmin}_{Y} \min_{M,C} \left\{ \|Y - X\|_2^2 + \tfrac{1}{2}\tau \operatorname{tr}(C) : \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \succeq \mathbf{0}\,,\ M \in \mathcal{M} \right\} \\ &= Q \cdot \operatorname*{argmin}_{R} \min_{M,C} \left\{ \|R - R_X\|_2^2 + \tfrac{1}{2}\tau \operatorname{tr}(C) : \begin{bmatrix} M & R^T \\ R & C \end{bmatrix} \succeq \mathbf{0}\,,\ M \in \mathcal{M} \right\}\end{aligned}$$

where R is restricted to be an upper triangular matrix and the new semidefinite matrix is of size $2m$ instead of $n + m$ that we had before. The above equality uses two facts. First,

$$\begin{bmatrix} I_m & \mathbf{0} \\ \mathbf{0} & Q^T \end{bmatrix} \begin{bmatrix} M & Y^T \\ Y & C \end{bmatrix} \begin{bmatrix} I_m & \mathbf{0} \\ \mathbf{0} & Q \end{bmatrix} = \begin{bmatrix} M & R^T \\ R & Q^TCQ \end{bmatrix} \succeq \mathbf{0} \tag{3.25}$$

where the right and left matrices in the multiplication are positive definite. Secondly, $\operatorname{tr}(C) = \operatorname{tr}(C')$ where $C' = Q^TCQ$ and assuming C' to be zero outside the first $m \times m$ block can only reduce the objective function. Therefore, we can ignore the last $n - m$ rows and columns of the above PSD matrix.

More generally, because of left unitarily invariance, the optimal Y's in all of the optimization problems in this section have the same column space as the input matrix X; otherwise, a rotation as in (3.25) produces a feasible Y with a smaller value for the objective function.

3.5 Algorithms for Optimization with VGF

In this section, we discuss optimization algorithms for solving convex minimization problems with VGF penalties in the form of (3.6). The proximal operators of VGFs we studied in the previous section are the key parts of proximal gradient methods (see, e.g., [15, 16, 126]).

More specifically, when the loss function $\mathcal{L}(X)$ is smooth, we can iteratively update the variables $X^{(t)}$ as follows:

$$X^{(t+1)} = \text{prox}_{\gamma_t \Omega}\big(X^{(t)} - \gamma_t \nabla \mathcal{L}(X^{(t)})\big), \qquad t = 0, 1, 2, \ldots,$$

where γ_t is a step size at iteration t. When $\mathcal{L}(X)$ is not smooth, then we can use subgradients of $\mathcal{L}(X^{(t)})$ in the above algorithm, or use the classical subgradient method on the overall objective $\mathcal{L}(X) + \lambda\Omega(X)$. In either case, we need to use diminishing step size and the convergence can be very slow. Even when the convergence is relatively fast (in terms of number of iterations), the computational cost of the proximal operator in each iteration can be very high.

In this section, we focus on loss functions that have a special form shown in (3.26). This form comes up in many common loss functions, some of which listed later in this section, and allows for faster algorithms. We assume that the loss function $\mathcal{L}$ in (3.6) has the following representation:

$$\mathcal{L}(X) = \max_{g \in \mathcal{G}} \ \langle X, \mathcal{D}(g) \rangle - \widehat{\mathcal{L}}(g)\,, \tag{3.26}$$

where $\widehat{\mathcal{L}} : \mathbb{R}^p \to \mathbb{R}$ is a convex function, $\mathcal{G}$ is a convex and compact subset of $\mathbb{R}^p$, and $\mathcal{D} : \mathbb{R}^p \to \mathbb{R}^{n \times m}$ is a linear operator. This is also known as a Fenchel-type representation (see, e.g., [94]). Moreover, consider the infimal post-composition [13, Def. 12.33] of $\widehat{\mathcal{L}} : \mathcal{G} \to \mathbb{R}$ by $\mathcal{D}(\cdot)$, defined as

$$(\mathcal{D} \triangleright \widehat{\mathcal{L}})(Y) = \inf \{\widehat{\mathcal{L}}(G) : \ \mathcal{D}(G) = Y\,,\ G \in \mathcal{G}\}\,.$$

Then, the conjugate to this function is equal to $\mathcal{L}$. In other words, $\mathcal{L}(X) = \widehat{\mathcal{L}}^\star(\mathcal{D}^\star(X))$ where $\widehat{\mathcal{L}}^\star$ is the conjugate function and $\mathcal{D}^\star$ is the adjoint operator. The composition of a nonlinear convex loss function and a linear operator is very common for optimization of linear predictors in machine learning (e.g., [79]), which we will demonstrate with several examples later in this section.

With the variational representation of $\mathcal{L}$ in (3.26), we can write the VGF-penalized loss minimization problem (3.6) as a convex-concave saddle-point optimization problem:

$$J_{\text{opt}} = \min_{X} \max_{M \in \mathcal{M} \cap \mathbb{S}_+,\, g \in \mathcal{G}} \left\{ \langle X, \mathcal{D}(g) \rangle - \widehat{\mathcal{L}}(g) + \lambda \operatorname{tr}(X M X^T) \right\}. \tag{3.27}$$

If $\widehat{\mathcal{L}}$ is smooth (while $\mathcal{L}$ may be nonsmooth) and the sets $\mathcal{G}$ and $\mathcal{M}$ are simple (e.g., admitting simple projections), we can solve problem (3.27) using the *mirror-prox* algorithm [125, 94]. In section 3.5.1, we present a variant of the miror-prox algorithm equipped with an adaptive line search scheme. Then in Section 3.5.2, we present a preprocessing technique to transform problems of the form (3.27) into smaller dimensions, which can be solved more efficiently under favorable conditions.

Before diving into the algorithmic details, we examine some common loss functions and derive the corresponding representation (3.26) for them. This discussion will provide intuition about the linear operator $\mathcal{D}$ and set $\mathcal{G}$ in relation with data and prediction.

Norm loss Given a norm $\|\cdot\|$ and its dual $\|\cdot\|^\star$, consider the squared norm loss

$$\mathcal{L}(x) = \tfrac{1}{2}\|Ax - \mathbf{b}\|^2 = \max_{g} \left\{ \langle g, Ax - \mathbf{b} \rangle - \tfrac{1}{2}(\|g\|^\star)^2 \right\}.$$

In terms of the representation in (3.26), here we have $\mathcal{D}(g) = A^T g$ and $\widehat{\mathcal{L}}(g) = \frac{1}{2}(\|g\|^\star)^2 + \mathbf{b}^T g$. Similarly, a norm loss can be represented as

$$\mathcal{L}(x) = \|Ax - \mathbf{b}\| = \max_{g} \{ \langle x, A^T g \rangle - \mathbf{b}^T g : \ \|g\|^\star \le 1 \},$$

where we have $\mathcal{D}(g) = A^T g$, $\widehat{\mathcal{L}}(g) = \mathbf{b}^T g$ and $\mathcal{G} = \{g : \|g\|^* \le 1\}$.

ε-insensitive (deadzone) loss Another variant of the absolute loss function is called the ε-insensitive loss (e.g., see [121, Section 14.5.1] for more details and applications) and can be represented, similar to (3.26), as

$$\mathcal{L}_\varepsilon(x) = (|x| - \varepsilon)_+ = \max_{\alpha,\beta} \{ \alpha(x - \varepsilon) + \beta(-x - \varepsilon) : \ \alpha, \beta \ge 0, \ \alpha + \beta \le 1 \}.$$

Hinge loss for binary classification In binary classification problems, we are given a set of training examples $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_N, b_N)$, where each $\mathbf{a}_s \in \mathbb{R}^p$ is a feature vector and $b_s \in \{+1, -1\}$ is a binary label. We would like to find $x \in \mathbb{R}^p$ such that the linear function $\mathbf{a}_s^T x$ can predict the sign of label b_s for each $s = 1, \ldots, N$. The hinge loss $\max\{0, 1 - b_s(\mathbf{a}_s^T x)\}$ returns 0 if $b_s(\mathbf{a}_s^T x) \geq 1$ and a positive loss growing with the absolute value of $b_s(\mathbf{a}_s^T x)$ when it is negative. The average hinge loss over the whole data set can be expressed as

$$\mathcal{L}(x) \;=\; \frac{1}{N}\sum_{s=1}^{N} \max\left\{0, 1 - b_s(\mathbf{a}_s^T x)\right\} \;=\; \max_{g \in \mathcal{G}} \langle g, \mathbf{1} - \mathbf{D}x \rangle \,.$$

where $\mathbf{D} = [b_1\mathbf{a}_1, \ldots, b_N\mathbf{a}_N]^T$. Here, in terms of (3.26), we have, $\mathcal{G} = \{g \in \mathbb{R}^N : \; 0 \leq g_s \leq 1/N\}$, $\mathcal{D}(g) = -\mathbf{D}^T g$, and $\widehat{\mathcal{L}}(g) = -\mathbf{1}^T g$.

Multi-class hinge loss For multiclass classification problems, each sample $\mathbf{a}_s$ has a label $b_s \in \{1, \ldots, m\}$, for $s = 1, \ldots, N$. Our goal is to learn a set of classifiers $x_1, \ldots, x_m$, that can predict the labels b_s correctly. For any given example $\mathbf{a}_s$ with label b_s, we say the prediction made by $x_1, \ldots, x_m$ is correct if

$$x_i^T \mathbf{a}_s \geq x_j^T \mathbf{a}_s \quad \text{for all } (i, j) \in \mathcal{I}(b_s), \tag{3.28}$$

where $\mathcal{I}_k$, for $k = 1, \ldots, m$, characterizes the required comparisons to be made for any example with label k. Here are two examples.

1. *Flat multiclass classification:* $\mathcal{I}(k) = \{(k, j) : j \neq k\}$. In this case, the constraints in (3.28) are equivalent to the label $b_s = \text{argmax}_{i \in \{1,\ldots,m\}} \, x_i^T \mathbf{a}_s$; see [161].

2. *Hierarchical classification.* In this case, the labels $\{1, \ldots, m\}$ are organized in a tree structure, and each $\mathcal{I}(k)$ is a special subset of the edges in the tree depending on the class label k; see Section 3.6 and [54, 169] for further details.

Given the labeled data set $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_N, b_N)$, we can optimize $X = [x_1, \ldots, x_m]$ to minimize the averaged multi-class hinge loss

$$\mathcal{L}(X) = \frac{1}{N}\sum_{s=1}^{N} \max\left\{0, 1 - \max_{(i,j) \in \mathcal{I}(b_s)} \{x_i^T \mathbf{a}_s - x_j^T \mathbf{a}_s\}\right\}, \tag{3.29}$$

which penalizes the amount of violation for the inequality constraints in (3.28).

In order to represent the loss function in (3.29) in the form of (3.26), we need some more notations. Let $p_k = |\mathcal{I}(k)|$, and define $E_k \in \mathbb{R}^{m\times p_k}$ as the incidence matrix for the pairs in $\mathcal{I}_k$; i.e., each column of E_k, corresponding to a pair $(i,j) \in \mathcal{I}_k$, has only two nonzero entries: -1 at the ith entry and $+1$ at the jth entry. Then the p_k constraints in (3.28) can be summarized as $E_k^T X^T \mathbf{a}_s \le \mathbf{0}$. It can be shown that the multi-class hinge loss $\mathcal{L}(X)$ in (3.29) can be represented in the form (3.26) via

$$\mathcal{D}(g) = -A\,\mathcal{E}(g), \qquad \text{and} \qquad \widehat{\mathcal{L}}(g) = -\mathbf{1}^T g,$$

where $A = [\mathbf{a}_1 \ \cdots \ \mathbf{a}_N]$ and $\mathcal{E}(g) = [E_{b_1} g_1 \ \cdots \ E_{b_N} g_N]^T \in \mathbb{R}^{N\times m}$. Moreover, the domain of maximization in (3.26) is defined as

$$\mathcal{G} = \mathcal{G}_{b_1} \times \ldots \times \mathcal{G}_{b_N} \quad \text{where} \quad \mathcal{G}_k = \{g \in \mathbb{R}^{p_k} \,:\, g \ge 0\,,\ \mathbf{1}^T g \le 1/N\}\,. \tag{3.30}$$

Combining the above variational form for multi-class hinge loss and a VGF as penalty on X, we can reformulate the nonsmooth convex optimization problem $\min_X \{\mathcal{L}(X) + \lambda\Omega_{\mathcal{M}}(X)\}$ as the convex-concave saddle point problem

$$\min_X \max_{M\in\mathcal{M}\cap\mathbb{S}_+,\, g\in\mathcal{G}} \left\{\mathbf{1}^T g - \langle X, A\,\mathcal{E}(g)\rangle + \lambda\,\mathrm{tr}(XMX^T)\right\}. \tag{3.31}$$

3.5.1 Mirror-prox algorithm with adaptive line search

The mirror-prox (MP) algorithm was proposed by Nemirovski [125] for approximating the saddle points of smooth convex-concave functions and solutions of variational inequalities with Lipschitz continuous monotone operators. It is an extension of the extra-gradient method [96], and more variants are studied in [93]. In this section, we first present a variant of the MP algorithm equipped with an adaptive line search scheme. Then explain how to apply it to solve the VGF-penalized loss minimization problem (3.27).

We describe the MP algorithm in the more general setup of solving variational inequality problems. Let Z be a convex compact set in Euclidean space E equipped with inner product

$\langle \cdot, \cdot \rangle$, and $\| \cdot \|$ and $\| \cdot \|_*$ be a pair of dual norms on E, i.e., $\|\xi\|_* = \max_{\|z\| \leq 1} \langle \xi, z \rangle$. Let $F : Z \to E$ be a Lipschitz continuous monotone mapping, i.e.,

$$\forall\, z, z' \in Z : \ \ \|F(z) - F(z')\|_* \leq L\|z - z'\|, \ \text{ and } \ \langle F(z) - F(z'), z - z' \rangle \geq 0\,. \tag{3.32}$$

The goal of the MP algorithm is to approximate a (strong) solution to the variational inequality associated with (Z, F):

$$\langle F(z^\star), z - z^\star \rangle \geq 0, \quad \forall\, z \in Z.$$

Let $\phi(x, y)$ be a smooth function that is convex in x and concave in y, and X and Y be closed convex sets. Then the convex-concave saddle point problem

$$\min_{x \in X} \max_{y \in Y} \ \phi(x, y), \tag{3.33}$$

can be posed as a variational inequality problem with $z = (x, y)$, $Z = X \times Y$ and

$$F(z) = \begin{bmatrix} \nabla_x \phi(x, y) \\ -\nabla_y \phi(x, y) \end{bmatrix}. \tag{3.34}$$

The setup of the Mirror-Prox algorithm requires a distance-generating function $h(z)$ which is compatible with the norm $\| \cdot \|$. In other words, $h(z)$ is subdifferentiable on the relative interior of Z, denoted Z^o, and is strongly convex with modulus 1 with respect to $\| \cdot \|$, i.e.,

$$\forall\, z, z' \in Z : \qquad \langle \nabla h(z) - \nabla h(z'), z - z' \rangle \ \geq \ \|z - z'\|^2. \tag{3.35}$$

For any $z \in Z^o$ and $z' \in Z$, we can define the Bregman divergence at z as

$$V_z(z') = h(z') - h(z) - \langle \nabla h(z), z' - z \rangle,$$

and the associated proximity mapping as

$$\mathrm{P}_z(\xi) \ \ = \ \ \mathrm{argmin}_{z' \in Z} \left\{ \langle \xi, z' \rangle + V_z(z') \right\} \ \ = \ \ \mathrm{argmin}_{z' \in Z} \left\{ \langle \xi - \nabla h(z), z' \rangle + h(z') \right\}.$$

Algorithm: Mirror-Prox($z_1, \gamma_1, \varepsilon$)

repeat

$t := t + 1$

repeat

$\gamma_t := \gamma_t / c_{\text{dec}}$

$w_t := \mathrm{P}_{z_t}(\gamma_t F(z_t))$

$z_{t+1} := \mathrm{P}_{z_t}(\gamma_t F(w_t))$

until $\delta_t \leq 0$

$\gamma_{t+1} := c_{\text{inc}} \gamma_t$

until $V_{z_t}(z_{t+1}) \leq \varepsilon$

return $\bar{z}_t := (\sum_{\tau=1}^{t} \gamma_\tau)^{-1} \sum_{\tau=1}^{t} \gamma_\tau w_\tau$

Figure 3.2: Mirror-Prox algorithm with adaptive line search. Here $c_{\text{dec}} > 1$ and $c_{\text{inc}} > 1$ are parameters controlling the decrease and increase of the step size γ_t in the line search trials. The stopping criterion for the line search is $\delta_t \leq 0$ where $\delta_t = \gamma_t \langle F(w_t), w_t - z_{t+1} \rangle - V_{z_t}(z_{t+1})$.

With these definitions, we are now ready to present the MP algorithm in Figure 3.2. Compared with the original MP algorithm [125, 93], our variant employs an adaptive line search procedure to determine the step sizes γ_t, for $t = 1, 2, \ldots$. We can exit the algorithm whenever $V_{z_t}(z_{t+1}) \leq \epsilon$ for some $\epsilon > 0$. Under the assumptions in (3.32), the MP algorithm in Figure 3.2 enjoys the same $O(1/t)$ convergence rate as the one proposed in [125], but performs much faster in practice. The proof requires only simple modifications of the proof in [125, 93].

To solve the saddle-point problem in (3.27), assuming $\widehat{\mathcal{L}}$ is smooth and $\mathcal{M} \subset \mathbb{S}_+$ we can apply the MP algorithm directly. Notice that $\mathcal{M} \subset \mathbb{S}_+$ (or the iterations of the MP being PSD) is required for the objective to be convex in X. We remark on such guarantee after Corollary 37. Moreover, the gradient mapping in (3.34) becomes

$$F(X, M, g) = \begin{bmatrix} \mathrm{vec}(2\lambda XM + \mathcal{D}(g)) \\ -\lambda \mathrm{vec}(X^T X) \\ \mathrm{vec}(\nabla \widehat{\mathcal{L}}(g) - \mathcal{D}^\star(X)) \end{bmatrix},$$

where $\mathcal{D}^\star(\cdot)$ is the adjoint operator to $\mathcal{D}(\cdot)$. Assuming $g \in \mathbb{R}^p$, computing F requires $O(nm^2 + nmp)$ operations for matrix multiplications. In the next section, we present a method that can potentially reduce the problem size by replacing n with $\min\{mp, n\}$. In the case of SVM with the hinge loss as in our real-data numerical example, one can replace n by $\min\{N, mp, n\}$, where N is the number of samples.

3.5.2 A Kernel Trick (Reduced Formulation)

As we discussed earlier, when the loss function has the structure (3.26), we can write the VGF-penalized minimization problem as a convex-concave saddle point problem

$$J_{\mathrm{opt}} = \min_{X \in \mathbb{R}^{n \times m}} \max_{g \in \mathcal{G}} \ \left\{ \langle X, \mathcal{D}(g) \rangle - \widehat{\mathcal{L}}(g) + \lambda\, \Omega(X) \right\}. \tag{3.36}$$

Since $\mathcal{G}$ is compact, Ω is convex in X, and $\widehat{\mathcal{L}}$ is convex in g, we can use the minimax theorem to interchange the max and min. Then, for any orthogonal matrix Q we have

$$\begin{aligned} J_{\text{opt}} &= \max_{g\in\mathcal{G}} \min_{X} \ \Big\{ \langle X, \mathcal{D}(g)\rangle - \widehat{\mathcal{L}}(g) + \lambda\,\Omega(X) \Big\} \\ &= \max_{g\in\mathcal{G}} \min_{X} \ \Big\{ \langle Q^T X, Q^T\mathcal{D}(g)\rangle - \widehat{\mathcal{L}}(g) + \lambda\,\Omega(Q^T X) \Big\} \\ &= \max_{g\in\mathcal{G}} \min_{X} \ \Big\{ \langle X, Q^T\mathcal{D}(g)\rangle - \widehat{\mathcal{L}}(g) + \lambda\,\Omega(X) \Big\} \end{aligned} \tag{3.37}$$

where the second equality is due to the left unitarily invariance of Ω, and we renamed the variable X to get the third equality. Observe that Q is an arbitrary orthogonal matrix in (3.37) and can be chosen in a clever way to simplify $\mathcal{D}$ as described in the sequel. Since $\mathcal{D}(g)$ is linear in g, consider a representation as

$$\mathcal{D}(g) = [D_1 g \ \cdots \ D_m g] = [D_1 \ \cdots \ D_m](I_m \otimes g) = \mathbf{D}(I_m \otimes g), \tag{3.38}$$

for some $D_i \in \mathbb{R}^{n\times p}$ and $\mathbf{D} \in \mathbb{R}^{n\times mp}$. Then, express $\mathbf{D}$ as the product of an orthogonal matrix and a residue matrix, such as in QR decomposition $\mathbf{D} = QR$, where provided that $n > mp$, only the first mp rows of R can be nonzero (will be denoted by R_1). Define $\mathcal{D}'(g) = R_1(I_m \otimes g) \in \mathbb{R}^{q\times m}$ for $q = \min\{mp, n\}$. Plugging the above choice of Q in (3.37) gives

$$J_{\text{opt}} = \max_{g\in\mathcal{G}} \min_{X_1, X_2} \ \left\{ \langle \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \begin{bmatrix} \mathcal{D}'(g) \\ \mathbf{0} \end{bmatrix} \rangle - \widehat{\mathcal{L}}(g) + \lambda\,\Omega(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}) \right\}.$$

Observe that setting X_2 to zero does not increase the value of Ω which allows for restricting the above to the subspace $X_2 = 0$ and getting

$$J_{\text{opt}} = \min_{X\in\mathbb{R}^{q\times m}} \max_{g\in\mathcal{G}} \langle X, \mathcal{D}'(g)\rangle - \widehat{\mathcal{L}}(g) + \lambda\Omega(X) \tag{3.39}$$

whose X variable has $q = \min\{mp, n\}$ rows compared to n rows in (3.6).

Notice that while the evaluation of J_{opt} via (3.39) can potentially be more efficient, we are interested in recovering an *optimal point* X in (3.36) which is different from the optimal

points in (3.39). However, tracing back the steps we took to get (3.39) from (3.36), we get

$$X_{\text{opt}}^{(3.36)} = Q \begin{bmatrix} X_{\text{opt}}^{(3.39)} \\ \mathbf{0} \end{bmatrix} .$$

The special case of regularization with squared Euclidean norm has been understood and used before; e.g., see [140]. However, the above derivations show that we can get similar results when the regularization can be represented as a maximum of squared weighted Euclidean norms.

It is worth mentioning that the reduced formulation in (3.39) can be similarly derived via a dual approach; one has to take the dual of the loss-regularized optimization problem (e.g., see Example 11.41 in [137]), use the left unitarily invariance of the conjugate VGF to reduce $\mathcal{D}$ to $\mathcal{D}'$, and dualize the problem again, to get (3.39).

3.5.3 A Representer Theorem

A general loss-regularized optimization problem as in (3.6) where the loss admits a Fenchel-type representation and the regularizer is a strongly convex VGF (including all squared vector norms) enjoys a representer theorem (see, e.g., [140]). More specifically, the optimal solution is linearly related to the linear operator $\mathcal{D}$ in the representation of the loss. As mentioned before, for many common loss functions, $\mathcal{D}$ encodes the samples, which reduces the following proposition to the usual representer theorem.

Proposition 47 *For a loss-regularized minimization problem as in* (3.6)*, i.e.,*

$$\underset{X \in \mathbb{R}^{n \times m}}{\text{minimize}} \quad \mathcal{L}(X) + \lambda\, \Omega_{\mathbb{M}}(X)\,,$$

where $\Omega_{\mathbb{M}}$ is strongly convex and $\mathcal{L}$ admits a Fenchel-type representation as

$$\mathcal{L}(X) = \max_{g \in \mathcal{G}} \ \langle X, \mathcal{D}(g) \rangle - \widehat{\mathcal{L}}(g) = \max_{g \in \mathcal{G}} \ \langle X, \mathbf{D}(I_m \otimes g) \rangle - \widehat{\mathcal{L}}(g)\,,$$

the optimal solution X_{opt} admits a representation of the form

$$X_{\text{opt}} = \mathbf{D}C$$

where the coefficient matrix C is given by $C = -\frac{1}{2\lambda} M_{\text{opt}}^{-1} \otimes g_{\text{opt}}$ (optimal solutions of (3.27)).

Proof. Denote the optimal solution of (3.27) by $(X_{\text{opt}}, g_{\text{opt}}, M_{\text{opt}})$, which shares $(X_{\text{opt}}, g_{\text{opt}})$ with (3.36). Consider the optimality condition as $-\frac{1}{\lambda}\mathcal{D}(g_{\text{opt}}) \in \partial\Omega(X_{\text{opt}})$ which implies

$$X_{\text{opt}} \in \partial\Omega^{\star}(-\tfrac{1}{\lambda}\mathcal{D}(g_{\text{opt}}))\,.$$

Now, suppose $\mathcal{M} \subset \mathbb{S}_{+}^{m}$ which is equivalent to assuming that $\Omega_{\mathcal{M}}$ is strongly convex. Considering the characterization of subdifferential for $\Omega^{\star}$ from Proposition 42 as well as the representation of $\mathcal{D}(g)$ in (3.38) we get

$$X_{\text{opt}} = -\tfrac{1}{2\lambda}\mathcal{D}(g_{\text{opt}})M_{\text{opt}}^{-1} = -\tfrac{1}{2\lambda}\mathbf{D}(I_m \otimes g_{\text{opt}})M_{\text{opt}}^{-1} = -\tfrac{1}{2\lambda}\mathbf{D}(M_{\text{opt}}^{-1} \otimes g_{\text{opt}})\,.$$

■

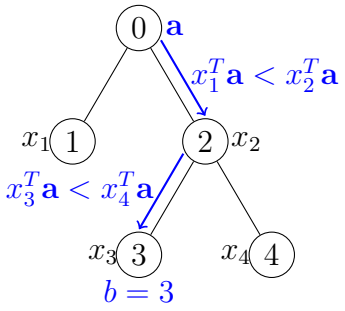
This representer theorem allows us to apply our methods in more general reproducing kernel Hilbert spaces (RKHS) by choosing a problem specific reproducing kernel; e.g. see [140, 169].

3.6 Numerical Example

In this section, we discuss the application of VGFs in hierarchical classification to demonstrate the effectiveness of the presented algorithms in a real data experiment. More specifically, we compare the modified mirror-prox algorithm with adaptive line search presented in Section 3.5.1 with the variant of Regularized Dual Averaging (RDA) method used in [169] in the text categorization application discussed in [169].

Let $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_N, b_N)$ be a set of labeled data where each $\mathbf{a}_i \in \mathbb{R}^n$ is a feature vector and the associated $b_i \in \{1, \ldots, m\}$ is a class label. The goal of multi-class classification is to learn a classification function $f : \mathbb{R}^n \to \{1, \ldots, m\}$ so that, given any sample $\mathbf{a} \in \mathbb{R}^n$ (not necessarily in the training set), the prediction $f(\mathbf{a})$ attains a small classification error compared with the true label.

In hierarchical classification, the class labels $\{1, \ldots, m\}$ are organized in a category tree, where the root of the tree is given the fictious label 0 (see Figure 3.3a). For each node



(a)

$$f(\mathbf{a}) = \left\{ \begin{array}{l} \textbf{initialize } i := 0 \\ \textbf{while } \mathcal{C}(i) \text{ is not empty} \\ \qquad i := \operatorname{argmax}_{j \in \mathcal{C}(i)} \; x_j^T \mathbf{a} \\ \textbf{return } i \end{array} \right\}$$

(b)

Figure 3.3: (a) An example of hierarchical classification with four class labels $\{1, 2, 3, 4\}$. The instance $\mathbf{a}$ is classified recursively until it reaches the leaf node $b = 3$, which is its predicted label. (b) Definition of the hierarchical classification function.

$i \in \{0, 1, \ldots, m\}$, let $\mathcal{C}(i)$ be the set of children of i, $\mathcal{S}(i)$ be the set of siblings of i, and $\mathcal{A}(i)$ be the set of ancestors of i excluding 0 but including itself. A hierarchical linear classifier $f(\mathbf{a})$ is defined in Figure 3.3b, which is parameterized by the vectors $x_1, \ldots, x_m$ through a recursive procedure. In other words, an instance is labeled sequentially by choosing the category for which the associated vector outputs the largest score among its siblings, until a leaf node is reached. An example of this recursive procedure is shown in Figure 3.3a. For the hierarchical classifier defined above, given an example $\mathbf{a}_s$ with label b_s, a correct prediction made by $f(\mathbf{a})$ implies that (3.28) holds with

$$\mathcal{I}(k) = \big\{(i, j) \; : \; j \in \mathcal{S}(i), \; i \in \mathcal{A}(k)\big\}.$$

Given a set of examples $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_N, b_N)$, we can train a hierarchical classifier parametrized by $X = [x_1, \ldots, x_m]$ by solving the problem $\min_X \big\{\mathcal{L}(X) + \lambda\Omega(X)\big\}$, with the loss function $\mathcal{L}(X)$ defined in (3.29) and an appropriate VGF penalty function $\Omega(X)$. As discussed in Section 3.5, the training optimization problem can be reformulated as a convex-concave saddle point problem of the form (3.27) and solved by the mirror-prox algorithm described in Section 3.5.1. In addition, we can use the reduction procedure discussed in Section 3.5.2 to

reduce computational cost.

As discussed in [169], one can assume a model where classification at different levels of the hierarchy rely on different features or different combination of features. Therefore, authors in [169] proposed regularization with $|x_i^T x_j|$ whenever $j \in \mathcal{A}(i)$. A convex formulation of such a regularization function can be written in the form (3.4) with

$$\mathcal{M} = \{M : \ M_{ii} = \overline{M}_{ii} \ , \ |M_{ij}| = |\overline{M}_{ij}|\} \tag{3.40}$$

where the nonzero pattern of $\overline{M}$ corresponds to the pairs of ancestor-descendant nodes. According to (3.15), we have $\mathcal{M} \subset \mathbb{S}_+^m$ provided that $\lambda_{\min}(\widetilde{M}) \geq 0$; see Figure 3.1.

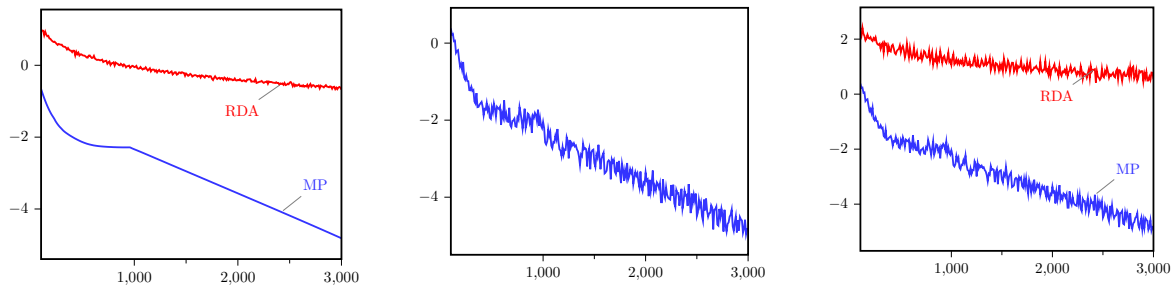
As a real-world example, we consider the classification dataset Reuters Corpus Volume I, RCV1-v2 [101], which is an archive of over 800,000 manually categorized newswire stories and is available in libSVM. A subset of the hierarchy of labels in RCV1-v2, with $m = 23$ labels (18 leaves), is called ECAT and is used in our experiments. The samples and the classifiers are of dimension $n = 47236$. Lastly, there are 2196 training, and 69160 test samples available.

We solve the same loss-regularized problem as in [169], but using mirror-prox (discussed in Section 3.5.1) instead of regularized dual averaging (RDA). The regularization function is a VGF and is given in (3.4). A reformulation of the whole problem as a smooth convex-concave problem is given in (3.31). To obtain comparable results, we use the same matrix $\overline{M}$ and regularization parameter $\lambda = 1$ as in [169]. Since we are solving the same problem as [169], the prediction error on test data by the estimated classifiers will be the same as the error reported in this reference. Note that in this experiment, $n = 47236$ while $m = 23$ and $p > 2196$, so the kernel trick is not particularly useful since n is not larger than mp.

In the setup of the mirror-prox algorithm, we use the ℓ_2 norm as the mirror map which requires the least knowledge about the optimization problem (see [93] for the requirements when combining a number of mirror maps corresponding to different constraint sets in the saddle point optimization problem). With this mirror map, the steps of mirror-prox only require orthogonal projection onto $\mathcal{G}$ and $\mathcal{M}$. The projection onto $\mathcal{G}$ in (3.30) boils down

to separate projections onto N scaled full-dimensional simplexes (where the summation of entries is bounded by 1 and not necessarily equal to 1). Each projection amounts to zeroing out the negative entries followed by a projection onto the ℓ_1 unit norm ball (e.g., using the simple process described in [59]).

The variant of RDA proposed in [169] has a convergence rate of $O(\ln(t)/\sigma t)$ for the objective value, where σ is the strong convexity parameter of the objective. On the other hand, mirror-prox enjoys a convergence rate of $O(1/t)$ as given in [125]. Although there is a clear advantage to the MP method compared to RDA in terms of the theoretical guarantee, one should be aware of the difference between the notions of gap for the two methods. Figure 3.4a compares $\|X_t - X_{\text{final}}\|_F$ for MP and RDA using each one's own final estimate X_{final}. In terms of the runtime, we epirically observe that each iteration of MP takes about 3 times more time compared to RDA. However, as evident from Figure 3.4a, MP is still much faster in generating a fixed-accuracy solution. Figure 3.4b illustrates the decay in the value of the gap for mirror-prox method, $V_{z_t}(z_{t+1})$, which confirms the theoretical convergence rate of $O(1/t)$.



(a) (b) (c)

Figure 3.4: Convergence behavior for mirror-prox and RDA in our numerical experiment. (a) Average error over the m classifiers between each iteration and the final estimate, $\|X_t - X_{\text{final}}\|_F$, for the MP and RDA algorithms, on a logarithmic scale. (b) $V_{z_t}(z_{t+1})$ on a logarithmic scale. (c) The value of loss function, relative to the final value, on a logarithmic scale for MP. For visualization purposes, the plots show data points at every 10 iterations.

3.7 Discussion

In this work, we introduce variational Gram functions, which include many existing regularization functions as well as important new ones. Convexity properties of this class, conjugate functions, subdifferentials, semidefinite representability, proximal operators, and other convex analysis properties are studied. By adapting the mirror-prox method [125], we provide a general and efficient optimization algorithm for VGF-regularized loss minimization problems. We establish a general kernel trick and a representer theorem for such problems. Finally, the effectiveness of VGF regularization as well as the efficiency of our optimization approach is illustrated by a numerical example on hierarchical classification for text categorization.

There are numerous directions for future research on this class of functions. One issue to address is how to systematically pick an appropriate set $\mathcal{M}$ when defining a new VGF for some new application. Statistical properties of VGFs, for example the corresponding sample complexity, are of interest from a learning theory perspective. The presented kernel trick (which uses the left unitarily invariance property of VGFs) can be potentially extended to other invariant regularizers. And last but not least, it is interesting to see if there is a variational Gram representation for any squared left unitarily invariant norm. We elaborate on these directions and add some more in Chapter 5.2.

Chapter 4
COMMUNITY DETECTION

The Stochastic Block Model (SBM) is a widely used random graph model for networks with communities. Despite the recent burst of interest in community detection under the SBM from statistical and computational points of view, there are still gaps in understanding the fundamental limits of recovery. In this chapter, we consider the SBM in its full generality, where there is no restriction on the number and sizes of communities or how they grow with the number of nodes, as well as on the connectivity probabilities inside or across communities. For such stochastic block models, we provide guarantees for exact recovery via a semidefinite program as well as upper and lower bounds on SBM parameters for exact recoverability. Our results exploit the tradeoffs among the various parameters of heterogenous SBM and provide recovery guarantees for many new interesting SBM configurations.

4.1 Introduction

A fundamental problem in network science and machine learning is to discover structures in large, complex networks (e.g., biological, social, or information networks). Community or cluster detection underlies many decision tasks, as a basic step that uses pairwise relations between data points in order to understand more global structures in the data. Applications include recommendation systems [163], image segmentation [144, 110], learning gene network structures in bioinformatics, e.g., in protein detection [44] and population genetics [92].

In spite of a long history of heuristic algorithms (see, e.g., [99] for an empirical overview), as well as strong research interest in recent years on the theoretical side as briefly reviewed below, there are still gaps in understanding the fundamental information theoretic limits of recoverability (i.e., if there is enough information to reveal the communities) and compu-

tational tractability (if there are efficient algorithms to recover them). This is particularly true in the case of sparse graphs (that test the limits of recoverability), graphs with heterogeneous communities (communities varying greatly in size and connectivity), graphs where the number of communities grows with the number of nodes, and partially observed graphs (with various observation models).

4.1.1 Exact Recovery for Heterogenous Stochastic Block Model

The stochastic block model (SBM), first introduced and studied in mathematical sociology by Holland, Laskey and Leinhardt in 1983 [81], can be described as follows. Consider n vertices partitioned into r *communities* $V_1, V_2, \ldots, V_r$, of sizes $n_1, n_2, \ldots, n_r$. We endow the kth community with an Erdős-Rényi random graph model $\mathcal{G}(n_k, p_k)$ and draw an edge between pairs of nodes in different communities independently with probability q; i.e., for any pair of nodes i and j, if $i, j \in V_k$ for some $k \in \{1, \ldots, r\}$ we draw an edge with probability p_k, and draw an edge with probability q if they are in different communities. We assume $q < \min_k p_k$ in order for the idea of communities to make sense. This defines a distribution over random graphs known as the stochastic block model. In this chapter, we assume the above model while allowing the number of communities to grow with the number of nodes (similar to [50, 76, 138]). We refer to this model as the *heterogeneous stochastic block model* to contrast our study of this general setting with previous works on special cases of SBM such as 1) homogenous SBM where the communities are *equivalent* (they are of the same size and the connectivity probabilities are equal,) e.g., as in [47], or, 2) SBM with linear-sized communities, where the number of communities is *fixed* and all community sizes are $\mathrm{O}(n)$; e.g., as in [2].

4.1.2 Statistical and Computational Regimes

What we can infer about the community structure from a single draw of the random graph varies based on the regime of model parameters. Often, the following scenarios are considered.

1. *Recovery,* where the proportion of misclassified nodes is negligible; either 0 or asymptotically negligible as the number of nodes grow, corresponding to the subregimes below.

   1a) *Exact Recovery (or Recovery with Strong Consistency).* In this regime it is possible to recover all labels, with high probability. That is, an algorithm has been proved to do so, whether in polynomial time or not. For example, [47, 2] studied the exact recovery problem for special cases of SBM.

   1b) *Almost Exact Recovery (or Recovery with Weak Consistency).* In this case, algorithms exist to recover a proportion $1 - o(1)$ of the nodes, but not all of them. See [138] for early works on weakly consistent recovery, [119] for the case of binary SBM, [165] for finite number of linear-sized communities, and [166, 69] for a growing number of approximately same-sized communities.

2. *Approximation,* where a finite fraction (bounded away from 1) of the vertices is recovered.

   2a) *Partial Recovery (or Approximation) Regime.* Only a *fraction* of vertices, i.e. $(1-\alpha)n$ for some $0 < \alpha < 1$, can be guaranteed to be recovered correctly. e.g. see [50, 53]. A series of works have provided partial recovery conditions for the cases of two equivalent communities [118, 107, 116, 117], finite number of linear-sized communities [2], and heterogenous SBM [76, 98].

   2b) *Detectability.* One may construct a partition of the graph which is correlated with the true partition (which in this context means doing better than guessing), but one cannot *guarantee* any kind of quantitative improvement over random guessing. This happens in very sparse regimes when some p_k's and q are of the same, small, order; e.g. see [118, 3].

Both recovery and approximation can be studied from statistical and computational points of view.

Statistically, one can ask about the parameter regimes for which the model can be recovered or approximated. Such characterizations are specially important when an information-theoretical lower bound (below which recovery is not possible with high probability) is shown to be achievable with an algorithm (with high probability), hence characterizing a *phase transition* in model parameters. Recently, there has been significant interest in identifying such *sharp thresholds* for various parameter regimes.

Computationally, one might be interested to study algorithms for recovery or approximation. In the older approach, algorithms were studied to provide upper bounds on the parameter regimes for recovery or approximation. See [45] or [2, Section 5] for a summary of such results. More recently, the paradigm has shifted towards understanding the limitations and strengths of tractable methods (e.g. see [114] on semidefinite programming based methods) and assessing whether successful retrieval can be achieved by tractable algorithms at the sharp statistical thresholds or there is a *gap*. So far, it is understood that there is no such gap in the case of exact recovery (weak and strong) and approximation of binary SBM as well as the exact recovery of linear-sized communities [2]. However, this is still an open question for more general cases; e.g., see [3] and the list of unresolved conjectures therein.

The statistical-computational picture for SBM with only two equivalent communities has been fully characterized in a series of recent papers [53, 118, 116, 107, 117, 119, 1, 77]. Apart from the binary SBM, the best understood cases are where there is a finite number r of equivalent or linear-sized communities. Outside of the settings described above, the full picture has not yet emerged and many questions are unresolved.

4.1.3 This Chapter

The community detection problem studied in this chapter is stated as: given the adjacency matrix of a graph generated by the heterogenous stochastic block model, for what SBM parameters we can recover the labels of *all* vertices, with high probability, using an algorithm that has been proved to do so. We consider a convex program in (4.4) and an estimator similar to the maximum likelihood estimator in (4.5) and characterize parts of the model

space for which exact recovery is possible via these algorithms. Theorems 48 and 49 provide sufficient conditions (in terms of the SBM parameters) for exact recovery via the convex recovery program and Theorem 50 provides sufficient conditions for exact recovery via the modified maximum likelihood estimator. In Section 4.2.3, we extend the above bounds to the case of partial observations, i.e., when each entry of the matrix is observed uniformly with some probability γ and the results are recorded. We also provide an information-theoretic lower bound, describing an impossibility regime for exact recovery in heterogenous SBM in Theorem 51. All of our results hold with high probability, as this is the best one can hope for; with tiny probability the model can generate graphs like the complete graph where the partition is unrecoverable.

The results of this chapter provide a clear improvement in the understanding of stochastic block models by exploiting tradeoffs among SBM parameters. We identify a key parameter (or summary statistic), defined in (4.1) and referred to as *relative density*, which shows up in our results and provides improvements in the statistical assessment and efficient computational approaches for certain configurations of heterogenous SBM; examples are given in in Section 4.3 to illustrate a number of such beneficial tradeoffs that lead to the following

- semidefinite programming can successfully recover communities of size $\mathrm{O}(\sqrt{\log n})$ under mild conditions on other communities (see Example 3 for details) while $\log n$ has long been believed to be the threshold for the smallest community size.

- The sizes of the communities can be wide spread, or the inter- and intra-community probabilities can be very close, and the model still be efficiently recoverable, while existing methods (e.g. *peeling strategy* [4]) providing false negatives.

While these results are a step towards understanding the information-computational picture about the heterogenous SBM with a growing number of communities, we cannot comment on phase transitions or a possible information-computational gap (see Section 4.1.2) in this setup based on the results of this chapter.

4.2 Main Results

Consider the heterogenous stochastic block model described above. In the proofs, we can allow for isolated nodes (communities of size 1) which are omitted from the model here to simplify the presentation. Denote by $\mathcal{Y}$ the set of admissible adjacency matrices according to a community assignment as above, i.e.,

$$\mathcal{Y} := \{Y \in \{0,1\}^{n\times n} : \ Y \text{ is a valid community matrix w.r.t. } V_1, \ldots, V_r \text{ where } |V_k| = n_k\} .$$

Define the *relative density of a community* as

$$\rho_k = (p_k - q)n_k \tag{4.1}$$

which gives $\sum_{k=1}^{r} \rho_k = \sum_{k=1}^{r} p_k n_k - qn$. Define $n_{\min}$ and $n_{\max}$ as the minimum and maximum of $n_1, \ldots, n_k$ respectively. The total variance over the kth community is defined as $\sigma_k^2 = n_k p_k(1-p_k)$, and we let $\sigma_0^2 = nq(1-q)$. Moreover, consider

$$\sigma_{\max}^2 = \max_{k=1,\ldots,r} \ \sigma_k^2 = \max_{k=1,\ldots,r} \ n_k p_k(1-p_k) . \tag{4.2}$$

A Bernoulli random variable with parameter p is denoted by $\mathrm{Ber}(p)$, and a Binomial random variable with parameters n and p is denoted by $\mathrm{Bin}(n,p)$. The Neyman Chi-square divergence between the two discrete random variables $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$ is given by

$$\widetilde{D}(p,q) := \frac{(p-q)^2}{q(1-q)} . \tag{4.3}$$

Chi-square divergence is an instance of a more general family of divergence functions called f-divergences or Ali-Silvey distances. This family also includes KL-divergence, total variation distance, Hellinger distance and Chernoff distance as special cases. Moreover, the divergence used in [2] is an f-divergence.

Lastly, log denotes the natural logarithm (base e), and the notation $\theta \gtrsim 1$ is equivalent to $\theta \geq \mathrm{O}(1)$.

4.2.1 Convex Recovery

Inspired by the success of semidefinite programs in community detection (e.g., see [76, 114]) we consider a natural convex relaxation of the maximum likelihood estimator, similar to the one used in [47], for exact recovery of the heterogeneous SBM with a growing number of communities. Assuming that $\sum_{k=1}^{r} n_k^2$ is known, we solve

$$\begin{array}{rcll} \widehat{Y} & = & \arg\max\limits_{Y} & \sum_{i,j} A_{ij} Y_{ij} \\ & & \text{subject to} & \|Y\|_\star \leq \|Y^\star\|_\star = n \ , \ \sum_{i,j} Y_{ij} = \sum_k n_k^2 \ , \ 0 \leq Y_{ij} \leq 1 \, . \end{array} \tag{4.4}$$

where $\|\cdot\|_\star$ denotes the nuclear norm (the sum of singular values of the matrix) and $Y^\star \in \mathcal{Y}$ is the true community matrix.

We prove two theorems giving conditions under which the above convex program outputs the true community matrix with high probability. In establishing these performance guarantees, we follow the standard *dual certificate* argument in convex analysis while utilizing strong matrix concentration results from random matrix theory, e.g., [43, 151, 158, 12]. These results allow us to bound the spectral radius of the matrix $A - \mathbb{E}(A)$ where A is an instance of adjacency matrix generated under heterogenous SBM. The proofs for both theorems along with the matrix concentration bounds are given in Appendix A.

Theorem 48 *Under the heterogenous stochastic block model, the output of the semidefinite program in* (4.4) *coincides with* $Y^\star$ *with high probability, provided that*

$$\rho_k^2 \gtrsim \sigma_k^2 \log n_k \ \ , \ \ \widetilde{D}(p_{\min}, q) \gtrsim \frac{\log n_{\min}}{n_{\min}} \ \ , \ \ \rho_{\min}^2 \gtrsim \max\{\sigma_{\max}^2, \, nq(1-q), \, \log n\}$$

and $\sum_{k=1}^{r} n_k^{-\alpha} = \mathrm{o}(1)$ *for some* $\alpha > 0$ *.*

Proof Sketch. For $Y^\star$ to be the unique solution of (4.4), we need to show that for any feasible $Y \neq Y^\star$, the following quantity

$$\langle A, Y^\star - Y \rangle = \langle \mathbb{E}(A), Y^\star - Y \rangle + \langle A - \mathbb{E}(A), Y^\star - Y \rangle$$

is strictly positive. In bounding the second term above, we make use of the constraint $\|Y\|_\star \leq \|Y^\star\|_\star$ by constructing a *dual certificate* from $A - \mathbb{E}(A)$. This is where the bounds on the spectral norm (dual norm for the nuclear norm) of $A - \mathbb{E}(A)$ enter and we use matrix concentration bounds (see Lemma 7 in Appendix A).

The assumption $\sum_{k=1}^{r} n_k^{-\alpha} = \mathrm{o}(1)$ above is tantamount to saying that the number of tiny communities cannot be too large (e.g., the number of polylogarithmic-size communities cannot be a power of n). In other words, one needs to have mostly large communities (growing like n^ϵ, for some $\epsilon > 0$) for this assumption to be satisfied. Note, however, that the condition does *not* restrict the number of communities of size n^ϵ for any fixed $\epsilon > 0$. In fact, Theorem 48 allows us to describe a regime in which *tiny* communities of size $\mathrm{O}(\sqrt{\log n})$ are recoverable provided that they are very dense and that only few tiny or small communities exist; see Example 3. The second theorem imposes more stringent conditions on the relative density, hence only allowing for communities of size down to $\log n$, but relaxes the condition that only a small number of nodes can be in small communities.

Theorem 49 *Under the heterogenous stochastic block model, the output of the semidefinite program in* (4.4) *coincides with $Y^\star$, with high probability, provided that*

$$\rho_k^2 \gtrsim \sigma_k^2 \log n \quad , \quad \widetilde{D}(p_{\min}, q) \gtrsim \tfrac{\log n}{n_{\min}} \quad , \quad \rho_{\min}^2 \gtrsim \max\{\sigma_{\max}^2 \, , \, nq(1-q)\} \, .$$

The proof of Theorem 49 is similar to the proof of Theorem 48 except that we use a different matrix concentration bound (see Lemma 10 in Appendix A).

4.2.2 Recoverability Lower and Upper Bounds

Next, we consider an estimator, inspired by maximum likelihood estimation, and identify a subset of the model space which is exactly recoverable via this estimator. The proposed estimation approach is unlikely to be computationally tractable and is only used to examine the conditions for which exact recovery is possible. For a fixed $Y \in \mathcal{Y}$ and an observed

matrix A, the likelihood function is given by

$$\mathbb{P}_Y(A) = \prod_{i<j} p_{\tau(i,j)}^{A_{ij}Y_{ij}} (1-p_{\tau(i,j)})^{(1-A_{ij})Y_{ij}} q^{A_{ij}(1-Y_{ij})} (1-q)^{(1-A_{ij})(1-Y_{ij})},$$

where $\tau : \{1,\dots,n\}^2 \to \{1,\dots,r\}$ and $\tau(i,j) = k$ if and only if $i,j \in V_k$, and arbitrary in $\{1,\dots,r\}$ otherwise. The log-likelihood function is given by

$$\log \mathbb{P}_Y(A) = \sum_{i<j} \log \frac{(1-q)p_{\tau(i,j)}}{q(1-p_{\tau(i,j)})} A_{ij}Y_{ij} + \sum_{i<j} \log \frac{1-p_{\tau(i,j)}}{1-q} Y_{ij} + \text{ terms not involving } \{Y_{ij}\}.$$

Maximizing the log-likelihood involves maximizing a weighted sum of $\{Y_{ij}\}$'s where the weights depend on the (usually unknown) values of $q, p_1, \dots, p_r$. To be able to work with less information, we will use the following modification of maximum likelihood estimation, which only uses the knowledge of $n_1, \dots, n_r$,

$$\widehat{Y} = \arg\max_{Y \in \mathcal{Y}} \sum_{i,j=1}^{n} A_{ij}Y_{ij}. \tag{4.5}$$

Theorem 50 *Suppose $n_{\min} \geq 2$ and $n \geq 8$. Under the heterogenous stochastic block model, if*

$$\rho_{\min} \geq 4(17+\eta)\left(\frac{1}{3} + \frac{p_{\min}(1-p_{\min}) + q(1-q)}{p_{\min}-q}\right)\log n,$$

for some choice of $\eta > 0$, then the optimal solution $\widehat{Y}$ of the non-convex recovery program in (4.5) coincides with $Y^\star$, with a probability not less than $1 - 7\frac{p_{\max}-q}{p_{\min}-q}n^{2-\eta}$.

Similar to the proof of Theorem 48, we establish $\langle A, Y^\star - Y\rangle > 0$ for any $Y \in \mathcal{Y}$, while this time, we use a counting argument (see Lemma 11 in Appendix B) similar to the one in [47]. The complete proofs for this Theorem and the next one are given in Appendix B. Notice that $\rho_{\min} = \min_{k=1,\dots,r} n_k(p_k - q)$ and $p_{\min} = \min_{k=1,\dots,r} p_k$ do not necessarily correspond to the same community.

Finally, to provide a better picture of community detection for heterogenous SBM we provide the following sufficient conditions for when the exact recovery is impossible.

Theorem 51 *If any of the following conditions holds,*

(1) $2 \leq n_k \leq n/e$ *, and* $4\sum_{k=1}^{r} n_k^2 \widetilde{D}(p_k, q) \leq \frac{1}{2}\sum_k n_k \log \frac{n}{n_k} - r - 2$

(2) $n \geq 128$ *,* $r \geq 2$ *and* $\max_k \left\{ n_k \widetilde{D}(p_k, q) + n_k \widetilde{D}(q, p_k) \right\} \leq \frac{1}{12}\log(n - n_{\min})$

then $\inf_{\widehat{Y}} \sup_{Y^\star \in \mathcal{Y}} \mathbb{P}(\widehat{Y} \neq Y^\star) \geq \frac{1}{2}$ *where the infimum is taken over all measurable estimators* $\widehat{Y}$ *based on the realization* A *generated according to the heterogenous stochastic block model.*

4.2.3 Partial Observations

In the general stochastic block model, we assume that the entries of a symmetric adjacency matrix $A \in \{0,1\}^{n \times n}$ have been generated according to a combination of Erdős-Rényi models with parameters that depend on the true community matrix. In the case of partial observations, we assume that the entries of A has been observed independently with probability γ. In fact, every entry of the input matrix falls into one of these categories: *observed as one* denoted by Ω_1, *observed as zero* denoted by Ω_0, and *unobserved* which corresponds to Ω^c where $\Omega = \Omega_0 \cup \Omega_1$. If an estimator only takes the observed part of the matrix as the input, one can revise the underlying probabilistic model to incorporate both the stochastic block model and the observation model; i.e. a revised distribution for entries of A as

$$A_{ij} = \begin{cases} \text{Ber}(\gamma p_k) & i,j \in V_k \text{ for some } k \\ \text{Ber}(\gamma q) & i \in V_k \text{ and } j \in V_l \text{ for } k \neq l\,. \end{cases}$$

yields the same output from an estimator that only takes in the observed values. Therefore, the estimators in (4.4) and (4.5), as well as the results of Theorems 48, 49, 50, can be easily adapted to the case of partially observed graphs.

4.3 Tradeoffs in Heterogenous SBM

As it can be seen from the results presented in this chapter, and the main summary statistics they utilize (the relative densities $\rho_1, \ldots, \rho_r$), the parameters of SBM can vary significantly

and still satisfy the same recoverability conditions. In the following, we examine a number of such tradeoffs which lead to recovery guarantees for interesting SBM configurations. Here, a *configuration* is a list of community sizes n_k, their connectivity probabilities p_k, and the inter-community connectivity probability q. A triple (m, p, k) represents k communities of size m each, with connectivity parameter p. We do not worry about whether m and k are always integers; if they are not, one can always round up or down as needed so that the total number of vertices is n, without changing the asymptotics. Moreover, when the $\mathrm{O}(\cdot)$ notation is used, we mean that appropriate constants can be determined. A detailed list of computations for the examples in this section are given in Appendix D.

Table 4.1: A summary of examples in Section 4.3. Each row gives the important aspect of the corresponding example as well as whether, under appropriate regimes of parameters, it would satisfy the conditions of the theorems proved in this chapter.

	importance	convex recovery by Thm. 48	convex recovery by Thm. 49	recoverability by Thm. 50
Ex. 1	relative densities $\{\rho_k\}$ vs $(p_{\min}, n_{\min})$	×	×	✓
Ex. 2	relative densities $\{\rho_k\}$ vs $(p_{\min}, n_{\min})$	✓	✓	✓
Ex. 3	$n_{\min} = \sqrt{\log n}$	✓	×	×
Ex. 4	many small communities, $n_{\max} = \mathrm{O}(n)$	✓	✓	✓
Ex. 5	$n_{\min} = \mathrm{O}(\log n)$, spread in sizes	×	✓	✓
Ex. 6	small $p_{\min} - q$	✓	✓	✓

Better Summary Statistics It is intuitive that using summary statistics such as $(p_{\min}, n_{\min})$, for a heterogenous SBM where n_k's and p_k's are allowed to take very different values, can be very limiting. Examples 1 and 2 are intended to give configurations that are guaranteed to

be recoverable by our results but fail the existing recoverability conditions in the literature.

Example 1 Suppose we have two communities of sizes $n_1 = n - \sqrt{n}$, $n_2 = \sqrt{n}$, with $p_1 = n^{-2/3}$ and $p_2 = 1/\log n$ while $q = n^{-2/3-0.01}$. The bound we obtain here in Theorem 50 makes it clear that this case is theoretically solvable (the modified maximum likelihood estimator successfully recovers it). By contrast, Theorem 3.1 in [29] (specialized for the case of no outliers), requiring

$$n_{\min}^2 (p_{\min} - q)^2 \gtrsim (\sqrt{p_{\min} n_{\min}} + \sqrt{nq})^2 \log n\,, \tag{4.6}$$

would fail and provide no guarantee for recoverability.

Example 2 Consider a configuration as

$$(n - n^{2/3},\, n^{-1/3+\epsilon},\, 1) \;\;,\;\; (\sqrt{n},\, \mathrm{O}(\tfrac{1}{\log n}),\, n^{1/6}) \;\;,\;\; q = n^{-2/3+3\epsilon}$$

where ϵ is some small quantity, e.g., $\epsilon = 0.1$. Either of Theorems 48 and 49 verify that this case is recoverable via the semidefinite program (4.4) with high probability. By contrast, using the $p_{\min} = n^{-1/3+\epsilon}$ and $n_{\min} = \sqrt{n}$ heuristic, neither the condition of Theorem 3.1 in [29] (given in (4.6)) nor the condition of Theorem 2.5 in [47] is fulfilled, hence providing no recovery guarantee for this configuration.

4.3.1 Small communities can be efficiently recovered

Most algorithms for clustering the SBM run into the problem of small communities [46, 23, 109], often because the models employed do not allow for enough parameter variation to identify the key quantities involved. The next three examples attempt to provide an idea of how small the community sizes can be, how many small communities are allowed, and how wide the spread of community sizes can be, as characterized by our results.

Example 3 (smallest community size for convex recovery) Consider a configuration as

$$(\sqrt{\log n},\, \mathrm{O}(1),\, m) \;\;,\;\; (n_2,\, \mathrm{O}(\tfrac{\log n}{\sqrt{n}}),\, \sqrt{n}) \;\;,\;\; q = \mathrm{O}(\tfrac{\log n}{n})$$

where $n_2 = \sqrt{n} - m\sqrt{\log n/n}$ to ensure a total of n vertices. Here, we assume $m \leq n/(2\sqrt{\log n})$ which implies $n_2 \geq \sqrt{n}/2$. It is straightforward to verify the conditions of Theorem 48.

To our knowledge, *this is the first example in the literature for which semidefinite programming based recovery works and allows the recovery of (a few) communities of size smaller than* $\log n$. Previously, $\log n$ was considered to be the standard bound on the community size for exact recovery, as illustrated by Theorem 2.5 of [47] in the case of equivalent communities. We have thus shown that it is possible, in the right circumstances (when sizes are spread and the smaller the community the denser it is), to recover very small communities (up to $\sqrt{\log n}$ size), *if there are just a few of them (at most polylogarithmic in* n*).* The significant improvement we made in the bound on the size of the smallest community is due to the fact that we were able to perform a closer analysis of the semidefinite program by utilizing stronger matrix concentration bounds, mainly borrowed from [43, 151, 158, 12]. For more details, see Appendix A.2.

Notice that the condition of Theorem 50 is *not* satisfied. This is not an inconsistency (as Theorem 50 gives only an upper bound for the threshold), but indicates the limitation of this theorem in characterizing all recoverable cases.

4.3.2 Spreading the sizes

As mentioned before, while Theorem 48 allows for going lower than the standard $\log n$ bound on the community size for exact recovery, it requires the number of very small communities to be relatively small. On the other hand, Theorem 49 provides us with the option of having many small communities but requires the smallest community to be of size $\mathrm{O}(\log n)$. We explore two cases with many small communities in the following.

Example 4 Consider a configuration where small communities are dense and we have one big community,

$$(\tfrac{1}{2}n^{\epsilon},\, \mathrm{O}(1),\, n^{1-\epsilon}) \;\;,\;\; (\tfrac{1}{2}n,\, n^{-\alpha}\log n,\, 1) \;\;,\;\; q = \mathrm{O}(n^{-\beta}\log n)$$

with $0 < \epsilon < 1$ and $0 < \alpha < \beta < 1$. We are interested to see how large the number of small communities can be. Then the conditions of Theorems 48 and 49 both require that

$$\tfrac{1}{2}(1-\alpha) < \epsilon < 2(1-\alpha) \quad , \quad \epsilon > 2\alpha - \beta \tag{4.7}$$

and are depicted in Figure 4.1. Since we have not specified the constants in our results, we only consider strict inequalities.

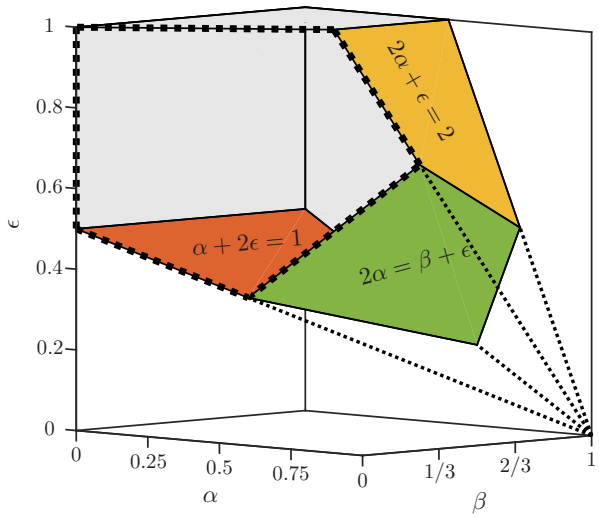


Figure 4.1: The space of parameters in Equation 4.7. The face defined by $\beta = \alpha$ is shown with dotted edges. The three gray faces in the back correspond to $\beta = 1$, $\alpha = 0$ and $\epsilon = 1$ respectively. The green plane (corresponding to the last condition in (4.7)) comes from controlling the intra-community interactions uniformly (interested reader is referred to Equations (A.8) and (A.9) in the supplement material) which might be only an artifact of our proof and can be possibly improved.

Notice that the small communities are as dense as can be, but the large one is not necessarily very dense. By picking ϵ to be just over $1/4$, we can make α just shy of $1/2$, and β very close to 1. As far as we can tell, there are no results in the literature surveyed that cover such a case, although the clever "peeling" strategy introduced in [4] would recover the largest community. The strongest result in [4] that seems applicable here is Corollary 4 (which works for non-constant probabilities). The [4] algorithm works to recover a large community (larger than $\mathrm{O}(\sqrt{n}\log^2 n)$), subject to existence of a gap in the community sizes (roughly, there should be no community sizes between $\mathrm{O}(\sqrt{n})$ and $\mathrm{O}(\sqrt{n}\log^2 n)$). Therefore, in this example, after a single iteration, the algorithm will stop, despite the continued existence of a gap, as there is no community with size above the gap. Hence the "peeling" strategy on this example would fail to recover all the communities.

Example 5 Consider a configuration with many small dense communities of size $\log n$. We are interested to see how large the spread of community sizes can be for the semidefinite program to work. As required by Theorems 48 and 49 and to control $\sigma_{\max}$ (defined in (4.2)), the larger a community the smaller its connectivity probability should be; therefore we choose the largest community at the threshold of connectivity (required for recovery). Consider the following community sizes and probabilities:

$$(\log n,\ \mathrm{O}(1),\ \tfrac{n}{\log n} - m\sqrt{\tfrac{n}{\log n}})\ \ ,\ \ (\sqrt{n\log n},\ \mathrm{O}(\sqrt{\tfrac{\log n}{n}}),\ m)\ \ ,\ \ q = \mathrm{O}(\tfrac{\log n}{n})$$

where m is a constant. Again, we round up or down where necessary to make sure the sizes are integers and the total number of vertices is n. All the conditions of Theorem 49 are satisfied and exact convex recovery is possible via the semidefinite program.

Note that the last condition of Theorem 48 is not satisfied since there are too many small communities. Also note that alternative methods proposed in the literature surveyed would not be applicable; in particular, the gap condition in [4] is not satisfied for this case from the start.

4.3.3 Weak communities are efficiently recoverable

The following examples illustrate how small $p_{\min} - q$ can be in order for the recovery, respectively, the convex recovery algorithms to still be guaranteed to work. When some p_k is very close to q, the Erdős-Rényi model $\mathcal{G}(n_k, p_k)$ looks very similar to the ambient edges from $\mathcal{G}(n, q)$. Again, we are going to exploit the possible tradeoffs in the parameters of SBM to guarantee recovery. Note that the difference in $p_{\min} - q$ for the two types of recovery is noticeable, indicating that there is a significant difference between what we know to be recoverable and what we can recover efficiently by our convex method. We consider both dense graphs (where $p_{\min}$ is $\mathrm{O}(1)$) and sparse ones.

Example 6 Consider a configuration where all of the probabilities are of $\mathrm{O}(1)$ and

$$(n_1,\, p_{\min},\, 1) \;\;,\;\; (n_{\min},\, p_2,\, 1) \;\;,\;\; (n_3,\, p_3,\, \tfrac{n-n_1-n_{\min}}{n_3}) \;\;,\;\; q = \mathrm{O}(1)$$

where $p_2 - q$ and $p_3 - q$ are $\mathrm{O}(1)$. On the other hand, we assume $p_{\min} - q = f(n)$ is small. For recoverability by Theorem 50, we need $f(n) \gtrsim (\log n)/n_{\min}$ and $f^2(n) \gtrsim (\log n)/n_1$. Notice that, since $n \gtrsim n_1 \gtrsim n_{\min}$, we should have $f(n) \gtrsim \sqrt{\log n/n}$. For the convex program to recover this configuration (by Theorem 48 or 49), we need $n_{\min} \gtrsim \sqrt{n}$ and $f^2(n) \gtrsim \max\{n/n_1^2, \log n/n_{\min}\}$, while all the probabilities are $\mathrm{O}(1)$.

Note that if all the probabilities, as well as $p_{\min} - q$, are $\mathrm{O}(1)$, then by Theorem 50 all communities down to a logarithmic size should be recoverable. However, the success of convex recovery is guaranteed by Theorems 48 and 49 when $n_{\min} \gtrsim \sqrt{n}$.

For a similar configuration to Example 6, where the probabilities are not $\mathrm{O}(1)$, recoverability by Theorem 50 requires $f(n) \gtrsim \max\{\sqrt{p_{\min}(\log n)/n}, n^{-c}\}$ for some appropriate $c > 0$.

4.4 Discussion

We have provided a series of extensions to prior works (especially [47, 2]) by considering the exact recovery for stochastic block model in its full generality with a growing number of communities. By capturing the tradeoffs among the various parameters of SBM, we have

identified interesting SBM configurations that are efficiently recoverable via semidefinite programs. However there are still interesting problems that remain open. Sharp thresholds for recovery or approximation of heterogenous SBM, models for partial observation (non-uniform, based on prior information, or adaptive as in [165]), overlapping communities (e.g., [2]), as well as considering outlier nodes (e.g., [29]) are important future directions.

Chapter 5

DISCUSSIONS AND FUTURE DIRECTIONS

5.1 Simultaneously Structured Models

In Chapter 2, we considered the problem of recovery of a simultaneously structured object from limited measurements. It is common in practice to combine known norm penalties corresponding to the individual structures (also known as regularizers in statistics and machine learning applications), and minimize this combined objective in order to recover the object of interest. The common use of this approach motivated us to analyze its performance, in terms of the smallest number of generic measurements needed for correct recovery. We showed that, under a certain assumption on the norms involved, the combined penalty requires more generic measurements than one would expect based on the degrees of freedom of the desired object. Our lower bounds on the required number of measurements implies that the combined norm penalty cannot perform significantly better than the best individual norm.

These results raise several interesting questions, and lead to directions for future work. We briefly outline some of these directions, as well as connections to some related problems.

Defining new atoms for simultaneously structured models Our results show that combinations of individual norms do not exhibit a strong recovery performance. On the other hand, the seminal paper [42] proposes a general construction for an appropriate penalty given a set of atoms. Can we revisit a simultaneously structured recovery problem, and define new atoms that capture all structures at the same time? And can we obtain a new norm penalty induced by the convex hull of the atoms? Abstractly, the answer is yes, but such convex hulls may be hard to characterize, and the corresponding penalty may not be efficiently

computable. It is interesting to find special cases where this construction can be carried out and results in a tractable problem. Recent developments in this direction include the "square norm" proposed by [120] for the low-rank tensor recovery which provably outperforms (2.2) for Gaussian measurements and the (k, q)-trace norm introduced by Richard et al to estimate sparse and low-rank matrices [132].

Algorithms for minimizing combination of norms Despite the limitation in their theoretical performance, in practice one may still need to solve convex relaxations that combine the different norms, i.e., problem (2.3). Consider the special case of sparse and low-rank matrix recovery. All corresponding optimization problems mentioned in Theorem 10 can be expressed as a semidefinite program and solved by standard solvers; for example, for the numerical experiments in Section 2.7 we used the interior-point solver SeDuMi [147]. However, faster and more scalable algorithms can be studied and it is an active area of research.

Other simultaneously structured models The presented results in Chapter 2 can be used for any set of simultaneous structures in theory. We would like to identify other cases considering new simultaneously structured models, beyond simultaneously sparse and low rank matrices, and study the corresponding estimation problems. For example, signals that are simultaneously sparse and smooth (entries vary slowly, i.e., the signal can be approximated by a piecewise constant function) have been considered in [149], and signals that are simultaneously sparse and have only a few distinct nonzero values have been considered in [87].

Measurement scenarios that are compatible with the structures. Our results indicate a fundamental limitation in utilizing combinations of norms for recovery of simultaneously structured models from *generic* measurements (see Section 2.4). On the other hand, Bahmani and Romberg in [10] consider *a nested measurement scenario* and study a

simple two-stage algorithm (consecutively minimizing $\ell_{1,2}$ and nuclear norms) that is nearly minimax optimal in recovery from such nested measurements. More specifically, given a simultaneously low-rank and row-sparse matrix X_0, they assume observations of the form $\mathcal{W}(\Psi X_0)$ where $\mathcal{W}$ is a restricted isometry for low-rank matrices and Ψ is a restricted isometry for row-sparse matrices, with high probability. Such a specific measurement scenario has allowed for providing a nearly minimax optimal recovery procedure. It is of great interest to understand the benefits and limitations of this approach and whether we can achieve recovery from optimal number of measurements by devising more practical measurement scenarios.

5.2 Variational Gram Functions

In Chapter 3, we introduce variational Gram functions which include many existing regularization functions as well as important new ones. Convexity properties of this class, conjugate functions, subdifferentials, semidefinite representability, proximal operators, and other convex analysis properties are studied. By adapting the mirror-prox method [125], we provide a general and efficient optimization algorithm for VGF-regularized loss minimization problems. We establish a general kernel trick and a representer theorem for such problems. Finally, the effectiveness of VGF regularization as well as the efficiency of our optimization approach is illustrated by a numerical example on hierarchical classification for text categorization. There are numerous directions for future research on this class of functions. In the following, we discuss some of these directions.

Left unitary invariance and VGFs Every variational Gram function is left unitary invariant as it is a function of the Gram matrix. Recall that every VGF is also a squared seminorm. It is not clear whether every left unitary invariant squared seminorm can be represented as a variational Gram function. Notice that we are not asking for deriving the representation (i.e., the corresponding set $\mathcal{M}$) but just assessing the possibility for such a representation.

It is not even clear whether every conjugate VGF can be represented as a VGF. More

specifically, we ask: given a set $\mathcal{M}$ and its associated VGF $\Omega_{\mathcal{M}}$, whether there exists a set $\mathcal{M}'$ for which $\Omega^{\star}_{\mathcal{M}} \equiv \Omega_{\mathcal{M}'}$? Notice that a positive answer to the first question automatically provides a positive answer to the second one. Observe that, in the simple case of $n = 1$ when X is a row vector, every squared vector semi-norm can be represented both as a VGF and as a conjugate VGF; see Section 3.2 for the former and [9] for the latter.

The interaction between the loss and invariances of the penalty In Section 3.5.2, we were able to derive a reduced problem by using the Fenchel-type representation of the loss as well as the left unitary invariance of the penalty. Exploiting other invariance properties of regularization functions in conjunction with the variational representations of the loss functions can provide us with new reduced forms and is a subject of future work.

Statistical guarantees for VGFs Statistical properties of VGFs, for example the corresponding sample complexity, are of interest from a learning theory perspective.

New variational penalties for tensors It is interesting to use the variational penalties machinery (or the composite penalties) for tensor recovery and study new structures for tensors.

Recall the definition of a variational Gram function as $\Omega_{\mathcal{M}}(X) = \sup_{M \in \mathcal{M}} \operatorname{tr}(XMX^T)$. This definition can be expressed in terms of $x = \operatorname{vec}(X)$ as

$$\Omega_{\mathcal{M}}(x) = \sup_{M \in \mathcal{M}} x^T (M \otimes I) x \tag{5.1}$$

where $\otimes$ denotes the Kronecker product. With this expression, we can view VGFs from the point of view of the specific quadratic functions they employ; quadratic forms $x^T Q x$ with matrices from

$$\left\{ Q : \; Q = M \otimes I \; , \; M \in \mathcal{M} \right\}.$$

Other structured sets of quadratic forms provide us with other functions. Such an approach, of vectorizing and employing certain weight matrices, can be taken for tensors depending on the applications of interest.

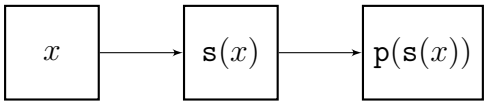


Figure 5.1: Composite penalties. `s`, often a smooth map, transforms the original structured model to one with another structure. `p`, often a non-smooth real-valued map, penalizes `s` according to the latter structure, hence providing a penalization scheme for the original structure. Such a decomposition provides us with more efficient oracles to be used in devising optimization algorithms.

Other composite regularizers We use the term *composite penalty* for a regularizer of the form

$$\Omega(x) = \mathtt{p}(\mathtt{s}(x)) \tag{5.2}$$

where

- `p` is the *outer penalty function*, and,
- `s` is the *structure mapping*, a smooth mapping to change the notion of parsimony.

Besides the linear composite penalties and the variational Gram functions, it is interesting to see which examples come up in applications. In the following, we list a number of instances of the penalty functions that can be constructed via the definition in (5.2). While (5.2) allows for a unified view towards some existing penalties, it also allows for construction of new ones in a systematic manner with efficient optimization tools.

Composition of a norm with a linear transformation is very common in statistical learning. The case of ℓ_1 norm,

$$f(x) = \|Ax\|_1 \,, \tag{5.3}$$

covers many prominent examples such as fused lasso [149] (or total variation norm) for smoothing, generalized lasso [150] (e.g., trend filtering [95]), or when dealing with heteroscedasticity of the noise [17]. Interestingly, the OSCAR norm [21] can also be represented similarly, using $\mathtt{p}(x) = \|x\|_1$, as

$$\|x\|_{\mathrm{OSCAR}} = \sum_{i<j} \max\{\, |x_i| \,,\, |x_j| \,\} = \|D_{\mathrm{OSCAR}} x\|_1 \tag{5.4}$$

where D_{OSCAR} is an appropriately scaled incidence matrix.

The composition of a unitarily invariant norm with a linear transformation of a matrix,

$$f(X) = \|\mathcal{A}(X)\| \,, \tag{5.5}$$

has also been shown to be powerful in different applications. As an example, [60] composes the nuclear norm with a linear transformation, as $f(X) = \|X(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\|_\star$, which is used in multitask learning. Another important example of such regularization has been used for recovery of complex exponential signals from time domain samples, where by converting spectral sparsity in the model into the low-rank structure of a Hankel matrix, one can use $\|H(x)\|_\star$ for regularization; $[H(x)]_{j,k} = x_{j+k}$. Such an approach has been taken in [28] for the robust recovery of a superposition of a number of distinct complex exponential functions from a few random Gaussian projections, and in [48] for devising an enhanced matrix completion algorithm (EMaC) for frequency estimation from limited time samples, among others.

Such a composite structure provides us with new opportunities for efficient optimization. For example, [8] compute the proximity operator for linear composite regularizers from the solution of a certain fixed point problem. See [65] for optimization techniques for low-rank Hankel matrix recovery. If A satisfies $AA^T = \nu I$ (for some $\nu > 0$) then the proximal mapping has a closed form solution [51, Table I].

The composition of a penalty $\mathtt{p}$ with the Gram matrix mapping $\mathtt{s}(X) = X^T X$ provides the variational Gram functions (VGF) introduced in Chapter 3. By representing a number of seemingly complicated penalties (e.g., those with a proximal mapping that cannot be evaluated in closed form or by a cheap iterative algorithm) as a VGF (e.g., see Section 3.2),

the composite structure can be exploited for deriving efficient optimization techniques; e.g., see Section 3.5.1.

The term "composite regularizer" is sometimes used to refer to linear or other combinations of a number of regularizers, i.e., $f(x) = \sum_k w_k r_k(x)$, which are out of the scope of our discussion. For example, overlapping group lasso [167], the combination of ℓ_1 and nuclear norm in [133] for recovery of simultaneously sparse and low rank matrices, or the approach in [120] for low rank tensor recovery, can be described as above. There are many different algorithms for optimizing such composite regularizers; e.g. see [168]. More complicated combinations of regularizers have also been considered in the statistical learning literature. Composite Absolute Penalties (CAP) [167] map a vector to an array of different norms of its blocks and computes a norm of the resulting array of values. CAP penalties include hierarchical norms, norms for overlapping groups with a hierarchical structure (see [9, 91] for other variants). Similarly, a common approach in learning simultaneously structured models is to combine suitable regularizers for each structure. Such combinations, in the form of $h\big(\|\cdot\|_{(1)}, \ldots, \|\cdot\|_{(\tau)}\big)$ where h is any increasing function, has been analyzed in Chapter 2.

5.3 Community Detection in Heterogenous Stochastic Block Model

We have provided a series of extensions to prior works (especially [47, 2]) by considering the exact recovery for stochastic block model in its full generality with a growing number of communities. By capturing the tradeoffs among the various parameters of SBM, we have identified interesting SBM configurations that are efficiently recoverable via semidefinite programs. However there are still interesting problems that remain open. Sharp thresholds for recovery or approximation of heterogenous SBM (such as the ones in [2, 3] for special cases of SBM), models for partial observation (non-uniform, based on prior information, or adaptive as in [165]), overlapping communities (e.g., [2]), as well as considering outlier nodes (e.g., [29]) are important future directions. We elaborate on these directions in the sequel.

Models for partial observation We considered the case where a subset of the edges in the underlying graph were observed uniformly at random. In practice, however, the observed edges are often not uniformly sampled, and care will be needed to model the effect of nonuniform sampling. Also, in many practical problems, the observed edges may be chosen by the algorithm based on some prior information (non-adaptive), or based on observations made so far (adaptive); e.g., see Yun and Proutiere [165]. It will be interesting to examine what the algorithms can achieve in these scenarios.

Overlapping communities SBMs with overlapping communities represent a more realistic model than the non-overlapping case; it has been shown that the large social and information network community structure is quite complex and that very large communities tend to have significant overlap. Only a few references in the literature have considered this problem (e.g., [2]), and there are many open questions on recovery regimes and algorithms. It would be interesting to develop a convex optimization-based algorithm for recovery of models generated by SBM with overlapping communities.

Outlier nodes A practically important extension to the SBM is to allow for adversarial outlier nodes. Cai and Li in [29] proposed a semidefinite program that can recover the communities in an SBM in the presence of outliner nodes connected to other nodes in an arbitrary way, provided that the number of outliers is small enough. Their result is comparable to the best known results in the case of equal-sized communities and equal probabilities. However, their complexity results are still parametrized by $p_{\min}$ and $n_{\min}$, which excludes useful examples, as discussed in Section 4.1.3. Extending our results to the setting of [29] is a direction for future work.

BIBLIOGRAPHY

[1] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory*, 62(1):471–487, 2016.

[2] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015.

[3] Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015.

[4] Nir Ailon, Yudong Chen, and Huan Xu. Breaking the small cluster barrier of graph clustering. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, volume 28 of *JMLR Proceedings*, pages 995–1003. JMLR.org, 2013.

[5] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

[6] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, page iau005, 2014.

[7] Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems*, pages 1457–1465, 2012.

[8] Andreas Argyriou, Charles A Micchelli, Massimiliano Pontil, Lixin Shen, and Yuesheng Xu. Efficient first order methods for linear composite regularizers. *arXiv preprint arXiv:1104.1436*, 2011.

[9] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[10] Sohail Bahmani and Justin Romberg. Near-optimal estimation of simultaneously sparse and low-rank matrices from nested linear measurements. *Information and Inference*, 2016.

[11] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 1993.

[12] Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506, 2016.

[13] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. With a foreword by Hédy Attouch.

[14] Amir Beck and Yonina C Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[15] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[16] Amir Beck and Marc Teboulle. Gradient-based algorithms with applications to signal-recovery problems. In *Convex optimization in signal processing and communications*, pages 42–88. Cambridge Univ. Press, Cambridge, 2010.

[17] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root Lasso in nonparametric regression. *Ann. Statist.*, 42(2):757–788, 2014.

[18] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2009.

[19] Dimitri P Bertsekas, Angelia Nedić, and Asuman E Ozdaglar. *Convex analysis and optimization*. Athena Scientific Belmont, 2003.

[20] Badri Narayan Bhaskar and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 261–268. IEEE, 2011.

[21] Howard D. Bondell and Brian J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123, 322–323, 2008.

[22] Kathleen H. V. Booth and D. R. Cox. Some systematic supersaturated designs. *Technometrics*, 4:489–495, 1962.

[23] Ravi B Boppana. Eigenvalues and graph bisection: An average-case analysis. In *Foundations of Computer Science, 1987., 28th Annual Symposium on*, pages 280–285. IEEE, 1987.

[24] Jonathan M Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

[25] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[26] James V. Burke and Tim Hoheisel. Matrix support functionals for inverse problems, regularization, and learning. *SIAM J. Optim.*, 25(2):1135–1159, 2015.

[27] Fennell Burns, David Carlson, Emilie Haynsworth, and Thomas Markham. Generalized inverse formulas using the schur complement. *SIAM Journal on Applied Mathematics*, 26(2):254–259, 1974.

[28] Jian-Feng Cai, Xiaobo Qu, Weiyu Xu, and Gui-Bo Ye. Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction. *Applied and Computational Harmonic Analysis*, 2016.

[29] T. Tony Cai and Xiaodong Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Statist.*, 43(3):1027–1059, 2015.

[30] Xinzhong Cai and Xinmao Wang. A note on the positive semidefinite minimum rank of a sign pattern matrix. *Electron. J. Linear Algebra*, 26:345–356, 2013.

[31] Emmanuel J Candès, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM J. Imaging Sci.*, 6(1):199–225, 2013.

[32] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[33] Emmanuel J Candès and Yaniv Plan. A probabilistic and ripless theory of compressed sensing. *Information Theory, IEEE Transactions on*, 57(11):7235–7254, 2011.

[34] Emmanuel J Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

[35] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[36] Emmanuel J Candès and Benjamin Recht. Simple bounds for recovering low-complexity models. *Math. Program.*, 141(1-2, Ser. A):577–589, 2013.

[37] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

[38] Emmanuel J Candès, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[39] Emmanuel J Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[40] Emmanuel J Candès and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[41] Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1610–1613. IEEE, 2010.

[42] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[43] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 2015.

[44] Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

[45] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. *J. Mach. Learn. Res.*, 15:2213–2238, 2014.

[46] Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In *Advances in neural information processing systems*, pages 2204–2212, 2012.

[47] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *Journal of Machine Learning Research*, 17(27):1–57, 2016.

[48] Yuxin Chen and Yuejie Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Trans. Inform. Theory*, 60(10):6576–6601, 2014.

[49] Ching-Shui Cheng. $E(s^2)$-optimal supersaturated designs. *Statist. Sinica*, 7(4):929–939, 1997.

[50] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combin. Probab. Comput.*, 19(2):227–284, 2010.

[51] Patrick L. Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, volume 49 of *Springer Optim. Appl.*, pages 185–212. Springer, New York, 2011.

[52] Noel Cressie and Timothy RC Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 440–464, 1984.

[53] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[54] Ofer Dekel, Joseph Keshet, and Yoram Singer. Large margin hierarchical classification. In *Proceedings of the 21st International Conference on Machine Learning*, pages 27–34, 2004.

[55] David L Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995.

[56] David L Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[57] David L Donoho and Matan Gavish. Minimax risk of matrix denoising by singular value thresholding. *Ann. Statist.*, 42(6):2413–2440, 2014.

[58] David L Donoho and Jared Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9446–9451, 2005.

[59] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

[60] John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.*, 10:2899–2934, 2009.

[61] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[62] Yonina C Eldar and Shahar Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[63] Yonina C Eldar and Moshe Mishali. Robust recovery of signals from a structured union of subspaces. *Information Theory, IEEE Transactions on*, 55(11):5302–5316, 2009.

[64] Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.*, 6:615–637, 2005.

[65] Maryam Fazel, Ting Kei Pong, Defeng Sun, and Paul Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM J. Matrix Anal. Appl.*, 34(3):946–977, 2013.

[66] James R Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

[67] Rina Foygel and Lester Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *Information Theory, IEEE Transactions on*, 60(2):1223–1247, 2014.

[68] Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[69] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.

[70] Ralph W Gerchberg. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237, 1972.

[71] Mohammad Golbabaee and Pierre Vandergheynst. Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 2741–2744. Ieee, 2012.

[72] Mohammad Golbabaee and Pierre Vandergheynst. Joint trace/tv norm minimization: A new efficient approach for spectral compressive imaging. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 933–936. IEEE, 2012.

[73] Yehoram Gordon. *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$. Springer, 1988.

[74] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[75] Lars Grasedyck, Daniel Kressner, and Christine Tobler. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78, 2013.

[76] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *arXiv preprint arXiv:1411.4686*, 2014.

[77] Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*, 2014.

[78] Robert W Harrison. Phase problem in crystallography. *JOSA A*, 10(5):1046–1055, 1993.

[79] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[80] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms. I.* Springer-Verlag, Berlin, 1993. Fundamentals.

[81] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[82] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[83] Norman E Hurt. *Phase Retrieval and Zero Crossings: Mathematical Methods in Image Reconstruction*, volume 52. Springer, 2001.

[84] Rishabh K Iyer and Jeff A Bilmes. Submodular point processes with applications to machine learning. In *AISTATS*, 2015.

[85] Laurent Jacob, Francis Bach, and Jean-Philippe Vert. Clustered multi-task learning: A convex formulation. In *NIPS*, volume 21, pages 745–752, 2008.

[86] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Recovery of sparse 1-d signals from the magnitudes of their fourier transform. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium On*, pages 1473–1477. IEEE, 2012.

[87] Amin Jalali and Maryam Fazel. A convex method for learning d-valued models. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 1123–1126. IEEE, 2013.

[88] Amin Jalali, Qiyang Han, Ioana Dumitriu, and Maryam Fazel. Relative density and exact recovery in heterogeneous stochastic block models. *arXiv preprint arXiv:1512.04937*, 2015.

[89] Amin Jalali, Lin Xiao, and Maryam Fazel. Variational gram functions: Convex analysis and optimization. *arXiv preprint arXiv:1507.04734*, 2015.

[90] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1629–1636. IEEE, 2014.

[91] Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12:2777–2824, 2011.

[92] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–1386, 2004.

[93] Anatoli Juditsky and Arkadi Nemirovski. First-order methods for nonsmooth convex large-scale optimization, II: Utilizing problems's structure. In *Optimization for Machine Learning*, chapter 6, pages 149–184. The MIT Press, 2011.

[94] Anatoli Juditsky and Arkadi Nemirovski. Solving variational inequalities with monotone operators on domains given by linear minimization oracles. *Math. Program.*, 156(1-2, Ser. A):221–256, 2016.

[95] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. l_1 trend filtering. *SIAM Rev.*, 51(2):339–360, 2009.

[96] G. M. Korpelevič. An extragradient method for finding saddle points and for other problems. *Èkonom. i Mat. Metody*, 12(4):747–756, 1976.

[97] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

[98] Can M Le, Elizaveta Levina, and Roman Vershynin. Sparse random graphs: regularization and concentration of the laplacian. *arXiv preprint arXiv:1502.03049*, 2015.

[99] Jure Leskovec, Kevin J Lang, and Michael Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web*, pages 631–640. ACM, 2010.

[100] Adrian S. Lewis. The convex analysis of unitarily invariant matrix functions. *J. Convex Anal.*, 2(1-2):173–183, 1995.

[101] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[102] Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

[103] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):208–220, 2013.

[104] Yue M Lu and Martin Vetterli. Sparse spectral factorization: Unicity and reconstruction algorithms. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5976–5979. IEEE, 2011.

[105] Jonathan Malkin and Jeff Bilmes. Ratio semi-definite classifiers. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4113–4116. IEEE, 2008.

[106] Bernard Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Française Informat. Recherche Opérationnelle*, 4(Ser. R-3):154–158, 1970.

[107] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

[108] Andrew M McDonald, Massimiliano Pontil, and Dimitris Stamos. New perspectives on k-support and cluster norms. *arXiv preprint arXiv:1403.1481*, 2014.

[109] Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[110] Marina Meila and Jianbo Shi. A random walks view of spectral segmentation. 2001.

[111] Charles A. Micchelli, Jean M. Morales, and Massimiliano Pontil. Regularizers for structured sparsity. *Adv. Comput. Math.*, 38(3):455–489, 2013.

[112] RP Millane. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.

[113] Leonid Mirsky. A trace inequality of John von Neumann. *Monatsh. Math.*, 79(4):303–306, 1975.

[114] Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs. *arXiv preprint arXiv:1504.05910*, 2015.

[115] Jean-Jacques Moreau. Décomposition orthogonale dun espace hilbertien selon deux cônes mutuellement polaires. *CR Acad. Sci. Paris*, 255:238–240, 1962.

[116] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. arXiv preprint arXiv:1311.4115, 2013.

[117] Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *JMLR: Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 356–370, 2014.

[118] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, pages 1–31, 2014.

[119] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 69–75. ACM, 2015.

[120] Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870*, 2013.

[121] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[122] D Needell and R Ward. Near-optimal compressed sensing guarantees for total variation minimization. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 22(10):3941, 2013.

[123] Deanna Needell and Rachel Ward. Stable image reconstruction using total variation minimization. *SIAM Journal on Imaging Sciences*, 6(2):1035–1058, 2013.

[124] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 2012.

[125] Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, 15(1):229–251 (electronic), 2004.

[126] Yu. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1, Ser. B):125–161, 2013.

[127] Henrik Ohlsson, Allen Y Yang, Roy Dong, and Shankar S Sastry. Compressive phase retrieval from squared output measurements via semidefinite programming. *arXiv preprint arXiv*, 1111, 2011.

[128] Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C. Eldar, and Babak Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Trans. Inform. Theory*, 61(5):2886–2908, 2015.

[129] Samet Oymak, Amin Jalali, Maryam Fazel, and Babak Hassibi. Noisy estimation of simultaneously structured models: Limitations of convex relaxation. In *52nd IEEE Conference on Decision and Control*, pages 6019–6024. IEEE, 2013.

[130] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized lasso: A precise analysis. *arXiv preprint arXiv:1311.0830*, 2013.

[131] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[132] Emile Richard, Guillaume R Obozinski, and Jean-Philippe Vert. Tight convex relaxations for sparse matrix factorization. In *Advances in neural information processing systems*, pages 3284–3292, 2014.

[133] Emile Richard, Pierre-Andre Savalle, and Nicolas Vayatis. Estimation of simultaneously sparse and low rank matrices. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1351–1358, 2012.

[134] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[135] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization*, 14(5):877–898, 1976.

[136] R Tyrrell Rockafellar. *Convex analysis*. Princeton university press, 1997.

[137] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

[138] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[139] Bernardino Romera-Paredes, Andreas Argyriou, Nadia Berthouze, and Massimiliano Pontil. Exploiting unrelated tasks in multi-task learning. In *International Conference on Artificial Intelligence and Statistics*, pages 951–959, 2012.

[140] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Computational learning theory (Amsterdam, 2001)*, volume 2111 of *Lecture Notes in Comput. Sci.*, pages 416–426. Springer, Berlin, 2001.

[141] Oguz Semerci, Ning Hao, Misha E Kilmer, and Eric L Miller. Tensor-based formulation and nuclear norm regularization for multi-energy computed tomography. *arXiv preprint arXiv:1307.5348*, 2013.

[142] Yoav Shechtman, Amir Beck, and Yonina C. Eldar. GESPAR: efficient phase retrieval of sparse signals. *IEEE Trans. Signal Process.*, 62(4):928–938, 2014.

[143] Yoav Shechtman, Yonina C Eldar, Alexander Szameit, and Mordechai Segev. Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. *Optics express*, 19(16):14807–14822, 2011.

[144] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[145] Pablo Sprechmann, Ignacio Ramirez, Guillermo Sapiro, and Yonina C Eldar. C-hilasso: A collaborative hierarchical sparse modeling framework. *Signal Processing, IEEE Transactions on*, 59(9):4183–4198, 2011.

[146] Mihailo Stojnic, Farzad Parvaresh, and Babak Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *Signal Processing, IEEE Transactions on*, 57(8):3075–3085, 2009.

[147] Jos F Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization methods and software*, 11(1-4):625–653, 1999.

[148] A Szameit, Yoav Shechtman, E Osherovich, E Bullkich, P Sidorenko, H Dana, S Steiner, Ernst B Kley, S Gazit, T Cohen-Hyams, et al. Sparsity-based single-shot subwavelength coherent diffractive imaging. *Nature materials*, 11(5):455–459, 2012.

[149] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005.

[150] Ryan J. Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *Ann. Statist.*, 39(3):1335–1371, 2011.

[151] Dan-Cristian Tomozei and Laurent Massoulié. Distributed user profiling via spectral methods. *Stoch. Syst.*, 4(1):1–43, 2014.

[152] Joel A Tropp. On the conditioning of random subdictionaries. *Applied and Computational Harmonic Analysis*, 25(1):1–24, 2008.

[153] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[154] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[155] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

[156] Roman Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling theory, a renaissance*, Appl. Numer. Harmon. Anal., pages 3–66. Birkhäuser/Springer, Cham, 2015.

[157] Kevin Vervier, Pierre Mahé, Alexandre D'Aspremont, Jean-Baptiste Veyrieras, and Jean-Philippe Vert. On learning matrices with orthogonal columns or disjoint supports. In *Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

[158] Van Vu. A simple svd algorithm for finding hidden partitions. *arXiv:1404.3918*, 2014.

[159] Adriaan Walther. The question of phase retrieval in optics. *Journal of Modern Optics*, 10(1):41–49, 1963.

[160] G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.

[161] Jason Weston and Chris Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the 6th European Symposium on Artificial Neural Networks (ESANN)*, pages 219–224, 1999.

[162] John Wright, Arvind Ganesh, Kerui Min, and Yi Ma. Compressive principal component pursuit. *Information and Inference*, 2(1):32–68, 2013.

[163] Jiaming Xu, Rui Wu, Kai Zhu, Bruce Hajek, R Srikant, and Lei Ying. Jointly clustering rows and columns of binary matrices: Algorithms and trade-offs. In *The 2014 ACM international conference on Measurement and modeling of computer systems*, pages 29–41. ACM, 2014.

[164] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[165] Se-Young Yun and Alexandre Proutiere. Community detection via random and adaptive sampling. In *Proceedings of The 27th Conference on Learning Theory*, pages 138–175, 2014.

[166] Anderson Y Zhang and Harrison H Zhou. Minimax rates of community detection in stochastic block model. *arXiv preprint arXiv:1507.05313*, 2015.

[167] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.*, 37(6A):3468–3497, 2009.

[168] Wenliang Zhong and James Tin-Yau Kwok. Accelerated stochastic gradient method for composite regularization.

[169] Dengyong Zhou, Lin Xiao, and Mingrui Wu. Hierarchical classification via orthogonal transfer. *Proceedings of the 28th International Conference on Machine Learning (ICML)*, June 2011.

Appendix A

SUPPLEMENT TO CHAPTER 2

A.1 Proofs for Section 2.3.2

Using the general framework of Section 2.3.1, we now prove Theorem 10, which states various convex and nonconvex recovery results for the S&L models. We start with proofs of the convex recovery.

A.1.1 Convex Recovery Results for S&L

In this section, we prove the statements of Theorem 10 regarding convex approaches, using Theorem 6 and Proposition 28. We make use of the decomposable norms to obtain better lower bounds. Hence, we first state a result on the sign vectors and the supports of the S&L model following Lemma 30. The proof is provided in Appendix A.3.

Lemma 52 *Denote the norm* $\|X^T\|_{1,2}$ *by* $\|\cdot^T\|_{1,2}$. *Given a matrix* $X_0 \in \mathbb{R}^{d_1 \times d_2}$, *let* $\mathbf{E}_\star, \mathbf{E}_c, \mathbf{E}_r$ *and* $T_\star, T_c, T_r$ *be the sign vectors and supports for the norms* $\|\cdot\|_\star$, $\|\cdot\|_{1,2}$, $\|\cdot^T\|_{1,2}$ *respectively. Then,*

- $\mathbf{E}_\star, \mathbf{E}_r, \mathbf{E}_c \in T_\star \cap T_c \cap T_r$,
- $\langle \mathbf{E}_\star, \mathbf{E}_r \rangle \geq 0$, $\langle \mathbf{E}_\star, \mathbf{E}_c \rangle \geq 0$, *and* $\langle \mathbf{E}_c, \mathbf{E}_r \rangle \geq 0$.

Proof of Theorem 10: Convex cases

Proof of (a1) We use the functions $\|\cdot\|_{1,2}, \|\cdot^T\|_{1,2}$ and $\|\cdot\|_\star$ without the cone constraint, i.e., $\mathcal{C} = \mathbb{R}^{d_1 \times d_2}$. We will apply Proposition 28 with $\mathcal{R} = T_\star \cap T_c \cap T_r$. From Lemma 52 all the sign vectors lie on $\mathcal{R}$ and they have pairwise nonnegative inner products. Consequently,

applying Proposition 32

$$\rho(\mathcal{R}, \partial f(X_0))^2 \geq \frac{1}{3}\min\{\frac{k_1}{d_1}, \frac{k_2}{d_2}, \frac{r}{\min\{d_1, d_2\}}\}.$$

If $m < \dim(\mathcal{R})$, then we have failure with probability 1. Hence, assume $m \geq \dim(\mathcal{R})$. Now, apply Proposition 28 with the given m_{low}.

Proof of (b1) In this case, we apply Lemma 62. We choose $\mathcal{R} = T_\star \cap T_c \cap T_r \cap \mathbb{S}^n$, the norms are the same as in the general model, and $\upsilon \geq \frac{1}{\sqrt{2}}$. Also, pairwise inner products are positive, hence, using Proposition 32, $\rho(\mathcal{R}, \partial f(X_0))^2 \geq \frac{1}{4}\min\{\frac{k}{d}, \frac{r}{d}\}$. Again, we may assume $m \geq \dim(\mathcal{R})$. Finally, based on Corollary 60, for the PSD cone we have $\bar{\mathbf{D}}(\mathcal{C}) \geq \frac{\sqrt{3}}{2}$. The result follows from Proposition 28 with the given m_{low}.

Proof of (c1) For the PSD cone, $\bar{\mathbf{D}}(\mathcal{C}) \geq \frac{\sqrt{3}}{2}$ and we simply use Theorem 6 to obtain the result by using $\kappa_{\ell_1}^2 = \frac{\|\bar{X}_0\|_1^2}{d^2}$ and $\kappa_\star^2 = \frac{\|\bar{X}_0\|_\star^2}{d}$.

Proof of Corollary 12

To show this, we use Theorem 3 and substitute κ's corresponding to ℓ_1 and the nuclear norm. $\kappa_\star = \frac{\|\bar{X}_0\|_\star}{\sqrt{d}}$ and $\kappa_{\ell_1} = \frac{\|\bar{X}_0\|_{\ell_1}}{d}$. Also observe that, $\lambda_{\ell_1} L_{\ell_1} = \beta d$ and $\lambda_\star L_\star = (1-\beta)d$. Hence, $\sum_{i=1}^2 \bar{\lambda}_i \kappa_i = \alpha\|\bar{X}_0\|_1 + (1-\alpha)\|\bar{X}_0\|_\star\sqrt{d}$. We then use Proposition 16 to conclude with sufficiently small $c_1, c_2 > 0$.

A.1.2 Nonconvex Recovery Results for S&L

While Theorem 10 states the result for Gaussian measurements, we prove the nonconvex recovery for the more general sub-gaussian measurements. We first state a lemma that will be useful in proving the nonconvex results. The proof is provided in Appendix A.4 and uses standard arguments.

Lemma 53 *Consider the set of matrices M in $\mathbb{R}^{d_1 \times d_2}$ that are supported over an $s_1 \times s_2$ submatrix with rank at most q. There exists a constant $c > 0$ such that whenever $m \geq$*

$c\min\{(s_1+s_2)q, s_1\log\frac{d_1}{s_1}, s_2\log\frac{d_2}{s_2}\}$, *with probability* $1-2\exp(-cm)$, $\mathcal{A}(\cdot):\mathbb{R}^{d_1\times d_2}\to\mathbb{R}^m$ *with i.i.d. zero-mean and isotropic sub-gaussian rows satisfies*

$$\mathcal{A}(X)\neq 0, \quad \textit{for all} \quad X\in M. \tag{A.1}$$

Proof of Theorem 10: Nonconvex cases

Denote the sphere in $\mathbb{R}^{d_1\times d_2}$ with unit Frobenius norm by $\mathcal{S}^{d_1\times d_2}$.

Proof of (a2) Observe that the function $f(X)=\frac{\|X\|_{0,2}}{\|X_0\|_{0,2}}+\frac{\|X^T\|_{0,2}}{\|X_0^T\|_{0,2}}+\frac{\mathrm{rank}(X)}{\mathrm{rank}(X_0)}$ satisfies the triangle inequality and we have $f(X_0)=3$. Hence, if all null space elements $\mathbf{W}\in\mathrm{Null}(\mathcal{A})$ satisfy $f(\mathbf{W})>6$, we have

$$f(X)\geq f(X-X_0)-f(-X_0)>3,$$

for all feasible X which implies X_0 being the unique minimizer.

Consider the set M of matrices, which are supported over a $6k_1\times 6k_2$ submatrix with rank at most $6r$. Observe that any $\mathbf{Z}$ satisfying $f(\mathbf{Z})\leq 6$ belongs to M. Hence ensuring $\mathrm{Null}(\mathcal{A})\cap M=\{0\}$ would ensure $f(\mathbf{W})>6$ for all $\mathbf{W}\in\mathrm{Null}(\mathcal{A})$. Since M is a cone, this is equivalent to $\mathrm{Null}(\mathcal{A})\cap(M\cap\mathcal{S}^{d_1\times d_2})=\emptyset$. Now, applying Lemma 53 with set M and $s_1=6k_1$, $s_2=6k_2$, $q=6r$ we find the desired result.

Proof of (b2) Observe that due to the symmetry constraint,

$$f(X)=\frac{\|X\|_{0,2}}{\|X_0\|_{0,2}}+\frac{\|X^T\|_{0,2}}{\|X_0^T\|_{0,2}}+\frac{\mathrm{rank}(X)}{\mathrm{rank}(X_0)}\,.$$

Hence, the minimization is the same as **(a2)**, the matrix is rank r contained in a $k\times k$ submatrix and we additionally have the positive semidefinite constraint which can only reduce the amount of required measurements compared to **(a2)**. Consequently, the result follows by applying Lemma 53, similar to **(a2)**.

Proof of (c2) Let $C = \{X \neq 0 \big| f(X) \leq f(X_0)\}$. Since $\text{rank}(X_0) = 1$, if $f(X) \leq f(X_0) = 2$, then $\text{rank}(X) = 1$. With the symmetry constraint, this means $X = \pm xx^T$ for some l-sparse x. Observe that $X - X_0$ has rank at most 2 and is contained in a $2k \times 2k$ submatrix as $l \leq k$. Let M be the set of matrices that are symmetric and whose support lies in a $2k \times 2k$ submatrix. Using Lemma 53 with $q = 2$, $s_1 = s_2 = 2k$, whenever $m \geq ck \log \frac{n}{k}$, with desired probability all nonzero $\mathbf{W} \in M$ will satisfy $\mathcal{A}(\mathbf{W}) \neq 0$. Consequently, any $X \in C$ has $\mathcal{A}(X) \neq \mathcal{A}(X_0)$, so that X_0 is the unique minimizer.

A.1.3 Existence of a Matrix with Large κ's

We now argue that, there exists an $S\&L$ matrix that has large $\kappa_{\ell_1}, \kappa_{\ell_{1,2}}$ and $\kappa_\star$ simultaneously. We provide have a deterministic construction that is close to optimal. Our construction is based on Hadamard matrices. $\mathbf{H}_n \in \mathbb{R}^{n \times n}$ is called a Hadamard matrix if it has ± 1 entries and orthogonal rows. Hadamard matrices exist for n that is an integer power of 2.

Using $\mathbf{H}_n$, our aim is to construct a $d_1 \times d_2$ $S\&L$ (k_1, k_2, r) matrix X_0 that satisfies $\|\bar{X}_0\|_1^2 \approx k_1 k_2$, $\|\bar{X}_0\|_\star^2 \approx r$, $\|\bar{X}_0\|_{1,2}^2 \approx k_2$ and $\|\bar{X}_0^T\|_{1,2}^2 \approx k_1$. To do this, we construct a $k_1 \times k_2$ matrix and then embed it into a larger $d_1 \times d_2$ matrix. The following lemma summarizes the construction.

Lemma 54 *Without loss of generality, assume $k_2 \geq k_1 \geq r$. Let $\mathbf{H} := \mathbf{H}_{\lfloor \log_2 k_2 \rfloor}$. Let $X \in \mathbb{R}^{k_1 \times k_2}$ be such that its ith row is equal to $[i - 1 \pmod r] + 1$'th row of $\mathbf{H}$ followed by 0's for $1 \leq i \leq k_1$. Then,*

$$\|\bar{X}_0\|_1^2 \geq \frac{k_1 k_2}{2}\ ,\ \|\bar{X}_0\|_\star^2 \geq \frac{r}{2}\,,$$
$$\|\bar{X}_0\|_{1,2}^2 \geq \frac{k_2}{2}\ ,\ \|\bar{X}_0^T\|_{1,2}^2 = k_1\,.$$

In particular, if $k_1 \equiv 0 \pmod r$ and k_2 is an integer power of 2, then,

$$\|\bar{X}_0\|_1^2 = k_1 k_2,\ \|\bar{X}_0\|_\star^2 = r,$$
$$\|\bar{X}_0\|_{1,2}^2 = k_2,\ \|\bar{X}_0^T\|_{1,2}^2 = k_1.$$

Proof. The left $k_1 \times 2^{\lfloor \log_2 k_2 \rfloor}$ entries of X are ± 1, and the remaining entries are 0. This makes the calculation of ℓ_1 and $\ell_{1,2}$ and Frobenius norms trivial.

In particular, $\|X_0\|_F^2 = \|X_0\|_1 = k_1 2^{\lfloor \log_2 k_2 \rfloor}$, $\|X_0\|_{1,2} = \sqrt{k_1} 2^{\lfloor \log_2 k_2 \rfloor}$ and $\|X_0^T\|_{1,2} = k_1 2^{\frac{\lfloor \log_2 k_2 \rfloor}{2}}$. Substituting these yield the results for these norms.

To lower bound the nuclear norm, observe that, each of the first r rows of $\mathbf{H}$ are repeated at least $\lfloor \frac{k_1}{r} \rfloor$ times in X. Combined with orthogonality, this ensures that each singular value of X that is associated with the jth row of $\mathbf{H}$ is at least $\sqrt{2^{\lfloor \log_2 k_2 \rfloor} \lfloor \frac{k_1}{r} \rfloor}$ for all $1 \leq j \leq r$. Consequently

$$\|X\|_\star \geq r \sqrt{2^{\lfloor \log_2 k_2 \rfloor} \lfloor \frac{k_1}{r} \rfloor}.$$

Hence,

$$\|\bar{X}\|_\star \geq \frac{r\sqrt{\lfloor \frac{k_1}{r} \rfloor 2^{\lfloor \log_2 k_2 \rfloor}}}{\sqrt{k_1 2^{\lfloor \log_2 k_2 \rfloor}}} = r \sqrt{\frac{1}{k_1} \lfloor \frac{k_1}{r} \rfloor}.$$

Finally, we use the fact that $\lfloor \frac{k_1}{r} \rfloor \geq \frac{k_1}{2r}$ as $k_1 \geq r$. ∎

If we are allowed to use complex numbers, then one can apply the same idea with the Discrete Fourier Transform (DFT) matrix. Similar to $\mathbf{H}_n$, DFT has orthogonal rows and its entries have the same absolute value. However, it exists for any $n \geq 1$ which makes the argument more concise.

A.2 Properties of Cones

In this appendix, we state some results regarding cones which are used in the proof of general recovery. Recall the definitions of polar and dual cones from Section 2.2.

Theorem 55 (Moreau's decomposition theorem, [115]) *Let $\mathcal{C}$ be a closed and convex cone in $\mathbb{R}^n$. Then, for any $x \in \mathbb{R}^n$, we have*

- $x = \mathcal{P}_{\mathcal{C}}(x) + \mathcal{P}_{\mathcal{C}^\circ}(x)$.
- $\langle \mathcal{P}_{\mathcal{C}}(x), \mathcal{P}_{\mathcal{C}^\circ}(x) \rangle = 0$.

Lemma 56 (Projection is nonexpansive) *Let $\mathcal{C} \in \mathbb{R}^n$ be a closed and convex set and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ be vectors. Then,*

$$\|\mathcal{P}_{\mathcal{C}}(\mathbf{a}) - \mathcal{P}_{\mathcal{C}}(\mathbf{b})\|_2 \leq \|\mathbf{a} - \mathbf{b}\|_2.$$

Corollary 57 *Let $\mathcal{C}$ be a closed convex cone and $\mathbf{a}, \mathbf{b}$ be vectors satisfying $\mathcal{P}_{\mathcal{C}}(\mathbf{a} - \mathbf{b}) = 0$. Then*

$$\|\mathbf{b}\|_2 \geq \|\mathcal{P}_{\mathcal{C}}(\mathbf{a})\|_2.$$

Proof. Using Lemma 56, we have $\|\mathcal{P}_{\mathcal{C}}(\mathbf{a})\|_2 = \|\mathcal{P}_{\mathcal{C}}(\mathbf{a}) - \mathcal{P}_{\mathcal{C}}(\mathbf{a} - \mathbf{b})\|_2 \leq \|\mathbf{b}\|_2$. ∎

The unit sphere in $\mathbb{R}^n$ will be denoted by $\mathcal{S}^{n-1}$ for the following theorems.

Theorem 58 (Escape through a mesh, [73]) *For a given set $\mathcal{D} \in \mathcal{S}^{n-1}$, define the* Gaussian width *as*

$$\omega(\mathcal{D}) = \mathbb{E}\left[\sup_{x \in \mathcal{D}} \langle x, g \rangle\right],$$

in which $g \in \mathbb{R}^n$ has i.i.d. standard Gaussian entries. Given m, let $d = \sqrt{n - m} - \frac{1}{4\sqrt{n-m}}$. Provided that $\omega(\mathcal{D}) \leq d$ a random $m-$dimensional subspace which is uniformly drawn w.r.t. Haar measure will have no intersection with $\mathcal{D}$ with probability at least

$$1 - 3.5 \exp(-(d - \omega(\mathcal{D}))^2). \tag{A.2}$$

Theorem 59 *Consider a random Gaussian map $\mathcal{G} : \mathbb{R}^n \to \mathbb{R}^m$ with i.i.d. entires and the corresponding adjoint operator $\mathcal{G}^*$. Let $\mathcal{C}$ be a closed and convex cone and recalling Definition 5, let*

$$\zeta(\mathcal{C}) := 1 - \bar{\mathbf{D}}(\mathcal{C}),\ \gamma(\mathcal{C}) := 2\sqrt{\frac{1 + \bar{\mathbf{D}}(\mathcal{C})}{1 - \bar{\mathbf{D}}(\mathcal{C})}}.$$

where $\bar{\mathbf{D}}(\mathcal{C}) = \frac{\mathbf{D}(\mathcal{C})}{\sqrt{n}}$. Then, if $m \leq \frac{7\zeta(\mathcal{C})}{16} n$, with probability at least $1 - 6\exp(-(\frac{\zeta(\mathcal{C})}{4})^2 n)$, for all $z \in \mathbb{R}^n$ we have

$$\|\mathcal{G}^*(z)\|_2 \leq \gamma(\mathcal{C}) \|\mathcal{P}_{\mathcal{C}}(\mathcal{G}^*(z))\|_2. \tag{A.3}$$

Proof. For notational simplicity, let $\zeta = \zeta(\mathcal{C})$ and $\gamma = \gamma(\mathcal{C})$. Consider the set

$$\mathcal{D} = \left\{ x \in \mathcal{S}^{n-1} : \ \|x\|_2 \geq \gamma \|\mathcal{P}_{\mathcal{C}}(x)\|_2 \right\}.$$

and we are going to show that with high probability, the range of $\mathcal{G}^*$ misses $\mathcal{D}$. Using Theorem 55, for any $x \in \mathcal{D}$, we may write

$$\begin{aligned} \langle x, g \rangle &= \langle \mathcal{P}_{\mathcal{C}}(x) + \mathcal{P}_{\mathcal{C}^\circ}(x), \mathcal{P}_{\mathcal{C}}(g) + \mathcal{P}_{\mathcal{C}^\circ}(g) \rangle \\ &\leq \langle \mathcal{P}_{\mathcal{C}}(x), \mathcal{P}_{\mathcal{C}}(g) \rangle + \langle \mathcal{P}_{\mathcal{C}^\circ}(x), \mathcal{P}_{\mathcal{C}^\circ}(g) \rangle \qquad \text{(A.4)} \\ &\leq \|\mathcal{P}_{\mathcal{C}}(x)\|_2 \|\mathcal{P}_{\mathcal{C}}(g)\|_2 + \|\mathcal{P}_{\mathcal{C}^\circ}(x)\|_2 \|\mathcal{P}_{\mathcal{C}^\circ}(g)\|_2 \\ &\leq \gamma^{-1} \|\mathcal{P}_{\mathcal{C}}(g)\|_2 + \|\mathcal{P}_{\mathcal{C}^\circ}(g)\|_2 \end{aligned}$$

where in (A.4) we used the fact that elements of $\mathcal{C}$ and $\mathcal{C}^\circ$ have nonpositive inner products and $\|\mathcal{P}_{\mathcal{C}}(x)\|_2 \leq \|x\|_2$ is by Lemma 56. Hence, from the definition of Gaussian width,

$$\begin{aligned} \omega(\mathcal{D}) &= \mathbb{E}\left[\sup_{x \in \mathcal{D}} \langle x, g \rangle \right] \\ &\leq \gamma^{-1} \mathbb{E}\left[\|\mathcal{P}_{\mathcal{C}}(g)\|_2 \right] + \mathbb{E}\left[\|\mathcal{P}_{\mathcal{C}^\circ}(g)\|_2 \right] \\ &\leq \sqrt{n}(\gamma^{-1} \bar{\mathbf{D}}(\mathcal{C}^\circ) + \bar{\mathbf{D}}(\mathcal{C})) \leq \frac{2-\zeta}{2}\sqrt{n}. \end{aligned}$$

Where we used the fact that $\gamma \geq \frac{2\bar{\mathbf{D}}(\mathcal{C}^\circ)}{1-\bar{\mathbf{D}}(\mathcal{C})}$; which follows from $\bar{\mathbf{D}}(\mathcal{C})^2 + \bar{\mathbf{D}}(\mathcal{C}^\circ)^2 \leq 1$ (see Theorem 55 above). Hence, whenever,

$$m \leq \frac{7\zeta}{16} n \leq (1 - (\frac{4-\zeta}{4})^2) n = m',$$

using the upper bound on $\omega(\mathcal{D})$, we have,

$$(\sqrt{n-m} - \omega(\mathcal{D}) - \frac{1}{4\sqrt{n-m}})^2 \geq (\sqrt{n-m} - \omega(\mathcal{D}))^2 - \frac{1}{2} \geq (\frac{\zeta}{4})^2 n - \frac{1}{2}.$$

Now, using Theorem 58, the range space of $\mathcal{G}^*$ will miss the undesired set $\mathcal{D}$ with probability at least $1 - 3.5\exp(-(\frac{\zeta}{4})^2 n + \frac{1}{2}) \geq 1 - 6\exp(-(\frac{\zeta}{4})^2 n)$. ∎

Lemma 60 *Consider the cones $\mathbb{S}^d$ and $\mathbb{S}^d_+$ in the space $\mathbb{R}^{d \times d}$. Then, $\bar{\mathbf{D}}(\mathbb{S}^d) < \frac{1}{\sqrt{2}}$ and $\bar{\mathbf{D}}(\mathbb{S}^d_+) < \frac{\sqrt{3}}{2}$.*

Proof. Let $\mathbf{G}$ be a $d \times d$ matrix with i.i.d. standard normal entries. Set of symmetric matrices $\mathbb{S}^d$ is an $\frac{d(d+1)}{2}$ dimensional subspace of $\mathbb{R}^{d\times d}$. Hence, $\mathbb{E}\|\mathcal{P}_{\mathbb{S}^d}(\mathbf{G})\|_F^2 = \frac{d(d+1)}{2}$ and $\mathbb{E}\|\mathcal{P}_{(\mathbb{S}^d)^\circ}(\mathbf{G})\|_F^2 = \frac{d(d-1)}{2}$. Hence,

$$\bar{\mathbf{D}}(\mathbb{S}^d) = \sqrt{\frac{d(d-1)}{2d^2}} < \frac{1}{\sqrt{2}}.$$

To prove the second statement, observe that projection of a matrix $A \in \mathbb{R}^{d\times d}$ onto $\mathbb{S}_+^d$ is obtained by first projecting A onto $\mathbb{S}^d$ and then taking the matrix induced by the positive eigenvalues of $\mathcal{P}_{\mathbb{S}^d}(A)$. Since, $\mathbf{G}$ and $-\mathbf{G}$ are identically distributed and $\mathbb{S}_+^d$ is a self dual cone, $\mathcal{P}_{\mathbb{S}_+^d}(\mathbf{G})$ is identically distributed as $-\mathcal{P}_{\mathbb{S}_-^d}(\mathbf{G})$ where $\mathbb{S}_-^d = (\mathbb{S}_+^d)^\circ$ stands for negative semidefinite matrices. Hence,

$$\mathbb{E}\|\mathcal{P}_{\mathbb{S}_+^d}(\mathbf{G})\|_F^2 = \frac{\mathbb{E}\|\mathcal{P}_{\mathbb{S}^d}(\mathbf{G})\|_F^2}{2} = \frac{d(d+1)}{4},$$
$$\mathbb{E}\|\mathcal{P}_{(\mathbb{S}_+^d)^\circ}(\mathbf{G})\|_F^2 = \frac{d(3d-1)}{4}.$$

Consequently, $\bar{\mathbf{D}}(\mathbb{S}_+^d) = \sqrt{\frac{3}{4} - \frac{1}{4d}} < \sqrt{\frac{3}{4}}$. ∎

A.3 Norms in Sparse and Low-rank Model

Relevant notation for the proofs Let $[k]$ denote the set $\{1, 2, \ldots, k\}$. Let S_c, S_r denote the indexes of the nonzero columns and rows of X_0 so that nonzero entries of X_0 lies on $S_r \times S_c$ submatrix. $\mathcal{S}_c, \mathcal{S}_r$ denotes the k_1, k_2 dimensional subspaces of vectors whose nonzero entries lie on S_c and S_r respectively.

Let X_0 have singular value decomposition $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ such that $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ and columns of $\mathbf{U}, \mathbf{V}$ lies on $\mathcal{S}_c, \mathcal{S}_r$ respectively.

A.3.1 Proof of Lemma 52

Proof. Observe that $T_c = \mathbb{R}^d \times \mathcal{S}_c$ and $T_r = \mathcal{S}_r \times \mathbb{R}^d$ hence $T_c \cap T_r$ is the set of matrices that lie on $S_r \times S_c$. Hence, $\mathbf{E}_\star = \mathbf{U}\mathbf{V}^T \in T_c \cap T_r$. Similarly, $\mathbf{E}_c$ and $\mathbf{E}_r$ are the matrices obtained

by scaling columns and rows of X_0 to have unit size. As a result, they also lie on $S_r \times S_c$ and $T_c \cap T_r$. $\mathbf{E}_\star \in T_\star$ by definition.

Next, we may write $\mathbf{E}_c = X_0\mathbf{D}_c$ where $\mathbf{D}_c$ is the scaling nonnegative diagonal matrix. Consequently, $\mathbf{E}_c$ lies on the range space of X_0 and belongs to $T_\star$. This follows from definition of $T_\star$ in Lemma 30 and the fact that $(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{E}_c = 0$.

In the exact same way, $\mathbf{E}_r = \mathbf{D}_r X_0$ for some nonnegative diagonal $\mathbf{D}_r$ and lies on the range space of X^T and hence lies on $T_\star$. Consequently, $\mathbf{E}_\star, \mathbf{E}_c, \mathbf{E}_r$ lies on $T_c \cap T_r \cap T_\star$.

Now, consider

$$\langle \mathbf{E}_c, \mathbf{E}_\star \rangle = \langle X_0\mathbf{D}_c, \mathbf{U}\mathbf{V}^T \rangle = \mathrm{tr}(\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{D}_c) = \mathrm{tr}(\mathbf{V}\mathbf{\Sigma}\mathbf{V}^T\mathbf{D}^c) \geq 0.$$

since both $\mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$ and $\mathbf{D}^c$ are positive semidefinite matrices. In the exact same way, we have $\langle \mathbf{E}_c, \mathbf{E}_\star \rangle \geq 0$. Finally,

$$\langle \mathbf{E}_c, \mathbf{E}_r \rangle = \langle X_0\mathbf{D}_c, \mathbf{D}_r X_0 \rangle = \mathrm{tr}(\mathbf{D}_c X_0^T \mathbf{D}_r X_0) \geq 0,$$

since both $\mathbf{D}_c$ and $X_0^T\mathbf{D}_r X_0$ are PSD matrices. Overall, the pairwise inner products of $\mathbf{E}_r, \mathbf{E}_c, \mathbf{E}_\star$ are nonnegative. ∎

A.3.2 Results on the PSD Constraint

Lemma 61 *Assume $X, Y \in \mathbb{S}_+^d$ have eigenvalue decompositions $X = \sum_{i=1}^{\mathrm{rank}(X)} \sigma_i \mathbf{u}_i\mathbf{u}_i^T$ and $Y = \sum_{i=1}^{\mathrm{rank}(Y)} c_i \mathbf{v}_i\mathbf{v}_i^T$. Further, assume $\langle Y, X \rangle = 0$. Then, $\mathbf{U}^T Y = 0$ where $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_{\mathrm{rank}(X)}]$.*

Proof. Observe that,

$$\langle Y, X \rangle = \sum_{i=1}^{\mathrm{rank}(X)} \sum_{j=1}^{\mathrm{rank}(Y)} \sigma_i c_j |\mathbf{u}_i^T \mathbf{v}_j|^2.$$

Since $\sigma_i, c_j > 0$, right hand side is 0 if and only if $\mathbf{u}_i^T\mathbf{v}_j = 0$ for all i, j. Hence, the result follows. ∎

Lemma 62 *Assume $X_0 \in \mathbb{S}_+^d$ so that in Section A.3, $S_c = S_r$, $T_c = T_r$, $k_1 = k_2 = k$ and $\mathbf{U} = \mathbf{V}$. Let $\mathcal{R} = T_c \cap T_r \cap T_\star \cap \mathbb{S}^d$, $S_\star = T_\star \cap \mathbb{S}^d$, and,*

$$\mathcal{Y} = \{Y \big| Y \in (\mathbb{S}_+^d)^*, \ \langle Y, X_0 \rangle = 0\},$$

Then, the following statements hold.

- $S_\star \subseteq \operatorname{span}(\mathcal{Y})^\perp$. *Hence, $\mathcal{R} \subseteq S_\star$ and is orthogonal to $\mathcal{Y}$.*
- $\mathbf{E}_\star \in \mathcal{R}$, $\frac{\|\mathcal{P}_\mathcal{R}(\mathbf{E}_c)\|_F}{\|\mathbf{E}_c\|_F} = \frac{\|\mathcal{P}_\mathcal{R}(\mathbf{E}_r)\|_F}{\|\mathbf{E}_r\|_F} \geq \frac{1}{\sqrt{2}}$.

Proof. The dual of $\mathbb{S}_+^d$ with respect to $\mathbb{R}^{d\times d}$ is the set sum of $\mathbb{S}_+^d$ and Skew^d where Skew^d is the set of skew-symmetric matrices. Now, assume, $Y \in \mathcal{Y}$ and $X \in S_\star$. Then, $\langle Y, X\rangle = \left\langle \frac{\mathbf{Z}}{2}, X\right\rangle$ where $\mathbf{Z} = Y + Y^T \in \mathbb{S}_+^d$ and $\langle \mathbf{Z}, X_0\rangle = 0$. Since X_0, $\mathbf{Z}$ are both PSD, applying Lemma 61, we have $\mathbf{U}^T\mathbf{Z} = 0$ hence $(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{Z}(\mathbf{I} - \mathbf{U}\mathbf{U}^T) = \mathbf{Z}$ which means $\mathbf{Z} \in T_\star^\perp$. Hence, $\langle \mathbf{Z}, X\rangle = \langle Y, X\rangle = 0$ as $X \in S_\star \subset T_\star$. Hence, $\operatorname{span}(\mathcal{Y}) \subseteq S_\star^\perp$.

For the second statement, let $T_\cap = T_\star \cap T_c \cap T_r$. Recalling Lemma 52, observe that $\mathbf{E}_\star \in T_\cap$. Since $\mathbf{E}_\star$ is also symmetric, $\mathbf{E}_\star \in \mathcal{R}$. Similarly, $\mathbf{E}_c, \mathbf{E}_r \in T_\cap$, $\langle \mathbf{E}_c, \mathbf{E}_r\rangle \geq 0$ and $\|\mathcal{P}_\mathcal{R}(\mathbf{E}_c)\| = \|\frac{\mathbf{E}_c+\mathbf{E}_r}{2}\|_F \geq \frac{\|\mathbf{E}_c\|_F}{\sqrt{2}}$. Similar result is true for $\mathbf{E}_r$. ∎

A.4 Results on Non-convex Recovery

Next two lemmas are standard results on sub-gaussian measurement operators.

Lemma 63 (Properties of sub-gaussian mappings) *Assume X is an arbitrary matrix with unit Frobenius norm. A measurement operator $\mathcal{A}(\cdot)$ with i.i.d. zero-mean isotropic subgaussian rows (see Section 2.4) satisfies the following:*

- $\mathbb{E}\|\mathcal{A}(X)\|_2^2 = m$.
- *There exists an absolute constant $c > 0$ such that, for all $1 \geq \varepsilon \geq 0$, we have*

$$\mathbb{P}\big(|\|\mathcal{A}(X)\|_2^2 - m| \geq \varepsilon m\big) \leq 2\exp\big(-c\varepsilon^2 m\big).$$

Proof. Observe that, when $\|X\|_F = 1$, entries of $\mathcal{A}(X)$ are zero-mean with unit variance. Hence, the first statement follows directly. For the second statement, we use the fact that square of a sub-gaussian random variable is sub-exponential and view $\|\mathcal{A}(X)\|_2^2$ as a sum of m i.i.d. subexponentials with unit mean. Then, result follows from Corollary 5.17 of [155]. ■

For the consequent lemmas, $\mathcal{S}^{d_1 \times d_2}$ denotes the unit Frobenius norm sphere in $\mathbb{R}^{d_1 \times d_2}$.

Lemma 64 *Let $\mathcal{D} \in \mathbb{R}^{d_1 \times d_2}$ be an arbitrary cone and $\mathcal{A}(\cdot) : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ be a measurement operator with i.i.d. zero-mean and isotropic sub-gaussian rows. Assume that the set $\bar{\mathcal{D}} = \mathcal{S}^{d_1 \times d_2} \cap \mathcal{D}$ has ε-covering number bounded above by $\eta(\varepsilon)$. Then, there exists constants $c_1, c_2 > 0$ such that whenever $m \geq c_1 \log \eta(1/4)$, with probability $1 - 2\exp(-c_2 m)$, we have*

$$\mathcal{D} \cap \mathrm{Null}(\mathcal{A}) = \{0\}.$$

Proof. Let $\eta = \eta(\frac{1}{4})$, and $\{X_i\}_{i=1}^{\eta}$ be a $\frac{1}{4}$-covering of $\bar{\mathcal{D}}$. With probability at least $1 - 2\eta\exp(-c\varepsilon^2 m)$, for all i, we have

$$(1-\varepsilon)m \leq \|\mathcal{A}(X_i)\|_2^2 \leq (1+\varepsilon)m.$$

Now, let $X_{\sup} = \arg\sup_{X \in \bar{\mathcal{D}}} \|\mathcal{A}(X)\|_2$. Choose $1 \leq a \leq \eta$ such that $\|X_a - X_{\sup}\|_2 \leq 1/4$. Then:

$$\|\mathcal{A}(X_{\sup})\|_2 \leq \|\mathcal{A}(X_a)\|_2 + \|\mathcal{A}(X_{\sup} - X_a)\|_2 \leq (1+\varepsilon)m + \frac{1}{4}\|\mathcal{A}(X_{\sup})\|_2 \,.$$

Hence, $\|\mathcal{A}(X_{\sup})\|_2 \leq \frac{4}{3}(1+\varepsilon)m$. Similarly, let $X_{\inf} = \arg\inf_{X \in \bar{\mathcal{D}}} \|\mathcal{A}(X)\|_2$. Choose $1 \leq b \leq \eta$ satisfying $\|X_b - X_{\inf}\| \leq 1/4$. Then,

$$\|\mathcal{A}(X_{\inf})\|_2 \geq \|\mathcal{A}(X_b)\|_2 - \|\mathcal{A}(X_{\inf} - X_b)\|_2 \geq (1-\varepsilon)m - \frac{1}{3}(1+\varepsilon)m \,.$$

This yields $\|\mathcal{A}(X_{\inf})\|_2 \geq \frac{2-4\varepsilon}{3}m$. Choosing $\varepsilon = 1/4$ whenever $m \geq \frac{32}{c}\log(\eta)$ with the desired probability, $\|\mathcal{A}(X_{\inf})\|_2 > 0$. Equivalently, $\bar{\mathcal{D}} \cap \mathrm{Null}(\mathcal{A}) = \emptyset$. Since $\mathcal{A}(\cdot)$ is linear and $\mathcal{D}$ is a cone, the claim is proved. ■

The following lemma gives a covering number of the set of low rank matrices.

Lemma 65 (Candes and Plan, [34]) *Let M be the set of matrices in $\mathbb{R}^{d_1 \times d_2}$ with rank at most r. Then, for any $\varepsilon > 0$, there exists a covering of $\mathcal{S}^{d_1 \times d_2} \cap M$ with size at most $(\frac{c_3}{\varepsilon})^{(d_1+d_2)r}$ where c_3 is an absolute constant. In particular, $\log(\eta(1/4))$ is upper bounded by $C^{(d_1+d_2)r}$ for some constant $C > 0$.*

Now, we use Lemma 65 to find the covering number of the set of simultaneously low rank and sparse matrices.

A.4.1 Proof of Lemma 53

Proof. Assume M has $\frac{1}{4}$-covering number N. Then, using Lemma 64, whenever $m \geq c_1 \log N$, (A.1) will hold. What remains is to find N. To do this, we cover each individual $s_1 \times s_2$ submatrix and then take the union of the covers. For a fixed submatrix, using Lemma 65, $\frac{1}{4}$-covering number is given by $C^{(s_1+s_2)q}$. In total there are $\binom{d_1}{s_1} \times \binom{d_2}{s_2}$ distinct submatrices. Consequently, by using $\log \binom{d}{s} \approx s \log \frac{d}{s} + s$, we find

$$\log N \leq \log \left(\binom{d_1}{s_1} \times \binom{d_2}{s_2} C^{(s_1+s_2)q} \right) \leq s_1 \log \frac{d_1}{s_1} + s_1 + s_2 \log \frac{d_2}{s_2} + s_2 + (s_1 + s_2) q \log C \,,$$

and obtain the desired result. ∎

Appendix B

SUPPLEMENT TO CHAPTER 3

B.1 Proofs

B.1.1 Proof of Lemma 39

Proof. First, assume that Ω is convex. By plugging in X and $-X$ in the definition of convexity for Ω we get $\Omega(X) \geq 0$, so the square root is well-defined. We show the triangle inequality $\sqrt{\Omega(X+Y)} \leq \sqrt{\Omega(X)} + \sqrt{\Omega(Y)}$ holds for any X, Y. If $\Omega(X+Y)$ is zero, the inequality is trivial. Otherwise, for any $\theta \in (0,1)$ let $A = \frac{1}{\theta}X$, $B = \frac{1}{1-\theta}Y$, and use the convexity and second-order homogeneity of Ω to get

$$\Omega(X+Y) = \Omega(\theta A + (1-\theta)B) \leq \theta\Omega(A) + (1-\theta)\Omega(B) = \tfrac{1}{\theta}\,\Omega(X) + \tfrac{1}{1-\theta}\,\Omega(Y). \qquad \text{(B.1)}$$

If $\Omega(X) \geq \Omega(Y) = 0$, set $\theta = (\Omega(X) + \Omega(X+Y))/(2\Omega(X+Y)) > 0$. Assuming $\theta < 1$, from (B.1) we get

$$\Omega(X+Y) \leq \tfrac{1}{\theta}\,\Omega(X) = \frac{2\Omega(X+Y)\Omega(X)}{\Omega(X+Y)+\Omega(X)},$$

which is equivalent to $\theta \geq 1$ and is a contradiction. Therefore, $\theta \geq 1$ which gives the desired inequality.

And if $\Omega(X), \Omega(Y) \neq 0$, set $\theta = \sqrt{\Omega(X)}/(\sqrt{\Omega(X)} + \sqrt{\Omega(Y)}) \in (0,1)$ to get

$$\Omega(X+Y) \leq \tfrac{1}{\theta}\,\Omega(X) + \tfrac{1}{1-\theta}\,\Omega(Y) = (\sqrt{\Omega(X)} + \sqrt{\Omega(Y)})^2\,.$$

Since $\sqrt{\Omega}$ satisfies the triangle inequality and absolute homogeneity, it is a semi-norm. Notice that $\Omega(X) = 0$ does not necessarily imply $X = 0$, unless Ω is strictly convex.

Now, suppose that $\sqrt{\Omega}$ is a semi-norm; hence convex. The function f defined by $f(x) = x^2$ for $x \geq 0$ and $f(x) = 0$ for $x \leq 0$ is non-decreasing, so the composition of these two functions is convex and equal to Ω.

One can alternatively use Corollary 15.3.2 of [134] to prove the first part of the lemma. ■

B.1.2 Proof of Proposition 42

Proof. We use the results on subdifferentiation in parametric minimization [137, Section 10.C]. First, let's fix some notation. Throughout the proof, we denote $\frac{1}{2}\Omega$ by Ω, and $2\Omega^\star$ by $\Omega^\star$. Denote by $\iota_{\mathcal{M}}(M)$ the indicator function of the set $\mathcal{M}$ which is 1 when $M \in \mathcal{M}$ and $+\infty$ otherwise. We use $\mathcal{M}$ instead of $\mathcal{M} \cap \mathbb{S}_+$ to simplify the notation. Considering

$$f(Y, M) := \begin{cases} \frac{1}{2}\operatorname{tr}(Y M^\dagger Y^T) & \text{if } \operatorname{range}(Y^T) \subseteq \operatorname{range}(M) \\ +\infty & \text{otherwise} \end{cases}$$

we have $\Omega^\star(Y) = \inf_M \ f(Y, M) + \iota_{\mathcal{M}}(M)$. For such a function, we can use results in [26, Theorem 4.8] to show that

$$\partial f(Y, M) = \operatorname{conv}\left\{(Z, -\tfrac{1}{2}Z^T Z) : \ Z = Y M^\dagger + W\,,\ \operatorname{range}(W^T) \subseteq \ker(M)\right\}\,.$$

Since $g(Y, M) := f(Y, M) + \iota_{\mathcal{M}}(M)$ is convex, we can use the second part of Theorem 10.13 in [137]: for any choice of M_0 which is optimal in the definition of $\Omega^\star(Y)$,

$$\partial\Omega^\star(Y) = \{Z : \ (Z, \mathbf{0}) \in \partial g(Y, M_0)\}\,.$$

Therefore, for any $Z \in \partial\Omega^\star(Y)$ we have

$$\tfrac{1}{2}Z^T Z \in \partial\iota_{\mathcal{M}}(M_0) = \{G : \ \langle G, M' - M_0\rangle \le 0\,,\ \forall\, M' \in \mathcal{M}\}$$

(Here $\partial\iota_{\mathcal{M}}(M_0)$ is the normal cone of $\mathcal{M}$ at M_0.) This implies

$$\tfrac{1}{2}\operatorname{tr}(Z M' Z^T) \le \tfrac{1}{2}\operatorname{tr}(Z M_0 Z^T)$$

for all $M' \in \mathcal{M}$. Taking the supremum of the left hand side over all $M' \in \mathcal{M}$, we get

$$\Omega(Z) = \tfrac{1}{2}\operatorname{tr}(Z M_0 Z^T) = \tfrac{1}{2}\operatorname{tr}(Y M_0^\dagger Y^T) = \Omega^\star(Y)\,,$$

where the second equality follows from range$(W^T) \subseteq$ ker(M_0) (which is equivalent to $M_0 W^T = \mathbf{0}$). Alternatively, for any matrix Z from the right hand side of (3.22) (after adjustment to our rescaling of definition of Ω by $\frac{1}{2}$), and any $Y' \in \mathbb{R}^{n \times m}$ we have

$$\Omega^\star(Y') \geq \langle Y', Z \rangle - \Omega(Z) = \langle Y', Z \rangle - \Omega^\star(Y) = \langle Y' - Y, Z \rangle + \Omega^\star(Y)$$

where we used Fenchel's inequality, as well as the characterization of Z. Therefore, $Z \in \partial\Omega^\star(Y)$. This finishes the proof. Notice that for an achieving M, ker$(M) \subseteq$ ker(Y) (or equivalently, range$(Y^T) \subseteq$ range(M)) has to hold for the conjugate function to be defined. ■

B.1.3 Alternative Proof for Corollary 45

Another proof of Corollary 45, in the case where $\sqrt{\Omega_{\mathbb{M}}}$ is a norm and not a semi-norm, is given as Lemma 66.

Lemma 66 *Consider any norm $\|\cdot\|$. Then, $|||\cdot|||$ is a norm itself if and only if we have $|||x||| = \min_{y \geq |x|} \|y\|$.*

Proof. First, suppose $\|\cdot\|_a := |||\cdot|||$ is a norm; hence it is an absolute norm and is monotonic as well by definition. Therefore, for any $y \geq |x|$ we have $\|y\|_a \geq \|x\|_a$ which gives $\min_{y \geq |x|} \|y\|_a \geq \|x\|_a$. Since $|x|$ is feasible in this optimization, and $|||x|||_a = \|x\|_a$ we get the desired result; $|||x||| = \|x\|_a = \min_{y \geq |x|} \|y\|$.

On the other hand, consider $f(\cdot) := \min_{y \geq |x|} \|y\|$. We show that it is a norm. Clearly, f is nonnegative and homogenous, and $f(x) = 0$ implies that $\|y\| = 0$ for some $y \geq |x| \geq 0$ which implies $x = 0$. The triangle inequality is also verified as follows,

$$\begin{aligned} f(x+z) = \min_{y \geq |x+z|} \|y\| &\leq \min_{y \geq |x|+|z|} \|y\| = \min_{y_1 \geq |x|\,,\, y_2 \geq |z|} \|y_1 + y_2\| \\ &\leq \min_{y_1 \geq |x|\,,\, y_2 \geq |z|} \|y_1\| + \|y_2\| = f(x) + f(z)\,. \end{aligned}$$

■

Appendix C

SUPPLEMENT TO CHAPTER 4

C.1 Proofs for Convex Recovery

In the following, we present the proofs of Theorems 48 and 49. The matrix concentration bounds play an important role in these proofs, and are given as Lemmas 70 and 73.

C.1.1 Notation

In this chapter, we consider the heterogenous stochastic block model described in Section 4.1.1. Consider a partition of the n nodes into $V_0, V_1, \ldots, V_r$, where $|V_k| = n_k$, $k = 0, 1, \ldots, r$. Consider $\bar{n} = \sum_{k=1}^{r} n_k$ and denote the number of isolated nodes by n_0 ; hence, $n_0 + \bar{n} = n$. Ignoring n_0, we further define $n_{\min}$ and $n_{\max}$ as the minimum and maximum of $n_1, \ldots, n_k$ respectively. The nodes in V_0 are isolated and the nodes in V_k form the community $\mathcal{C}_k = V_k \times V_k$, for $k = 1, \ldots, r$. The union of communities is denoted by $\mathcal{C} = \cup_{k=1}^{r} \mathcal{C}_k$ and $\mathcal{C}^c$ denotes the complement; i.e. $\mathcal{C}^c = \{(i,j) : \ (i,j) \notin \mathcal{C}_k \text{ for any } k = 1, \ldots, r, \text{ and } i, j = 1, \ldots, n\}$. Denote by $\mathcal{Y}$ the set of admissible adjacency matrices according to a community assignment as above, i.e.

$$\mathcal{Y} := \{Y \in \{0,1\}^{n \times n} : \ Y \text{ is a valid community matrix w.r.t. } V_0, V_1, \ldots, V_r \text{ where } |V_k| = n_k\}.$$

We will denote by $\mathbf{1}_C \in \mathbb{R}^{n \times n}$ a matrix which is 1 on $C \subset \{1, \ldots, n\}^2$ and zero elsewhere. log denotes the natural logarithm (base e), and the notation $\theta \gtrsim 1$ is equivalent to $\theta \geq \mathrm{O}(1)$. A Bernoulli random variable with parameter p is denoted by $\mathrm{Ber}(p)$, and a Binomial random variable with parameters n and p is denoted by $\mathrm{Bin}(n,p)$. $\|\cdot\|_\star$ denotes the matrix nuclear norm or trace norm, i.e., the sum of singular values of the matrix. The dual to the nuclear norm is the spectral norm, denoted by $\|\cdot\|$.

Given a single graph drawn from the heterogenous stochastic block model, the goal is to recover the underlying community matrix $Y^\star \in \mathcal{Y}$ exactly. We will need the following definitions:

Define the *relative density of a community* as

$$\rho_k = (p_k - q)n_k$$

which gives $\sum_{k=1}^r \rho_k = \sum_{k=1}^r p_k n_k - qn$. Define the total variance $\sigma_k^2 = n_k p_k(1-p_k)$ over the kth community, and let $\sigma_0^2 = nq(1-q)$. Also, define

$$\sigma_{\max}^2 = \max_{k=1,\ldots,r} \sigma_k^2 = \max_{k=1,\ldots,r} n_k p_k(1-p_k).$$

The Neyman Chi-square divergence (e.g., see [52]) between the two discrete random variables $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$ is given by

$$\widetilde{D}(p,q) := \frac{(p-q)^2}{q(1-q)}$$

and we have $\widetilde{D}(p,q) \geq D_{\mathrm{KL}}(p,q) := D_{\mathrm{KL}}(\mathrm{Ber}(p), \mathrm{Ber}(q))$; see (C.18). Chi-square divergence is an instance of a more general family of divergence functions called f-divergences or Ali-Silvey distances [5]. This family also has KL-divergence, total variation distance, Hellinger distance and Chernoff distance as special cases. Moreover, the divergence used in [2] is an f-divergence.

C.1.2 Proof of Theorem 48

We are going to prove that under the heterogenous stochastic block model (HSBM), with high probability, the output of the convex recovery program in (4.4) coincides with the underlying community matrix $Y^\star = \sum_{k=1}^r \mathbf{1}_{V_k}\mathbf{1}_{V_k}^T$ provided that

$$\begin{aligned} \rho_k^2 &\gtrsim n_k p_k(1-p_k)\log n_k \\ (p_{\min}-q)^2 &\gtrsim q(1-q)\tfrac{\log n_{\min}}{n_{\min}} \\ \rho_{\min}^2 &\gtrsim \max\Big\{\max_k\ n_k p_k(1-p_k),\ nq(1-q),\ \log n\Big\} \end{aligned}$$

as well as $\sum_{k=1}^{r} n_k^{-\alpha} = \mathrm{o}(1)$ for some $\alpha > 0$.

Notice that $p_k(1-p_k)n_k \gtrsim \log n_k$, for all $k = 1, \ldots, r$, is implied by the first condition, as mentioned in Remark 67.

Remark 67 *For exact recovery to be possible, we need all communities (but at most one) to be connected. Therefore, in each subgraph, which is generated by $\mathcal{G}(n_k, p_k)$, we need $p_k n_k > \log n_k$, for $k = 1, \ldots, r$. Observe that this connectivity requirement is implicit in the first condition of Theorems 48, 49: for example, the first condition of Theorem 48 can be equivalently expressed as $n_k \widetilde{D}(q, p_k) \gtrsim \log n_k$. Moreover, for $q < p$, when both p and q/p are bounded away from 1, we have*

$$\widetilde{D}(q,p) = p\frac{(1-q/p)^2}{1-p} \approx p\,.$$

Before proving Theorem 48, we state a crucial result from random matrix theory that allows us to bound the spectral radius of the matrix $A - \mathbb{E}(A)$ where A is an instance of adjacency matrix under HSBM. This result appears, for example, as Theorem 3.4 in [43][1]. Although Lemma 2 from [151] appears to state a weaker version of this result, the proof presented there actually supports the version we give below in Lemma 68. Finally, Lemma 8 from [158] states the same result and presents a very brief sketch of the proof idea, along the lines of the proof presented fully in [151].

Lemma 68 *Let $A = \{a_{ij}\}$ be a $n \times n$ symmetric random matrix such that each a_{ij} represents an independent random Bernoulli variable with $\mathbb{E}(a_{ij}) = p_{ij}$. Assume that there exists a constant C_0 such that $\sigma^2 = \max_{i,j} p_{ij}(1-p_{ij}) \geq C_0 \log n/n$. Then for each constant $C_1 > 0$ there exists $C_2 > 0$ such that*

$$\mathbb{P}\left(\|A - \mathbb{E}(A)\| \geq C_2\sigma\sqrt{n}\right) \;\leq\; n^{-C_1}\,.$$

As an immediate consequence of this, we have the following corollary.

[1] As a more general result about the norms of rectangular matrices, but with the slightly stronger growth condition $\sigma^2 \geq \log^{6+\epsilon} n/n$.

Corollary 69 *Let $A = \{a_{ij}\}$ be a $n \times n$ symmetric random matrix such that each a_{ij} represents an independent random Bernoulli variable with $\mathbb{E}(a_{ij}) = p_{ij}$. Assume that there exists a constant C_0 such that $\sigma^2 = \max_{i,j} p_{ij}(1-p_{ij}) \leq C_0 \log n/n$. Then for each constant $C_1 > 0$ there exists $C_3 > 0$ such that such that*

$$\mathbb{P}\left(\|A - \mathbb{E}(A)\| \geq C_3\sqrt{\log n}\right) \;\leq\; n^{-C_1}\,.$$

Proof. The corollary follows from Lemma 68, by replacing the $(1,1)$ entry of A with a Bernoulli variable of probability $p_{11} = C_0 \log n/n$. Given that the old $(1,1)$ entry and the new $(1,1)$ entry are both Bernoulli variables, this can change $\|A - \mathbb{E}(A)\|$ by at most 1. The new maximal variance is equal to $\max_{i,j} p_{ij}(1-p_{ij}) = C_0 \log n/n$. Therefore Lemma 68 is applicable to the new matrix and the conclusion holds. ∎

We use Lemma 68 to prove the following result.

Lemma 70 *Let A be generated according to the heterogenous stochastic block model (HSBM). Suppose*

(1) $p_k(1-p_k)n_k \gtrsim \log n_k$, for $k = 1, \ldots, r$, and

(2) there exists an $\alpha > 0$ such that $\sum_{k=0}^{r} n_k^{-\alpha} = \mathrm{o}(1)$.

Then with probability at least $1 - \mathrm{o}(1)$ we have

$$\|A - \mathbb{E}(A)\| \;\lesssim\; \max_i \sqrt{p_i(1-p_i)n_i} + \sqrt{\max\{q(1-q)n\,,\,\log n\}}\,.$$

Proof. We split the matrix A into two matrices, B_1 and B_2. B_1 consists of the block-diagonal projection onto the clusters, and B_2 is the rest. Denote the blocks on the diagonal of B_1 by C_1, $C_2, \ldots, C_r$, where C_i corresponds to the ith cluster. Then $\|B_1 - \mathbb{E}(B_1)\| = \max_i \|C_i - \mathbb{E}(C_i)\|$, and for each i, $\|C_i - \mathbb{E}(C_i)\| \gtrsim \sqrt{p_i(1-p_i)n_i}$ with probability at most $n_i^{-\alpha}$, by Lemma 68. By assumptions (1) and (2) of Lemma 70 and applying a union bound, we conclude that

$$\|B_1 - \mathbb{E}(B_1)\| \lesssim \max_i \sqrt{p_i(1-p_i)n_i}$$

with probability at least $1-\sum_{i=1}^{r} n_i^{-\alpha} = 1-\mathrm{o}(1)$. We shall now turn our attention to B_2. Let $\sigma^2 = \max\{q(1-q), \log n/n\}$. By Corollary 69, $\|B_2 - \mathbb{E}(B_2)\| \lesssim \max\{\sqrt{q(1-q)n}, \sqrt{\log n}\}$, with high probability. Putting the two norm estimates together, the conclusion of Lemma 70 follows. ∎

We are now in the position to prove Theorem 48.

Proof. [of Theorem 48] We need to show that for any feasible $Y \neq Y^\star$, we have $\Delta(Y) := \langle A, Y^\star - Y\rangle > 0$. Rewrite $\Delta(Y)$ as

$$\Delta(Y) = \langle A, Y^\star - Y\rangle = \langle \mathbb{E}(A), Y^\star - Y\rangle + \langle A - \mathbb{E}(A), Y^\star - Y\rangle\,. \tag{C.1}$$

Note that $\sum_{i,j} Y^\star_{ij} = \sum_{i,j} Y_{ij} = \sum_{k=1}^r n_k^2$, thus $\sum_{i,j}(Y^\star_{ij} - Y_{ij}) = 0$. Express this as

$$\sum_{k=1}^{r} \sum_{i,j\in V_k} (Y^\star - Y)_{ij} = -\sum_{k'\neq k''} \sum_{i\in V_{k'},\, j\in V_{k''}} (Y^\star - Y)_{ij}\,.$$

Then we have

$$\begin{aligned}\langle \mathbb{E}(A), Y^\star - Y\rangle &= \sum_{k=1}^{r} \sum_{i,j\in V_k^\star} p_k (Y^\star - Y)_{ij} + \sum_{k'\neq k''} \sum_{i\in V_{k'},\, j\in V_{k''}} q(Y^\star - Y)_{ij} \\ &= \sum_{k=1}^{r} \sum_{i,j\in V_k} (p_k - q)(Y^\star - Y)_{ij}\,.\end{aligned}$$

Finally, since $0 \le (Y^\star - Y)_{ij} \le 1$ for $i, j \in V_k$, we can write

$$\langle \mathbb{E}(A), Y^\star - Y\rangle = \sum_{k=1}^{r} \sum_{i,j\in V_k} (p_k - q)\|(Y^\star - Y)_{\mathcal{C}_k}\|_1\,. \tag{C.2}$$

Next, recall that the subdifferential (i.e., the set of all subgradients) of $\|\cdot\|_\star$ at $Y^\star$ is given by

$$\partial\|Y^\star\|_\star = \{UU^T + Z \;\big|\; U^T Z = ZU = 0\,,\ \|Z\| \le 1\}$$

where $Y^\star = UKU^T$ is the singular value decomposition for $Y^\star$ with $U \in \mathbb{R}^{n\times r}$, $K = \mathrm{diag}(n_1, \ldots, n_r)$, and $U_{ik} = 1/\sqrt{n_k}$ if node i is in cluster $\mathcal{C}_k$ and $U_{ik} = 0$ otherwise.

Let $M := A - \mathbb{E}(A)$. Since conditions (1) and (2) of Lemma 70 are verified, there exists $C_1 > 0$ such that $\|M\| \le \lambda$, with probability $1 - \mathrm{o}(1)$, where

$$\lambda := C_1 \left(\max_i \sqrt{p_i(1-p_i)n_i} + \sqrt{\max\{q(1-q)n, \log n\}}\right)\,. \tag{C.3}$$

Furthermore, let the projection operator onto a subspace T be defined by

$$\mathcal{P}_T(M) := UU^T M + MUU^T - UU^T MUU^T \,,$$

and also $\mathcal{P}_{T^\perp} = \mathcal{I} - \mathcal{P}_T$, where $\mathcal{I}$ is the identity map. Since $\|\mathcal{P}_{T^\perp}(M)\| \leq \|M\| \leq \lambda$ with high probability, $UU^T + \frac{1}{\lambda}\mathcal{P}_{T^\perp}(M) \in \partial\|Y^\star\|_\star$ with high probability. Now, by the constraints of the convex program, we have

$$\begin{aligned} 0 &\geq \|Y\|_\star - \|Y^\star\|_\star \\ &\geq \langle UU^T + \tfrac{1}{\lambda}\mathcal{P}_{T^\perp}(M), Y - Y^\star\rangle \\ &= \langle UU^T - \tfrac{1}{\lambda}\mathcal{P}_T(M), Y - Y^\star\rangle + \tfrac{1}{\lambda}\langle M, Y - Y^\star\rangle \,, \end{aligned}$$

which implies $\langle M, Y^\star - Y\rangle \geq \langle \mathcal{P}_T(M) - \lambda UU^T, Y^\star - Y\rangle$. Combining (C.1) and (C.2) we get,

$$\begin{aligned} \Delta(Y) &\geq \sum_{k=1}^{r}(p_k - q)\|(Y^\star - Y)_{\mathcal{C}_k}\|_1 + \langle \mathcal{P}_T(M) - \lambda UU^T, Y^\star - Y\rangle \\ &\geq \sum_{k=1}^{r}(p_k - q)\|(Y^\star - Y)_{\mathcal{C}_k}\|_1 \\ &\quad - \sum_{k=1}^{r} \underbrace{\|(\mathcal{P}_T(M) - \lambda UU^T)_{\mathcal{C}_k}\|_\infty}_{(\mu_{kk})} \|(Y^\star - Y)_{\mathcal{C}_k}\|_1 \\ &\quad - \sum_{k' \neq k''} \underbrace{\|(\mathcal{P}_T(M) - \lambda UU^T)_{V_{k'} \times V_{k''}}\|_\infty}_{(\mu_{k'k''})} \|(Y^\star - Y)_{V_{k'} \times V_{k''}}\|_1 \end{aligned} \tag{C.4}$$

where we have made use of the fact that an inner product can be bounded by a product of dual norms. We now derive bounds for the quantities μ_{kk} and $\mu_{k'k''}$ marked above. Note that the former indicates sums over the clusters, while the latter indicates sums outside the clusters.

For μ_{kk}, if $(i,j) \in \mathcal{C}_k$ then

$$\begin{aligned} \big(\mathcal{P}_T(M) - \lambda UU^T\big)_{ij} &= \big(UU^T M + MUU^T - UU^T MUU^T - \lambda UU^T\big)_{ij} \\ &= \frac{1}{n_k}\sum_{l \in \mathcal{C}_k} M_{lj} + \frac{1}{n_k}\sum_{l \in \mathcal{C}_k} M_{il} - \frac{1}{n_k^2}\sum_{l,l' \in \mathcal{C}_k} M_{ll'} - \frac{\lambda}{n_k} \,. \end{aligned}$$

Recall Bernstein's inequality (e.g. see Theorem 1.6.1 in [153]):

Proposition 71 *(Bernstein Inequality) Let $S_1, S_2, \ldots, S_n$ be independent, centered, real random variables, and assume that each one is uniformly bounded:*

$$\mathbb{E}(S_k) = 0 \quad \textit{and} \quad |S_k| \leq L \quad \textit{for each } k = 1, \ldots, n\,.$$

Introduce the sum $Z = \sum_{k=1}^{n} S_k$, and let $\nu(Z)$ denote the variance of the sum:

$$\nu(Z) = \mathbb{E}(Z^2) = \sum_{k=1}^{n} \mathbb{E}(S_k^2)\,.$$

Then

$$\mathbb{P}(\ |Z| \geq t\ ) \ \leq\ 2\exp\left(\frac{-t^2/2}{\nu(Z) + Lt/3}\right) \quad \textit{for all } t \geq 0\,.$$

We will apply it to bound the three sums in μ_{kk}, using the fact that each of the sums contains only centered, independent, and bounded variables, and that the variance of each entry in the sum is $p_k(1-p_k)$. For the first two sums, we can use $t \sim \sqrt{n_k p_k(1-p_k)\log n_k}$ to obtain a combined failure probability (over the entire cluster) of $\mathrm{O}(n_k^{-\alpha})$. Finally, for the third sum, we may choose $t \sim n_k\sqrt{p_k(1-p_k)\log n_k}$, again for a combined failure probability over the whole cluster of no more than $\mathrm{O}(n_k^{-\alpha})$.

We have thusly

$$\begin{aligned}\mu_{kk} &\leq |\tfrac{1}{n_k}\sum_{l\in\mathcal{C}_k} M_{lj}| + |\tfrac{1}{n_k}\sum_{l\in\mathcal{C}_k} M_{il}| + |\tfrac{1}{n_k^2}\sum_{l,l'} M_{l,l'}| + \frac{\lambda}{n_k} \\ &\lesssim \sqrt{\frac{p_k(1-p_k)}{n_k}\log n_k} + \frac{\sqrt{p_k(1-p_k)\log n_k}}{n_k} + \frac{\lambda}{n_k}\ ,\end{aligned}$$

for all $i, j \in \mathcal{C}_k$, with probability $1 - \mathrm{O}(n_k^{-\alpha})$. Note that in the inequality above, the second term is much smaller in magnitude than the first, so we can disregard it; using (C.3), we obtain

$$\mu_{kk} \lesssim \frac{1}{n_k}\left(\sqrt{n_k p_k(1-p_k)\log n_k} + \max_i \sqrt{p_i(1-p_i)n_i} + \sqrt{\max\{q(1-q)n, \log n\}}\right) \quad \text{(C.5)}$$

and by taking a union bound over k we can conclude that the probability that any of these bounds fail is $\mathrm{o}(1)$. Similarly, for $\mu_{k'k''}$, for $k' \neq k''$, we can calculate that

$$\begin{aligned}\mu_{k'k''} &\leq |\tfrac{1}{n_{k'}} \sum_{l \in \mathcal{C}_{k'}} M_{lj}| + |\tfrac{1}{n_{k''}} \sum_{l \in \mathcal{C}_{k''}} M_{il}| + |\tfrac{1}{n_{k'} n_{k''}} \sum_{l' \in \mathcal{C}_{k'},\, l'' \in \mathcal{C}_{k''}} M_{l',l''}| \\ &\lesssim \sqrt{q(1-q)(\frac{\log n_{k'}}{n_{k'}} + \frac{\log n_{k''}}{n_{k''}})} + \frac{\sqrt{q(1-q)\log(n_{k'} n_{k''})}}{\sqrt{n_{k'} n_{k''}}} \ ,\end{aligned}$$

with failure probability over all $i \in \mathcal{C}_{k'}$, $j \in \mathcal{C}_{k''}$ of no more than $\mathrm{O}(n_{k'}^{-\alpha} n_{k''}^{-\alpha})$. We do this by taking $t \sim \sqrt{n_{k'} q(1-q)\log(n_{k'} n_{k''})}$, respectively $t \sim \sqrt{n_{k''} q(1-q)\log(n_{k'} n_{k''})}$ in the first two sums. For the third, we just take $t \sim \sqrt{n_{k'} n_{k''} q(1-q)\log(n_{k'} n_{k''})}$. As before, note that the second term is much smaller in magnitude than the first, and hence we can disregard it to obtain

$$\mu_{k'k''} \lesssim \max_k \sqrt{\frac{q(1-q)\log n_k}{n_k}} = \sqrt{\frac{q(1-q)\log n_{\min}}{n_{\min}}} := \mu_{\text{off}} \ , \tag{C.6}$$

as the function $\log x / x$ is strictly increasing if $x \geq 3$, with the probability that all of the above are simultaneously true being $1 - \mathrm{o}(1)$. Since the bound on $\mu_{k'k''}$ is independent of k' and k'' we can rewrite (C.4) as

$$\begin{aligned}\Delta(Y) &\geq \sum_{k=1}^{r} (p_k - q) \|(Y^\star - Y)_{\mathcal{C}_k}\|_1 - \sum_{k=1}^{r} \mu_{kk} \|(Y^\star - Y)_{\mathcal{C}_k}\|_1 - \sum_{k' \neq k''} \mu_{k'k''} \|(Y^\star - Y)_{V_{k'} \times V_{k''}}\|_1 \\ &\geq \sum_{k=1}^{r} \left(p_k - q - \mu_{kk} - \mu_{\text{off}}\right) \|(Y^\star - Y)_{\mathcal{C}_k}\|_1\end{aligned}$$

where we use the fact that $\sum_{k' \neq k''} \|(Y^\star - Y)_{V_{k'} \times V_{k''}}\|_1 = \sum_{k=1}^{r} \|(Y^\star - Y)_{\mathcal{C}_k}\|_1$. Finally, the conditions of theorem guarantee the nonnegativity of the right hand side, hence the optimality of $Y^\star$ as the solution to the convex recovery program in (4.4). ∎

C.1.3 Proof of Theorem 49

We use a different result than Lemma 70, which we state below.

Lemma 72 (Corollary 3.12 in [12]) *Let X be an $n \times n$ symmetric matrix whose entries X_{ij} are independent symmetric random variables. Then there exists for any $0 < \epsilon \leq \frac{1}{2}$ a universal constant c_ϵ such that for every $t \geq 0$*

$$\|X\| \leq 2(1+\epsilon)\tilde{\sigma} + t\,,$$

with probability at least $1 - n\exp(\frac{-t^2}{c_\epsilon \tilde{\sigma}_\star^2})$, where

$$\tilde{\sigma} = \max_i \sqrt{\sum_j \mathbb{E}(X_{ij}^2)}\,, \quad \tilde{\sigma}_\star = \max_{i,j} \|X_{ij}\|_\infty\,.$$

We specialize Lemma 72 to HSBM to get the following result.

Lemma 73 *Let A be generated according to the heterogenous stochastic block model (HSBM). Then there exists for any $0 < \epsilon \leq \frac{1}{2}$ a universal constant c_ϵ such that*

$$\|A - \mathbb{E}(A)\| \leq 4(1+\epsilon)\max\{\sigma_{\max}, \sigma_0\} + \sqrt{2c_\epsilon \log n}$$

with probability at least $1 - n^{-1}$.

We can now present the proof for Theorem 49.

Proof. The proof follows the same lines as the proof of Theorem 48. Given the similarities between the proofs, we will only describe here the differences between the tools employed, and how they affect the conditions in Theorem 49. The proof proceeds identically as before, up to the definition of λ, which–since we use Lemma 73 rather than 70–becomes

$$\lambda := C_2 \max\{\sigma_{\max},\, \sigma_0,\, \sqrt{\log n}\}\,, \tag{C.7}$$

where C_2 was chosen as a good upper bounding constant for Lemma 73.

The other two small changes come from the fact that we will need to make sure that the failure probabilities for the quantities μ_{kk} and $\mu_{k'k''}$ are polynomial in $1/n$, which leads to the replacement of $\log n_k$ in either of them by a $\log n$. The rest of the proof proceeds exactly in the same way. ∎

C.2 Proofs for Recoverability and Non-recoverability

We use the same notation as in the main chapter and in Appendix C.1.1.

C.2.1 Proofs for Recoverability

Proof. [of Theorem 50] For $\Delta(Y) := \langle A, Y^\star - Y\rangle$, we have to show that for any feasible $Y \neq Y^\star$, we have $\Delta(Y) > 0$. For simplicity we assume $Y_{ii} = Y^\star_{ii} = 0$ for all $i \in \{1, \dots, n\}$. Consider an splitting as

$$\Delta(Y) = \langle A, Y^\star - Y\rangle = \langle \mathbb{E}(A), Y^\star - Y\rangle + \langle A - \mathbb{E}(A), Y^\star - Y\rangle . \tag{C.8}$$

Notice that $Y^\star = \sum_{k=1}^r \mathbf{1}_{\mathcal{C}_k}$ and $\mathbb{E}(A) = q\mathbf{1}\mathbf{1}^T + \sum_{k=1}^r (p_k - q)\mathbf{1}_{\mathcal{C}_k}$. Considering $d_k(Y) = \langle Y^\star_{\mathcal{C}_k}, Y^\star - Y\rangle$, the number of entries on $\mathcal{C}_k$ on which Y and $Y^\star$ do not match, we get

$$\langle \mathbb{E}(A), Y^\star - Y\rangle = \sum_{k=1}^r (p_k - q) d_k(Y) \tag{C.9}$$

where we used the fact that $Y, Y^\star \in \mathcal{Y}$ and have the same number of ones and zeros, hence $\sum_{i,j} Y_{ij} = \sum_{i,j} Y^\star_{ij}$. On the other hand, the second term in (C.8) can be represented as

$$T(Y) := \langle A - \mathbb{E}(A), Y^\star - Y\rangle \; = \sum_{Y^\star_{ij}=1, Y_{ij}=0} (A - \mathbb{E}(A))_{ij} \; + \sum_{Y^\star_{ij}=0, Y_{ij}=1} (\mathbb{E}(A) - A)_{ij}$$

where each term is a centered Bernoulli random variable bounded by 1. Observe that the total variance for all the summands in the above is given by

$$\sigma^2 = \sum_{k=1}^r d_k(Y) p_k(1 - p_k) + q(1 - q) \sum_{k=1}^r d_k(Y) .$$

Then, combining (C.8) and (C.9), and applying the Bernstein inequality yields

$$\begin{aligned}
\mathbb{P}(\Delta(Y) \leq 0) &= \mathbb{P}\bigg(T(Y) \leq -\sum_k (p_k - q) d_k(Y) \bigg) \\
&\leq \exp\bigg(-\frac{t^2}{2\sigma^2 + 2t/3} \bigg) \\
&= \exp\bigg(-\frac{\sum_k (p_k - q) d_k(Y)}{2\nu(Y) + 2/3} \bigg)
\end{aligned}$$

where $t = \sum_k (p_k - q) d_k(Y)$ and

$$\begin{aligned}
\nu(Y) &= \frac{\sigma^2}{t} \\
&= \frac{\sum_{k=1}^{r} (p_k(1-p_k) + q(1-q)) d_k(Y)}{\sum_k (p_k - q) d_k(Y)} \\
&\leq \max_k \frac{p_k(1-p_k) + q(1-q)}{p_k - q} \\
&= \frac{p_{\min}(1-p_{\min}) + q(1-q)}{p_{\min} - q} := \bar{\nu}_0 \,.
\end{aligned}$$

Considering $\bar{\nu} := 2\bar{\nu}_0 + 2/3$ and $\theta_k := \lfloor \frac{p_k - q}{p_{\min} - q} \rfloor$, we get

$$\mathbb{P}(\Delta(Y) \leq 0) \leq \exp\left(-\tfrac{1}{\bar{\nu}} \sum_k (p_k - q) d_k(Y) \right) \leq \exp\left(-\tfrac{1}{\bar{\nu}} (p_{\min} - q) \sum_k \theta_k d_k(Y) \right) \quad \text{(C.10)}$$

which can be bounded using the next lemma which is a direct extension of Lemma 4 in [47].

Lemma 74 *Given the values of θ_k and n_k, for $k = 1, \ldots, r$, and for each integer value $\xi \in [\min \theta_k(2n_k - 1),\ \sum_k \theta_k n_k^2]$, we have*

$$\left| \{[Y] \subset \mathcal{Y} : \sum_{k=1}^{r} \theta_k d_k(Y) = \xi \} \right| \leq \left(\frac{5\xi}{\tau} \right)^2 n^{16\xi/\tau} \quad \text{(C.11)}$$

where $\tau := \min_k\ \theta_k n_k$, and $[Y] = \{Y' \in \mathcal{Y} :\ Y'_{ij} Y^{\star}_{ij} = Y_{ij} Y^{\star}_{ij}\}$.

Now plugging in the result of Lemma 74 into (C.10) yields,

$$\begin{aligned}
\mathbb{P}\Big(\exists Y \in \mathcal{Y} : Y \neq Y^{\star}, \Delta(Y) \leq 0 \Big) &\leq \sum_{\xi} \mathbb{P}\big(\exists Y \in \mathcal{Y} : \sum_k \theta_k d_k(Y) = \xi \,,\ \Delta(Y) \leq 0 \big) \\
&\leq 2 \sum_{\xi} \left(\frac{5\xi}{\tau} \right)^2 n^{16\xi/\tau} \exp\left(-\tfrac{1}{\bar{\nu}} (p_{\min} - q) \xi \right) \\
&= 50 \sum_{\xi} \left(\frac{\xi}{\tau} \right)^2 \exp\left((16 \log n - \tfrac{1}{\bar{\nu}} (p_{\min} - q) \tau) \frac{\xi}{\tau} \right) \\
&\leq 50 \sum_{\xi} \left(\frac{\xi}{\tau} \right)^2 \exp\left((16 \log n - \tfrac{1}{2\bar{\nu}} \rho_{\min}) \frac{\xi}{\tau} \right) \quad \text{(C.12)}
\end{aligned}$$

In order to have a meaningful bound for the above probability, we need the exponential term in (C.12) to be decreasing. Hence, we require $\rho_{\min} \geq 64\bar{\nu}\log n$. Moreover, the function in (C.12) is a decreasing function of ξ/τ for

$$\frac{\xi}{\tau} \geq \frac{4\bar{\nu}}{\rho_{\min} - 32\bar{\nu}\log n}\,.$$

Since $\xi \geq \min \theta_k(2n_k - 1) \geq \min \theta_k n_k = \tau$, requiring the following condition (for some $\eta > 0$ which will be determined later),

$$\rho_{\min} \geq 2(16+\eta)\bar{\nu}\log n + 4\bar{\nu}\,, \tag{C.13}$$

implies

$$\frac{\xi}{\tau} \geq 1 \geq \frac{4}{4 + 2\eta\log n} \geq \frac{4\bar{\nu}}{\rho_{\min} - 32\bar{\nu}\log n}$$

and allows us to bound the summation in (C.12) with the largest term (corresponding to the smallest value of ξ/τ, or an even smaller value, namely 1) times the number of summands (which is bounded by $\sum \theta_k n_k^2$ since θ_k's are integers); i.e.,

$$\begin{aligned}
\text{(C.12)} &\leq 50\,\big(\sum \theta_k n_k^2\big)\,\exp\big(16\log n - \tfrac{1}{2\bar{\nu}}\rho_{\min}\big) \\
&\leq 50\sum \theta_k n_k^2 \exp(-2 - \eta\log n) \\
&\leq 7\,\theta_{\max} n^{2-\eta} \\
&\leq 7\,\tfrac{p_{\max}-q}{p_{\min}-q} n^{2-\eta}\,,
\end{aligned}$$

or, similarly,

$$\text{(C.12)} \leq 50\sum \theta_k n_k^2 \exp(-2-\eta\log n) \leq 7\frac{\sum_{k=1}^{r}\rho_k}{p_{\min}-q} n^{1-\eta}\,.$$

Hence, if the condition in (C.13) holds we get the optimality of $Y^\star$ with a probability at least equal to the above. Finally, $n \geq 8$ implies $\log n \geq 2$ and (C.13) follows from

$$\rho_{\min} \geq 4(17+\eta)\left(\frac{1}{3} + \frac{p_{\min}(1-p_{\min}) + q(1-q)}{p_{\min}-q}\right)\log n\,.$$

■

Proof. [of Lemma 74] We extend the proof of Lemma 4 in [47] to our case. Fix a $Y \in \mathcal{Y}$ with $\sum_{k=1}^{r} \theta_k d_k(Y) = \xi$ and consider the corresponding r clusters as well as the set of isolated nodes. Notice that for any $Y' \in [Y]$ we also have $\sum_{k=1}^{r} \theta_k d_k(Y') = \xi$. In the following, we will construct an ordering for the clusters of Y based on $Y^\star$. Denote the clusters of $Y^\star$ by $V_1^\star, \ldots, V_r^\star$, and $V_{r+1}^\star$.

Consider the set of values of cluster sizes $\{n_1, \ldots, n_r\} = \{\eta_1, \ldots, \eta_s\}$ where $\eta_1, \ldots, \eta_s$ are distinct, and define $\mathcal{I}_\ell = \{k : \ n_k = \eta_\ell\} \subset \{1, \ldots, r\}$ for $\ell = 1, \ldots, s$. For any ℓ with $|\mathcal{I}_\ell| = 1$, the cluster in $Y \in \mathcal{Y}$ of size η_ℓ can be uniquely assigned to a cluster among $V_1^\star, \ldots, V_r^\star$ of similar size. We now define an ordering for the remaining clusters. Consider a ℓ with $|\mathcal{I}_\ell| > 1$, and restrict the attention to clusters V of size η_ℓ and clusters $V_k^\star$ for $k \in \mathcal{I}$ (all clusters in $Y^\star$ of size η_ℓ). This is similar to the case in [47] where all sizes are equal: For each new cluster V of size η_ℓ, if there exists a $k \in \mathcal{I}_\ell$ such that $|V \cap V_k^\star| > \frac{1}{2}\eta_\ell$ then we label this cluster as V_k; this label is unique. The remaining unlabeled clusters are labeled arbitrarily by a number in $\mathcal{I}_\ell$.

Hence, we labeled all the clusters of Y according to the clusters of $Y^\star$. For each $(k, k') \in \{1, \ldots, r\} \times \{1, \ldots, r+1\}$, we use $\alpha_{kk'} := |V_k^\star \cap V_{k'}|$ to denote the sizes of intersections of clusters of Y and $Y^\star$. We observe that the new clusters $(V_1, \ldots, V_{r+1})$ have the following properties:

(A1) $(V_1, \ldots, V_{r+1})$ is a partition of $\{1, \ldots, n\}$ with $|V_k| = n_k$ for all $k = 1, \ldots, r$; since $Y \in \mathcal{Y}$.

(A2) For $\ell \in \{1, \ldots, s\}$ with $|\mathcal{I}_\ell| = 1$, we have $\alpha_{kk} = n_k$ for the index $k \in \mathcal{I}_\ell$.

(A3) For $\ell \in \{1, \ldots, s\}$ with $|\mathcal{I}_\ell| > 1$, consider any $k \in \mathcal{I}_\ell$. Then, exactly one of the following is true: (1) $\alpha_{kk} > \frac{1}{2}n_k$; (2) $\alpha_{kk'} \leq \frac{1}{2}n_k$ for all $k' \in \mathcal{I}_\ell$.

(A4) For $d_k(Y) = \langle Y^\star_{\mathcal{C}_k}, Y^\star - Y \rangle$, where $k = 1 \ldots, r$, we have

$$\begin{aligned} d_k(Y) &= |\{(i,j) : \ (i,j) \in \mathcal{C}^\star_k \, , \, Y_{ij} = 0\}| \\ &= |\{(i,j) : \ (i,j) \in \mathcal{C}^\star_k \, , \, i,j \in V_{r+1}\}| \\ &+ \sum_{k' \neq k''} |\{(i,j) : \ (i,j) \in \mathcal{C}^\star_k \, , \, (i,j) \in V_{k'} \times V_{k''}\}| \\ &= \alpha^2_{k(r+1)} + \sum_{k' \neq k''} \alpha_{kk'} \alpha_{kk''} \, , \end{aligned}$$

which implies

$$\xi = \sum_{k=1}^{r} \theta_k d_k(Y) = \sum_{k=1}^{r} \theta_k \alpha^2_{k(r+1)} + \sum_{k=1}^{r} \sum_{k' \neq k''} \theta_k \alpha_{kk'} \alpha_{kk''} \, .$$

Unless specified otherwise, all the summations involving k' or k'' are over the range $1, \ldots, r+1$.

We showed that the ordered partition for a $Y \in \mathcal{Y}$ with $\sum_{k=1}^{r} \theta_k d_k(Y) = \xi$ satisfies the above properties. Therefore,

$$|\{[Y] \in \mathcal{Y} : \ \sum_{k=1}^{r} \theta_k d_k(Y) = \xi\}| \leq |\{(V_1, \ldots, V_{r+1}) \text{ satisfying the above conditions}\}| \, .$$

Next, we upper bound the right hand side of the above.

Fix an ordered clustering $(V_1, \ldots, V_{r+1})$ which satisfies the above conditions. Define,

$$m_1 := \sum_{k' \neq 1} \alpha_{1k'}$$

as the number of nodes in $V^\star_1$ that are misclassified by Y ; hence $m_1 + \alpha_{11} = n_1$. Consider the following two cases:

- if $\alpha_{11} > n_1/4$ we have

$$\sum_{k' \neq k''} \alpha_{1k'} \alpha_{1k''} \geq \alpha_{11} \sum_{k'' \neq 1} \alpha_{1k''} > \tfrac{1}{4} n_1 m_1$$

- if $\alpha_{11} \leq n_1/4$ we have $m_1 \geq 3n_1/4$, which from the aforementioned properties, we must have $\alpha_{1k'} \leq n_1/2$ for all $k' = 1, \ldots, r$. Then,

$$\begin{aligned}\sum_{k' \neq k''} \alpha_{1k'}\alpha_{1k''} + \alpha_{1(r+1)}^2 &\geq \sum_{1 \neq k' \neq k'' \neq 1} \alpha_{1k'}\alpha_{1k''} + \alpha_{1(r+1)}^2 \\ &= m_1^2 - \sum_{k'=2}^{r} \alpha_{1k'}^2 \\ &\geq m_1^2 - \tfrac{1}{2} n_1 m_1 \\ &\geq \tfrac{1}{4} n_1 m_1\end{aligned}$$

Therefore,

$$d_1(Y) = \sum_{k' \neq k''} \alpha_{1k'}\alpha_{1k''} + \alpha_{1(r+1)}^2 \geq \tfrac{1}{4} n_1 m_1$$

which holds for all other indices $k \neq 1$ as well. This yields

$$\xi \geq \tfrac{1}{4} \sum_{k=1}^{r} \theta_k n_k m_k \geq \tfrac{1}{4} (\min_k \theta_k n_k) \sum_{k=1}^{r} m_k \quad \Longrightarrow \quad \bar{w} := \sum_{k=1}^{r} m_k \leq \frac{4\xi}{\min_k \theta_k n_k} := M$$

where $\bar{w}$ is the number of misclassified non-isolated nodes. Since one misclassified isolated node produces one misclassified non-isolated node, we have $w_0 \leq \bar{w} \leq M$ where w_0 is the number of misclassified isolated nodes.

- The pair of numbers $(\bar{w}, w_0)$ can take at most $(M+1)^2$ different values.

- For each such pair of numbers, there are at most $\bar{n}^{2M}$ ways to choose the identity of the misclassified nodes.

- Each misclassified non-isolated node can be assigned to one of $r - 1 \leq \bar{n}$ different clusters or be left isolated, and each misclassified isolated node can be assigned to one of $r \leq \bar{n}$ clusters.

All in all,

$$\begin{aligned}|\{[Y] \in \mathcal{Y} : \sum_{k=1}^{r} \theta_k d_k(Y) = \xi\}| &\leq (M+1)^2 \bar{n}^{4M} \\ &= \left(\frac{4\xi}{\min_k \theta_k n_k} + 1\right)^2 \exp\left(\frac{16\xi}{\min_k \theta_k n_k} \log \bar{n}\right) \\ &\leq \left(\frac{5\xi}{\min_k \theta_k n_k}\right)^2 \exp\left(\frac{16\xi}{\min_k \theta_k n_k} \log \bar{n}\right) .\end{aligned}$$

■

C.2.2 Proofs for impossibility of recovery

We prove a more comprehensive version of Theorem 51.

Theorem 75 *If any of the following conditions holds,*

(1) $2 \leq n_k \leq n/e$ *, and*

$$4 \sum_{k=1}^{r} n_k^2 \widetilde{D}(p_k, q) \leq \tfrac{1}{2} \sum_k n_k \log \tfrac{n}{n_k} - r - 2$$

(2) $2 \leq n_k \leq n/e$ *, and*

$$\tfrac{1}{2} r + \log \tfrac{1-p_{\min}}{1-p_{\max}} + 1 + \sum_k n_k^2 p_k \leq (\tfrac{1}{4} n - \sum n_k^2 p_k) \log n + \sum (n_k p_k - \tfrac{1}{4}) n_k \log n_k$$

(3) $n \geq 128$ *,* $r \geq 2$ *and*

$$\max_k \; n_k \left(\widetilde{D}(p_k, q) + \widetilde{D}(q, p_k)\right) \leq \tfrac{1}{12} \log(n - n_{\min})$$

then

$$\inf_{\widehat{Y}} \sup_{Y^\star \in \mathcal{Y}} \mathbb{P}(\widehat{Y} \neq Y^\star) \geq \frac{1}{2}$$

where the infimum is taken over all measurable estimators $\widehat{Y}$ *based on the realization* A *generated according to the heterogenous stochastic block model.*

Proof. [of cases 1 and 2 of Theorem 75] Let $\mathbb{P}_{(Y^\star,A)}$ be the joint distribution of $Y^\star$ and A, where $Y^\star$ is sampled uniformly from $\mathcal{Y}$ and A is generated according to the heterogenous stochastic block model conditioning on $Y^\star$. Note that

$$\inf_{\widehat{Y}} \sup_{Y^\star \in \mathcal{Y}} \mathbb{P}(\widehat{Y} \neq Y^\star) \geq \inf_{\widehat{Y}} \mathbb{P}_{(Y^\star,A)}[\widehat{Y} \neq Y^\star]\,.$$

By Fano's inequality we have,

$$\mathbb{P}_{(Y^\star,A)}(\widehat{Y} \neq Y^\star) \geq 1 - \frac{I(Y^\star; A) + 1}{\log|\mathcal{Y}|}, \tag{C.14}$$

where $I(X;Z)$ is the mutual information, and $H(X)$ is the Shannon entropy for X. By counting argument we find that $|\mathcal{Y}| = \binom{n}{\bar{n}} \frac{\bar{n}!}{n_1!\ldots n_r!}$. Using $\sqrt{n}(n/e)^n \leq n! \leq e\sqrt{n}(n/e)^n$ and $\binom{n}{\bar{n}} \geq (n/\bar{n})^{\bar{n}}$, it follows that

$$|\mathcal{Y}| \geq \frac{n^{\bar{n}}\sqrt{\bar{n}}}{e^r\sqrt{n_1\ldots n_r}n_1^{n_1}\ldots n_r^{n_r}}$$

which gives

$$\log|\mathcal{Y}| \geq \sum_{i=1}^r n_i\big(\log\frac{n}{n_i} - \frac{\log n_i}{2n_i}\big) - r \geq \frac{1}{2}\sum_{i=1}^r n_i \log\frac{n}{n_i} - r\,.$$

On the other hand, note that $H(A) \leq \binom{n}{2} H(A_{12})$ by chain rule, the fact that $H(X|Y) \leq H(X)$, and the symmetry among identically distributed A_{ij}'s. Furthermore A_{ij}'s are conditionally independent and hence $H(A|Y^\star) = \binom{n}{2} H(A_{12}|Y^\star_{12})$. Now it follows that

$$I(Y^\star; A) = H(A) - H(A|Y^\star) \leq \binom{n}{2} I(Y^\star_{12}; A_{12}).$$

Observe that

$$\mathbb{P}(Y^\star_{12} = 1, (1,2) \in \mathcal{C}_i) = \frac{\binom{n-2}{n_i-2}\binom{n-n_i}{n_1,\ldots,n_{i-1},n_{i+1},\ldots,n_r,n_0}}{|\mathcal{Y}|} = \frac{n_i(n_i-1)}{n(n-1)} := \alpha_i\,.$$

Using the properties of KL-divergence, we have $\mathbb{P}(A_{12}=1) = \sum_{i=1}^r \alpha_i p_i + (1-\sum_i \alpha_i)q := \beta\,.$ Therefore,

$$I(Y^\star_{12}, A_{12}) = \sum_{i=1}^r \alpha_i D_{\mathrm{KL}}(p_i, \beta) + (1 - \sum_i \alpha_i) D_{\mathrm{KL}}(q, \beta) \tag{C.15}$$

$$= H(\beta) - \sum \alpha_i H(p_i) - (1 - \sum \alpha_i) H(q) \tag{C.16}$$

Since $I(Y^\star; A) \leq \binom{n}{2} I(Y_{12}^\star; A_{12})$, plugging in the following condition in Fano's inequality (C.14),

$$\tfrac{1}{2}\sum_i n_i \log \frac{n}{n_i} - r \geq 2 + 2\binom{n}{2} I(Y_{12}^\star; A_{12})\,, \tag{C.17}$$

guarantees $\mathbb{P}_{(Y^\star,A)}(\widehat{Y} \neq Y^\star) \geq \frac{1}{2}$. In the following, we bound $I(Y_{12}^\star; A_{12})$ in two different ways to derive conditions 1 and 2 of Theorem 75. Throughout the proof we use the following inequality from [47] for the Kullback-Leibler divergence of Bernoulli variables,

$$D_{\mathrm{KL}}(p,q) := D_{\mathrm{KL}}(\mathrm{Ber}(p), \mathrm{Ber}(q)) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \leq \frac{(p-q)^2}{q(1-q)}\,, \tag{C.18}$$

where the inequality is established by $\log x \leq x - 1$, for any $x \geq 0$.

- From (C.15), we have

$$I(Y_{12}^\star, A_{12}) \leq \sum_{i=1}^{r} \frac{4\alpha_i (p_i - q)^2}{q(1-q)} \leq \frac{4\sum_{i=1}^r n_i^2 (p_i-q)^2}{n(n-1)q(1-q)}$$

  where we assumed $\sum n_i^2 \leq \frac{1}{2}n^2$. Now, the right hand side of C.17 can be bounded as

$$2\binom{n}{2} I(Y_{12}^\star; A_{12}) \leq \frac{4\sum_{i=1}^r n_i^2 (p_i-q)^2}{q(1-q)} = 4\sum_{i=1}^r n_i^2 \widetilde{D}(p_i, q)$$

  and gives the sufficient condition 1 of Theorem 75.

- Again from (C.15), we have

$$\begin{aligned} I(Y_{12}^\star; A_{12}) &= \sum_i \alpha_i \left(p_i \log\frac{p_i}{\beta} + (1-p_i)\log\frac{1-p_i}{1-\beta}\right) + \big(1 - \sum_i \alpha_i\big) D_{\mathrm{KL}}(q,\beta) \\ &\leq \sum \alpha_i p_i \log\frac{1}{\alpha_i} + \log c + \big(1 - \sum_i \alpha_i\big)\frac{(q-\beta)^2}{\beta(1-\beta)} \end{aligned}$$

  where the first term is bounded via $\beta \geq \sum_i \alpha_i p_i \geq \alpha_i p_i$, the second term is bounded via $\beta \leq p_{\max}$ and $c = (1-p_{\min})/(1-p_{\max})$, and we used (C.18) for the last term. Since $1 - \beta = 1 - q - \sum_i \alpha_i (p_i - q) \geq (1 - \sum_i \alpha_i)(1-q)$, the last term can be bounded

as

$$\begin{aligned}\big(1-\sum_i \alpha_i\big)\frac{(q-\beta)^2}{\beta(1-\beta)} &\le \big(1-\sum_i \alpha_i\big)\frac{\big(\sum_i \alpha_i(p_i-q)\big)^2}{\big(\sum_i \alpha_i p_i\big)\big(1-\sum_i \alpha_i\big)(1-q)}\\ &\le \sum_i \alpha_i(p_i-q)\\ &\le \sum_i \alpha_i p_i\,.\end{aligned}$$

This implies

$$I(Y_{12}^\star; A_{12}) \le \sum_i \alpha_i p_i \log\tfrac{1}{\alpha_i} + \sum_i \alpha_i p_i + \log c \le \sum_i \alpha_i p_i \log\tfrac{e}{\alpha_i} + \log c\,.$$

Since $n_i \ge 2$, $\alpha_i = \frac{n_i(n_i-1)}{n(n-1)} \ge \frac{n_i^2}{en^2}$. Hence

$$\begin{aligned}2\binom{n}{2} I(Y_{12}^\star; A_{12}) &\le n(n-1)\sum_i \frac{n_i(n_i-1)}{n(n-1)} p_i \log\frac{e^2n^2}{n_i^2} + 2\log c\\ &\le 2\sum_i n_i^2 p_i \log\frac{en}{n_i} + 2\log c\end{aligned}$$

which gives the sufficient condition 2 of Theorem 75.

■

Proof. [of case 3 in Theorem 75] Without loss of generality assume $n_1 \le n_2 \le \ldots \le n_r$. Let $M := \bar{n} - n_{\min} = \bar{n} - n_1$, and $\bar{\mathcal{Y}} := \{Y_0, Y_1, \ldots, Y_M\}$. Y_0 is the clustering matrix with clusters $\{\mathcal{C}_\ell\}_{\ell=1}^r$ that correspond to $V_1 = \{1, \ldots, n_1\}$, $V_\ell = \{\sum_{i=1}^{\ell-1} n_i + 1, \ldots, \sum_{i=1}^{\ell} n_i\}$ for $\ell = 2, \ldots, r$. Other members of $\bar{\mathcal{Y}}$ are given by swapping an element of $\cup_{\ell=2}^r V_\ell$ with an element of V_1. Let $\mathbb{P}_i$ be the distributional law of the graph A conditioned on $Y^\star = Y_i$.

Since $\mathbb{P}_i$ is product of $\frac{1}{2}n(n-1)$ Bernoulli random variables, we have

$$\begin{aligned}
I(Y^\star; A) &= \mathbb{E}_Y \left(D_{\mathrm{KL}} \left(\mathbb{P}(A|Y), \mathbb{P}(A) \right) \right) \\
&= \tfrac{1}{M+1} \sum_{i=0}^{M} D_{\mathrm{KL}} \big(\mathbb{P}_i, \tfrac{1}{M+1} \sum_{j=0}^{M} \mathbb{P}_j \big) \\
&\leq \tfrac{1}{(M+1)^2} \sum_{i,j=0}^{M} D_{\mathrm{KL}} (\mathbb{P}_i, \mathbb{P}_j) \\
&\leq \max_{i,j=0,\ldots,M} \; D_{\mathrm{KL}} (\mathbb{P}_i, \mathbb{P}_j) \\
&\leq \max_{i_1,i_2,i_3=1,\ldots,r} \sum_{j=1}^{3} \left(\frac{n_{i_j}(p_{i_j}-q)^2}{q(1-q)} + \frac{n_{i_j}(p_{i_j}-q)^2}{p_{i_j}(1-p_{i_j})} \right) \\
&\leq 3 \max_{i=1,\ldots,r} \left(\frac{n_i(p_i-q)^2}{q(1-q)} + \frac{n_i(p_i-q)^2}{p_i(1-p_i)} \right)
\end{aligned}$$

where the third line follows from the convexity of KL-divergence, and the line before the last follows from the construction of $\bar{\mathcal{Y}}$ and (C.18). Now if the condition of the theorem holds, then $I(Y^\star; A) \leq \frac{1}{4} \log(n - n_{\min}) = \frac{1}{4} \log |\bar{\mathcal{Y}}|$. Note that for $n \geq 128$ we get $\log |\bar{\mathcal{Y}}| = \log(n - n_{\min}) \geq \log(n/2) \geq 4$. The conclusion follows by Fano's inequality in (C.14) restricting the supremum to be taken over $\bar{\mathcal{Y}}$. ∎

C.3 Detailed Computations for Examples in Section 4.3

In the following, we present the detailed computations for the examples in Section 4.3 and summarized in Table 4.1. When there is no impact on the final result, quantities are approximated as denoted by $\approx$.

First, we repeat the conditions of Theorems 48 and 49. The conditions of Theorem 48 can be equivalently stated as

- $\rho_k^2 \gtrsim n_k p_k (1-p_k) \log n_k = \sigma_k^2 \log n_k$
- $(p_{\min} - q)^2 \gtrsim q(1-q) \frac{\log n_{\min}}{n_{\min}}$
- $\rho_{\min}^2 \gtrsim \max \{ \log n, nq(1-q), \max_k \; n_k p_k (1-p_k) \}$

- $\sum_{k=1}^{r} n_k^{-\alpha} = \mathrm{o}(1)$ for some $\alpha > 0$.

Notice that $n_k p_k(1-p_k) \gtrsim \log n_k$, for $k = 1, \ldots, r$, is implied by the first condition, as mentioned in Remark 67. The conditions of Theorem 49 can be equivalently stated as

- $\rho_k^2 \gtrsim n_k p_k(1-p_k) \log n$
- $(p_{\min} - q)^2 \gtrsim q(1-q)\frac{\log n}{n_{\min}}$
- $\rho_{\min}^2 \gtrsim \max\big\{nq(1-q), \max_k\, n_k p_k(1-p_k)\big\}$.

Remark 76 *Provided that both p_k and q/p_k are bounded away from 1, we have*

$$\widetilde{D}(q, p_k) = p_k \frac{(1-q/p_k)^2}{1-p_k} \approx p_k \quad , \quad \frac{\rho_k^2}{\sigma_k^2} = \frac{(1-q/p_k)^2}{1-p_k}\, n_k p_k \approx n_k p_k \,.$$

This simplifies the first condition of Theorem 48 to a simple connectivity requirement. Hence, we can rewrite the conditions of Theorems 48, 49 as

- *Theorem 48:* $n_k p_k \gtrsim \log n_k$, $\widetilde{D}(p_{\min}, q) \gtrsim \frac{\log n_{\min}}{n_{\min}}$, $\rho_{\min}^2 \gtrsim \max\{\sigma_{\max}^2, nq(1-q), \log n\}$, $\sum_{k=1}^{r} n_k^{-\alpha} = \mathrm{o}(1)$ *for some* $\alpha > 0$.
- *Theorem 49:* $n_k p_k \gtrsim \log n$, $\widetilde{D}(p_{\min}, q) \gtrsim \frac{\log n}{n_{\min}}$, $\rho_{\min}^2 \gtrsim \max\{\sigma_{\max}^2, nq(1-q)\}$.

Example 1: In a configuration with two communities $(n-\sqrt{n},\, n^{-2/3},\, 1)$ and $(\sqrt{n},\, \frac{1}{\log n},\, 1)$ with $q = n^{-2/3-0.01}$, we have $n_{\min} = \sqrt{n}$ and $p_{\min} = n^{-2/3}$. We have,

$$\widetilde{D}(p_{\min}, q) \approx n^{-2/3+0.01}$$

which does not exceed either $\frac{\log n_{\min}}{n_{\min}} \approx \frac{\log n}{\sqrt{n}}$ or $\frac{\log n}{n_{\min}} \approx \frac{\log n}{\sqrt{n}}$, and we get no recovery guarantee from Theorems 48 and 49 respectively. However, as $p_{\min} - q$ is not much smaller than q, while $\rho_{\min} \approx n^{1/3}$ grows much faster than $\log n$, the condition of Theorem 50 trivially holds.

Here are the related quantities for this configuration:

$$\rho_1 = n_1(p_1 - q) = (n - \sqrt{n})(n^{-2/3} - n^{-2/3-0.01}) \approx n^{1/3}$$
$$\rho_2 = n_2(p_2 - q) = \sqrt{n}(\tfrac{1}{\log n} - n^{-2/3-0.01}) \approx \tfrac{\sqrt{n}}{\log n}$$

which gives $\rho_{\min} \approx n^{1/3}$. Furthermore,

$$\sigma_1^2 = n_1 p_1(1-p_1) \approx n^{1/3} \quad , \quad \sigma_2^2 = n_2 p_2(1-p_2) = \tfrac{\sqrt{n}}{\log n} \,,$$

which gives $\sigma_{\max} = \frac{\sqrt{n}}{\log n}$. On the other hand $nq(1-q) \approx n^{1/3-0.01}$ which is smaller than $\sigma_{\max}^2$.

Example 2: Consider a configurations with $(n - n^{2/3},\, n^{-1/3+\epsilon},\, 1)$ and $(\sqrt{n},\, \frac{c}{\log n},\, n^{1/6})$ and $q = n^{-2/3+3\epsilon}$. Since all p_k's and q/p_k's are much less than 1, the first condition of both Theorems 48 and 49 can be verified by Remark 76. Moreover, $n_{\min} = \sqrt{n}$ and $p_{\min} = n^{-1/3+\epsilon}$ which gives

$$\widetilde{D}(p_{\min}, q) = n^{-\epsilon}$$

and verifies $\widetilde{D}(p_{\min}, q) \gtrsim \frac{\log n_{\min}}{n_{\min}}$ for 48, as well as $\widetilde{D}(p_{\min}, q) \gtrsim \frac{\log n}{n_{\min}}$ for 49. Moreover, $\rho_1 \approx n^{2/3+\epsilon}$ and $\rho_2 \approx \frac{\sqrt{n}}{\log n}$ which gives $\rho_{\min} \approx \frac{\sqrt{n}}{\log n} \gtrsim \sqrt{\log n}$. On the other hand, $\sigma_1^2 \approx n^{2/3+\epsilon}$ and $\sigma_2^2 \approx \sqrt{n}/\log n$ which gives

$$\max\{\sigma_{\max}^2 \,,\, nq(1-q)\} \approx n^{2/3+\epsilon} \,.$$

Thus all conditions of Theorems 48 and 49 are satisfied. Moreover, as $p_{\min} - q$ is not much smaller than q, while $\rho_{\min} \approx \frac{\sqrt{n}}{\log n}$ is growing much faster than $\log n$, the condition of Theorem 50 trivially holds.

Example 3: Consider a configurations with $(\sqrt{\log n},\, \mathrm{O}(1),\, m)$ and $(n_2,\, \mathrm{O}(\frac{\log n}{\sqrt{n}}),\, \sqrt{n})$ and $q = \mathrm{O}(\log n/n)$, where $n_2 = \sqrt{n} - m\sqrt{\log n/n}$. Here, we assume $m \le n/(2\sqrt{\log n})$ which implies $n_2 \ge \sqrt{n}/2$. Since all p_k's and q/p_k's are much less than 1, we can use Remark 76: the first condition of Theorem 48 holds as $n_1 p_1 \approx \sqrt{\log n} \gtrsim \log n_1 \approx \log\log n$ and $n_2 p_2 \approx \log n \gtrsim \log n_2$. However, $n_1 p_1 \approx \sqrt{\log n} \not\gtrsim \log n$ and Theorem 49 does not offer a guarantee for this configuration.

Moreover, $n_{\min} = \sqrt{\log n}$ and $p_{\min} = \mathrm{O}(\frac{\log n}{\sqrt{n}})$ which gives

$$\widetilde{D}(p_{\min}, q) = \log n$$

and verifies $\widetilde{D}(p_{\min}, q) \gtrsim \frac{\log n_{\min}}{n_{\min}} \approx \frac{\log\log n}{\sqrt{\log n}}$ for 48, as well as $\widetilde{D}(p_{\min}, q) \gtrsim \frac{\log n}{n_{\min}} = \sqrt{\log n}$ for 49. Moreover, $\sigma_1^2 = \sqrt{\log n}$ (also ρ_1) and $\sigma_2^2 = \log n$ (also ρ_2) which gives

$$\max\{\sigma_{\max}^2\,,\, nq(1-q)\} \approx \log n$$

and $\rho_{\min}^2 \approx \log n$. For the last condition of Theorem 48 we need

$$m(\log n)^{-\alpha/2} + \sqrt{n}(\sqrt{n} - k\sqrt{\tfrac{\log n}{n}})^{-\alpha} = \mathrm{o}(1)$$

for some $\alpha > 0$ which can be guaranteed provided that m grows at most polylogarithmically in n. All in all, we verified the conditions of Theorem 48 while the first condition of 49 fails. Observe that $\rho_{\min}$ fails the condition of Theorem 50.

Alternatively, consider a configuration with $(\sqrt{\log n},\, \mathrm{O}(1),\, m)$ and $(\sqrt{n},\, \mathrm{O}(\frac{\log n}{\sqrt{n}}),\, m')$ and $q = \mathrm{O}(\frac{\log n}{n})$, where $m' = \sqrt{n} - m\sqrt{\log n/n}$ to ensure a total of n vertices. Here, we assume $m \le n/(2\sqrt{\log n})$ which implies $m' \ge \sqrt{n}/2$. Similarly, all conditions of Theorem 48 can be verified provided that m grows at most polylogarithmically in n. Moreover, the conditions of Theorems 49 and 50 fail to satisfy.

Example 4: Consider a configuration with $(\frac{1}{2}n^{\epsilon},\, \mathrm{O}(1),\, n^{1-\epsilon})$ and $(\frac{1}{2}n,\, n^{-\alpha}\log n,\, 1)$ and $q = n^{-\beta}\log n$, where $0 < \alpha < \beta < 1$ and $0 < \varepsilon < 1$.

We have $\rho_1 \approx n^{\varepsilon}$ and $\rho_2 \approx n^{1-\alpha}\log n$. Since $\rho_{\min}^2 \gtrsim \log n$, the last condition of Theorem 48 holds, and $\log n_{\min} \approx \log n$, we need to check for similar conditions to be able to use Theorems 48 and 49. Using Remark 76, the first condition of both Theorems holds because of $n_1 p_1 \approx n^{\varepsilon} \gtrsim \log n$ and $n_2 p_2 \approx n^{1-\alpha}\log n \gtrsim \log n$. Moreover, the condition

$$\widetilde{D}(p_{\min}, q) \approx n^{\beta-2\alpha}\log n \gtrsim \tfrac{\log n}{n_{\min}} \approx \tfrac{\log n}{n^{\varepsilon}}$$

is equivalent to $\beta + \varepsilon > 2\alpha$. Furthermore, $\sigma_1^2 = n^{\varepsilon}$ and $\sigma_2^2 = n^{1-\alpha}\log n$, and for the last condition we need

$$\min\{n^{2\varepsilon}\,,\, n^{2-2\alpha}\log^2 n\} \gtrsim \max\{n^{\varepsilon}\,,\, n^{1-\alpha}\log n\,,\, n^{1-\beta}\log n\}$$

which is equivalent to $2\varepsilon + \alpha > 1$ and $\varepsilon + 2\alpha < 2$. Notice that $\beta + 1 > 2\alpha$ is automatically satisfied when we have $\beta + \varepsilon > 2\alpha$ from the previous part.

Example 5: Consider a configuration with

$$(\log n,\, \mathrm{O}(1),\, \tfrac{n}{\log n} - m\sqrt{\tfrac{n}{\log n}}\,)$$

and $(\sqrt{n \log n},\, \mathrm{O}(\sqrt{\tfrac{\log n}{n}}),\, m)$ and $q = \mathrm{O}(\tfrac{\log n}{n})$. All of ρ_1, ρ_2, σ_1^2, σ_2^2, and $nq(1-q)$, are approximately equal to $\log n$. Thus, the first and third conditions of Theorems 48 and 49 are satisfied. Moreover,

$$\widetilde{D}(p_{\min}, q) \approx 1 \gtrsim \frac{\log n_{\min}}{n_{\min}} \approx \frac{\log \log n}{\log n}$$

which establishes the conditions of Theorem 49. On the other hand, the last condition of Theorem 48 is not satisfied as one cannot find a constant value $\alpha > 0$ for which

$$\sum_{k=1}^{r} n_k^{\alpha} = \left(\tfrac{n}{\log n} - m\sqrt{\tfrac{n}{\log n}}\right) \log^{-\alpha} n + m(n \log n)^{-\alpha/2}$$

is $\mathrm{o}(1)$ while n grows.

Example 6: For the first configuration, Theorem 48 requires

$$f^2(n) \gtrsim \max\{\tfrac{\log n_1}{n_1}\,,\, \tfrac{\log n_{\min}}{n_{\min}}\,,\, \tfrac{n}{n_1^2}\}$$

while Theorem 49 requires

$$f^2(n) \gtrsim \max\{\tfrac{\log n_1}{n_1}\,,\, \tfrac{\log n}{n_{\min}}\,,\, \tfrac{n}{n_1^2}\}$$

and both require $n_{\min} \gtrsim \sqrt{n}$. Therefore, both set of requirements can be written as

$$f^2(n) \gtrsim \max\{\tfrac{\log n}{n_{\min}}\,,\, \tfrac{n}{n_1^2}\} \quad , \quad n_{\min} \gtrsim \sqrt{n}\,.$$

C.4 Recovery by a Simple Counting Algorithm

In Section 4.2.1, we considered a tractable approach for exact recovery of (partially) observed models generated according to the heterogenous stochastic block model. However, in the interest of computational effort, one can further characterize a subset of models that are

recoverable via a much simpler method than the convex program. The following algorithm is a proposal to do so. Moreover, the next theorem provides a characterization for models for which this simple thresholding algorithm is effective for exact recovery. Here, we allow for isolated nodes as described in Section 4.2.

Algorithm 1 Simple Thresholding Algorithm

1: (Find isolated notes) For each node v, compute its degree d_v. Declare i as isolated if

$$d_v < \min_k \frac{(n_k - 1)(p_k - q)}{2} + (n-1)q.$$

2: (Find all communities) For every pair of nodes (v, u), compute the number of common neighbors $S_{vu} := \sum_{w \neq v,u} A_{vw} A_{uw}$. Declare v, u as in the same community if

$$S_{vu} > nq^2 + \frac{1}{2}\left(\min_k \left((n_k - 2)p_k^2 - n_k q^2 \right) + q \cdot \max_{i \neq j} \left(\rho_k - p_k + \rho_l - p_l \right) \right)$$

where $\rho_k = n_k(p_k - q)$.

Theorem 77 *Under the stochastic block model, with probability at least $1 - 2n^{-1}$, the simple counting algorithm 1 find the isolated nodes provided*

$$\min_k (n_k - 1)^2 (p_k - q)^2 \geq 19(1-q)\left(\max_k n_k p_k + nq \right) \log n \,. \tag{C.19}$$

Furthermore the algorithm finds the cluster if

$$\left[\min_k \left\{ (n_k - 2)p_k^2 + (n - n_k)q^2 \right\} - q \max_{k \neq l} \left\{ (n_k - 1)p_k + (n_l - 1)p_l + (n - n_k - n_l)q \right\} \right]^2$$
$$\geq 26(1 - q^2)\left(\max_k n_k p_k^2 + nq^2 \right) \log n \,, \tag{C.20}$$

while the term inside the bracket (which is squared) is assumed to be non-negative.

We remark that the following is a slightly more restrictive condition than (C.20)

$$\left[\min_k n_k (p_k^2 - q^2) - 2q\rho_{\max} \right]^2 \geq 26(1 - q^2)\left[nq^2 + \max_k n_k p_k^2 \right] \log n \,.$$

with better interpretability.

Proof. [of Theorem 77] For node v, let d_v denote its degree. Let $\bar{V} = \cup_{i=1}^{r} V_i$ denote the set of nodes which belong to one of the clusters, and V_0 be isolated nodes. If $v \in V_i$ for some $i = 1, \ldots, r$, then d_v is distributed as a sum of independent binomial random variables $\mathrm{Bin}(n_i - 1, p_i)$ and $\mathrm{Bin}(n - n_i, q)$. If $v \in V_0$, then d_v is distributed as $\mathrm{Bin}(n-1, q)$. Hence we have,

$$\mathbb{E}(d_v) = \begin{cases} (n_i - 1)p_i + (n - n_i)q & v \in V_i \subset \bar{V} \\ (n-1)q & v \in V_0\,, \end{cases}$$

and

$$\mathrm{Var}(d_v) = \begin{cases} (n_i - 1)p_i(1-p_i) + (n - n_i)q(1-q) & v \in V_i \subset \bar{V} \\ (n-1)q(1-q) & v \in V_0\,. \end{cases}$$

Let $\kappa_0^2 := \max_i n_i p_i(1-q) + nq(1-q)$, and $t = \min_i \frac{(n_i-1)(p_i-q)}{2} \le \frac{\kappa_0^2}{2}$. Then $\mathrm{Var}(d_v) \le \kappa_0^2$ for any $v \in V_0 \cup \bar{V}$. By Bernstein's inequality we get

$$\mathbb{P}\big(|d_v - \mathbb{E}(d_v)| > t\big) \le 2\exp\bigg(-\frac{t^2}{2\kappa_0^2 + 2t/3}\bigg) \le 2\exp\bigg(-\frac{3\min_i(n_i-1)^2(p_i-q)^2}{28\kappa_0^2}\bigg) \le 2n^{-2},$$

where the last inequality follows from the condition (C.19). Now by union bound over all nodes, with probability at least $1 - 2n^{-1}$, for node $v \in V_i \subset \bar{V}$ we have,

$$d_v \ge (n_i - 1)p_i + (n - n_i)q - t > \min_i \frac{(n_i-1)(p_i-q)}{2} + (n-1)q\,,$$

and for node $v \in V_0$,

$$d_v \le (n-1)q(1-q) + t < \min_i \frac{(n_i-1)(p_i-q)}{2} + (n-1)q\,.$$

This proves the first statement of the theorem, and all the isolated nodes are correctly identified. For the second statement, let S_{vu} denote the common neighbor for nodes $v, u \in \bar{V}$. Then

$$S_{vu} \sim_d \begin{cases} \mathrm{Bin}(n_i - 2, p_i^2) + \mathrm{Bin}(n - n_i, q^2) & (v,u) \in V_i \times V_i \\ \mathrm{Bin}(n_i - 1, p_i q) + \mathrm{Bin}(n_j - 1, p_j q) + \mathrm{Bin}(n - n_i - n_j, q^2) & (v,u) \in V_i \times V_j\,,\ i \ne j \end{cases}$$

where $\sim_d$ denotes equality in distribution and $+$ denotes the summation of independent random variables. Hence

$$\mathbb{E}(S_{vu}) = \begin{cases} (n_i - 2)p_i^2 + (n - n_i)q^2 & (v,u) \in V_i \times V_i \\ (n_i - 1)p_i q + (n_j - 1)p_j q + (n - n_i - n_j)q^2 & (v,u) \in V_i \times V_j\,,\ i \neq j \end{cases}$$

and

$$\mathrm{Var}(S_{vu}) = \begin{cases} (n_i - 2)p_i^2(1 - p_i^2) + (n - n_i)q^2(1 - q^2) & (v,u) \in V_i \times V_i \\ (n_i - 1)p_i q(1 - p_i q) + (n_j - 1)p_j q(1 - p_j q) & \\ \qquad +(n - n_i - n_j)q^2(1 - q^2) & (v,u) \in V_i \times V_j\,,\ i \neq j \end{cases}$$

Let

$$\begin{aligned} \Delta &= \min_i \big((n_i - 2)p_i^2 + (n - n_i)q^2\big) - \max_j \big(2(n_j - 1)p_j q + (n - 2n_j)q^2\big) \\ &= \min_i \big((n_i - 2)p_i^2 - n_i q^2\big) - \max_j \big(2(n_j - 1)p_j q - 2n_j q^2\big)\,, \end{aligned}$$

Let $\kappa_1^2 := 2\max_i n_i p_i^2(1-q^2) + nq^2(1-q^2)$. Then $\mathrm{Var}(S_{vu}) \leq \kappa_1^2$ for all v, u. Then $\Delta \leq \kappa_1^2/2$. Bernstein's inequality with $t = \Delta/2$ yields

$$\mathbb{P}\big(|S_{vu} - \mathbb{E}(S_{vu})| > t\big) \leq 2\exp\left(-\frac{t^2}{2\kappa_1^2 + 2t/3}\right) \leq 2\exp\left(-\frac{3\Delta^2}{26\kappa_1^2}\right) \leq 2n^{-3},$$

where the last line follows from assumption (C.20). By union bound over all pair of nodes (v, u), we get with probability at least $1 - 2n^{-1}$, $S_{vu} > \Gamma$ for all v, u in the same cluster and $S_{vu} < \Gamma$ otherwise. Here

$$\Gamma := \frac{1}{2}\left(\min_i \big((n_i - 2)p_i^2 + (n - n_i)q^2\big) + \max_{i \neq j} \big((n_i - 1)p_i q + (n_j - 1)p_j q + (n - n_i - n_j)q^2\big)\right).$$

■

VITA

Amin Jalali received B.Sc. degrees in electrical engineering and in pure mathematics both from Sharif University of Technology, Tehran, Iran, in 2010. He is currently a Ph.D. candidate in the Department of Electrical Engineering at University of Washington, Seattle, WA, USA. He has been awarded the University of Washington Graduate School Top Scholar Fellowship, Paul C. Leach Fellowship, and the Yang Research Award for Outstanding Doctoral Student in the Department of Electrical Engineering at University of Washington.